



ELSEVIER

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Data on the nucleotide composition of the first codons encoding the complementary determining region 3 (CDR3) in immunoglobulin heavy chains



Linnea Thörnqvist, Mats Ohlin*

Dept. of Immunotechnology, Lund University, Lund, Sweden

ARTICLE INFO

Article history:

Received 22 February 2018

Accepted 30 April 2018

Available online 4 May 2018

ABSTRACT

The highly variable complementary determining region 3 (CDR3) of antibodies is generated through recombination of immunoglobulin heavy chain variable (IGHV), diversity, and joining genes. The codons encoding the first residues of CDR3 may be derived directly from the IGHV germline gene but they may also be generated as part of the rearrangement process. Data of the nucleotide composition of these codons of rearranged genes, an indicator of the degree of contribution of the IGHV gene to CDR3 diversity, are presented in this article. Analyzed data are presented for two unrelated sets of raw sequence data. The raw data sets consisted of sequences of antibody heavy chain-encoding transcripts of six allergic subjects (European Nucleotide Archive accession number PRJEB18926), and paired antibody heavy and light chain variable region-encoding transcripts of memory B cells of three subjects (European Nucleotide Archive accession numbers SRX709625, SRX709626, and SRX709627). The nucleotide compositions of the corresponding 5'-ends of sequences encoding the CDR3 are presented for transcripts with an origin in 47 different IGHV alleles. These data have been used (Thörnqvist and Ohlin, 2018) [1] to demonstrate the extent of incorporation of the 3' most bases of IGHV germline genes into rearranged immunoglobulin encoding sequences, and the extent whereby any difference in incorporation affects the specificity of inference of the 3'-end of IGHV genes from immunoglobulin-encoding transcripts. They have also been used

DOI of original article: <https://doi.org/10.1016/j.molimm.2018.02.013>

* Corresponding author.

E-mail address: mats.ohlin@immun.lth.se (M. Ohlin).<https://doi.org/10.1016/j.dib.2018.04.125>2352-3409/© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

to assess the effect of observed gene differences on the composition of the ascending strand of CDR3 associated to antibodies with an origin in different IGHV genes (Thörnqvist and Ohlin, 2018) [1].

© 2018 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
More specific subject area	Immunobiology
Type of data	Figures, table
How data was acquired	Next generation sequencing (MiSeq, Illumina)
Data format	Analyzed
Experimental factors	Extraction of peripheral blood mononuclear cell RNA, construction of libraries encoding antibody heavy chain variable domains
Experimental features	Analysis of the nucleotide composition in the three most 5' codons of the CDR3 of immunoglobulin heavy chain
Data source location	Lund, Sweden
Data accessibility	Analyzed data are available within this article. Raw data generated by us [2,3] are available in the European Nucleotide Archive, with accession number PRJEB18926 (www.ebi.ac.uk/ena/data/view/PRJEB18926). Additional raw data [4] also analyzed as part of this study are available from the European Nucleotide Archive (accession numbers SRX709625, SRX709626 and SRX709627).

Value of the data

- These data are useful for further development of processes used to infer the immunoglobulin gene repertoire of an individual, and for interpretation of the results of such analyses.
- These data are useful for further development of processes used to infer new germline gene sequences.
- These data are useful to investigators of antibody repertoire as they suggest avenues to identify the existence of, to this date, unrecognized alleles of immunoglobulin germline genes.
- These data are useful for interpretation of sequence diversity in the ascending strand of CDR3 of naïve and antigen-specific immune repertoires.

1. Data

This article present data of nucleotide composition in antibody heavy transcripts originating in 47 different immunoglobulin heavy chain variable (IGHV) germline genes/alleles (Fig. 1) [1]. The data is limited to the three most 5' codons (codon 105–107, according to IMGT numbering [5]) that encode the sequence of the complementary determining region 3 (CDR3). For transcripts originating in germline genes that encodes also the first base of the fourth codon of CDR3 (codon 108), the nucleotide composition at this position is also presented. The location of, and polar interactions potentially mediated by, the side chain of amino acid residue 107 in a set of antibody structures is shown (Fig. 2). The number of subjects that contributed sequence information for the generation of Fig. 1 is summarized in Table 1.

