frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 09 February 2024
DOI 10.3389/fimmu.2023.1330153

OPEN ACCESS

EDITED BY
Amy L Kenter,
University of Illinois Chicago, United States

REVIEWED BY
Jean-Philippe Bürckert,
Independent Researcher, Boston, MA, United States
Michel Cogne,
University of Rennes 1, France

*CORRESPONDENCE
Andrew M. Collins
a.collins@unsw.edu.au
William D. Lees
william@lees.org.uk

†The list of endorsing members of the AIRR Community is provided as a supplementary document (Supplementary Data Sheet 1)

RECEIVED 30 October 2023
ACCEPTED 27 December 2023
PUBLISHED 09 February 2024

CITATION
Collins AM, Ohlin M, Corcoran M, Heather JM, Ralph D, Law M, Martínez-Barnetche J, Ye J, Richardson E, Gibson WS, Rodriguez OL, Peres A, Yaari G, Watson CT and Lees WD (2024) AIRR-C IG Reference Sets: curated sets of immunoglobulin heavy and light chain germline genes.
Front. Immunol. 14:1330153.
doi: 10.3389/fimmu.2023.1330153

AIRR-C IG Reference Sets: curated sets of immunoglobulin heavy and light chain germline genes

Andrew M. Collins[1*], Mats Ohlin[2], Martin Corcoran[3], James M. Heather[4,5], Duncan Ralph[6], Mansun Law[7], Jesus Martínez-Barnetche[8], Jian Ye[9], Eve Richardson[10], William S. Gibson[11], Oscar L. Rodriguez[11], Ayelet Peres[12], Gur Yaari[12], Corey T. Watson[11], William D. Lees[13,14*] and for The AIRR-Community[†]

[1]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia, [2]Department of Immunotechnology, and SciLifeLab, Lund University, Lund, Sweden, [3]Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden, [4]Mass General Cancer Center, Massachusetts General Hospital, Charlestown, MA, United States, [5]Department of Medicine, Harvard Medical School, Boston, MA, United States, [6]Fred Hutchinson Cancer Research Center, Seattle, WA, United States, [7]Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, United States, [8]Centro de Investigación Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, Cuernavaca, Morelos, Mexico, [9]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, [10]La Jolla Institute for Immunology, San Diego, CA, United States, [11]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Louisville, Louisville, KY, United States, [12]Bioengineering Program, Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, [13]Institute of Structural and Molecular Biology, Birkbeck College, London, United Kingdom, [14]Human-Centered Computing and Information Science, Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal

Introduction: Analysis of an individual's immunoglobulin (IG) gene repertoire requires the use of high-quality germline gene reference sets. When sets only contain alleles supported by strong evidence, AIRR sequencing (AIRR-seq) data analysis is more accurate and studies of the evolution of IG genes, their allelic variants and the expressed immune repertoire is therefore facilitated.

Methods: The Adaptive Immune Receptor Repertoire Community (AIRR-C) IG Reference Sets have been developed by including only human IG heavy and light chain alleles that have been confirmed by evidence from multiple high-quality sources. To further improve AIRR-seq analysis, some alleles have been extended to deal with short 3' or 5' truncations that can lead them to be overlooked by alignment utilities. To avoid other challenges for analysis programs, exact paralogs (e.g. IGHV1-69*01 and IGHV1-69D*01) are only represented once in each set, though alternative sequence names are noted in accompanying metadata.

Results and discussion: The Reference Sets include less than half the previously recognised IG alleles (e.g. just 198 IGHV sequences), and also include a number of novel alleles: 8 IGHV alleles, 2 IGKV alleles and 5 IGLV alleles. Despite their

10.3389/fimmu.2023.1330153

smaller sizes, erroneous calls were eliminated, and excellent coverage was achieved when a set of repertoires comprising over 4 million V(D)J rearrangements from 99 individuals were analyzed using the Sets. The version-tracked AIRR-C IG Reference Sets are freely available at the OGRDB website (https://ogrdb.airr-community.org/germline_sets/Human) and will be regularly updated to include newly observed and previously reported sequences that can be confirmed by new high-quality data.

KEYWORDS

immunoglobulin, heavy chain, light chain, IGHV genes, IGHD, IGHJ

Introduction

Cellular and humoral immune responses are key components of our defence against external threats from a range of pathogens (1, 2), as well as internal threats like oncogenic transformation and cancer (3). Repertoire analysis of the B and T cell transcriptome is now a part of many studies of the immune response, and such analyses have deepened our understanding of immune protection against, for example, Type 1 Human Immunodeficiency Virus (4), influenza viruses (5), and SARS-CoV-2 (6). Such analyses can guide the development of novel vaccine strategies (7), and can lead to the identification of highly functional antibodies with the potential to be translated into anti-viral drugs of clinical utility (8–10).

An important aspect of IG repertoire analysis is the identification of the germline V, D and J genes that have contributed to each rearranged V(D)J gene sequence. Germline gene reference sets - made up of known IG allelic variants - are critical for these kinds of analyses. Knowledge of germline gene sequences has accumulated very slowly over time, since their first reporting in 1980 (11, 12). It was only in the late 1990s that mapping of the complete human IGH locus allowed the allelic relationships between reported gene sequences to be gradually defined (13, 14). The resulting collections of genes and allelic variants allowed proper analysis of the human antibody repertoire to begin.

Comprehensive reference sets allow the accurate identification of the germline genes that contribute to the formation of particular V(D)J gene sequences. This in turn allows the clonal relationships between different sequences to be identified (15–18). Analysis of V(D)J rearrangements also allows somatic point mutations within the gene rearrangements to be determined, giving insights into the development of antibody specificities and affinity-driven selection (19–21) as well as antibody isotype functions (22–24). If these kinds of analyses are to be improved, the available reference sets must also be improved.

Previously reported reference sets have been compiled from sequences that have been reported over several decades. Over this time, as sequencing technology has changed, the risks of sequencing errors have also changed. Today errors may arise, for example, from the annotation of poor genome assemblies (25). In the past, sequencing itself was so error-prone that it is likely that many sequences reported in the 1980s and 1990s included sequencing errors (26).

Rigorous studies are now transforming our knowledge of genetic variation in the human IG loci. Significant structural variation including gene duplications has recently been reported in the IGH locus (27), and in the light chain IGL and IGK loci (28, 29). These and other studies over the last decade have also substantially increased the number of accurately reported germline sequences (30–34). Together this has likely led to the identification of most common allelic variants that are found in well-studied populations, but a large number of variants from less-studied populations probably remain to be found (28, 35, 36). Rare alleles remain to be documented in all populations (31), and population coverage must be increased (37). Fortunately, next-generation sequencing of the IG loci should soon provide this population coverage, but challenges associated with new sequencing technologies mean that IG gene discovery must always be approached with great care (25, 38).

In recent years most reports of newly-discovered alleles have come from the analysis of IG and T-cell receptor (TR) gene repertoire sequencing data (AIRR-seq data). Since 2018, 34 inferred IGHV allelic variants, as well as 6 IGLV and 3 IGKV variants, have been validated and assigned temporary names (e.g. IGHV1-69*i04) by the Inferred Allele Review Committee (IARC) of the Adaptive Immune Receptor Repertoire Community (AIRR-C) (39). The IARC is now one of two Review Committees operating at the direction of the T-cell receptor (TR) and Immunoglobulin (IG) Nomenclature Sub-committee of the International Union of Immunological Societies' (IUIS') Nomenclature Committee. The second review committee, the TR-IG Nomenclature Review Committee (TR-IG NRC), is responsible for the official naming of IG genes. To date, 33 of the 43 inferred sequences affirmed by the IARC have been officially named by IUIS after referral from the IARC (40). The nomenclature used is referred to here as the IUIS nomenclature to reflect IUIS responsibility for naming. The

nomenclature was developed by Marie-Paule Lefranc, the founder of the ImMunoGeneTics (IMGT) group, and the former chair of the IUIS IG, TR and MH Nomenclature Subcommittee (41–43). Other sequences that have not yet been assigned official names are referred to here by their temporary IARC names, or by unofficial names indicating their differences from other named sequences (e.g. IGLV3-25*03_G74C).

Although the inferred IG sequences have not been mapped to any reported human genome assembly, their temporary names are modelled on the official IUIS human IG nomenclature – a positional nomenclature. The general principle of the IARC naming process has been that when a putative variant sequence aligns to a known allele of a single gene with high sequence identity, and with substantially lower identity to the alleles of other genes, the sequence can reasonably be assigned as a variant of that gene. In many instances, however, an assignment cannot be made with high certainty based on sequence identity. There may be several similar genes that could all be assigned as the germline gene of the novel allele. The IARC continues to assign temporary names, but their most recent reports make clear that the genomic locations of some new alleles must be considered uncertain.

