Data on haplotype-supported immunoglobulin germline gene inference

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

A R T I C L E   I N F O

Article history:
Received 21 March 2017
Received in revised form
26 May 2017
Accepted 19 June 2017
Available online 27 June 2017

Keywords:
Antibody
Gene inference
Germline repertoire
Immunoglobulin germline gene
Transcriptome
Validation

A B S T R A C T

Data that defines IGHV (immunoglobulin heavy chain variable) germline gene inference using sequences of IgM-encoding transcriptomes obtained by Illumina MiSeq sequencing technology are described. Such inference is used to establish personalized germline gene sets for in-depth antibody repertoire studies and to detect new antibody germline genes from widely available immunoglobulin-encoding transcriptome data sets. Specifically, the data has been used to validate (Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery (DOI: 10.1016/j.molimm.2017.03.012) (Kirik et al., 2017) [1]) the inference process. This was accomplished based on analysis of the inferred germline genes’ association to the donors’ different haplotypes as defined by their different, expressed IGHJ alleles and/or IGHD genes/alleles. The data is important for development of validated germline gene databases containing entries inferred from immunoglobulin-encoding transcriptome sequencing data sets, and for generation of valid, personalized antibody germline gene repertoires.

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
**Specifications Table**

| Subject area               | Biology, Medicine                        |
|----------------------------|------------------------------------------|
| More specific subject area | Immunobiology                            |
| Type of data               | Sequence reads, tables, figures          |
| How data was acquired      | Next generation sequencing using Illumina MiSeq technology; analysis using immunoglobulin repertoire inference software |
| Data format                | Raw data, analyzed data                  |
| Experimental factors       | Data processing was performed using pRESTO, Change-O, TIgGER, IgDiscover, GlgGle |
| Experimental features      | Immunoglobulin M heavy chain variable domain-encoding genes were amplified by RT-PCR, sequenced by next generation sequencing technology, and analyzed by bioinformatics approaches. |
| Data accessibility         | FASTQ raw sequence data files are available from the European Nucleotide Archive, study accession number: PRJEB18926. Data is within this article. Code available at https://github.com/ukirik/giggle |

**Value of the data**

- The data is valuable for development of computational inference approaches that feature improved confidence in the outcomes of the inference process.
- The data is valuable for development of validated immunoglobulin germline gene databases.
- The data is valuable for validation of computational inference of personalized antibody germline gene repertoires.
- The data is valuable for the analytical process preceding studies of evolution of immune responses.

1. Data

The data of this article summarize the identity and accession numbers of sequencing data files (Table 1), the sizes of the sequence sets during the different stages of data processing (Table 2), and the outcome of validation of new inferred genes/alleles (Table 3), identified by use of IgDiscover and TIgGER. The frequencies of readily inferable [2] IGHD (Immunoglobulin heavy D-gene) genes used by the two haplotypes of five subjects are summarized (Table 4). Furthermore the data illustrate the effect of using a germline gene database that extends beyond codon 105 on gene inference (Fig. 1), and summarizes the outcome of TIgGER-based germline gene inference of six transcriptomes (Fig. 2). The data also illustrates how low sequencing quality scores are associated with some, but certainly not all, inferred germline gene alleles (Fig. 3), and summarizes IGHJ (Immunoglobulin heavy J-gene) alleles used by transcriptomes of six subjects (Fig. 4). The link between inferred IGHV (Immunoglobulin heavy V-gene) germline genes/alleles and different alleles of IGHJ6 in bone marrow (BM)- and peripheral blood (PB)-derived transcriptomes of two hetrozygous subjects is shown (Fig. 5). The data summarizes linkage of different IGHD genes to two different haplotypes defined by alleles of IGHJ6 or defined by hetrozygous IGHV genes (Fig. 6). The linkage of IGHV1-8, IGHV3-9, IGHV5-10-1, and IGHV3-64D germline genes to different haplotypes in subjects with two different IGHD gene-defined haplotypes (Fig. 7) is shown. Association of IGHV germline genes/alleles with particular IGHD genes in five subjects with different IGHD-defined haplotypes is shown (Fig. 8), as is the extent of association of alleles of IGHV4-59 to particular IGHD genes (Fig. 9). Finally, data describing assessment of alleles of IGHD genes detected in IgM-encoding transcriptomes of six subjects (Fig. 10), and of IGHV germline genes associated to the different alleles of IGHD genes in two subjects (Fig. 11) is shown.
Table 1
Summary of identity of sequenced samples of the study (European Nucleotide Archive (ENA) accession number PRJEB18926).\(^a\)