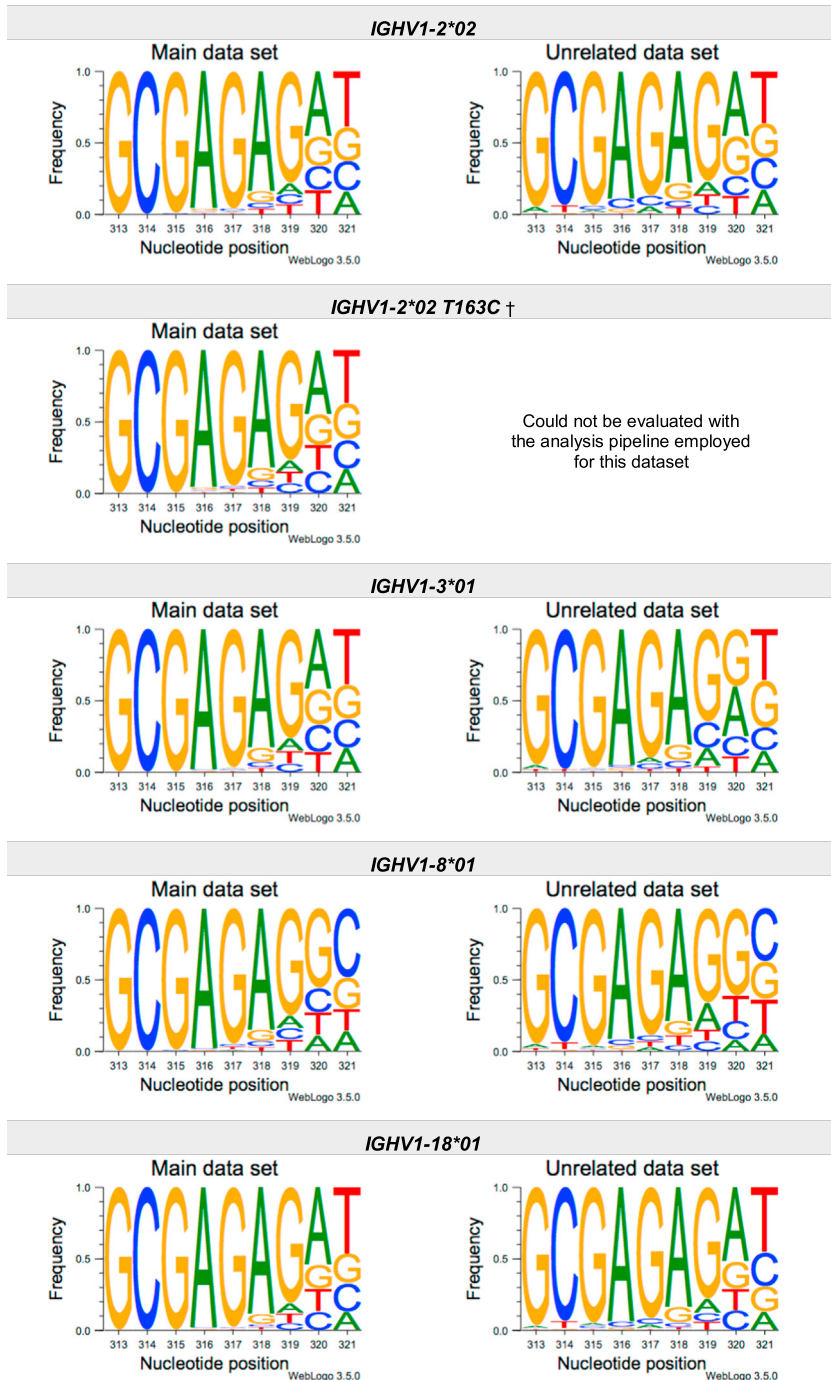


Fig. 1. Distribution of bases in the first three codons of 47 genes/alleles encoding CDR3 of antibody heavy chains in the main examined data set [2,3] and in an unrelated data set [4]. For the latter data set, only transcripts that were exclusively inferred to one germline gene/allele were used. *IGHV1-2*02 T163C* (†) would be inferred as either *IGHV1-2*02* or *IGHV1-2*05*, and could thus not be evaluated with the used method. *IGHV3-30*03* (‡) and *IGHV3-30*18* are identical in the part of the sequence that is inferred by the used approach, but differ in codon 106 where they carry an *AGA* and an *AAA* trimer, respectively. Hence, transcripts that herein have been inferred as derived from *IGHV3-30*03* more likely originates from *IGHV3-30*18*, since they predominantly incorporated an *AAA* trimer in codon 106. The number of subjects used for analysis varies between 3 and 6 in the main data set and 0 and 3 in the unrelated data set (Table 1).

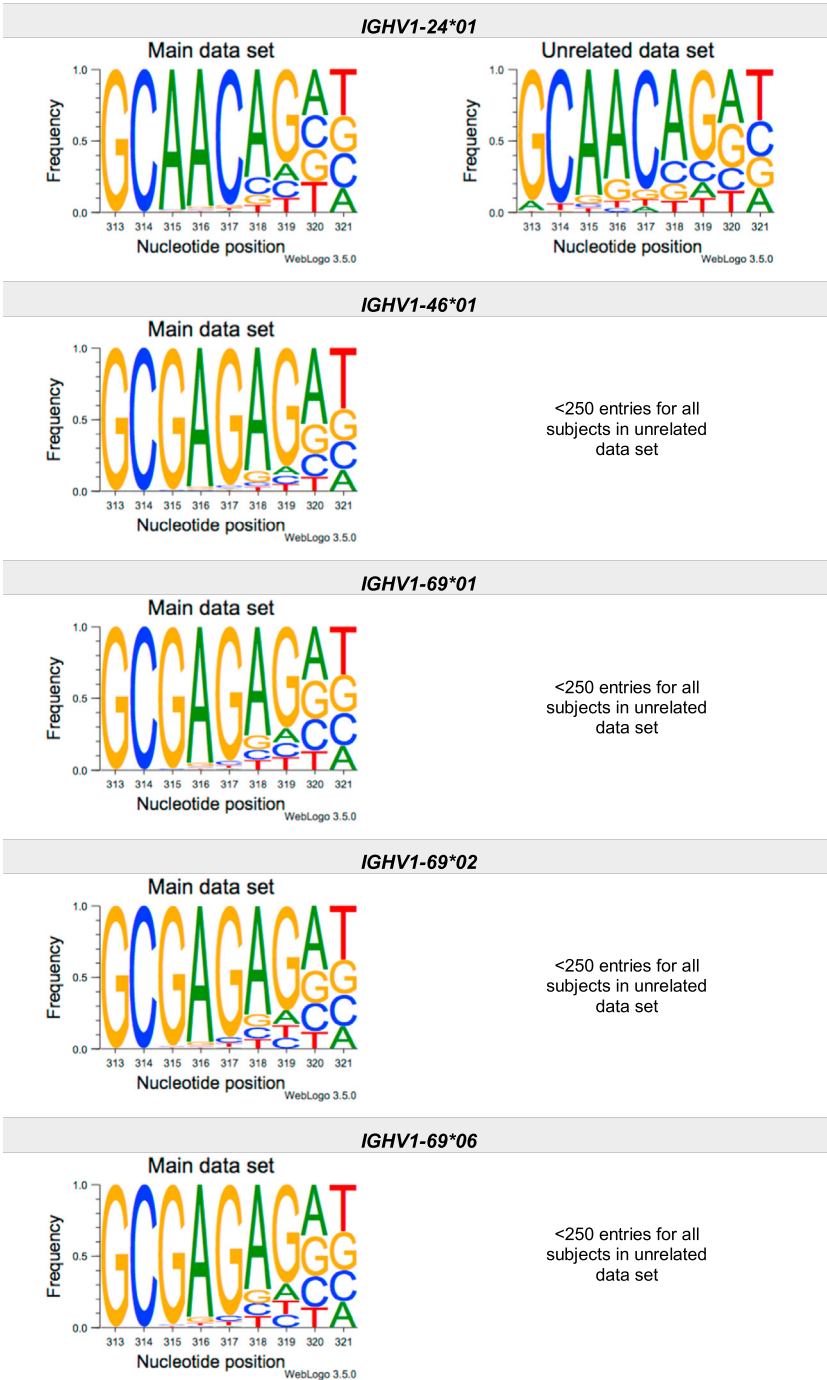


Fig. 1. (continued)

