Published online 24 March 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 12* ***e68***
https://doi.org/10.1093/nar/gkac190

Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information

James M. Heather [1,2,*], Matthew J. Spindler[3], Marta Herrero Alonso [1], Yifang Ivana Shui[1], David G. Millar [1,2], David S. Johnson[3], Mark Cobbold[1,2] and Aaron N. Hata [1,2,*]

[1]Massachusetts General Hospital Cancer Center, Charlestown, MA, USA, [2]Department of Medicine, Harvard Medical School, Boston, MA, USA and [3]GigaMune, Inc., South San Francisco, CA, USA

Received December 20, 2021; Revised February 18, 2022; Editorial Decision March 05, 2022; Accepted March 09, 2022

ABSTRACT

The study and manipulation of T cell receptors (TCRs) is central to multiple fields across basic and translational immunology research. Produced by V(D)J recombination, TCRs are often only recorded in the literature and data repositories as a combination of their V and J gene symbols, plus their hypervariable CDR3 amino acid sequence. However, numerous applications require full-length coding nucleotide sequences. Here we present Stitchr, a software tool developed to specifically address this limitation. Given minimal V/J/CDR3 information, Stitchr produces complete coding sequences representing a fully spliced TCR cDNA. Due to its modular design, Stitchr can be used for TCR engineering using either published germline or novel/modified variable and constant region sequences. Sequences produced by Stitchr were validated by synthesizing and transducing TCR sequences into Jurkat cells, recapitulating the expected antigen specificity of the parental TCR. Using a companion script, Thimble, we demonstrate that Stitchr can process a million TCRs in under ten minutes using a standard desktop personal computer. By systematizing the production and modification of TCR sequences, we propose that Stitchr will increase the speed, repeatability, and reproducibility of TCR research. Stitchr is available on GitHub.

INTRODUCTION

Alongside immunoglobulins, T cell receptors (TCRs) underly adaptive immunity in jawed vertebrates. They are the basis through which T cells initiate their major functions – detecting and responding to pathogens, cancers, and other threats – via recognition of peptides and other molecules displayed by MHC proteins. If congenitally absent, either from a lack of production of the proteins themselves or of the T cells that bear them, untreated individuals face high mortality risk during early infancy due to uncontrolled microbial infections (1,2). Conversely, inappropriate recognition of molecules by TCRs can lead to autoimmunity or allergies (3). A lack of sufficient TCR-based responses to neoantigens and other tumor antigens contributes to the development and progression of malignancies, as illustrated by the clinical success of checkpoint blockade therapies in recent years (4,5). It is hard to overstate the importance of T cell receptors within and beyond the field of immunology.

TCRs are produced through a process of somatic DNA recombination of multiple gene segments arrayed at the different TCR loci, named 'V(D)J recombination' after the variable (V), diversity (D), and joining (J) genes which constitute the different varieties of genes. During recombination, the intervening DNA between a single segment of each of the different types being recombined is imprecisely excised, and the once-separate coding portions are joined together (6). There are two conserved types of heterodimeric TCRs: alpha/beta (TRA/TRB) and gamma/delta (TRG/TRD) TCRs, with alpha/beta predominating in human T cells. Distinct TCR loci contain different types of genes: alpha and gamma chain TCRs consist of just V and J genes, while beta and delta chains are composed of V, D, and J genes, recombined together.

This process is capable of producing an incredible diversity of TCRs. Combinatorial diversity is generated both by merit of TCRs being heterodimers of two polypeptides, with each polypeptide chain produced through this process, and by there being multiple V(D)J genes to be selected from at each TCR locus. Even greater diversity is introduced by the non-templated deletion and addition of nucleotides at the rearranged junctions. Ultimately, this process produces the region of the TCR gene that encodes the complementarity determining region 3 (CDR3) (containing the entirety of the short D gene for beta and delta chains), the hypervariable section of the

[*]To whom correspondence should be addressed. Tel: +1 617 724 0104; Email: jheather@mgh.harvard.edu
Correspondence may also be addressed to Aaron N. Hata. Tel: +1 617 724 3442; Email: ahata@mgh.harvard.edu

TCR which contacts the antigen. Cumulatively, this system is estimated to be capable of producing $\sim 1 \times 10^{15}$ unique alpha/beta TCRs in humans (7,8), orders of magnitude greater than the number of T cells that an individual body could contain (9). Even though this 'TCR space' is not evenly utilized, there is still tremendous inter- and intra-individual receptor diversity, providing a substantial barrier to study. Investigation of the functional behavior of TCRs is additionally complicated by the diversity of their ligands. Alpha/beta TCRs bind to short peptide fragments (derived from potentially any protein that finds its way into the body) presented in the groove of other cells' MHC proteins, which are among the most polymorphic genes in vertebrate genomes (10), while gamma/delta TCR ligands can include both non-MHC and non-presented antigens (when they are known at all) (11). Mechanistic studies on TCRs are therefore both extremely important, yet complicated due to the molecular diversity of the system.

TCR chains are frequently only reported as annotated rearrangements, consisting of the involved V and J genes plus the CDR3 sequence spanning the hypervariable region. In principle, assuming that the TCR has been correctly annotated, this contains all of the necessary information to reproduce the entire coding sequence (with the D gene sequence being contained within the CDR3 for beta and delta chains). In order to perform experiments that require TCR expression, there is a need to reliably convert these concise TCR descriptions into full-length coding nucleotide sequences. To the best of our knowledge, no computational tool exists to do this. Instead, the traditional approach requires manual assembly of the V/J/CDR3 combinations of interest using germline TCR database repositories (such as IMGT-GENE/DB (12)) and a text editor or DNA software tool. While this approach produces valid results, it is (1) slow and labor intensive, thus scaling poorly; (2) vulnerable to human error, leading to (3) poor repeatability (by one user) and reproducibility (by others) (13).

To overcome these limitations we developed Stitchr, the first software capable of automatic generation of full-length coding TCR nucleotide sequences from minimally reported V/J/CDR3 information. Stitchr produces a nucleotide sequence encoding the CDR3 inserted in-frame between the provided V and J genes segments, adding properly spliced upstream leader and downstream constant region sequences. Stitchr can be used to produce TCRs of any of the four conserved loci – alpha, beta, gamma, and delta – across all species for which IMGT-GENE/DB stores sufficient data. The modularity of its approach also allows users to substitute in different, even non-natural, TCR gene segments, resulting in rapid sequence generation for protein expression and engineering experiments. We demonstrate that Stitchr produces the expected TCR sequences and verify the approach by synthesizing expression vectors for TCRs of known specificity and demonstrating their activity in Jurkat cells. We also report Thimble, a companion script which allows users to run Stitchr on single or paired-chain TCR repertoires, capable of processing a million TCRs in under ten minutes using a standard desktop personal computer. Finally, we illustrate case examples where high-throughput TCR datasets can be accurately converted to full-length coding equivalents. We propose that Stitchr and related tools will accelerate the pace of TCR research at the interface between experimental validation and high-throughput bioinformatic analysis.

MATERIALS AND METHODS

Stitchr implementation

Stitchr was written in Python, tested exhaustively on Python 3.6.9 on Ubuntu and Python 3.7.7 on Mac OS. The only non-standard package needed for its operation is PySimpleGUI (version ≥4.45.0), if users elect to use the graphical user interface script. It is also provided with Thimble, a companion wrapper script which allows users to supply TCRs in a tab separated spreadsheet file, for the simultaneous stitching of multiple receptors (which, when retained alongside the germline data used for TCR generation, provide full documentation for good data provenance practices). All scripts and data necessary to run Stitchr can be found on the GitHub repository: https://github.com/JamieHeather/stitchr.

By default, Stitchr uses germline sequences downloaded from IMGT/GENE-DB (12) (last updated on 2022-02-09, from IMGT release number 20225-7). It takes as input a TCR rearrangement described by the V and J genes used, plus the CDR3 junction sequence, which can be provided as a nucleotide or amino acid sequence (running inclusively from the cysteine to the phenylalanine or equivalent residue) or as a longer nucleotide sequence extending further into the V/J genes.

The exact mode of determining the CDR3 junction depends on the input format. If an amino acid CDR3 is provided, Stitchr looks for sequence matches between its N terminus and the C terminus of the translated relevant V gene, incrementally deleting germline V gene residues until it finds a match. The non-V portion of the CDR3 is then used to similarly search the translated J gene until the minimal CDR3 contribution is met, leaving the portion of the CDR3 junction which is not feasibly encoded by either germline gene (which will include any remaining residues from the D gene in the case of beta and delta chains). The nucleotide sequences for the wholly-germline encodable V/J residues are then produced by trimming the provided reference sequences, and the intervening non-templated sequence is generated by selecting the most common codon for that residue from a species-specific frequency table. Users can also provide exact CDR3 junctions as nucleotides, which are first translated and then processed in a similar manner up until the non-templated region is determined, at which point the corresponding segment of the provided nucleotide sequence is used in place of a codon-optimized selection.

If users have additional nucleotide context beyond the junction, they can also use the 'seamless' nucleotide stitching option (abbreviated 'SL-NT' in the figures) for more faithful replication of the nucleotide sequence. In this mode the overlap detection occurs at the nucleotide level: the 5′ of the provided junctional sequence is searched against the incrementally deleted 3′ of the V gene, and the 3′ of the non-V portion is searched against the 5′ of the J. Once the sites of overlap between the V, the junction, and the J are determined, the extraneous sections of the germline

V/J genes are removed, and the sequences are joined. The leader and constant regions necessary for expression are then added. By default, these sequences will be inferred from the V and J genes chosen respectively, defaulting to the prototypical allele's (*01) leader sequence if IMGT does not record one for the specified allele, and taking the relevant TRBC gene per TRBJ cluster (although both values can be optionally overridden). Note that for non-human/non-mouse species (which may differ in their TCR loci architecture) the constant region to be used must also be specified. Users can also specify any additional sequences to the 5′ and 3′ of the total rearrangement (e.g. to add Kozak sequences, stop codons, restriction enzyme sites, or primer binding sites, if TCRs are to be synthesized). If Thimble or the graphical user interface script is used, the individually stitched chains of a heterodimer can be linked together with any desired sequence (e.g. a 2A self-cleaving peptide sequence) for bicistronic vector expression.

