



# Poorly Expressed Alleles of Several Human Immunoglobulin Heavy Chain Variable Genes are Common in the Human Population

Mats Ohlin\*

Department of Immunotechnology, Lund University, Lund, Sweden

## OPEN ACCESS

### Edited by:

Shiv Pillai, Harvard Medical School,  
United States

### Reviewed by:

Peter Daniel Burrows,  
University of Alabama at Birmingham,  
United States

Felix Breden,  
Simon Fraser University, Canada

### \*Correspondence:

Mats Ohlin  
mats.ohlin@immun.lth.se

### Specialty section:

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

**Received:** 08 September 2020

**Accepted:** 08 December 2020

**Published:** 24 February 2021

### Citation:

Ohlin M (2021) Poorly Expressed Alleles of Several Human Immunoglobulin Heavy Chain Variable Genes are Common in the Human Population. *Front. Immunol.* 11:603980. doi: 10.3389/fimmu.2020.603980

Extensive diversity has been identified in the human heavy chain immunoglobulin locus, including allelic variation, gene duplication, and insertion/deletion events. Several genes have been suggested to be deleted in many haplotypes. Such findings have commonly been based on inference of the germline repertoire from data sets covering antibody heavy chain encoding transcripts. The inference process operates under conditions that may limit identification of genes transcribed at low levels. The presence of rare transcripts that would indicate the existence of poorly expressed alleles in haplotypes that otherwise appear to have deleted these genes has been assessed in the present study. Alleles IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 were all identified as being expressed from multiple haplotypes, but only at low levels, haplotypes that by inference often appeared not to express these genes at all. These genes are thus not as commonly deleted as previously thought. An assessment of the 5' untranslated region (up to and including the TATA-box), the signal peptide-encoding part of the gene, and the 3'-heptamer suggests that the alleles have no or minimal sequence difference in these regions in comparison to highly expressed alleles. This suggests that they may be able to participate in immunoglobulin gene rearrangement, transcription and translation. However, all four poorly expressed alleles harbor unusual sequence variants within their coding region that may compromise the functionality of the encoded products, thereby limiting their incorporation into the immunoglobulin repertoire. Transcripts based on IGHV7-4-1\*01 that had undergone somatic hypermutation and class switch had mutated the codon that encoded the unusual residue in framework region 3 (cysteine 92; located far from the antigen binding site). This finding further supports the poor compatibility of this unusual residue in a fully functional protein product. Indications of a linkage disequilibrium were identified as IGHV1-2\*05 and IGHV4-4\*01 co-localized to the same haplotypes. Furthermore, transcripts of two of the poorly expressed alleles (IGHV1-3\*02 and IGHV4-4\*01) mostly do not encode in-frame, functional products, suggesting

that these alleles might be essentially non-functional. It is proposed that the functionality status of immunoglobulin genes should also include assessment of their ability to encode functional protein products.

**Keywords:** adaptive immune receptor repertoire, allelic diversity, antibody heavy chain, germline gene, haplotype, immunoglobulin, inference, next generation sequencing

## INTRODUCTION

The specificity-defining variable domains of human immunoglobulin heavy chains are encoded by genes [IGHV (immunoglobulin heavy chain variable), IGHD (immunoglobulin heavy chain diversity), and IGHJ (immunoglobulin heavy chain joining) genes] located on chromosome 14. These genes rearrange during B cell development in a largely stochastic process to form the complete genes that encode one of the two polypeptide chain types that make up the mature antibody. As a result of the availability of several genes of each type (IGHV, IGHD, and IGHJ) and the stochasticity of this process, a vast diversity of antibodies is produced by human B cells, proteins that are able to protect us from bacteria, viruses and other threats. In addition, several alleles have been associated with most of the IGHV genes and substantial differences between subjects exist with respect to which genes/alleles are available to mount an immune response. Despite this vast diversity and a stochastic process that generate the genes encoding antibodies, it is well established that particular genes and alleles may play a role in the generation of functional adaptive immunity (1–3), highlighting the importance of proper analysis in studies of antibody responses in relation to health and disease.

Immunoglobulin germline genes are described in the IMGT database (4), a source of information that is used by a range of bioinformatics tools to define the genes and the downstream hypermutation processes involved in establishment and evolution of particular antibodies. Such processes require precise definition of the germline IGHV, IGHD, and IGHJ genes that have been used to generate the genes that encode antibodies of interest. This analysis is complicated by the fact that germline gene databases are incomplete and even contain allelic sequences in error (5). Extensive efforts are in place to describe new germline gene alleles and to define immunoglobulin genotypes/haplotypes of multiple subjects. Genomic sequencing, in particular the introduction of long-read sequence technology, will likely offer important insight in this field in the future (6). Germline gene inference based on the information content of next generation sequencing (NGS) data sets has, however, emerged as an important approach to define personalized germline gene allele repertoires available to individuals to generate their antibody responses (7–11). These approaches allow for better gene/allele assignment and tracking of hypermutation pathways that have resulted in antibody sequences of the subject under investigation. The germline gene/allele repertoire of large numbers of individuals has now been described by use of this approach. The VDJbase server (<https://www.vdjbase.org/>) (12) allows public access to such

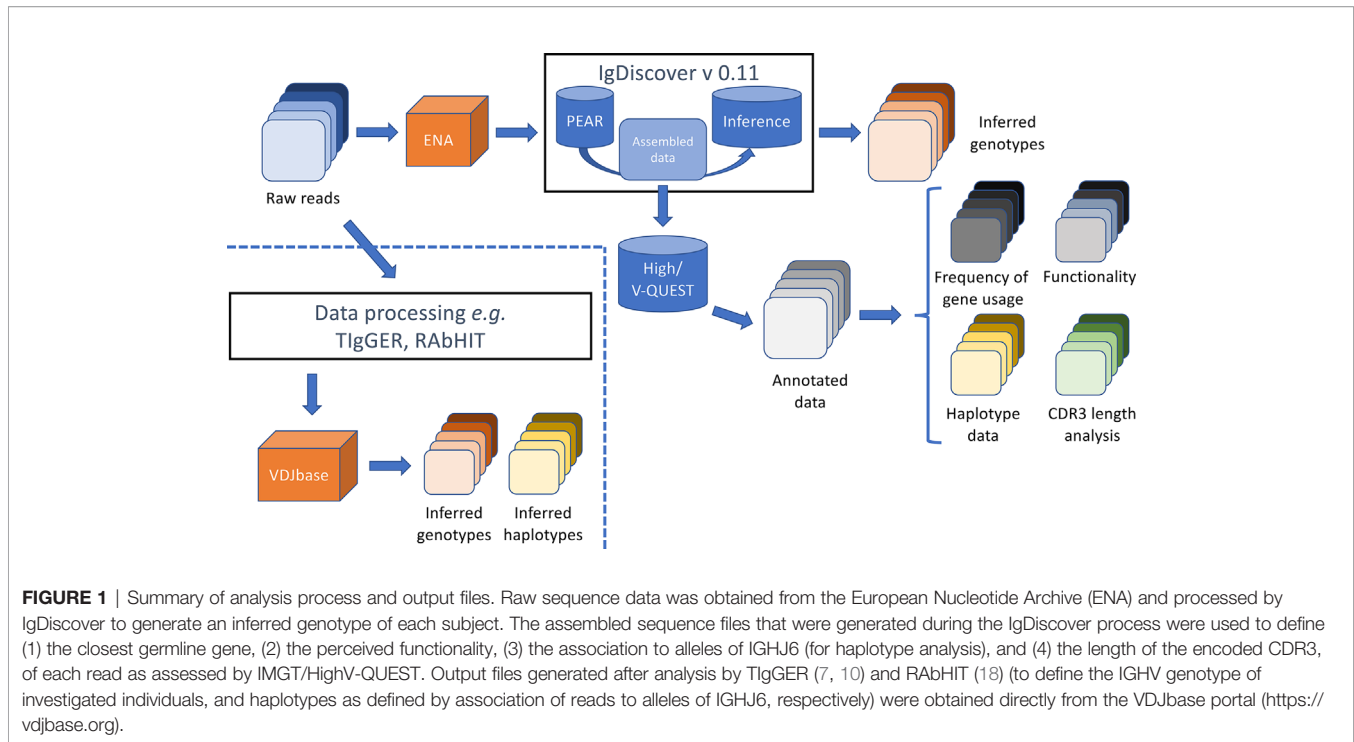
processed information. The fact that numerous subjects carry different alleles of IGHJ allows further refinement and validation of gene assignments through haplotyping (13–16) and a deeper understanding of antibody-encoding genes that can be generated through gene rearrangement and somatic hypermutation. That approach has furthermore enabled identification of gene deletion and insertion events and other complex genetic events only some of which had been previously characterized (17). It is for instance possible to readily identify haplotypes that carry either the IGHV1-8/IGHV3-9 or the IGHV3-64D/IGHV5-10-1 genes, or those that have incorporated well-recognized insertions close to e.g., IGHV3-30, IGHV3-43, IGHV1-69, and IGHV2-70 genes. It has also been possible to identify allele usage bias and mosaic patterns of deleted genes in a large set of subjects (16), as well as functional deletions of large parts of the IGHV locus (15, 16).