| Subject | Sample origin\(^b\) | Replicate | Sequencing sample ID | Isotypes | ENA sample accession number | ENA experiment accession number |
|---------|---------------------|-----------|----------------------|----------|-----------------------------|-------------------------------|
| 1 BM    | 1                   | P1882_1001 | IgA, IgE, IgG, IgM   | ERS1531209 | ERX1875309                  |
| 1 BM    | 2                   | P1882_1002 | IgA, IgE, IgG, IgM   | ERS1531210 | ERX1875310                  |
| 1 PB    | 1                   | P1882_1007 | IgA, IgE, IgG, IgM   | ERS1531215 | ERX1875315                  |
| 1 PB    | 2                   | P1882_1008 | IgA, IgE, IgG, IgM   | ERS1531216 | ERX1875316                  |
| 2 BM    | 1                   | P1882_1003 | IgA, IgE, IgG, IgM   | ERS1531211 | ERX1875311                  |
| 2 BM    | 2                   | P1882_1004 | IgA, IgE, IgG, IgM   | ERS1531212 | ERX1875312                  |
| 2 PB    | 1                   | P1882_1009 | IgA, IgE, IgG, IgM   | ERS1531217 | ERX1875317                  |
| 2 PB    | 2                   | P1882_1010 | IgA, IgE, IgG, IgM   | ERS1531218 | ERX1875318                  |
| 3 BM    | 1                   | P1882_1005 | IgA, IgE, IgG, IgM   | ERS1531213 | ERX1875313                  |
| 3 BM    | 2                   | P1882_1006 | IgA, IgE, IgG, IgM   | ERS1531214 | ERX1875314                  |
| 3 PB    | 1                   | P1882_1011 | IgA, IgE, IgG, IgM   | ERS1531219 | ERX1875319                  |
| 3 PB    | 2                   | P1882_1012 | IgA, IgE, IgG, IgM   | ERS1531220 | ERX1875320                  |
| 4 BM    | 1                   | P1882_1013 | IgA, IgE, IgG, IgM   | ERS1531221 | ERX1875321                  |
| 4 BM    | 2                   | P1882_1014 | IgA, IgE, IgG, IgM   | ERS1531222 | ERX1875322                  |
| 4 PB    | 1                   | P1882_1019 | IgA, IgE, IgG, IgM   | ERS1531227 | ERX1875327                  |
| 4 PB    | 2                   | P1882_1020 | IgA, IgE, IgG, IgM   | ERS1531228 | ERX1875328                  |
| 5 BM    | 1                   | P1882_1015 | IgA, IgE, IgG, IgM   | ERS1531223 | ERX1875323                  |
| 5 BM    | 2                   | P1882_1016 | IgA, IgE, IgG, IgM   | ERS1531224 | ERX1875324                  |
| 5 PB    | 1                   | P1882_1021 | IgA, IgE, IgG, IgM   | ERS1531229 | ERX1875329                  |
| 5 PB    | 2                   | P1882_1022 | IgA, IgE, IgG, IgM   | ERS1531230 | ERX1875330                  |
| 6 BM    | 1                   | P1882_1017 | IgA, IgE, IgG, IgM   | ERS1531225 | ERX1875325                  |
| 6 BM    | 2                   | P1882_1018 | IgA, IgE, IgG, IgM   | ERS1531226 | ERX1875326                  |
| 6 PB    | 1                   | P1882_1023 | IgA, IgE, IgG, IgM   | ERS1531231 | ERX1875331                  |
| 6 PB    | 2                   | P1882_1024 | IgA, IgE, IgG, IgM   | ERS1531232 | ERX1875332                  |

\(^a\) Read numbers representing each sample/isotype are available in Supplementary Table EIV of Ref. [3].
\(^b\) BM: bone marrow; PB: peripheral blood.
\(^c\) No PCR product was derived using IgE-specific 3'-primers.