Benchmarking Thimble

Large TCR V/J/CDR3 datasets were obtained by randomly picking five samples from the Emerson *et al.* Adaptive Biotechnologies' dataset (DOI: 10.21417/B7001Z, samples HIP02873, HIP02805, HIP02811, HIP02820, and HIP02855) (14) with the Python random.choice function, and downloading the entire VDJdb database (using the 'vdjdb_slim' file, accessed on 2021-02-02) (15). The Adaptive Biotechnologies data are beta chain sequences produced with a unified experimental and analytical pipeline from healthy donor peripheral blood mononuclear cell (PBMC) gDNA, while the VDJdb TCRs have been extracted and annotated from the literature and thus represent a diverse array of input cell sources and TCR identification processes, featuring both alpha and beta chains. Adaptive Biotechnologies uses custom non-standard identifiers to refer to TCR genes, so these were first converted to standard IMGT nomenclature and Adaptive Immune Receptor Repertoire Community (AIRR-C) format (16) with the Python script immunoseq2airr (version 1.2.0, DOI: 10.5281/zenodo.5224597). This was run with the following non-default parameters to account for input file formats and filter non-productive and non-interpretable rearrangements: -nd, -a, -or, -pf, -mf, and -p (pointing to the Emerson parameter conversion file provided in the repository). Rearrangements from both datasets were filtered to keep only in-frame potentially productive chains with both a V and J gene call. For ambiguous cases with >1 V gene call the first gene provided was used. Non-human TCRs were discarded from VDJdb. Note that Thimble successfully produces stitched sequences for >99% of all input V/J/CDR3 combinations, with the vast majority of those that fail lacking complete CDR3 junctions (i.e. they do not run from the conserved C to F residues, inclusively). In order to establish a wide dynamic range, variable numbers of TCRs were randomly drawn from these datasets (pooling all of the Adaptive samples into a single file) using the Python random.choices function, up- or down-sampling to 1e2, 1e3, 1e4, 1e5, or 1e6 rows, three times per dataset.

Generating simulated TCR data with immuneSIM

50,000 AIRR-C compliant human TCR alpha and beta recombinations were simulated using the R package immuneSIM (17) (version 0.8.7), with a minimum CDR3 length of eight residues (and all other settings default). Rearrangements with CDR3 junctions not ending in one of the three conserved terminating residues found in human predicted-functional TRAJ/TRBJ genes (phenylalanine, tryptophan, or cysteine) were filtered and removed.

High-quality long read TCR-seq dataset generation and analysis

In order to obtain 'real' full-length V domain TCR repertoire sequences, we leveraged published datasets produced in part by one of the authors previously, in which αβ TCR-seq was performed on RNA extracted from whole blood taken from 16 healthy volunteers (18,19). These data were produced using a ligation-based 5′RACE strategy, in which random unique molecular identifier (UMI) barcodes were added to TCR cDNA prior to amplification, allowing for error-correction downstream. However, in contrast to previous studies, in which TCR rearrangements were annotated using only the constant region-proximal read of the paired-end sequencing (R1) prior to error correction, raw FASTQ were first merged using FLASH (version 1.2.11, default parameters) (20). This identifies paired end reads with 3′ overlap and combines them into a FASTQ with longer complete amplicons. Merged reads which share UMIs were then collapsed and error-corrected using stringent criteria: each UMI had to contain at least three reads, and the calls of those reads must all have quality scores ≥Q25. The consensus base at each position was then determined, with the abundance of a given consensus reported by counting the number of associated barcodes, and output as FASTA files. While this greatly reduces the number of available reads in a repertoire file, the remaining reads are more likely to cover the entire variable domain and much less likely to contain PCR or sequencing errors. For the purposes of testing Stitchr/Thimble, these donor/locus specific FASTA files were combined into one large repertoire file.

Extended reads were then analyzed using a modified version of the TCR annotation software Decombinator (21,22), called autoDCR. Like Decombinator, autoDCR uses short (20 nt) 'tag' sequences to populate an Aho-Corasick trie (or search tree) for efficient string-matching based V/J gene identification. However, unlike the original Decombinator implementation in which single CDR3-proximal tags are selected for their unique occurrence in single genes, autoDCR tiles 20-mer tags overlapping 10 nt across the entirety of every allele of every gene, making V and J gene calls based on the presence of multiple tag matches. While the much larger trie takes longer to search each read, it outputs V and J gene calls with allele-level accuracy. This allows determination of sequence across the length of the rearrangement, enabling selection of reads which include the start of the V gene. Technically, this is achieved by using the '-jv' flag, which outputs the 'jump' values indicating the furthest positions of V/J tag matches: filtering on v_jump values = 0 selects reads where the first tag match corresponds to the start of the V-REGION. Additionally, TCRs with ambiguous gene calls (>1 allele), or those using alleles for which only partial nucleotide information is available in IMGT, were filtered out.

This feature of detecting overlapping tags was also utilized to perform rudimentary detection of novel TCR alleles, inspired by earlier studies in TCRs (23) and immunoglobulins (24), guided by the hypothesis that (a) individuals with alleles not present in IMGT will exist in our cohort and (b) most of these are likely just single nucleotide variants (SNV). TCRs with fully sequenced variable domains from each donor were screened for potential novel alleles, indicated by multiple rearrangements using the same V gene but which all share a two-tag mismatch with the reference (as a SNV relative to a recorded germline gene will result in two consecutive overlapping tags failing to match), with the same sequence spanning the break. To distinguish potential novel allele variants from PCR/sequencing errors, sequences had to meet the following criteria: (i) be present in a V gene with ≥10 distinct recombinations with unambiguous gene calls; (ii) account for ≥5% of the reads belonging to that V gene; (iii) be found in ≥3 unique recombinations; (iv) account for ≥10% of reads for that V gene which contain a break of two tag matches. As an additional check, we called potential novel alleles only if they occurred in the top two most abundant sequences for that gene, which would then constitute the genotype for that gene in that donor. Inferred potential novel alleles were assigned an identifier indicating their variant suffixed to their original reference allele match (e.g. TRAV27*01_A233G) and output as FASTA reads in IMGT format, and then either appended to the IMGT database to re-generate tags for autoDCR, or supplied to Stitchr by including them in the 'additional-genes.fasta' file.

In vitro TCR validation

Full-length TCRα and TCRβ coding regions generated by Stitchr were used to generate TCRαβ lentiviral expression constructs with the BioXp 3200 system (Codex DNA) as previously described (25). In brief, the TCRα and TCRβ sequences were generated with Stitchr (using the same TRBC2 constant region and TRAV21/TRBV6-5 leader sequences for reliable expression) and cloned into a pReceiver-based lentiviral vector (GeneCopoeia) that contained an EF1α promoter and puromycin resistance gene. The bicistronic TCRα-TCRβ coding region incorporated a P2A ribosomal skip motif (26) to generate independent TCR polypeptide chains. Lentivirus was packaged into VSV-G pseudotyped particles using the third-generation ViraSafe Lentiviral Packaging System (Cell Biolabs) and Lenti-Pac 293Ta cells (GeneCopoeia) (27). Fresh lentiviral supernatant was used to transduce TCRβ-deficient (ΔTCRβ) Jurkat cells (J.RT3-T3.5; ATCC TIB-153), which were previously engineered to stably express human CD8 (lentiviral construct layout: PGK promoter – CD8A-P2A-CD8B(M-1) – IRES-blasticidin resistance gene) (28,29). Transduced ΔTCRβ Jurkat cells were selected with puromycin for 14 days and introduced TCR surface expression confirmed by antibody staining for CD3 and TCRαβ surface expression.

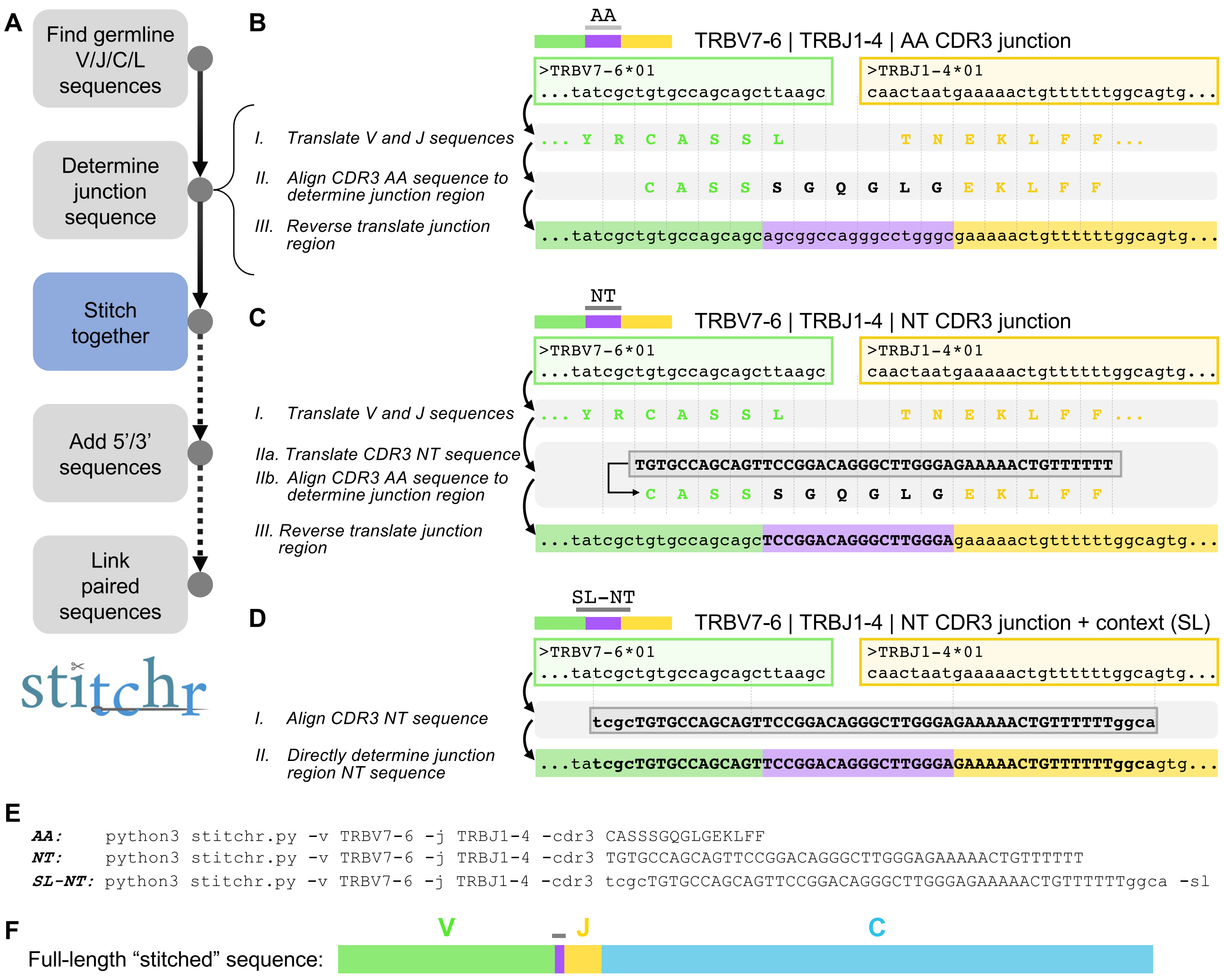
TCR-engineered Jurkat cells and target cell lines (Supplementary Tables S1 and S2) were cultured in RPMI 1640 supplemented with 10% FBS, 2 mM glutamine, and penicillin/streptomycin. Target cell line HLA type information was obtained from the TRON Cell Line Portal (30). PBS-washed target cells were stained with 1 μM CFSE (eBioscience) for 10 min at RT, before being quenched with 5× volumes of complete media, incubated on ice for 10 min, and repeatedly washed in media. Stained cells were then peptide pulsed with 10, 1, or 0 μg/ml of the relevant peptide epitope (GenScript, ≥85% purity) for 1 hour at RT in complete media, washed twice, and co-cultured overnight (18–22 h) with Jurkat cells expressing the relevant TCR at an effector-to-target ratio of 2:1. Experiments were set up in round-bottomed 96-well plates, with 2e5 Jurkats and/or 1e5 target cells per well, with each experimental condition in triplicate. Plates contained additional no target cell negative control and anti-CD3 antibody (CD3-2, Mabtech) positive control wells. Following co-incubation, cells were washed in PBS and then FACS buffer (PBS with 2% FBS and 1 mM EDTA). Cells were then stained in the dark on ice for 30 min with 0.5 μl anti-CD62L and anti-CD69 antibodies conjugated to PE and APC respectively (BioLegend, clones DREG-56 and FN50) in 100 μl FACS buffer per well. After washing twice, cells were resuspended in 100 μl of FACS buffer with 2 μl 7AAD (BioLegend) and incubated in the dark on ice for a further 5 min. Antibody staining was quantified on an Accuri C6 Plus flow cytometer (BD Biosciences) with a CSampler. FCS data were analyzed with FlowJo version 10.7.1.