Several genes including IGHV1-3, IGHV4-4, and IGHV7-4-1, and to some extent also IGHV1-2, proximal to the IGHD gene locus were shown to be subject to a lack of perceived expression from one or both haplotypes of many subjects under investigation (16). This effect may be caused by an absence of the gene in question, or complete silencing of gene transcription/expression, but also by filtering strategies in inference programs that in the interest of analysis specificity sacrifice the ability of the tool to detect alleles expressed at low levels, or at low levels compared to other alleles of the same gene that are present in the genotype.

In this study, it has been demonstrated that alleles of four genes (IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01) are expressed at such low level that they are not detected by inference tools and incorrectly identified as deleted. It was observed that all these alleles encode polypeptide chains with one or several atypical amino acid residues. It is proposed that two of the alleles (IGHV1-3\*02 and IGHV4-4\*01) are largely non-functional as rearranged genes derived from these alleles typically cannot encode a functional polypeptide. It was possible to demonstrate that rearranged, hypermutated, class-switched sequences derived from one of the alleles (IGHV7-4-1\*01) typically had substituted an atypical residue in its framework region. It is hypothesized that such substitutions may improve the biophysical properties of the encoded product.

## MATERIAL AND METHODS

An analysis approach, summarized in **Figure 1**, was designed to identify and characterize IGHV repertoires in 98 high throughput sequencing data sets by use of a range of different software tools and approaches.



## Immunoglobulin Transcriptome Data Sets

All 100 immunoglobulin variable domain-encoding transcriptome data sets of project PRJEB26509 were downloaded from the European Nucleotide Archive. All data sets of this project, with the exception of ERR2567273 and ERR2567275 that were limited in size and sequence diversity, were used for the present analysis. These data sets represent immunoglobulin IgM and IgD heavy-chain encoding transcriptomes and light chain-encoding transcriptomes of sorted (CD19<sup>+</sup>, CD27<sup>-</sup>, IgD<sup>+</sup>, IgA<sup>-</sup>, and IgG<sup>-</sup>) human naïve B-cells derived from healthy subjects and subjects with celiac disease in Norway (16). The libraries had been generated using 5'-RACE technology and sequenced on the MiSeq sequencing platform. Ninety-four of the data sets, including 33 sets that could be further studied using haplotyping based on IGHJ6 heterozygosity, were used in a past study of deletion patterns in IGHV repertoires (16). Ninety-six of these data sets, including 34 sets that could be further studied using haplotyping based on IGHJ6 heterozygosity, are also described in VDJbase (<https://www.vdjbase.org/>) (12) as study P1.

Data sets of IgG and IgA-encoding transcriptomes of peripheral blood and bone marrow of two subjects (donors 2 and 4) (19), previously shown by inference of the IGHV gene repertoire using the subjects' IgM-encoding transcriptome data to carry the IGHV7-4-1\*01 but not the IGHV7-4-1\*02 allele (14, 15), were used to illustrate evolution of IGHV7-4-1\*01 during somatic hypermutation and selection. These samples (European Nucleotide Archive project PRJEB18926) had been obtained as part of a study of antibody repertoires in allergic subjects. Pre-processing of the data, as well as subsequent IgDiscover-based

germline gene inference, and IMGT/HighV-QUEST analysis have been previously described (15, 19).

## Germline Gene Inference

Heavy chain immunoglobulin germline genes were inferred from NGS data sets of project PRJEB26509 using IgDiscover 0.11 (8). Inference (1 iteration) was performed using settings summarized in Supplementary Methods. A set (Supplementary Methods) of all alleles of human germline heavy chain genes that had at least one allele considered to be functional by the international ImMunoGeneTics information system<sup>®</sup> (IMGT) was downloaded from <http://www.imgt.org> and used as the starting database for the inference process. Bases of IGHV genes are numbered according to the standard IMGT numbering system (20). Population based diversity in base positions associated to particular alleles, and linkage disequilibrium involving SNPs (single nucleotide polymorphisms) were investigated by tools available in the ENSEMBL Genome Browser (<https://www.ensembl.org>) (release 101) (21).

## Haplotyping

An IGHV gene/allele will be able to rearrange to IGHD and IGHJ genes carried on the very same chromosome 14 that carries the IGHV gene/allele in question. If the two haplotypes (each representing one of the two different chromosomes that carry the IGHV locus) carries different IGHD or IGHJ genes or alleles it will be possible to associate a specific IGHV gene to one of the two haplotypes through the existence of such rearrangements. Allelic diversity in IGHJ6 is particularly useful in this respect (15) as this gene is used at high frequency in gene rearrangements.

Sufficient numbers of rearrangements to one of the two alleles of IGHJ6 are thus commonly encountered even if a poorly expressed gene is being assessed. Through such assessment it is possible to demonstrate that a sequence that represents a poorly expressed allele was not generated as a consequence of sequence reads derived through a hypermutation event, or a PCR or sequencing error, but rather represents a true identification of a particular allele of a germline gene, an allele that segregates to a haplotype that does not carry a highly expressed allele of the same gene. Such analysis was carried out based on the output of IMGT/HighV-QUEST analysis (see Section 2.4).

### ImMunoGeneTics Information System®/ HighV-QUEST Analysis

Immunoglobulin germline gene assignment was performed using IMGT/HighV-QUEST (22) on data sets that could be haplotyped based on association to different alleles of IGHJ6 (16). IMGT/HighV-QUEST, in contrast to IgDiscover and TIGGER/VDJbase will carry out such annotation without limiting the annotation to those germline genes/alleles that have been perceived as being part of the subject's genotype. This approach occasionally/frequently (depending on the mutational status of the sequences) assigns multiple possible alleles to a single read (even to alleles that cannot be part of the genotype), but also allows assignments of germline gene origin that could not be inferred by tools like IgDiscover or TIGGER, a factor of importance for the analytical design used in this study. The assay was performed using IMGT/V-QUEST program version 3.5.18 or 3.5.19, and IMGT/V-QUEST reference directory release 202011-3 or 202031-2. Reads, as scored by the IMGT/HighV-QUEST tool, that were unequivocally assigned to a single germline gene allele were used for analysis of allele expression levels, association to one of two alleles of IGHJ6, and perceived functionality. The length of the part of the rearranged genes that encoded complementarity determining region 3 (CDR3) was also determined. This part of the genes is generated during the rearrangement process itself. The process allows for generation of a diversity of lengths of CDR3, but the number of bases encoding CDR3 must be evenly divisible by three to generate a functional sequence. This length thus serves as a measure to understand the extent of occurrence of functional rearrangements. In addition, the length distribution of the CDR3-encoding part of rearranged genes also serves as an indicator of the polyclonality of the rearrangements under investigation.

### Structures of Antibodies With an Origin in IGHV7-4-1

The structures of five antibodies based on in IGHV7-4-1, for which high resolution (<2.5 Å) structures are available, as defined in the IMGT/3Dstructure-DB and IMGT/2Dstructure-DB (23), were investigated. Coordinates for PDB entries 4D9Q, 4EOW, 5CGY, 5ZMJ, and 6B5R were downloaded from Protein Data Bank (<https://www.rcsb.org/>) and visualized using PyMOL 2.0.5 (The PyMOL Molecular Graphics System, Schrödinger, LLC).

### Analysis of Upstream Regions to Poorly Expressed Immunoglobulin Heavy Chain Variable Genes

5'-untranslated regions (5'UTR) and sequences encoding signal peptides (the leader sequence) were assessed in transcripts associated with IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01 and IGHV7-4-1\*01 based on reads unequivocally assigned to these alleles by IMGT/HighV-QUEST. These sequences were compared to relevant genomic annotations as described in the IMGT database and by some additional genomic sequences found in GenBank. These sequences were also compared to those defined in a separate study (24) studying the data sets also investigated in the present study, and to the most prevalent 5'UTR-leader sequence variants reported in a study of unrelated data sets (25).

