ImmunoInformatics 20 (2025) 100058

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

The gremlin in the works: why T cell receptor researchers need to pay more attention to germline reference sequences

James M. Heather [a,*], Ayelet Peres [b], Gur Yaari [c], William Lees [d]

[a] Center for Cancer Research, Massachusetts General Hospital, Boston, USA
[b] Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel
[c] Department of Pathology, Yale School of Medicine, New Haven, CT, USA
[d] Clareo Biosciences, KY, USA

ARTICLE INFO

Keywords:
T cell receptor
TCRseq
Germline
Reference
AIRR-seq

ABSTRACT

The rise of T cell receptor (TCR) sequencing technologies is driving both new understandings of the immune system and the development of novel clinical platforms. Such analyses rely on comparing recombined TCR sequences to unrearranged germline reference sequences during V(D)J annotation. In this study we observed that, despite the importance of this step in TCR analysis, most published studies do not properly report the reference used. We use public datasets to illustrate why references should be explicitly specified: using IMGT/GENE-DB as an example, we document how the reference set changes over time. Furthermore we illustrate how prescriptivist interpretations of reference metadata may be obscuring rather than illuminating TCR biology, and demonstrate the need to perform full V gene sequencing in order to unambiguously determine the final translated TCR polypeptide sequence. In summary, we argue that in order to ensure the accuracy and reproducibility of TCR sequencing – an ever more pressing task as more TCR-based diagnostics and therapeutics are developed – we should all take more care with the development, use, and reporting of the TCR germline references used in our science.

Introduction

T cell receptors (TCRs) are the means by which T cells detect antigen, forming one of the cornerstones of adaptive immunology in jawed vertebrates, alongside immunoglobulins expressed by B cells. Through their recognition of peptide-MHC complexes, T cells are able to detect cells that may be infected, cancerous, or otherwise dangerous. The system relies on a process of somatic DNA recombination to generate a large, anticipatory repertoire of receptors that can cover a wide potential 'antigen space', protecting us from diverse hazards.

TCRs are produced through V(D)J recombination, named after the genes involved: variable, diversity, and joining regions. This occurs in both polypeptide chains in TCRs of the two common vertebrate TCR gene lineages: αβ (alpha/beta) and γδ (gamma/delta). Beta and delta chains undergo VDJ recombination, while alpha and gamma chains (which lack D genes) just undergo VJ recombination. The molecular mechanics however are conserved: the recombinase activating gene (RAG)-complex binds to conserved recognition signal sequences (RSS) flanking the V(D)J genes. RAG then recruits other factors and mediates the process of excising the intervening DNA, and ligating the once-distant regions together into a contiguous rearranged gene [1]. The system can generate large numbers of receptors as there are many V, D, and J genes per locus, which undergo further diversification through non-templated deletion and addition of nucleotides at the rearranging edges. This produces an enormous potential 'sequence space', making it unlikely that any two developing T cells produce the same combination of rearrangements [2]. TCRs can thus be used to identify mature T cells derived from the same original T cell (a 'clonotype').

As T cells activate, differentiate, and expand in response to antigen sensed through their TCRs – which do not change – measuring TCRs can provide insights into immune responses. This includes predicting which antigens an individual may have been exposed to or inferring a disease state [3], tracking minimal residual disease in T cell-derived malignancies [4], and permitting the production of therapeutics, as TCRs specific to a cancer antigen can redirect a patient's immune cells to kill their tumours [5]. TCR analysis is therefore hugely important across basic, translational, and clinical immunology research.

Central to such study is the sequencing and analysis of TCR chains

* Corresponding author.
E-mail address: jheather@mgh.harvard.edu (J.M. Heather).

https://doi.org/10.1016/j.immuno.2025.100058
Received 14 April 2025; Received in revised form 11 July 2025; Accepted 26 August 2025
Available online 28 August 2025

themselves, which has been reviewed thoroughly elsewhere [6–9]. However, we believe there's an aspect of TCR research which has not received sufficient attention: TCR germline referencing. A germline reference is simply the set of pre-recombination TCR gene sequences used in a given analysis, to which rearranged sequences are compared for annotation. This may differ depending on the research question: for example D genes (which are frequently highly deleted in TCRs and not unambiguously detectable [10,11]) or leader and constant regions (which are spliced onto rearranged gene sections, and may not be targeted in a given sequencing reaction) may be omitted [12,13]. Similarly some references may differ in which alleles or genes are represented, depending on their source data, and production/curation pipeline. The germline reference is therefore the foundation upon which an analysis is built: differences in the reference – such as absent or different sequences – will likely produce differences in the results. Thus for a given TCRseq experiment to be fully understood or replicated one needs to know exactly which reference was used. In this study we explore how the field is currently describing the TCR references we use in our analyses, and give examples of how germline data considerations may be impacting our research.

Results

The current state of TCR germline reference reporting

In order to determine the current state of TCR germline reference reporting, we conducted a systematic literature review of recent TCR sequencing (TCRseq) studies (Supplementary Figure 1). As some studies perform multiple kinds of TCRseq, potentially analysed with different references, we have quantified these experimental configurations ('setups'), rather than studies. What constitutes an appropriate germline reference relies on the wider experimental context, so we have recorded the relevant species, loci, sequence production and analysis details alongside reference details, shown in Fig. 1. Reference details were recorded at different 'depths': the ultimate *source* (i.e. the body or website responsible), the specific *resource* (the named database or dataset), and further specific *version* details, necessary for exactly determining to which sequences a given reference refers (e.g. version, release number, identifier, or date of access).

As expected, most recent TCRseq experiments involved studying αβ TCRs (~90%), in humans (~76%), primarily using commercial TCRseq kits and services. Irrespective of method, most experimental setups strikingly did not explicitly detail any germline reference details ('not provided'/'NP', 325/459 = ~71%). Of the 134 setups with at least a germline source recorded, 74 (~55%) gave further details as to which resource was used; of those 36 (~49%) gave sufficient details for unambiguous resource determination. Overall, <8% of experimental configurations explicitly recorded the germline reference they used.

The frequency of germline reference reporting varied by TCRseq methodology, making direct comparison difficult. For example, publications describing data produced with Adaptive Biotechnologies (analysed with their ImmunoSEQ Analyzer platform) never described the germline reference used, as this involves a proprietary dataset that the end users cannot see or interact with. Another notable example is MiXCR, the most popular TCR analysis tool in this review not fixed to a specific commercial TCRseq product. Despite MiXCR being capable of analysing TCRs using either the MiLaboratories proprietary reference (which was recently published and made accessible [14]) or alternatives, only a quarter of its usage had a germline source listed (typically IMGT).

Unfortunately many papers exclusively reported their reference as some variant on "the IMGT database", or simply "IMGT". This does not allow determination of even the resource used, as IMGT maintains multiple sequence databases (including IMGT/GENE-DB [15], IMGT/LIGM-DB [16], and the IMGT/V-QUEST reference directory sets). While often not stated, some undeclared entries can be inferred to have used one of these, due to the field's reliance on IMGT: IMGT's own tools (IMGT/V-QUEST [17] and IMGT/HighV-QUEST [18]) use