Data analysis and visualization

All non-FCS analysis and data visualization was carried out in Python ≥3.6.9, using a combination of matplotlib (version 3.3.2) (31), pandas (1.1.2) (32) and seaborn (0.11.0) (33). PDB TCR amino acid sequences were aligned using Clustal Omega (34) version 1.2.4, accessed via the web portal in April 2021. Low-throughput V/J/CDR3 annotation for all non-human and/or non-TRA/TRB analysis (including immunoglobulins) was performed using IMGT/V-QUEST version 3.5.28, using default parameters (35).

RESULTS

We designed Stitchr to take the minimal V gene, J gene, and CDR3 junction sequences typically used to report TCR chains and generate a complete functional coding nucleotide sequence (Figure 1A). Stitchr accommodates multiple CDR3 formats: the junction sequence (running inclusively between the conserved V gene cysteine and J gene phenylalanine) can be supplied in either amino acid or nucleotide form. Stitchr determines which residues of the translated CDR3 can be encoded wholly by germline sequences, and then all 'non-templated' nucleotide sequences between those portions (which also contains D gene contributions for the beta and delta chains) will either be generated from a species-specific codon frequency table (for amino acid 'AA' input – Figure 1B) or will be cropped out of the provided nucleotide sequence (for nucleotide


A
Find germline V/J/C/L sequences
Determine junction sequence
Stitch together
Add 5′/3′ sequences
Link paired sequences
stitchr
B
AA
TRBV7-6 | TRBJ1-4 | AA CDR3 junction
>TRBV7-6*01
...tatcgctgtgccagcagcttaagc
>TRBJ1-4*01
caactaatgaaaaactgtttttttggcagtg...
I. Translate V and J sequences
II. Align CDR3 AA sequence to determine junction region
III. Reverse translate junction region
... Y R C A S S L T N E K L F F ...
C A S S S G Q G L G E K L F F
...tatcgctgtgccagcagcagcggccagggcctgggcgaaaaactgtttttttggcagtg...
C
NT
TRBV7-6 | TRBJ1-4 | NT CDR3 junction
>TRBV7-6*01
...tatcgctgtgccagcagcttaagc
>TRBJ1-4*01
caactaatgaaaaactgtttttttggcagtg...
I. Translate V and J sequences
IIa. Translate CDR3 NT sequence
IIb. Align CDR3 AA sequence to determine junction region
III. Reverse translate junction region
... Y R C A S S L T N E K L F F ...
TGTGCCAGCAGTTCCGGACAGGGCTTGGGAGAAAAACTGTTTTTT
C A S S S G Q G L G E K L F F
...tatcgctgtgccagcagcTCCGGACAGGGCTTGGGAGaaaaactgtttttttggcagtg...
D
SL-NT
TRBV7-6 | TRBJ1-4 | NT CDR3 junction + context (SL)
>TRBV7-6*01
...tatcgctgtgccagcagcttaagc
>TRBJ1-4*01
caactaatgaaaaactgtttttttggcagtg...
I. Align CDR3 NT sequence
II. Directly determine junction region NT sequence
tcgcTGTGCCAGCAGTTCCGGACAGGGCTTGGGAGAAAAACTGTTTTTTggca
...tatcgcTGTGCCAGCAGTTCCGGACAGGGCTTGGGAGAAAAACTGTTTTTTggcagtg...
E
F
Full-length “stitched” sequence:
V
J
C

```
AA:     python3 stitchr.py -v TRBV7-6 -j TRBJ1-4 -cdr3 CASSSGQGLGEKLFF
NT:     python3 stitchr.py -v TRBV7-6 -j TRBJ1-4 -cdr3 TGTGCCAGCAGTTCCGGACAGGGCTTGGGAGAAAAACTGTTTTTT
SL-NT: python3 stitchr.py -v TRBV7-6 -j TRBJ1-4 -cdr3 tcgcTGTGCCAGCAGTTCCGGACAGGGCTTGGGAGAAAAACTGTTTTTTggca -sl
```

Figure 1. Schematic of Stitchr algorithm. (**A**) Overview of Stitchr modules. Stitchr first obtains germline V gene, J gene, constant region (C), and leader (L) sequences from IMGT/GENE-DB. Next, the junction-spanning sequence is determined, depending on input mode (see B–D), and the complete TCR sequence is assembled. Complete single chain rearrangements can subsequently have arbitrary user-provided sequences appended to the 5′ or 3′ of the TCR, and finally paired chains can be combined (e.g. via a 2A self-cleaving peptide sequence) into a bicistronic single expression sequence. (**B**) When an amino acid (AA) CDR3 junction sequence is provided, the V and J genes are translated (I), aligned, and ‘deleted’ back from the CDR3-proximal edge until the longest possible overlap with the appropriate side of the junction is found (II), i.e. the longest suffix of the V that matches the prefix of the CDR3, or vice versa for the J. The remaining residues which cannot be encoded by the germline genes are then ‘reverse translated’ using a codon frequency table (III), and the trimmed germline genes and non-templated residues are concatenated. Vertical dotted lines show codons. (**C**) If provided a nucleotide (NT) CDR3 junction sequence (depicted by bold/capitalized font), the germline genes are again translated (I), as well as the CDR3 sequence (IIa). The amino acid sequences are aligned and the germline contributions to the CDR3 are determined (IIb). The AA sequence is then converted to NT, however instead of assigning codons for the non-templated residues based on a codon usage table, the nucleotides in the provided CDR3 are used (III, bold text indicates retained original NT sequence). (**D**) If the provided junction sequence includes additional nucleotide sequence context that extends beyond the CDR3 (depicted by lowercase text), the ‘seamless’ (SL) option can be used. In this mode, V and J germline genes are again deleted to the edge of the overlapping NT sequence (vertical dotted lines), allowing Stitchr to seamlessly combine germline V and J with the provided CDR3-spanning sequence (II). (**E**) Examples of actual Stitchr commands used to run the examples shown in B (AA), C (NT), and D (NT-SL). Note that Stitchr defaults to human TCRs, thus the species flag doesn't need to be set here. All three options produce a full-length TCR sequence (F) that encodes the same amino acid sequence, with the seamless option reproducing the identical nucleotide sequence (assuming the correct V and J alleles were provided).

‘NT’ input – Figure 1C). Alternatively, if a nucleotide sequence that extends beyond the edges of the CDR3 junction is supplied, Stitchr seamlessly (‘SL-NT’) integrates it into the germline V and J genes after computationally ‘deleting’ from the germline genes and looking for overlapping sequences (Figure 1D). Example code used to generate the illustrated TCRs is shown in Figure 1E. Complete TCR nucleotide sequences (Figure 1F) are output after splicing on 5′ leader and 3′ constant region sequences, with the option to add arbitrary sequences to either side of a gene (e.g. for gene regulatory or PCR/cloning purposes). Stitchr can be run as a command line tool for a single chain recombination, or via a wrapper script (discussed below) for larger numbers of sequences, which also supports the production of bicistronic sequences for paired αβ or γδ chains (e.g. for making expression vectors). Alternatively, we designed a graphical user interface that can be used to generate single or paired chain receptors (Supplementary Figure S1).

To test the capabilities of Stitchr, we used the command line interface to generate full-length human alpha/beta TCR sequences from four published receptors that have rigorous data demonstrating epitope recognition, including solved structures of the TCR–peptide–MHC complex, and which cover a range of V/J gene combinations and HLA restrictions (Figure 2A, Supplementary Tables S1 and S2) (36–39). We then downloaded amino acid sequences for each of these alpha-beta TCRs from the Protein Data Bank (PDB) (40) and aligned them against translations of the Stitchr-generated sequences (Figure 2B). With one exception, the variable domain sequences produced by Stitchr aligned perfectly with the PDB TCR structures. The exception – in the alpha chain V gene segment of the MAG-IC3 TCR – is explained by that TCR having been engineered to include a non-germline modification in order to increase affinity (39). As Stitchr makes use of a modular approach in its selection of genes, alternative and additional sequences can be added to expand the types of TCRs that can be produced. When we supplied Stitchr with a suitable reference for this altered V gene, it accurately reproduced the PDB amino acid sequence. Thus, Stitchr faithfully replicates TCR sequences, at least when the templated sections are present in the 'germline' genes supplied to it.