Table 2
Number of IgM-encoding sequences at different stages of the analysis process.

| Donor | Tissue\(^b\) | \# of reads after filtering\(^c\) | \# of sequences after PRESTO pipeline\(^c\) | \# of unique sequences\(^d\) | \# of unique sequences with V_errors=0\(^d\) | \# of unique sequences with V_errors=0 & \(D\text{-}\text{coverage} > 35\%\)\(^d\) |
|-------|--------------|---------------------------------|---------------------------------|-----------------|---------------------------------|---------------------------------|
| 1 BM  | 1            | 258,988                         | 261,967                         | 86,135          | 47,233                          | 43,006                          |
| 2 BM  | nd           | 194,555                         | 197,949                         | 90,181          | 58,685                          | 52,815                          |
| 3 BM  | 1            | 278,426                         | 281,711                         | 70,515          | 28,827                          | 26,400                          |
| 4 BM  | nd           | 339,935                         | 345,021                         | 91,511          | 45,510                          | 40,850                          |
| 5 BM  | nd           | 318,207                         | 324,269                         | 106,047         | 63,924                          | 57,998                          |
| 6 BM  | 1            | 406,893                         | 412,689                         | 152,125         | 85,956                          | 77,603                          |

\(^a\) BM: bone marrow; PB peripheral blood
\(^b\) Number of sequences used for initiation of the workflow towards TiGGER-based analysis.
\(^c\) Number of sequences used for initiation of the workflow towards IgDiscover-based analysis.
\(^d\) Number of unique sequences in the final filtered output obtained using IgDiscover as inference method.
2. Experimental design, materials and methods

IgM heavy chain variable domain-encoding gene repertoires were isolated by RT-PCR from transcriptomes of PB and BM collected out of season of most seasonal allergens from six allergic subjects [3]. Ethical approval and informed consent had been obtained from all donors. Sequencing was performed using the 2 × 300 bp MiSeq technology (Illumina, Inc., San Diego, CA, USA) at the National Genomics Infrastructure (SciLifeLab, Stockholm, Sweden) [3]. Details of sequence output and availability are outlined in Table 1. Data was pre-processed using pRESTO [4] and Change-O [5] as summarized in Fig. 1 in Ref. [1]. Germline gene inference was performed using TiGER [6] and IgDiscover [7]. Additional bioinformatics analysis was performed as outlined elsewhere [1] including analysis performed using GlgGle (release 0.2) that is available under Apache License at https://github.com/ukirik/giggle. Immunoglobulin gene names and sequence numbering complies with the nomenclature defined by the International ImMunoGeneTics information system® (IMGT) (http://www.imgt.org) [8,9].