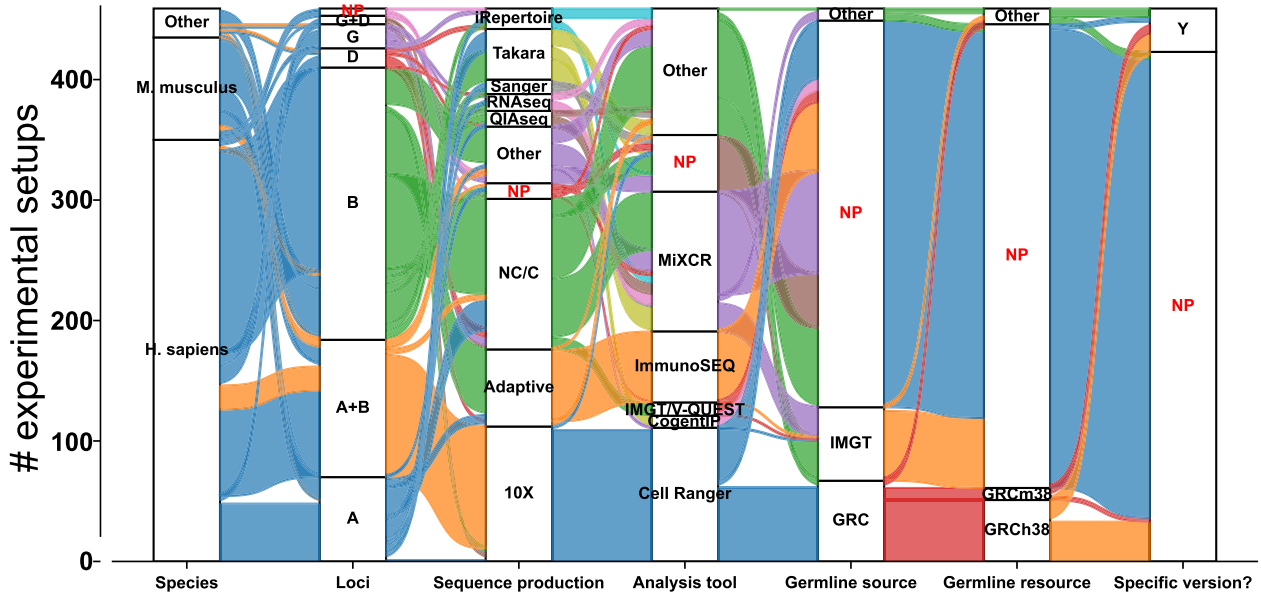

Fig. 1. Alluvial plot illustrating the sparsity of complete germline referencing of TCR analyses in recent TCRseq publications. A systematic review of TCR sequencing related papers since 2023 was conducted, reviewing papers in the web-based tool Covidence. 320 papers met the full inclusion criteria, and the data of different experimental setups (i.e. TCRseq data produced and analysed with specific techniques and tools, either for a single locus or pair of loci for single-cell data) were numerated. General abbreviations used: NP (highlighted red) = not provided (data not stated in the source publication), NC/C = non-commercial/custom. Loci abbreviations: A = TRA (α/alpha), B = TRB (β/beta), G = TRG (γ/gamma), D = TRD (δ/delta), with X+Y indicating paired chain data generated via a single-cell technology. 'Other' indicates any value for an experimental setup that was found in under 10 experimental setups for a given parameter. 'Y' in the 'Specific version?' field is short for 'yes', i.e. there was some specific version detail provided (e.g. release number or date). Note that the small number of setups which lack a specific germline resource recorded but have a specific version (top right) are four setups from three studies, which all recorded that they used 'IMGT' as well as either a date or a release, but without specifying which specific IMGT database was used.

IMGT/GENE-DB, as do many non-profit- or academic-produced tools (e.g. IgBLAST [19], TRUST4 [20], Decombinator [21], and RTCR [12], which all feature under 'Other' in the fourth column of Fig. 1).

Conversely, over half of the 10X Chromium-produced single-cell TCR setups (analysed using Cell Ranger) reported both a germline source and resource, almost half of which provided specific version details. While this is among the best reference reporting for any technique featured, most of these used specific genome assemblies from the Genome Reference Consortium (GRC), particularly GRCh38 for humans [22] and GRCm38 for mice. These are high-quality accessible references with well-defined provenance: however, they effectively only detail sequences at the *gene* level. This is unlikely to matter when analysing repertoires from inbred mice of the appropriate strains (C57BL/6J-related lines), where polymorphism would not be expected [23]. However this approach could be extremely limiting for studies in either different strains of mice, or in humans. Being outbred, our TCR alleles will be expected to differ [24] (evidenced by the frequent detection of novel alleles when searched for [14,25–29]). Other variation, such as the presence of genes not seen in the subject sequenced for GRCh38, may also confound analyses [25].

Thus most TCRseq studies published recently are not adequately reporting the TCR germline reference used, to the point where readers can obtain that exact reference. There are of course many TCR-related analyses for which only gene-level discrimination is required – or is possible, in the case of shorter-read experiments – in which the choice of germline reference used likely matters less. This will however often not be true, particularly where accuracy is required, such as when performing highly-quantitative assignment of even closely-related V(D)J gene sequences, or when annotating TCRs for functional or clinical use.

TCR germline reference data change over time

One possible interpretation of the literature review in Fig. 1 is that many researchers presume that TCR germline details don't need to be recorded because they do not differ. We sought to illustrate that this is not the case, as TCR germline references – as with many genomic references – are continually being corrected, updated, and (theoretically) improved, as new data and techniques become available.

We chose IMGT/GENE-DB, which aims to be a comprehensive catalogue of TCR and immunoglobulin complexity, suitable for diverse immunogenetic research questions [15], as the best example reference for this purpose. This is one of the longest-running accessible TCR sequence databases, being the first official repository approved by the International Union of Immunological Societies. As it's free to academics for non-commercial purposes it has been widely adopted, driving TCRseq analysis tool development.

We compared downloads of different historical releases of IMGT/GENE-DB, asking whether individual genes' and alleles' identifiers are always present and referring to the same sequence. Looking first at human alpha/beta V genes, we observed that multiple genes and alleles appeared after the earliest found release, changed sequence after first appearing, or disappeared in later releases (Fig. 2A), often in batches likely representing significant database updates. This is not unique to humans, e.g. as can be observed for mice (Supplementary Figure 2). Quantification of V/J gene alterations across all loci in three of the most studied organisms (humans, mice, and Rhesus macaques) revealed such changes are not specific to human TRAV/TRBV genes (Supplementary Figure 3), as each locus/species combination underwent some changes over time.

These differences in alleles can impact the results of TCR analyses. For example, a recent release (202514-7) shows that TRBV15*__02__ differs from TRBV15*__01__ by a single nucleotide polymorphism (SNP), yet prior to an update in 2018 (201900-0) it also lacked five nt at its 3′ terminus. As many tools rely on alignment scores, this should affect annotation of rearrangements using TRB15*02 when fewer than 4 nt were deleted during recombination. We identified such an example, the public CDR3

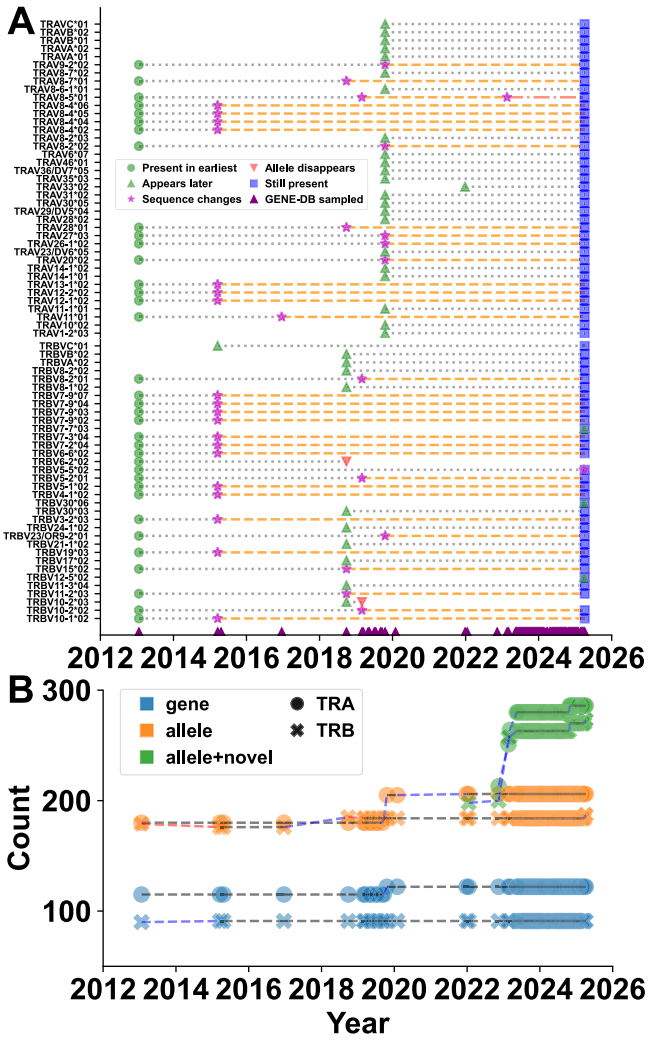

Fig. 2. Germline reference allele changes over time, using banked versions of IMGT/GENE-DB as an example (from 2013 to the present). **A**: Timeline of the human TRAV (upper) and TRBV (lower) alleles which differ depending on the version of GENE-DB used, by merit of them appearing (green up triangle) after the first record, by their sequence changing (purple star and line style change), or being removed (magenta down triangle), with green circles indicating alleles present in the earliest GENE-DB version collected in this effort (from 2013-01-20). **B**: Illustration of the cumulative gene (blue) and allele (orange) counts for human TRA (circle) and TRB (cross) V and J genes in IMGT/GENE-DB. Further cumulative novel alleles (added to those reported in IMGT/GENE-DB) reported in the literature are shown in green.