As the process through which TCRs are produced is conserved between alpha/beta and gamma/delta TCRs, Stitchr readily applies to gamma/delta chain sequences as well (Supplementary Figure S2A). Indeed, as long as Stitchr can be provided sufficient, suitably formatted sequence data (i.e. at least one each of leader, variable, joining, and constant region sequences in IMGT format) it can generate TCRs for any locus from any species. Including the four shown already for humans, there are currently 42 eligible loci across 14 species available in IMGT-GENE/DB (Supplementary Table S3). We validated an additional 19 of those, covering as many species/loci combinations for which we could find annotated rearranged mRNA sequences in GenBank (Supplementary Figure S2B). Furthermore, the process can even be extended to immunoglobulins, demonstrated in Supplementary Figure S3 with example human heavy and light chain recombinations.

Stitchr's potential for rational protein design was further explored by generating the beta chain of the anti-MART1 TCR DMF5 (41) in combination with different constant regions. Stitchr generated appropriate coding sequences correctly spliced onto human TRBC1, TRAC, TRDC, and TRGC1, and mouse TRBC1 constant regions (Supplementary Figure S4), thus supporting the use of Stitchr for generating even non-natural TCR sequences.

To test whether TCRs produced by Stitchr are functional, we generated expression constructs for the four TCRs from Figure 2B, plus an additional published TCR which expanded the range of TCR genes and HLA alleles covered (but which lacked structural data) (42) (Supplementary Tables S1 and S2). P2A-linked bicistronic TCR constructs were stably expressed in ΔTCRβ Jurkat cells and co-cultured with cognate peptide-pulsed cancer cell lines that express the relevant HLA allele. Jurkat activation (CD69+/CD62L-negative) was assayed by flow cytometry (Supplementary Figure S5). We observed dose-dependent peptide-induced activation of TCR-Jurkat cells when cultured with target cells bearing the appropriate HLA allele, but not with HLA-mismatched target cells (Figure 2C). These results confirm that Stitchr generates functional TCRs that reproduce the antigen specificity of the rearrangements they replicate.

With the rise of high-throughput TCR sequencing technologies, large TCR datasets are increasingly available. However, many such studies do not sequence the entire variable domain, and an even smaller fraction sequence the entire transcript. Even when the entire chain is sequenced, it is not often reported nor sufficient raw data provided to extract it, which can limit usefulness for certain applications. To overcome this limitation, we developed Thimble, a companion wrapper script which allows Stitchr to be applied to multiple TCRs in a single command. We benchmarked Thimble using large, published datasets (see Materials and Methods), focusing on human alpha/beta sequences as these have the most data available. This revealed that run-time scales linearly with number of input TCRs, taking under ten minutes to process a million TCRs on a standard desktop personal computer when provided with amino acid CDR3s (Figure 3A), successfully stitching ≥99.9% of all input rearrangements (Supplementary Figure S6A). The Emerson *et al.* data (14) also contained the original TCR-seq reads used for TCR gene annotation, which extend 20-40 nucleotides beyond the CDR3 into the V gene: when used as input for seamless mode stitching this resulted in ∼10× times slower run-time relative to the amino acid sequence input (Supplementary Figure S6B).

While these published datasets allowed us to benchmark basic run results, they do not contain complete variable domains, and were thus unable to confirm that Stitchr and Thimble were generating the correct nucleotide sequences. To generate gold-standard TCR sequences spanning the entirety of the variable domain necessary to rigorously assess this, we used the immuneSIM tool (17) to simulate V(D)J recombination, creating known TCR sequences from the IMGT germline reference database and predetermined generation probabilities (Figure 3B). This produces a repertoire with a normal distribution of variable domain lengths (Supplementary Figure S7A). These were then converted to Thimble input files, submitting the CDR3 junction as either amino acids (AA), nucleotides (NT), or as nucleotides with different 5′ and 3′ lengths (10/10, 20/20, 30/30, or 200/30 nt) for seamless (SL) integration. As expected, the seamless options took longer to run, scaling with longer nucleotide contexts (Figure 3C), but all options were completely 'stitchable' (Supplementary Figure S7B). By comparing the nucleotide and translated sequences produced by Stitchr to the original sequences generated by immuneSIM, we observed that Stitchr's accuracy was very high, perfectly recapitulating AA sequences in all modes, NT sequences in all seamless modes, and almost 99% correct NT residues even when supplying CDR3s as AA input (Figure 3D and Supplementary Figure S7C). Examination of the distribution of these mismatches

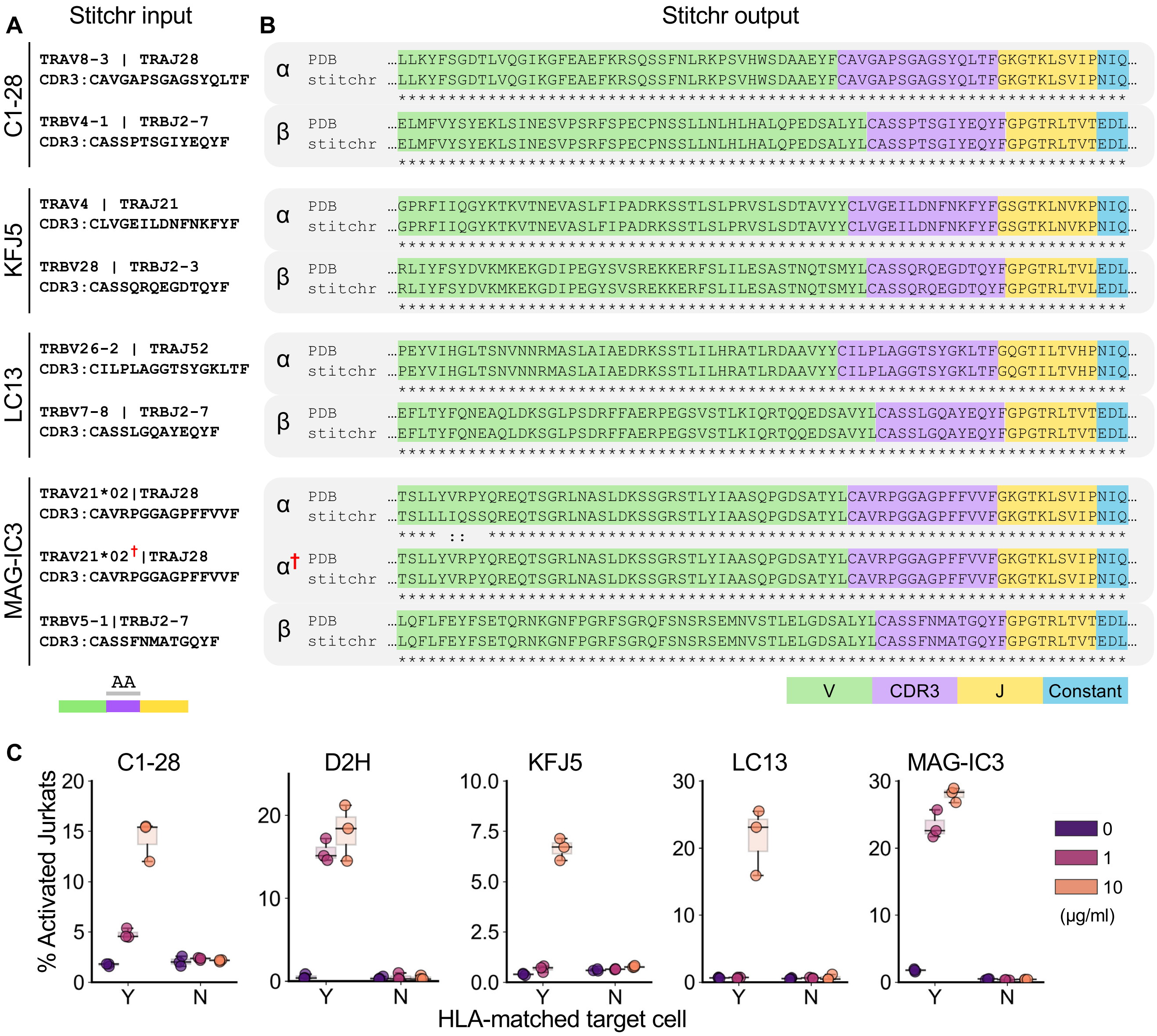


Figure 2. Validation of Stitchr-generated TCR sequences. Amino acid CDR3 sequences of four TCR heterodimers were used as Stitchr input (**A**) and stitched output sequences were aligned (**B**) to the rearranged sequences extracted from the corresponding PDB structures (using 'ATG' in place of leaders omitted from the crystallized structures), showing the correct incorporation of junction sequence and constant region. MAG-IC3 'α†' sequence indicates Stitchr output using a modified TRAV21*02 gene to replicate the engineered amino acid sequence used in the PDB structure. (**C**) Functional validation of Stitchr-produced TCR sequences using a Jurkat activation assay. CD8-positive, TCRb-negative Jurkat cells were transduced with one of five different TCRs and co-cultured with peptide pulsed (10, 1, or 0 μg/ml) HLA-matched or mis-matched target cell lines. Data shown are triplicate technical replicates from one experiment and are representative of at least two independent biological repeats.

between input and output sequences revealed that they were all confined around the relative positions 0.8–0.9 along the variable domain, corresponding to the expected location of the CDR3 where Stitchr generates codon-optimized sequences to fill in non-templated regions (Figure 3E). Note that providing the junction as a NT sequence still infrequently produces differences, as sometimes V(D)J recombination will delete only part of the codon(s) at the recombining edges before adding alternate nucleotides that still encode the same amino acid, while Stitchr defaults to using the germline-encoded sequence (e.g. Figure 1C, where the last residue of the 'CASS' motif was encoded by 'AGT' in the rearrangement but the 'AGC' found in the germline gene gets used).