Despite the growth of historical reference sets in recent years (44, 45), many nucleotide and structural level polymorphisms (large insertions, duplications and deletions) remain to be identified in the human population. The ideal reference set for the study of an individual's immune response is therefore a personalized one, determined by genomic sequencing prior to V(D)J sequence annotation. Alternatively, a personalized repertoire developed from AIRR-seq data (expressed V(D)J gene rearrangements) using inference technology can be used for high-quality sequence annotation (15, 46). In practice, most repertoire studies rely on standard publicly available reference sets for their analyses (38, 44, 45), and even the inference of a personalized repertoire begins with analysis using a standard set.

Sufficient published research now exists to create human reference sets in which each sequence is validated by multiple lines of evidence, providing confidence that erroneous sequences have been excluded. In this report we describe how well documented alleles that were reported in the past have been combined with newly discovered IGH and IG light chain (lambda, IGL; kappa, IGK) alleles to produce the high-quality AIRR-C IGH_VDJ, IGKappa_VJ and IGLambda_VJ Reference Sets. Together, these three reference sets are referred to here as the AIRR-C IG Reference Sets.

The AIRR-C IG Reference Sets include a small number of sequences that have artificial extensions, for some reported germline sequences appear to have short truncations, usually at their 3' ends. These truncations usually arose because of mistakes, in early studies, identifying the boundary between the end of an IGHV gene and the highly similar first nucleotides of its downstream Recombination Signal Sequence (RSS). Extensions of these truncated sequences in the AIRR-C IG Reference Sets will help ensure that they are not overlooked by sequence alignment utilities, and the presence of extensions is noted in their sequence metadata. The sets also include modifications to deal with issues arising with sequences that appear in the genome as exact paralogs (duplicates). In AIRR-seq analysis, the presence of identical sequences in the reference set can cause downstream analytical pipelines to overlook both copies of the sequence. A single copy of each sequence pair is therefore retained in the AIRR-C IG Reference Sets with the missing sequence being recorded as a paralog in the sequence metadata. A script is available at the OGRDB website that facilitates the production of reference sets that have the details of exact paralogs in the FASTA header (https://airr-community.github.io/receptor-germline-tools/_build/html/introduction.html).

The AIRR-C IG Reference Sets are published under the FAIR guiding principles for scientific data management (47), using a recently described schema (40), with a minimally restricted licence. The sets will grow over time by the inclusion of newly identified allelic variants, as well as by the inclusion of previously reported sequences whose existence may come to be confirmed from the accumulation of new evidence. The nature of these changes to the sets and evidence in support of those changes will be clearly documented at the OGRDB website (https://ogrdb.airr-community.org) (48). The sets will be strictly versioned, and versioned sets will be referenceable via Digital Object Identifiers (DOIs). This will mean that proper documentation in the literature of the use of the reference sets will allow readers to easily understand the germline gene data that has been used in an analysis.

Method

Evaluation of the effects of short 3' truncations of reference set sequences upon AIRR-seq analysis

The way that AIRR-seq analysis is influenced by changes in the lengths of sequences in reference sets was investigated using transcriptome data sets of project PRJEB26509 from the European Nucleotide Archive (43), generated from FACS-sorted (CD19+ CD27- IgD+ IgA- IgG-) naïve B cells. These datasets can be expected to express high frequencies of unmutated IG-encoding V(D)J genes. Six samples were selected for analysis based upon their carriage of different alleles of the IGHV4-38-2 gene. (A different analysis elsewhere in this study involves the complete PRJEB26509 dataset, where it is referred to as the Gidoni-VDJbase dataset.) IG heavy chain-encoding reads were assembled using PEAR v0.9.6. Annotation was performed within the IgDiscover v0.15.1 framework using IgBLAST (version 1.17). Filtering was performed using the default IgDiscover settings that removed reads with no J gene assigned, with stop codons, with V gene coverage <90%, with J gene coverage <60%, and/or with V gene E-value >1E-3. Reference sets including an archived IMGT IGHV Reference Set (accessed January 2020) and a modified set that incorporated an extension of IGHV4-38-2*01 by two nucleotides to resolve a truncation of its 3'-end. The addition of the -GA extension (changing the 3' end from -TGCGAGA to -TGCGAGAGA) was based upon the genomic sequence ON052084.

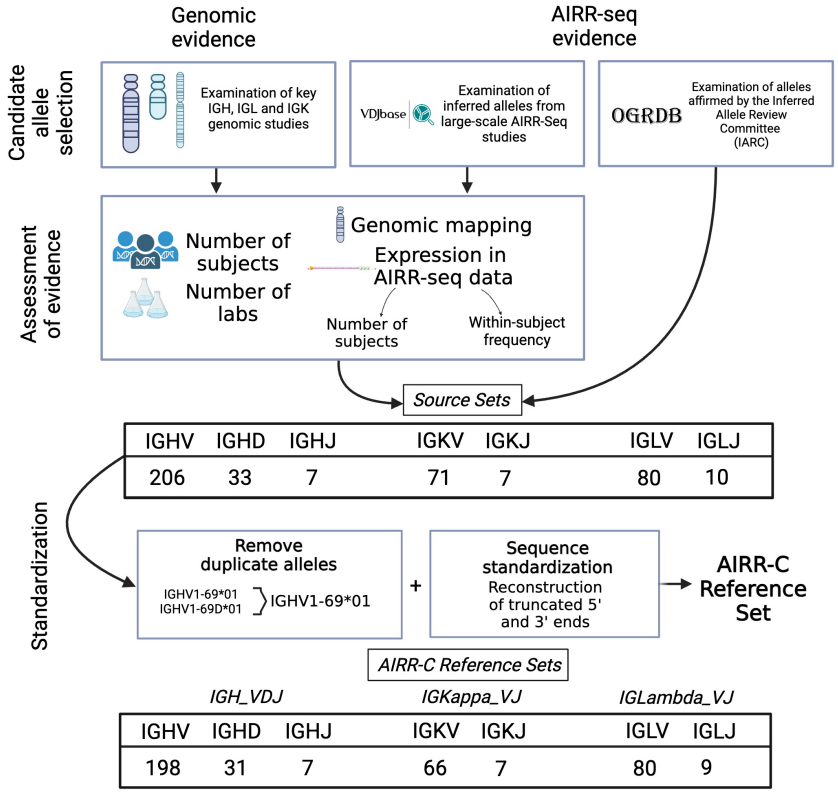

FIGURE 1
The compilation of Reference Sets began with the identification of candidate sequences from a limited number of key genomic and AIRR-seq studies, as well as reports from the Inferred Allele Review Committee (IARC). Evaluation of evidence in support of each candidate sequence led to the preparation of Source Sets. These include all supported sequences as officially recognized. Reference Sets were then prepared from the Source Sets by retention of just a single copy of duplicated sequences, and by the artificial extension of sequences that appear to be truncated at their 3′ or 5′ ends.

Compilation and evaluation of data relating to IGHV, IGKV and IGLV genes

The AIRR-C Reference Sets were produced using a protocol outlined in Figure 1. Sets of candidate V, D and J sequences were first compiled either directly from a number of key reports, or by investigation of GenBank submissions associated with those reports. This included the GRCh38.p14 genome reference sequence and IGLV and IGLJ assemblies produced using similar sequencing and assembly methods (28) that are more fully described below. Orphons, which are non-functional IG-like genes found outside the IG loci, were not considered for inclusion. V sequences with apparent 5' or 3' truncations of 25 or more nucleotides were also excluded. Candidate sequences were then restricted to those genes for which at least one allele has been reported that is an Open Reading Frame (ORF) and that includes both the first and the second conserved cysteines of the sequence. In this way, many pseudogenes were excluded from consideration. Of the remaining sequences, a candidate V sequence was ultimately included in what is referred to here as a Source Set if it could be confirmed from at least two of these independent reports, or if the reported sequence could be confirmed by additional evidence from particular AIRR-seq datasets that had been evaluated for the quality of the methods used in their generation. The Source Sets were finally used to produce IGH_VDJ, IGKappa_VJ and IGLambda_VJ Reference Sets by the extension of (likely) truncated sequences, and after dealing with exact paralogs as outlined below.

The key reports used to compile the set of candidate IGHV sequences included the historically important genomic studies of the IGH locus by Matsuda and colleagues (13, 49) and Watson and colleagues (34). The recent genomic study of 154 individuals of diverse ethnicity, by Rodriguez and colleagues, was also included (27). Candidate IGKV and IGLV sequences were compiled by reference to critical studies from the Zachau group (50–52), the Winter/Tomlinson group (53–55), the Watson group (28, 33), and the Shimizu group (56, 57). Sequences from the key studies were identified directly from the literature or from BLAST searches of GenBank. Genomic IGLV and IGLJ sequences from a recent study (28) were accessed at the VDJbase website (VDJbase.org). In this study, the authors generated haplotyped assemblies of the IGL locus from 16 individuals using a combination of long-read whole-genome sequencing, fosmid technology and capture-probe long-read sequencing. The assemblies were annotated using IGenotyper (58). The authors also deposited at VDJbase annotations of 32 long-read assemblies created by the Human Pangenome Project (59). This data is collectively referred to here as the Gibson-VDJbase dataset.