CATSRGQGYEQYF (found in 2000+ rearrangements/717 repertoires/25 studies, according to an iReceptor [30] search in June 2025), with a reported anti-viral specificity (a CMV pp65 epitope on HLA-A2) [31]. IgBLAST [19] provided with a current IMGT/GENE-DB reference can correctly distinguish TRBV15/CATSRGQGYEQYF/TRBJ2–7 rearrangements regardless of which TRBV15 allele was used. When supplying a pre-2018 reference however (e.g. 201311–0), rearrangements using either allele both get called as *01, as the additional three bases provide additional shared identity. Notably the SNP that differentiates these two TRBV15 alleles encodes different amino acids, so efforts to functionally validate such a receptor may be affected based upon the date of the reference download. TRBV10–2 is another noteworthy example: the TRBV10–2*__03__ allele appeared in 2018, but disappeared in the next banked release (201908–0), coincident with a sequence change in TRBV10–2*__02__. These seemingly are the same allele: presumably a truncated form was deposited, then the full-length version was

discovered and given a new identifier (*03), before being consolidated later as *02. A TRBV10–2*02 TCR annotated with a 2018 release would've likely been annotated as TRBV10–2*03, potentially causing confusion for later researchers to whom that allele no longer appeared to exist. Analysis of pseudogenes requires particular attention, as the lack of canonical features complicates their annotation, exemplified by TRAV8–5*01: its V region length changes dramatically, from 357 nt (2013), to 1355 nt (2020), falling back to 84 nt (presently).

These data potentially underestimate the true history of alterations in this database, as the snapshots collected are not exhaustive, particularly among the older releases which relied on archival recovery. One should also consider that databases differ in aims and scope. IMGT/GENE-DB aims to be comprehensive, hence the inclusion of partial sequences (which may need to be filtered), or those inferred from potentially sub-optimal sources (which may require updating or deletion). Conversely, GRC references do not aim cover all TCR alleles, but instead provide stable representative high-quality assemblies suitable for basic research needs. However, even these undergo some revision such that details need to be specified for unambiguous data interpretation. For example, (as reported in [26]) the GRCh37 TRB assembly has three ~20 kb sections (including five functional TRBV genes) that are not present in the primary GRCh38 assembly, instead found only in an alternative haplotype.

Furthermore, any reference that aims for coverage is only as exhaustive as the data available to it: rare alleles, or those that aren't commonly observed in TCRseq studies (potentially due to populations screened or technologies used), are unlikely to be well-documented. Several recent papers that have performed full V region-spanning sequencing of either rearranged RNA or unrearranged genomic DNA have inferred or discovered multiple putative novel alleles which are not yet featured in IMGT/GENE-DB [14,25–29]. As shown in Fig. 2B, these efforts add a substantial number of alleles: ~80 per locus for TRA/TRB. While not all of these alleles have been validated, many are detectable across techniques and in multiple donors (half described in more than one publication), suggestive of relatively high-frequency alleles that are likely missed or misassigned even in current pipelines using the most up-to-date IMGT set.

TCR germline reference metadata may be misleading

We have observed that sometimes assumptions about the biology of a TCR gene may lead to methodologies which adversely affect the results of analysis. In particular, the metadata describing expected gene functionality may be interpreted as prescriptive, rather than descriptive. That is, such metadata may be interpreted to mean that certain sequences either will not appear in their data, or won't contribute biologically; however, these fields actually intend to convey likelihoods based on reasonable assumptions, based on the best knowledge of the time. This commonly manifests in the omission of genes, particularly those predicted to not be functional (i.e. not capable of forming a viable receptor), from amplification or annotation protocols. There are two classes of such genes under IMGT nomenclature: pseudogenes ('P') which are expected to never be expressed (i.e. those containing stop codons, frameshift mutations, or V genes lacking start codons), and 'ORFs', genes which have intact open reading frames but other reasons to suspect they might not function (e.g. alterations in conserved recombination, splice, regulatory, or functional sites).

Some protocols exclude some or all of these, especially those using multiplex primer mixes, from some of the earliest studies which later produced commercial services [32,33] through to the cutting-edge of modern spatial TCR profiling [34]. However the recombination and expression of multiple P/ORF genes has been described in both gDNA and cDNA by many groups (e.g. in [35,36]) and is readily detectable in many datasets. We have taken the recent large human cohort published by Mikelov *et al.* [14]. as an exemplary TRA/TRB dataset, with deep V-REGION spanning 5′RACE coverage to illustrate this.

Fig. 3A (right) shows the average frequencies with which different TCR V/J genes are used in productive rearrangements in the 134 individuals in the cohort. Multiple genes lacking a functional 'F' label are frequently found: 7/11 detected ORF and 4/8 P genes were observed in $\geq$ 50% of donors, at frequencies comparable to many F genes. (This likely underestimates the frequency of non-functional gene usage, as rearrangements using stop codon-containing pseudogenes have been filtered out.)

The justification for the labelling of several TRAJ (which occur at reasonable frequencies) as ORFs might be expected to reduce, but not eliminate, the chances of rearranging and expressing a functional TCR. Productively rearranged examples of the most common are highlighted in Fig. 3B: TRAJ58 (which contains a 5′ stop codon that VJ recombination can remove), TRAJ25 (which has a non-canonical RSS heptamer), and TRAJ61 (which lacks a canonical splice donor site, which can rarely splice correctly as shown here).

Another group of rearrangements perhaps overlooked due to prescriptive metadata interpretations are delta V gene-containing alpha chains. The TRD locus resides inside the TRA locus, using a partially overlapping pool of V genes, of which: three are supposedly TRD-only (TRDV1, TRDV2, and TRDV3), five can feature in both chains (TRAV/DV), and several dozen are TRA-only (TRAV). However each of the three TRDV genes are readily detectable in alpha chain repertoires (Fig. 3A, left): TRDV1 rearrangements appear in all 134 donors at frequencies within the range of regular TRAV genes, while TRDV2 and TRDV3 appear in fewer donors (52 and 68 respectively), at frequencies comparable to the least-expressed TRAVs. These are rearranged with TRAJ genes, and spliced on to TRAC (e.g. Fig. 3C), and therefore theoretically as capable of being expressed as a TRA chain as any TRAV or TRAV/DV gene. Moreover there is a record of researchers describing TRDV1-TRAJ recombinations dating back to the 1990s, when they were observed at both protein and nucleotide levels, using diverse TRAJ genes, and capable of classic HLA-restricted αβ-TCR immune responses [37–40]. Frequent TRDV1-TRAJ discoveries continued as DNA sequencing technologies developed [41–44], prompting at least one group to suggest reclassifying TRDV1 as a TRAV/DV gene [45]. Despite these data, several popular tools featured in Fig. 1 are unable to detect TRDV when searching for TRA rearrangements. This currently includes IMGT/V-QUEST (which is likely to either fail to detect a rearranged alpha, or will assign an incorrect TRAV gene at low confidence) and Cell Ranger (which will result either in no rearranged TRA detected, or potentially an irrelevant second expressed chain in dual TCR cells being called as the only paired chain). While a recent publication has proposed a workaround for Cell Ranger [46], and IMGT/V-QUEST can be made to identify the V correctly by searching for TRD rearrangements instead of TRA, we expect that end users would be better served by common and potentially-functional rearrangements being detectable with default settings.