Synthetic sequences do not necessarily reflect the true complexity of empirically sequenced repertoires. We therefore leveraged a published TCR-seq dataset generated with a unique molecular index (UMI)-barcoded 5′ RACE protocol, allowing production and allele-level annotation of stringently error-corrected TCR rearrangements, running from the start of the V gene region to the end of the J (see Materials and Methods). This produced ∼365,000

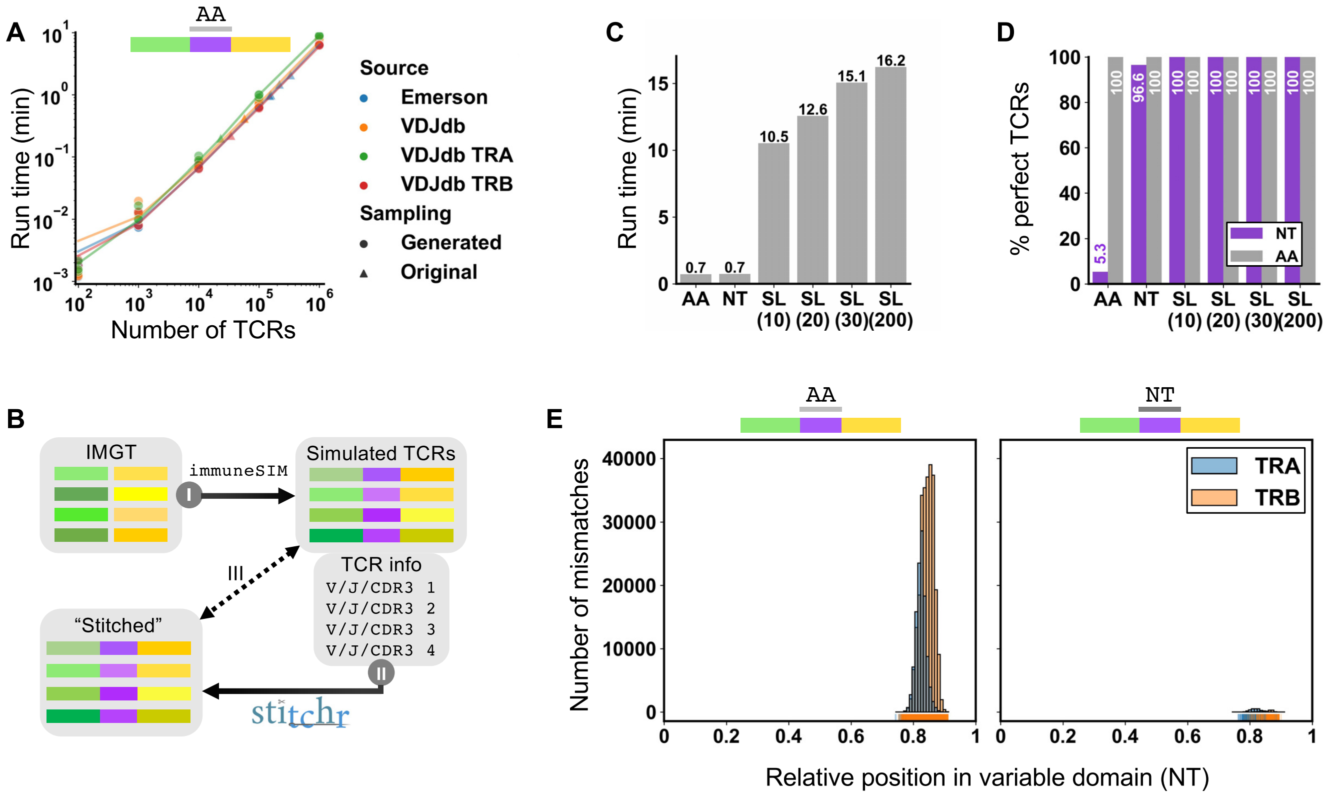


Figure 3. Application of Stitchr to high-throughput TCR datasets using the companion script Thimble. (**A**) To benchmark the speed of Thimble, large TCR datasets with amino acid CDR3s provided were downloaded either from bulk beta chain TCR-seq datasets (14), or from the curated antigen-associated TCR database VDJdb (15) (processed both all together and by each chain individually). Thimble, the high-throughput interface to Stitchr, was run on these original files (triangle markers), and from files containing 100–1,000,000 TCRs generated by randomly re-sampling these files (dot markers), with each repertoire size randomly produced 3 times. Connecting lines indicate bootstrapped locally weighted linear regressions. (**B**) Overview of sequence-level Stitchr validation. TCRs with known V/J/CDR3 information and nucleotide sequence were produced by *in silico* recombination of IMGT-stored germline genes using immuneSIM (I). V/J genes and CDR3 information (taken as exact junctions in nucleotide or amino acid forms, or as nucleotides with additional padding sequences for seamless mode) were input to Stitchr (via Thimble) (II). TCR variable domain sequences produced by Stitchr were then compared against the corresponding parental simulated TCR sequences (III). (**C**) Run time duration of Thimble applied to 50,000 α and β TCRs generated by immuneSIM, comparing different formats of junction region input: amino acid (AA), nucleotide (NT), nucleotide with padding nucleotides 5′ and 3′ for seamless (SL) integration, either 10, 20, 30, or 200 nt (200 5′, 30 3′). (**D**) Percentage of TCRs produced by Stitchr for which the variable region (start of V gene to end of J gene) perfectly matched the input sequence generated by immuneSIM, at both the nucleotide (NT, purple) and translated (AA, grey) levels. (**E**) Histogram of positional mismatches between simulated and stitched sequences for NT and AA junction input modes. Histograms were generated with 111 bins, so each bar corresponds approximately to one codon (given the variable domain length distribution of ∼333 nucleotides, Supplementary Figure S7A).

TCRs of known sequence that were then submitted to Stitchr/Thimble, providing the CDR3 junction in different formats as with immuneSIM (Supplementary Figure S8A). Basic Stitchr results were broadly similar to those seen for the simulated data (Supplementary Figure S7D–F). Inspecting the accuracy profiles revealed that as expected, the majority of cases produced correct amino acid sequences, with some expected nucleotide mismatches when providing CDR3 junctions as AA or NT (Figure 4A, B). However, there were some additional nucleotide mismatches even when providing extended junctions for seamless integration Figure 4A, top two rows) that occurred outside of the region included in the padded junction (which gets integrated into the stitched sequence and is thus perfectly matched). While some of these mismatches are likely technical errors accumulated during the TCR-seq protocol (during RT, PCR, or sequencing) that survived the error-correction process, some appeared at markedly higher frequencies than others (Figure 4A, middle row). We hypothesized that some of these peaks may represent novel TCR alleles present in our cohort that are not represented in the IMGT reference database. We performed a novel allele inference analysis on each of our donor repertoires (see Methods and Supplementary Figure S8B) and introduced the potential novel alleles inferred from that process both to our TCR annotation software and Stitchr reference files. Potential novel alleles inferred by this process are listed in Supplementary Table S4. When we re-ran the analysis, the largest mismatch peaks at both the NT (Figure 4A, bottom row) and AA (Supplementary Figure S7G) levels were no longer present. When we restricted the analysis to TCRs that were found to incorporate a potentially novel allele, we observed an increase to near-completely perfect amino acid TCR replication when using the combined IMGT + inferred allele reference database (Figure 4C). Collectively, these results demonstrate that Stitchr can be

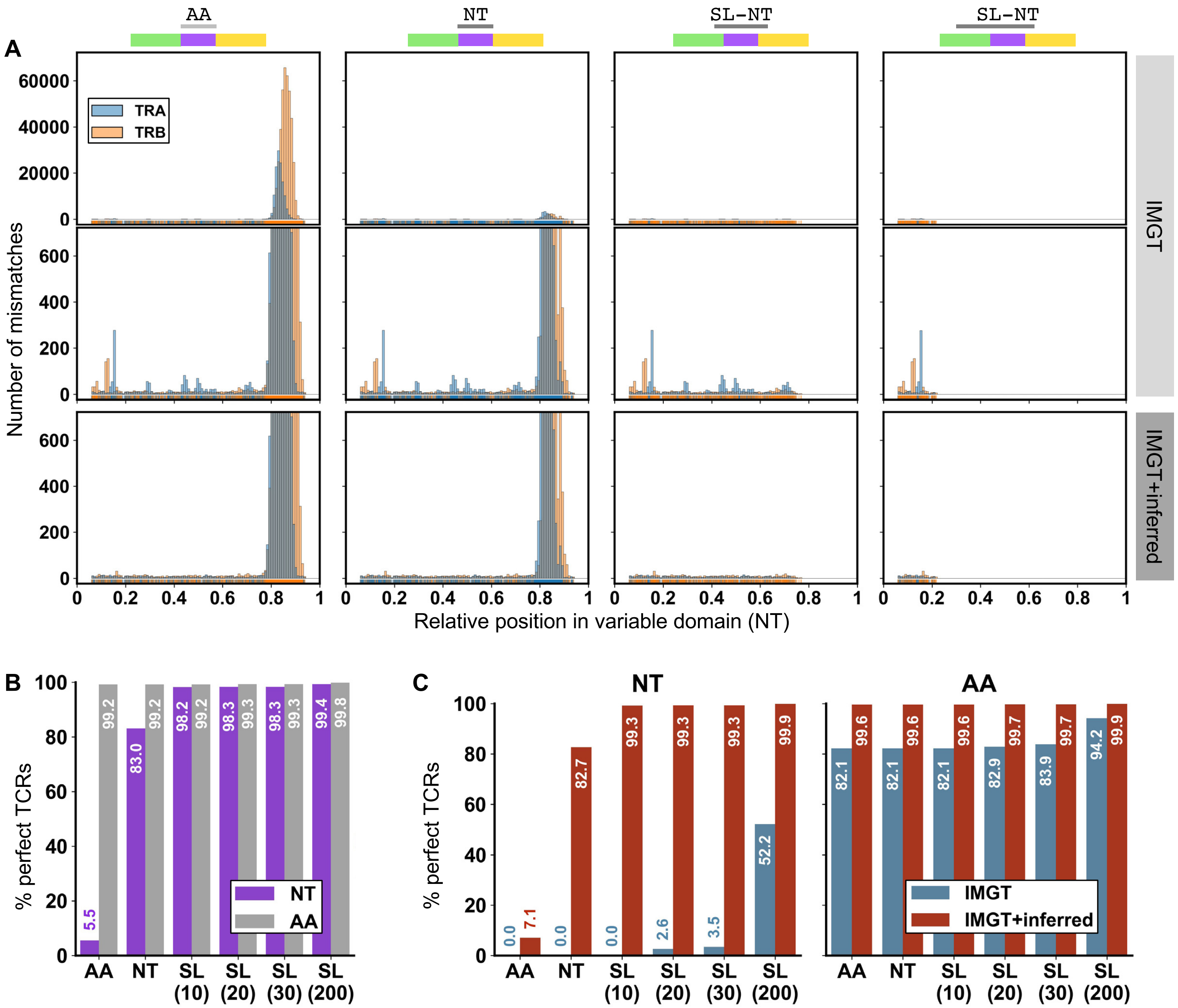


Figure 4. Assessment of Stitchr/Thimble accuracy on high-throughput TCR-seq data. (**A**) Relative positional mismatches between Thimble-generated sequences and original input sequenced TCRs for different junction inputs: (columns left-to-right): AA, NT, SL (20), SL (200). Top row shows errors when using IMGT-provided TCR germline genes only. Middle row shows the same analysis with expanded Y-axis to highlight the bottom hundredth of the mismatch range. Bottom row shows mismatches upon rerunning Stitchr/Thimble when providing additional novel TCR alleles inferred from the individual donor repertoires. (**B**) Percentage of all TCRs produced by Stitchr/Thimble that match perfectly to the original input sequences, using the IMGT reference. (**C**) Percentage of only those TCRs that use a potentially novel inferred V gene allele and agree perfectly between TCR-seq and Stitchr/Thimble output, before (blue) and after (red) including those alleles in the reference dataset, at the nucleotide (left) or amino acid (right) level.

