



Antibody Heavy Chain Variable Domains of Different Germline Gene Origins Diversify through Different Paths

Ufuk Kirik^{1†}, Helena Persson², Fredrik Levander^{1,3}, Lennart Greiff^{4,5} and Mats Ohlin^{1,6,7*}

¹Department of Immunotechnology, Lund University, Lund, Sweden, ²Science for Life Laboratory, Drug Discovery and Development Platform, School of Biotechnology, KTH Royal Institute of Technology, Stockholm, Sweden,

³National Bioinformatics Infrastructure Sweden (NBIS), Science for Life Laboratory, Department of Immunotechnology, Lund University, Lund, Sweden, ⁴Department of Clinical Sciences, Lund University, Lund, Sweden, ⁵Department of Otorhinolaryngology, Head and Neck Surgery, Skåne University Hospital, Lund, Sweden, ⁶Science for Life Laboratory, Drug Discovery and Development Platform, Human Antibody Therapeutics, Lund University, Lund, Sweden,

⁷U-READ, Lund School of Technology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:

Jacob Glanville,
Distributed Bio, Inc.,
United States

Reviewed by:

Claude-Agnes Reynaud,
Institut national de la santé et de la
recherche médicale, France
Michael Zemlin,
Universitätsklinikum
des Saarlandes, Germany

*Correspondence:

Mats Ohlin
mats.ohlin@immun.lth.se

†Present address:

Ufuk Kirik,
Disease Systems Biology Program,
University of Copenhagen,
Copenhagen, Denmark

Specialty section:

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

Received: 28 June 2017

Accepted: 16 October 2017

Published: 13 November 2017

Citation:

Kirik U, Persson H, Levander F,
Greiff L and Ohlin M (2017) Antibody
Heavy Chain Variable Domains of
Different Germline Gene Origins
Diversify through Different Paths.
Front. Immunol. 8:1433.
doi: 10.3389/fimmu.2017.01433

B cells produce antibodies, key effector molecules in health and disease. They mature their properties, including their affinity for antigen, through hypermutation events; processes that involve, e.g., base substitution, codon insertion and deletion, often in association with an isotype switch. Investigations of antibody evolution define modes whereby particular antibody responses are able to form, and such studies provide insight important for instance for development of efficient vaccines. Antibody evolution is also used *in vitro* for the design of antibodies with improved properties. To better understand the basic concepts of antibody evolution, we analyzed the mutational paths, both in terms of amino acid substitution and insertions and deletions, taken by antibodies of the IgG isotype. The analysis focused on the evolution of the heavy chain variable domain of sets of antibodies, each with an origin in 1 of 11 different germline genes representing six human heavy chain germline gene subgroups. Investigated genes were isolated from cells of human bone marrow, a major site of antibody production, and characterized by next-generation sequencing and an in-house bioinformatics pipeline. Apart from substitutions within the complementarity determining regions, multiple framework residues including those in protein cores were targets of extensive diversification. Diversity, both in terms of substitutions, and insertions and deletions, in antibodies is focused to different positions in the sequence in a germline gene-unique manner. Altogether, our findings create a framework for understanding patterns of evolution of antibodies from defined germline genes.

Keywords: antibody germline gene, antibody sequence, somatic hypermutation, immunoglobulin, insertion and deletion, substitution

INTRODUCTION

Antibodies, central components of humoral immunity, are crucial to our survival. The immune system allows antibodies to evolve in efforts to enhance their ability to mediate protection against disease. The biochemistry and mechanism of this complex evolution process has been extensively studied at the molecular level (1). Technological advances in sequencing and single cell analysis

Abbreviations: BM, bone marrow; CDR, complementarity determining region; FR, framework region; H, heavy; V, variable; VH, heavy chain variable; VL, light chain variable.

technology have recently allowed us to, at great depth, study antibody sequences as they develop *in vivo* (2). Indeed, various sequencing strategies and bioinformatics pipelines have been generated to allow such analysis (3, 4).

Studies of the development of humoral immune responses require knowledge of the repertoire of genes that are available in the genome. Such information allows us to properly analyze germline gene rearrangement events and hypermutation, as exemplified by extensive studies of the response against the envelope protein of HIV-1 (5, 6). Databases and associated analysis tools, like IMGT/IMGT V-QUEST/IMGT HighV-QUEST (7), have consequently been built to allow efficient analysis of antibody-encoding sequences, their genetic origin, and their evolution. Common concepts, like standardized framework and complementarity determining regions (FR and CDR, respectively), the latter of which is considered to represent the antigen-contacting part of the antibody, are commonly used in such analysis. However, numerous definitions of these regions exist in parallel (8–13), highlighting the difficulties associated with the establishment of a clear-cut definition of these regions. We hypothesized that a thorough understanding of the ways through which antibodies derived from different germline genes evolve as a consequence of somatic mutation processes will aid the establishment of such definitions. Such understanding will also aid a proper mutational analysis of clones that populate immune responses.