## RESULTS

### Germline Allele IGHV1-2\*05 Is Expressed at Low Levels Compared to Other Common Alleles of This Gene

In a past study using the TIGGER inference tool, it was suggested that IGHV1-2 was deleted in at least 7/66 haplotypes of 33 subjects (16). VDJbase (12) report common occurrences of IGHV1-2\*02 (n = 66), IGHV1-2\*04 (n = 63), and IGHV1-2\*06 (n = 18), and one case of IGHV1-2\*07, among the 96 subjects of the study as documented in the database, but no cases of IGHV1-2\*01 and IGHV1-2\*05. Furthermore, the gene is reported to be deleted in 6/68 haplotypes of 34 subjects in VDJbase (**Supplementary Figure 1**). When reanalyzing publicly available data sets of this study (PRJEB26509) with the IgDiscover tool the common expression of alleles IGHV1-2\*02, IGHV1-2\*04, and IGHV1-2\*06 in these subjects was confirmed. IGHV1-2\*07, a recently recognized allele that was not present in the database used to initiate the inference process, was not inferred by IgDiscover but in-depth examination of the reads associated with IGHV1-2 in data set ERR2567201 confirmed the presence of this allele in association with one of the haplotypes of this subject (data not shown). Neither IGHV1-2\*01 nor IGHV1-2\*05 were reported by IgDiscover to be present in the final inferred genotypes of any of these subjects (**Table 1**, **Supplementary Figure 2**). IGHV1-2\*05 was however implicated as being present at low levels in five data sets although the inference was not maintained following the final filtering step (**Table 1**). Genomic data suggests that SNPs relevant for the presence of IGHV1-2\*05 in European populations have been identified (**Supplementary Figure 3**). This prompted further detailed assessment of the available data.

Data sets representing the IGHV transcriptome of the 35 subjects, for which haplotyping based on expression of two different alleles of IGHJ6 was possible, were subjected to IMGT/HighV-QUEST analysis. Transcripts derived from rearrangements involving IGHV1-2\*05 were present in one haplotype in six subjects but always at a level ( $0.092 \pm 0.017\%$  per haplotype of all IGHV-encoding reads) substantially

**TABLE 1** | Alleles of IGHV1-2, IGHV1-3, IGHV4-4, and IGHV7-4-1 of 35 haplotypable (based on heterozygosity of IGHJ6) data sets used for the present analysis.

	Data set (run accession number)	Reads assigned by IMGT/HighV-QUEST to IGHV after preprocessing of data set	IGHV1-2		IGHV1-3		IGHV4-4		IGHV7-4-1	
			IgDiscover	High/V-Quest	IgDiscover	High/V-Quest	IgDiscover	High/V-Quest	IgDiscover	High/V-Quest
1	ERR2567187	225406	*04 <sup>†</sup>	*04, *05	*01	*01	*02	*01, *02	*02	*01, *02
2	ERR2567189	224871	*02 <sup>‡</sup>	*02, *04	*01	*01, *02	*02, *07	*02, *07		*01
3	ERR2567192	212646	*02, *04	*02, *04	*01	*01	*02	*02	*01	*01
4	ERR2567199	240716	*02, *04	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01
5	ERR2567200	180844	*02, *04	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01
6	ERR2567201	215909	*02 <sup>§</sup>	*02, *07		*02	*07	*07		
7	ERR2567204	265106	*04 *06	*04 *06	*01	*01	*02	*02	*02	*01, *02
8	ERR2567206	262069	*06	*05, *06	*01	*01	*02	*01, *02	*02	*02
9	ERR2567213	211280	*02	*02		*02	*07	*07		
10	ERR2567214	276041	*04	*04	*01	*01	*02	*02	*01	*01
11	ERR2567215	220448	*02, *04	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01
12	ERR2567217	182276	*02	*02		*02	*07	*07		
13	ERR2567220	151436	*02 <sup>‡</sup>	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01
14	ERR2567221	164821	*02 <sup>‡</sup>	*02, *04	*01	*01	*02, *07	*02, *07		
15	ERR2567223	156379	*02, *06	*02, *06	*01	*01, *02	*02, *07	*02, *07	*02	*02
16	ERR2567226	117706	*02	*02		*02	*02, *07	*02, *07		*01
17	ERR2567230	183580	*02	*02	*02	*02	*02, *07	*02, *07		*01
18	ERR2567231	240347	*02 <sup>†</sup>	*02, *05	*01	*01, *02	*07	*01, *07	*02	*02
19	ERR2567232	192854	*02, *06	*02, *06	*01	*01, *02	*02, *07	*02, *07	*02	*02
20	ERR2567240	149185	*02	*02		*02	*07	*07		
21	ERR2567242	348123	*02 <sup>‡</sup>	*02, *04	*01	*01, *02	*02, *07	*02, *07		*01
22	ERR2567243	253195	*04 <sup>†</sup>	*04, *05	*01	*01	*02	*01, *02	*02	*01, *02
23	ERR2567246	223410	*04	*04	*01	*01	*02	*02	*01	*01
24	ERR2567249	223137	*04 <sup>†</sup>	*04, *05	*01	*01	*02	*01, *02	*02	*01, *02
25	ERR2567254	241355	*04	*04	*01	*01	*02	*02	*01	*01
26	ERR2567259	189273	*02 <sup>†</sup>	*02, *05	*01	*01, *02	*07	*01, *07	*02	*02
27	ERR2567261	308875	*04	*04	*01	*01	*02	*02	*01	*01
28	ERR2567263	312067	*02	*02	*02	*02	*07	*07		
29	ERR2567264	226845	*02, *04	*02, *04	*01	*01, *02	*02	*02	*01	*01
30	ERR2567265	194775	*02, *04	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01
31	ERR2567266	321732	*04	*04	*01	*01	*02, *07	*02, *07		*01
32	ERR2567271	168997	*06 <sup>‡</sup>	*04 *06	*01	*01	*02	*02	*02	*01, *02
33	ERR2567274	184738	*04	*04	*01	*01	*02	*02		*01
34	ERR2567276	148621	*02, *06	*02, *06	*01	*01, *02	*02, *07	*02, *07	*02	*02
35	ERR2567277	193732	*02	*02, *04	*01	*01, *02	*02, *07	*02, *07	*01	*01

<sup>†</sup>IgDiscover tentatively inferred IGHV1-2\*05 in this data set but the allele did not feature in the more strictly filtered V\_expressed output.

<sup>‡</sup>IgDiscover tentatively inferred IGHV1-2\*04 in this data set but the allele did not feature in the more strictly filtered V\_expressed output.

<sup>§</sup>Inference using a starting database that also incorporated IGHV1-2\*07 also inferred IGHV1-2\*07 in this data set.

Inferred alleles as defined by IgDiscover in the V\_expressed file (**Supplementary Figure 2**) and alleles that are primarily scored by IMGT/HighV-QUEST are shown.

