frontiers in Immunology

REVIEW
published: 18 March 2019
doi: 10.3389/fimmu.2019.00435

Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming

Mats Ohlin[1]*, Cathrine Scheepers[2,3], Martin Corcoran[4], William D. Lees[5], Christian E. Busse[6], Davide Bagnara[7], Linnea Thörnqvist[1], Jean-Philippe Bürckert[8], Katherine J. L. Jackson[9], Duncan Ralph[10], Chaim A. Schramm[11], Nishanth Marthandan[12], Felix Breden[13], Jamie Scott[14], Frederick A. Matsen IV[10], Victor Greiff[15], Gur Yaari[16], Steven H. Kleinstein[17], Scott Christley[18], Jacob S. Sherkow[19], Sofia Kossida[20], Marie-Paule Lefranc[20], Menno C. van Zelm[21], Corey T. Watson[22] and Andrew M. Collins[23]*

[1] Department of Immunotechnology, Lund University, Lund, Sweden, [2] Center for HIV and STIs, National Institute for Communicable Diseases, Johannesburg, South Africa, [3] Faculty of Health Sciences, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa, [4] Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden, [5] Institute of Structural and Molecular Biology, Birkbeck College, University of London, London, United Kingdom, [6] Division of B Cell Immunology, German Cancer Research Center, Heidelberg, Germany, [7] Department of Experimental Medicine, University of Genoa, Genoa, Italy, [8] BISC Global Inc., Boston, MA, United States, [9] Immunology Division, The Garvan Institute of Medical Research, Darlinghurst, NSW, Australia, [10] Fred Hutchinson Cancer Research Center, Seattle, WA, United States, [11] Vaccine Research Center, National Institutes of Health, Washington, DC, United States, [12] Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada, [13] Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada, [14] Department of Molecular Biology and Biochemistry, Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada, [15] Department of Immunology, Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [16] Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel, [17] Department of Pathology, Yale University, New Haven, CT, United States, [18] Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, United States, [19] Innovation Center for Law and Technology, New York Law School, New York, NY, United States, [20] IMGT®, The International ImMunoGenetics information system® (IMGT), Laboratoire d'ImmunoGénétique Moléculaire (LIGM), CNRS, Institut de Génétique Humaine, Université de Montpellier, Montpellier, France, [21] Department of Immunology and Pathology, Central Clinical School, The Alfred Hospital, Monash University, Melbourne, VIC, Australia, [22] Department of Biochemistry and Molecular Genetics, University of Louisville, Louisville, KY, United States, [23] School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Patrick C. Wilson,
University of Chicago, United States

Reviewed by:
Jiwon Lee,
University of Texas at Austin,
United States
Masaki Hikida,
Akita University, Japan

***Correspondence:**
Mats Ohlin
mats.ohlin@immun.lth.se
Andrew M. Collins
a.collins@unsw.edu.au

Specialty section:
This article was submitted to
B Cell Biology,
a section of the journal
Frontiers in Immunology

Received: 10 December 2018
Accepted: 19 February 2019
Published: 18 March 2019

Citation:
Ohlin M, Scheepers C, Corcoran M, Lees WD, Busse CE, Bagnara D, Thörnqvist L, Bürckert J-P, Jackson KJL, Ralph D, Schramm CA, Marthandan N, Breden F, Scott J, Matsen FA IV , Greiff V, Yaari G, Kleinstein SH, Christley S, Sherkow JS, Kossida S, Lefranc M-P, van Zelm MC, Watson CT and Collins AM (2019) Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. Front. Immunol. 10:435. doi: 10.3389/fimmu.2019.00435

Immunoglobulins or antibodies are the main effector molecules of the B-cell lineage and are encoded by hundreds of variable (V), diversity (D), and joining (J) germline genes, which recombine to generate enormous IG diversity. Recently, high-throughput adaptive immune receptor repertoire sequencing (AIRR-seq) of recombined V-(D)-J genes has offered unprecedented insights into the dynamics of IG repertoires in health and disease. Faithful biological interpretation of AIRR-seq studies depends upon the annotation of raw AIRR-seq data, using reference germline gene databases to identify the germline genes within each rearrangement. Existing reference databases are incomplete, as shown by recent AIRR-seq studies that have inferred the existence of many previously unreported polymorphisms. Completing the documentation of genetic variation in germline gene databases is therefore of crucial importance. Lymphocyte receptor genes and alleles are currently assigned by the Immunoglobulins, T cell Receptors and Major Histocompatibility Nomenclature Subcommittee of the International Union of Immunological Societies (IUIS) and managed in IMGT®, the international ImMunoGeneTics information system® (IMGT). In 2017, the IMGT Group reached agreement with a group of AIRR-seq researchers on the principles of a streamlined process for identifying and naming inferred allelic sequences, for their incorporation into IMGT®. These researchers represented the AIRR Community, a network of over 300 researchers whose objective is to promote all aspects of immunoglobulin and T-cell receptor repertoire studies, including the standardization of experimental and computational aspects of AIRR-seq data generation and analysis. The Inferred Allele Review Committee (IARC) was established by the AIRR Community to devise policies, criteria, and procedures to perform this function. Formalized evaluations of novel inferred sequences have now begun and submissions are invited via a new dedicated portal (https://ogrdb.airr-community.org). Here, we summarize recommendations developed by the IARC—focusing, to begin with, on human IGHV genes—with the goal of facilitating the acceptance of inferred allelic variants of germline IGHV genes. We believe that this initiative will improve the quality of AIRR-seq studies by facilitating the description of human IG germline gene variation, and that in time, it will expand to the documentation of TR and IG genes in many vertebrate species.

Keywords: immunoglobulin, allelic variation, inference, AIRR-seq, IGHV, V(D)J rearrangement

INTRODUCTION

Immunoglobulins (IG) are the main antigen receptors and effector molecules of the B cell lineage, and are expressed either as a component of the membrane-bound B cell receptor (BCR) or as secreted antibodies. They are encoded by large numbers of variable (V), diversity (D), and joining (J) genes, which recombine in developing B cells to generate rearranged V-(D)-J genes. This process, referred to as V-(D)-J rearrangement, occurs at the DNA level and leads to an IG V domain repertoire of immense diversity. The study of such repertoires has recently been revolutionized by high-throughput sequencing (1–4), and this is termed Adaptive Immune Receptor Repertoire (AIRR) sequencing (AIRR-seq). The technical and biological interpretation of AIRR-seq data is facilitated by databases containing reference sequences of all known germline genes (**Figure 1**), but AIRR-seq studies have demonstrated that these databases are presently far from complete (5–8). This compromises analysis of AIRR-seq data in many ways. For example, it can lead to the inaccurate determination of gene utilization frequencies, and the extent to which sequences have been affected by the process of somatic point mutation.

The first complete nucleotide sequence of a human germline heavy chain variable gene was reported in 1980 (9). In 1989 at the Human Gene Mapping (HGM) (10) Workshop in New Haven, starting with the human T cell receptor gamma (TRG) locus genes as a paradigm, the variable, diversity and joining IG and TR genes were officially acknowledged as "genes" just like conventional genes, and under the HGM auspices, IMGT®, the international ImMunoGeneTics information system® (IMGT) was created by University of Montpellier and the Centre National de la Recherche Scientifique (CNRS) (10). Ten years of IMGT biocuration on sequences from human genomic cosmid and artificial chromosome libraries were key to the assembly of the IG loci and their annotation (11–13). The IG and TR gene names, available on the IMGT web site since 1995, were approved by the HUGO Nomenclature Committee (HGNC) in 1999 and are managed by the IMGT Nomenclature Committee (IMGT-NC), the IG, TR and MH nomenclature subcommittee of the International Union of Immunological Societies (IUIS). The functional and open reading frame (ORF) of approved human genes were published with their alleles (203 IG and 168 TR) in two FactsBooks in 2001 (14, 15), and the number of sequences now cataloged by IMGT is shown in **Table 1**.

With this description of the human IG germline genes, the gene identification and mutation description became an integral part of the study of V-(D)-J gene rearrangements. Over the next 20 years, hundreds of thousands of expressed V-(D)-J genes were reported, and dedicated tools and databases were established to facilitate research (10, 16, 17). It soon became possible to compile datasets of hundreds of rearranged human V-(D)-J gene sequences that could be used to analyse the process of V-(D)-J recombination (18, 19). These analyses also demonstrated that such datasets could be used to identify previously unreported allelic variants of known germline IG genes (20).

In 2009, AIRR-seq data were reported for the first time (21, 22). Even in the earliest AIRR-seq studies, thousands of independent V-(D)-J rearrangements could be identified from each subject investigated, and this facilitated the detection of previously unreported polymorphisms (5–8) (**Figure 2**). New allelic variants of IGHV genes were detectable in these AIRR-seq data because the crucial nucleotides that defined these alleles showed up as conspicuous patterns of shared mismatches within alignments to the known germline V gene sequences.

