Molecular Immunology 87 (2017) 12–22

Contents lists available at ScienceDirect

Molecular Immunology

journal homepage: www.elsevier.com/locate/molimm

Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery

Ufuk Kirik[a], Lennart Greiff[b,c], Fredrik Levander[a], Mats Ohlin[a,*]

[a] Dept. of Immunotechnology, Lund University, Lund, Sweden
[b] Dept. of Clinical Sciences, Division of Otorhinolaryngology, Head and Neck Cancer, Lund University, Sweden
[c] Dept. of Otorhinolaryngology, Skåne University Hospital, Lund, Sweden

ARTICLE INFO

Article history:
Received 16 December 2016
Received in revised form 7 March 2017
Accepted 8 March 2017

Keywords:
Antibody
Antibody repertoire
Bioinformatics
Germline gene
Germline gene inference
Haplotype
Heavy chain variable domain

ABSTRACT

Analysis of antibody repertoire development and specific antibody responses important for e.g. autoimmune conditions, allergy, and protection against disease is supported by high throughput sequencing and associated bioinformatics pipelines that describe the diversity of the encoded antibody variable domains. Proper assignment of sequences to germline genes are important for many such processes, for instance in the analysis of somatic hypermutation. Germline gene inference from antibody-encoding transcriptomes, by using tools such as TIgGER or IgDiscover, has a potential to enhance the quality of such analyses. These tools may also be used to identify germline genes not previously known. In this study, we exploited such software for germline gene inference and define aspects of analysis settings and pre-existing knowledge of germline genes that affect the outcome of gene inference. Furthermore, we demonstrate the capacity of IGHJ and IGHD haplotype inference, whenever subjects are heterozygous with respect to such genes, to lend support to IGHV gene inference in general, and to the identification of novel alleles presently not recognized by germline gene reference directories. We propose that such haplotype analysis shall, whenever possible, be used in future best practice to support the outcome of germline gene inference. IGHJ-directed haplotype inference was also used to identify haplotypes not expressing some IGHV germline genes. In particular, we identified a haplotype that did not express several major germline genes such as IGHV1-8, IGHV3-9, IGHV3-15, IGHV1-18, IGHV3-21, and IGHV3-23. We envisage that haplotype analysis will provide an efficient approach to identify subjects for further studies of the link between the available immunoglobulin repertoire and outcomes of immune responses.

© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Antibodies are critical components in higher organisms', including mans', defence against bacteria, viruses, and other threats. Consequently, they have been studied intensively since their discovery more than a century ago. Recent advances in cell culture technology, microdroplet technology, structural biology, and next generation sequencing (Georgiou et al., 2014) have substantially enhanced our understanding of immunoglobulins and their recognition of antigen and development following encounter with antigen. The establishment of large germline gene databases and accompanying analysis tools (Alamyar et al., 2012; Giudicelli et al., 2005; Retter et al., 2005; Gaeta et al., 2007) now allows detailed analysis of antibody sequence information. However, germline gene databases are not complete, and analysis of many species' antibody germline gene repertoire is lacking in depth. For instance, sequencing of immunoglobulin genes still identifies new alleles in several cases (Scheepers et al., 2015; Watson and Breden, 2012). The very recent identification of extensive differences in germline gene repertoires of Balb/c and C57BL/6 mouse strains further highlights the lack of completeness of germline databases (Collins et al., 2015). Furthermore, databases contain erroneous entries (Wang et al., 2008) that, depending on the research question, may affect downstream analysis. Although sequencing of the immunoglobulin gene loci of every subject of a study to allow proper germline gene assignment of identified transcripts is a possibility, it represents a substantial obstacle. Therefore, such an approach may not be a realistic option in many investigations. Inference of individual subjects' germline gene repertoires from their immunoglobulin-encoding transcripts using bioinformatics approaches is a more realistic

Abbreviations: BM, bone marrow; CDR, complementarity determining region; H, heavy; PB, peripheral blood; V, variable.

* Corresponding author at: Dept. of Immunotechnology, Lund University, Medicon Village building 406, S-223 81 Lund, Sweden.
E-mail address: mats.ohlin@immun.lth.se (M. Ohlin).

http://dx.doi.org/10.1016/j.molimm.2017.03.012
0161-5890/© 2017 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

alternative. Consequently, software tools and pipelines have been developed for this type of inference and they are utilized to facilitate new discoveries in the field of antibody repertoires (Gadala-Maria et al., 2015; Elhanati et al., 2015; Boyd et al., 2010; Kidd et al., 2012; Corcoran et al., 2016; Zhang et al., 2016). The present investigation was designed to assess how analysis is affected by a variety of factors related to e.g. composition of germline gene databases, prior knowledge, and somatic hypermutation/PCR and sequencing errors, in an effort to support future development of best practice in the field of germline gene inference. We specifically exploited inferred haplotypes of IGHJ and IGHD genes to lend support to inferred IGHV genes. Finally, we used the outcome of germline gene and haplotype inference to define deletions in individual haplotypes. In particular we defined one deletion spanning many of the major IGHV genes, in an individual that proved to be heterozygous with respect to alleles of IGHJ6.

2. Materials and methods

2.1. Antibody heavy chain variable domain transcriptomes

Six allergic donors had been recruited for studies of immunoglobulin-encoding repertoires, a study that had been approved by the regional ethical review board at Lund University (Levin et al., 2017). Duplicate samples of both bone marrow (BM) and peripheral blood (PB) cells were obtained out of season of exposure to seasonal allergens (Levin et al., 2017). Transcripts encoding the heavy (H) chain variable (V) domains of different isotypes found in cells of these samples were individually amplified by PCR using primers based on the Biomed2 primer set annealing to the sequence encoding V domain framework region (FR) 1 (van Dongen et al., 2003) and those annealing to the sequence encoding the first constant (C) domain. After addition of barcodes and sequencing adaptors the PCR products were sequenced using the 2×300 bp MiSeq technology (Illumina, Inc., San Diego, CA, USA) at the National Genomics Infrastructure (SciLifeLab, Stockholm, Sweden). FASTQ sequence files (study accession number: PRJEB18926) are available from the European Nucleotide Archive.

2.2. Initial data processing

Paired end reads of IgM-encoding transcripts of one BM sample of each donor was processed using the pRESTO pipeline (Vander Heiden et al., 2014) and sequences were binned based on the isotype-specific 3′ primer as reported in Supplementary Methods and Supplementary Table EIV in Reference Levin et al. (2017). This set was used as such for gene inference using IgDiscover (Fig. 1). Furthermore, sequences not carrying an amplified gene sequence identical to a part of the sequence encoding the first constant domain of the isotype were removed. Sequences were also subsequently analysed (Levin et al., 2017) using IMGT HighV-QUEST tool (Alamyar et al., 2012) and further adapted for use in TIgGER using the Change-O pipeline (Gupta et al., 2015) (Fig. 1). IgM-encoding sequences of duplicate PB samples were pooled and used together for some studies using IgDiscover. The numbers of sequences of each sample at different steps in the analysis pipeline are summarized in Table 2 in Reference Kirik et al. (2017).

2.3. Germline gene inference using TIgGER

TIgGER (Gadala-Maria et al., 2015) uses the output of IMGT/HighV-QUEST (Alamyar et al., 2012) analysis of a set of transcripts, analyses the observed mutational pattern to compute a likely germline gene database, a database that is used to re-analyse the germline gene assignments, the output of which is used to define a likely germline gene set. TIgGER version 0.2.7 (downloaded via CRAN (https://cran.r-project.org)) was used for analysing sequences and inferring genotype information. For each donor, a two-step procedure was carried out, in compliance with the documentation of the software. IGHV gene database was retrieved from IMGT (Lefranc et al., 2015) homepage on 2016-08-16. As a first step novel alleles were searched. At this stage, the range of nucleotides to be considered by the algorithm was set to 79–312, numbering according to IMGT definitions, based on the location of the primers. All other settings were left to defaults, except the number of processors to be used for calculation. In the second step, an IGHV germline genotype inference was carried out for each donor, both with and without filtration of mutated sequences. Furthermore, the gene_cutoff value was set to 1e-3, meaning that a gene must be observed at least 1/1000 of the total allele calls to be included in the genotype.