In this study, we have focused our attention to human IgG encoded by unsorted cells in bone marrow (BM) (14), a major site of antibody production, to define how evolution proceeds in antibody heavy (H) chains derived from 11 commonly used germline genes. The advent of high throughput next-generation sequencing methodology, and its application to studies of antibody gene sequences (2), allowed us to decipher the mutability of antibodies of different origins in ways not possible in the recent past. The analysis was highly enhanced by germline gene inference technology (15) that defined the germline gene/allele repertoires of the donors under study, thereby minimizing errors originating from inappropriate gene assignment. We now demonstrate how antibodies of different germline gene origins evolve residues and introduce insertions and deletions into CDRs and FRs. This information has implications for our understanding and interpretation of human immune responses.

MATERIALS AND METHODS

Antibody-Encoding Transcriptomes

Antibody-encoding transcripts were isolated from unsorted cells of BM of six subjects diagnosed with allergic rhinitis, examined out of season of most seasonal pollen allergens (14). Transcripts encoding H chain variable (V) domains of different antibody isotypes were individually amplified by PCR, barcoded, and sequenced using Illumina MiSeq technology (14). Sequences are available from the European Nucleotide Archive accession number PRJEB18926. Reads were processed by pRESTO (16) and transcripts encoding each isotype were analyzed by IMGT HighV-QUEST (17) as previously described (14). A summary of the number of sequences at different stages

of the analysis pipeline is provided in Supplementary Table EIV in Levin et al. (14).

Germline Gene Repertoire

The germline gene repertoire of the donors have been inferred using IgDiscover (15) using the IgM-encoding transcriptomes of the donors' BM, and has, when possible, been quality controlled by haplotype-based analysis (18, 19). Eleven commonly expressed germline genes (IGHV1-8, IGHV1-18, IGHV2-5, IGHV3-7, IGHV3-11, IGHV3-21, IGHV3-23, IGHV4-39, IGHV4-59, IGHV5-51, and IGHV6-1) (Table 1 in Supplementary Material) mostly encoded by a single or a few highly related alleles, representing six germline gene subgroups, were selected for further analysis (Table 1). Full-length (all codons from 1 to 105) sequences of functional germline genes were downloaded from the IMGT database¹ (release 201718-0). Sequence similarity between these genes/alleles was determined after alignment using the ClustalW algorithm (20) as implemented in MacVector 15.5.0 (MacVector, Inc., Apex, NC, USA). Hot-spots for mutation of individual germline genes/alleles were identified by analysis through use of IMGT V-QUEST (21).

Analysis of Diversification of Residues Encoded Proteins

Data defining productive sequences with an origin in investigated germline genes were retrieved following IMGT HighV-QUEST-based analysis (17). Only sequences not showing evidence of

¹<http://www.imgt.org>.

TABLE 1 | Examples of germline gene allele repertoire of the six lymphocyte donors as assessed using the IgM-encoding transcriptome.

Germline	Allele composition of each donor ^a					
	Donor 1	Donor 2	Donor 3	Donor 4	Donor 5	Donor 6
IGHV1-2 ^b	*02, *p06	*02, *04	*02	*02	*04, *p06	*02, *p06
IGHV1-8	*01	*01	*01	*01	*01	*01
IGHV1-18	*01	*01	*01	*01	*01	*01
IGHV2-5	*02	*01, *02	*02	*01, *02	*02	*01, *02
IGHV3-7 ^c	*01, *02	*01	*01, *02	*01, *02	*01, *03	*01
IGHV3-11	*01, *03	*01	*01, *03	*01	*01, *06	*01
IGHV3-21	*01	*01	*01	*01	*01	*01
IGHV3-23	*01	*01	*01	*01	*01	*01
IGHV4-39 ^d	*01	*01	*01, *07	*01	*07	*01
IGHV4-59 ^e	*01, *08	*01	*01, *08	*01	*01	*01, *08
IGHV5-51	*01	*01	*01	*01	*01	*01
IGHV6-1	*01	*01	*01	*01	*01	*01

Sequence sets not used in the analysis due to an inability to precisely define the germline origin of each protein sequence in at least three donors are showed with a gray background.