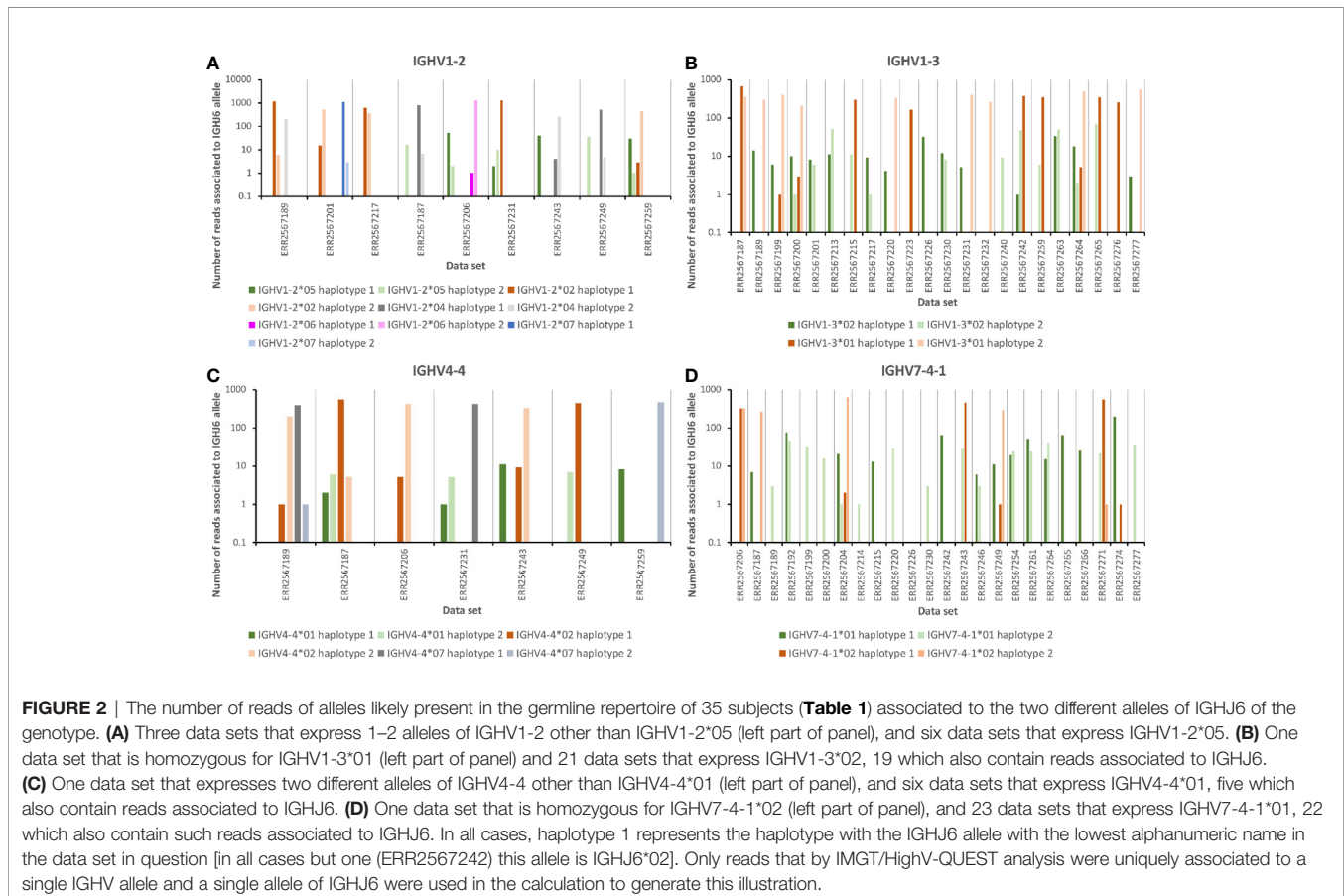
lower than those of IGHV1-2\*02 [ $2.28 \pm 0.52\%$  (present in 29 haplotypes)], IGHV1-2\*04 [ $0.63 \pm 0.22\%$  (present in 28 haplotypes)], IGHV1-2\*06 [ $1.86 \pm 0.8\%$  (present in 6 haplotypes)], or IGHV1-2\*07 [ $2.50\%$  (present in 1 haplotype)]. These frequencies were calculated as the frequency of reads unequivocally assigned to a single germline gene allele in relation to the total number of sequences present in the input data file used for IMGT/HighV-QUEST analysis (**Figure 1**). The relatively low levels of transcripts of IGHV1-2\*04 in some data sets had resulted in its exclusion from the final genotype proposed by IgDiscover (**Table 1**) through the standard filtering feature used in this study. If both haplotypes of IGHV1-2 were occupied by alleles other than IGHV1-2\*05, only very few reads ( $0.0015 \pm 0.0026\%$ ) associated to this allele were observed. Thus, reads associated to IGHV1-2\*05 were rarely incorrectly identified in these data sets as a consequence of antibody evolution, or PCR or sequencing errors of reads originating from rearrangements involving other alleles of IGHV1-2. The frequency of such background reads was, however, about 10-fold higher in subjects that harbored IGHV1-2\*06 (the allele of IGHV1-2 that is most similar to IGHV1-2\*05) in the genotype ( $0.0062 \pm 0.003\%$ ), as compared to those that only had other alleles of IGHV1-2 ( $0.0005 \pm 0.0007\%$ ), illustrating the effect that other alleles may have on the analysis of rarely expressed genes/alleles. Although the

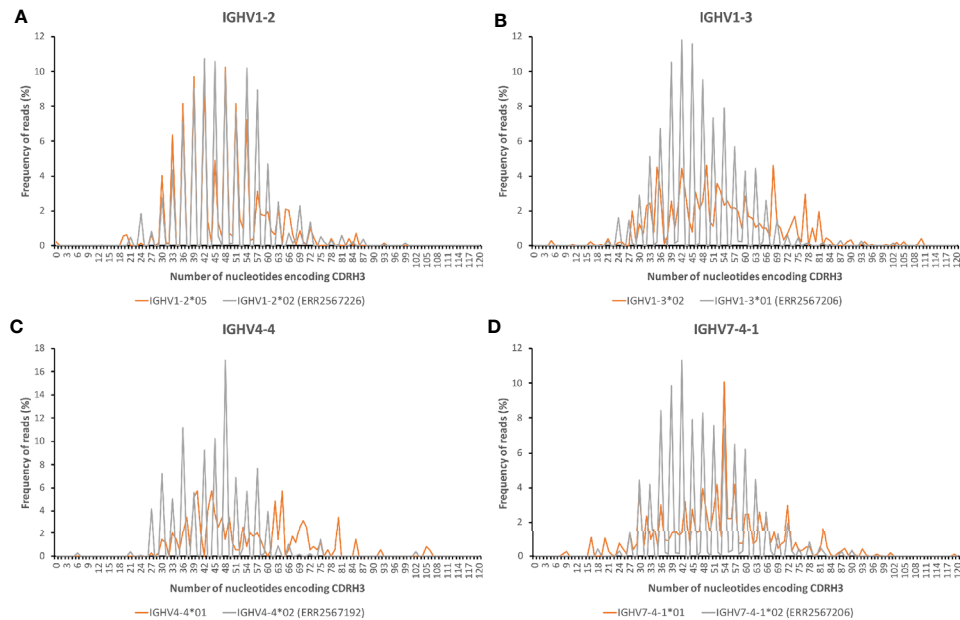
number of reads of IGHV1-2\*05 was low, it was in all cases observed that reads encoding IGHV1-2\*05 were primarily associated to the IGJ6-defined haplotype not associated to reads of the other allele found in the subject under investigation (**Figure 2A**), further supporting the validity of the observation of these reads based on rearrangements that involved IGHV1-2\*05. Altogether, these observations support the presence of IGHV1-2\*05 in 6/70 haplotypes although the allele had not been inferred by TIgGER or IgDiscover. The presence of IGHV1-2\*05 in these genotypes, argues that the gene had not been deleted in any of the investigated haplotypes.

IGHV1-2\*05 is, as observed above, expressed at low levels. 68% of the reads derived from IGHV1-2\*05 were, however, considered to be productive by IMGT/HighV-QUEST. The IGHV-IGHD-IGHJ rearrangements were mostly in-frame (**Figure 3A**). These observations suggest that many reads derived from IGHV1-2\*05 are able to encode a functional product.

### Germline Allele IGHV1-3\*02 is Expressed at Low Levels Compared to Another Common Allele of This Gene

In a past study using the TIgGER inference tool, it was suggested that IGHV1-3 was deleted in 26/66 haplotypes of 33 subjects for which haplotypes had been determined (16). VDJbase indicated





**FIGURE 3** | The number of nucleotides of the CDR3-encoding part of reads derived from IGHV1-2\*05 (A), IGHV1-3\*02 (B), IGHV4-4\*01 (C), and IGHV7-4-1\*01 (D) extracted from all donors that use these genes. The distribution of lengths of bases are compared to those of IGHV1-2\*02 (data set ERR2567226), IGHV1-3\*01 (data set ERR2567206), IGHV4-4\*02 (data set ERR2567192), and IGHV7-4-1\*02 (data set ERR2567206), respectively.

that the gene was present as IGHV1-3\*01 or IGHV1-3\*01 T35A (now officially recognized as IGHV1-3\*05) in 77/96 subjects of this dataset. No instances of IGHV1-3\*02 had been inferred. The ENSEMBL database (21) however suggests that variants of bases 6, 12, 167, 208, 291, and 296, bases indicative of IGHV1-3\*02, are all present at a frequency of about 40% in many populations, including in European populations (Supplementary Figure 4). Reanalysis of publicly available sequencing data sets by IgDiscover indicated that IGHV1-3\*02 was present in the genotype of 11/98 subjects (in all cases without simultaneous detection of IGHV1-3\*01). Further analysis, using IMGT/HighV-QUEST, of the 35 datasets that could be haplotyped using haplotype-defining allelic variation in IGJ6 indicated that 12/35 datasets were homozygous with respect to IGHV1-3\*01. Reads derived from rearrangements involving IGHV1-3\*01 were present at a frequency of  $1.78 \pm 0.15\%$  in these subjects. The frequency of reads assigned to IGHV1-3\*02 in these data sets was, as expected, very low,  $0.0007 \pm 0.0009\%$ . Thus, reads associated to IGHV1-3\*02 had not been artificially generated in these data sets through antibody evolution, or through PCR or sequencing errors of reads based on rearrangements involving IGHV1-3\*01. Subjects ( $n = 16$ ) that had IGHV1-3\*01 associated to only one of its haplotypes expressed such transcripts at a frequency of  $0.90 \pm 0.22\%$ . Reads associated to IGHV1-3\*02 were found in 14 of these subjects at a frequency of  $0.035 \pm 0.017\%$ , i.e., a level of reads 50-fold higher than that observed in subjects that were homozygous for IGHV1-3\*01 and thus could not express IGHV1-3\*02 unless the gene would have been duplicated. Subjects ( $n = 7$ ) that did not express antibody-encoding genes derived from IGHV1-3\*01 (frequency  $0.002 \pm$

$0.002\%$  of all reads), also expressed sequences derived from IGHV1-3\*02 ( $0.065 \pm 0.19\%$ ). In all subjects but two these transcripts were associated to both haplotypes (Figure 2B). Altogether, IGHV1-3\*01 was associated to 40/70 haplotypes while IGHV1-3\*02 was associated to at least 26/70 haplotypes. Only 4/70 haplotypes (defined by reads associated to either allele of IGJ6) could not be identified in any reads, possibly suggesting a deletion of IGHV1-3, a frequency of deletion much lower than that previously suggested.