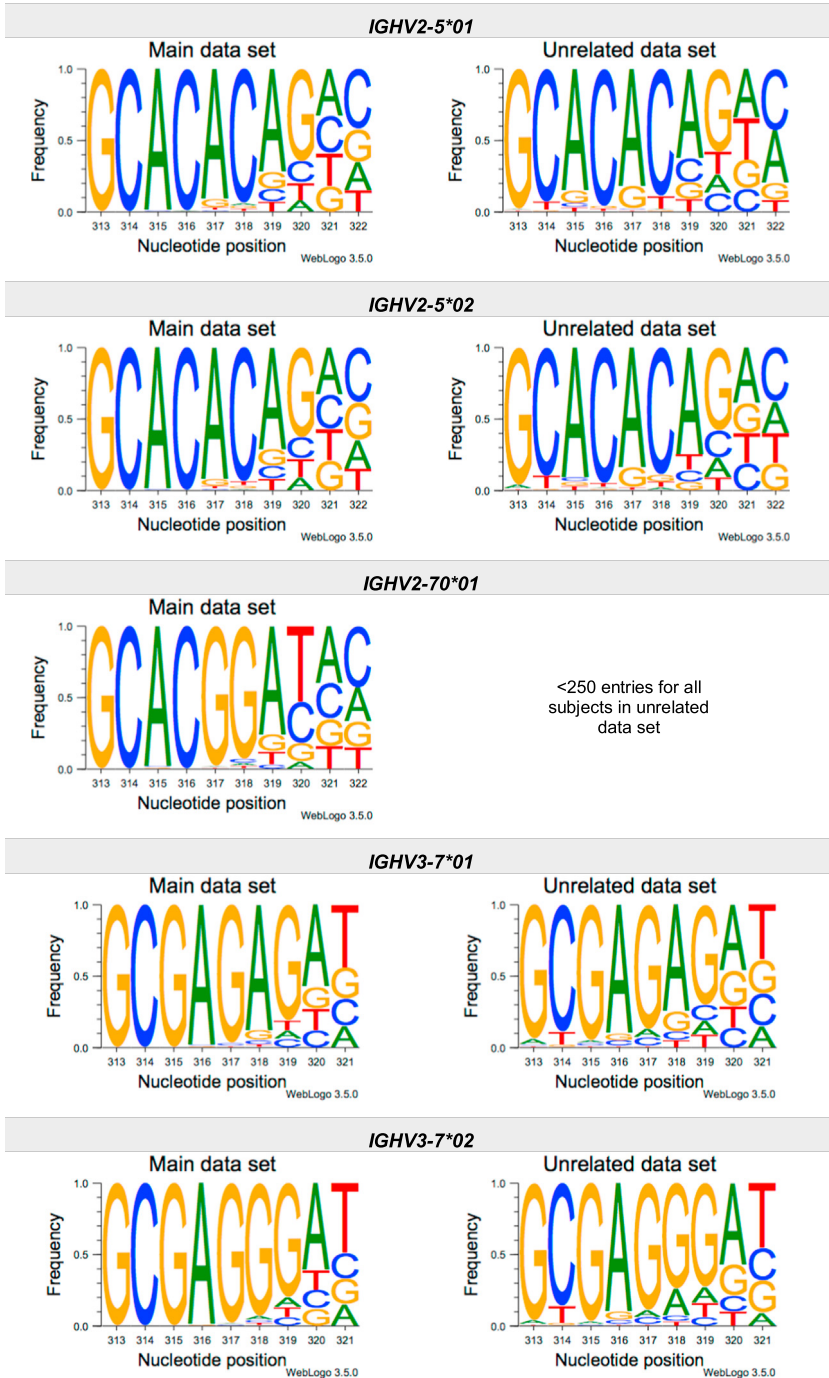


Fig. 1. (continued)

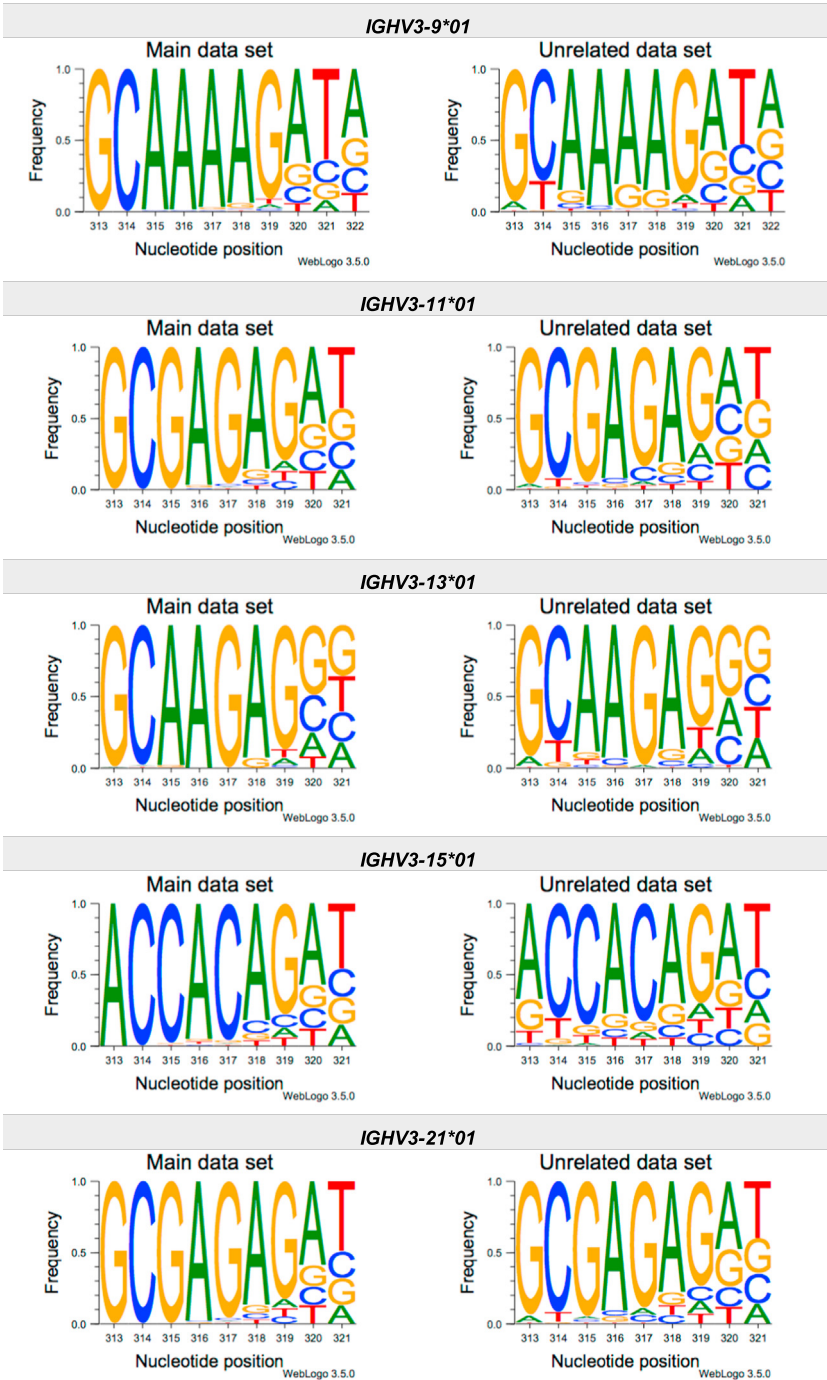


Fig. 1. (continued)