2.4. Germline gene inference using IgDiscover

IgDiscover (Corcoran et al., 2016) infers germline genes by an iterative process that involves initial assignment of sequences using IgBlast (Ye et al., 2013) to a pre-existing germline gene database, a cluster identification process, and subsequent filtering steps to generate a new germline gene database. IgDiscover version 0.5 was downloaded via the Bioconda channel (https://bioconda.github.io) together with all the dependencies. For consistency in analysis between the TIgGER and IgDiscover, the same pre-processing pipeline was used whenever possible. Thus the input files to IgDiscover consisted of merged paired reads filtered and sorted based on isotype primer by the pRESTO tool. The same list of genes from IMGT was used as reference database, but nucleotides corresponding to amino acids 1–25 were initially removed. Following initial studies (Sections 3.1–3.2 below), nucleotides corresponding to residues 106–107 were also removed from database entries since IgDiscover does not allow for defining a range in V-gene analysis. The gene lists were processed by a Perl script to shorten the headers and remove the gaps, in compliance with IgBLAST requirements.

For testing the influence of the reference database on the inference outcomes, two smaller versions of the reference database containing a representative sequence for different clans were generated, referred to as V123 (containing IGHV1-18*01, IGHV2-5*01, and IGHV3-23*01) and V345 (containing IGHV3-15*01, IGHV4-39*07, and IGHV5-51*01), respectively.

The input data was analysed with the differences parameter (henceforth referred to as diff) set to 0, 1, and 2 (default) for both the pre-germline and germline filters. This parameter controls the number of differences allowed between a sequence and a reference gene for the sequence to be assigned to that genes cluster. In other words, setting this parameter to 0, 1, and 2 allows for separation of alleles that differ by a single, two, and three nucleotides, respectively. Additionally, the minimal number of unique CDR3s parameter of the germline filter was set to 50 in order to weed out potential PCR and sequencing artefacts.

In order to assess sequence read quality related to specific inferred genes/alleles we identified sequence read identities from the output from IMGT/HighV-QUEST analysis. The relevant sequences were to deviate from the gene/allele by not more than one base (script default: >99.4% sequence identity) and to carry a gene/allele-defining sequence motif (e.g. GCTTGAGTGGATGGGA[CT]GGATCAACCCTAACAG of IGHV1-2; bases within square brackets represent the alternatives of the allele-differentiating nucleotide, in this case nucleotide 163 of IGHV1-2). The corresponding quality-defining entities were retrieved from the FASTQ file obtained after pRESTO processing. Average read qualities of bases of the motif were calculated and plotted. The quality control analysis of the sequences was carried

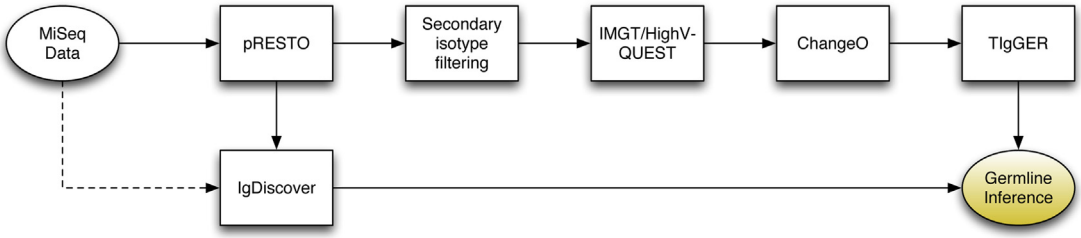

Fig. 1. Schematic presentation of the bioinformatics workflow employed in the present investigation.

out by a custom script, which is available on Github, see Section 2.6.

2.5. IGHJ and IGHD gene usage and IGHV germline gene inference based on haplotype

In donors heterozygous for alleles of a given IGHJ gene it is possible to infer each IGHV haplotype through linkage to the different alleles of IGHJ (Kidd et al., 2012). We analyzed the frequency of each allele of IGHJ as defined in the IMGT database. This script calculates and plots numbers of each IGHV gene, as found in a tab-delimited file generated by IgDiscover, linked to alleles of an IGHJ gene in those cases when more than two alleles of the IGHJ gene are present at a frequency above a given threshold (default: 25%). Sequences that match the V_errors = 0 and D_covered >35 criteria were used for analysis of IGHV gene linkage to IGHD genes/alleles apparently differentially expressed by different haplotypes. Haplotype analysis was carried out by custom scripts, available on Github, see Section 2.6.

2.6. Code availability

All custom scripts are written in Python 3 and they are available open source in the GIgGle repository (release 0.2) under Apache License at https://github.com/ukirik/giggle.

3. Results and discussion

To investigate the effects of approaches to germline gene inference, IgM-encoding transcripts in BM derived from 6 individuals were used as a starting point for such analysis. Our study employed germline gene inference tools TIgGER (Gadala-Maria et al., 2015) and IgDiscover (Corcoran et al., 2016). The sequences used for this investigation were originally obtained as part of a past study (Levin et al., 2017). They had been amplified using the standard Biomed2 primer set, primers that anneal to IGHV FR1. In this analysis process, a number of aspects that have implications for future use of software intended for germline gene inference applications were observed, as outlined in Sections 3.1–3.7.

3.1. Considerations of germline gene inference linked to codon 106

The immunoglobulin germline gene rearrangement process does not use a precise site for rearrangement. Consequently, bases may be removed at the 3′ terminus of IGHV genes during rearrangement. Diversity may thus be introduced in these codons by processes that are not encoded in the germline but instead introduced by the rearrangement process itself. It is conceivable that, depending on the inference methodology, such diversity may be mistaken for germline diversity. A germline gene inference software may thus incorrectly infer germline genes alleles based on such diversity if this region is also assessed during the inference process. Indeed, we repeatedly identified putative novel alleles with diversity in codon 106 using IgDiscover with a diff = 0 setting, a setting necessary for separate identification of very similar germline alleles. For instance, IGHV1-18 and IGHV3-21 were implicated as to have such alternative inferred alleles differing only in codon 106, variants that occurred even at frequencies above 10% of the commonly recognised allele present in the database (Fig. 1 in Reference (Kirik et al., 2017)). To minimize this problem, researchers should consequently consider not to analyse this codon or beyond in a germline gene inference process. If analysis of codon 106 and beyond is implemented, further in-depth analysis of the result, and potentially conformational studies like genome sequencing or haplotype-based analysis, as outlined in Section 3.5–3.6, should be considered.

3.2. Considerations of germline gene inference linked to the variable 3′-end

Germline gene databases (like the IMGT IGHV database) may contain germline genes with different endings, i.e. they carry different number of bases that hypothetically encode residues in CDRH3. Depending on germline gene inference software design, this may have implications for germline gene calling. For instance, sequences in the IMGT database representing IGHV4-38-2*01 and *02 include 0 (GenBank: Z12367), and 0 (GenBank: X56365) or 2 (GenBank: AC233755) bases, respectively, of codon 107. Such differences may affect germline gene inference analysis. Under situations where the amplified product (as in the case of applying Biomed2 primers for amplification of IGHV4-38-2) only differ by the number of bases encoding CDRH3, inference of alleles may define differences based solely on these bases, which are likely removed in many rearrangements due to the imprecise rearrangement process. Indeed, IgDiscover separately inferred these two alleles based on these differences. TIgGER similarly inferred both alleles in some donors (Fig. 2 in Reference Kirik et al. (2017)). We suggest that investigators ought to consider eliminating length differences of members of the pre-existing germline gene database so as to minimize artefacts in the inference process.

3.3. Considerations of germline gene inference linked to prior knowledge

It is conceivable that inference of a germline gene repertoire may depend on the content of the database used as starting point for the inference. For instance, detection of IGHV1-2*p06 in the presence of transcripts derived from IGHV1-2*02 (two genes that differ in a single base) may be problematic (Gadala-Maria et al., 2015). In our hands, TIgGER identified this variant but did not always (depending on runtime parameters) list it in its final graphic output (Fig. 2 in Reference Kirik et al. (2017)). Similarly, IGHV3-7*01 and IGHV3-7*02, two genes that also differ by a single base, were not always separately inferred, depending on software settings. In contrast, IGHV1-2*p06 was readily inferred and shortlisted by TIgGER in the

Table 1
Alleles in IgM repertoires of 6 subjects inferred† depending on alleles of IGHV1-2 present in the initiating germline gene database and the setting of the collapsing parameter diff.