Utilities have now been developed to streamline the identification of allelic variants, and to assign measures of

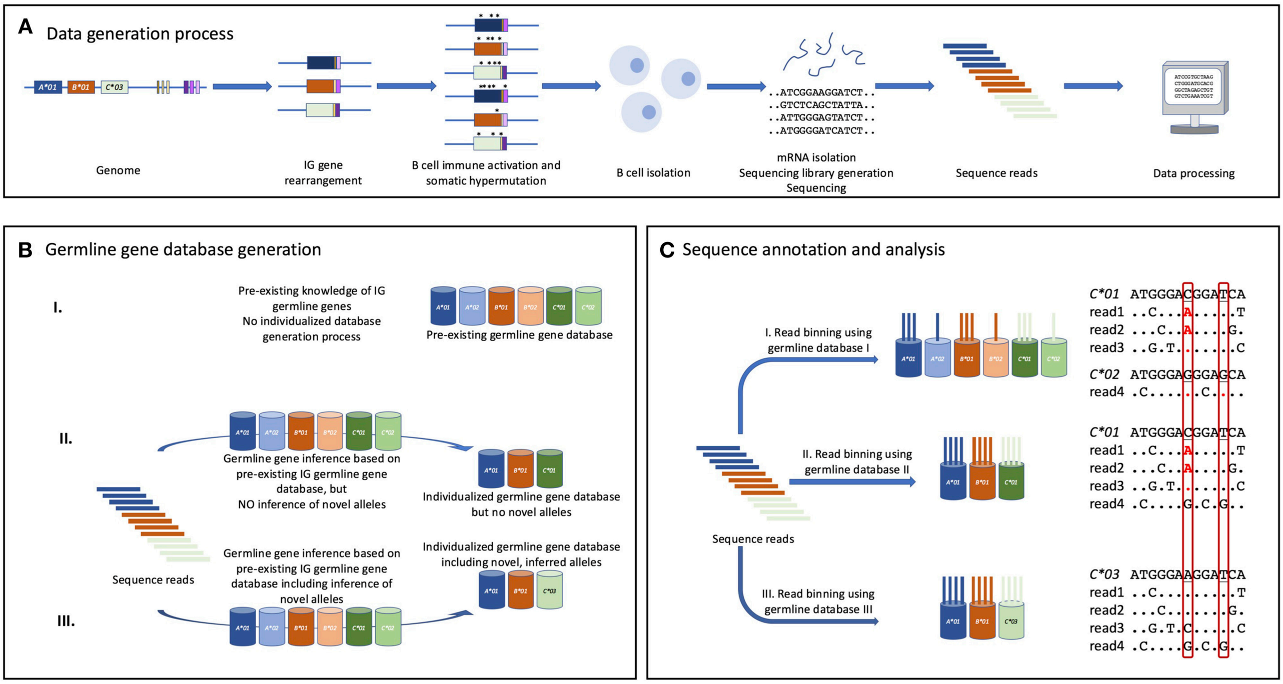
A Data generation process
A*01
B*01
C*03
Genome
IG gene rearrangement
B cell immune activation and somatic hypermutation
B cell isolation
..ATCGGAAGGATCT..
..GTCTCAGCTATTA..
..ATTGGGAGTATCT..
..ATGGGGATCATCT..
mRNA isolation
Sequencing library generation
Sequencing
Sequence reads
Data processing
B Germline gene database generation
I.
Pre-existing knowledge of IG germline genes
No individualized database generation process
Pre-existing germline gene database
II.
Germline gene inference based on pre-existing IG germline gene database, but NO inference of novel alleles
Individualized germline gene database but no novel alleles
Sequence reads
Germline gene inference based on pre-existing IG germline gene database including inference of novel alleles
Individualized germline gene database including novel, inferred alleles
III.
A*01
A*02
B*01
B*02
C*01
C*02
C*03
C Sequence annotation and analysis
I. Read binning using germline database I
II. Read binning using germline database II
III. Read binning using germline database III
Sequence reads
C*01 ATGGGACGGATCA
read1 ..C...A.....T
read2 ...C..A....G.
read3 ..G.T.......C
C*02 ATGGGAGGGAGCA
read4 .C.......C...
C*01 ATGGGACGGATCA
read1 ..C...A.....T
read2 ...C..A....G.
read3 ..G.T.......C
read4 .C....G.C.G..
C*03 ATGGGAAGGATCA
read1 ..C.........T
read2 ...C.......G.
read3 ..G.T.C.....C
read4 .C....G.C.G..

FIGURE 1 | The value of germline IGHV gene inference for detailed AIRR-seq annotation and analysis. **(A)** Germline genes of an individual [here represented by a very limited set of three IGHV genes (A, B, C), and a small number of IGHD (yellow/brown) and IGHJ (purple) genes] are rearranged in cells of the B cell lineage. Following stimulation with antigen many sequences undergo somatic hypermutation and acquire base substitutions (marked *) that may impact subsequent data analysis. An investigated subject's B and plasma cells are collected and typically the cells' transcriptomes are sequenced (e.g., using Illumina MiSeq technology) to generate reads that can be computationally processed. **(B)** A germline IGHV gene database [here represented only by three genes (A, B, C)] will facilitate data analysis, though it is possible to infer genes and alleles without reference to a starting database. The database could be a collection of all known germline IGHV gene alleles (I), or an individualized subset of these (II) that best fits the set of sequence reads that are to be analyzed. Finally, computationally inferred novel germline IGHV gene alleles can be introduced into the individualized germline gene database (III) to even better account for the diversity observed in the experimentally generated sequence dataset. **(C)** Each sequence read is binned to the most appropriate germline gene/allele available in the used germline gene database. If germline gene alleles are present in the database but not in the subject's genotype, some reads will be binned to them as a consequence of base changes introduced by somatic hypermutation (or sequencing errors), resulting in a partial incorrect assignment of germline gene allele origin and consequently of the associated analysis of the mutational pattern. Detailed annotations of part of the sequences are provided for reads binned to alleles of gene *C*. In this example, the investigated subject has an allele (*C*03*) of this gene that is not represented in the original pre-existing germline IGHV gene database. Two bases that differ between one or several alleles in the database and C*03, and thus may be misinterpreted in mutational analysis, are boxed. Unless valid inference of novel germline genes is also performed, the mutational analysis will substantially misinterpret the mutational pattern (highlighted in red letters/dots) targeting this gene. Dots indicate identity to the germline gene to which it is compared.

confidence to each inference (23–27). These utilities employ a variety of inference methodologies, as they have been designed for the analysis of different kinds of data. IgDiscover, for example, is best suited to the analysis of relatively unmutated sequences (23), whereas TIgGER (24) and partis (26) are specifically designed to analyse data that include both unmutated and mutated sequences. To date, 58 sequences have been inferred in this way (see **Table 1**), and can be found in the Immunoglobulin Polymorphism database (IgPdb) (http://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/).

The identification of these previously unreported polymorphisms has remained unknown to many researchers because such variants lie outside the scope of the widely-used IMGT/V-QUEST reference directory of germline sequences (28). This emerged as an early concern of the AIRR Community (https://www.antibodysociety.org/the-airr-community/), a grassroots organization that was founded in 2015 to address the challenges surrounding the generation, analysis and use of AIRR-seq data (29). In 2018, this community formally joined The Antibody Society, a non-profit trade association dedicated to the field of antibody research and immunotherapeutics.

In 2017, the AIRR Community and IMGT agreed to an approach for evaluating the veracity of inferred germline-gene sequences, and for the incorporation of validated sequences into the IMGT Reference Directory. The Germline Database (GLDB) Working Group of the AIRR Community was formed to develop the necessary policies and procedures, and the Inferred Allele Review Committee (IARC) was formed to critically evaluate submitted inferences.

Here we present challenges faced in inferring novel IGHV sequences from AIRR-seq data, and outline strategies for their mitigation. The process for submitting inferred sequences to the IARC is also described. It is our aim that this initiative of the AIRR Community will contribute to a more complete description of human genetic variation, thereby improving the quality of AIRR-seq studies. Human IGHV genes are the focus of this

TABLE 1 | Numbers of human IGHV genes and alleles reported in the IMGT repertoire and in the IgPdb database of inferred alleles.