scaled to accurately generate full-length TCR sequences for high-throughput datasets, using a variety of input CDR3 formats.

DISCUSSION

TCRs have been intensely studied since their discovery in the 1980s, drawing on many innovative approaches to overcome the challenges presented by their complexity. In particular, the advent of high-throughput sequencing technologies (TCR-seq) has allowed the identification of many orders of magnitude more rearranged TCRs than were possible with traditional techniques (43). More recently, microfluidic and other single-cell technologies have enabled high-throughput pairing of alpha-beta chain information through sequencing (44,45), and even high-throughput functional cloning of screenable libraries (25). From these efforts, and those of other experimental approaches, there now exist various databases of deep-sequenced TCR repertoires (46–49), antigen- or pathology-associated TCRs (15,50,51), and structurally determined TCR-pMHC interactions (52–54). These provide a wealth of information for other researchers to build upon. Moreover, the translational potential of TCRs is increasingly being explored, particularly for anti-cancer treatments, as TCRs are capable of directly targeting proteins other than those expressed on the cell membrane (in contrast to monoclonal antibodies, for example). This includes a range of both cellular (TCR-T) and soluble (e.g. ImmTACs and other TCR fusions) TCR therapies undergoing clinical trials (55–57). The ability of TCR research to effect change in basic immunology and in the clinic has never been greater.

Despite these advances, barriers remain both within and between the sub-fields of TCR biology. Many of these relate in some way to one of two problems: (1) researchers are often not working with full-length TCR sequences, and (2) the methodologies of different fields tend to work at different scales. The former issue typically arises as researchers work off sequencing reads shorter in length than TCR variable domains (often just targeting the CDR3 and the regions immediately adjacent), or from reported TCRs in which the sequence has been converted to the detected or inferred V and J genes and CDR3 sequence used (58) (thus lacking nucleotide sequence information entirely). While it is possible to cover the entire variable region with available sequencing technologies, this often comes as a trade-off with depth (or cost). Full-length variable domain sequencing therefore tends to be restricted to validation of small numbers of experimentally important TCRs, while most high-throughput TCR-seq datasets numbering in the thousands to millions of receptors narrow their focus to the CDR3.

Both issues can present a barrier to many subsequent experimental and computational applications, such as synthesizing and expressing TCRs for validation of antigen specificity, or computationally investigating the contribution of different regions of the TCR to certain biological properties. TCRs proposed to enter pre-clinical testing require empirical validation of their specificities, due to effects like bystander activation (59), non-specific MHC multimer reagent staining (60,61), and cross-reactivity (62,63). Moreover several bioinformatic strategies to predict antigen specificity from sequence make use of information encoded at regions outside those typically sequenced in CDR3-centric protocols (64–66).

Here, we introduce Stitchr, a Python script that uses the V/J/CDR3 TCR information commonly used to report TCR identity and tables of germline sequences to generate corresponding full-length coding nucleotide sequences. We show that the sequences produced by Stitchr faithfully reproduce the amino acid sequences of the TCRs they aim to replicate. Stitchr can also be used to assemble TCRs with non-natural sequences, as may be desirable in TCR engineering applications. Moreover, through use of the companion script Thimble, we demonstrate that Stitchr is capable of processing a million sequences in ten minutes on a standard desktop personal computer, meaning that deep-sequencing data covering the CDR3 portion of TCRs can be converted into full-length sequences in a high-throughput manner.

The modular approach by which Stitchr reads in and assembles TCR sequences provides an effective way to generate edited or non-natural TCR sequences. A common modification is the use of alternative constant regions, typically to reduce the likelihood of mispairing with endogenous TCRs when introducing an exogenous TCR. This can take the form of swapping either whole (67,68) or partial (69,70) constant regions with their orthologous equivalents from different species (e.g. swapping human for mouse sequences), swapping whole or partial sequences between loci within a species (e.g. swapping whole (71) or partial (72) alpha/beta constant region sequences), or swapping alpha/beta chain regions for gamma/delta TCR equivalents (73). Constant region domain swaps or supplementation of modified variable region genes are simply performed in Stitchr by including the wanted sequence in the reference data (Supplementary Figure S4 and Figure 2A respectively). Arbitrary non-TCR sequences can also be appended to either end of a TCR rearrangement, or even used to bridge paired TCRs into a bicistronic sequence, to facilitate molecular manipulations and transgene expression (as used for the TCRs tested in Figure 2C). It is also increasingly appreciated that the IMGT database of germline TCR alleles is incomplete – for example, a recent publication reports discovery of 38 novel TRBV alleles mined from public datasets, two of which were also found among this study's inferred alleles (23) – which raises concerns about applicability given the over-representation of certain geographic populations in these public datasets (74). Much as with non-natural modifications, we show that Stitchr can be used to include novel inferred alleles in the sequences it produces, improving the fidelity of the TCRs it generates (Figure 4).

One potential limitation of Stitchr is that it can only produce TCRs using the available component sequences (V and J genes plus leader and constant regions). While additional sequences can easily be provided, it is also possible that the V/J/CDR3 information being used as input may not faithfully reflect the 'true' sequenced TCR, e.g. if the TCR contained polymorphisms which were not captured by CDR3-targeted sequencing. V gene polymorphisms can impact upon antigen recognition (75) and surface expression levels (76), and could theoretically be recognized as foreign antigens themselves (62), thus such differences could prove functionally relevant depending on the intended application. Therefore, we recommend that wherever faithful reproduction of TCR sequences is required, full-length variable domain sequencing should be performed. Advanced users may draw on their own sequencing data, along with databases of inferred TCR alleles like OGRDB (77) and VDJbase (78) to supplement or replace the provided germline sequences as needed. This consideration becomes particularly important for non-human species which may have far less well-studied loci, and thus likely fewer annotated V/J gene polymorphisms. Users wishing to apply Stitchr to immunoglobulins must take particular care: in addition to greater germline polymorphism these loci undergo somatic hypermutation (24), meaning that genes will often effectively need to undergo a sequence inference process at the level of the clonotype, rather than just the individual. Use of the seamless mode to provide Stitchr with CDR3-spanning TCR-seq reads will also maximize the likelihood that output sequences accurately reproduce the intended TCR.

While we have demonstrated the production and validation of a small number of TCR expression constructs here, we anticipate that Stitchr and Thimble will prove useful in large-scale TCR gene synthesis and validation efforts. The domain switching illustrated in Supplementary Figure S4 could be adapted and expanded to any number of TCR-related efforts, e.g. to convert TCRs to soluble forms by adding appropriate constant region sequences (79). As screens for cancer- or pathogen-recognizing clonotypes increase (80), and more engineered TCR assays and immunotherapies are developed (55,56), we believe that a tool like Stitchr stands to benefit the field by reducing the time and effort currently spent manually assembling full-length TCR expression construct sequences. Moreover, by effectively converting TCR design into a programmatic process it becomes exquisitely repeatable and reproducible, consistently producing the same output given the same input. This will contribute to rigor in TCR research, accelerating the pace and minimizing the chances of mistakes in TCR sequence production, both in basic research and potentially in the clinic.

DATA AVAILABILITY

Healthy donor UMI-barcoded merged FASTQ files are available on SRA under the accession PRJNA359580. Raw flow cytometry data from Jurkat validation experiments are available from FlowRepository, under the experiment IDs #4949–4958 inclusively.

Stitchr is available on GitHub under a BSD 3-Clause License here: https://github.com/JamieHeather/stitchr. The immunoseq2airr script for converting Adaptive Biotechnologies data to standard IMGT nomenclature and AIRR-C format (https://github.com/JamieHeather/immunoseq2airr) and the autoDCR script for TCR annotation (https://github.com/JamieHeather/autoDCR/) are similarly available. The code underlying the analyses shown in this manuscript (including low-throughput Stitchr commands used in the featured example TCRs, plus additional input data required for high-throughput analyses) is available here: https://github.com/JamieHeather/stitchr-paper-analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

JMH would like to thank David Pinto (under the pseudonym Carnë Draug) and Margaret Axelrod for their early feedback on and contributions to the code on GitHub, and both Margaret and Justin Balko for their encouragement and feedback in adapting Stitchr for murine TCRs.

Author contributions: J.M.H.: conceived/wrote software, identified/verified TCRs, drafted manuscript. M.H.A.: verified TCRs. M.J.S.: identified/verified TCRs. Y.I.S., D.G.M.: guided experimental strategy, performed preliminary experiments. D.J., M.J.S., M.C., A.N.H.: project oversight, supervision and sourcing funding. All authors: contributed to and critically appraised experimental design and manuscript.

FUNDING

NIH/NCI [R01CA164273]; NIH/NIAID [R43AI120313-01 to D.S.J.]; NIH/NCI [R43CA232942 to M.J.S.]; Emily Venanzi Fund for Innovation in Lung Cancer Research; Be a Piece of the Solution, and Targeting a Cure for Lung Cancer Research Fund at MGH (Boston, MA). Funding for open access charge: Emily Venanzi Fund for Innovation in Lung Cancer Research.