	Alleles of IGHV1-2 in database: IGHV1-2*01, *02, *03, *04, *05			Allele of IGHV1-2 in database: IGHV1-2*p06		
Donor	diff =0	diff =1	diff =2	diff =0	diff =1	diff =2
1	*02, *p06	*02	*02	*02, *p06	*p06	*p06
2	*02, *04	*02, *04	*02, *04	*02, *04	*02	*02
3	*02	*02	*02	*02	*02	*02
4	*02	*02	*02	*02	*02	*02
5	*04, *p06	*04, *p06	*04	*04, *p06	*04, *p06	*p06
6	*02, *p06	*02	*02	*02, *p06	*p06	*p06

†Values that correspond independently of database composition are highlighted on a grey background.

presence of transcripts derived from rearrangements originating from IGHV1-2*04 (Fig. 2 in Reference Kirik et al. (2017)), two alleles that differ by two bases rather than one. When germline genes were inferred using IgDiscover, it was observed that analysis settings substantially affected the output depending on the germline repertoire present in a given individual (Table 1). Using the IMGT database as a starting point for gene inference, IGHV1-2*p06 (an allele not present in the IMGT database) was only detected in subjects also using IGHV1-2*02 if argument diff = 0. Furthermore, it was only detected in the presence of IGHV1-2*04 if diff $\leq$ 1. When IGHV1-2*p06 was not inferred (i.e. if diff $\neq$ 0 or diff >1, respectively), the programme only reported the presence of the allele already present in the database used to carry out the analysis. The reciprocal was also true, i.e. when we created a germline database not containing the alleles of IGHV1-2 (*01–*05) recognized by IMGT but instead only featured IGHV1-2*p06, this latter allele was readily recognized while alleles *02 and *04 were only recognized at low settings of the diff parameter (Table 1). These differences may substantially affect many downstream analysis efforts, for instance if mutational events are recorded for analysis of antigen-driven selection (Yaari et al., 2012), as exemplified by IGHV1-2*02, *04, and *p06, alleles that encode products that differ in 1–2 residues.

Germline gene inference tools may be used to identify new germline genes in organisms for which limited genomic sequence data is available. Such analysis has to be accomplished using a limited set of available germline genes or even germline genes derived from other organisms as a starting point for the computational process (Corcoran et al., 2016). Given the fact that an input V gene database may affect the inferred gene repertoire, as illustrated above, we investigated how analysis of human V gene repertoires by IgDiscover was affected by the composition of such a starting database. Databases were constructed that each contained three sequences each encoding H chain V domains belonging to clan I (encoded by genes of the IGHV1, 5, and 7 subgroups), clan II (IGHV2, 4, and 6), or clan III (IGHV3), respectively. In general, these small databases with members derived from IGHV1, IGHV2, and IGHV3 (V123) and IGHV3, IGHV4, and IGHV5 (V345), respectively, were able to infer most genes of their own subgroups correctly (Table 2). However, the germline V123 database poorly inferred genes belonging to IGHV4 and IGHV5 subgroups while V345 poorly inferred genes and alleles belonging to the IGHV1, IGHV6, and IGHV7 subgroups (Table 2). There were also differences in the ability of the two small sets of genes to be used for inference of some alleles.

In summary, the content of the V gene database, information that defines prior knowledge, that is used to initiate the inference process will affect the analysis output. Depending on the research question, this factor may affect the outcomes of downstream analysis, a fact that must be taken into account during experimental design and data interpretation.

3.4. Considerations of inference specificity vs. selectivity

IgM-encoding transcriptomes in humans, and other species that do not employ hypermutation as a way to evolve the naïve antibody repertoires, are likely to harbour large numbers of unmutated rearranged genes derived from antigen-unexperienced B cells. Nevertheless, unless antigen-stimulated B cells are removed prior to generation of libraries for sequencing, the sequence population will contain those that have undergone hypermutation. Mutated bases residing in mutation hotspots (Neuberger, 2008), or insertions and deletions created during hypermutation (Wilson et al., 1998; Ohlin and Borrebaeck, 1998), may thereby form a nucleus for inference of novel alleles. Bases systematically targeted by PCR/sequencing artefacts may similarly affect the inference process. Immunoglobulin germline gene inference tools may offer opportunities to reduce improper allele inference caused by such somatic hypermutation events in hotspot regions, but such approaches may affect allele detection sensitivity, as outlined in Section 3.3. When assessing our data sets using IgDiscover, it was evident that variants of IGHV4-39*01 were inferred at frequencies of about 5% of those of the generally recognized sequence variant in individuals (donor 1–4, 6) that also had the regular IGHV4-39*01 gene. In particular the A143C mutation (resulting in a Lys/Thr difference) that resides in a mutation hotspot (WA) was identified in these five donors (Table 3 in Reference Kirik et al. (2017)). Identified sequences were diverse in origin. Should this particular sequence variant be accepted as a rarely expressed allele (or rarely expressed separate gene highly similar to the IGHV4-39 gene), or should it be disregarded as a mutational or PCR/sequencing artefact? To address this matter we also analysed IgG-transcriptomes derived from BM of these donors, which in general, are more mutated that their IgM-encoding counterpart (Levin et al., 2017). These sequences did not carry higher frequencies of the A143C transversion in rearranged sequences derived from IGHV4-39 ($5.3 \pm 0.9\%$ vs. $4.4 \pm 0.8\%$; $p = 0.087$ using the one-tailed Mann-Whitney test). Instead A- > G transitions were more common in base 143 in the mutated IgG-encoding transcriptome ($4.7 \pm 1.1\%$ vs. $1.9 \pm 1.0\%$; $p = 0.0066$) suggesting that the mutational machinery rather favours this type of mutation. It thus appears that a high mutation rate introducing C in this position was not a likely contributor to observed finding. Furthermore, the quality of those base calls (C143) of position 143 that specifically result in the definition of this inferred sequence variant of IGHV4-39 was substantially lower than the quality of reads of other bases in the

Table 2
Germline genes inferred by IgDiscover from IgM-encoding transcripts of BM samples of six donors, as a function of the inference-initiating germline gene database either as derived from IMGT germline gene collection, or comprising three germline genes that are derived from germline gene subgroups 1, 2, and 3 (V123) or 3, 4, and 5 (V345).

	Donor 1			Donor 2			Donor 3			Donor 4			Donor 5			Donor 6		
Germline database	IMGT	V123	V345	IMGT	V123	V345	IMGT	V123	V345	IMGT	V123	V345	IMGT	V123	V345	IMGT	V123	V345
IGHV1-18*01	*	*		*	*		*	*		*	*		*	*		*	*	
IGHV1-2*02	*	*		*	*		*	*		*	*					*	*	
IGHV1-2*p06	*	*											*	*		*	*	
IGHV1-2*04				*	*								*	*				
IGHV1-24*01	*			*			*			*			*			*		
IGHV1-3*01	*	*		*	*								*	*		*	*	
IGHV1-46*01 †§	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV1-58*01				*	*								*	*		*	*	
IGHV1-58*02																*		
IGHV1-69*01 †	*	*		*	*		*	*		*	*		*	*		*	*	
IGHV1-69*02							*	*		*	*					*	*	
IGHV1-69*02 †											*					*		
IGHV1-69*04 †	*	*								*	*							
IGHV1-69*06 †	*	*		*	*								*	*				
IGHV1-8*01	*	*		*	*		*	*		*	*		*	*		*	*	
IGHV2-26*01	*	*	*	*	*	*	*	*	*	*			*	*	*	*	*	*
IGHV2-5*01				*	*	*				*	*	*				*	*	*
IGHV2-5*02	**	*	*	*	*	*	*	*	*	*	*	*	**	*	*	**	*	*
IGHV2-70*01	*	*	*	*	*	*	*						*	*	*	*	*	*
IGHV2-70*11	*	*	*				*	*	*							*	*	*
IGHV2-70D*04				*	*	*							*	*	*			
IGHV3-11*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-11*03 †	*	*	*				*	*	*									
IGHV3-11*06													*	*	*			
IGHV3-13*01	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-13*04	*	*	*				*	*	*									
IGHV3-13*05													*	*	*			
IGHV3-15*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-15*07													*	*	*			
IGHV3-20*01 †	**	***	***	*	*	*	***	***	***	*	*	*	*	*	*	*	*	*
IGHV3-21*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-23*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-30-3*01	*	*	*				*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-30*03 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-30*04 †				*	*	*												
IGHV3-33*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-43*01	*	*	*				*	*	*							*	*	*
IGHV3-43D*01													*	*	*	***	***	***
IGHV3-48*01	*	*	*	*	*	*	*	*	*	*	*	*						
IGHV3-48*02	*	*	*				*	*	*	*	*	*				*	*	*
IGHV3-48*03				*	*	*							*	*	*	*	*	*
IGHV3-48*04													*	*	*			
IGHV3-49*03 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-49*04				*	*	*							*	*	*	*	*	*
IGHV3-53*01 †	*	*	*	**	**	**	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-64*01	*	*	*				*	*	*	*	*	*				*	*	*
IGHV3-64D*06	*	*	*				*	*	*				*	*	*			
IGHV3-66*01 †	*	*	*				*	*	*							*	*	*
IGHV3-66*02										*	*	*						
IGHV3-66*03													*	*	*			
IGHV3-7*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-7*02	*	*	*				*			*	*	*						
IGHV3-7*03													*	*	*			
IGHV3-72*01				*	*	*				*	*	*				*	*	*
IGHV3-73*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-74*01 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV3-9*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV4-30-2*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV4-30-4*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV4-31*02 †	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV4-34*01 †	**		*	*		*	*		*	**		*	*		*	**		*
IGHV4-38-2*01 †				*		*							**		**	**		*
IGHV4-39*01 †	**		*	**		*	**		*	**		*				**		*
IGHV4-39*07							*		*				**		*			
IGHV4-4*02 †	*	*	*	*	*	*	*	*	*	*	*	*	*		*	*		*
IGHV4-4*07				*		*	*		*	*		*				*		*
IGHV4-59*01 †	*		*	*		*	*		*	*		*	*		*	*		*
IGHV4-59*08	*		*				*		*							*		*
IGHV4-61*01	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
IGHV4-61*02										*	*	*						
IGHV5-10-1*01 †	*		*				*		*				*		*			
IGHV5-51*01 †	*		*	*		*	*		*	*		*	*		*	*		*
IGHV6-1*01 †	*	*		*	*		*	*		*	*		**			**	*	
IGHV7-4-1*01				*	*					*	*		*					
IGHV7-4-1*02	*	*											*	*		*	*	