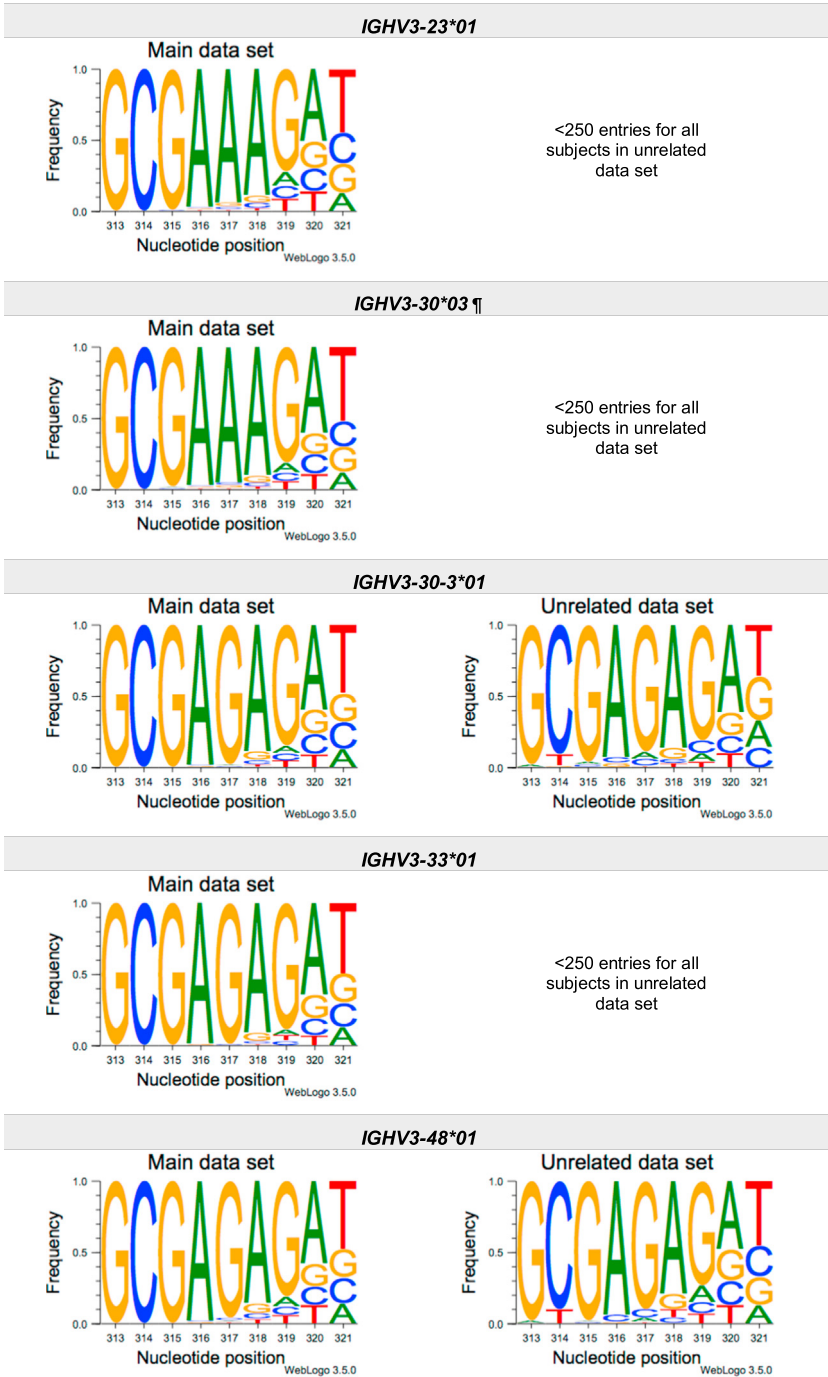


Fig. 1. (continued)

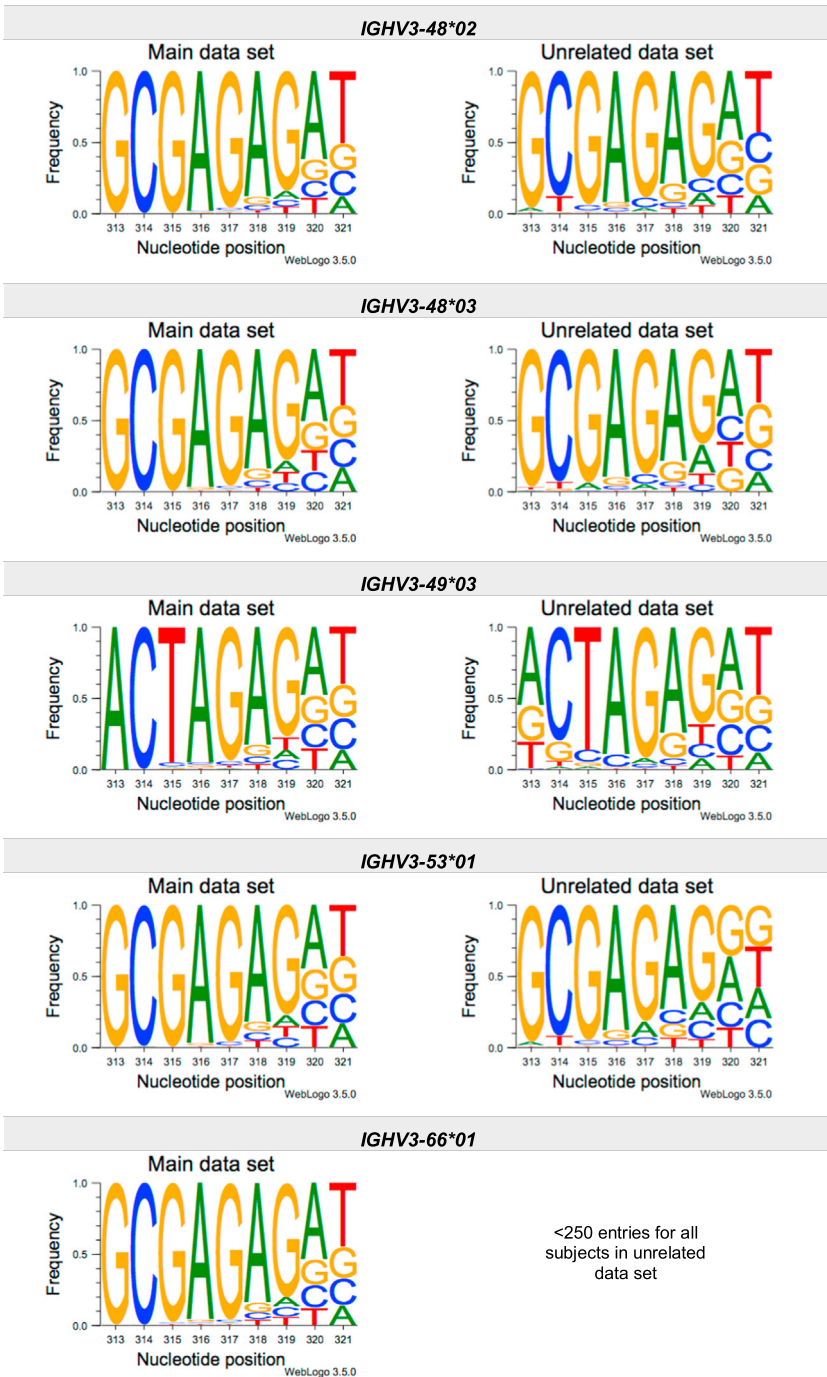


Fig. 1. (continued)