Reads derived from genes rearranged from IGHV1-3\*01 were mostly (for instance 83 and 84% in ERR2567187 and ERR2567204, respectively) productive, that is it is conceivable that they encode a protein product. In contrast, a much larger fraction of reads derived from transcripts based on IGHV1-3\*02 were non-productive and cannot be perceived to encode a protein due to the presence of stop codons or out-of-frame rearrangements at the IGHV-IGHD-IGHJ junction. Of 1797 reads that also contained a detectable IGJ sequence, only 14% were, based on the nucleotide sequence alone, considered to be productive. Many rearrangements were not in frame as evidenced by the length of the sequence that encoded CDR3 (Figure 3B). Consequently, most observed transcripts derived from IGHV1-3\*02 cannot encode a functional product.

### Germline Allele IGHV4-4\*01 is Expressed at Low Levels Compared to Other Common Alleles of This Gene

In a past study using the TIGER inference tool, it was suggested that IGHV4-4 was deleted in 18/94 genotypes and at least 29/66 haplotypes of 33 subjects (16). However, a recent analysis of the

material, as presented in VDJbase, suggests that IGHV4-4 is not absent in any of the genotypes of this set (IGHV4-4\*02 and IGHV4-4\*07 are found in 78/96 and 63/96 genotypes, respectively). However, it is reported to be absent in 6/68 haplotypes of haplotypable genomes, while IGHV4-4\*02 and IGHV4-4\*07 are found in 36/68 and 26/68 haplotypes, respectively (**Supplementary Figure 1**). IGHV4-4\*01 is not suggested by VDJbase to be expressed in any of the 96 subjects for which data is available in the database. In the present study, IGHV4-4\*01 was not inferred by IgDiscover in any of the 98 data sets. IGHV4-4\*02 and IGHV4-4\*07 were found in 37 and 27 haplotypes of 35 subjects that could be haplotyped based on differential expression of alleles of IGHJ6 (**Table 1; Supplementary Figure 2**). In subjects with alleles of IGHV4-4\*02 and/or IGHV4-4\*07 in both haplotypes the frequency of reads assigned by IMGT/HighV-QUEST to IGHV4-4\*01 was, as expected, very low ( $0.0007 \pm 0.0007\%$ ). When assessing individual reads from the 6 subjects that did not carry IGHV4-4\*02 or IGHV4-4\*07 on one of its haplotypes, it was possible to find evidence of transcripts of IGHV4-4\*01 at a statistically significant higher level ( $0.026 \pm 0.07\%$ ) than if both sites of IGHV4-4 were occupied by any of the other alleles [ $p = 0.0001$  (Mann-Whitney one-sided test)]. In 5/6 subjects a few reads based on rearrangements utilizing IGHV4-4\*01 associated to IGHJ6 were found. These appropriately associated to the allele of IGHJ6 that was not used by the other allele of IGHV4-4 present in the genome (**Figure 2C**). The frequency of transcripts derived from IGHV4-4\*01 when found in the genome was, however, estimated to be 38-fold and 37-fold lower than the frequency of transcripts derived from a single copy of IGHV4-4\*02 and IGHV4-4\*07, respectively. Collectively the data suggests that IGHV4-4\*01 is present in many haplotypes that do not express other common alleles of this gene but that its level of expression is low.

Only 7% of reads assigned to IGHV4-4\*01 were considered by IMGT/HighV-QUEST to represent productive sequences, with a large number of sequences having out-of frame IGHV-IGHD-IGHJ rearrangements (**Figure 3C**). Thus, most observed transcripts derived from IGHV4-4\*01 cannot encode a functional product.

Alleles of IGHV4-4, IGHV4-59, and IGHV4-61 are in many cases similar and their exact location in the genotype may not always be known. One such example is IGHV4-59\*08 that often appears to reside in gene IGHV4-61 (26). The inferred genotype of each subject was assessed by IGHJ6-based haplotyping in the six samples that encoded IGHV4-4\*01. In all cases, alleles of IGHV4-59 and IGHV4-61 were assigned to both haplotypes of these subjects (data not shown) while a highly expressed allele of IGHV4-4 was associated to one of these subjects' haplotypes. It is thus hypothesized that IGHV4-4\*01 was indeed located to gene IGHV4-4, and not to any other gene, on one of each subject's haplotypes.

Sequence features of IGHV4-4\*01 at base 46 (C) and 308 (G) differentiate this allele from most alleles of IGHV4-4, IGHV4-59, and IGHV4-61 (**Supplementary Figure 5**). SNP analysis suggests that these variants are present at a frequency of 3–4%

in European populations and at a higher frequency in many other populations (**Supplementary Figure 5**). Altogether, this data lend support to the likely identification of IGHV4-4\*01 in large datasets like the one investigated here.

### Germline Allele IGHV7-4-1\*01 is Expressed at Low Levels Compared to Another Common Allele of This Gene

In a past study using the TIGGER inference tool, it was suggested that IGHV7-4-1 was deleted in 60/94 genotypes, and at least 49/66 (74%) haplotypes of 33 subjects (16). A recent analysis of the material, as presented in VDJbase, suggests that IGHV7-4-1 is deleted in 56/96 genotypes including in 19/34 genotypes that can be haplotyped based on heterozygosity of IGHJ6 (**Supplementary Figure 1**). Furthermore, VDJbase data shows that IGHV7-4-1\*02 is present in 37/96 genotypes including 13 that can be haplotyped (in two of these cases (I15 and I92) the evidence for the presence of this allele shows very low confidence), while IGHV7-4-1\*01 is present in only 9/96 genotypes, two of which can be haplotyped. In contrast, SNP analysis suggests that the allele-defining sequence variant of IGHV7-4-1\*01 is found at a higher frequency in most populations (**Supplementary Figure 6**). When reanalyzing the publicly available, haplotypable data sets of project PRJEB26509 ( $n = 35$ ) with IgDiscover, IGHV7-4-1\*02 was inferred at high frequency in 11 transcriptomes, but also IGHV7-4-1\*01 at low frequency in 12 transcriptomes (**Table 1; Supplementary Figure 2**). IGHV7-4-1\*01 was only inferred in transcriptomes that did not simultaneously express IGHV7-4-1\*02.

To better understand the occurrence of IGHV7-4-1\*01, data sets were individually analyzed using IMGT HighV-QUEST. One sample (ERR2567206) encoded transcripts derived from IGHV7-4-1\*02 from both haplotypes. The level of transcripts perceived as originating from IGHV7-4-1\*01 in this sample was as expected very low (0.0015% of all reads). Ten samples carried IGHV7-4-1\*02 on one of its haplotypes. Eleven samples that expressed genes derived from IGHV7-4-1\*02 from none or one of its haplotypes expressed similarly low levels ( $0.001 \pm 0.001\%$ ) of IGHV7-4-1\*01-derived transcripts. Higher levels ( $0.056 \pm 0.051\%$ ) of IGHV7-4-1\*01-derived transcripts were seen in 23/35 samples. The frequency of transcripts of IGHV7-4-1\*02 from a single copy of the allele was, however, significantly higher [ $p < 0.0001$  (Mann-Whitney one-sided test)] and determined to represent  $0.96 \pm 0.32\%$  of all reads. The substantial difference in expression of the two alleles of IGHV7-4-1 agrees with previous observations (27). Although the number of reads was low, IGHV7-4-1\*01 was suggested to associate to at least 28/70 haplotypes. Such reads could be assigned to one or both haplotypes in a subject except in one data set as no reads were associated to IGHJ6. Given the low number of reads associated to IGHV7-4-1\*01 it is conceivable that additional haplotypes may carry the allele although no rearrangements to IGHJ6 were found in the transcriptome to confirm such an association. The frequency of IGHV7-4-1/haplotype, taking both IGHV7-4-1\*01 and IGHV7-4-1\*02 into account, was thus at least 57%. If



IGHV7-4-1\*02 was expressed in the same sample as IGHV7-4-1\*01, the two alleles were as expected, by IGHJ6-based haplotyping, not primarily associated to the same haplotype whenever rearrangements of IGHV7-4-1\*01 associated to IGHJ6 were actually found in the data (**Figure 2D**). Among the 34 samples, only 6 (17%) showed no evidence of expression of either IGHV7-4-1\*01 or IGHV7-4-1\*02. Altogether, in-depth analysis of the underlying data identified expression of IGHV7-4-1\*01 in multiple samples, expression that was not always readily detected by inference technology.