†One of several possible alleles.
§Alleles *01 and *03 differing only in base 315 (codon 105) are not treated separately here.
*Wildtype sequence called by analys.
**Wildtype and mutant sequence called by analysis.
***Mutant sequence called by analysis.

vicinity of nucleotide 143, or of those reads that represented A143 itself (Fig. 3A–F in Reference Kirik et al. (2017)). Similar concerns are associated with the inference of the IGHV6-1*01 A85C sequence variant (Fig. 3G, H in Reference Kirik et al. (2017)). This is in contrast to sequencing reads that represented allele-differentiating reads of new inferred alleles like IGHV1-2*02 T163C (IGHV1-2*p06) and IGHV3-53*01 G88A, that did not show evidence of poorer base calls in the relevant part of the sequence (Fig. 3I–L in Reference Kirik et al. (2017)). It is thus conceivable that sequencing problems might contribute to inference of some (like IGHV4-39*01 A143C), but certainly not all (IGHV1-2*p06), new alleles. There will likely be a continuum of results representing these kinds of outcomes and defining a criterion that objectively separates these situations will be challenging. Identifying such variants in multiple individuals will not be sufficient if the methodology as such has a tendency to systematically target some bases with inferable sequence diversity. Detailed confirmation of analysis outputs, for instance using haplotype inference, as further outlined in Sections 3.5 and 3.6, or genome sequencing, is thus required to eliminate improperly inferred alleles, observed as a consequence of the need to maintain a high degree of inference sensitivity required for proper inference of some alleles such as those of IGHV1-2.

3.5. Inferred alleles and the use of IGHJ for haplotype analysis

The use of germline gene inference tools identifies a number of candidate genes for further evaluation. The present study identified a set of fragments of germline genes (considering the fact that the 5′-Biomed2 primers used for amplification resided in FR1) not corresponding to genes annotated by the IMGT reference directory. Some of these, e.g. IGHV1-2*02 T163C (also referred to as IGHV1-2*p06 by IgPdb (http://cgi.cse.unsw.edu.au/~ihmmune/IgPdb/index.php)), IGHV3-20*01 C307T (IGHV3-20*p02), IGHV3-43D*01 C195A (IGHV3-43*p04), and IGHV3-53*01 G88A (new putative allele IGHV3-53*p07), were present at frequencies similar to or greater than those of the inferred canonical IMGT allele, while others sequences were identified at numbers lower (e.g. IGHV4-38-2 A83C, IGHV6-1*01 A85C, and IGHV6-1*01 A104C) or substantially lower than the corresponding canonical IMGT-defined sequence. The latter sets, in particular, require considerable scrutiny.

In many cases TIgGER and IgDiscover similarly identified a particular gene as being homozygous and heterozygous with respect to allele use (Table 3; Fig. 2 in Reference Kirik et al. (2017)). Analysis settings, however, affected which allele is included in the final inferred repertoire (e.g. the inference of IGHV5-51 in Fig. 2 in Reference Kirik et al. (2017)). In other cases, different software may differently define a gene as being represented by one or two alleles (Table 3). Such cases also need specific attention as the selection of gene inference tool may affect the outcomes of subsequent, downstream analysis.

By taking advantage of heterozygous IGHJ genes in the genotype of some individuals, it is possible to link each IGHJ gene to a set of IGHV genes and thereby to infer the haplotypes (Kidd et al., 2012), information that we consider may support gene and allele inference analysis. Two of the donors of the present study (subjects number 4 and 5) were demonstrated to use two different IGHJ6 alleles (IGHJ6*02 and *03) (Fig. 4 in Reference Kirik et al. (2017)). Analysis of inferred IGHV alleles following sequencing of BM transcripts linked to each of these IGHJ alleles, as defined by IgDiscover, was performed (Fig. 2). This procedure identified a number of genomic events in the IGHV locus, such as an introduction of IGHV5-10-1 and IGHV3-64D in replacement of IGHV1-8 and IGHV3-9 (as reported in the past (Watson and Breden, 2012; Watson et al., 2013)). Deletions were also observed such as of IGHV4-30-2, IGHV3-30-3, IGHV4-30-4, and IGHV4-39 (also observed in the past (Kidd et al., 2012; Chimge et al., 2005; Watson et al. 2013)) in one haplotype of donor 5, of IGHV4-34 in the other haplotype of donor 5, of IGHV7-4-1 (also described in the past (Kidd et al., 2012; Chimge et al., 2005; Watson et al., 2013)) in one haplotype of donor 4, of IGHV3-64 in one haplotype of donor 4 (in agreement with past studies (Kidd et al., 2012)), and in both haplotypes of donor 5, and of IGHV3-66 in one haplotype of donor 4. A particularly extensive deletion was found in one haplotype of donor 4, a deletion that spanned the entire region in-between (but not including) IGHV3-7 and IGHV3-30-5 or IGHV3-30 (two genes that cannot be differentiated by the present sequencing strategy). Similar analysis of transcripts of peripheral blood confirmed these findings (Fig. 5 in Reference Kirik et al. (2017)). Individuals homozygous for this latter haplotype would lack ability to produce antibodies derived from a range of major IGHV germline genes, including IGHV3-23 and IGHV3-15. These particular genes are for instance used by antibodies of stereotyped (Henry Dunand and Wilson, 2015) antibody responses against Haemophilus influenzae polysaccharide (Adderson et al., 1993; Lucas et al., 2003). It will be of substantial interest to identify the extent by which such deletions in homozygous individuals affect their ability to develop protective immunity against this pathogen, in similarity to the inability of individuals that do not express the *01 allele of IGKV2D-29, that is also required for this stereotyped immune response (Nadel et al., 1998). If such deletions are common in some populations, particular strategies for vaccination (such as employment of conjugated vaccines (Santosham et al., 1992)) may be required to allow for efficient induction of humoral immune responses against this antigen.

The inference of two sequences supposed to represent allelic variants of a single gene onto a single haplotype is indicative either of their independent presence in different genes (for instance through gene duplication) or of an incorrect inference of one of the sequence variants. In contrast, their association to different haplotypes suggests that they are not artefacts (e.g. not caused by PCR/sequencing errors or hypermutation) of the analysis but rather truly represent two different genomic sequences. Analysis of IGHV gene usage demonstrated that many alleles were expressed only in combination with a single IGHJ6 allele (Fig. 2). This was true for instance in the case of IGHV1-2*p06, an inferred allele not represented in the IMGT database, and IGHV1-2*04 in subject 5. Such findings made possible by IGHJ gene heterozygozity of some subjects substantially support the inference of an allele. The simultaneous expression of more than one inferred allele/haplotype of IGHV1-69 likely includes the expression of an allele of gene IGHV1-69D (Watson et al., 2013) that cannot be separately inferred by this sequencing approach from many alleles of IGHV1-69. For instance, a possible interpretation of the analysis (Fig. 2) suggests the presence of IGHV1-69*04 and IGHV1-69D*01 on one chromosome and IGHV1-69*02 on the other chromosome of donor 4. In contrast, analysis of the IGHJ6 linkage of putative, inferred allelic variants IGHV2-5*02 A100C, IGHV4-34*01 A103C, IGHV4-38-2*01 A83C, IGHV4-39*01/07 A91C, IGHV4-39*01/07 A143C, IGHV6-1*01 A104C, and IGHV6-1*01 A85C in BM samples all illustrated that the presence of such transcripts were not only much lower than those of the canonical sequence defined by the IMGT database, but also that both they and their canonical counterpart were inferred to be expressed in combination with the same IGHJ6 allele(s). The same was true for inferred variants that carried differences in codon 106, as described in Section 3.1 (Fig. 1C, D in Reference Kirik et al. (2017)). The actual presence of such inferred sequence variants in the genome would rather imply that gene duplication events had occurred. These findings cast substantial doubts on the relevance of their inference and suggest that they are artefacts introduced by somatic hypermutation, PCR/sequencing errors, and/or (in the case of codon 106) diverse sites of IGHV-IGHD gene rearrangement events. Certainly, additional studies involving e.g. genomic sequencing would be required to confirm

Table 3
Examples of different calls with respect to the inferred presence of one or more allele of germline genes in BM-derived transcriptomes as defined by TIgGER and IgDiscover.