| Sequence variant | Difference from IMGT sequence in mutational hot spot | Shortlisted by IgDiscover (ifEF<0) | Comment | Inferred allele composition of gene as proposed by IgDiscover (ifEF<0) / TiGER inferred genotype | Shortlisted by TiGER (donor #) |
|-----------------|-----------------------------------------------|-------------------------------|---------|---------------------------------------------------|----------------------------|
| IGHV1-2*02 T163C or IGHV1-2*065 T299C (IgPdh: IGHV1-2*06) | No | Subjects 1, 5, 6 | Transcripts present at levels similar to subject’s other allele (*02 or *04). Multiple independent rearrangements identified. Expressed from a different haplotype as compared to the other IGHV1-2 gene present in these subjects, as demonstrated by linkage to alleles of IGHJ6 (donor 5) and particular IGHD genes (donor 1, 5, and 6). The read quality of the allele-differentiating base defining the T163C variant is shown in Figures 3I-K. | Yes (1, 5, 6) (implicated in final result depending on assay setting) |
| IGHV1-69*02 A112C | Yes (VA) | Subject 6 | Transcripts present at levels substantially lower than subject’s other alleles (*01 and *02). | Yes (4, 6) (not implicated in final result) |
| IGHV2-5*02 A87C | No | Subject 1 | Transcripts present at levels substantially lower than subject’s other allele (*02). | No |
| IGHV2-5*02 A108C | No | Subjects 5, 6 | Transcripts present at levels substantially lower than subject’s other allele (*02) and (for subject 6) *01. Haplotype inference indicated presence of IGHV2-5*02 in both haplotypes of donor 5. | No |
| Sequence variant | Difference from IMGT sequence in mutational hot spot | Shortlisted by IgDiscover (diff=0) | Comment | Inferred allele composition of gene as proposed by IgDiscover (diff=0) / TlgGER inferred genotype | Shortlisted by TlgGER (donor #) |
|------------------|-----------------------------------------------------|----------------------------------|---------|-------------------------------------------------|---------------------------|
| IGHV3-26*01 C307T (IgPAb: IGHV3-20*p02) | No | Subjects 1, 3 | Transcripts inferred at levels higher than subject’s other allele (*01 for donor 1). Inspection of IMGT/High-VQUEST analysis also identified transcripts derived from IGHV3-20*01 in donor 3 but at a level approximately 3-fold lower than IGHV3-20*01 C307T. IgDiscover similarly identified both alleles in the larger P9-derived data set of donor 3 (data not shown) with IGHV3-20*01 C307T as the dominant component. Transcripts of IGHV3-20*03 C307T is represented by multiple CDRH3, multiple CDRH3 lengths, and they were associated with all IGH genes ([Gilij1]-6). The read quality of the allele-differentiating base defining the C307T variant is shown Figures S3, N. | IgDiscover: IGHV3-20*01: blue; IGHV3-20*01 C307T: green | Yes (1) (implicated in final result depending on assay setting) |
| IGHV3-43*01 C195A (IgPAb: IGHV3-43*p04 - but more similar in sequence to IGHV3-43D) | No | Subject 6 | Analysis of 140 sequences of donor 6 assigned by IMGT/High-VQUEST to IGHV3 43D*01 with 1 nucleotide difference (none were completely identical), all demonstrated the C195A difference. These represented 98 different CDRH3, 19 different CDRH3 lengths, and they were associated with all IGH genes ([Gilij1]-6). The sequence variant differs from IGHV3-43*01, a gene that was present in twice as many transcripts, by 3 nucleotides. The read quality of the base defining the C195A variant is shown in Figure S3B. | IgDiscover: IGHV3-43D*01: green; IGHV3-43D*01 C195A: blue | No |
| IGHV3-53*01 G88A (IgPAb: new putative allele IGHV3-53*p07) | No | Subject 2 | Transcripts present at levels similar to subject’s other allele (*01). Multiple independent rearrangements identified. The read quality of the allele-differentiating base defining the G88A variant is shown in Figure S1A. | IgDiscover: IGHV3-53*01: blue; IGHV3-53*01 G88A: green | Yes (2) (implicated in final result depending on assay setting) |
| IGHV4-31*02 A91C | No | Not inferred | | IgDiscover: IGHV4-31*02: blue | Yes (5) (not implicated in final result) |
### Table 3 (continued)

| Sequence variant | Difference from IMGT sequence in mutational hot spot | Shortlisted by TrGer (d.f.<0) | Comment | Inferred allele composition of gene as proposed by IgDiscover (d.f.<0) | Shortlisted by TrGer (donor #) |
|------------------|-----------------------------------------------|-----------------------------|---------|------------------|-----------------------------|
| IGHV4-34*01 A110C | No | Subjects 4, 6 | Transcripts present at levels substantially lower than subject's other allele ("01"). Haplotype inference furthermore indicated presence of IGHV4-34*01 in both haplotypes. | No | |
| IGHV4-34*01 A110C | Yes (WA) | Subject 1, 6 | Transcripts present at levels substantially lower than subject's other allele ("01"). Haplotype inference furthermore indicated presence of IGHV4-34*01 in both haplotypes. | No | |
| IGHV4-38-2*01 A83C | Yes (WA) | Subjects 5, 6 | Transcripts present at levels lower than subject's other allele ("01"). Multiple independent rearrangements identified. However, haplotype inference indicated presence of IGHV4-38-2*01 in both haplotypes of donor 5. IMGT/High-VQUEST analysis of the entire amplicon (including short amplified part of FR1 not analysed by IgDiscover as employed in this study) was able to identify both alleles "01 and "02 in subject 5. A fraction of transcripts of both alleles incorporated the ABC modification in transcripts of both IGHV4-38-2*01 and "02. | Yes (6) [not implicated in final result] | |
| IGHV4-39*01 A91C | No | Subject 4 | Transcripts present at levels substantially lower than subject's other allele ("01"). Haplotype inference indicated presence of IGHV4-39*01 in both haplotypes of donor 4. | No | |
| IGHV4-39*01 A143C | Yes (WA) | Subjects 1, 4, 6 | Transcripts present at levels substantially lower than subject's other allele(s) ("01 and "07 (subject 3)). Haplotype inference indicated presence of IGHV4-39*01 in both haplotypes of donor 4. The low read quality of the allele-differentiating base defining the A143C variant is shown in Figures 3A-D, E. | No | |
| IGHV4-39*01 A91C | No | Subject 5 | Transcripts present at levels substantially lower than subject's other allele ("07). Haplotype inference indicated presence of IGHV4-39*01 and the variant in the same haplotype of donor 5. | No | |
| IGHV4-39*01 A143C | Yes (WA) | Subject 5 | Transcripts present at levels substantially lower than subject's other allele ("07). Haplotype inference indicated presence of IGHV4-39*01 and the variant in the same haplotype of donor 5. The low read quality of the allele-differentiating base defining the A143C variant is shown in Figure 3E. | No | |
| IGHV6-1*01 A83C | No | Subjects 5, 6 | Transcripts present at levels substantially lower than subject's other allele ("01). Furthermore, haplotype inference indicated presence of IGHV6-1*01 in both haplotypes of donor 5. The low read quality of the allele-differentiating base defining the A83C variant is shown in Figures 3G, H. | Yes (5, 6) [not implicated in final result] | |
| IGHV6-1*01 A194C | Yes (WA) | Subject 5 | Transcripts present at levels substantially lower than subject's other allele ("01). Furthermore, haplotype inference indicated presence of IGHV6-1*01 in both haplotypes of donor 5. | No | |