Heavy and light chain sequences identified in the genomic studies of Mikocziova and colleagues (30, 31), genomic studies of IGH genes in African and Melanesian individuals (35, 36) and

heavy chain sequences identified by genomic validation (individual D19) in the study of Narang and colleagues (60) were then included as candidate sequences.

Additional IG sequences inferred from AIRR-seq data and affirmed by the IARC were also added to the set of candidate sequences (39). These IARC-affirmed sequences have not yet been officially named by the IUIS TR-IG NRC, but the sequences are publicly available at the OGRDB website https://ogrdb.airr-community.org/sequences/Human (48).

Finally, reports were compiled from the VDJbase website of genes that were inferred from AIRR-seq data in the study of Gidoni and colleagues (61), which was chosen because it is a particularly high-quality study of FACS-sorted naïve ($CD19^+$ $CD27^-$ IgD^+ IgA^- IgG^-) B cells. The VDJbase pipeline uses the TIgGER genotype inference utility (46), and was used to produce the Gidoni-VDJbase dataset of 99 individual heavy and light chain IG genotypes. These genotypes were additionally reviewed and confirmed against VDJbase haplotyping data, whenever possible.

For a candidate IGHV sequence to be included in the IGH Source Set, at least two independent reports of that sequence were required. Such reports could, for example, be provided by two key studies, or a single key study could be supported by Gidoni-VDJbase AIRR-seq data or Gibson-VDJbase genomic sequence data. A report of a genomic sequence from a particular laboratory was not considered suitable confirmation of a separate report of that sequence by the same laboratory, unless the genomic data was clearly reported from multiple individuals. If the multiple origins of genomic germline sequences were unequivocal, as was often the case with sequences from the Wang, Gibson and Rodriguez studies (27, 28, 36), this was accepted as sufficient confirmation of the reported sequence. Multiple reports from different individuals in the Gidoni-VDJBase AIRR-seq dataset were also considered to be adequate confirmation of the reality of a particular sequence, if the sequence was also supported by a suitable GenBank entry that reported an unrearranged sequence.

For most sequences, key studies provided details of genomic mapping. The general location of other sequences within the genome was inferred by careful analysis of selected AIRR-seq datasets. Expression data is unable to distinguish between identical sequences, such as IGHV1-69*01 and IGHV1-69D*01. In the evaluation of such cases, a sequence was confirmed at a particular genomic location if it had been mapped at that locus by at least one of the key genomic studies and if the sequence had been observed in AIRR-seq data.

Compilation and evaluation of data relating to IGHD, IGHJ, IGKJ and IGLJ genes

The set of candidate IGHD and IGHJ sequences was compiled by identifying all sequences reported in the historically important genomic studies of Mattila (62), Corbett (63) and Watson (34), as well as from the recent study of Rodriguez and colleagues (27). IGHD sequences from the Matsuda (13) study were also included. Candidate IGKJ and IGLJ sequences were similarly compiled from studies of the Leder group (64), the Watson group (33) and the Shimizu group (56), as well as from Gidoni-VDJbase AIRR-seq and Gibson-VDJbase genomic sequence datasets.

For a candidate IGHD, IGHJ, IGKJ or IGLJ gene to be included in the Source Sets, it had to be observed in at least two of these key genomic studies, or in a key study and at least one inferred or directly sequenced genotype from the Gidoni-VDJbase or Gibson-VDJbase datasets. Inferred genotypes were manually reviewed to confirm the validity of any candidate sequence that was observed in just a few individuals, or that was present in an individual's dataset at low frequency.

Gidoni-VDJbase IGHD genotypes are mostly based upon the identification of partial IGHD genes within VDJ rearrangements. Exonuclease trimming of IGHD gene ends means that rearrangements rarely include full-length IGHD genes. This makes it almost impossible to identify some full-length IGHD genes from VDJ rearrangements with certainty. In this study, genotypes based upon AIRR-seq data were used to confirm critical centrally-located gene-defining and allele-defining nucleotides in selected IGHD sequences. Expression data was not used to confirm the four highly similar genes of the IGHD1 gene family (IGHD1-1*01, IGHD1-7*01, IGHD1-14*01 and IGHD1-20*01), the alleles of IGHD2-2 that are distinguished from one another by terminal nucleotides, or the extremely short IGHD7-27 gene. And while expression data can identify the identical IGHD4-4*01/IGHD4-11*01 and IGHD5-5*01/IGHD5-18*01gene pairs, it cannot be used as evidence in support of the existence of any of those individual genes.

Development of the AIRR-C IG Reference Sets from the Source Sets

Truncated sequences in the Source Sets have been extended in the AIRR-C IGH_VDJ, IGKappa_VJ and IGLambda_VJ Reference Sets. IGHV, IGKV and IGLV alleles in the Source Sets were first identified that were shorter than other allelic variants of their genes, or shorter than other highly similar genes. This included most IARC-affirmed alleles. The IARC does not usually affirm the terminal nucleotides of a sequence, but their reports may include recommendations regarding the terminal nucleotides, based upon analysis of the gene ends in the AIRR-seq data. Where such recommendations have been made, they were used as suitable extensions. Recent genomic sequences were also used as a source of many extensions (27, 60), and for some sequences, Gidoni-VDJbase AIRR-seq data (61) was reviewed to identify the most likely gene ends. If no evidence of these kinds were available, extensions of one or two nucleotides were based upon the endings of similar sequences.

The AIRR-C IG Reference Sets also include modifications to deal with exact paralogs. Where a nucleotide sequence was present as two or more identical entries in a Source Set, only a single entry was retained in the matching Reference Set. The names of any exact paralogs are noted in the metadata associated with the retained sequences.

No investigations of functionality were conducted here. The AIRR-C IG Reference Sets are restricted to those genes that are known to include alleles that are ORFs. Some genes include alleles that are ORFs and other alleles that are not. A small number of

sequences in the Reference Sets that are not ORFs are allelic variants of these genes. The Reference Sets may also include some pseudogenes that are non-functional for reasons that lie outside their coding regions. It is noted within the OGRDB database whether or not each sequence is an ORF, and expression data is available via VDJbase.

GenBank entries were found for all IG genes identified in the key studies. GenBank Accession numbers were identified by BLAST searches for the small number of previously reported sequences that were only identified by reference to VDJbase data. For all but one of these sequences, BLAST searches identified one or more suitable GenBank Accession numbers. The only reported sequence in GenBank matching IGLV2-8*03 was the truncated sequence Y12418. This sequence is therefore reported in the Reference Set metadata. For all other sequences, accession numbers of full-length unrearranged sequences are reported in the sequence metadata, and a single sequence, selected where possible to include full sequences of the flanking regions, was used as the basis of OGRDB annotation. This 'representative' sequence is reported in the associated notes.

Results

We present optimized germline AIRR-C IG Reference Sets for IG heavy (IGH_VDJ), kappa (IGKappa_VJ) and lambda (IGLambda_VJ) genes in which evidence in support of each sequence included in the database has been evaluated and strict criteria have been met. The possible consequences of inclusion of truncated sequences in a reference set is first demonstrated.

Demonstration of the consequences of sequence truncations on AIRR-seq analysis

IgBLAST-based annotation of AIRR-seq data sets of six subjects shown by inference to carry either one of the two alleles of IGHV4-38-2 in their genotypes was assessed. When data sets were annotated using the reference set with the truncated IGHV4-38-2*01, four individuals who only carried the IGHV4-38-2*01 allele were suggested to be heterozygous at the IGHV4-38-2 locus (Figure 2A). About a third of reads were software-annotated as being derived from IGHV4-38-2*02, with all these alignments including at least 1 difference from the germline sequence (data not shown). Extension of the IGHV4-38-2*01 sequence allowed IgDiscover to annotate essentially all reads in these four individuals as being derived from IGHV4-38-2*01 (Figure 2B). In contrast, annotation of reads derived from IGHV4-38-2 in two subjects who through inference had been shown to express only IGHV4-38-2*02 resulted, as expected, in the vast majority of reads being identified as IGHV4-38-2*02. This was independent of the length of the 3' end of the IGHV4-38-2*01 sequence. There was a slight increase in misalignments to IGHV4-38-2*01 in these datasets, after extension of the IGHV4-38-2*01 sequence. Review of IgBLAST alignments showed these misidentified VDJ rearrangements to have identical sequence identity to both of the alleles, but the reads were assigned to IGHV4-38-2*01 by downstream analysis programs. In summary, truncated sequences in a reference set can negatively affect the quality of sequence annotation.