	TIgGER – filtered†		TIgGER – unfiltered§		IgDiscover@	
	Homozygous	Heterozygous	Homozygous	Heterozygous	Homozygous	Heterozygous
IGHV1-2	3, 4	1, 2, 5, 6*	1, 3, 4, 6	2, 5	3, 4	1, 2, 5, 6*
IGHV1-46	1–3, 5, 6	4	1–6		1–3, 5, 6	4
IGHV2-70		1, 2, 5, 6	1, 2, 5	3, 6	2, 5	1, 3, 6
IGHV3-7	1, 2, 3, 4, 6	5	2, 6	1, 3, 4, 5	2, 6	1, 3, 4, 5
IGHV4-61	1–3, 5, 6	4	2, 3	1, 4–6	1–3, 5, 6	4
IGHV3-20	2, 3, 5, 6	1**	1–6		2–6***	1**
IGHV3-30	1, 3–6	2	3–6	1, 2	1, 3–6	2
IGHV3-53	1, 3–6	2****	1–6		1, 3–6	2****
IGHV4-30-4	1–6		1–3, 5, 6	4	1–6	
IGHV4-38-2	2, 6	5	2	5, 6	2	5, 6

† Find_unmutated = true.
§ Find_unmutated = false.
@ Diff = 0.
* Donors 1, 5, and 6 are inferred as heterozygous including variant IGHV1-2*p06.
** Donor 1 is inferred as heterozygous including variant IGHV3-20 C307T (IGHV1-20*p02).
*** Donor 3 is inferred as only encoding variant IGHV3-20 C307T (IGHV3-20*p02).
**** Donor 2 is inferred as heterozygous including variant IGHV3-53*01 G88A (IGHV3-53*p07).

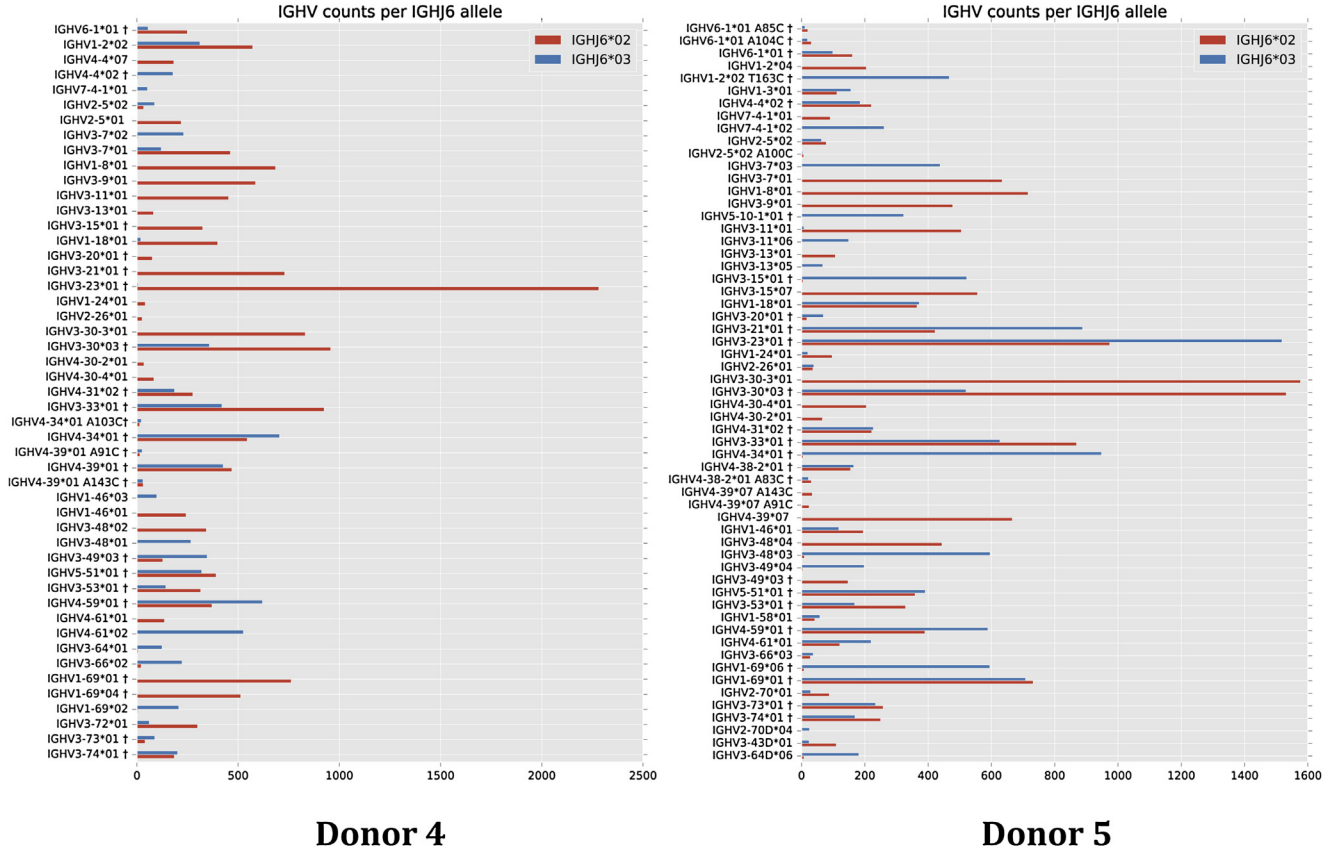

Fig. 2. Association of inferred IGHV germline genes with different alleles of IGHJ6 as indicators of the two haplotypes, as inferred by IgDiscover using the diff = 0 argument from the transcriptomes of donors 4 and 5. The approach identifies a large potential deletion in the haplotype associated with IGHJ6*03 in donor 4 and segregates most alleles (including those of IGHV1-2) of one gene into two haplotypes. Several inferred but rare sequence variants, however, do not segregate in this way. † defines that the name of only one of a set of different alleles of the gene that cannot be differentiated by the analysis approach is shown. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

their existence. It is noted that these sequences all represent A –> C modifications, a common substitution error in sequencing using MiSeq technology (Schirmer et al., 2015; Fuellgrabe et al., 2015), and (as outlined in Section 3.4) that some of these inferred alleles were associated to sequence reads with lower quality scores (Fig. 3 in Reference Kirik et al. (2017)). Several additional examples of such sequence variants were inferred using datasets derived from PB lymphocytes (Fig. 5 in Reference Kirik et al. (2017)). These findings strongly suggest that inferred variants carrying A –> C modifications and identified at frequencies substantially lower than the canonical allele must undergo particular scrutiny.

In summary, the present findings suggest that the presence of different alleles of IGHJ genes in the two haplotypes of a given individual may be used to support gene inference and to reject some inferred alleles or to call for further validation of their existence.