We also note that the metadata associated with a germline reference is also subject to revision, as exemplified by the human TRAJ35 gene. This was originally recorded as an ORF, as it encodes a non-canonical junction-terminating residue (Cys instead of Phe). However repeated evidence of frequent rearrangements led IMGT to relabel this as a functional gene in 2019 [47], albeit one that forms CDR3s with an uncommon end.

Germline reference utility corresponds to V gene coverage

Fig. 1 covers a tremendous breadth of TCR research, including the development of novel tools permitting a broader range of immunological hypothesis testing, by ever more researchers. Some of those studies' analyses however might be considered suboptimal from a germline perspective, as they inferred sequence beyond their capacity to do so. This primarily includes short-read bulk data where V regions are not completely sequenced, or single-cell data (which frequently has full V sequence available) mapped to a reference lacking any allelic

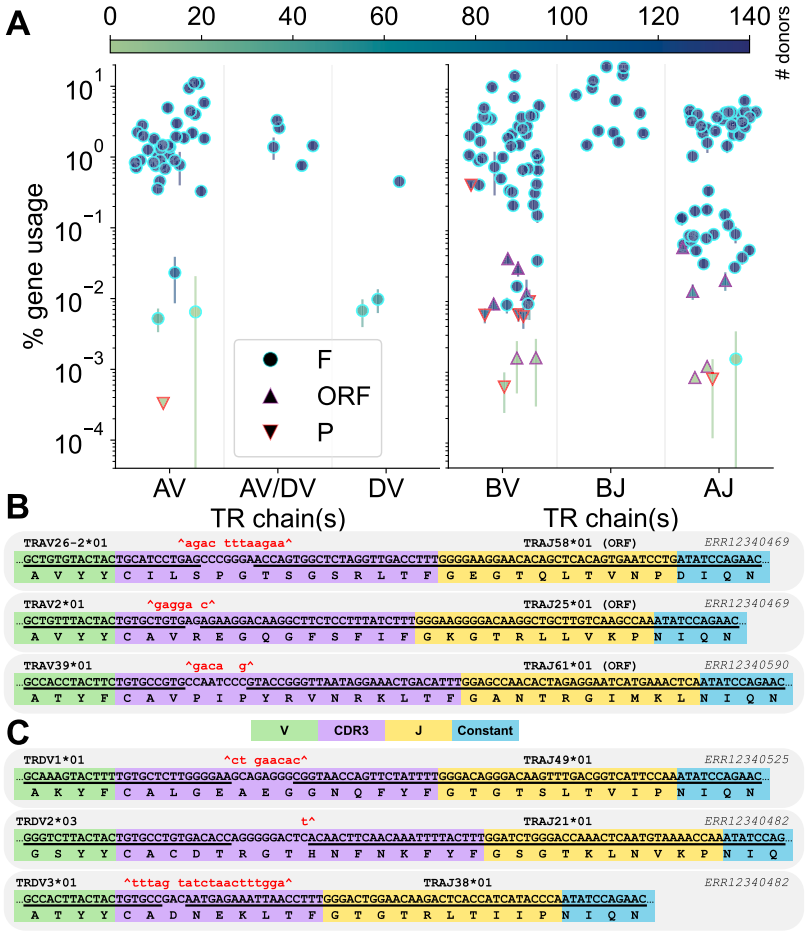

Fig. 3. **A**: TCR gene usage frequencies in the publicly available Mikelov *et al.* TRA and TRB 5′RACE TCRseq dataset [14], highlighting usage of all TRDV genes in TRA rearrangements (**left**), and common, high-frequency ORF and P gene usage in TRBV, TRBJ, and TRAJ (**right**). Paired-end reads were merged, rearrangements annotated, and the percentage of V and J gene usage of productive rearrangements (i.e. those lacking a stop codon between their V- and J-REGIONs, with an in-frame CDR3 junction) were calculated per donor and averaged. Vertical lines indicate 95% confidence intervals, marker edge colour denotes IMGT-reported gene functionality of prototypical (*01) alleles: F = functional (aqua-rimmed circles), ORF = open reading frame (purple-edged up triangles), P = pseudogene (red-outlined down triangles). Marker face colour indicates the number of the 134 donors that gene was featured in (see colour bar scale above). **B**: Examples identified in this dataset of productive rearrangements using TRAJ genes labelled as ORFs. Coloured by TCR region (see colour legend), with underlined nucleotides indicating perfect shared identity with the indicated reference allele, and red text indicating terminal nucleotides present in the germline allele which have been trimmed off during recombination (with the caret character '^' indicating the relative orientation, i.e. ^X = V gene nt, X^ = J gene nt). Originating sample SRA accession numbers are shown in the top right of each example (in the format 'ERRxxxxxxxx'). **C**: As in **B**, but showing examples of each TRDV gene in alpha chain recombinations (recombined with a TRAJ gene, spliced onto the alpha chain constant region).

information. Some TCR analyses – like tracking the presence of specific clonotypes, or comparing global repertoire metrics – likely gain little benefit from a full understanding of the TCR alleles involved, and are agnostic to the TCRs' antigen-specificities. However analyses involving inferring or testing functional activity, or comparing rearrangements between people, should arguably be performed with allelic resolution, as TCR polymorphisms can result in the loss of antigen specificity [48], differential peripheral TCR gene expression [49], and impact surface expression levels [50].

Inspired by Omer and Peres *et al.* [26], which visualised the distance between different TRB alleles assuming amplification with different primer sets, we sought to illustrate the degree of uncertainty one has about the true sequence of a V gene, as a function of how far into it one reads. We therefore measured how confident one can be calling the sequence correctly, having taken increasing length substrings from the 3′ of each V allele. By moving the start site 5′-wards, we mimicked sequencing into the V from a rearrangement. We applied this to a recent IMGT/GENE-DB release (supplemented with novel alleles from the literature as featured in Fig. 2B, conservatively filtering down to those found in >1 study), and measured how confident one can be of identifying the correct sequence at various resolutions from each substring.

Fig. 4 shows the results for all four human TCR loci. As expected (being the rationale upon which most TCRseq to date is predicated), on average the nature of the *gene* from which a given substring derives can be readily determined with relatively short 'reads', reaching >90% confidence when covering only the 3′ end of FR3 for all loci. The ability to correctly determine the exact *allele* however lags behind, with an average confidence ranging from 65–80% when reading halfway into the V, in FR2. That is, one might obtain 150 nt of V region sequence, and still on average only be able to narrow down the allele to one of 3–5. The ability to unambiguously infer alleles (100% confidence) mostly only occurs when sequencing near-complete V regions – and sometimes not even then, as some loci have genes with identical alleles (e.g. TRBV6–2*01 and TRBV6–3*01). Applying these findings to query error rates in existing studies is complicated by issues of read-length and uncertain references, but consider that TRB primers from the BIOMED-2 protocol [51] anneal around CDR1/FR2 (corresponding to an average ~80% confidence in calling the correct allele), while those from Adaptive Biotech [32] bind FR3 (averaging only ~60–70% confidence).

Perhaps more surprising is the observation that the ability to resolve different translated V region polypeptide sequences closely follows the allele (and not gene) confidence average. The amino acid confidence means are slightly higher, but have extremely overlapping confidence intervals for all loci, suggestive of many alleles having unique translations. Indeed, when we checked the translations we observed that on average 80% of human TCR V gene alleles encode unique polypeptides.

Discussion

In this study, we used publicly available datasets to illustrate properties of TCR germline biology that we believe current methodologies do not always accurately consider, which may be limiting the biological accuracy and reproducibility of TCR analysis.