Genes of the IGH locus

A set of 295 candidate IGHV gene sequences was compiled from published sources, as outlined. 67 sequences were rejected because they were associated with genes for which no alleles which are ORFs have been seen, because of major truncations, or because the sequences lacked the conserved cysteines. Based on the criteria outlined in the Methods section, we found evidence providing confidence in the existence of 206 of the remaining IGHV sequences, including a number of exact paralogs. The 22 candidate sequences that were not confirmed for inclusion in the IGH Source Set are documented in Supplementary Table 1. This will aid their

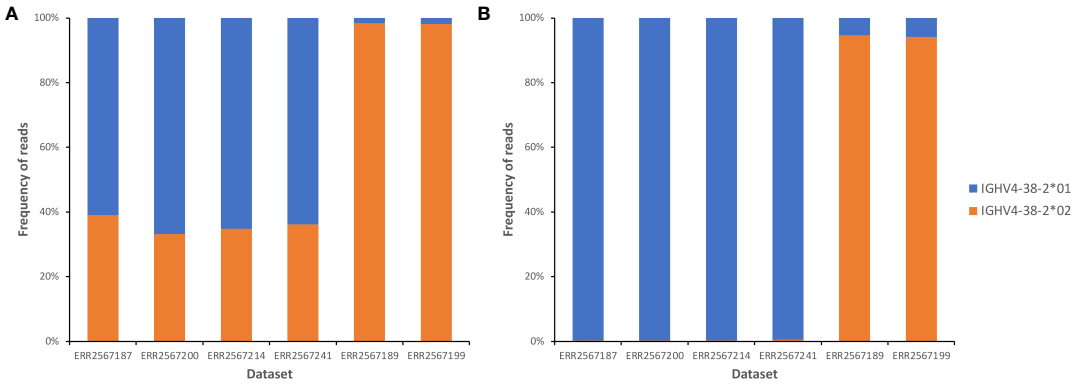

FIGURE 2
Frequencies of alignments to each of two IGHV4-38-2 alleles, in analysis of datasets from six individuals who each carried just one IGHV4-38-2 allele, before **(A)** and after **(B)** the extension of the IGHV4-38-2*01 allele by two nucleotides at its 3' end. In panel **(B)**, review of the small number of alignments to IGHV4-38-2*01 in the ERR2567189 and ERR2567199 datasets showed them to involve VDJ rearrangements that had identical sequence identity with both of the alleles. In such circumstances, many downstream analytical tools assign the rearrangements to the allele with the lowest allele number.

evaluation for possible inclusion in future versions of the reference sets, if additional supportive evidence becomes available.

The genomic locations of 167 of the 206 IGHV sequences in the IGH Source Set were confirmed by the key studies or by analysis of Gidoni-VDJbase AIRR-seq data. This analysis focused upon individuals who carried a sequence in question, but also carried other mapped alleles of that particular gene, and were heterozygous at the IGHJ6 locus. Haplotype analysis of AIRR-seq data can show that an unmapped sequence is always associated with a particular J allele, in individuals who are heterozygous at that J locus, while a mapped allele of the same gene is always associated with the alternative J allele (Figure 3). In such cases, the general location of the unmapped sequence can be inferred with confidence.

Careful review of genomic data (27), numerous AIRR-seq datasets (e.g. Gidoni-VDJBase P1_I10_S1) and previous studies (60, 65) suggest that the IGHV4-59*08 sequence is usually found at the IGHV4-61 locus. As the sequence was observed by Rodriguez and colleagues at the IGHV4-59 locus in one individual (27), the IGHV4-59*08 sequence was included in the IGH Source Set. The identical sequence at the IGHV4-61 locus was given the temporary name IGHV4-NL1*01, and was also included in the Source Set. As the IGHV4-61 locus is the more likely source of the sequence, IGHV4-NL1*01 was selected for inclusion in the IGH_VDJ Reference Set.

Twenty-nine of the IGHV gene sequences had apparent short 3' truncations of one or two nucleotides. All these sequences are extended in the IGH_VDJ Reference Set, and this is noted in the metadata. Only one sequence was extended (by two nucleotides) based solely upon the endings of other similar full-length alleles: IGHV5-51*06.

Many pairs of apparently identical IGHV sequences were identified in the Rodriguez study (27). Twelve alleles of the IGHV1-69 gene, four alleles of the IGHV2-70 gene and three alleles of the IGHV3-23 gene were also observed as exact paralogs at the IGHV1-69D, IGHV2-70D and IGHV3-23D loci (27). For the moment, this is not reflected in the IGH_VDJ Reference Set, and these observations are not included in counts of sequences in the Source Sets or Reference Sets. Most of these sequences are only included in the IGH Source Set and the IGH Reference Set as alleles of the IGHV1-69, IGHV2-70 and IGHV3-23 genes. Four sequences (IGHV1-69D*01, IGHV2-70D*04, IGHV2-70D*14 and IGHV3-23D*01) are included in the Source Set and are noted as exact paralogs in the metadata associated with IGHV1-69*01, IGHV2-70*04, IGHV2-70*14 and IGHV3-23*01. Evaluation of the genomic location of the other reported IGHV1-69D, IGHV2-70D and IGHV3-23D sequences awaits the publication of assemblies produced using longer genomic sequences.

Four other sequence pairs with names that do not hint at their shared sequence identity are recognized: IGHV3-30*02 and IGHV3-30-5*02; IGHV3-30*04 and IGHV3-30-3*03; IGHV3-30*18 and IGHV3-30-5*01; IGHV4-30-4*09 and IGHV4-31*03. In the IGH_VDJ Reference Set, the IGHV3-30, IGHV3-33 and IGHV4-31 alleles are included, and the exact paralogs at alternative loci are noted in the Reference Set metadata. The data of Rodriguez and colleagues suggests that there are many other paralogs of IGHV3-30, IGHV3-33, IGHV3-30-3 and IGHV3-30-5 genes (27). These Rodriguez sequences also await formal validation.

Thirty-four candidate IGHD sequences were considered for inclusion in the Reference Set. Three reported IGHD sequences, IGHD3-3*02, IGHD3-10*02 and IGHD3-16*01, were not included as candidate sequences, as they were not reported by any of the key genomic studies and were not identified in the Gidoni-VDJbase IGHD genotypes. Confirmation of most of the candidate IGHD sequences came from their reporting in two or more of the key

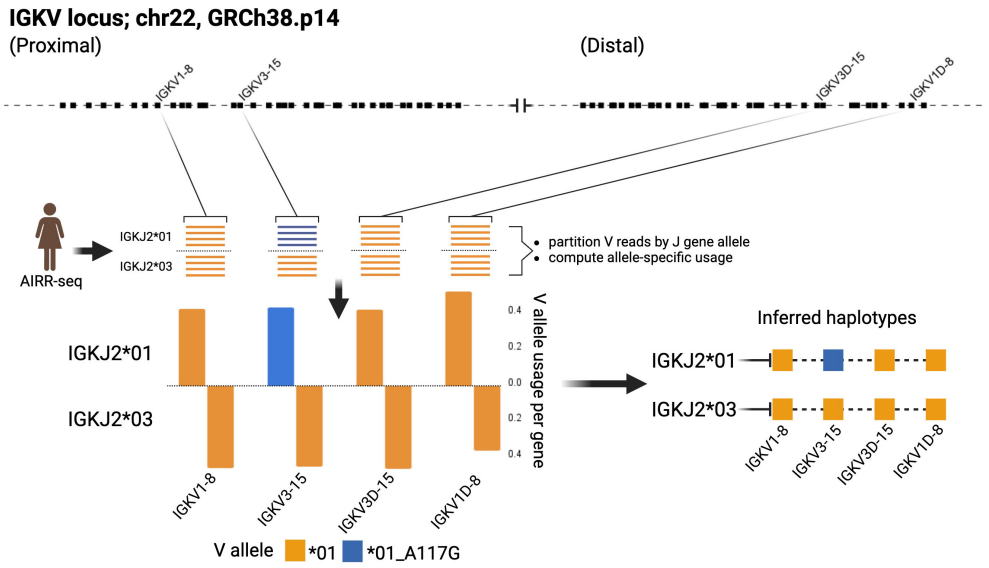

FIGURE 3
The use of AIRR-seq data to infer the existence and likely location of novel alleles. Data from Gidoni-VDJbase IGK dataset P1_I24 provided strong support for the existence of the novel IGKV sequence IGKV3-15*01_A117G, and of its location in the proximal locus. Inferred haplotypes were determined using the IGKJ2 locus in this heterozygous individual. Created with BioRender.com.

10.3389/fimmu.2023.1330153

genomic studies. IGHD2-2*03 was only documented in a single key genomic study and could not be validated with AIRR-seq data. It is therefore not included in the Reference Set. IGHD2-8*02 and IGHD2-21*01 were seen in the genomic sequences of Rodriguez and were confirmed by AIRR-seq alignments. Three candidate IGHD sequences were only seen in the Rodriguez study - IGHD3-10*03, IGHD3-16*03 and IGHD5-18*02. These sequences were unofficially assigned these names when they were recently incorporated into the IMGT database, and they have been included in the Reference Set using these names. Two IGHD gene pairs of identical sequences were seen: IGHD4-4*01/IGHD4-11*01 and IGHD5-5*01/IGHD5-18*01. IGHD4-4*01 and IGHD5-5*01 represent the two pairs in the IGH_VDJ Reference Set.