3.6. Inferred alleles and the use of IGHD for haplotype analysis

We hypothesized that particular IGHV genes might also be associated to particular chromosomes if these differed in the IGHD genes they carried or utilized. Differences in IGHD gene presence between chromosomes have been observed in the past (Kidd et al., 2016). Investigation of IGHD gene usage in combination with IGHV genes that were inferred as two different alleles in a single donor was therefore performed. In all donors but subject 2 we observed differences in the functional IGHD haplotypes, as summarized in Table 4. Haplotypes exploited 9–17 different IGHD genes, each of which code for >1% of H chain V domains. As donor 1 was homozygous with respect to deletion of a set of IGHD genes he only used 12 IGHD genes in total as a source of diversity in CDR3 of the H chain. Some haplotypes had putative deletions of genes in the span IGHD3-3–IGHD2-8, IGHD3-22–IGHD2-26 (also reported by Kidd et al. (2016) and Boyd et al. (2010)), or IGHD3-10 (Fig. 6 in Reference Kirik et al. (2017)). As reported in the past (Kidd et al., 2016), those haplotypes that did not have functional genes from IGHD3-3–IGHD2-8 overexpressed IGHD3-10 ($24.1 \pm 1.1\%$ vs. $8.1 \pm 1.1\%$ of unique sequences) (Fig. 6 in Reference Kirik et al. (2017)). Based on these findings we investigated the possibility to exploit the linkage of germline genes/alleles present in the samples to these IGHD genes and to IGHD genes (like IGHD6-13 and IGHD6-19) linked to both alleles of these genes (Fig. 6 in Reference Kirik et al. (2017)). Such analysis demonstrated that IGHV1-8/IGHV3-9 and IGHV3-64D/IGHV5-10-1, pairs of genes that have been reported as mutually exclusive in their presence on a single chromosome (Watson et al. 2013; Kidd et al., 2012), indeed showed different linkage to IGHD genes in subjects 1, 3, and 5 (Fig. 7 in Reference Kirik et al. (2017)), suggesting their presence in different haplotypes of these individuals. It also confirmed the assignment of IGHV1-69D*01, IGHV1-69*02, and IGHV1-69*04 identified by linkage to different IGHJ6 alleles in donor 4, as outlined above, and suggested linkage of IGHV1-69D*01 and IGHV1-69*06 on one chromosome, and IGHV1-69*04 on the other chromosome of donor 1 (Fig. 8 in Reference Kirik et al. (2017)). Furthermore, we demonstrated this way that inferred alleles, not present in the IMGT reference directory, of IGHV1-2 (donors 1, 5, and 6) and IGHV3-20 (donors 1 and 3), were indeed associated with IGHD genes of only one of the two haplotypes of these donors (Fig. 3), thereby supporting the appropriateness of the inference. We also observed potential absence of e.g. IGHV3-66 on one chromosome of donor 1, IGHV3-30-3, IGHV3-33, IGHV4-31, and IGHV4-61 on one chromosome of donor 3, IGHV3-66 on one chromosome of donor 4, IGHV4-34 on one chromosome of donor 5, and of IGHV4-39 on one chromosome of donor 6 (Fig. 8 in Reference Kirik et al. (2017)). Moreover, the analysis suggested that the alleles of IGHV4-59 (*01 and *08) in donors 3 and 6 were present together on one of these subjects' haplotypes (Fig. 9 in Reference Kirik et al. (2017)), possibly suggesting that the gene had undergone a duplication event. Such duplication of this gene has been reported to in the past (Kidd et al., 2012), in support of the present inference. Finally, the finding that two inferred alleles (IGHV4-39*01 A143C and IGHV6-1*01 A104C) present at low frequency as compared to the already recognized alleles (IGHV4-39*01 and IGHV6-1*01) showed the same linkage to IGHD gene sets as their more common counterparts (Fig. 8 in Reference Kirik et al. (2017)), suggests that they are incorrectly inferred, or that their existence requires that gene duplication has occurred. The use of IGHD-supported gene inference in these cases thus does not support these particular gene inferences and strongly suggests that additional studies have to be carried out before these inferred alleles could be considered to be included into a germline gene database.

Some IGHD genes may exist as different alleles, and the existence of such alleles is recognized by the IMGT germline gene collection. IGHV sequences derived from donors heterozygous for such alleles may be segregated onto different haplotypes based on their association to such alleles. Such analysis is complicated by the fact that only a fragment of the IGHD gene is usually incorporated into the H chain V domain-encoding gene. Depending on the location of bases differentiating alleles of IGHD genes these may be more or less likely to be incorporated and thus available for analysis. An analysis of IGHD allele involvement in the creation of repertoires identified genes that were interpreted as being present as different alleles. The analysis often inferred extensive usage of two alleles of IGHD3-16, alleles that differ from each other in some of its 3′-most bases, bases that are likely to be removed during the DJ rearrangement process. This may result in incorrect inference of allele *01 in subjects homozygous for allele *02. Indeed, linkage of alleles of IGHV to alleles of IGHD3-16 did not show any association between pairs of alleles (data not shown). In the case of IGHD2-8 and IGHD2-21, all donors were suggested to incorporate two alleles into their H chain V domain-encoding transcriptome. However, in most donors, the ratio between allele usages was large. However, in two donors (donors 1 and 5), this ration was substantially smaller, suggesting that one (donor 1) or both (donor 5) of them may be heterozygous with respect to these genes' alleles (Fig. 10 in Reference Kirik et al. (2017)). As the frequency of incorporation of these IGHD genes is low (Boyd et al., 2010) the number of reads representing each IGHV gene associated to these IGHD alleles is low. Nevertheless, in most cases alleles of IGHV genes were each associated to different alleles of IGHD genes (Fig. 11 in Reference Kirik et al. (2017)). This was true irrespective of whether transcriptomes encoding IgM found in BM or PB were analysed. This approach lend further support to the observation that IGHV1-2*p06 and IGHV1-2*02 were found to be present in different haplotypes of donor 1 and that IGHV1-2*p06 and IGHV1-2*04 were found in different haplotypes of donor 5. It also demonstrated that IGHV4-59*01 appeared to be present in both haplotypes of donor 1 while only one of them also harbours IGHV4-59*08, supporting an interpretation that also donor 1 harbours a duplication of this gene in one of its haplotypes. Importantly, the appropriateness of the IGHD allele calls was supported by detailed sequence analysis of the incorporated parts of the IGHD alleles. Such analysis confirmed that sequences that in most cases incorporated the allele-differentiating base were appropriately associated to one or the other of member of pairs of IGHV alleles. Some IGHD allele calls that did not support these conclusions were in fact shown not to incorporate such a defining base, a fact that resulted in an irrelevant allele assignment (Fig. 11E in Reference Kirik et al. (2017)). The use of IGHD allele-based haplotype inference also confirmed that there was no evidence that rare inferred allelic variants like IGHV4-39*07 A91C and A143C of donor 5 (Fig. 11A–D in Reference Kirik et al. (2017)) were correct as their presence were strongly linked to the same haplotype as their much more common, well-recognized allele. The acceptance of their existence would require further analysis like genome sequencing. Altogether, the findings suggest that the presence of different alleles of IGHD genes can be exploited to validate germline gene inference and to support or reject inferred novel alleles of IGHV genes. It will in particular be useful in subjects that carry allelic diversity in more commonly used IGHD genes or in data sets comprising even larger numbers of unique sequences than those used in this study.

In summary, IGHD genes and alleles of genes linked to individual haplotypes may be used to add confidence to IGHV gene inference and to rule out or question inferred genes that may be artefacts introduced by somatic hypermutation, PCR, and/or sequencing errors. This is particularly true in cases where the more reliable IGHJ gene inference approach cannot be used for this pur-

Table 4
Number of IGHD genes (present in at least 1% of unique sequences as defined by IgDiscover) in each haplotype and genes that were also combined >10 times more frequently with one of two inferred alleles of IGHV genes of 5 donors (see also Table 4 in Reference Kirik et al. (2017)).

Donor	Number of IGHD genes in each haplotype*		IGHD genes differentially utilized by alleles of the two haplotypes†	IGHV genes with two or more alleles potentially segregated onto different chromosomes
1	9	12	IGHD3-3, IGHD6-6, IGHD2-8	IGHV1-2, IGHV1-69, IGHV2-70, IGHV3-11, IGHV3-13, IGHV3-20, IGHV3-48, IGHV3-7, IGHV4-59
3	14	16	IGHD3-3, IGHD6-6	IGHV1−69, IGHV2-70, IGHV3-11, IGHV3-20, IGHV3-48, IGHV3-7, IGHV4-39, IGHV4-4§
4	10	17	IGHD3-3, IGHD6-6, IGHD2-8, IGHD3-22, IGHD1-26	IGHV1-46, IGHV1-69, IGHV2-5, IGHV3-48, IGHV3-7, IGHV4-4, IGHV4-61
5	12	16	IGHD3-22, IGHD1-26	IGHV1-2, IGHV1-69, IGHV3-11, IGHV3-48, IGHV3-49, IGHV3-7
6	16	17	IGHD3-10	IGHV1−2, IGHV1-69, IGHV2-5, IGHV3-48, IGHV3-49, IGHV4-4§

* Excluding those IGHD genes that cannot be inferred with confidence (Kidd et al., 2016).
† The gene should be present in transcripts derived from one allele of the IGHV genes at a level of >1% and at a frequency >10 times higher than those belonging to the other allele.
§ In donors 3 and 6 both IGHV4-59*01 and *08 were inferred but these sequences showed evidence that at least one copy of each allele was associated with one of the haplotypes.