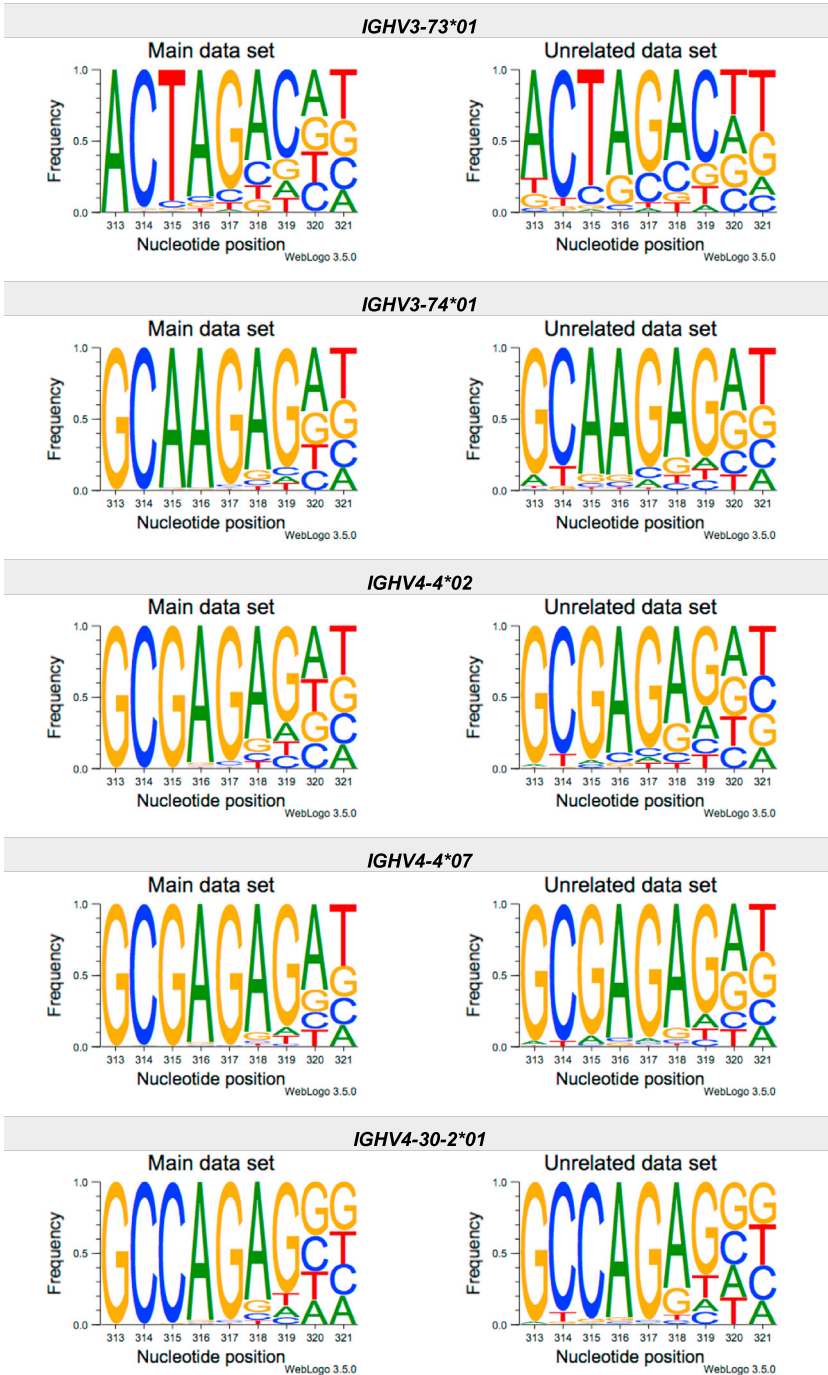


Fig. 1. (continued)