Only eight IGHJ gene sequences were candidates for inclusion in the IGH_VDJ Reference Set, as five other previously reported sequences (IGHJ3*01, IGHJ4*01, IGHJ4*03, IGHJ5*01, and IGHJ6*01) lacked evidentiary support. The reported IGHJ pseudogenes IGHJ1P, IGHJ2P and IGHJ3P were also not considered here. Seven of the eight candidate sequences were confirmed from multiple genomic reports. IGHJ6*04 was only seen in one Gidoni-VDJbase dataset (P1_I69), but as this observation lacked confirmation, the sequence was not included in the IGH_VDJ Reference Set. Recent sequencing data (27, 60) was used to provide the 3' terminal nucleotide for the IGHJ6*03 sequence in the Reference Set.

Genes of the Kappa and Lambda Light Chain Loci

102 candidate IGKV gene sequences were identified. Twenty-four sequences were rejected because they were associated with genes for which no alleles which are ORFs that include the conserved cysteines have been seen. Evidence was seen providing confidence in the existence of 71 of the remaining sequences. Seven candidate sequences that were excluded from the Reference Set because of a lack of confirmatory evidence are documented in Supplementary Table 2. The genomic locations of all but 8 of the sequences in the Reference Set were confirmed either by direct mapping studies, or by inference from AIRR-seq data. The IGKV inferred mapping analysis focused on individuals who are heterozygous at the IGKJ2 locus. The genomic location of two novel alleles as well as six genes of the duplicated regions of the IGK locus were confirmed with confidence after analysis of Gidoni-VDJbase haplotype-based data (Figure 3). Expression frequency data was also used. Genes of the proximal locus are known to be expressed at significantly higher frequencies than their corresponding genes of the distal locus (53). The two novel IGKV sequences were confirmed as alleles of IGKV3-15 rather than IGKV3D-15, for like IGKV3-15*01, these sequences were recorded at frequencies greater than 4% of the total repertoire (Figure 4). IGKV3D-15*01 was consistently seen at frequencies of 0.6% or less. The novel sequences were recently affirmed by the IARC and given the names IGKV3-15*i01 and IGKV3-15*i02.

Mapping studies have confirmed that five IGKV sequences of the proximal locus (IGKVare present as exact paralogs in the distal locus. In the IGKappa_VJ Reference Set, these sequences are listed as genes of the proximal locus, with their paralogs noted in the Reference Set metadata.

There were seven candidate IGKJ sequences, and they were all confirmed as suitable for the Source Set, and subsequently for the AIRR-C IGKappa_VJ Reference Set. Two previously reported IGKJ sequences (IGKJ2*02 and IGKJ4*02) were excluded. 71 IGKV and 7 IGKJ sequences therefore form the IGKappa_VJ Source Set, while 66 unique IGKV and 7 IGKJ sequences make up the AIRR-C IGKappa_VJ Reference Set.

115 candidate IGLV gene sequences were first identified, including 5 novel alleles. The novel alleles were all ORFs from the study of Gibson and colleagues (28) that were seen in six or more individuals. Other less-frequently observed novel alleles from the

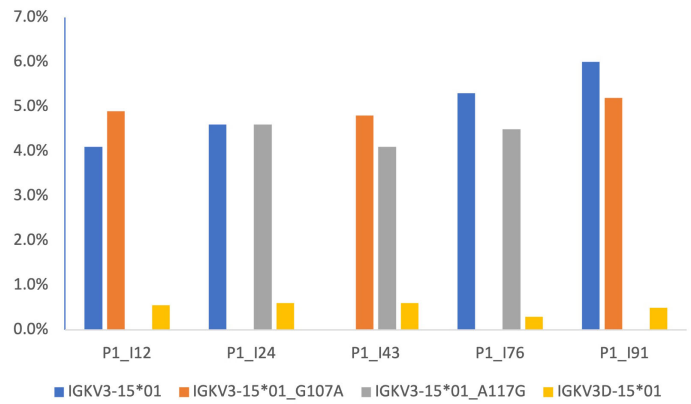

FIGURE 4
Frequencies of expression of IGKV genes, as percentages of the overall VJ repertoire, in five individuals identified in the Gidoni-VDJbase datasets. These datasets show novel alleles are expressed in three individuals who also express two IGKV3D-15*01 alleles. This suggest that the unmapped novel alleles are located in the proximal kappa locus, and this is also supported by haplotyping of P1_I12 and P1_I24 (data not shown). Expression frequencies give further support for the genomic location of the novel alleles, for genes of the proximal locus are known to be expressed at significantly higher frequencies than genes of the distal locus.

Gibson study await future assessment. Twenty candidate sequences were then rejected because they were associated with genes for which no alleles which are ORFs with the conserved cysteines have been seen. Confirmatory evidence was seen for 81 of the sequences, and the genomic locations of 52 of those sequences were confirmed by the studies used here in the compilation of the set. Haplotype-based mapping was not possible for IGLV genes as there is no commonly heterozygous IGLJ locus by which haplotyping can be performed. Three of the sequences (IGLV2-8*03, IGLV2-14*04 and IGLV3-9*01) had substantial truncations, and in the AIRR-C IGLambda_VJ Reference Set these sequences have been extended by reference to either Gibson-VDJbase or Gidoni-VDJbase datasets. The 14 candidate IGLV sequences that were excluded from the Reference Set because of a lack of confirmatory evidence are documented in Supplementary Table 3.

There were ten candidate IGLJ gene sequences and they were all confirmed as suitable for the Source Set. IGLJ2*01 and IGLJ3*01 are identical sequences and in the Reference Set only IGLJ2*01 is shown. Nine IGLJ and 81 IGLV sequences therefore form the AIRR-C IGLambda_VJ Reference Set.

The Reference Sets were compared to the IMGT Reference Directory that is used by the IMGT/V-QUEST alignment utility, focusing on genes for which ORFs are observed and the conserved cysteines are present (Table 1). In addition to gene pairs that are only represented once in the AIRR-C Reference Set, a large number of sequences are present in the IMGT Reference Directory but absent from the AIRR-C Set. However, analysis of the Gidoni-VDJbase dataset shows a very small number of alignments to sequences other than those in the Reference Set (Table 1).

Discussion

Analyses of AIRR-seq data, and even analyses of solitary V(D)J gene rearrangements, generally require the use of germline gene reference sets that are as complete as possible, and that contain as few erroneously reported sequences as possible. These conflicting demands posed a challenge in the preparation of the AIRR-C Reference Sets. We have erred here on the side of caution. By restricting our sources of reported germline genes to a small number of key studies, we have likely removed all sequences that were reported in error. Inevitably, some real sequences will also have been lost from the Reference Sets, for the moment, as a consequence.

We believe that analysis of new high-quality AIRR-seq expression data and new genomic data should allow most currently excluded but genuine sequences to be confirmed in the near future. Many other novel alleles have already been reported in the studies that were used to establish the Reference Sets (27, 61). The present study was only able to evaluate a handful of novel sequences for inclusion in the Reference Sets, and the evaluation of other sequences is a task for the future. Even before this happens, it is likely that these first versions of the Reference Sets include nearly all common IGHV, IGKV, IGLV, IGHD, IGHJ, IGKJ and IGLJ allelic variants and many less-common variants that are found in those populations that presently receive the most attention from AIRR-seq researchers.

We believe that a relatively small number of previously reported sequences that are absent from the Reference Sets are real functional alleles, and many of the 'missing' IGHV sequences have previously

TABLE 1 Number of IG alleles that are shared by the AIRR-C Reference Set and the IMGT V-QUEST Reference Directory (accessed 24/07/2023), the number of alleles that are unique to the AIRR-C Reference Set, and the number of alleles unique to the IMGT Reference Directory.

	Shared	AIRR-C alone[b]	IMGT alone	Non-AIRR-C Alignments[c]
IGHV[a]	190	8 (4128/2888407: 0.14%)	156	521/2888407 (0.02%)
IGHD	31	–	4	1621/2364574 (0.07%)
IGHJ	7	–	6	1372/2910350 (0.05%)
IGKV[a]	64	2 (5625/828127: 0.68%)	32	1990/828127 (0.24%)
IGKJ	7	–	2	0/1325749 (0.00%)
IGLV[a]	75	5 (2229/541380: 0.41%)	30	29/541380 (0.005%)[d]
IGLJ	9	–	–	0/326639 (0.00%)

[a]Tabulation is confined to V genes that include alleles that are ORFs and that include the conserved cysteines.
[b]The number of distinct sequences in the Gidoni-VDJbase dataset that align to AIRR-C alleles that are not in the IMGT Reference Directory, after genotyping, as a proportion of the total number of distinct sequences seen in each of the repertoires.
[c]The number of distinct sequences in the Gidoni-VDJbase dataset that align to IMGT alleles that are not in the AIRR-C Reference Set, after genotyping, as a proportion of the total number of distinct sequences seen in each of the repertoires.
[d]30312 alignments (5.60%) to IGLV2-23*01 and IGLV3-9*02, neither of which is in the Reference Set, have been removed from the calculation. We believe these are all misalignments resulting from the fact that they are longer than the alternative alleles.
Genes that exist as two or more exact copies in the genome are only represented once in the table. Orphons (e.g. IGHV1/OR15*01) are not included in the tabulation.

been flagged as likely resulting from sequencing errors in the 1980s and 1990s when accurate sequencing was particularly challenging (26). For example, IGHV3-30 is often described as the most polymorphic of the human IGHV genes, but most of the reported IGHV3-30 alleles came from a single study that amplified sequences from a single individual (66). No additional evidence has ever emerged in support of 14 of these IGHV3-30 alleles, and they are not included in the AIRR-C IGH_VDJ Reference Set. Most of the rejected IGHD and IGHJ genes have also previously been identified as having been reported in error (67, 68).