We first showed that the majority of recent TCRseq publications fail to provide sufficient details for readers to unambiguously know which TCR germline reference they used. While some such studies were either generally methodologically sparse or reliant on proxy descriptions (e.g. "conducted as in study X"), many were not. Indeed many had rich technical descriptions, often including versioned and/or dated references for *other* kinds of data, particularly genomes or transcriptomes for RNAseq analysis. The absence of TCR germline reference information is therefore seemingly not a product of authors not recognising the need to properly reference in general, but rather from being unaware that such exacting standards are also appropriate for TCR data.

Using IMGT/GENE-DB as the archetypal TCR germline reference, we demonstrated why it is important to specify the reference used in an analysis: because it changes over time, with alleles and genes appearing, disappearing, or changing sequence between releases. These changes are not an indictment of any particular resource, but an inevitable part of the process of producing a complex product from diverse datasets. We but highlight them to illustrate how a TCR analysis could give different results depending on when the reference was acquired.

We then explored how the metadata tied to a reference, even being built from sound biological principles applied in a logical fashion, can establish assumptions that lead to inappropriate analyses. Specifically, we demonstrate that all TRDV genes can be found used in transcribed human alpha chain rearrangements, and many ORFs and pseudogenes appear at high frequency, perhaps contrary to the received wisdom of the field. We do not seek to imply that frequently detectable genes are necessarily functionally relevant: some will of course never be able to encode a functional polypeptide (e.g. stop-codon containing pseudogenes, whose presence presumably represents a leakiness or failure of the allelic exclusion and nonsense-mediated decay mechanisms which downregulate such transcripts [52,53]). However rearrangements using some of these genes are theoretically able to be expressed and structurally viable; arbitrarily filtering them out may miss relevant biology. Regardless of functionality, omission of these genes from reference sets may result in rearrangements which do use them being misassigned, which may result in an incorrect assumption of functionality or productivity.

Finally we illustrated that the highest confidence about the V alleles used in rearrangements – and therefore their amino acid sequences, which dictates their ability to function and interact with antigen – is only attainable when sequencing the entire variable domain. As TCR research increasingly leads to the development of therapeutics, the need to

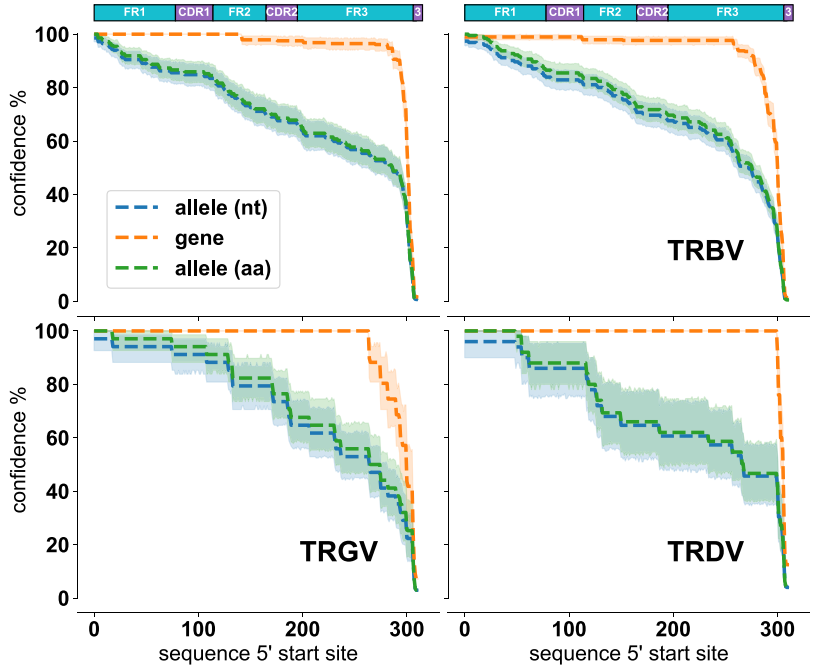

Fig. 4. Ability to distinguish TCR V gene sequence increases based on length of V region sequenced. Increasing length substrings were exhaustively generated from the 3′ end of germline V gene alleles from IMGT/GENE-DB plus novel alleles recently identified in the literature (see Methods), after trimming back to the codon encoding the second conserved cysteine that defines the beginning of the CDR3 (CYS-104), mimicking sequencing into the V region from a recombined junction. Substrings thus generated were then used to search the 3′ ends of all alleles of the corresponding locus (TRA top-left, TRB top-right, TRG bottom-left, TRD bottom-right), to see how well a given theoretical sequence featuring that substring would resolve the correct complete sequence. The x axis shows the nucleotide position (using IMGT gapped numbering) defining the 5′ start site of each substring for each gene, which relates to the different regions of the V genes shown in the schematic above the plots. The y axis shows the percentage 'confidence', calculated as 1 divided by the number of sequences that end in the specified string, x 100. Dashed lines show the mean confidence across all alleles, when looking for sequences at either the full allele level (blue, i.e. dividing by however many trimmed alleles' nucleotide sequences end with a given substring), at the gene level (orange, i.e. dividing by the number of TCR gene symbols represented among all substring matches), or as translated amino acid sequences (green, i.e. dividing by the number of unique translations across the matched allele sequences). Shaded regions indicate 95% confidence intervals. Schematics at the top of each plot show the different V regions, framework region ('FR') and complementarity determining region ('CDR') 1–3, based on their gapped sequence number.

accurately sequence, describe, and report these receptors has never been more crucial.

These analyses primarily considered alleles in the IMGT system, which radically improved TCR and immunoglobulin nomenclature consistency and interpretability after being sanctioned for use by various bodies – particularly the International Union of Immunological Societies (IUIS) – through the late 1990s/early 2000s, replacing multiple competing naming schemes [54]. Pockets of uncertainty regarding the mapping of variable immune gene or allele names to specific sequences however still remains, both within the IMGT system – e.g. the renaming of the Camelus dromedarius IGHV1 gene family to IGHV3 [55] – and without. Even now, two decades later, there are still papers published using deprecated pre-IMGT gene names, including a clonotype described as "AV22S1/AJ45/AC [and] BV7S3/BJ1S4/BC1" last year [56], and multiple publications using gene names containing Greek characters [57–60]. This latter group is presumably due to many V gene-specific reagents retaining older gene identifiers, which might be causing unnecessary confusion. For example, the 3C10 monoclonal antibody binds TRAV1–2, and is therefore used in flow cytometry experiments to help stain for MAIT cells (which use TRAV1–2 in the canonical 'invariant' rearrangement) [61]: however the product pages from three major suppliers refer to its target antigen primarily as 'Vα7.2', and fail to include the proper IUIS name even in the 'also known as' fields [62–64]. Adaptive Biotechnologies, one of largest commercial suppliers of TCRseq services, is similarly noteworthy for producing its own non-standard modified TCR gene nomenclature, adding letters, leading zeroes, and hyphenated gene sub-family indicators, even where none exist (e.g. turning 'TRBV9' into 'TCRBV09–01', or 'TRBJ1–2' into 'TCRBJ01–02', which are just some of many names from this scheme to have been used in recent publications [65,66]). While this system produces more alphabetically-sortable lists, it doesn't map to any external identifiers, and thus creates friction when comparing to or integrating results with datasets using official gene names. While correction of deprecated or custom gene names can be attempted manually with conversion tables [67,68] or automatically with tools like `tidytcells` [69], consistency in the field would likely be served best by modern official nomenclature being used in the first place.

Many of these issues also affect immunoglobulin research (particularly ambiguous allele identifiers), often to a greater extent due to more germline polymorphism and the existence of somatic hypermutation, as discussed in previous work from the Adaptive Immune Receptor Repertoire (AIRR)-Community (AIRR-C) [70,71], who work towards providing methods and standards to improve TCR and immunoglobulin research. It is tempting to speculate that these greater barriers may have encouraged immunoglobulin researchers into better practices, as mitigation strategies (like full-length variable domain sequencing, and accounting for under-sampled allelic diversity) are often addressed. Regardless, the principles promoted in this study should apply across variable receptor loci, just as those established in earlier immunoglobulin papers apply to TCRs.