**IgDiscover:** IGHV4-34*01: blue; IGHV4-34*01 A110C: orange; IGHV4-34*01 A110C: green

**TrGer:** Mutation patterns for polymorphism positions in IGHV sequences of IGHV6-1*01 A83C of donors 5 (top) and 6 (bottom).
Table 4
Estimated frequency* of use of readily identified IGHD germline genes[2] in haplotypes of five lymphocyte donors, and the ratio of estimated frequency† of these genes in the two haplotypes.

| D-gene | Donor 1 | Donor 2 | Donor 3 | Donor 4 | Donor 5 | Donor 6 |
|--------|---------|---------|---------|---------|---------|---------|
|        | Haploype #1 (%) | Haploype #2 (%) | Haploype #3 (%) | Haploype #4 (%) | Haploype #5 (%) | Haploype #6 (%) |
| IGHD6-2 | 0.8 | 5.3 | 149 | 5.2 | 7.8 | 74 |
| IGHD3-3 | 0.0 | 10.8 | 24 | 0.1 | 0.8 | 27 |
| IGHD4-6 | 0.1 | 6.2 | 17 | 0.1 | 2.4 | 21 |
| IGHD2-8 | 0.0 | 2.0 | 4 | 0.1 | 1.0 | 3 |
| IGHD3-9 | 9.4 | 3.9 | 261 | 5.8 | 1.8 | 319 |
| IGHD5-10 | 24.7 | 10.2 | 241 | 22.5 | 8.4 | 204 |
| IGHD5-12 | 4.9 | 6.2 | 88 | 3.0 | 2.4 | 124 |
| IGHD6-13 | 12.8 | 11.0 | 116 | 14.9 | 8.8 | 170 |
| IGHD1-144 | 0.2 | 0.3 | 56 | 0.2 | 0.3 | 71 |
| IGHD2-15 | 7.2 | 10.9 | 66 | 4.2 | 6.9 | 84 |
| IGHD3-16 | 2.6 | 3.4 | 77 | 1.8 | 1.8 | 99 |
| IGHD4-17 | 0.1 | 9.7 | 63 | 4.3 | 5.4 | 79 |
| IGHD6-19 | 12.1 | 10.5 | 115 | 10.3 | 10.7 | 97 |
| IGHD2-21 | 1.0 | 0.9 | 113 | 1.8 | 2.1 | 91 |
| IGHD3-22 | 0.0 | 0.2 | 285 | 12.1 | 18.5 | 73 |
| IGHD4-23 | 0.3 | 0.4 | 75 | 1.6 | 1.6 | 103 |
| IGHD5-24 | 0.2 | 0.3 | 52 | 1.4 | 2.5 | 51 |
| IGHD6-25 | 0.2 | 0.2 | 61 | 0.2 | 0.2 | 80 |
| IGHD1-16 | 0.2 | 0.2 | 121 | 4.5 | 5.5 | 96 |

Number of IGHD genes used at frequency >1%: 9

* IGHD germline genes used at a frequency >1% are highlighted on a red background.
† Haplotypes that differ >10-fold in estimated frequency of use are highlighted on a green background.
‡ Entries that fulfill both herein used criteria, frequency of use >1% and >10-fold difference in frequency of use between haplotypes. Some rarely used IGHD genes are unable to meet both criteria.