Other 'missing' IGHV sequences have only recently been reported, so they cannot be supported by older key studies. Forty-two human IGHV sequences have been added to the IMGT database in 2023, and 22 of these sequences are now included in the IMGT Reference Directory. The other sequences are described by IMGT as out-of-frame pseudogenes. Only five of the 22 sequences were seen in the Rodriguez study (27), and only four of these sequences are included in the AIRR-C Reference Set. It remains unclear whether the other sequences are rarely carried in the human populations that were the focus of the Rodriguez study, or whether they have been reported in error. In time, if these sequences have been accurately reported, accumulating evidence should lead them to be included in the AIRR-C Reference Set.

Many sequences that IMGT has described as IGHV, IGKV or IGLV pseudogenes are also absent from the AIRR-C Reference Sets. The failure to confirm the existence of these pseudogenes may partly reflect the inability of AIRR-seq data to provide such evidentiary support. If they are seen at all in AIRR-seq V(D)J rearrangements, pseudogenes are likely to be very rare. The Gidoni-VDJbase data was therefore generally unable to provide confirming evidence in support of pseudogenes. If real pseudogenes have been omitted for this reason, however, it will have little consequence for AIRR-seq analysis.

The three 'Not Located' sequences in the Reference Sets were given additional attention, for some might believe that the failure to map such distinctly different sequences to a genomic assembly is an indication that the sequences were reported in error. IGHV4-NL1*01 was named as part of this report, and should be considered 'misplaced' rather than being unlocatable in the genome. We believe this sequence will soon be officially named as an allelic variant of the IGHV4-61 gene.

BLAST searches identified IGKV1-NL1*01 in two FOSMID clones (AC253566 and AC215521) and a BAC clone (AC145029). The novel gene appears to be located between two genes at the far end of the kappa distal locus: IGKV3D-7 and IGKV1-D8. There can be no doubt that this gene exists.

The likely genomic location of IGHV3-NL*01 could not be investigated as it has not been reported from FOSMIDS or BAC clones. In fact, it has only been reported from Papua New Guinean samples. The sequence has been seen in genomic amplifications from buccal swabs, using IGHV RSS-based primers (36), and in cDNA amplified from PBMCs using IG constant region-based primers (69). The sequence was identified in multiple individuals from samples collected at different places and times (36, 69). The IMGT database notes that the sequence might be a result of chimeric amplification, though this possibility was ruled out in the original report of the sequence (36). We strongly believe that the sequence has been accurately reported, and its existence likely points to the significant structural variation that remains to be documented in different human populations.

The AIRR-C Reference Sets are different from other reference sets in several critical ways. Importantly, the data that underpins the reporting of each sequence can be accessed at the OGRDB website, and the site allows easy interrogation of the data. The reasons why sequences are present in, or sequences have been excluded from the Reference Sets are also clearly documented.

The AIRR-C Reference Sets report whether sequences in the Reference Sets are ORFs, but they do not attempt to identify sequences as being either Functional or Pseudogenes. The functionality of IG genes is not easily defined for there is still no comprehensive understanding of the impact of genetic variation in non-coding regions upon IG expression. There have been some studies of variations in elements such as Recombination Signal Sequences (RSSs) (70, 71), but it is unclear, for example, why some apparently functional IGHV genes (e.g. IGHV4-4*01) and IGHD genes (e.g. IGHD6-25*01) are present at such low frequencies in the expressed VDJ repertoire (67, 72) despite these genes reportedly being associated with functional RSSs (44). For this reason, there was no attempt to document functionality in the AIRR-C Reference Sets.

In contrast to other reference sets, the AIRR-C Reference Sets have been developed to address specific challenges with AIRR-seq analysis. While the Source Sets report sequences exactly as they have been reported in key studies, or as they have been affirmed from AIRR-seq data by the IARC, the AIRR-C Reference Sets include modifications to facilitate accurate analysis of AIRR-seq data.

Many reported sequences are truncated, and this is particularly true for inferred sequences. Although V sequences can be inferred from AIRR-seq data with great confidence for most of their length, the terminal 3' nucleotides are usually less certain (73). This is because of data loss resulting from exonucleolytic trimming of gene ends during V(D)J recombination. Other historical reports of germline genes describe sequences that now appear to have short 3' truncations as a result of problems identifying the boundary between the V gene exons and their RSS. We have demonstrated the consequences of truncations on alignments by analysis of VDJ rearrangements involving the IGHV4-38-2 gene. The consequences are also clear in a review of lambda VJ rearrangements identified as involving IGLV2-23*01 and IGLV3-9*02 when sequences were aligned against the IMGT Reference Directory (Table 1). These two sequences are not in the AIRR-C IGLambda_VJ Reference Set, and we believe that all these alignments are in error. We believe that these sequences all involve rearrangements of IGLV2-23*03 and IGLV3-9*01, but these shorter alleles have been overlooked by the alignment utility. In the AIRR-C Reference Sets, apparently truncated sequences have been extended, using IARC recommendations and recent genomic evidence. Extended alleles are listed in the Reference Set release notes.

The AIRR-C Reference Sets have also been developed to deal with analytical challenges that arise from the presence of exact paralogs in a reference set. To allow accurate analysis, the AIRR-C

Reference Sets list each duplicate sequence only once. This involves IGHV and IGHD gene pairs, as well as IGKV sequences that are present as identical copies in both the proximal and distal IGKV loci, and a single IGLJ gene pair. The names of all duplicate sequences in the Reference Sets are noted in the Reference Set metadata and in the Reference Set release notes. The OGRDB website provides access to a script that facilitates the production of Reference Sets with details of exact paralogs and sequence aliases in the FASTA header (https://airr-community.github.io/receptor-germline-tools/_build/html/introduction.html), should that be desired.

The AIRR-C Reference Sets will provide an avenue by which high quality sequences that lack official names can be rapidly brought to the attention of the research community. This is an issue of growing importance because of an emerging problem with gene reporting. The number of groups of genes that are recognized for their similarities has recently grown, with the documentation of previously unknown structural variation and gene duplication in the IGH locus (27). Until a new approach to the IG nomenclature is developed, there will likely be delays in the official naming of many newly discovered human alleles that appear to belong to these gene groups. The AIRR-C Reference Sets aim to provide early access to such sequences, if their existence is supported by multiple lines of evidence.

Updates to the Reference Sets will be managed by the IARC and the AIRR-C Germline Gene Working Group, and details of future reviews will be publicly available. Only sequences identified in studies of the highest quality will be considered for future inclusion in the sets. If additional inferred sequences are included in the Reference Sets prior to their consideration by the IUIS TR-IG NRC, sequences will be reported using IARC-assigned or other temporary names. These names will be recorded as 'aliases' when and if official names are given to the sequences. In the future, it will therefore be possible for reports of each sequence to be easily traced through the literature.

The AIRR-C IGH_VDJ, IGKappa_VJ and IGLambda_VJ Reference Sets, as well as the Source Sets from which they are derived, are available at the OGRDB website and can be directly accessed by users of IgBLAST. The sets are published using the AIRR Data Schema (74) which has recently been revised to allow for the definition of reference sets (40). The AIRR-C schema for GermlineSet is supported, and germline sets are downloadable in JSON format compliant with the schema, or in FASTA format. They can also be queried via a REST API. OGRDB manages versioning and change control, such that both users and curators can identify the addition, removal, or modification of sequences in a germline set, and drill down to individual records for each sequence to reveal details. Data published on OGRDB is provided under a minimally restrictive Creative Commons CC0 1.0 licence. OGRDB data is periodically archived at Zenodo (https://zenodo.org) for long-term storage, and each version of a germline set is also deposited at Zenodo and allocated a DOI (https://www.iso.org/standard/81599.html): hence users may cite a persistent identifier that uniquely references the set used in their work.