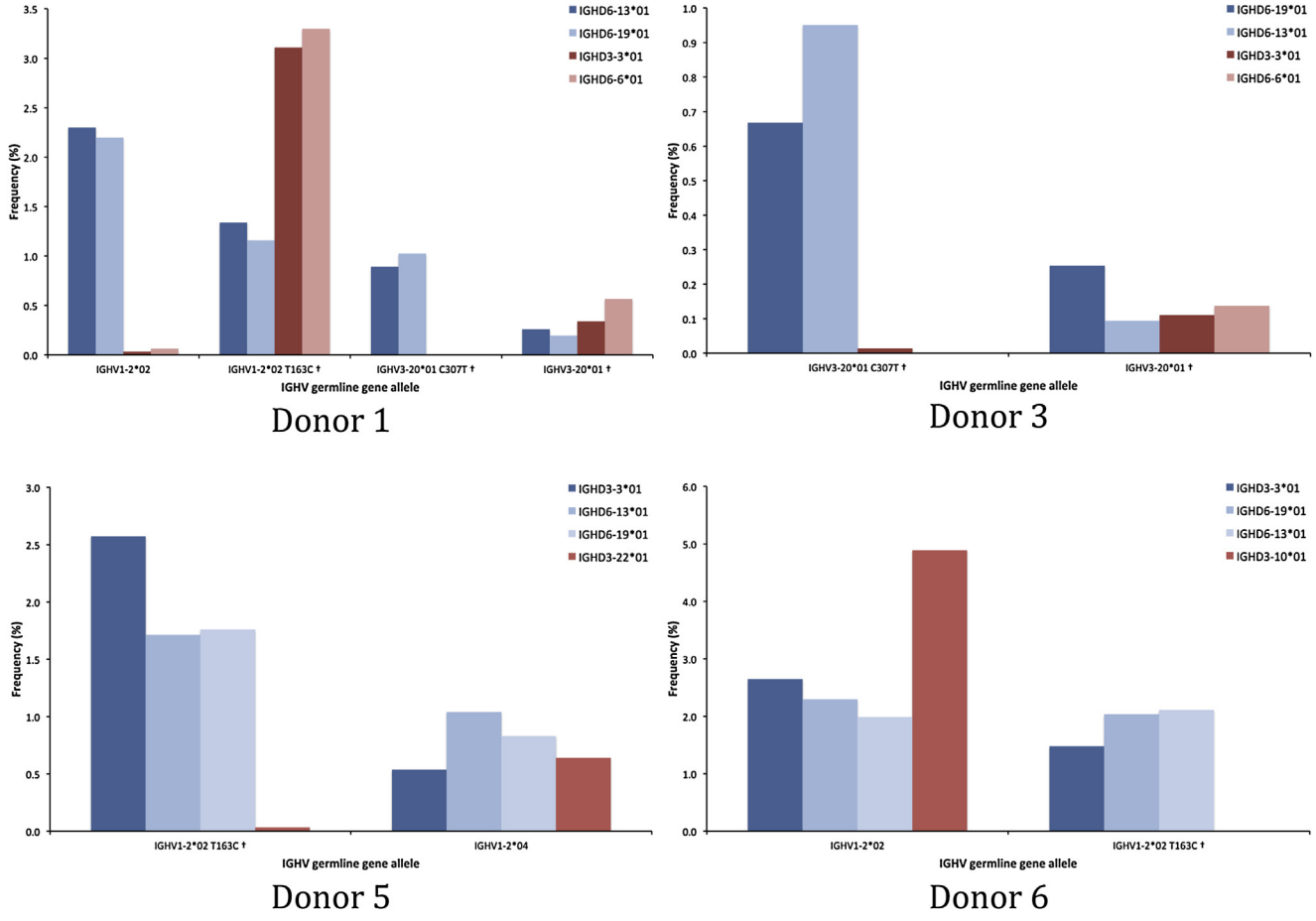

Fig. 3. Differential association of inferred IGHV alleles not present in the IMGT reference directory with different haplotypes of IGHD of donors 1, 3, 5, and 6. Sequences associated with IGHD genes present in only one of the haplotypes are shown in red while sequences associated to those IGHD genes that are present on both haplotypes are shown in blue. † defines that the name of only one of a set of different alleles of the gene that cannot be differentiated by the analysis approach is shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pose, as the subject is homozygous with respect to alleles of IGHJ genes.

3.7. Outlook beyond the present sequencing and analysis pipeline

Protocols for antibody gene repertoires are differently designed to address particular aspects of repertoires of interest in the context of the capacity of the available sequencing technology. Such protocols may focus on e.g. the sequence encoding the entire V domain, or on parts thereof in ways that may or may not ensure efficient assignment of individual sequences to particular antibody isotypes or even subclasses thereof. In either case the 5′ and/or 3′-primers will have to be placed at a distance from the V domain-encoding sequence. 5′-RACE may be used to minimize amplification bias, but this process also adds to the length of the genes to be sequenced (He et al., 2014). In some studies unique

molecular identifiers (UMI) (Shugay et al., 2014) are used to minimise the impact of PCR/sequencing errors on the analysis. Such approaches will improve the quality of sequences but incorporation of such barcodes also increases the length of the genes to be sequenced. In addition, the length of the encoded CDRH3 vary substantially between different transcripts and sequences encoding antibodies with long CDRH3 might more likely be lost in the quality pre-processing if barcodes or other sequences beyond those encoding the V domain are present. Given the maximum read length of currently used sequencing technology (usually Illumina MiSeq technology) any substantial increase in gene length may influence the quality of parts of the final, processed sequences used for analysis. In the context of haplotype-based analysis, such errors may influence our ability to properly assign each sequence to particular alleles. In particular, IGHD allele-based quality assessment (given the low frequency of use of some IGHD genes that display allelic diversity) may become problematic in experimental designs in which the IGHD-encoded sequence is in a region of lower read quality. Furthermore, any measure or procedure that substantially reduces the number of reads of each sample available for analysis will limit our ability to infer and validate genes/alleles expressed at low frequency, in particular when our quality analysis depends on IGHD and IGHJ alleles also expressed at low frequency. We envisage that future enhancement in sequencing quality, and read length may resolve this matter. Importantly, specific sequencing designs employing unique molecular identifiers have already been implemented that provide longer, high quality sequences (Turchaninova et al., 2016) likely suitable for germline gene inference and associated haplotype-based quality control. Furthermore, the possible influence of gene conversion (Duvvuri and Wu, 2012) on computational germline gene inference ought to be addressed in future studies to develop an understanding of this process' effect, if any, on the validity of inference procedures, including those performed in this study.

4. Conclusion

Germline gene inference has the potential to substantially enhance our understanding of the antibody repertoire in large numbers of subjects. Depending on specific features of the inference software particular attention should be paid to aspects, such as features of the inference-initiating germline database and software settings, that may affect assay specificity and sensitivity. We describe how assessment of sequencing quality can be used for assessment of inferred genes/alleles. We furthermore propose that haplotype assignment of specific germlines not only based on the presence of IGHJ gene alleles, as originally suggested by Kidd et al. (2012), but also on IGHD gene usage, adds substantial confidence to the analysis, both in terms of acceptance or rejection of inferred alleles as demonstrated in several cases. Such approaches will certainly add value to a best practice in germline gene inference technology. Although the present study was not designed to sequence the entire length of the IGHV gene, we were, through use of a diversity of inference methodology, able to lend support to the identification of alleles of IGHV1-2, IGHV3-20, IGHV3-43D, and IGHV3-53 currently not present in the IMGT reference directory. We were also able to, through use of inference methodology targeting IGHJ and/or IGHD genes, support identification of IGHV germline gene haplotypes, including one that had eliminated many major IGHV genes including IGHV3-23. We envisage that more frequent use of haplotype inference will identify important biological diversity in this respect and potentially identify subjects that carry haplotype variants worthy of further characterization through genomic locus sequencing or analysis of relationship of haplotype diversity to antibody functionality, in particular in cases where stereotyped antibody repertoires are important for antibody activity.

Conflicts of interest

The authors declare that there are no conflicts of interest in relation to this manuscript.

Role of funding source

The funding sources had no role in study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Acknowledgements

This study was supported by Lund University (ALF), the Swedish Research Council (grant number 2016-01720), and the Crafoord Foundation. We acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with NGS and access to the UPPMAX computational infrastructure.

References

Adderson, E.E., Shackelford, P.G., Quinn, A., Wilson, P.M., Cunningham, M.W., Insel, R.A., et al., 1993. Restricted immunoglobulin VH usage and VDJ combinations in the human response to Haemophilus influenzae type b capsular polysaccharide. Nucleotide sequences of monospecific anti-Haemophilus antibodies and polyspecific antibodies cross-reacting with self antigens. J. Clin. Invest. 91, 2734–2743, http://dx.doi.org/10.1172/JCI116514.

Alamyar, E., Duroux, P., Lefranc, M.P., Guidicelli, V., 2012. 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. 882, 569–604, http://dx.doi.org/10.1007/978-1-61779-842-9_32.

Boyd, S.D., Gaeta, B.A., Jackson, K.J., Fire, A.Z., Marshall, E.L., Merker, J.D., et al., 2010. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J. Immunol. 184, 6986–6992, http://dx.doi.org/10.4049/jimmunol.1000445.