	IMGT[a]		IgPdb[b]	
Subgroup	**Genes**	**Alleles**	**Genes**	**Alleles**
IGHV1	12	45	8	21
IGHV2	4	29	2	4
IGHV3	30	110	11	18
IGHV4	11	79	8	13
IGHV5	2	9	1	2
IGHV6	1	2	0	0
IGHV7	2	6	0	0

[a] *IMGT genes and allele counts include sequences reported as Functional sequences and Open Reading Frames. The IMGT repertoire was accessed on 11/02/2019.*

[b] *Sequences in IgPdb that have only been identified by genomic sequencing, and sequences that extend previously reported but truncated sequences are not included. Eleven sequences (IGHV1-2*05, IGHV1-2*06, IGHV1-8*03, IGHV1-69*15, IGHV1-69*17, IGHV2-70*15, IGHV3-11*05, IGHV3-11*06, IGHV3-13*05, IGHV3-43D*04 and IGHV3-64D*06) that were first discovered by inference but are now present in the IMGT repertoire are also not included here.*

discussion, though the challenges surrounding the inference of other IG and TR germline genes in human and non-human species are likely to be similar. We anticipate that over time this initiative will expand to the documentation of IG and TR genes in all vertebrate species.

GERMLINE GENE INFERENCES: CHALLENGES, AND STRATEGIES FOR MINIMIZING ERRONEOUS INFERENCES

Reports of inferred antibody sequences have not been immediately and universally accepted, in part because alternative explanations can account for observed nucleotide differences in IG genes (see **Figure 2**). Uniquely, IG genes within activated B cells undergo secondary diversification by somatic hypermutation (SHM) (30). During an immune response, an IGHV gene with a 300 bp length will commonly accumulate 15–20 somatic point mutations (31, 32) and much higher levels of mutations can be observed (33).

The datasets of Sanger sequences that underpinned the first inferred IGHV sequences were very small—in some cases, just six or seven sequences (20). This raised the possibility that these sequences were mutated versions of known alleles. Importantly though, many of the early inferences have now been confirmed by genomic sequencing (20, 34, 35), lending support to the validity of the inference process. Today, the availability of large AIRR-seq datasets gives much greater confidence in the inference process, but challenges remain. These challenges have their origins in the biology of the B cell and of the antibody repertoire, as well as in technical issues affecting the preparation and sequencing of recombined V-(D)-J gene libraries.

The following strategies and tests will aid in the identification of real allelic variants while minimizing the reporting of erroneous inferences.

- Inferences must be made from AIRR-seq data of the highest quality. Experimental strategies to ensure such quality in library generation and sequencing of IG transcripts are now well-established (36, 37), and the assessment of the quality of library generation and of sequencing, using synthetic mRNA spike-ins, is a strategy that can build confidence in inferences made from a dataset (38–41). Proof-reading enzymes with minimal error rates should always be used (42), and putative polymorphisms should be assessed in light of the different types of sequencing errors (base insertions, deletions and substitutions) that are associated with the different sequencing technologies (43). Such errors can be specifically enriched at particular sequence motifs (44), and if these motifs are present in a germline gene, the errors may suggest the existence of a novel allele (7, 45).
- A vital step in the pre-processing of raw sequence data is the removal of reads with a low average quality, but Phred scores should also be assessed for critical nucleotides in individual reads that have contributed to a particular inference. Poor read quality of single nucleotides may result in erroneous inferences (7, 45).
- Correction of sequencing errors and PCR artifacts can be achieved by the use of unique molecular identifiers (UMI). UMIs are introduced during library preparation, labeling each individual transcript prior to amplification. Subsequent consensus building of reads employing identical UMIs can largely remove erroneous bases (46). Technical or biological replicates can also be used to validate sequences and increase confidence that artifacts have been properly discarded.
- Incomplete PCR amplifications create problems. An incompletely amplified product generated in one cycle may later anneal to a similar but distinct template, resulting in the amplification of a hybrid sequence (**Figure 3**) (47, 48). Such chimeric amplification products are often observed in datasets of IG transcripts (49), and unless appropriate filters are applied to AIRR-seq data, these chimeras can masquerade as novel alleles. Preparing libraries with minimally detectable PCR bands helps reduce the problem of chimerism (49, 50), but this strategy is incompatible with some research objectives.
- The detection and elimination of chimeric sequences can be a valuable step in the pre-processing of data. Manual identification of chimeric sequences involves assessment of the distribution of apparent mutations along the length of a sequence. Chimeric sequences often appear to have somatic point mutations clustered at one or the other end of the sequence, and utilities have been developed to automate the detection of sequences with such a non-random distribution of apparent mutations (51).
- Very large AIRR-seq datasets are required if variants of some IGHV genes are to be identified. Reports from analysis of peripheral blood B cells show that usage frequencies of particular IGHV genes in V-(D)-J rearrangements can be as high as 20% for IGHV3-23*01 (52), but as low as 0.01% for rearranged genes incorporating IGHV3-13, IGHV4-28, or IGHV7-81 (5). Rarely utilized IGHV genes will only be present in convincing numbers in the very largest V-(D)-J

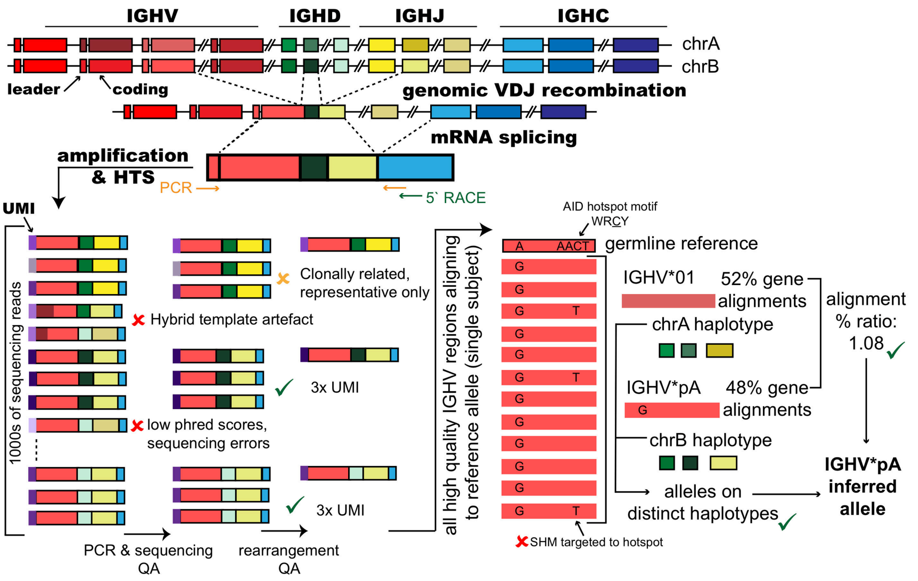

FIGURE 2 | Inference of polymorphic IGHV gene alleles from immunoglobulin repertoire sequencing datasets. The genes that encode immunoglobulin heavy chain variable regions are generated through genomic recombination of single genes of three different types; variable (IGHV), diversity (IGHD), and joining (IGHJ). Each of these gene types are present in the genome as a set of tandem genes that are both polymorphic and polygenic and include approximately 50 IGHV genes, 27 IGHD genes and 6 IGHJ genes. At the mRNA level, splicing joins the rearranged V-(D)-J gene with the IGH constant region genes that confer the isotype to the IG. Repertoire sequencing studies amplify mRNA transcripts, via cDNA, by methods such as targeted PCR using leader region forward primers coupled with reverse primers specific for the IGHC or by using 5′ RACE primed from the IGHC genes. To be suitable for inference, amplification protocols must capture the complete V-(D)-J rearranged coding region and cannot use primer that bind within the V region. Amplification strategies may add unique molecular identifiers (UMIs) as part of cDNA synthesis to tag individual RNA transcripts. High throughput sequencing of a V-(D)-J gene library can generate many thousands of reads for single subjects. Reads are quality assured (QA) to remove sequencing errors; for example, reads with low quality (phred) scores, and PCR artifacts such as hybrid amplicons that have resulted from single stranded DNA from previous PCR cycles acting as primers in future cycles creating hybrid or chimeric templates that are derived from two original amplicons. For inference, IGH from clonally expanded B cells, which have each originated from a single progenitor cell must be reduced to a single representative sequence (for example, the clone member with the fewest mutations) to prevent over-counting. IGH carrying the same UMI can greatly increase confidence that the V-(D)-J rearrangement was in the original pool. In the absence of UMIs, read counts for unique IGH can provide some confidence. Finally, V-(D)-Js are aligned to germline reference datasets that report alleles for the population. Shared "mismatches" relative to the closest germline gene among many sequences from the same subject at position(s) that are not motifs for somatic hypermutation (SHM) in sequences expected to be unmutated (naïve B cells) or have low mutation (IgM+) can be suggestive of the putative allelic variant in the subject's genotype. When a polymorphism is inferred at a heterozygous locus, confidence in the inference can be greatly increased if the putative allele is considered in the context of any other expressed alleles for the gene, with approximately equal expression of the two alleles, and the haplotype, which can phase the gene alleles to their respective chromosomes.

datasets. Large datasets are also needed if the final nucleotides of a germline IGHV sequence are to be determined. The uncertainties surrounding the nucleotides at the 3′ end of the sequence are a consequence of the variability of the gene ends, produced by the processes of exonuclease removal and N nucleotide addition. Biases in these processes can result in the generation of relatively common motifs that may be mistaken for germline-encoded nucleotides (53–56). Of special note, the last base of a germline sequence may not be the most common base in rearranged sequences.