We believe that together, these many features make the AIRR-C Reference Sets far superior to existing publicly available human IG reference sets, and the AIRR-C Germline Gene Working Group strongly recommends their use for all human AIRR-seq analysis.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

AC: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. MO: Conceptualization, Formal analysis, Visualization, Writing – review & editing. MC: Conceptualization, Writing – review & editing. JH: Conceptualization, Writing – review & editing. DR: Methodology, Writing – review & editing. ML: Writing – review & editing. JM-B: Investigation, Writing – review & editing. JY: Methodology, Writing – review & editing. ER: Visualization, Writing – review & editing. WG: Data curation, Writing – review & editing. OR: Data curation, Writing – review & editing. AP: Conceptualization, Data curation, Writing – review & editing. GY: Conceptualization, Data curation, Methodology, Writing – review & editing. CW: Investigation, Writing – review & editing, Conceptualization, Data curation, Methodology, Visualization. WL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. JY was supported by the National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health. MO was supported in part by the Swedish Research Council (grant number 2019-01042). ER was supported by NIH contract 75N93019C00001 (NIAID) and grant U24CA248138 (NCI). CTW, WSG, and OLR were funded in part by relevant grants from the National Institute of Allergy and Infectious Diseases (R21AI142590 and R24AI138963). WL was supported in part by the European Union's Horizon 2020 research and innovation program (grant number 825821).

Acknowledgments

We thank Christian E. Busse, German Cancer Research Center, Heidelberg, and Ivana Mikocziova, University of Turku, Finland for their valuable comments on earlier drafts of the manuscript.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1330153/full#supplementary-material

SUPPLEMENTARY TABLE 1
Evidence in support of the existence of human IGH genes that were candidates for inclusion in the AIRR-C IGH_VJ Reference Set, but which lacked sufficient evidence for inclusion.

SUPPLEMENTARY TABLE 2
Evidence in support of the existence of human IGKV genes that were candidates for inclusion in the AIRR-C IGKappa_VJ Reference Set, but which lacked sufficient evidence for inclusion.

SUPPLEMENTARY TABLE 3
Evidence in support of the existence of human IGLV genes that were candidates for inclusion in the AIRR-C IGLambda_VJ Reference Set, but which lacked sufficient evidence for inclusion.

References

1. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep* (2016) 6:20842. doi: 10.1038/srep20842

2. Nielsen SCA, Boyd SD. Human adaptive immune receptor repertoire analysis-Past, present, and future. *Immunol Rev* (2018) 284:9–23. doi: 10.1111/imr.12667

3. Schumacher TN, Thommen DS. Tertiary lymphoid structures in cancer. *Science* (2022) 375:eabf9419. doi: 10.1126/science.abf9419

4. Wu X, Zhou T, Zhu J, Zhang B, Georgiev I, Wang C, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science* (2011) 333:1593–602. doi: 10.1126/science.1207532

5. Sun X, Liu C, Lu X, Ling Z, Yi C, Zhang Z, et al. Unique binding pattern for a lineage of human antibodies with broad reactivity against influenza A virus. *Nat Commun* (2022) 13:2378. doi: 10.1038/s41467-022-29950-w

6. Cao Y, Su B, Guo X, Sun W, Deng Y, Bao L, et al. Potent neutralizing antibodies against SARS-coV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. *Cell* (2020) 182:73–84 e16. doi: 10.1016/j.cell.2020.05.025

7. Burton DR. Advancing an HIV vaccine; advancing vaccinology. *Nat Rev Immunol* (2019) 19:77–8. doi: 10.1038/s41577-018-0103-6

8. Akinosoglou K, Rigopoulos EA, Kaiafa G, Daios S, Karlafti E, Ztriva E, et al. Tixagevimab/cilgavimab in SARS-coV-2 prophylaxis and therapy: A comprehensive review of clinical experience. *Viruses* (2023) 15:118–36. doi: 10.3390/v15010118

9. Loo YM, McTamney PM, Arends RH, Abram ME, Aksyuk AA, Diallo S, et al. The SARS-CoV-2 monoclonal antibody combination, AZD7442, is protective in nonhuman primates and has an extended half-life in humans. *Sci Transl Med* (2022) 14:eabl8124. doi: 10.1126/scitranslmed.abl8124

10. Hammitt LL, Dagan R, Yuan Y, Baca Cots M, Bosheva M, Madhi SA, et al. Nirsevimab for prevention of RSV in healthy late-preterm and term infants. *N Engl J Med* (2022) 386:837–46. doi: 10.1056/NEJMoa2110275

11. Early P, Huang H, Davis M, Calame K, Hood L. An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH. *Cell* (1980) 19:981–92. doi: 10.1016/0092-8674(80)90089-6

12. Matthyssens G, Rabbitts TH. Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. *Proc Natl Acad Sci U. S. A* (1980) 77:6561–5. doi: 10.1073/pnas.77.11.6561

13. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. *J Exp Med* (1998) 188:2151–62. doi: 10.1084/jem.188.11.2151

14. Pallares N, Lefebvre S, Contet V, Matsuda F, Lefranc MP. The human immunoglobulin heavy variable genes. *Exp Clin Immunogenet* (1999) 16:36–60. doi: 10.1159/000019095

15. Corcoran MM, Phad GE, Vazquez Bernat N, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi: 10.1038/ncomms13642

16. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc Natl Acad Sci* (2015) 112:E862–70. doi: 10.1073/pnas.1417683112

17. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol* (2017) 87:12–22. doi: 10.1016/j.molimm.2017.03.012

18. Ralph DK, Matsen AT. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. *PloS Comput Biol* (2019) 15:e1007133. doi: 10.1371/journal.pcbi.1007133

19. Sheng Z, Schramm CA, Kong R, Program NCS, Mullikin JC, Mascola JR, et al. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol* (2017) 8:537. doi: 10.3389/fimmu.2017.00537

20. Schramm CA, Douek DC. Beyond hot spots: biases in antibody somatic hypermutation and implications for vaccine design. *Front Immunol* (2018) 9:1876. doi: 10.3389/fimmu.2018.01876

21. Kirik U, Persson H, Levander F, Greiff L, Ohlin M. Antibody heavy chain variable domains of different germline gene origins diversify through different paths. *Front Immunol* (2017) 8:1433. doi: 10.3389/fimmu.2017.01433

22. Collins AM. IgG subclass co-expression brings harmony to the quartet model of murine IgG function. *Immunol Cell Biol* (2016) 94:949–54. doi: 10.1038/icb.2016.65

23. Collins AM, Jackson KJ. A temporal model of human igE and igG antibody function. *Front Immunol* (2013) 4:235. doi: 10.3389/fimmu.2013.00235

24. Jackson KJ, Wang Y, Collins AM. Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol Cell Biol* (2014) 92:729–33. doi: 10.1038/icb.2014.44

25. Collins AM, Peres A, Corcoran MM, Watson CT, Yaari G, Lees WD, et al. Commentary on Population matched (pm) germline allelic variants of immunoglobulin (IG) loci: relevance in infectious diseases and vaccination studies in human populations. *Genes Immun* (2021) 22:335–8. doi: 10.1038/s41435-021-00152-6

26. Wang Y, Jackson KJ, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. *Immunol Cell Biol* (2008) 86:111–5. doi: 10.1038/sj.icb.7100144

27. Rodriguez OL, Safonova Y, Silver CA, Shields K, Gibson WS, Kos JT, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat Commun* (2023) 14:4419. doi: 10.1038/s41467-023-40070-x

28. Gibson WS, Rodriguez OL, Shields K, Silver CA, Dorgham A, Emery M, et al. Characterization of the immunoglobulin lambda chain locus from diverse populations reveals extensive genetic variation. *Genes Immun* (2023) 24:21–31. doi: 10.1038/s41435-022-00188-2

29. Engelbrecht E, Rodriguez OL, Shields K, Schultze S, Tieri D, Jana U, et al. Resolving haplotype variation and complex genetic architecture in the human immunoglobulin kappa chain locus in individuals of diverse ancestry. *bioRxiv* (2023) 2023:10. doi: 10.1101/2023.10.23.563321

30. Mikocziova I, Gidoni M, Lindeman I, Peres A, Snir O, Yaari G, et al. Polymorphisms in human immunoglobulin heavy chain variable genes and their upstream regions. *Nucleic Acids Res* (2020) 48:5499–510. doi: 10.1093/nar/gkaa310

31. Mikocziova I, Peres A, Gidoni M, Greiff V, Yaari G, Sollid LM. Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. *iScience* (2021) 24:103192. doi: 10.1016/j.isci.2021.103192

32. Vergani S, Korsunsky I, Mazzarello AN, Ferrer G, Chiorazzi N, Bagnara D. Novel method for high-throughput full-length IGHV-D-J sequencing of the immune repertoire from bulk B-cells with single-cell resolution. *Front Immunol* (2017) 8:1157. doi: 10.3389/fimmu.2017.01157

33. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes Immun* (2015) 16:24–34. doi: 10.1038/gene.2014.56

34. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet* (2013) 92:530–46. doi: 10.1016/j.ajhg.2013.03.004

35. Scheepers C, Shrestha RK, Lambson BE, Jackson KJ, Wright IA, Naicker D, et al. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J Immunol* (2015) 194:4371–8. doi: 10.4049/jimmunol.1500118