^aSeveral alleles with identical nucleotide sequence in the assessed parts of the gene may exist, in which case only the lowest allele number is shown.

^bIGHV1-2*p06 is a sequence variant (T163C) of IGHV1-2*02 that is not present in the IMGT germline gene database.

^cThere is no difference in amino acid sequence in the analyzed part of the sequence of IGHV3-7*01, IGHV3-7*02, and IGHV3-7*03.

^dThere is no difference in amino acid sequence in the analyzed part of the sequence of IGHV4-39*01 and IGHV4-39*07.

^eThere is no difference in amino acid sequence the analyzed part of the sequence of IGHV4-59*01 and IGHV4-59*08.

insertions and deletions were scored with respect to presence of substitutions. The frequency of each amino acid was calculated for each position within the range of codons from 27 to 104. Such analysis was performed only for donors that were homozygous for a given allele or heterozygous for alleles that encode identical protein products from the analyzed part of their unmutated sequence. Sequence variability was calculated as the number of amino acids encoded by more than 1% of all reads, divided by the fraction of reads encoding the most common residue. For comparison with real protein structures, examples of structures with an origin in IGHV1-18 (PDB: 3SDY) (22), IGHV1-8 (PDB 3X3G and 3U1S) (23, 24), IGHV2-5 (PDB: 3QRG), IGHV3 subgroup (PDB: 2R56 and 3FZU) (25, 26), IGHV4-39 (PDB: 5C6T) (27), IGHV4-59 (PDB: 3HI1) (28), and IGHV5-51 (PDB: 4BUH) (29) were identified using the IMGT/3Dstructure-DB web interface (30), and coordinates were downloaded from RCSB Protein Data Bank.² Structures were visualized using MacPyMOL v1.8.0.6. Sequence numbering and CDR and FR definitions are those defined by the IMGT nomenclature (13).

Insertions and Deletions

Somatic hypermutation not only involves base substitutions but also insertions and deletions in the coding sequence (31, 32). The positions of such productive (in-frame) modifications were scored in each read based on IMGT HighV-QUEST analysis (17).

Evidence of Selection

The 10 most highly expressed rearrangements (based on a defined CDRH3-encoding sequence) of six germline genes, IGHV1-18, IGHV2-5, IGHV3-23, IGHV4-39, IGHV5-51, and IGHV6-1, were investigated. Sequences (codons 27–104) were only retrieved from donors homozygous for a given allele to eliminate the risk of incorrect allele assignment that would contribute to perceived sequence diversification. The sequence with the highest number of counts was chosen so as to minimize the impact of random errors introduced by PCR and/or sequencing artifacts. Sequences showing evidence of insertion or deletion were not used, as the analysis pipeline is incompatible with such modes of antibody diversification. The resulting sequences of IGHV1-18 ($n = 60$ sequences), IGHV2-5 ($n = 30$ sequences), IGHV3-23 ($n = 60$ sequences), IGHV4-39 ($n = 50$ sequences), IGHV5-51 ($n = 60$ sequences), and IGHV6-1 ($n = 60$ sequences) were analyzed for evidence of positive and negative selection using Bayesian Estimation of Antigen-Driven Selection in Immunoglobulin Sequences (BASELINE, version 1.3) using a web-based interface.³ Focused selection statistics and the Human S5F somatic hypermutation targeting model were used for this assessment (33, 34).

RESULTS

Individual Germline Repertoires

Bone marrow had been obtained from six individuals with different germline gene repertoires, a material that has previously been used for assessment of antibody repertoires in allergic subjects