- Somatic point mutations accumulate in IG-encoding genes at a rate of about one mutation per 1,000 bp per cell division within the germinal center reaction (57, 58). The existence of mutational hotspots (59–61) that can target specific germline IGHV genes (62, 63) means that it is inevitable that there will be some shared mutations in any dataset that includes mutated sequences. Some IGHV genes have positions that can be mutated in >30% of class-switched sequences (24). Very high levels of mutation can occur at positions far removed from the regions encoding complementarity determining regions (CDR) of an IGHV sequence, and even at positions outside conventional mutational hotspots (62, 64). For these reasons, inferences of new germline IGHV genes using datasets of mutated sequences are more likely to be erroneous.

Somatic point mutations may be mistaken for germline-encoded nucleotides, but this issue is substantially reduced if sequences are derived from less-mutated cell populations. This can be achieved by the amplification of IgM-encoding transcripts through the use of constant region-specific primers. The issue is partially addressed by the amplification of sequences from sorted B cells displaying a naïve phenotype. More highly mutated datasets can, however, still be the source of reliable inferences if appropriate analytical tools are used. Both the TIgGER and partis software suites, for instance, are designed to use patterns of apparent mutation to infer novel alleles (24, 26). While taking different overall approaches, they both use regression-based statistical tests to identify polymorphisms at positions that

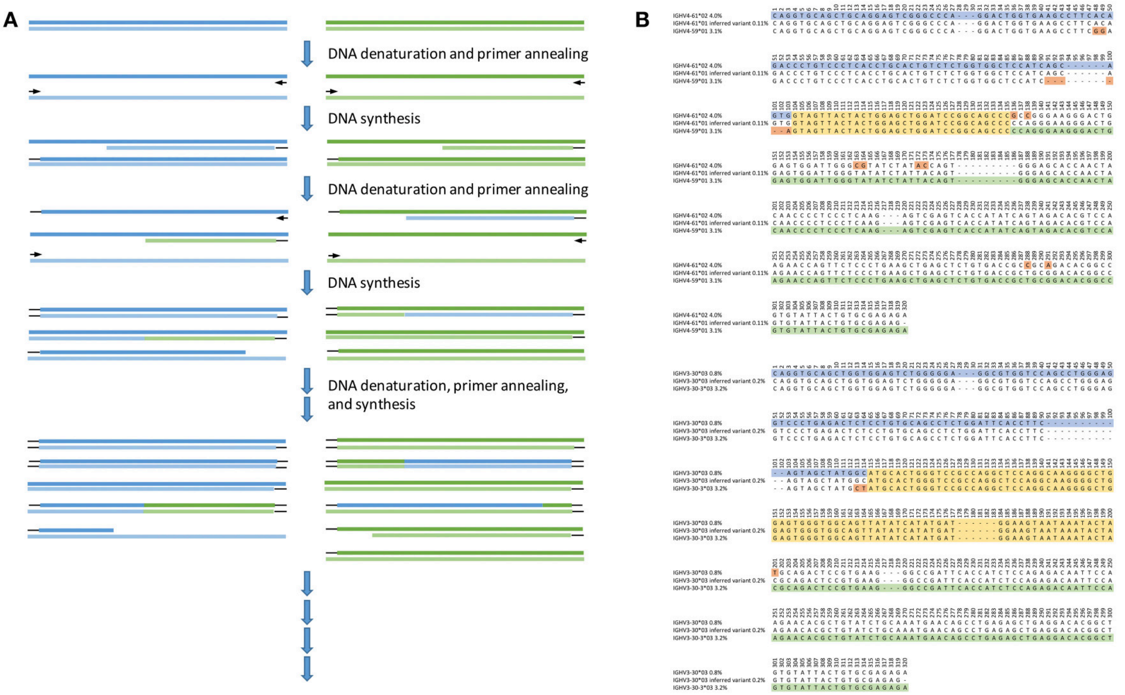
A
DNA denaturation and primer annealing
DNA synthesis
DNA denaturation and primer annealing
DNA synthesis
DNA denaturation, primer annealing, and synthesis
B

FIGURE 3 | Likely principle of generation of hybrid molecules through cross-over caused by incomplete products during previous PCR cycles **(A)**, and examples of inferred sequences **(B)** that may have an origin in cross-over events during the library generation process. Shown are the inferred sequences (middle) and the parts of other germline sequences that are inferred to be present in the same subject and that through an artifactual cross-over event may have created the inferred novel allele. The region in which cross-over events may have occurred is shown in yellow. The 5′- and 3′- sequences unique to one of the potential donor sequences are shown in blue and green, respectively. Mismatches between the two potential donor sequences are highlighted in orange. The frequency of reads associated with each inference is shown after the name, indicating that some but not all of these inferences may be removed by cut-offs defined in terms of allelic ratios. Base numbering is according to the IMGT numbering scheme.

appear to be recurrently mutated, in sequences that are otherwise relatively unmutated.

- A useful test of an inferred allelic variant is to consider the sequence in the context of other alleles of the gene that may be present in the dataset. For single-copy genes, only two alleles should be present in the genotype of an individual. If an inference suggests that three alleles of a particular gene are present in an individual genotype, the inference should be further investigated. More than two named variants of some genes can be present in an AIRR-seq dataset for gene names without genomic information on the haplotype copy number, as a result of copy number variation (CNV) (5, 6, 8, 35, 46, 65). It is highly likely, for example, that some named allelic variants of the IGHV1-69 gene are actually variants of the duplicate IGHV1-69D gene. Genotypes may also include three or more alleles of one or other of the highly similar IGHV4-4, IGHV4-59, and IGHV4-61 genes, for the genomic location of some sequences associated with these genes is uncertain (7, 45, 52, 66). It seems likely, for example, that the IGHV4-59*08 sequence in some subjects is actually a variant of the IGHV4-61 gene (66). In view of these complications, an evaluation of some inferences must be made with reference to alleles of several genes that may be present in the genotype of the individual. Genomic data of a single cell or individual will remain necessary to unambiguously assign expressed genes with CNV.
- Alleles at heterozygous loci are usually expressed at similar frequencies (52), while inferred sequences suggested by sequencing errors or somatic point mutations are usually present at relatively low frequencies. The calculation of the percentage of alignments to a gene that involve the inferred allele is therefore a simple test that can be used to identify false inferences. Although the IARC will affirm inferred alleles that are observed in just 10% of all alignments to a particular gene, additional analysis may be required to support inferences that imply expression at a low level. We recognize that this will make it more difficult to infer some alleles. CNV, mentioned above, will also complicate the interpretation of this measure of allele expression, whereas recombination signal (RS) sequence variation and other non-coding region variation could lead to abnormal allele expression levels (67). For all these reasons, the allele expression test has limitations.
- The validity of an inference can be demonstrated if all V-(D)-J sequences containing the inference in an AIRR-seq dataset are associated with just one of the two chromosomes of the individual. Such validation can be done using haplotype analysis as outlined in **Figure 4**. This is a method that was developed for human AIRR-seq studies (6, 53, 68), and it is