Conflicts of interest statement. A.N.H. has received research support from Pfizer, Amgen, Roche/Genentech, Eli Lilly, Blueprint Medicines, Bristol Myers Squibb, Nuvalent, Scorpion Therapeutics, BridgeBio and served as a paid consultant for Nuvalent, Engine Biosciences, Tolremo Therapeutics and TigaTx. M.J.S. and D.S.J. hold equity shares in GigaMune and M.J.S. is a salaried employee of GigaMune. M.C. is currently an employee of AstraZeneca.

REFERENCES

1. Buckley,R.H. (2004) Molecular defects in human severe combined immunodeficiency and approaches to immune reconstitution. *Annu. Rev. Immunol.*, **22**, 625–655.
2. Markert,M.L., Hummell,D.S., Rosenblatt,H.M., Schiff,S.E., Harville,T.O., Williams,L.W., Schiff,R.I. and Buckley,R.H. (1998) Complete digeorge syndrome: persistence of profound immunodeficiency. *J. Pediatr.*, **132**, 7.
3. Yin,L., Dai,S., Clayton,G., Gao,W., Wang,Y., Kappler,J. and Marrack,P. (2013) Recognition of self and altered self by T cells in autoimmunity and allergy. *Protein Cell*, **4**, 8–16.
4. Yi,M., Qin,S., Zhao,W., Yu,S., Chu,Q. and Wu,K. (2018) The role of neoantigen in immune checkpoint blockade therapy. *Exp. Hematol. Oncol.*, **7**, 28.
5. Zappasodi,R., Merghoub,T. and Wolchok,J.D. (2018) Emerging concepts for immune checkpoint blockade-based combination therapies. *Cancer Cell*, **33**, 581–598.
6. Alt,F.W., Oltz,E.M., Young,F., Gorman,J., Taccioli,G. and Chen,J. (1992) VDJ recombination. *Immunol. Today*, **13**, 306–314.
7. Davis,M.M. and Bjorkman,P.J. (1988) T-cell antigen receptor genes and T-cell recognition. *Nature*, **334**, 395–402.
8. Murugan,A., Mora,T., Walczak,A.M. and Callan,C.G. (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 16161–16166.
9. Lythe,G., Callard,R.E., Hoare,R.L. and Molina-París,C. (2016) How many TCR clonotypes does a body maintain? *J. Theor. Biol.*, **389**, 214–224.
10. Radwan,J., Babik,W., Kaufman,J., Lenz,T.L. and Winternitz,J. (2020) Advances in the evolutionary understanding of MHC polymorphism. *Trends Genet.*, **36**, 298–311.
11. Deseke,M. and Prinz,I. (2020) Ligand recognition by the $\gamma\delta$ TCR and discrimination between homeostasis and stress conditions. *Cell. Mol. Immunol.*, **17**, 914–924.
12. Giudicelli,V. (2004) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.*, **33**, D256–D261.
13. Vitek,J. and Kalibera,T. (2011) Repeatability, reproducibility, and rigor in systems research. In: *Proceedings of the Ninth ACM International Conference on Embedded Software - EMSOFT '11*. ACM Press, Taipei, Taiwan, p. 33.
14. Emerson,R.O., DeWitt,W.S., Vignali,M., Gravley,J., Hu,J.K., Osborne,E.J., Desmarais,C., Klinger,M., Carlson,C.S., Hansen,J.A. *et al.* (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.*, **49**, 659–665.
15. Shugay,M., Bagaev,D.V., Zvyagin,I.V., Vroomans,R.M., Crawford,J.C., Dolton,G., Komech,E.A., Sycheva,A.L., Koneva,A.E., Egorov,E.S. *et al.* (2018) VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.*, **46**, D419–D427.

16. Vander Heiden,J.A., Marquez,S., Marthandan,N., Bukhari,S.A.C., Busse,C.E., Corrie,B., Hershberg,U., Kleinstein,S.H., Matsen,F.A. IV, Ralph,D.K. *et al.* (2018) AIRR community standardized representations for annotated immune repertoires. *Front. Immunol.*, **9**, 2206.
17. Weber,C.R., Akbar,R., Yermanos,A., Pavlović,M., Snapkov,I., Sandve,G.K., Reddy,S.T. and Greiff,V. (2020) immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics*, **36**, 3594–3596.
18. Heather,J.M., Best,K., Oakes,T., Gray,E.R., Roe,J.K., Thomas,N., Friedman,N., Noursadeghi,M. and Chain,B. (2016) Dynamic perturbations of the T-Cell receptor repertoire in chronic HIV infection and following antiretroviral therapy. *Front. Immunol.*, **6**, 644.
19. Oakes,T., Heather,J.M., Best,K., Byng-Maddick,R., Husovsky,C., Ismail,M., Joshi,K., Maxwell,G., Noursadeghi,M., Riddell,N. *et al.* (2017) Quantitative characterization of the T cell receptor repertoire of naïve and memory subsets using an integrated experimental and computational pipeline which is robust, economical, and versatile. *Front. Immunol.*, **8**, 1267.
20. Magocˇ,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.
21. Thomas,N., Heather,J., Ndifon,W., Shawe-Taylor,J. and Chain,B. (2013) Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, **29**, 542–550.
22. Peacock,T., Heather,J.M., Ronel,T. and Chain,B. (2021) Decombinator V4: an improved AIRR compliant-software package for T-cell receptor sequence annotation. *Bioinformatics*, **37**, 876–878.
23. Omer,A., Peres,A., Rodriguez,O.L., Watson,C.T., Lees,W., Polak,P., Collins,A.M. and Yaari,G. (2022) T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Med.*, **14**, 2.
24. Ohlin,M., Scheepers,C., Corcoran,M., Lees,W.D., Busse,C.E., Bagnara,D., Thörnqvist,L., Bürckert,J.-P., Jackson,K.J.L., Ralph,D. *et al.* (2019) Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming. *Front. Immunol.*, **10**, 435.
25. Spindler,M.J., Nelson,A.L., Wagner,E.K., Oppermans,N., Bridgeman,J.S., Heather,J.M., Adler,A.S., Asensio,M.A., Edgar,R.C., Lim,Y.W. *et al.* (2020) Massively parallel interrogation and mining of natively paired human TCRαβ repertoires. *Nat. Biotechnol.*, **38**, 609–619.
26. Funston,G.M., Kallioinen,S.E., de Felipe,P., Ryan,M.D. and Iggo,R.D. (2008) Expression of heterologous genes in oncolytic adenoviruses using picornaviral 2A sequences that trigger ribosome skipping. *J. Gen. Virol.*, **89**, 389–396.
27. Zufferey,R., Dull,T., Mandel,R.J., Bukovsky,A., Quiroz,D., Naldini,L. and Trono,D. (1998) Self-Inactivating lentivirus vector for safe and efficient in vivo gene delivery. *J. Virol.*, **72**, 9873–9880.
28. Lyons,G.E., Moore,T., Brasic,N., Li,M., Roszkowski,J.J. and Nishimura,M.I. (2006) Influence of human CD8 on antigen recognition by T-Cell receptor–transduced cells. *Cancer Res.*, **66**, 11455–11461.
29. Thakral,D., Dobbins,J., Devine,L. and Kavathas,P.B. (2008) Differential expression of the human CD8β splice variants and regulation of the M-2 isoform by ubiquitination. *J. Immunol.*, **180**, 7431–7442.
30. Scholtalbers,J., Boegel,S., Bukur,T., Byl,M., Goerges,S., Sorn,P., Loewer,M., Sahin,U. and Castle,J.C. (2015) TCLP: an online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. *Genome Med.*, **7**, 118.
31. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
32. McKinney,W. (2010) Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. Austin, Texas, pp. 56–61.
33. Waskom,M., Botvinnik,O., O'Kane,D., Hobson,P., Lukauskas,S., Gemperline,D., Augspurger,T., Halchenko,Y., Cole,J., Warmenhoven,J. *et al.* (2021) seaborn: statistical data visualization. *J. Open Source Sci.*, **6**, 3021.
34. Sievers,F. and Higgins,D.G. (2018) Clustal omega for making accurate alignments of many protein sequences: clustal omega for many protein sequences. *Protein Sci.*, **27**, 135–145.
35. Giudicelli,V., Brochet,X. and Lefranc,M.-P. (2011) IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb. Protoc.*, **2011**, 695–715.
36. Culshaw,A., Ladell,K., Gras,S., McLaren,J.E., Miners,K.L., Farenc,C., van den Heuvel,H., Gostick,E., Dejnirattisai,W., Wangteeraprasert,A. *et al.* (2017) Germline bias dictates cross-serotype reactivity in a common dengue-virus-specific CD8+ T cell response. *Nat. Immunol.*, **18**, 1228–1237.
37. Chan,K.F., Gully,B.S., Gras,S., Beringer,D.X., Kjer-Nielsen,L., Cebon,J., McCluskey,J., Chen,W. and Rossjohn,J. (2018) Divergent T-cell receptor recognition modes of a HLA-I restricted extended tumour-associated peptide. *Nat. Commun.*, **9**, 1026.
38. Kjer-Nielsen,L., Clements,C.S., Purcell,A.W., Brooks,A.G., Whisstock,J.C., Burrows,S.R., McCluskey,J. and Rossjohn,J. (2003) A structural basis for the selection of dominant alpha-beta T cell receptors in antiviral immunity. *Immunity*, **18**, 53–64.
39. Raman,M.C.C., Rizkallah,P.J., Simmons,R., Donnellan,Z., Dukes,J., Bossi,G., Le Provost,G.S., Todorov,P., Baston,E., Hickman,E. *et al.* (2016) Direct molecular mimicry enables off-target cardiovascular toxicity by an enhanced affinity TCR designed for cancer immunotherapy. *Sci. Rep.*, **6**, 18851.
40. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
41. Johnson,L.A., Heemskerk,B., Powell,D.J., Cohen,C.J., Morgan,R.A., Dudley,M.E., Robbins,F. and Rosenberg,S.A. (2006) Gene transfer of tumor-reactive TCR confers both high avidity and tumor reactivity to nonreactive peripheral blood mononuclear cells and tumor-infiltrating lymphocytes. *J. Immunol.*, **177**, 6548–6559.
42. Shimizu,A., Kawana-Tachikawa,A., Yamagata,A., Han,C., Zhu,D., Sato,Y., Nakamura,H., Koibuchi,T., Carlson,J., Martin,E. *et al.* (2013) Structure of TCR and antigen complexes at an immunodominant CTL epitope in HIV-1 infection. *Sci. Rep.*, **3**, 3097.
43. Six,A., Mariotti-Ferrandiz,M.E., Chaara,W., Magadan,S., Pham,H.-P., Lefranc,M.-P., Mora,T., Thomas-Vaslin,V., Walczak,A.M. and Boudinot,P. (2013) The past, present, and future of immune repertoire biology – the rise of next-generation repertoire analysis. *Front. Immunol.*, **4**, 413.
44. Stubbington,M.J.T., Lönnberg,T., Proserpio,V., Clare,S., Speak,A.O., Dougan,G. and Teichmann,S.A. (2016) T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods*, **13**, 329–332.
45. De Simone,M., Rossetti,G. and Pagani,M. (2018) Single cell T cell receptor sequencing: techniques and future challenges. *Front. Immunol.*, **9**, 1638.
46. Corrie,B.D., Marthandan,N., Zimonja,B., Jaglale,J., Zhou,Y., Barr,E., Knoetze,N., Breden,F.M.W., Christley,S., Scott,J.K. *et al.* (2018) iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.*, **284**, 24–41.
47. Christley,S., Scarborough,W., Salinas,E., Rounds,W.H., Toby,I.T., Fonner,J.M., Levin,M.K., Kim,M., Mock,S.A., Jordan,C. *et al.* (2018) VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Front. Immunol.*, **9**, 976.
48. Chen,S.-Y., Yue,T., Lei,Q. and Guo,A.-Y. (2021) TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. *Nucleic Acids Res.*, **49**, D468–D474.
49. Zhang,W., Wang,L., Liu,K., Wei,X., Yang,K., Du,W., Guo,N., Ma,C., Luo,L., Wu,J. *et al.* (2019) PIRD: pan immune repertoire database. *Bioinformatics*, **36**, 897–903.
50. Tickotsky,N., Sagiv,T., Prilusky,J., Shifrut,E. and Friedman,N. (2017) McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, **33**, 2924–2929.
51. Mahajan,S., Vita,R., Shackelford,D., Lane,J., Schulten,V., Zarebski,L., Jespersen,M.C., Marcatili,P., Nielsen,M., Sette,A. *et al.* (2018) Epitope specific antibodies and T cell receptors in the immune epitope database. *Front. Immunol.*, **9**, 2688.
52. Gowthaman,R. and Pierce,B.G. (2019) TCR3d: the T cell receptor structural repertoire database. *Bioinformatics*, **35**, 5323–5325.