out of season of exposure to most environmental allergens (14). This dataset was now reanalyzed to assess antibody diversification. In any such material, allelic diversity will contribute to antibody diversity and will compromise computational analysis of antibody evolution unless correct allele assignment is made. In particular, as differences between alleles often are small, incorrect allele assignment of hypermutated genes cannot be avoided. Prior analysis of these donors' IgM repertoires, repertoires that carry large numbers of unmutated sequences, can be used to ensure proper downstream analysis of IgG repertoires. Such analysis (19) was performed using IgDiscover (15) to define the lymphocyte donors' IGHV germline gene and allele makeup. We furthermore used a haplotype quality-control approach to define the validity of many of the allele calls (18, 19). This approach also allowed us to validate novel alleles not present in the IMGT reference directory used for gene assignments (19). By using donors with defined germlines, we minimized the risk of introducing artifacts in our analysis of the mutational paths taken by antibodies of different germline gene origins. Importantly, this approach identified allele IGHV1-2*p06 (IGHV1-2*02 T163C) in three of six individuals (19) (Table 1), an allele that is not identified by standard IMGT HighV-QUEST or V-QUEST analysis. A failure to identify this allele would incorrectly have enhanced the perceived substitution frequency in one position of this gene by the approach taken in this study. As a consequence of the substantial, but difficult to detect, allelic diversity of IGHV1-2, it was not included in this study. We also made sure that the investigated germline genes were not extensively similar to alleles of other germline genes in the donors' repertoires, as such similarity may, following hypermutation, incorrectly relate products of other genes to the genes under investigation. IGHV3-23 and IGHV3-23D are in this context treated as one gene as they are identical in sequence. Among the other genes, only one investigated allele of IGHV4-59 had a highly similar allele defined by IMGT (>98% nucleotide identity) assigned to another gene location (Figure S1 in Supplementary Material), but this other allele (IGHV4-4*08) was not present in the repertoires of the investigated individuals.

Overall, several germline genes were highly diverse (different alleles used by different donors or presence of different alleles in a given individual), but others were not, some of which were used for this study (Table 1). In all, our analysis focused on a set of commonly used "core" genes (35) utilized in rearrangements obtained from individuals conceived to be homozygous for a given allele or heterozygous for alleles expected to encode identical protein sequences in their unmutated form. As the analysis focused on sequences (germline-encoded protein sequences are shown in Figure 1) from CDRH1 up to the end of FR3, alleles that encode identical protein products in this part of the H chain variable domain could be included in the study. Thus, for the purpose of analysis of selection associated with hypermutation, only sequences derived from individuals homozygous for a given gene sequence encoding CDRH1 to FR3 (codons 27–104) were used.

The IgG Population Encoded by BM Is Highly Somatic Evolved

We chose to study mutations in antibodies of the IgG isotype produced in BM as this is a major site of antibody production.

²<http://www.rcsb.org/pdb/>.

³<http://selection.med.yale.edu/baseline/>.