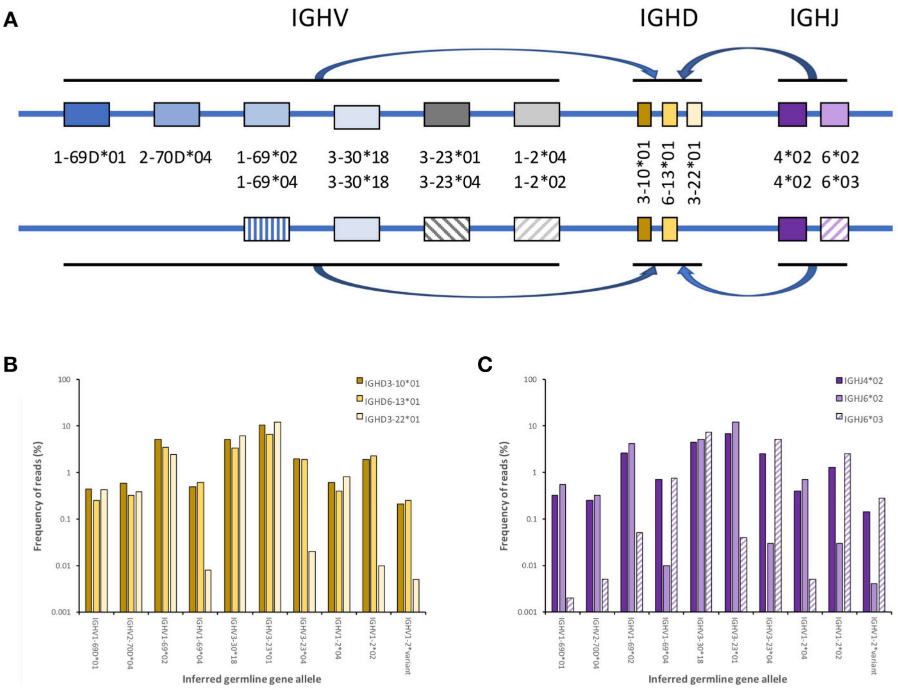

FIGURE 4 | Principle of inference haplotyping defining the AIRR-encoding genes associated with each of the chromosomes carrying the relevant locus, here demonstrated by genes encoding human IG heavy chains. Germline genes involved in V-(D)-J rearrangements are from genes harbored on each chromosome only (i.e., in *cis*), as illustrated with a small number of the genes that actually populate the IGH locus **(A)**. It is possible, using large sequence datasets, to computationally define the association of each IGHV allele to one of the two haplotypes using their association to e.g., different alleles of an IGHJ gene (commonly IGHJ6), if such different alleles are present in the genotype. These alleles serve as anchors in the haplotyping process. An IGHV allele that resides on both chromosomes will rearrange to both alleles of the heterozygous IGHJ gene, whereas IGHV alleles that reside on only one chromosome should primarily be found rearranged to one of the alleles of the heterozygous IGHJ gene **(C)**. In the case of haplotype differentiating expression of IGHD genes (or allelic differences in one or several IGHD genes) these differences can similarly be used as anchors to visualize IGHV allele distributions between haplotypes **(B)**. Such inference can be used to raise confidence in specific allele calls, as incorrectly inferred alleles are likely to associate with the same haplotype as another allele of the same (or a very similar allele) that also exist in the haplotype. This is exemplified here by the haplotyping of an artifactual inference of a novel allele (IGHV1-2*variant) that has a similar association to haplotypes as IGHV1-2*02. [For specific examples, see Kirik et al. (22, 53)].

increasingly being used to support reported inferences (7, 45, 52). Haplotyping is only possible for the validation of IGHV gene inferences in subjects who are heterozygous at IGH loci beyond the IGHV locus region. Anchors for the haplotype inference of IGHV genes are most commonly IGHJ6 alleles (IGHJ6*02 and IGHJ6*03), but heterozygosity at the IGHD2-8 and IGHD2-21 loci can also allow them to be used (7, 45, 52, 68). It is likely that novel long-read high-throughput sequencing platforms will soon make it possible to use IGH constant region genes as haplotype anchors as well.

SUBMISSION OF INFERENCES AND DATA DEPOSITION

IARC and the GLDB WG strive to provide the community with open, transparent and reusable information on inferred genes. To this end, a web-based service termed Open Germline Reference Database (OGRDB) has been set up to facilitate the submission and evaluation of inferences as well as the subsequent retrieval of inferred genes accepted by IARC. In addition, the inferred sequence and the NGS data supporting it have to be deposited in general purpose sequence repositories of the International Nucleotide Sequence Database Collaboration framework to allow re-analysis by third parties and ensure long-term availability of the data. The detailed workflow for data submission is available at OGRDB (https://ogrdb.airr-community.org). In brief, it covers the following steps:

- Verification that the complete raw data of the underlying experiment is available via the Sequence Read Archive (SRA). If possible, the SRA and associated metadata records should be compliant with the Minimal Information about Adaptive Immune Receptor Repertoire (MiAIRR) standard (69).
- Deposition of reads supporting the gene inference to SRA. Note that this submission is in addition to the publication of the complete read data of a given set of experiments.
- Submission of the inferred sequence to GenBank/TPA, depending on the origin of the data on which the inference is based:

  a. First-party data (the inference is performed on one's own datasets) is submitted to GenBank.
  b. Third-party data (inference performed on datasets produced by others) is submitted to GenBank's Third Party Annotation (TPA) section.

- Submission of the inferred sequence and the associated information about the inference procedure as well as the accession IDs of the INSDC submission to IARC via the OGRDB interface.

Each inference must be made from data that originates from a single individual. The standardized submission protocol incorporates metadata related to the individual, as well as to the generation, processing and analysis of the individual's sequences. It also provides data that gives the genotypic context in which an inference should be assessed, and helps identify confounding factors that should be considered.

Currently, data used for germline IGHV gene inference are often generated from PCR-amplified IG transcripts using Illumina's MiSeq technology, as it provides sufficient read length and depth. The IARC will, however, consider inferences and determinations made in other ways. The IMGT-NC requires genomic sequencing of IGHV genes, including the complete leader sequence and associated Recombination Signal sequence (V-RS). Genomic sequences that are not suitable for submission to IMGT-NC will be considered by the IARC if they include the complete IGHV coding region. Partial genomic sequences may also be considered by IARC as evidence in support of an inference from AIRR-seq data. Direct RNA sequencing (70) may also come to play an important role in defining germline IGHV genes in the future.

Inferences must be made from full-length sequencing reads. In contrast, many studies employ primers that anneal within the IGHV sequences themselves, such as the well-validated BIOMED-2 primer set (71). Although sequences generated in this way may be suitable for many research purposes, the partial sequences that can be inferred from such datasets are not suitable for submission to IARC. Submitted sequences must be full-length V-REGION sequences, from base 1 to at least base 318 of the IGHV sequence, according to the IMGT numbering system. Inferences generated using primers that anneal within the sequence should not be submitted to the IARC.

Inference may be carried out using a diversity of computational methodologies. The IARC is agnostic to the investigator's choice of inference methodology as long as it is validated, published, publicly available, and well-documented.

We believe that the identification of dependable, curated gene sets, to which this effort contributes, is a public good. To that end, affirmed sequences, and the submissions that support them are published by IARC under the Creative Commons CC0 license (https://creativecommons.org/publicdomain/zero/1.0/legalcode), allowing their use for any purpose without restriction under copyright or database law.

THE EVALUATION AND DECISION-MAKING PROCESS

The affirmation of submitted inferences requires the unanimous support of the IARC, and this may only be possible after the provision of additional information by the Submitter. The deliberations of this Committee may differ depending on the biological context in which particular sequences are observed and on the process of inference. Particular attention will be paid to:

- The frequency of V-(D)-J rearrangements that include the inferred sequence. Inferences that appear to be very rarely represented in the IG repertoire are at high risk of being incorrect inferences. To guard against this, inferences of sequences that are seen at a frequency of 0.05% or less will not generally be affirmed.
- The number and frequency of unmutated sequences representing the inferred sequence.
- The presence of the inferred IGHV sequence in a diversity of V-(D)-J rearrangements. The sequence needs to be seen in association with different IGHJ genes and in rearrangements with varying CDR3 lengths. This guards against the possibility that sequences that support the inference are clonally-related sequences.
- The number of alleles assigned to the relevant gene or to the set of highly similar genes.
- The distribution of reads between an inferred allele and other alleles of that particular gene, calculated using unmutated sequences. Inferences with low expression frequencies may require additional supporting evidence.
- The outcome of haplotype analysis, where such analysis is possible.
- Evidence that PCR artifacts, such as cross-over events involving other genes and alleles of the subject's genotype, do not explain the inference. Evidence could include a demonstration of the absence of cross-over effects in sequencing libraries of germline gene standards analyzed in parallel to the subject's expressed IG repertoire (38), or demonstration of the systematic identification and removal of sequences with evidence of cross-over effects prior to inference, or analysis of the extent of shared CDR3 sequences between different V-(D)-J gene rearrangements.
- Evidence supporting the reported 3′-end of an inferred germline IGHV gene. The final base of an IGHV gene sequence cannot be inferred with confidence (55, 56) unless additional investigations are undertaken. If a sequence is reported up to and including base 320, the final base will only be affirmed by IARC if supporting analysis is provided.
- Sequencing of part of an inferred allele, from non-B cell genomic DNA.

An assessment will result in one of three outcomes. If a sequence is affirmed as a valid inference, it will be assigned an IARC sequence name and a summary of evidence in support of the inference will be documented in an Inferred Sequence Documentation Sheet. This will be made publicly available at the AIRR community website. It will also be reported to IMGT-NC with an individual GenBank accession number for inclusion in the IMGT Reference Directory. When a sequence is affirmed for the first time, it will be reported as a Level 1 Sequence. If affirmed a second time, it will be reported as a Level 2 Sequence, and if affirmed a third time, it will be reported as a Level 3 Sequence. It is important that researchers continue to notify the IARC of later identification of Level 1 and Level 2 Sequences, so that they can rise to higher tiers. This will promote acceptance of the inferences within the research community. The IARC will not consider additional inferences of a sequence following its elevation to Level 3.

If evidence in support of a sequence does not reach the level of certainty required for immediate affirmation, the sequence may remain "under review". An Inferred Sequence Documentation Sheet will be completed, and the sequence will be assigned an IARC name, but it will not be publicly reported. Such sequences will be re-assessed if additional supporting information becomes available, or if identical inferences are later submitted to IARC. If a later inference supports the elevation of the sequence to Level 1, the original inference will be credited in the documentation of the sequence.