We expect that the wider field can do more to promote greater accuracy in data reporting and inter-operability. The frequency with which publications analysing 10X single-cell TCRseq data reported germline reference details suggests the Cell Ranger process or documentation may encourage it. Notably users must provide a path to a reference file: not only does this imply that alternatives might exist or be merited, the default example contains all pertinent information ("refdata-cellranger-vdj-GRCh38-alts-ensembl-7.1.0"). Efforts like the Observed T-cell Receptor Space database [72] and VDJBase [73] take a different approach, re-analysing many published studies' data using a consistent reference and pipeline, producing large accessible datasets that are intrinsically controlled for inconsistencies between releases.

The AIRR-C has multiple efforts aiming to improve the accuracy and reproducibility of repertoire research, several relating to germline referencing. Foremost among these is the production of minimal reporting standards for AIRR-seq data (MiAIRR) [74] with detailed schema [75], which already recommend the reporting of both the name and version or date of germline reference used. These have been used in the establishment of the AIRR Data Commons [76], which makes large amounts of AIRR-seq datasets annotated under MiARR standards available through platforms like iReceptor [30] and VDJServer [77]. In recent years we have produced germline reference set standards and guidelines, as well as immunoglobulin reference sets for human and mice [70,71], focusing on high-quality alleles with transparent histories and full-length sequence support. The AIRR-C Germline Database Working Group (GLDB-WG) and Inferred Allele Review Committee (IARC) are currently using these principles and schema to similarly generate new human TCR germline reference sets, which we hope will support researchers in their pursuit of accurate, repeatable, and reproducible TCR research.

All of these sets are made freely available from the Open Germline Receptor Database (OGRDB) [78] under a minimally-restrictive CC0 1.0 Universal license. This means that the same resource can be used by researchers across academia and industry, encouraging the development and adoption of interoperable tools and resources (unlike e.g. IMGT/GENE-DB, which uses a CC BY-NC-ND 4.0 license, requiring a financial arrangement for commercial use). We hope that this will incentivise companies to avoid the use of custom private databases (which offer little transparency or traceability to end users), or the provision of inappropriate references (e.g. the use of GRCh38 derivatives to annotate diverse human repertoires, simply because it is free to do so).

In summary, germline references – and the nomenclatures with which they are constructed – are the bedrock upon which TCR analyses are built. They provide context to the rearranged receptors which are the focus of our experiments, yet in many cases we simply do not know which references were used. Even when known, their idiosyncrasies may not be well understood, leading to either inappropriately- or suboptimally-analysed results. We hope that the vignettes presented here will help researchers better understand the importance of their germline sets, aiding them to produce and communicate ever more accurate and reproducible data, and useful and innovative tools.

Germline reference recommendations

For TCR researchers

- Consider how the scope and methodology of the reference will help address the biological questions being posed.
- When requiring greater precision (e.g. functional TCR validation or gene inference), consider sequencing complete V regions.
- Use a reference containing a suitable breadth of alleles, and tools capable of resolving them.
- Consider inferring personalised TCR genotypes, permitting annotation with the most accurate possible reference.
- Where possible, bank a version of the reference (potentially alongside analysed data), to ensure all necessary information is retained and analyses are repeatable.
- When publishing, report the reference name, date of access, and release or version number in the methods section, citing the relevant paper and/or resource.
- If using custom or curated germlines, ensure the process of reference generation is suitably reported.

For tool, product, and germline reference developers

- Include all TRDV alleles in your TRAV gene set, and make them detectable in alpha chain rearrangements.
- Don't omit ORF or P genes on the presumption that they will not be either used or relevant in the repertoire.

- Where licensing allows, make all historical germline reference sets available for download, sensibly and unambiguously identifiable, numbered following a documented nomenclature.
- Provide sufficient documentation, detailing not only how to use the provided reference(s), but how to report and cite them in publications.
- If possible, allow users to supply their own reference when annotating reads, to allow for up-to-date, patient-specific, or custom analyses.
- Ensure correct, IUIS-approved TCR nomenclature is used in both product descriptions and tool outputs. For gene-specific reagents, transparency can be further supported by providing citations or links to externally-hosted gene-specific summary resources.

For journals

- Encourage or require authors to unambiguously report the TCR germline reference used (i.e. name, version and/or release, and access date).

Methods

Code and data availability

All analyses were performed in Python (3.12.0), using the standard library plus: `pandas` [79] (2.1.3), `matplotlib` [80] (3.8.2), `seaborn` [81] (0.13.2), `numpy` [82] (1.26.2), `scipy` [83] (1.11.4), and `receptor_utils` (0.0.47, https://github.com/williamdlees/receptor_utils/). All code is available at http://github.com/JamieHeather/tcr-germline-paper. Other packages and publicly available datasets are referred to below.

TCRseq germline referencing systematic literature review

Papers were identified from PubMed, searching for those with free texts available, published between 2023 and the search date (2025-03-20), with these terms:

```
(Search: (("t cell receptor" [All Fields]) OR ("tcr" [All Fields])) OR (t-cell receptor) AND (("tcrseq" [All Fields]) OR ("tcr sequencing" [All Fields])) OR ("tcr repertoire" [All Fields])) NOT ("review" [Publication Type])
```

Papers were then manually screened to remove pre-prints, papers which did not perform TCRseq, or were re-analysing/re-using published datasets (to prevent duplicate contributions). Relevant TCRseq details were then extracted, recording the loci and species investigated, and TCRseq and analysis protocol details, in particular the germline reference details at several resolutions (source organisation, named database/resource, and specific version/release/date of access), using the Covidence systematic review software (Veritas Health Innovation, http://www.covidence.org). Details were recorded per 'experimental setup', comprising the combination of species and protocol used. The same publication may therefore contribute multiple setups if they performed e.g. bulk and single-cell TCRseq, or analysed multiple loci/species. The alluvial plot was constructed using the `pyalluvial` (version 0.0.0) package. For several fields, values present in fewer than ten setups were given a shared 'Other' field in the table for readability (with original values retained in columns suffixed 'all', available in the CSV file in the Github repository linked in the availability section).

Tracking historical changes in IMGT/GENE-DB

We used IMGT/GENE-DB [15] as the exemplar reference with which to demonstrate resource evolution. While this is available for download (from https://www.imgt.org/download/GENE-DB/ by non-profit/academic organisations, under a CC BY-NC-ND 4.0 license), to the best of our knowledge historical releases are not accessible. In March of 2023 we began a regular process of automatically downloading and banking for comparison. Pre-2023 releases were obtained either from having coincidentally banked earlier releases for contemporaneous analysis, or were downloaded from the Internet Archive's Wayback Machine (https://web.archive.org/), an internet scraping archival platform. While this collection is extremely incomplete (particularly before 2023, with the earliest Wayback Machine record being from 2013, ~9 years after IMGT/GENE-DB's publication), it provides a sampling of the reference resources available to researchers. The 'IMGTGENEDB-ReferenceSequences.fasta-nt-WithGaps-F+ORF+inframeP' FASTA files per release were parsed, and gene/allele identifiers for the specified species recorded and compared between timepoints. This aggregated resource is located at https://github.com/JamieHeather/genedb-releases (accessed for this analysis on 2025-04-06).

Novel alleles were gathered from published reports [14,25–29] (either inferred from V region-spanning TCRseq datasets or covered in long read genomic sequencing). These are tabulated and automatically compared against IMGT/GENE-DB, to ensure that alleles that were later added to that resource aren't counted among both sets. This collection is maintained at https://github.com/JamieHeather/novel-tcr-alleles/ (accessed on 2025-04-06).