Assessment by IMGT/HighV-QUEST of the perceived functionality of rearranged sequences derived from IGHV7-4-1\*01 suggested that 31% of reads that also contained an IGHJ-derived sequence were productive and they featured an in-frame IGHV-IGHD-IGHJ rearrangement (**Figure 3D**). To further assess the somatic evolution of rearrangements derived from IGHV7-4-1\*01, NGS data sets of IgA and IgG repertoires of two subjects known to express IGHV7-4-1\*01 but not IGHV7-4-1\*02 (15, 19) were investigated. These two genes differ in only one base in the part of the gene that encode the final protein product. IGHV7-4-1\*01 encodes an unusual, likely surface-exposed, cysteine residue at position 92 (C92), a residue located far from an antibody's paratope (**Supplementary Figure 7**). The number of reads and clones based on rearrangements involving IGHV7-4-1\*01 was low, but it was observed that all productive rearrangements of these data sets (**Supplementary Figure 8**) had mutated the germline-encoded C92 to a range of other residues (primarily serine, proline, tyrosine, histidine, leucine, phenylalanine, and asparagine). There thus appears to be a strong driving force to replace the unusual cysteine residue of IGHV7-4-1\*01 with another residue during somatic hypermutation and selection, despite the fact that codon 92 does not carry a RGYW/WRCY (28) or a WA/TA mutational hotspot (29).

### IGHV1-2\*05 and IGHV4-4\*01 Are Expressed on the Same Haplotype

Both IGHV1-2\*05 and IGHV4-4\*01 were identified in six genotypes (**Table 1, Figures 1A, C**). Despite their low overall frequency, these alleles were in all cases found in the very same genotypes. An analysis of the haplotypes confirmed that these alleles are present together on the same haplotype in these six genotypes (**Figures 2A, C**). The most likely conserved order of genes/alleles in all the six haplotypes of the present study that carry these poorly expressed alleles was: IGHV6-1\*01 – IGHV1-2\*05 – IGHV1-3\*01 – IGHV4-4\*01 – IGHV7-4-1\*02. The entire associated IGHV locus of these haplotypes were, however, not identical as this set of alleles were associated to alleles like IGHV3-11\*01, IGHV3-11\*05, or IGHV3-11\*06; IGHV3-48\*01, IGHV3-48\*02, or IGHV3-48\*04; IGHV3-49\*03, IGHV3-49\*04, or IGHV3-49\*05; IGHV1-69\*01 and IGHV1-69\*06, IGHV1-69\*02, IGHV1-69\*04, or IGHV1-69\*10. An analysis of SNPs separating IGHV1-2\*05 and IGHV4-4\*01 from other, common alleles of these genes, confirmed the existence of a possible linkage disequilibrium involving these SNPs in some but not all populations (**Supplementary Figure 9**).

### Sequences of Poorly Expressed Immunoglobulin Heavy Chain Variable Alleles Beyond the Coding Region

It is conceivable that poor expression of IGHV alleles would be caused by sequence diversity outside of the coding region itself. The 3'-heptamer is important for efficient rearrangement of the IGHV gene to previously rearranged IGHD-IGHJ genes. The heptamer is however not present in heavy chain-coding transcripts and sequencing libraries derived from them. Consequently, the analysis had to rely on genomic information of the alleles of the investigated gene collected in other studies. In all four instances, the reported heptamer of these poorly expressed alleles is identical to those of other alleles of the same gene (**Supplementary Figure 10**), suggesting that they are not the cause of the poor expression of the genes.

An analysis of the upstream region (part of the 5'UTR and the sequence encoding the leader sequence) of the current data was possible as the NGS-derived sequence information had been derived from libraries generated by 5'-RACE technology (16). This comparison could thus be carried out both using genomic and transcriptome-based, inferred (24, 25) sequence information. The intron sequence found within the signal peptide-encoding sequence, and the sequence upstream of the 5'-end of the mRNA had to be compared to genomic sequence information found in public repositories (**Supplementary Figure 10**). Such analysis demonstrated that these sequences in IGHV1-2\*05 were identical to those found in common, highly expressed alleles like IGHV1-2\*02, IGHV1-2\*04, and IGHV1-2\*06. Similarly, the upstream region of IGHV7-4-1\*01 was identical to that of IGHV7-4-1\*02. A single additional base found in the leader sequence of the primary reference sequence in the IMGT database was not replicated in a number of other germline gene entries suggesting that it may represent a sequencing error or an uncommon allelic variant. Altogether, it thus appears that the upstream region of these poorly expressed alleles is not the cause of their low level of transcription.

IGHV1-3\*02 shows a number of differences in the upstream region as compared to germline gene entries of IGHV1-3\*01 in the IMGT database. These differences include sequence differences in the 5'UTR, just 5' of the transcription initiation codon. However, the same 5'UTR sequence has recently been seen in a number of inferred 5'UTR sequences of highly expressed IGHV1-3\*01. Similarly, the common leader sequence of IGHV1-3\*02 (difference at base -8 from the most common, corresponding sequence of IGHV1-3\*01) has similarly been observed together with highly expressed allele IGHV1-3\*01 (Huang et al. manuscript in preparation). It is thus unlikely that these sequence variants are the cause of the poor expression of IGHV1-3\*02. A couple of base differences exist between the leader sequence intron of IGHV1-3\*01 and IGHV1-3\*02, respectively. Their effect on gene transcription can currently not be assessed.

The observed 5'UTR-leader sequences of the transcriptomes derived from IGHV4-4\*01 of the present investigation did not exactly match neither the corresponding sequence of the allele or of highly expressed alleles IGHV4-4\*02 or IGHV4-4\*07, as

recoded in the IMGT database. However, gene inference studies have identified such alternative 5'UTR-leader sequences in IGHV4-4\*02 in a number of subjects (25; Huang et al. manuscript in preparation). Furthermore, the leader sequence intron of genomic sequences of IGHV4-4\*01 and IGHV4-4\*02 are identical. Altogether, there is no immediate reason to believe that poor expression of IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 is a consequence of the sequences of their upstream regions, or the heptamer associated to them.

## DISCUSSION

Different subjects have access to highly different sets of alleles of germline genes to encode the naïve antibody population they may use to mount humoral immune responses against the hostile environment. Such differences may translate into differences in their ability to raise antibodies against some epitopes (2) as stereotypic responses highly dependent on the availability of particular germline genes are common features of humoral immunity (1). Efficient genomic sequencing of the immunoglobulin heavy and light chain loci has been particularly challenging and only lately been facilitated by long read sequencing technology. Consequently, our understanding of the complexity of immunoglobulin germline gene repertoires is limited. Fortunately, computational germline gene inference using information contained in large NGS data sets has allowed us to describe personal germline repertoires of many subjects, and to identify novel immunoglobulin germline gene alleles that were previously not defined (7, 15, 16, 30). A number of such new alleles have been inferred, reviewed, and documented using a standardized procedure (31). However, the immunoglobulin germline gene rearrangement process, somatic hypermutation, and PCR and sequencing artefacts complicate germline gene inference. Consequently, such tools commonly operate under a set of restrictions, to enhance the specificity of the algorithm at the cost of sensitivity. For instance, inferred alleles that are expressed at much lower levels than other alleles of the same gene in the genotype will commonly be removed during the process.

It has been established by genomic sequencing that gene insertion and deletion events as well as more complex events have generated highly different human immunoglobulin haplotypes. Recently the genotypes of almost 100 subjects were defined by an inference process (16). Importantly, haplotypic analysis could be performed on many of these data sets and multiple deletion events could be defined. As inference cannot efficiently detect alleles that are poorly expressed, an in-depth analysis, using complementary tools, of the sequences was performed in the present investigation, with a focus on genes located in proximity to the IGHD locus, genes that in past studies (16) had been suggested to be deleted in several haplotypes. Such analysis demonstrated the existence of reads providing evidence of alleles expressed at a level much lower than other alleles of the same gene. Such reads were not present in transcriptomes from genotypes that were known to have other alleles populating these

gene sites. The existence of such alleles was in several cases supported by SNPs described through population studies. Such studies are commonly based on short read sequencing, a technology that suffers from substantial caveats in particular in relation to sequencing of immunoglobulin loci (32). Nevertheless, it is reassuring that SNP analysis supports the existence of these alleles in samples obtained in Europe. Altogether, the common presence of the IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 alleles in a substantial fraction of haplotypes is now well established despite the fact that their level of expression is low. The results suggest that deletion at the genomic level is less common than envisaged from past analysis (16). Intriguingly, two of the alleles, IGHV1-2\*05 and IGHV4-4\*01, appear to be tightly co-inherited in this population and potentially also in other populations as assessed by genotype analysis. Although genotype analysis involving genes in immunoglobulin loci is complicated by the by a range of challenges (32), these finding collectively argues in favor of a need for further assessment of linkage disequilibria in the human IGHV locus. Poor expression of multiple genes in particular when it occurs from both haplotypes of an individual may amplify into an inability to mount effective humoral immunity to some target antigens (3) in particular when such responses are limited in breadth.