If there is insufficient evidence to allow a sequence to remain "under review", details of the submission will be retained by IARC, but the submission will not be a part of any future re-assessments.

Inferred alleles will be named using a modification of the IMGT nomenclature (72), incorporating:

- the gene locus (e.g., IGHV, IGKV, IGLV for genes of heavy, kappa light, and lambda light chain loci, respectively);
- the most similar gene at the time of submission in the IMGT/V-QUEST reference directory (28), or in the case of multiple, most-similar genes, using the name with the lowest alphanumeric value;
- an allele number, preceded by an "i" to indicate its discovery by inference. Assigned allele numbers for any gene will be consecutive, and the first inferred allele will be designated the *i01 allele (e.g., IGHV1-2*i01).

A given allele number for a specific gene will be uniquely associated with a specific sequence. If the sequence is incorporated into the IMGT Reference Directory, it will be assigned a new name by IMGT-NC based on the chronological rule and reported to the IUIS/IMGT Nomenclature Committee. The inferred allele name will not be reused and records of the inference will be permanently maintained. Similarly, if evidence emerges suggesting that a particular inference was made in error, the sequence will be removed from any listing of affirmed sequences, but the name and documentation sheets will remain permanently associated with the sequence.

Germline gene databases currently include entries that are incomplete at the 5′ and/or the 3′ end. The inference process could allow the extension of incomplete sequences, as is the case with the sequence IGHV4-4*i01 that is reported here (see **Figure 5**). A sequence of this kind could be a longer representation of the previously reported allele, or it could be a very similar sequence that varies from the original sequence at its ends. The IARC will not attempt to resolve this ambiguity and will simply assign an inferred allele name to the new sequence.

AFFIRMED NOVEL ALLELES

Using the recommendations and policies outlined above, as of August 31, 2018, the IARC has approved five novel alleles at Level 1 (**Figure 5**) and nine inferred alleles remain "under review" (data not shown). Four of the inferred alleles were affirmed from data submitted by the data-generating author (73), of which three were from one donor and one was from a second donor.

IGHV1-2*i01 differs from IGHV1-2*02, its closest matching allele from IMGT, by a single substitution (t163c), resulting in an amino acid change (W55R). Exact matches to the inference were seen in 2.19% of those donor sequences that were determined to be unmutated rearrangements. A second allele for IGHV1-2 (IGHV1-2*04) was observed within the subject's genotype, however IGHV1-2*i01 was seen in 71% of alignments to IGHV1-2. This sequence has been previously described in multiple subjects from AIRR-seq (5, 7, 24), and from genomic DNA (8) and it is listed in the IgPdb database as IGHV1-2*p06. Since this inference was affirmed by IARC, it has been confirmed using full-length genomic DNA sequencing and was recently accepted (24 July 2018) by IMGT-NC as IGHV1-2*06 (Report 2018-1-0724) (http://www.imgt.org/IMGTindex/IMGT-NC.php).

IGHV1-3*i01 was present in 1.17% of the donor's sequences, and differs from IGHV1-3*01 by a single nucleotide (g172a), resulting in an amino acid change (A58T). This sequence has not been observed previously.

IGHV4-30-4*i01 was observed in 1.3% of the donor's sequences, and also has a single nucleotide difference (t120c) compared to its closest matching IMGT allele, IGHV4-30-4*01, however this did not result in an amino acid change. It has been observed in multiple individuals from genomic DNA sequencing (8) and in a single individual from AIRR-seq (63). It was previously listed as IGHV4-30-4*p08 in the IgPDb database.

IGHV4-4*i01 was observed in 0.6% of the donor's sequences. It may be an extension of the existing IGHV4-4*03 allele described in IMGT, involving bases 312-319.

The last of the five affirmed alleles, IGHV3-43D*i01, was submitted as a third party annotation dataset (74) and although it was observed at a low frequency (0.07%) in the subject's repertoire, it could be accepted as a Level 1 sequence. It has been observed previously in multiple individuals from AIRR-seq studies (7), and as genomic DNA (8), and is listed as IGHV3-43*p04 in IgPdb. It has also been observed in a fosmid clone (GenBank: AC242184) that was not annotated in detail. At the time of its acceptance by IARC, this sequence differed from its closest matching IMGT sequence IGHV3-43D*01 (now renamed as IGHV3-43D*03) by a single nucleotide (c195a), however this does not result in an amino acid change. Since the affirmation by IARC of this novel inferred allele, it has been accepted (October 4, 2018) by IMGT as IGHV3-43D*04, based on genomic evidence.

For all five affirmed alleles, the genotype and allele frequencies were within the IARC guidelines. Where possible, haplotype analysis confirmed the validity of the inferences, and cross-over artifacts were ruled out. The Inference Documentation Sheets for these inferred alleles can be found at the OGRDB website (https://ogrdb.airr-community.org).

CONCLUSION

Germline IGHV, IGHD, and IGHJ genes constitute the building blocks of IG V domain diversity, and so have a direct bearing on the functional B cell immune response. The formation of

A

Name of inferred allele	Closest allele in IMGT database at the time of inference	Modification relative to closest IMGT allele	Other names associated with the sequence	Frequency in donor's repertoire	Allele Percentage
*IGHV1-2*i01*	*IGHV1-2*02*	T163C	*IGHV1-2*p06; IGHV1-2*06*	2.2%	71%
*IGHV1-3*i01*	*IGHV1-3*01*	G172A	none	1.2%	57%
*IGHV4-30-4*i01*	*IGHV4-30-4*01*	T120C	*IGHV4-30-4*p08*	1.3%	100%
*IGHV4-4*i01*	*IGHV4-4*03*	Extension of sequence with bases 312-319	none	0.6%	25%
*IGHV3-43D*i01*	*IGHV3-43D*01*	C195A	*IGHV3-43*p04; IGHV3-43D*04*	0.07%	100%