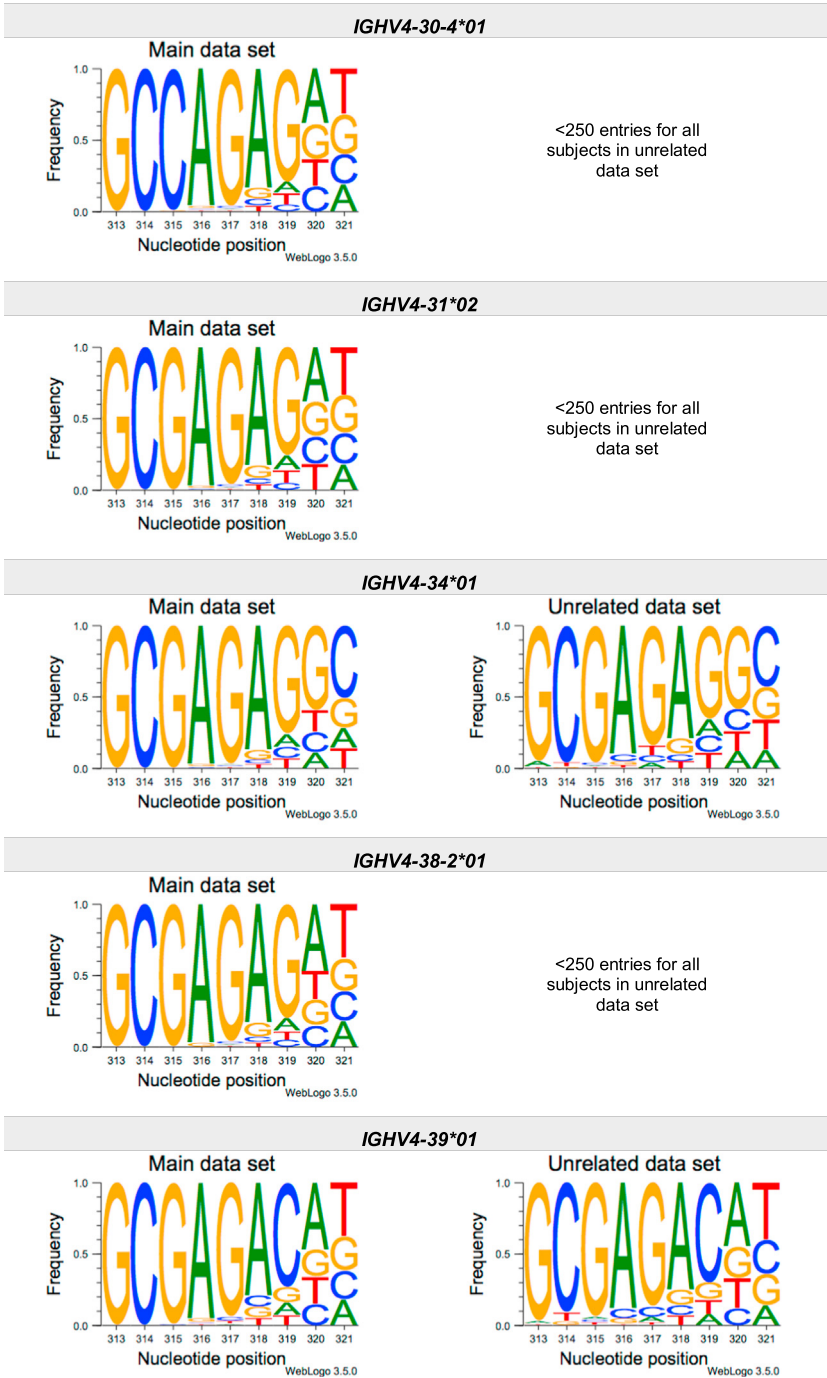


Fig. 1. (continued)

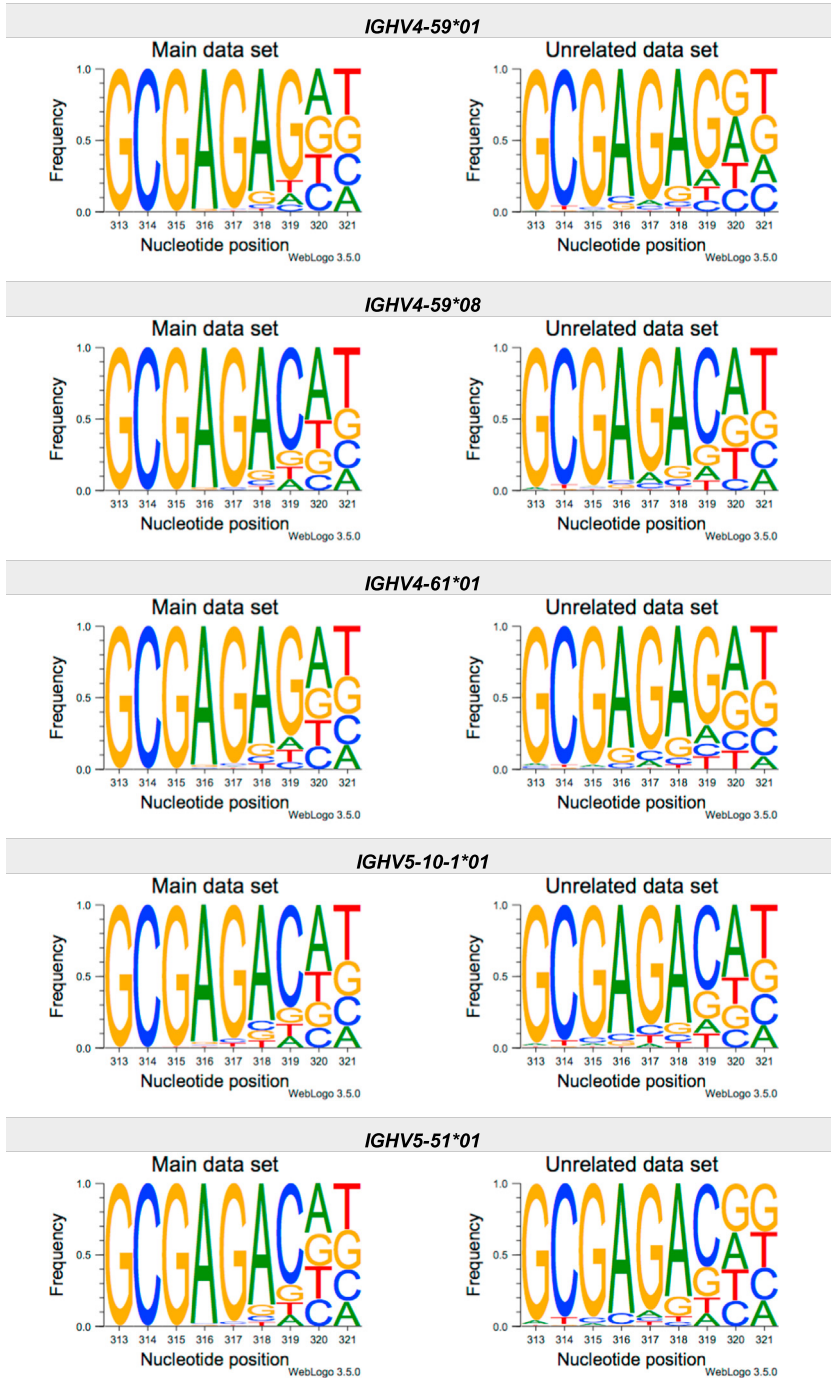


Fig. 1. (continued)

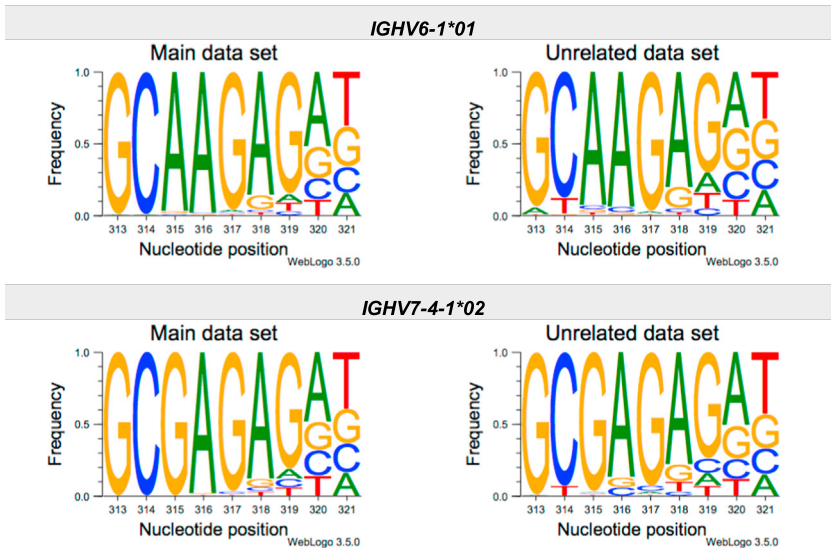


Fig. 1. (continued)

2. Experimental design, materials and methods

2.1. Sample collection, library construction and sequencing

Peripheral blood and bone marrow samples of six allergic subjects were collected (approved by the regional ethical review board at Lund University), and used to construct libraries of antibody H chain V domains, as previously described [2]. In brief, isolated mononuclear cells were divided into duplicate samples from which RNA was extracted. Subsequently, cDNA was produced from the RNA and amplified with Biomed2 primers [6] targeting sequences encoding the constant domain (isotype-specifically) and the first framework region of antibody H chains, respectively. The products were barcoded and subsequently sequenced at National Genomics Infrastructure (SciLifeLab, Stockholm, Sweden), using MiSeq technology (Illumina, Inc. San Diego, CA, USA) and a paired-end setting (2×300 bp) [2].