Although the alleles IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 are present at substantial frequencies in a population, they may, due to their low level of expression, have little impact on the diversity of the antibody repertoire. It is conceivable that their limited expression could be a result of an impairment in their ability to participate in the gene rearrangement processes, to undergo mRNA or protein synthesis, or to encode a functional product. The fact that upstream regions (up to and including the TATA-box), as well as the 3'-heptamer are identical or very highly similar to known, highly expressed alleles of the same gene strongly suggests that such factors are not likely to contribute to the poor expression of the alleles in question. It is now hypothesized that several of these alleles have limited ability to form a functional protein following rearrangement. IGHV1-3\*02 encodes products that carry residues S56, E70, and N99. Virtually all other germline genes of the IGHV1 subgroup encode products with I56 (or another hydrophobic residue), K70, and T99. Similarly, IGHV1-2\*05 encodes an atypical residue V100 in the domain's lower core, IGHV7-4-1\*01 encodes an atypical C92, and IGHV4-4\*01 encodes atypical residues P16 and C103. With access to NGS data sets defining IgG and IgA repertoires of two subjects previously shown to carry IGHV7-4-1\*01 but not the more highly expressed allele IGHV7-4-1\*02 in their genotypes, it was possible to assess how the immune system approached evolution of the allele encoding an unusual cysteine at position 92, a residue located far from the antigen-binding site. Indeed, this residue was invariably substituted in productively rearranged and mutated reads. Thus, there seems to be a strong incentive to remove this unusual residue during the hypermutation process. It is hypothesized that C92 compromises the structure or stability of the protein product and that its substitution improves the

functionality of the encoded product and thereby its ability to get an upper hand during the selection process. This agrees with past findings of evolution of IGHV1-18\*01, IGHV1-8\*01, and IGHV5-51\*01 that all carry unusual residues in their framework regions, residues that are also substituted at a high frequency during somatic hypermutation (33). These findings collectively support the hypothesis that it is the protein sequence of these variants that compromise the generation of clones derived from these alleles. Further assessment of the biophysical characteristics of proteins derived from the alleles is warranted to understand the role of these sequence variants in an antibody generation process.

The expression of other alleles carrying unusual amino acids may be affected to a lesser extent than the four alleles that are the focus of the present study. It was observed (data not shown) that rearranged transcripts of IGHV4-31\*01 in data sets of the present data sets were present at a level approximately 3.6-fold lower than those derived from IGHV4-31\*03. Intriguingly, IGHV4-31\*01 also carries an unusual residue, leucine 75, otherwise commonly an arginine that is located in the charge cluster common to folded antibodies. Interestingly, IGHV1-2\*04, a functional allele that encode an unusual tryptophane in this position, is also expressed at lower levels as compared to IGHV1-2\*02 and IGHV1-2\*06 alleles of this gene (see section 3.1 above), alleles that both encode arginine at this position. The arginine commonly establishes a core, stabilizing polar interaction with aspartate in position 98 (34). These residues in positions 75 and 98, respectively, are hardly ever substituted in hypermutated antibodies (33), illustrating the importance of this interaction. This stabilizing interaction cannot readily be formed by unmutated antibodies derived from the IGHV4-31\*01 germline gene. In any case, this sequence variant is apparently sufficiently tolerated to allow a substantially higher level of expression as compared to IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 in the naive repertoire. Altogether, these findings further support the need for analysis of the possible role of protein stability as a factor that contributes to repertoire development.

Most reads of IGHV1-3\*02 and IGHV4-4\*01 originated from rearranged genes that would not encode a full-length product, for instance as they carried out-of-frame rearrangements or stop codons. It is conceivable that the level of functionality of these alleles should be modified from Functional to ORF (open reading frame). It seems possible that many of the reads represent rearrangements that are merely passengers in cells that use a rearrangement made on the other chromosome to encode an antibody. Nevertheless, their existence highlights the fact that these alleles are present in the haplotypes of many subjects although specific measures have to be taken if they are to be detected by germline gene inference as performed on NGS-based transcriptome analysis. Importantly, their presence may aid our understanding of the antibody hypermutation process as they may serve as controls that have not undergone selection based on retained or improved antigen-binding properties. Furthermore, as we now realize the low level of transcription of some of these alleles, it will be possible to design computational approaches to

correctly assign reads derived from such alleles even in the presence of other, more highly expressed alleles, through an allele-specific filtering strategy.

In summary, IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01 are commonly present in human genotypes but they are poorly expressed and at least two of them do not commonly encode functional products. The identification of these alleles in many data sets through inference requires that computational processes are properly adapted to that task.

## DATA AVAILABILITY STATEMENT

Raw sequence data files of IgM-encoding transcriptomes are available from the European Nucleotide Archive as project PRJEB26509. TIgGER-calculated genotypes and haplotypes are available from VDJbase ([www.vdjbase.org](http://www.vdjbase.org)). Raw sequence data files that include sequences representing IgG and IgA-encoding transcriptomes, used for analysis of mutation patterns of IGHV7-4-1\*01, are available from the European Nucleotide Archive as part of project PRJEB18926.

## AUTHOR CONTRIBUTIONS

MO conceived the study, carried out the analysis, interpreted the data, and wrote the manuscript.

## FUNDING

This work was supported by a grant from The Swedish Research Council (grant number 2019-01042).

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at bioRxiv (<https://www.biorxiv.org/>) (35).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2020.603980/full#supplementary-material>

**Supplementary Figure 1** | Visualization of genotypes and haplotypes of 34 of the samples of the P1 study, also assessed in the present analysis, as defined by TIgGER inference technology and as displayed on the VDJbase website (<http://vdjbase.org>) (12). Illustrations were retrieved in July, 2020. This information was provided under a CC0 (Creative Commons 0) license.

**Supplementary Figure 2** | Visualization of expression levels, as inferred by IgDiscover technology (8), of IGHV genes found in 35 genotypes of subjects for which haplotyping, based on heterozygosity of IGHJ6, is possible.

**Supplementary Figure 3** | Allelic variants of IGHV1-2 as defined by IMGT are illustrated. Variability of some of the positions of these genes in samples obtained in different geographical locations as illustrated by the ENSEMBL browser (release 101, August 2020) (21) is shown. Only bases 163, 223, and 299 [IMGT numbering nomenclature (20)] of this gene display frequencies of variation >1% in the 1000 Genomes Project. The variant (SNP rs12588974) at base 299, indicative of the IGHV1-2\*01 or IGHV1-2\*05 alleles is present at about 5% in European populations. Bases 233 and 234 (SNPs rs782139757 and rs1425538657), that separate these two alleles remains as T and G, respectively, at very high frequency in most populations suggesting that IGHV1-2\*01 is not common in these populations (not shown). All sequence variants of the illustrations of SNPs are indicated as seen in the reversed strand, hence they are complementary to the base of the coding strand.

**Supplementary Figure 4** | Allelic variants of IGHV1-3 as defined by IMGT are illustrated. Variability of some of the positions of these genes in samples obtained in different geographical locations as illustrated by the ENSEMBL browser (release 101, August 2020) (21) is shown. Only bases 6, 12, 167, 208, 291 and 296 [IMGT numbering nomenclature (20)] of this gene display frequencies of variation >1% in the 1000 Genomes Project. Variants indicative of the IGHV1-3\*02 allele are present at about 40%. All sequence variants of the illustrations of SNPs are indicated as seen in the reversed strand, hence they are complementary to the base of the coding strand.

**Supplementary Figure 5** | Allelic variants of IGHV4-4 as defined by IMGT are illustrated. Variability of some of the positions of these genes in samples obtained in different geographical locations as illustrated by the ENSEMBL browser (release 101, August 2020) (21) is shown. Analysis of this gene is complicated by extensive similarity with alleles of IGHV4-59 and IGHV4-61, alleles of which are also shown. A few of the positions of IGHV4-4 that display frequencies of variation >1% in all populations in the 1000 Genomes Project are shown. Note that variants at bases 46 and 308 [IMGT numbering nomenclature (20)], indicative of the IGHV4-4\*01 allele are present at about 3-4% in European populations. All sequence variants of the illustrations of SNPs are indicated as seen in the reversed strand, hence they are complementary to the base of the coding strand.