53. Borrman,T., Cimons,J., Cosiano,M., Purcaro,M., Pierce,B.G., Baker,B.M. and Weng,Z. (2017) ATLAS: a database linking binding affinities with structures for wild-type and mutant TCR-pMHC complexes: linking TCR-pMHC affinities with structure. *Proteins Struct. Funct. Bioinforma.*, **85**, 908–916.
54. Leem,J., de Oliveira,S.H.P., Krawczyk,K. and Deane,C.M. (2018) STCRDab: the structural T-cell receptor database. *Nucleic Acids Res.*, **46**, D406–D412.
55. Jones,H.F., Molvi,Z., Klatt,M.G., Dao,T. and Scheinberg,D.A. (2021) Empirical and rational design of T cell receptor-based immunotherapies. *Front. Immunol.*, **11**, 585385.
56. He,Q., Jiang,X., Zhou,X. and Weng,J. (2019) Targeting cancers through TCR-peptide/MHC interactions. *J. Hematol. Oncol.J Hematol Oncol*, **12**, 139.
57. Zhang,J. and Wang,L. (2019) The emerging world of TCR-T cell trials against cancer: a systematic review. *Technol. Cancer Res. Treat.*, **18**, 1533033819831068.
58. Heather,J.M., Ismail,M., Oakes,T. and Chain,B. (2017) High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Brief. Bioinform.*, **19**, 554–565.
59. Bangs,S., Mcmichael,A. and Xu,X. (2006) Bystander T cell activation – implications for HIV infection and other diseases. *Trends Immunol.*, **27**, 518–524.
60. Chattopadhyay,P.K., Melenhorst,J.J., Ladell,K., Gostick,E., Scheinberg,P., Barrett,A.J., Wooldridge,L., Roederer,M., Sewell,A.K. and Price,D.A. (2008) Techniques to improve the direct ex vivo detection of low frequency antigen-specific CD8 $^+$ T cells with peptide-major histocompatibility complex class I tetramers: detection of low frequency antigen-specific CD8 $^+$ T cells. *Cytometry A*, **73A**, 1001–1009.
61. Dolton,G., Tungatt,K., Lloyd,A., Bianchi,V., Theaker,S.M., Trimby,A., Holland,C.J., Donia,M., Godkin,A.J., Cole,D.K. *et al.* (2015) More tricks with tetramers: a practical guide to staining T cells with peptide-MHC multimers. *Immunology*, **146**, 11–22.
62. Burrows,S.R. and Miles,J.J. (2013) Immune parameters to consider when choosing T-Cell receptors for therapy. *Front. Immunol.*, **4**, 229.
63. Sewell,A.K. (2012) Why must T cells be cross-reactive? *Nat. Rev. Immunol.*, **12**, 669–677.
64. Dash,P., Fiore-Gartland,A.J., Hertz,T., Wang,G.C., Sharma,S., Souquette,A., Crawford,J.C., Clemens,E.B., Nguyen,T.H.O., Kedzierska,K. *et al.* (2017) Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, **547**, 89–93.
65. Lanzarotti,E., Marcatili,P. and Nielsen,M. (2019) T-Cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. *Front. Immunol.*, **10**, 2080.
66. Ehrlich,R., Kamga,L., Gil,A., Luzuriaga,K., Selin,L.K. and Ghersi,D. (2021) SwarmTCR: a computational approach to predict the specificity of T cell receptors. *BMC Bioinformatics*, **22**, 422.
67. Xue,S.-A., Chen,Y., Voss,R.-H., Kisan,V., Wang,B., Chen,K.-K., He,F.-Q., Cheng,X.-X., Scolamiero,L., Holler,A. *et al.* (2020) Enhancing the expression and function of an EBV-TCR on engineered T cells by combining Sc-TCR design with CRISPR editing to prevent mispairing. *Cell. Mol. Immunol.*, **17**, 1275–1277.
68. Cohen,C.J., Zhao,Y., Zheng,Z., Rosenberg,S.A. and Morgan,R.A. (2006) Enhanced antitumor activity of murine-human hybrid T-Cell receptor (TCR) in human lymphocytes is associated with improved pairing and TCR/CD3 stability. *Cancer Res.*, **66**, 8878–8886.
69. Bialer,G., Horovitz-Fried,M., Ya'acobi,S., Morgan,R.A. and Cohen,C.J. (2010) Selected murine residues endow human TCR with enhanced tumor recognition. *J. Immunol.*, **184**, 6232–6241.
70. Sommermeyer,D. and Uckert,W. (2010) Minimal amino acid exchange in human TCR constant regions fosters improved function of TCR gene-modified T cells. *J. Immunol.*, **184**, 6223–6231.
71. Bethune,M.T., Gee,M.H., Bunse,M., Lee,M.S., Gschweng,E.H., Pagadala,M.S., Zhou,J., Cheng,D., Heath,J.R., Kohn,D.B. *et al.* (2016) Domain-swapped T cell receptors improve the safety of TCR gene therapy. *eLife*, **5**, e19095.
72. Voss,R.-H., Willemsen,R.A., Kuball,J., Grabowski,M., Engel,R., Intan,R.S., Guillaume,P., Romero,P., Huber,C. and Theobald,M. (2008) Molecular design of the cαβ interface favors specific pairing of introduced TCRαβ in human T cells. *J. Immunol.*, **180**, 391–401.
73. Tao,C., Shao,H., Zhang,W., Bo,H., Wu,F., Shen,H. and Huang,S. (2017) γδTCR immunoglobulin constant region domain exchange in human αβTCRs improves TCR pairing without altering TCR gene-modified T cell function. *Mol. Med. Rep.*, **15**, 1555–1564.
74. Peng,K., Safonova,Y., Shugay,M., Popejoy,A.B., Rodriguez,O.L., Breden,F., Brodin,P., Burkhardt,A.M., Bustamante,C., Cao-Lormeau,V.-M. *et al.* (2021) Diversity in immunogenomics: the value and the challenge. *Nat. Methods*, **18**, 588–591.
75. Gras,S., Chen,Z., Miles,J.J., Liu,Y.C., Bell,M.J., Sullivan,L.C., Kjer-Nielsen,L., Brennan,R.M., Burrows,J.M., Neller,M.A. *et al.* (2010) Allelic polymorphism in the T cell receptor and its impact on immune responses. *J. Exp. Med.*, **207**, 1555–1567.
76. Thomas,S., Mohammed,F., Reijmers,R.M., Woolston,A., Stauss,T., Kennedy,A., Stirling,D., Holler,A., Green,L., Jones,D. *et al.* (2019) Framework engineering to produce dominant T cell receptors with enhanced antigen-specific function. *Nat. Commun.*, **10**, 4451.
77. Lees,W., Busse,C.E., Corcoran,M., Ohlin,M., Scheepers,C., Matsen,F.A., Yaari,G., Watson,C.T., Community,TheAIRR, Collins,A. *et al.* (2020) OGRDB: a reference database of inferred immune receptor genes. *Nucleic Acids Res.*, **48**, D964–D970.
78. Omer,A., Shemesh,O., Peres,A., Polak,P., Shepherd,A.J., Watson,C.T., Boyd,S.D., Collins,A.M., Lees,W. and Yaari,G. (2020) VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res.*, **48**, D1051–D1056.
79. Robinson,R.A., McMurran,C., McCully,M.L. and Cole,D.K. (2021) Engineering soluble T-cell receptors for therapy. *FEBS J.*, **288**, 6159–6173.
80. Manfredi,F., Cianciotti,B.C., Potenza,A., Tassi,E., Noviello,M., Biondi,A., Ciceri,F., Bonini,C. and Ruggiero,E. (2020) TCR redirected T cells for cancer treatment: achievements, hurdles, and goals. *Front. Immunol.*, **11**, 1689.