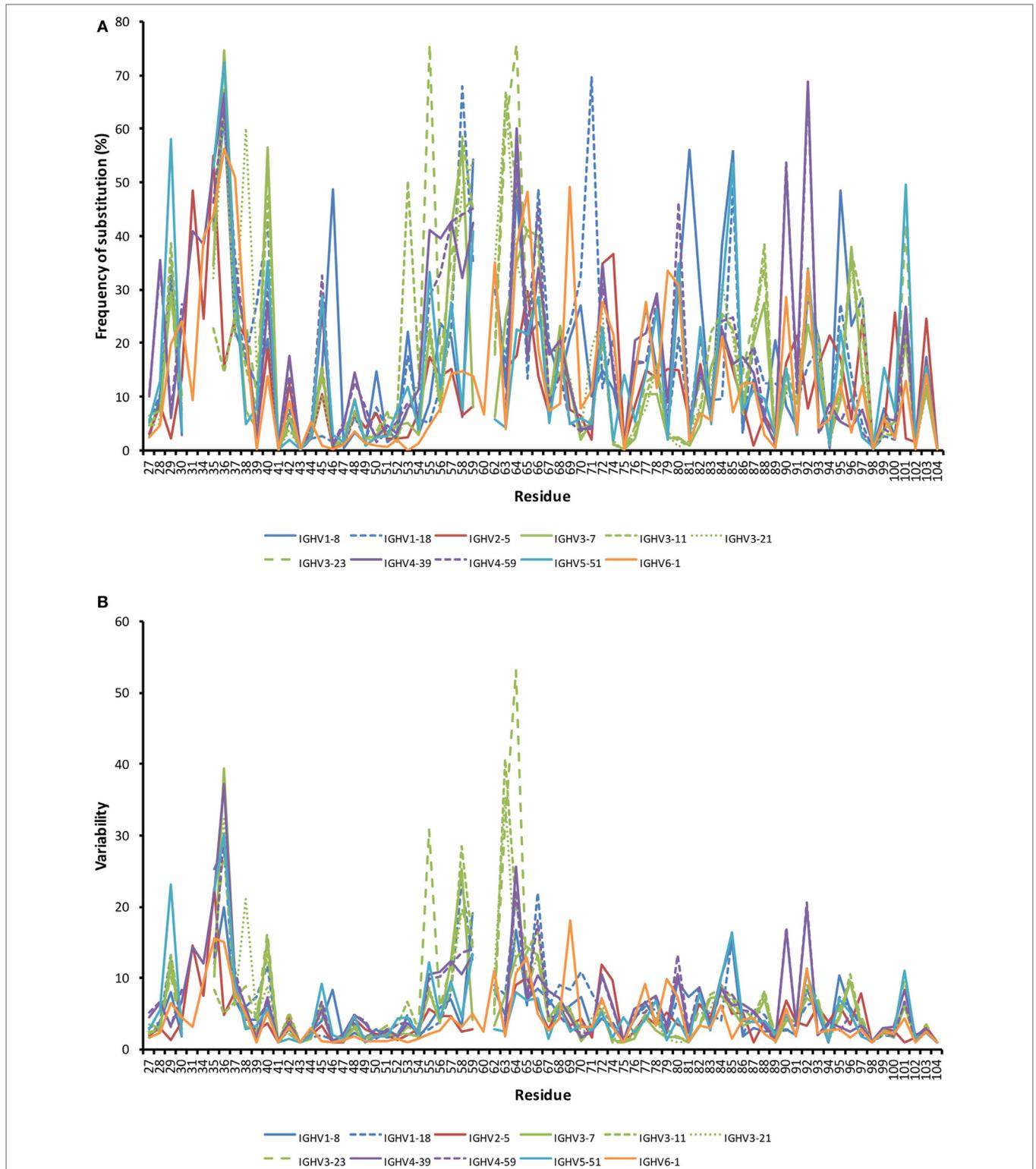
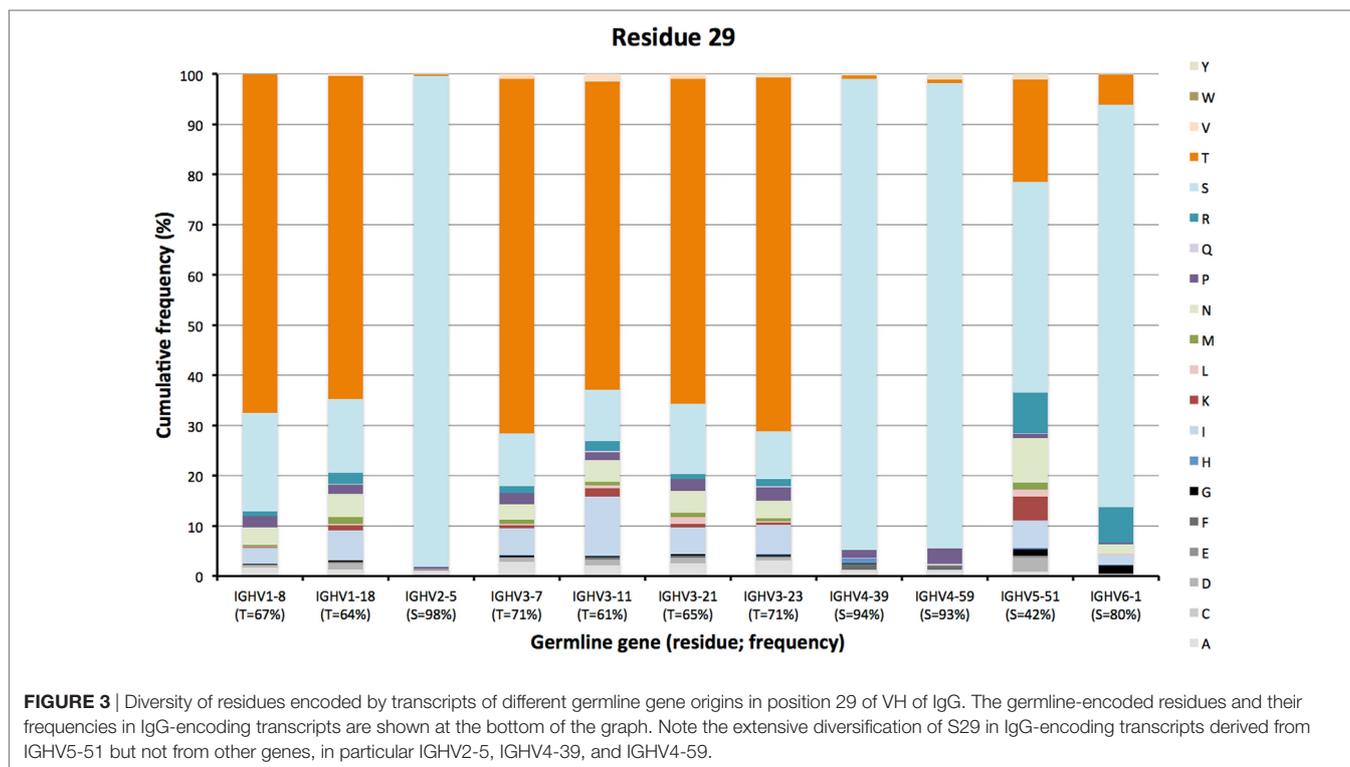