**Supplementary Figure 6** | Allelic variants of IGHV7-4-1 as defined by IMGT are illustrated. Variability of some of the positions of these genes in samples obtained in different geographical locations as illustrated by the ENSEMBL browser (release 101, August 2020) (21) is shown. Sequence variation in base 274 [IMGT numbering nomenclature (20)] suggests that the base associated to IGHV7-4-1\*01 is more common than the base associated to other alleles of this gene in most populations. All sequence variants of the illustrations of SNPs are indicated as seen in the reversed strand, hence they are complementary to the base of the coding strand.

**Supplementary Figure 7** | High resolution structures of five antibodies with a heavy chain variable domain encoded by a gene derived from IGHV7-4-1. Heavy chain CDR3 is shown at the top of each structure in red. The side chain of residue 92 (in all cases a serine), located far from the antibody binding site is shown in green (carbon) and red (oxygen). Structures include PDB entries 4D9Q (A), 4EOW (B), 5CGY (C), 5ZMJ (D), and 6B5R (E).

**Supplementary Figure 8** | Translated sequences of productive IgA and IgG-encoding reads derived from NGS data sets of two subjects (donors 2 and 4) that both have IGHV7-4-1\*01 but not IGHV7-4-1\*02 in their genotype (15). The sequencing protocol (19) allowed for determination of the sequence from the end of framework 1 and extended into the first constant domain of the heavy chain. The sequences encoded by IGHV7-4-1\*01 and IGHV7-4-1\*02 are shown on top of the figure. Residue 92 is highlighted by an arrow.

**Supplementary Figure 9** | Linkage equilibrium involving SNPs associated to IGHV1-2\*05 and IGHV4-4\*01 is identified in many populations [ENSEMBL browser (release 101, August 2020) (21)]. SNP rs12588974 (base 299 of IGHV1-2) separates IGHV1-2\*05 from other commonly used alleles of the gene (IGHV1-2\*02, IGHV1-2\*04, IGHV1-2\*06) (Supplementary Figure 3) while SNPs rs150123115 (base 308 of IGHV4-4) (A) and rs201063945 (base 46 of IGHV4-4) (B) separate IGHV4-4\*01 from other commonly used alleles of the gene (IGHV4-4\*02 and IGHV4-4\*07) (Supplementary Figure 5).

**Supplementary Figure 10** | Genomic germline gene sequences and inferred sequences representing the 5'- and 3'-end of major alleles and the poorly expressed alleles of IGHV1-2 (A), IGHV1-3 (B), IGHV4-4 (C), and IGHV7-4-1 (D). The TATA box (when identified in the IMGT database) and the ATG start codon are highlighted with a gray and yellow box, respectively. The intron sequence within the leader sequence is indicated by an orange line. Most of the sequence encoding the mature heavy chain variable domain (the end of which is indicated by an arrow) is not shown. Only bases starting from codon 102 and beyond, as well as the 3'-heptamer (indicated by a blue line) are shown. Genomic sequences were mostly retrieved from the IMGT website (<http://www.imgt.org>) but in a few cases directly from GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>). The IMGT/LIGM-DB reference sequences is indicated as primary while other sequences derived from the IMGT website are indicated as secondary. Inferred sequences were obtained from publications by Zhu et al. (25) (of which only the more common variants are shown) and Mikocziova et al. (24). The number after these entries indicate the number of cases in which precisely this sequence had been inferred in these studies. At the bottom, the corresponding 5'UTR- and leader sequences of IGHV1-2\*05, IGHV1-3\*02, IGHV4-4\*01, and IGHV7-4-1\*01, as identified in data sets in the present study, are shown.

## REFERENCES

- Henry Dunand CJ, Wilson PC. Restricted, canonical, stereotyped and convergent immunoglobulin responses. *Philos Trans R Soc Lond B Biol Sci* (2015) 370:20140238. doi: 10.1098/rstb.2014.0238
- 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
- Collins AM, Yaari G, Shepherd AJ, Lees W, Watson CT. Germline immunoglobulin genes: disease susceptibility genes hidden in plain sight? *Curr Opin Syst Biol* (2020) 24:100–8. doi: 10.1016/j.coisb.2020.10.011
- Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res* (2005) 33:D256–261. doi: 10.1093/nar/gki010
- 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
- Ford M, Haghshenas E, Watson CT, Sahinalp SC. Genotyping and Copy Number Analysis of Immunoglobulin Heavy Chain Variable Genes Using Long Reads. *iScience* (2020) 23:100883. doi: 10.1016/j.isci.2020.100883
- 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 U S A* (2015) 112: E862–870. doi: 10.1073/pnas.1417683112
- 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
- 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) 7:457:457. doi: 10.3389/fimmu.2016.00457
- 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) 10:129:129. doi: 10.3389/fimmu.2019.00129
- Ralph DK, Matsen F. Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data. *PLoS Comput Biol* (2019) 15:e1007133. doi: 10.1371/journal.pcbi.1007133
- Omer A, Shemesh O, Peres A, Polak P, Shepherd AJ, Watson CT, et al. VDJbase: an adaptive immune receptor genotype and haplotype database. *Nucleic Acids Res* (2020) 48:D1051–6. doi: 10.1093/nar/gkz872

13. 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
14. Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supported immunoglobulin germline gene inference. *Data Brief* (2017a) 13:620–40. doi: 10.1016/j.dib.2017.06.031
15. 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* (2017b) 87:12–22. doi: 10.1016/j.molimm.2017.03.012
16. 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
17. 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
18. Peres A, Gidoni M, Polak P, Yaari G. RAbHIT: R Antibody Haplotype Inference Tool. *Bioinformatics* (2019) 35:4840–2. doi: 10.1093/bioinformatics/btz481
19. Levin M, Levander F, Palmason R, Greiff L, Ohlin M. Antibody-encoding repertoires of bone marrow and peripheral blood—a focus on IgE. *J Allergy Clin Immunol* (2017) 139:1026–30. doi: 10.1016/j.jaci.2016.06.040
20. Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) 2011:633–42. doi: 10.1101/pdb.ip85
21. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. *Nucleic Acids Res* (2020) 48:D682–8. doi: 10.1093/nar/gkz966
22. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi: 10.1007/978-1-61779-842-9\_32
23. Ehrenmann F, Kaas Q, Lefranc MP. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhSF. *Nucleic Acids Res* (2010) 38:D301–307. doi: 10.1093/nar/gkp946
24. 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
25. Zhu Y, Yang X, Wu J, Tang H, Wang Q, Guan J, et al. Antibody upstream sequence diversity and its biological implications revealed by repertoire sequencing. *BioRxiv* (2020) 2020.09.02.280396. doi: 10.1101/2020.09.02.280396
26. 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
27. 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
28. Dörner T, Foster SJ, Farner NL, Lipsky PE. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur J Immunol* (1998) 28:3384–96. doi: 10.1002/(SICI)1521-4141(199810)28:10<3384::AID-IMMU3384<3.0.CO;2-T
29. Zhao Y, Gregory MT, Biertumpfel C, Hua YJ, Hanaoka F, Yang W. Mechanism of somatic hypermutation at the WA motif by human DNA polymerase  $\epsilon$ . *Proc Natl Acad Sci U S A* (2013) 110:8146–51. doi: 10.1073/pnas.1303126110
30. Vazquez Bernat N, Corcoran M, Hardt U, Kaduk M, Phad GE, Martin M, et al. High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Front Immunol* (2019) 10:660:660. doi: 10.3389/fimmu.2019.00660
31. 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) 10:435:435. doi: 10.3389/fimmu.2019.00435
32. Watson CT, Matsen F, Jackson KJL, Bashir A, Smith ML, Glanville J, et al. Comment on “A database of human immune receptor alleles recovered from population sequencing data”. *J Immunol* (2017) 198:3371–3. doi: 10.4049/jimmunol.1700306
33. 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* (2017c) 8:1433:1433. doi: 10.3389/fimmu.2017.01433
34. Kugler M, Stein C, Schwenkert M, Saul D, Vockentanz L, Huber T, et al. Stabilization and humanization of a single-chain Fv antibody fragment specific for human lymphocyte antigen CD19 by designed point mutations and CDR-grafting onto a human framework. *Protein Eng Des Sel* (2009) 22:135–47. doi: 10.1093/protein/gzn079
35. Ohlin M. Poorly expressed alleles of several human immunoglobulin heavy chain variable (IGHV) genes are common in the human population. *bioRxiv* (2020) 2020.09.05.284257. doi: 10.1101/2020.09.05.284257

**Conflict of Interest:** MO is a member of the Adaptive Immune Receptor Repertoire (AIRR) Community’s Germline Database Working Group, and its Inferred Allele Review Committee. The Committee defines processes for approval of alleles of immunoglobulin gene alleles identified through computational inference, and that also approves inferences of such alleles.

Copyright © 2021 Ohlin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.