B

```
                   1   2   3   4   5   6   7   8   9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  35  36  37  38
                  Q   V   Q   L   V   Q   S   G   A   E   V   K   K   P   G   A   S   V   K   V   S   C   K   A   S   G   Y   T   F   T   G   Y   Y
IGHV1-2*02       CAG GTG CAG CTG GTG CAG TCT GGG GCT GAG GTG AAG AAG CCT GGG GCC TCA GTG AAG GTC TCC TGC AAG GCT TCT GGA TAC ACC TTC ACC GGC TAC TAT
IGHV1-2*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  62  63  64  65  66  67  68  69  70  71  72  74
                  M   H   W   V   R   Q   A   P   G   Q   G   L   E   W   M   G  W/R  I   N   P   N   S   G   G   T   N   Y   A   Q   K   F   Q   G
IGHV1-2*02       ATG CAC TGG GTG CGA CAG GCC CCT GGA CAA GGG CTT GAG TGG ATG GGA TGG ATC AAC CCT AAC AGT GGT GGC ACA AAC TAT GCA CAG AAG TTT CAG GGC
IGHV1-2*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... C.. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
                  R   V   T   M   T   R   D   T   S   I   S   T   A   Y   M   E   L   S   R   L   R   S   D   D   T   A   V   Y   Y   C   A   R
IGHV1-2*02       AGG GTC ACC ATG ACC AGG GAC ACG TCC ATC AGC ACA GCC TAC ATG GAG CTG AGC AGG CTG AGA TCT GAC GAC ACG GCC GTG TAT TAC TGT GCG AGA GA
IGHV1-2*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .-
```

```
                   1   2   3   4   5   6   7   8   9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  35  36  37  38
                  Q   V   Q   L   V   Q   S   G   A   E   V   K   K   P   G   A   S   V   K   V   S   C   K   A   S   G   Y   T   F   T   S   Y   A
IGHV1-3*01       CAG GTC CAG CTT GTG CAG TCT GGG GCT GAG GTG AAG AAG CCT GGG GCC TCA GTG AAG GTT TCC TGC AAG GCT TCT GGA TAC ACC TTC ACT AGC TAT GCT
IGHV1-3*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  62  63  64  65  66  67  68  69  70  71  72  74
                  M   H   W   V   R   Q   A   P   G   Q   R   L   E   W   M   G   W   I   N  A/T  G   N   G   N   T   K   Y   S   Q   K   F   Q   G
IGHV1-3*01       ATG CAT TGG GTG CGC CAG GCC CCC GGA CAA AGG CTT GAG TGG ATG GGA TGG ATC AAC GCT GGC AAT GGT AAC ACA AAA TAT TCA CAG AAG TTC CAG GGC
IGHV1-3*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... A.. ... ... ... ... ... ... ... ... ... ... ... ... ...
                  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
                  R   V   T   I   T   R   D   T   S   A   S   T   A   Y   M   E   L   S   S   L   R   S   E   D   T   A   V   Y   Y   C   A   R
IGHV1-3*01       AGA GTC ACC ATT ACC AGG GAC ACA TCC GCG AGC ACA GCC TAC ATG GAG CTG AGC AGC CTG AGA TCT GAA GAC ACG GCT GTG TAT TAC TGT GCG AGA GA
IGHV1-3*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .-
```

```
                   1   2   3   4   5   6   7   8   9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  34  35  36
                  Q   V   Q   L   Q   E   S   G   P   G   L   V   K   P   S   Q   T   L   S   L   T   C   T   V   S   G   G   S   I   S   S   G   D
IGHV4-30-4*01    CAG GTG CAG CTG CAG GAG TCG GGC CCA GGA CTG GTG AAG CCT TCA CAG ACC CTG TCC CTC ACC TGC ACT GTC TCT GGT GGC TCC ATC AGC AGT GGT GAT
IGHV4-30-4*i01   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  63  64  65  66  67  68  69  70  71  72
                  Y   Y   W   S   W   I   R   Q   P   P   G   K   G   L   E   W   I   G   Y   I   Y   Y   S   G   S   T   Y   Y   N   P   S   L   K
IGHV4-30-4*01    TAC TAC TGG AGT TGG ATC CGC CAG CCC CCA GGG AAG GGC CTG GAG TGG ATT GGG TAC ATC TAT TAC AGT GGG AGC ACC TAC TAC AAC CCG TCC CTC AAG
IGHV4-30-4*i01   ... ... ... ..C ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
                  S   R   V   T   I   S   V   D   T   S   K   N   Q   F   S   L   K   L   S   S   V   T   A   A   D   T   A   V   Y   Y   C   A   R
IGHV4-30-4*01    AGT CGA GTT ACC ATA TCA GTA GAC ACG TCC AAG AAC CAG TTC TCC CTG AAG CTG AGC TCT GTG ACT GCC GCA GAC ACG GCC GTG TAT TAC TGT GCC AGA GA
IGHV4-30-4*i01   ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .-
```

```
                   1   2   3   4   5   6   7   8   9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  35  36  37
                  Q   V   Q   L   Q   E   S   G   P   G   L   V   K   P   P   G   T   L   S   L   T   C   A   V   S   G   G   S   I   S   S   S   N
IGHV4-4*03       CAG GTG CAG CTG CAG GAG TCG GGC CCA GGA CTG GTG AAG CCT CCG GGG ACC CTG TCC CTC ACC TGC GCT GTC TCT GGT GGC TCC ATC AGC AGT AGT AAC
IGHV4-4*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  63  64  65  66  67  68  69  70  71  72  74
                  W   W   S   W   V   R   Q   P   P   G   K   G   L   E   W   I   G   E   I   Y   H   S   G   S   T   N   Y   N   P   S   L   K   S
IGHV4-4*03       TGG TGG AGT TGG GTC CGC CAG CCC CCA GGG AAG GGG CTG GAG TGG ATT GGG GAA ATC TAT CAT AGT GGG AGC ACC AAC TAC AAC CCG TCC CTC AAG AGT
IGHV4-4*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106
                  R   V   T   I   S   V   D   K   S   K   N   Q   F   S   L   K   L   S   S   V   T   A   A   D   T   A   V   Y   Y   C   A   R
IGHV4-4*03       CGA GTC ACC ATA TCA GTA GAC AAG TCC AAG AAC CAG TTC TCC CTG AAG CTG AGC TCT GTG ACC GCC GCG GAC ACG GCC GTG TAT TAC TG- --- --- -
IGHV4-4*i01      ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..T GCG AGA G
```

```
                   1   2   3   4   5   6   7   8   9  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  35  36  37  38
                  E   V   Q   L   V   E   S   G   G   V   V   V   Q   P   G   G   S   L   R   L   S   C   A   A   S   G   F   T   F   D   D   Y   A
IGHV3-43D*01     GAA GTG CAG CTG GTG GAG TCT GGG GGA GTC GTG GTA CAG CCT GGG GGG TCC CTG AGA CTC TCC TGT GCA GCC TCT GGA TTC ACC TTT GAT GAT TAT GCC
IGHV3-43D*i01    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
                  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  62  63  64  65  66  67  68  69  70  71  72  74
                  M   H   W   V   R   Q   A   P   G   K   G   L   E   W   V   S   L   I   S   W   D   G   G   S   T   Y   Y   A   D   S   V   K   G
IGHV3-43D*01     ATG CAC TGG GTC CGT CAA GCT CCG GGG AAG GGT CTG GAG TGG GTC TCT CTT ATT AGT TGG GAT GGT GGT AGC ACC TAC TAT GCA GAC TCT GTG AAG GGT
IGHV3-43D*i01    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ..A ... ... ... ... ... ... ... ...
                  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 102 103 104 105 106 107
                  R   F   T   I   S   R   D   N   S   K   N   S   L   Y   L   Q   M   N   S   L   R   A   E   D   T   A   L   Y   Y   C   A   K   D
IGHV3-43D*01     CGA TTC ACC ATC TCC AGA GAC AAC AGC AAA AAC TCC CTG TAT CTG CAA ATG AAC AGT CTG AGA GCT GAG GAC ACC GCC TTG TAT TAC TGT GCA AAA GAT A
IGHV3-43D*i01    ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... -
```

FIGURE 5 | Affirmed inferred alleles. **(A)** Table of inferred alleles. Shown are the names given to the inferred sequences by IARC; the closest matching IMGT alleles; the genetic differences observed in the inferred allele relative to the IMGT allele; any other name that has previously been associated with this sequence, if previously identified; the genotype frequency of the inferred sequences within the donor's genotype and the allele percentage of the inferred allele based on all of the alleles within the donors genotype for that particular gene. **(B)** Alignment of each inferred sequence relative to the closest matching IMGT allele with the differences between the sequences highlighted in orange. Numbering of the alignments are according to IMGT numbering.

IARC, and the establishment of processes for the evaluation of inferred sequences provides an important new avenue for cataloging germline gene variation at the population level. Ultimately, this should provide insights into how germline gene diversity influences functional immunity (75, 76). Here, we describe the prerequisites, procedures and potential outcome of the IARC-based review and evaluation process, and as proof of principle, we report five novel alleles.

The establishment of the IARC review process should help the research community to chart germline IGHV gene variation across human ethnicities and patient groups. This is an achievable goal if studies increasingly infer the germline gene repertoires of each of their study subjects. Such personalized references databases will also improve AIRR-seq studies, through the improved germline gene annotation and confidence in identification of SHMs that will result (**Figure 1**).

The AIRR Community and the IMGT group have attempted to provide a robust roadmap and conceptual framework for germline gene inference, but the challenge will now be to encourage the incorporation of germline gene inference software into preprocessing and data analytical workflows. This has not yet been widely adopted by the community of researchers who generate and analyze AIRR-seq data. To facilitate this, IARC aims to create detailed step-by-step experimental and bioinformatics tutorials, and will document case studies showing the manifold advantages that lie in this approach. To minimize human intervention and subjectivity, we will also work to further automate the evaluation process of putative germline gene alleles, and to improve the data submission toolchains to INSDC repositories and to IMGT. Finally, in the future, we intend to partner with other researchers, to extend this initiative to the validation of other adaptive immune receptor gene loci. Other IG and TR genes in humans and species of medical importance may be an early focus, but in time we anticipate that the process of inference can be used to extend our knowledge of antigen receptor genes in all vertebrate species.

Putative novel alleles may now be submitted to the IARC-managed web portal for evaluation.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://ogrdb.airr-community.org/.

AUTHOR CONTRIBUTIONS

The authors are all members of the Germline Database Working Group of the AIRR Community of the Antibody Society. All authors contributed to the development of the policies and procedures described. MO and AC drafted the manuscript, and all authors contributed to the editing of the manuscript.

FUNDING

DB was supported in part by the Fondazione Umberto Veronesi. SC was supported in part by an NIAID grant (R01-AI097403). SK was supported in part by an NIH grant (R01-AI104739). FM and DR were supported in part by NIH grants (R01-GM113246, R01-AI120961, R01-AI138709). DR was also supported by grants from the NIH and University of Washington/Fred Hutchinson CFAR (U19-AI117891, P30-AI027757). MO and LT were supported in part by a grant from the Swedish Research Council (2016-01720). CW was supported in part by NIAID grants (R24-AI138963, R21-AI142590). GY was supported in part by a grant from the United States-Israel Binational Science Foundation (2017253). MvZ was supported in part by a grant from the NHMRC (GNT1117687).