Plotting TCR V/J gene usage

We downloaded the Mikelov *et al.* [14] dataset of 134 donors' combined TRA/TRB TCR repertoires, accessible from EBI (accession E-MTAB-13593). This is a 5′ rapid amplification of cDNA ends (5′ RACE) protocol, reverse transcribed from the 5′ of the constant region, and MiSeq 2x300 bp paired-end sequenced, such that many reads can be overlapped giving full V coverage. FASTQ files were downloaded, and overlapping read 1 and 2 files merged using `FLASH` [84] (1.2.11), and TCRs annotated using the version 0.2.7 release of `autoDCR` (available from https://github.com/JamieHeather/autoDCR), a modified version of Decombinator [13,21] capable of resolving TCR alleles (as described in Heather *et al.* [27]) with default parameters. This used a filtered version of IMGT/GENE-DB (release 202410–7, downloaded on 2024-03-13) as a reference, containing just human TRA and TRB alleles supplemented with all TRDV genes. Unique V/J/CDR3 descriptors were then counted, and the mean frequency with which each TCR gene (inclusive of all alleles) was calculated across all donor samples.

Plotting TCR V gene confidence

The most recent IMGT/GENE-DB release as of the analysis (release 202514–7, downloaded on 2025-04-06) was used, supplemented with novel alleles as described above (filtered on those observed in ≥2 studies). Gapped sequences (where gaps are introduced to align V regions with differing numbers of residues, per the IMGT unique numbering system [85]) were generated for each sequence using `receptor_utils` modules: `extract_refs` applied to the 'IMGTGENEDB-ReferenceSequences.fasta-nt-WithGaps-F+ORF+inframeP' FASTA file for this release established locus-specific references, which were then combined into one pan-locus reference, which was used with `gap_sequences` to generate gapped versions of each allele. The gapped sequences were then used to trim each sequence back to the nucleotides encoding the second conserved cysteine residue which defines the start of CDR3 (CYS-104), discarding those that were too short (due to only partial sequence coverage in the reference).

All alleles within a locus were then iterated over, and starting from the 3′-most nucleotide, substrings of increasing length incrementing by 1 nt were banked. Each iteration, gaps (indicated with a '.') were removed from the substrings, which were then used to find all CYS-104-trimmed alleles containing an exact match. Percentage confidence was calculated as $1 \div num\_matches \times 100$. Gene- and amino acid-level confidence is calculated similarly, but instead of dividing by the number of *alleles*

matched, it uses the number of unique TCR *gene identifiers*, or translated *amino acid sequences* of matched alleles, respectively. Framework and complementarity-determining regions were labelled according to IMGT numbering.

CRediT authorship contribution statement

James M. Heather: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Ayelet Peres:** Writing – review & editing, Methodology, Conceptualization. **Gur Yaari:** Writing – review & editing, Methodology, Conceptualization. **William Lees:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

GY is the Editor-in-Chief of Immunoinformatics, but had no involvement in the peer review or editorial decision-making for this manuscript. He also serves as an advisor to Clareo Biosciences Inc. and holds equity in the company. WL is an officer and shareholder of Clareo Biosciences Inc. The other authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Acknowledgements

The authors would like to thank the current and past members of two Adaptive Immune Receptor Repertoire Community (AIRR-C) affiliated groups for formative discussions on and around this topic: the Inferred Allele Review Committee (IARC) and the Germline Database (GLDB) Working Group. JMH would also like to thank autocorrect for (repeatedly) providing him with the inspiration for this manuscript's title. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2025.100058.

Data availability

Most data analysed are publicly available. New data and all code are available in the referenced Github repositories.

References

[1] Alt FW, et al. VDJ recombination. Immunol Today 1992;13:306–14.
[2] Dupic T, Marcou Q, Walczak AM, Mora T. Genesis of the αβ T-cell receptor. PLoS Comput Biol 2019;15:e1006874.
[3] Zaslavsky ME, et al. Disease diagnostics using machine learning of B cell and T cell receptor sequences. Science 2025;387:eadp2407.
[4] Wu D, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med 2012;4.
[5] Klebanoff CA, Chandran SS, Baker BM, Quezada SA, Ribas A. T cell receptor therapeutics: immunological targeting of the intracellular cancer proteome. Nat Rev Drug Discov 2023;22:996–1017.
[6] Six A, et al. The past, present, and future of Immune repertoire biology – The rise of next-generation repertoire analysis. Front Immunol 2013;4.
[7] Rosati E, et al. Overview of methodologies for T-cell receptor repertoire analysis. BMC Biotechnol 2017;17:61.
[8] Heather JM, Ismail M, Oakes T, Chain B. High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. Brief Bioinform 2017;19:554–65.
[9] Valkiers S, et al. Recent advances in T-cell receptor repertoire analysis: bridging the gap with multimodal single-cell RNA sequencing. ImmunoInformatics 2022;5:100009.
[10] Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res 2009;19:1817–24.
[11] Murugan A, Mora T, Walczak AM, Callan CG. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. Proc Natl Acad Sci 2012;109:16161–6.
[12] Gerritsen B, Pandit A, Andeweg AC, De Boer RJ. R.TCR: a pipeline for complete and accurate recovery of T cell repertoires from high throughput sequencing data. Bioinformatics 2016;32:3098–106.
[13] Thomas N, Heather J, Ndifon W, Shawe-Taylor J, Chain B. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. Bioinformatics 2013;29:542–50.
[14] Mikelov A, et al. Ultrasensitive allele inference from immune repertoire sequencing data with MiXCR. Genome Res 2024;34:2293–303.
[15] Giudicelli V. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 2004;33:D256–61.
[16] Giudicelli V. IMGT/LIGM-DB, the IMGT(R) comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res 2006;34:D781–4.
[17] Giudicelli V, Brochet X, Lefranc M-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. Cold Spring Harb Protoc 2011:695–715. 2011.
[18] Li S, et al. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. Nat Commun 2013;4:2333.
[19] Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 2013;41:W34–40.
[20] Song L, et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat Methods 2021;18:627–30.
[21] Peacock T, Heather JM, Ronel T, Chain B. Decombinator V4: an improved AIRR compliant-software package for T-cell receptor sequence annotation. Bioinformatics 2021;37:876–8.
[22] Schneider VA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 2017;27(5):849–64.
[23] Sarsani VK, et al. The genome of C57BL/6J "Eve", the mother of the laboratory mouse genome reference strain. G3: Genes|Genomes|Genetics 2019;9:1795–805.
[24] Peng K, et al. Diversity in immunogenomics: the value and the challenge. Nat Methods 2021;18:588–91.
[25] Rodriguez OL, Silver CA, Shields K, Smith ML, Watson CT. T.argeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor alpha, delta, and beta loci. Cell Genomics 2022;2:100228.
[26] Omer A, et al. T cell receptor beta germline variability is revealed by inference from repertoire data. Genome Med 2022;14:2.
[27] Heather JM, et al. Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. Nucleic Acids Res 2022;50(12):e68.
[28] Lin M-J, et al. Profiling genes encoding the adaptive immune receptor repertoire with gAIRR Suite. Front Immunol 2022;13:922513.
[29] Corcoran M, et al. Archaic humans have contributed to large-scale variation in modern human T cell receptor genes. Immunity 2023;56:635–52. e6.
[30] Corrie BD, et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev 2018;284:24–41.
[31] Chen G, et al. Sequence and structural analyses reveal distinct and highly diverse human CD8+ TCR repertoires to immunodominant viral antigens. Cell Rep 2017;19:569–83.
[32] Robins HS, et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. Blood 2009;114:4099–107.
[33] Wang C, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. Proc Natl Acad Sci 2010;107:1518–23.
[34] Hudson WH, Sudmeier LJ. Localization of T cell clonotypes using the Visium spatial transcriptomics platform. STAR Protoc 2022;3:101391.
[35] Kitaura K, Shini T, Matsutani T, Suzuki R. A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) α and β repertoires and identifying potential new invariant TCR α chains. BMC Immunol 2016;17:38.
[36] Shi B, et al. Compositional characteristics of human peripheral TRBV pseudogene rearrangements. Sci Rep 2018;8:5926.
[37] Miossec C, et al. Further analysis of the T cell receptor gamma/delta+ peripheral lymphocyte subset. The V delta 1 gene segment is expressed with either C alpha or C delta. J Exp Med 1990;171:1171–88.
[38] Miossec C, et al. Molecular characterization of human T cell receptor α chains including a Vδ 1-encoded variable segment. Eur J Immunol 1991;21:1061–4.
[39] Castelli C, Mazzocchi A, Salvi S, Anichini A, Sensi M. Use of the Vδ1 variable region in the functional T-cell receptor α chain of a WT31+ cytotoxic T lymphocyte clone which specifically recognizes HLA-A2 molecule. Scand J Immunol 1992;35:487–94.
[40] Bank I, et al. Expansion of a unique subpopulation of cytotoxic T cells that express a C alpha V delta 1 T-cell receptor gene in a patient with severe persistent neutropenia. Blood 1992;80:3157–63.
[41] Bank I, et al. Aberrant T-cell receptor signalling of interferon-γ- and tumour necrosis factor-α-producing cytotoxic CD8+ Vδ1/Vβ16 T cells in a patient with chronic neutropenia. Scand J Immunol 2003;58:89–98.
[42] Ueno T, Tomiyama H, Fujiwara M, Oka S, Takiguchi M. HLA class I-restricted recognition of an HIV-derived epitope peptide by a human T cell receptor α chain having a Vδ1 variable segment. Eur J Immunol 2003;33:2910–6.
[43] Sun X, et al. Unbiased analysis of TCRα/β chains at the single-cell level in Human CD8+ T-cell subsets. PLoS ONE 2012;7:e40386.
[44] Sun X, et al. Superimposed epitopes restricted by the same HLA molecule drive distinct HIV-specific CD8+ T cell repertoires. J Immunol 2014;193:77–84.