2.2. Processing of sequencing data

FASTQ raw data files (available at the European Nucleotide Archive with accession number PRJEB18926) generated in our laboratory, constituted the main data set. They were processed as previously described [2]. The sequences were filtered, trimmed, paired, assembled and divided in isotype specific FASTA files using pRESTO 0.4.4 [7], and the isotype annotation were confirmed through evaluation of the presence of isotype-specific sequences. Any sequences lacking such were discarded [2]. Germline genes were inferred for IgM encoding sequences using IgDiscover [8], as previously described [9]. Germline gene libraries retrieved from IMGT [10] were used, but with the IGHV library adjusted to cover no more than codon 25–105. Finally, sequences were filtered so that only those that encoded at least eight amino acids in the CDR3, that covered at least 99% of the inferred IGHV germline gene and that lacked errors compared with the inferred IGHV gene were further analysed.

Another, unrelated set of raw sequence data was downloaded from the European Nucleotide Archive (accession numbers SRX709625, SRX709626 and SRX709627) [4], and prepared for analysis. The data set contained transcripts from peripheral blood memory B cells encoding paired H chain V domain and light chain V domain in three subjects, and were generally processed as described above,

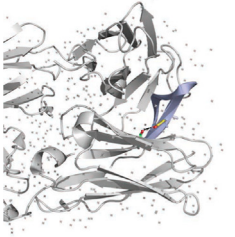
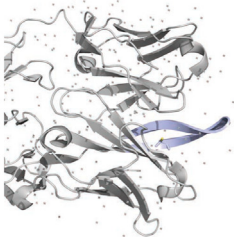
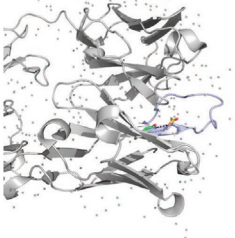

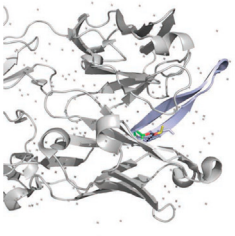
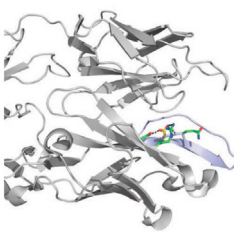
IGHV germline gene	Germline-encoded residue in position 40	Structure and potential polar interactions made by side chain of H chain V domain residue 107	
IGHV4-59	Ser		
		PDB: 3MLY Asp107 O δ 1->Ser40 O γ 2.7 Å Asp107 O δ 2->Ser40 O γ 3.2 Å	PDB: 5FHB no hydrogen bonds by side chain of residue 107 (alanine)
IGHV3-21	Asn		
		PDB: 3H42 Asp107 O δ 1->Asn40 N δ 2 2.8 Å	PDB: 3INU Glu107 O ϵ 1->Arg38 N ϵ 3.1 Å Glu107 O ϵ 1->Asn40 N δ 2 2.9 Å Glu107 O ϵ 2->Arg38 N η 2 2.7 Å Glu107 O ϵ 2->Asp113 N 3.2 Å
IGHV3-15	Asn or Ser		
		PDB: 4G6F Thr107 O γ 1->Thr40 O γ 1 2.7 Å Thr107 O γ 1->Ala105 O 2.9 Å	PDB: 4UV4 Asp107 O δ 1->Gln108 N 3.0 Å Asp107 O δ 1->Arg55 N η 2 3.5 Å Asp107 O δ 2->Asn40 N δ 2 2.7 Å

Fig. 2. Examples of position of and potential polar interactions made by the side chain of H chain V domain residue 107. Carbon atoms of the side chain of residue 107 are highlighted in yellow and those of the side chain of other residues are highlighted in green. The backbone of H chain CDR3 is shown in light blue.

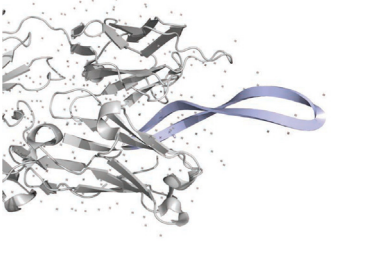
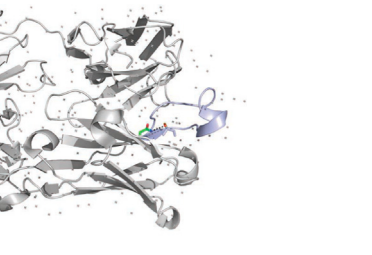
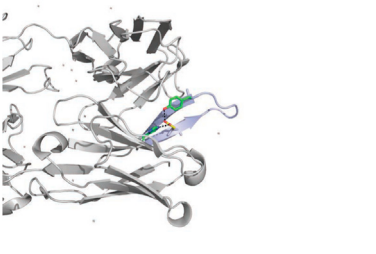
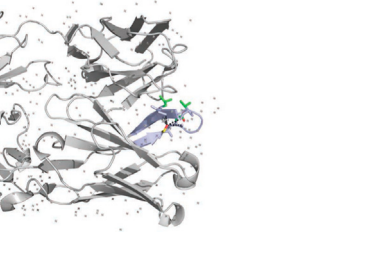
IGHV germline gene	Germline-encoded residue in position 40	Structure and potential polar interactions made by side chain of H chain V domain residue 107	
IGHV1-8	Asn		
		PDB: 3U1S no hydrogen bonds by side chain of residue 107 (glycine)	PDB: 3X3F Ser107 O γ ->Asn40 N δ 2 3.1 Å
IGHV3-13	His	Structure and potential polar interactions made by side chain of H chain V domain residue 107	
IGHV3-13	His		
		PDB: 4U6V Asp107 O δ 2->His40 N ϵ 2 2.7 Å Asp107 O δ 1->Tyr112 OH 2.7 Å	PDB: 5FHA Ser 107O γ ->Asp108 O 3.6 Å Ser 107O γ ->Gly113 N 3.3 Å Ser 107O γ ->Val111 O 3.2 Å Ser 107O γ ->Leu114 N 3.4 Å Ser 107O γ ->Leu 114 C α 3.5 Å

Fig. 2. (continued)

but using pRESTO 0.5.4 [7]. As the isotype encoded by the transcripts was unknown, no dividing of sequences with regard to isotype were performed. Consequently, IgDiscover [8], which mostly are designed for IgM analysis, could not be used for germline genes inference. Instead, duplicate sequences were removed using the pRESTO 0.5.4 CollapseSeq tool [7] and IGHV gene were subsequently inferred using IMGT HighV-QUEST [11]. For further analysis, only sequences inferred as productive to one single allele of an IGHV gene and that had at least eight amino acids in the CDR3 were used.