Fig. 1. Germline gene variants of IGHV1-18 and IGHV3-21 inferred by IgDiscover when a starting germline database extending beyond codon 105 was used to initiate the process. The number of sequence counts (A) and unique CDRH3 (B) are shown. Examples (IGHV1-18, IGHV3-21, IGHV3-33, IGHV3-48, and IGHV4-59) of germline genes with new inferred variants, mostly in codon 106, and their similar association to the two different alleles of IGHJ6 of donor 4 (C) and donor 5 (D) are shown. Segregation of different established alleles of IGHV3-48 to the two alleles of IGHJ6 is also shown for comparison. † 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.
Fig. 2. Genotype inferred by TlgGER using IgM-encoding transcripts of BM. Note difference in the calling of IGHV1-2. Heterozygous state of IGHV1-2 (*02/*p06) is inferred in subjects 1 and 6 only when argument find_unmutated = true while it is inferred in subject 2 (*02/*04) independently of the setting of find_unmutated. Heterozygous state of IGHV3-7 (*01/*02) is inferred in subjects 1, 3, and 4 only when argument find_unmutated = false while it is inferred in subject 5 (*01/*03) independently of the setting of find_unmutated. Heterozygous state of IGHV3-20 (*01/*01 C307T) is inferred in subject 1 only when argument find_unmutated = true and the allele variant is not at all inferred in donor 3. Heterozygous state of IGHV3-64 is inferred in donors 1, 3, 4, and 6 when argument find_unmutated = false and in donor 1 when argument find_unmutated = true.
Fig. 2. (continued)
Fig. 3. Quality score of sequencing reads representing germline genes inferred by IgDiscover. Sequence reads representing sequence variant A143C of IGHV4-39 show lower read quality (donors 1 (A), 2 (B), 3, (C), 4 (D), 5 (E), and 6 (F)) of the nucleotide representing the allele-differentiating base as opposed to reads defining the corresponding unmutated alleles (IGHV4-39*01 and *07). Similarly, inferred allele IGHV6-1 A85C shows low read quality of the allele-differentiating base (donor 5 (G), donor 6 (H)). Sequence reads representing parts of the sequences of alleles of IGHV1-2*02 and IGHV1-2*04 (represented by nucleotide T163) and IGHV1-2*06 (C163) of donors 1 (I), 5 (J), and 6 (K) show highly similar read quality. Sequences representing IGHV3-53*01 and IGHV3-53*01 G88A of donor 2 (L), IGHV3-20*01 and IGHV3-20*01 C307T of donor 1 (M) and donor 3 (N), and IGHV3-43D*01 C195A of donor 6 (O) show high quality of the allele-differentiating base calls. The analysed sequence is shown above each graph and the allele-differentiating base is highlighted within square brackets.
Fig. 3. (continued)
Fig. 3. (continued)
Fig. 4. Perceived frequency of IGHJ gene usage in transcripts derived from donors 1–6, as analysed by IgDiscover.
Fig. 5. Summary of linkage of inferred germline genes/alleles of donors 4 (A) and 5 (B) to IGHJ6*02 and *03, as indicators of the donors’ two haplotypes, after analysis of transcripts found in bone marrow (BM) (also shown in Fig. 2 in Ref. [1]) and peripheral blood (PB). † 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.
Fig. 6. Association of IGHD gene expression IGHJ expression in unique IgM-encoding transcripts (at \( V_{\text{errors}} = 0 \) and \( D_{\text{coverage}} > 35 \) as defined by IgDiscover) derived from PB of donor 4 (A) and donor 5 (C). Association of IGHD gene expression (average \( \pm \) SD) to that of IGHV genes inferred as being present as two different alleles in transcripts derived from PB donors 1 (E), 3 (F), 4 (B), 5 (D), and 6 (G). A summary of IGHD gene usage (irrespective of allele call) based on association to expression of IGHV genes is shown (H).
Fig. 7. Linkage of IGHV1-8*01, IGHV3-64D*06, IGHV3-9*01, and IGHSV5-10-1*01 to different IGHD genes in transcripts of donor 1, 3, and 5. While germline genes IGHV1-8*01 and IGHV3-9*01 were linked to the haplotype also carrying IGHD genes not present on both haplotypes, IGHV3-64D*06 and IGHSV5-10-1*01 were not.
Fig. 8. Association of IGHV genes/alleles of donors 1 (A), and 3–6 (B–E) with different IGHD genes as indicators of association with different haplotypes represented by IGHD. Analysis was performed on sequences found in cells of PB using the final filtered output of IgDiscover \( \text{diff} = 0 \). Only IGHV genes/alleles represented by at least 50 sequences with \( \text{V_errors} = 0 \) and \( \text{D_coverage} \geq 35 \) in the IGHD gene set shown in dark blue are shown. The frequencies of IGHV sequences associated to IGHD genes found in both haplotypes are shown in blue while the corresponding frequencies of IGHV sequences associated to IGHD genes expressed from only one of the inferred haplotypes are shown in red. † 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.
**Fig. 9.** Differential association of inferred alleles of IGHV4-59 with different haplotypes of IGHD of donors 1 (A), 3 (B), and 6 (C). The frequencies of sequences associated to IGHD genes apparently expressed from both haplotypes are shown in blue while the frequencies of sequences associated to IGHD genes apparently expressed from only one of the haplotypes are shown in red. The fraction of reads represented by IGHV4-59*01 (blue) and *08 (green) in all three subjects is shown (fraction of sequences to the left and fraction of unique CDR3 to the right) (D). † 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.
Fig. 10. Apparent utilization of alleles of IGHD genes in IgM-encoding transcripts of BM of donors 1 (A), 2 (B), 3 (C), 4 (D), 5 (E), and 6 (F), as annotated by IgDiscover.
Fig. 11. Immunoglobulin IGHV gene haplotype analysis based on heterozygous presence of IGHD alleles of donor 1 (A, B) and donor 5 (C, D). Transcripts found in BM (A, C) and PB (B, D) were analysed. The analysis of transcripts derived from PB employing IGHD2-21 was not included due to the low number of such sequences. Detailed sequence analysis (E) may be used to define whether or not IGHD allele assignments are appropriate. The rare association of reads of IGHV1-2*02 to IGHD2-21*01 (grey) instead of the expected IGHD2-21*02 (black) in some BM-derived transcripts of donor 1 (see A) does not cover the base within the IGHD that defines the individual alleles. IGHD2-21 allele calls for both alleles of IGHV4-59*01 include the allele-differentiating base, and rearrangements involving IGHV4-59*08 include the base identifying IGHD2-21*02. The arrow indicates the only base that differentiate IGHD2-21*01 and *02. Mutated bases within the sequences derived from IGHD genes are spelled out. † 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.
Acknowledgements