[45] Liu P, et al. Characterization of human αβTCR repertoire and discovery of D-D fusion in TCRβ chains. Protein Cell 2014;5:603–15.
[46] Volkmar M, et al. Identification of TRDV-TRAJ V domains in human and mouse T-cell receptor repertoires. Front Immunol 2023;14:1286688.
[47] Dominique Scaviner, Joumana Jabado-Michaloud, Géraldine Folch, & Viviane Nguefack Ngoune. IMGT repertoire: gene table: human (Homo sapiens) TRAJ, 2020. https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=human&group=TRAJ, accessed on 2025-05-27.
[48] Gras S, et al. Allelic polymorphism in the T cell receptor and its impact on immune responses. J Exp Med 2010;207:1555–67.
[49] Brennan RM, et al. Missense single nucleotide polymorphisms in the human T cell receptor loci control variable gene usage in the T cell repertoire. Br J Haematol 2014;166:148–52.
[50] Thomas S, et al. Framework engineering to produce dominant T cell receptors with enhanced antigen-specific function. Nat Commun 2019;10:4451.
[51] Van Dongen JJM, et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. Leukemia 2003;17:2257–317.
[52] Brady BL, Steinel NC, Bassing CH. Antigen receptor allelic exclusion: an update and reappraisal. J Immunol 2010;185:3801–8.
[53] Weischenfeldt J, et al. NMD is essential for hematopoietic stem and progenitor cells and for eliminating by-products of programmed DNA rearrangements. Genes Dev 2008;22:1381–96.
[54] Lefranc M-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors report. Immunogenetics 2007;59:899–902.
[55] Joumana Jaabado-Michaloud & Nathalie Bosc. IMGT Repertoire: gene table: arabian camel (Camelus dromedarius) IGHV. https://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=arabian_camel&group=IGHV, accessed on 2025-05-27.
[56] Barisic S, et al. Regression of renal cell carcinoma by T cell receptor-engineered T cells targeting a human endogenous retrovirus. J Immunother Cancer 2024;12:e009147.
[57] Peruzzi B, Bencini S, Caporale R. TCR Vβ Evaluation by Flow Cytometry. Methods Mol Biol 2021;2285:99–109. https://doi.org/10.1007/978-1-0716-1311-5_8.
[58] Andrlová H, et al. MAIT and Vδ2 unconventional T cells are supported by a diverse intestinal microbiome and correlate with favorable patient outcome after allogeneic HCT. Sci Transl Med 2022;14. eabj2829.
[59] Pangli BK, Braddock SR, Knutsen AP. O.menn syndrome in a 10-month-old male with athymia and VACTERL association. J Allergy Clin Immunol Glob 2023;2:100153.
[60] Giorgetti OB, Haas-Assenbaum A, Boehm T. Probing TCR specificity using artificial In vivo diversification of CDR3 regions. Eur J Immunol 2025;55. e202451434.
[61] Held K, Beltrán E, Moser M, Hohlfeld R, Dornmair K. T-cell receptor repertoire of human peripheral CD161hiTRAV1-2+ MAIT cells revealed by next generation sequencing and single cell analysis. Hum Immunol 2015;76:607–14.
[62] Beckman Coulter. Inc. TCR Vα7.2 antibodies - Beckman Coulter Life sciences. https://www.beckman.com/reagents/coulter-flow-cytometry/antibodies-and-kits/single-color-antibodies/tcr-va-7-2,; accessed on 2025-05-27.
[63] BioLegend, Inc. Purified anti-human TCR Valpha7.2 antibody anti-TCR V alpha7.2 - 3C10. https://www.biolegend.com/en-gb/products/purified-anti-human-tcr-valpha7-2-antibody-7122?GroupID=BLG9339, accessed on 2025-05-27.
[64] Miltenyi Biotec. TCR Vα7.2 antibody, anti-human, REAfinityTM | Miltenyi Biotec. https://www.miltenyibiotec.com/UN-en/products/tcr-va7-2-antibody-anti-human-reafinity-rea179.html, accessed on 2025-05-27.
[65] Miller D, et al. Immunosequencing and profiling of T cells at the maternal–Fetal interface of women with preterm labor and chronic chorioamnionitis. J Immunol 2023;211:1082–98.
[66] Sigdel TK, et al. Perturbations of the T-cell immune repertoire in kidney transplant rejection. Front Immunol 2022;13:1012042.
[67] Arden B, Clark Stephen P, Kabelitz D, Mak Tak W. Human T-cell receptor variable gene segment families. Immunogenetics 1995;42.
[68] Lefranc M. Nomenclature of the Human T cell receptor genes. Curr Protoc Immunol 2000;40.
[69] Nagano Y, Chain B. tidytcells: standardizer for TR/MH nomenclature. Front Immunol 2023;14:1276106.
[70] Lees WD, et al. AIRR community curation and standardised representation for immunoglobulin and T cell receptor germline sets. ImmunoInformatics 2023;10:100025.
[71] Collins AM, et al. AIRR-C IG Reference Sets: curated sets of immunoglobulin heavy and light chain germline genes. Front Immunol 2024;14:1330153.
[72] Raybould MIJ, et al. The Observed T Cell Receptor Space database enables paired-chain repertoire mining, coherence analysis, and language modeling. Cell Rep 2024;43:114704.
[73] Omer A, et al. VDJbase: an adaptive immune receptor genotype and haplotype database. Nucleic Acids Res 2020;48:D1051–6.
[74] The AIRR Community. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol 2017;18:1274–8.
[75] Vander Heiden JA, et al. AIRR community standardized representations for annotated immune repertoires. Front Immunol 2018;9:2206.
[76] Christley S, et al. The ADC API: a web API for the programmatic query of the AIRR data commons. Front Big Data 2020;3:22.
[77] Christley S, et al. VDJServer: a cloud-based analysis portal and data commons for Immune repertoire sequences and rearrangements. Front Immunol 2018;9.
[78] Lees W, et al. OGRDB: a reference database of inferred immune receptor genes. Nucleic Acids Res 2020;48:D964–70.
[79] McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python In Science Conference 56–61; 2010. https://doi.org/10.25080/Majora-92bf1922-00a.
[80] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–5.
[81] Waskom M. seaborn: statistical data visualization. J Open Source Sci 2021;6:3021.
[82] Harris CR, et al. Array programming with NumPy. Nature 2020;585:357–62.
[83] SciPy. 1.0 Contributors et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17:261–72.
[84] Magoč T, Salzberg SL. F.LASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 2011;27:2957–63.
[85] Lefranc M-P, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 2003;27:55–77.