2.3. Analysis of nucleotide composition in CDR3 codons encoded by IGHV germline gene

The nucleotide composition of the first three codons of the CDR3 region, which are encoded by the IGHV gene, were analysed for each donor of both the main and the unrelated data set. In total, transcripts originating in 47 different alleles of IGHV genes were studied, each of them having at least 500 transcripts in at least three of the donors of the main data set. Mean frequency of nucleotide bases at each examined position were calculated for both data sets separately. For the main data set, only values from subjects with at least 500 transcripts originating in a certain allele of an IGHV gene were considered. For the unrelated data set, this limit was set to 250 transcripts. The number of subjects for which these conditions were fulfilled is summarized for each allele in Table 1. The mean

Table 1

Number of subjects in which the number of transcript entries exceeded the cut-off value.

	Main Data Set	Unrelated Data Set
	Number of subjects with > 500 entries	Number of subjects with > 250 entries
<i>IGHV1-2*02</i>	5	3
<i>IGHV1-2*02 T163C</i>	3	Not evaluated
<i>IGHV1-3*01</i>	4	2
<i>IGHV1-8*01</i>	6	2
<i>IGHV1-18*01</i>	6	3
<i>IGHV1-24*01</i>	4	3
<i>IGHV1-46*01</i>	6	0
<i>IGHV1-69*01</i>	6	0
<i>IGHV1-69*02</i>	3	0
<i>IGHV1-69*06</i>	3	0
<i>IGHV2-5*01</i>	3	1
<i>IGHV2-5*02</i>	6	1
<i>IGHV2-70*01</i>	3	0
<i>IGHV3-7*01</i>	6	3
<i>IGHV3-7*02</i>	3	2
<i>IGHV3-9*01</i>	6	3
<i>IGHV3-11*01</i>	6	2
<i>IGHV3-13*01</i>	3	1
<i>IGHV3-15*01</i>	6	3
<i>IGHV3-21*01</i>	6	3
<i>IGHV3-23*01</i>	6	0
<i>IGHV3-30*03</i>	6	0
<i>IGHV3-30-3*01</i>	5	1
<i>IGHV3-33*01</i>	6	0
<i>IGHV3-48*01</i>	4	2
<i>IGHV3-48*02</i>	4	2
<i>IGHV3-48*03</i>	3	1
<i>IGHV3-49*03</i>	5	2
<i>IGHV3-53*01</i>	5	2
<i>IGHV3-66*01</i>	3	0
<i>IGHV3-73*01</i>	3	1
<i>IGHV3-74*01</i>	6	3
<i>IGHV4-4*02</i>	6	3
<i>IGHV4-4*07</i>	4	2
<i>IGHV4-30-2*01</i>	4	2
<i>IGHV4-30-4*01</i>	6	0
<i>IGHV4-31*02</i>	6	0
<i>IGHV4-34*01</i>	6	3
<i>IGHV4-38-2*01</i>	3	0
<i>IGHV4-39*01</i>	5	3
<i>IGHV4-59*01</i>	6	3
<i>IGHV4-59*08</i>	3	1
<i>IGHV4-61*01</i>	6	2
<i>IGHV5-10-1*01</i>	3	1
<i>IGHV5-51*01</i>	6	2
<i>IGHV6-1*01</i>	5	2
<i>IGHV7-4-1*02</i>	3	1

The cut-off value was set to 500 entries for the main data set [2,3] and to 250 entries for the unrelated data set [4]. For the latter, only transcripts that were exclusively inferred to a single germline allele were used.

frequency values were used to construct the illustrations presented in Fig. 1, using WebLogo 3.5.0 [12].

Most of the studied IGHV genes may contribute to nucleotides of the first three codons that encode the CDR3 (codon 105–107, as defined by the IMGT numbering system [5]). Hence, these are the codons for which the nucleotide composition generally was analysed. Four of the germline genes/alleles (*IGHV2-5*01*, *IGHV2-5*02*, *IGHV2-70*01*, and *IGHV3-9*01*) may however also encode the first

base of codon 108. Thereby, the nucleotide composition was analysed also at this position for transcripts originating in any of these four germline genes/alleles.

2.4. Protein structures

Example structures of antibodies encoded by genes with a particular germline gene origin were identified using IMGT/3Dstructure-DB [13]. Protein structure coordinates were downloaded from the Protein Data Bank (<https://www.rcsb.org>). The structures were visualized using MacPyMol 1.8.0.6 (The PyMOL Molecular Graphics System, Schrödinger, LLC).

Acknowledgements

This study was supported by the Swedish Research Council (Grant no. 2016-01720). We acknowledge support from Science for Life Laboratory, the Knut and Alice Wallenberg Foundation, the National Genomics Infrastructure (NGI) funded by the Swedish Research Council and Uppsala Multidisciplinary Center for Advanced Computational Science for providing assistance with NGS and access to the UPPMAX computational infrastructure.

Transparency document. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.dib.2018.04.125>.

References

- [1] L. Thörnqvist, M. Ohlin, The functional 3'-end of immunoglobulin heavy chain variable (IGHV) genes, *Mol. Immunol.* 96 (2018) 61–68.
- [2] 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.
- [3] 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.
- [4] B.J. DeKosky, T. Kojima, A. Rodin, W. Charab, G.C. Ippolito, A.D. Ellington, G. Georgiou, In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire, *Nat. Med.* 21 (2015) 86–91.
- [5] 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.
- [6] J.J. van Dongen, A.W. Langerak, M. Bruggemann, P.A. Evans, M. Hummel, F.L. Lavender, E. Delabesse, F. Davi, E. Schuurung, R. Garcia-Sanz, J.H. van Krieken, J. Droese, D. Gonzalez, C. Bastard, H.E. White, M. Spaargaren, M. Gonzalez, A. Parreira, J. L. Smith, G.J. Morgan, M. Kneba, E.A. Macintyre, 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 (2003) 2257–2317.
- [7] 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.
- [8] M.M. Corcoran, G.E. Phad, N. Vazquez Bernat, 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.
- [9] U. Kirik, H. Persson, F. Levander, L. Greiff, M. Ohlin, Antibody heavy chain variable domains of different germline gene origins diversify through different paths, *Front. Immunol.* 8 (2017) 1433.
- [10] M.P. Lefranc, IMGT, the international ImMunoGeneTics information system, *Cold Spring Harb. Protoc.* 2011 (2011) 595–603.
- [11] E. Alamyar, P. Duroux, M.P. Lefranc, V. Giudicelli, 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 (2012) 569–604.
- [12] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, *Genome Res.* 14 (2004) 1188–1190.
- [13] F. Ehrenmann, M.P. Lefranc, IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA), *Cold Spring Harb. Protoc.* 2011 (2011) 750–761.