The collection and analysis of the data set was supported by Lund University (ALF), the Swedish Research Council (Grant number 2016-01720), the Crafoord Foundation. 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 assisted with NGS and access to the UPPMAX computational infrastructure.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.06.031.

References

[1] U. Kirik, L. Greiff, F. Levander, M. Ohlin, Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery, Mol. Immunol. 87 (2017) 12–22.
[2] M.J. Kidd, K.J. Jackson, S.D. Boyd, A.M. Collins, D.J. pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes, J. Immunol. 196 (2016) 1158–1164.
[3] M. Levin, F. Levander, R. Palmason, L. Greiff, M. Ohlin, Antibody-encoding repertoires of bone marrow and peripheral blood— a focus on IgE, J. Allergy Clin. Immunol. 139 (2017) 1026–1030.
[4] J.A. Vander Heiden, G. Yaari, M. Uduman, J.N. Stern, K.C. O’Connor, D.A. Hafler, F. Vigneault, S.H. Kleinstein, pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires, Bioinformatics 30 (2014) 1930–1932.
[5] N.T. Gupta, J.A. Vander Heiden, M. Uduman, D. Gadala-Maria, G. Yaari, S.H. Kleinstein, Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data, Bioinformatics 31 (2015) 3356–3358.
[6] D. Gadala-Maria, G. Yaari, M. Uduman, S.H. Kleinstein, Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles, Proc. Natl. Acad. Sci. USA 112 (2015) E862–E870.
[7] M.M. Corcoran, G.E. Phad, V.B. Nestor, C. Stahl-Hennig, N. Sumida, M.A. Persson, M. Martin, G.B. Karlsson Hedestam, Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity, Nat. Commun. 7 (2016) 13642.
[8] M.P. Lefranc, 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) 633–642.
[9] M.P. Lefranc, From I.M.G.T.—O.N.T.O.L.O.G.Y. CLASSIFICATION Axiom, to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR), Cold Spring Harb. Protoc. 2011 (2011) 627–632.