REFERENCES

1. 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
2. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol*. (2014) 32:158–68. doi: 10.1038/nbt.2782
3. Wardemann H, Busse CE. Novel approaches to analyze immunoglobulin repertoires. *Trends Immunol*. (2017) 38:471–82. doi: 10.1016/j.it.2017.05.003
4. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front Immunol*. (2018) 9:224. doi: 10.3389/fimmu.2018.00224
5. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol*. (2010) 184:6986–92. doi: 10.4049/jimmunol.1000445
6. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol*. (2012) 188:1333–40. doi: 10.4049/jimmunol.1102097
7. 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
8. 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
9. Matthyssens G, Rabbitts TH. Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. *Proc Natl Acad Sci USA*. (1980) 77:6561-5.
10. Lefranc M-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT® and the birth and rise of immunoinformatics. *Front Immunol*. (2014) 5:22. doi: 10.3389/fimmu.2014.00022
11. 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.
12. 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.
13. 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)31:43.3.CO;2-V
14. Lefranc M-P, Lefranc G. *The Immunoglobulin FactsBook*. London, UK: Academic Press (2001), pp 1–458.
15. Lefranc M-P, Lefranc G. *The T Cell Receptor FactsBook*. London, UK: Academic Press (2001), pp 1–398.
16. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res*. (2015) 43:D413–22. doi: 10.1093/nar/gku1056
17. 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

18. Jackson KJ, Gaeta B, Sewell W, Collins AM. Exonuclease activity and P nucleotide addition in the generation of the expressed immunoglobulin repertoire. *BMC Immunol.* (2004) 5:19. doi: 10.1186/1471-2172-5-19
19. Ohm-Laursen L, Nielsen M, Larsen SR, Barington T. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*. (2006) 119:265–77. doi: 10.1111/j.1365-2567.2006.02431.x
20. 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
21. Weinstein JA, Jiang N, White RA, III, Fisher DS, Quake SR. High-throughput sequencing of the zebrafish antibody repertoire. *Science*. (2009) 324: 807–10. doi: 10.1126/science.1170020
22. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, et al. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci USA*. (2009) 106:20216–21. doi: 10.1073/pnas.0909775106
23. 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
24. 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 USA*. (2015) 112:E862–70. doi: 10.1073/pnas.1417683112
25. Zhang W, Wang IM, Wang C, Lin L, Chai X, Wu J, et al. IMPre: An accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. *Front Immunol.* (2016) 7:457. doi: 10.3389/fimmu.2016.00457
26. Ralph DK, Matsen IV FA. *Per Sample Immunoglobulin Germline Inference From B cell Receptor Deep Sequencing Data*. arXiv. Available online at: https://arxiv.org/abs/1711.05843 (Accessed August 15, 2018).
27. Wendel BS, He C, Crompton PD, Pierce SK, Jiang N. A streamlined approach to antibody novel germline allele prediction and validation. *Front Immunol.* (2017) 8:1072. doi: 10.3389/fimmu.2017.01072
28. Brochet X, Lefranc M-P, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucl Acids Res.* (2008) 36:W503–508. doi: 10.1093/nar/gkn316
29. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front Immunol*. (2017) 8:1418. doi: 10.3389/fimmu.2017.01418
30. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Ann Rev Biochem*. (2007) 76:1–22. doi: 10.1146/annurev.biochem.76.061705.090740
31. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. *Nature*. (1991) 354:389–92. doi: 10.1038/354389a0
32. Zheng NY, Wilson K, Jared M, Wilson PC. Intricate targeting of immunoglobulin somatic hypermutation maximizes the efficiency of affinity maturation. *J Exp Med*. (2005) 201:1467–78. doi: 10.1084/jem.20042483
33. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, Seaman MS, et al. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*. (2010) 329:856–61. doi: 10.1126/science.1187659
34. 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
35. 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
36. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med*. (2015) 7:121. doi: 10.1186/s13073-015-0243-2
37. Greiff V, Miho E, Menzel U, Reddy ST. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trend Immunol*. (2015) 36:738–49. doi: 10.1016/j.it.2015.09.006
38. Friedensohn S, Lindner JM, Cornacchione V, Iazeolla M, Miho E, Zingg A, et al. Synthetic standards combined with error and bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naive and memory B cells. *Front Immunol*. (2018) 9:1401. doi: 10.3389/fimmu.2018.01401
39. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Commun.* (2013) 4:2680. doi: 10.1038/ncomms3680
40. Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv*. (2016) 2:e1501371. doi: 10.1126/sciadv.1501371
41. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. *Nat Methods*. (2014) 11:653–5. doi: 10.1038/nmeth.2960
42. McInerney P, Adams P, Hadi MZ. Error rate comparison during polymerase chain reaction by DNA polymerase. *Mol Biol Int*. (2014) 2014:287430. doi: 10.1155/2014/287430
43. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. *Eur J Immunol*. (2012) 42:3073–83. doi: 10.1002/eji.201242517
44. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*. (2016) 17:125. doi: 10.1186/s12859-016-0976-y
45. Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supported immunoglobulin germline gene inference. *Data Brief*. (2017) 13: 620–40. doi: 10.1016/j.dib.2017.06.031
46. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci USA*. (2013) 110:13463–8. doi: 10.1073/pnas.1312146110
47. Meyerhans A, Vartanian JP, Wain-Hobson S. DNA recombination during PCR. *Nucleic Acids Res*. (1990) 18:1687–91.
48. Judo MS, Wedel AB, Wilson C. Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res*. (1998) 26:1819–25.
49. Zylstra P, Rothenfluh HS, Weiller GF, Blanden RV, Steele EJ. PCR amplification of murine immunoglobulin germline V genes: strategies for minimization of recombination artefacts. *Immunol Cell Biol*. (1998) 76:395–405. doi: 10.1046/j.1440-1711.1998.00772.x
50. Kalle E, Kubista M, Rensing C. Multi-template polymerase chain reaction. *Biomol Det Quant*. (2014) 2:11–29. doi: 10.1016/j.bdq.2014.11.002
51. Gupta NT, vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein, SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*. (2015) 31:3356–8. doi: 10.1093/bioinformatics/btv359
52. Gidoni M, Snir O, Peres A, Polak P, Lindeman I, Clouser C, et al. Mosaic deletion patterns of the human antibody heavy chain gene locus. *Nat Commun*. (2019) 10:628. doi: 10.1038/s41467-019-08489-3
53. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Jr., Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci*. (2015) 370:1676. doi: 10.1098/rstb.2014.0243
54. Ralph DK, Matsen IV FA. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput Biol*. (2016) 12:e1004409. doi: 10.1371/journal.pcbi.1004409
55. 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
56. Thornqvist L, Ohlin M. Data on the nucleotide composition of the first codons encoding the complementary determining region 3 (CDR3) in immunoglobulin heavy chains. *Data Brief*. (2018) 19:337–52. doi: 10.1016/j.dib.2018.04.125
57. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol*. (2003) 171:4639–49. doi: 10.4049/jimmunol.171.9.4639

58. McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc Natl Acad Sci USA*. (1984) 81:3180–4.
59. Chang B, Casali P. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunol Today*. (1994) 15:367–73. doi: 10.1016/0167-5699(94)90175-9
60. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol*. (2004) 172:3382–4. doi: 10.4049/jimmunol.172.6.3382
61. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat Immunol*. (2001) 2:530–6. doi: 10.1038/88732
62. 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
63. 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
64. 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
65. Luo S, Yu JA, Song YS. Estimating copy number and allelic variation at the immunoglobulin heavy chain locus using short reads. *PLoS Comput Biol*. (2016) 12:e1005117. doi: 10.1371/journal.pcbi.1005117
66. 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
67. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type b disease. *J Clin Invest*. (1996) 97:2277–82. doi: 10.1172/JCI118669
68. Kidd MJ, Jackson KJ, Boyd SD, Collins AM. DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *J Immunol*. (2016) 196:1158–64. doi: 10.4049/jimmunol.1501401
69. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol*. (2017) 18:1274–8. doi: 10.1038/ni.3873
70. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*. (2018) 15:201–6. doi: 10.1038/nmeth.4577
71. van Dongen JJ, Langerak AW, Bruggemann M, Evans PA, Hummel M, Lavender FL, 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. doi: 10.1038/sj.leu.2403202
72. Lefranc MP. From IMGT-ontology classification axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). *Cold Spring Harbor Protocols*. (2011) 2011:627–32. doi: 10.1101/pdb.ip84
73. 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
74. Thornqvist L, Ohlin M. Critical steps for computational inference of the 3′-end of novel alleles of immunoglobulin heavy chain variable genes - illustrated by an allele of IGHV3-7. *Mol Immunol*. (2018) 103:1–6. doi: 10.1016/j.molimm.2018.08.018
75. Watson CT, Glanville J, Marasco WA. The Individual and Population Genetics of Antibody Immunity. *Trend Immunol*. (2017) 38:459–70. doi: 10.1016/j.it.2017.04.003
76. 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

Conflict of Interest Statement: 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.