36. Wang Y, Jackson KJ, Gaeta B, Pomat W, Siba P, Sewell WA, et al. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics* (2011) 63:259–65. doi: 10.1007/s00251-010-0510-8

37. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med* (2022) 28:243–50. doi: 10.1038/s41591-021-01672-4

38. Khatri I, Berkowska MA, Akker EBvd, Teodosio C, Reinders MJT, Dongen JJMv. Population matched (pm) germline allelic variants of immunoglobulin (IG) loci: Relevance in infectious diseases and vaccination studies in human populations. *Genes Immun* (2021) 22:172–86. doi: 10.1038/s41435-021-00143-7

39. Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, et al. Inferred allelic variants of immunoglobulin receptor genes: A system for their evaluation, documentation, and naming. *Front Immunol* (2019) 10:435. doi: 10.3389/fimmu.2019.00435

40. Lees WD, Christley S, Peres A, Kos J, Corrie B, Ralph D, et al. AIRR community curation and standardised representation for immunoglobulin and T cell receptor germline sets. *ImmunoInformatics* (2023) 10:10025. doi: 10.1016/j.immuno.2023.100025

41. Lefranc MP. Nomenclature of the human immunoglobulin heavy (IGH) genes. *Exp Clin Immunogenet* (2001) 18:100–16. doi: 10.1159/000049189

42. Lefranc MP. Nomenclature of the human immunoglobulin lambda (IGL) genes. *Exp Clin Immunogenet* (2001) 18:242–54. doi: 10.1159/000049203

43. Lefranc MP. Nomenclature of the human immunoglobulin kappa (IGK) genes. *Exp Clin Immunogenet* (2001) 18:161–74. doi: 10.1159/000049195

44. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res* (2006) 34:D781–4. doi: 10.1093/nar/gkj088

45. Retter I, Althaus HH, Munch R, Muller W. VBASE2, an integrative V gene database. *Nucleic Acids Res* (2005) 33:D671–4. doi: 10.1093/nar/gki088

46. Gadala-Maria D, Gidoni M, Marquez S, Vander Heiden JA, Kos JT, Watson CT, et al. Identification of subject-specific immunoglobulin alleles from expressed repertoire sequencing data. *Front Immunol* (2019) 10:129. doi: 10.3389/fimmu.2019.00129

47. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* (2016) 3:160018. doi: 10.1038/sdata.2016.18

48. Lees W, Busse CE, Corcoran M, Ohlin M, Scheepers C, Matsen FA, et al. OGRDB: a reference database of inferred immune receptor genes. *Nucleic Acids Res* (2020) 48:D964–70. doi: 10.1093/nar/gkz822

49. Matsuda F, Shin EK, Nagaoka H, Matsumura R, Haino M, Fukita Y, et al. Structure and physical map of 64 variable segments in the 3'0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nat Genet* (1993) 3:88–94. doi: 10.1038/ng0193-88

50. Huber C, Huber E, Lautner-Rieske A, Schable KF, Zachau HG. The human immunoglobulin kappa locus. Characterization of the partially duplicated L regions. *Eur J Immunol* (1993) 23:2860–7. doi: 10.1002/eji.1830231120

51. Huber C, Schable KF, Huber E, Klein R, Meindl A, Thiebe R, et al. The V kappa genes of the L regions and the repertoire of V kappa gene sequences in the human germ line. *Eur J Immunol* (1993) 23:2868–75. doi: 10.1002/eji.1830231121

52. Schable KF, Zachau HG. The variable genes of the human immunoglobulin kappa locus. *Biol Chem Hoppe Seyler* (1993) 374:1001–22.

53. Cox JP, Tomlinson IM, Winter G. A directory of human germ-line V kappa segments reveals a strong bias in their usage. *Eur J Immunol* (1994) 24:827–36. doi: 10.1002/eji.1830240409

54. Williams SC, Frippiat JP, Tomlinson IM, Ignatovich O, Lefranc MP, Winter G. Sequence and evolution of the human germline Vl repertoire. *J Mol Biol* (1996) 264:220–32. doi: 10.1006/jmbi.1996.0636

55. Williams SC, Winter G. Cloning and sequencing of human immunoglobulin V lambda gene segments. *Eur J Immunol* (1993) 23:1456–61. doi: 10.1002/eji.1830230709

56. Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, Asakawa S, et al. Evolutionary dynamics of the human immunoglobulin kappa locus and the germline repertoire of the Vkappa genes. *Eur J Immunol* (2001) 31:1017–28. doi: 10.1002/1521-4141(200104)31:4<1017::AID-IMMU1017>3.0.CO;2-3

57. Kawasaki K, Minoshima S, Nakato E, Shibuya K, Shintani A, Schmeits JL, et al. One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res* (1997) 7:250–61. doi: 10.1101/gr.7.3.250

58. Rodriguez OL, Gibson WS, Parks T, Emery M, Powell J, Strahl M, et al. A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front Immunol* (2020) 11:2136. doi: 10.3389/fimmu.2020.02136

59. Wang T, Antonacci-Fulton L, Howe K, Lawson HA, Lucas JK, Phillippy AM, et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* (2022) 604:437–46. doi: 10.1038/s41586-022-04601-8

60. Narang S, Kaduk M, Chernyshev M, Karlsson Hedestam GB, Corcoran MM. Adaptive immune receptor genotyping using the corecount program. *Front Immunol* (2023) 14:1125884. doi: 10.3389/fimmu.2023.1125884

61. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Mikocziova I, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat Commun* (2019) 10:628. doi: 10.1038/s41467-019-08489-3

62. Mattila PS, Schugk J, Wu H, Makela O. Extensive allelic sequence variation in the J region of the human immunoglobulin heavy chain gene locus. *Eur J Immunol* (1995) 25:2578–82. doi: 10.1002/eji.1830250926

63. Corbett SJ, Tomlinson IM, Sonnhammer EL, Buck D, Winter G. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J Mol Biol* (1997) 270:587–97. doi: 10.1006/jmbi.1997.1141

64. Hieter PA, Maizel JV Jr., Leder P. Evolution of human immunoglobulin kappa J region genes. *J Biol Chem* (1982) 257:1516–22. doi: 10.1016/S0021-9258(19)68223-2

65. Parks T, Mirabel MM, Kado J, Auckland K, Nowak J, Rautanen A, et al. Association between a common immunoglobulin heavy chain allele and rheumatic heart disease risk in Oceania. *Nat Commun* (2017) 8:14946. doi: 10.1038/ncomms14946

66. Olee T, Yang PM, Siminovitch KA, Olsen NJ, Hillson J, Wu J, et al. Molecular basis of an autoantibody-associated restriction fragment length polymorphism that confers susceptibility to autoimmune diseases. *J Clin Invest* (1991) 88:193–203. doi: 10.1172/JCI115277

67. Lee CE, Gaeta B, Malming HR, Bain ME, Sewell WA, Collins AM. Reconsidering the human immunoglobulin heavy-chain locus: 1 An evaluation of the expressed human IGHD gene repertoire. *Immunogenetics* (2006) 57:917–25. doi: 10.1007/s00251-005-0062-5

68. Lee CE, Jackson KJ, Sewell WA, Collins AM. Use of IGHJ and IGHD gene mutations in analysis of immunoglobulin sequences for the prognosis of chronic lymphocytic leukemia. *Leuk. Res* (2007) 31:1247–52. doi: 10.1016/j.leukres.2006.10.013

69. Wang Y, Jackson KJ, Chen Z, Gaeta BA, Siba PM, Pomat W, et al. IgE sequences in individuals living in an area of endemic parasitism show little mutational evidence of antigen selection. *Scand J Immunol* (2011) 73:496–504. doi: 10.1111/j.1365-3083.2011.02525.x

70. Nagawa F, Kodama M, Nishihara T, Ishiguro K, Sakano H. Footprint analysis of recombination signal sequences in the 12/23 synaptic complex of V(D)J recombination. *Mol Cell Biol* (2002) 22:7217–25. doi: 10.1128/MCB.22.20.7217-7225.2002

71. Arnal SM, Holub AJ, Salus SS, Roth DB. Non-consensus heptamer sequences destabilize the RAG post-cleavage complex, making ends available to alternative DNA repair pathways. *Nucleic Acids Res* (2010) 38:2944–54. doi: 10.1093/nar/gkp1252

72. Ohlin M. Poorly expressed alleles of several human immunoglobulin heavy chain variable genes are common in the human population. *Front Immunol* (2020) 11:603980. doi: 10.3389/fimmu.2020.603980

73. Thornqvist L, Ohlin M. The functional 3'-end of immunoglobulin heavy chain variable (IGHV) genes. *Mol Immunol* (2018) 96:61–8. doi: 10.1016/j.molimm.2018.02.013

74. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, et al. AIRR community standardized representations for annotated immune repertoires. *Front Immunol* (2018) 9:2206. doi: 10.3389/fimmu.2018.02206