Chimge, N.O., Pramanik, S., Hu, G., Lin, Y., Gao, R., Shen, L., et al., 2005. Determination of gene organization in the human IGHV region on single chromosomes. Genes Immun. 6, 186–193, http://dx.doi.org/10.1038/sj.gene.6364176.

Collins, A.M., Wang, Y., Roskin, K.M., Marquis, C.P., Jackson, K.J.L., 2015. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20140236, http://dx.doi.org/10.1098/rstb.2014.0236.

Corcoran, M.M., Phad, G.E., Nestor, V.B., Stahl-Hennig, C., Sumida, N., Persson, M.A., et al., 2016. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat. Commun. 7, 13642, http://dx.doi.org/10.1038/ncomms13642.

Duvvuri, B., Wu, G.E., 2012. Gene conversion-like events in the diversification of human rearranged IGHV3-23*01 gene sequences. Front. Immunol. 3, 158 http://dx.doi.org/10.3389/fimmu.2012.00158.

Elhanati, Y., Sethna, Z., Marcou, Q., Callan Jr., C.G., Mora, T., Walczak, A.M., 2015. Inferring processes underlying B-cell repertoire diversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20140243, http://dx.doi.org/10.1098/rstb.2014.0243.

Fuellgrabe, M.W., Herrmann, D., Knecht, H., Kuenzel, S., Kneba, M., Pott, C., et al., 2015. High-throughput, amplicon-based sequencing of the CREBBP gene as a tool to develop a universal platform-independent assay. PLoS One 10, e0129195, http://dx.doi.org/10.1371/journal.pone.0129195.

Gadala-Maria, D., Yaari, G., Uduman, M., Kleinstein, S.H., 2015. 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. 112, E862–70, http://dx.doi.org/10.1073/pnas.1417683112.

Gaeta, B.A., Malming, H.R., Jackson, K.J., Bain, M.E., Wilson, P., Collins, A.M., 2007. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. Bioinformatics 23, 1580–1587, http://dx.doi.org/10.1093/bioinformatics/btm147.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., Quake, S.R., 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. 32, 158–168, http://dx.doi.org/10.1038/nbt.2782.

Giudicelli, V., Chaume, D., Lefranc, M.P., 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33, D256–61, http://dx.doi.org/10.1093/nar/gki010.

Gupta, N.T., Vander Heiden, J.A., Uduman, M., Gadala-Maria, D., Yaari, G., Kleinstein, S.H., 2015. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics 31, 3356–3358, http://dx.doi.org/10.1093/bioinformatics/btv359.

He, L., Sok, D., Azadnia, P., Hsueh, J., Landais, E., Simek, M., et al., 2014. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Sci. Rep. 4, 6778, http://dx.doi.org/10.1038/srep06778.

Henry Dunand, C.J., Wilson, P.C., 2015. Restricted, canonical, stereotyped and convergent immunoglobulin responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, http://dx.doi.org/10.1098/rstb.2014.0238.

Kidd, M.J., Chen, Z., Wang, Y., Jackson, K.J., Zhang, L., Boyd, S.D., et al., 2012. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. J. Immunol. 188, 1333–1340, http://dx.doi.org/10.4049/jimmunol.1102097.

Kidd, M.J., Jackson, K.J., Boyd, S.D., Collins, A.M., 2016. DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. J. Immunol. 196, 1158–1164, http://dx.doi.org/10.4049/jimmunol.1501401.

Kirik, U., Greiff, L., Levander, F., Ohlin, M., Data on haplotype-supported immunoglobulin germline gene inference. *Data Brief* 2017 (in press).

Lefranc, M.P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., et al., 2015. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. Nucleic Acids Res. 43, D413–22, http://dx.doi.org/10.1093/nar/gku1056.

Levin, M., Levander, F., Palmason, R., Greiff, L., Ohlin, M., 2017. Antibody-encoding repertoires of bone marrow and peripheral blood-a focus on IgE. J. Allergy Clin. Immunol., http://dx.doi.org/10.1016/j.jaci.2016.06.040.

Lucas, A.H., McLean, G.R., Reason, D.C., O'Connor, A.P., Felton, M.C., Moulton, K.D., 2003. Molecular ontogeny of the human antibody repertoire to the Haemophilus influenzae type B polysaccharide: expression of canonical variable regions and their variants in vaccinated infants. Clin. Immunol. 108, 119–127.

Nadel, B., Tang, A., Lugo, G., Love, V., Escuro, G., Feeney, A.J., 1998. Decreased frequency of rearrangement due to the synergistic effect of nucleotide changes in the heptamer and nonamer of the recombination signal sequence of the V kappa gene A2b, which is associated with increased susceptibility of Navajos to Haemophilus influenzae type b disease. J. Immunol. 161, 6068–6073.

Neuberger, M.S., 2008. Antibody diversification by somatic mutation: from Burnet onwards. Immunol. Cell Biol. 86, 124–132, http://dx.doi.org/10.1038/sj.icb.7100160.

Ohlin, M., Borrebaeck, C.A.K., 1998. Insertions and deletions in hypervariable loops of antibody heavy chains contribute to molecular diversity. Mol. Immunol. 35, 233–238, http://dx.doi.org/10.1016/S0161-5890(98)00030-3.

Retter, I., Althaus, H.H., Munch, R., Muller, W., 2005. VBASE2, an integrative V gene database. Nucleic Acids Res. 33, D671–4, http://dx.doi.org/10.1093/nar/gki088.

Santosham, M., Rivin, B., Wolff, M., Reid, R., Newcomer, W., Letson, G.W., et al., 1992. Prevention of Haemophilus influenzae type b infections in Apache and Navajo children. J. Infect. Dis. 165 (Suppl. 1), S144–51.

Scheepers, C., Shrestha, R.K., Lambson, B.E., Jackson, K.J., Wright, I.A., Naicker, D., et al., 2015. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. J. Immunol. 194, 4371–4378, http://dx.doi.org/10.4049/jimmunol.1500118.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43, e37, http://dx.doi.org/10.1093/nar/gku1341.

Shugay, M., Britanova, O.V., Merzlyak, E.M., Turchaninova, M.A., Mamedov, I.Z., Tuganbaev, T.R., et al., 2014. Towards error-free profiling of immune repertoires. Nat. Methods 11, 653–655, http://dx.doi.org/10.1038/nmeth.2960.

Turchaninova, M.A., Davydov, A., Britanova, O.V., Shugay, M., Bikos, V., Egorov, E.S., et al., 2016. High-quality full-length immunoglobulin profiling with unique molecular barcoding. Nat. Protoc. 11, 1599–1616, http://dx.doi.org/10.1038/nprot.2016.093.

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N., O'Connor, K.C., Hafler, D.A., et al., 2014. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics 30, 1930–1932, http://dx.doi.org/10.1093/bioinformatics/btu138.

van Dongen, J.J., Langerak, A.W., Bruggemann, M., Evans, P.A., Hummel, M., Lavender, F.L., et al., 2003. 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 17, 2257–2317, http://dx.doi.org/10.1038/sj.leu.2403202.

Wang, Y., Jackson, K.J., Sewell, W.A., Collins, A.M., 2008. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol. Cell Biol. 86, 111–115, http://dx.doi.org/10.1038/sj.icb.7100144.

Watson, C.T., Breden, F., 2012. The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun. 13, 363–373, http://dx.doi.org/10.1038/gene.2012.12.

Watson, C.T., Steinberg, K.M., Huddleston, J., Warren, R.L., Malig, M., Schein, J., et al., 2013. 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. 92, 530–546, http://dx.doi.org/10.1016/j.ajhg.2013.03.004.

Wilson, P.C., de Bouteiller, O., Liu, Y.J., Potter, K., Banchereau, J., Capra, J.D., et al., 1998. Somatic hypermutation introduces insertions and deletions into immunoglobulin genes. J. Exp. Med. 187, 59–70, http://dx.doi.org/10.1084/jem.187.1.59.

Yaari, G., Uduman, M., Kleinstein, S.H., 2012. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res. 40, e134, http://dx.doi.org/10.1093/nar/gks457.

Ye, J., Ma, N., Madden, T.L., Ostell, J.M., 2013. IgBLAST. an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 41, W34–40, http://dx.doi.org/10.1093/nar/gkt382.

Zhang, W., Wang, I.M., Wang, C., Lin, L., Chai, X., Wu, J., et al., 2016. IMPre: an accurate and efficient software for prediction of T- and B-cell receptor germline genes and alleles from rearranged repertoire data. Front. Immunol. 7, 457, http://dx.doi.org/10.3389/fimmu.2016.00457.