FIGURE 2 | Frequency of substitutions **(A)** and degree of variability **(B)** of residues encoded by transcripts in bone marrow (BM) with an origin in the 11 investigated germline genes. Residues 27–38 code for complementarity determining region (CDR) 1, while residues 56–65 code for CDR2. Variability was calculated as (the number of amino acids encoded by more than 1% of all reads)/(fraction of reads encoding the most common residue). Only sequences showing no evidence of insertions and deletions were included in the analysis. Numerous residues differed substantially between germline genes in the degree of targeting by substitution. Note the substantial degree of substitution also in some residues of framework region (FR). Frequency of substitution and variability for individual genes are shown in Figure S2 in Supplementary Material.



mutation. Indeed, five residues within the region CDRH1-FR3 of VH were substituted in <2.5% of all human IgG transcripts independently of their germline gene origin. These included W41 and C104 in the domain's central core, residues R43 and D98 in the domain's charge cluster, and residue Y102 in the lower core of the domain (Figure 6). Nevertheless, as described below, we observed several other residues belonging to the core regions that are diversified more extensively during somatic antibody evolution.

The lower core is shielded from the upper core by the highly conserved central core (including, e.g., W41 and C104) and a more direct influence by substitutions in this core on the binding site may be limited. Instead, lower core residues may affect the biophysical properties of the domain (36). Several of these residues (of which residues 53, 54, 71, 76, 89, 91, 94, 100, and 102 were assessed in this study) were essentially untouched by somatic diversification, while others, depending on germline gene origin, were diversified. In particular, residue 53 of IGHV3-11 (but not IGHV3-7, IGHV3-21, and IGHV3-23) and residue 71 of IGHV1-18, were prone to diversification (Figures 6 and 7). IGHV1-18 encodes a L at position 71, while other genes of the IGHV1 subgroup encode F in this position. Mutation of this codon in IGHV1-18 introduced F in this position at a high frequency (55%). Mutation of residue 53 incorporated conserved hydrophobic substitutions in place of the germline-encoded residue. In the case of IGHV3-11, V was mainly substituted by L or I. Similarly, mutation of residue 76 largely introduced conservative, hydrophobic substitutions (Figure 7). In summary, some residues of the VH domain's lower core are targets for conservative hypermutation.

A cluster of charged residues is situated close to the lower core of VH (36). It involves residues at positions 43, 51, 75, 95, 97, and 98 (37). Of these, 0–2 residues, mostly residue 95 and 97, were targeted by substitutions at frequencies above 10% (Figure 6). Of note, 48% of all sequences with an origin in IGHV1-8 were targeted by substitution at residue R95 (a codon not associated with a mutational hotspot), while only 9% of sequences with an origin in IGHV3-7 was diversified in this position. Limited diversity (T or K) dominated the diversity introduced at this position. Similarly, substitution at position 97 (a codon not associated with a mutational hotspot in any of the investigated germline genes) in IGHV3 gene subgroup members was dominated by a conservative E → D mutation, while substitution of V97 in IGHV2-5, an unusual, hydrophobic side chain in this cluster, introduced a range of modifications although mainly to A (Figure 8). Altogether, there is room for diversification in the charge cluster in a germline-directed manner, modifications that for instance may affect the biophysical properties of the domain.

Substitutions of residues that belong to the VH/VL interface may affect binding site architecture. We investigated the tendency for substitution in five residues (40, 42, 50, 52, and 103), the surfaces (Figure 6) of which are substantially buried by formation of the VH/VL dimer (11). Among these, residues 50 and 52 were rarely mutated [each below 10% of transcripts, except in the case of residue 50 of IGHV1-18 (15%)] while, in particular residue 40 (G, N, or S) but also residues 42 (I or V) and 103 (Y) were frequently substituted (Figure 6), although mostly in a restricted manner (Figure 5; Figure S4 in Supplementary Material). Residue 40 of some germline gene origins showed substantial levels of substitution (even above 50%)

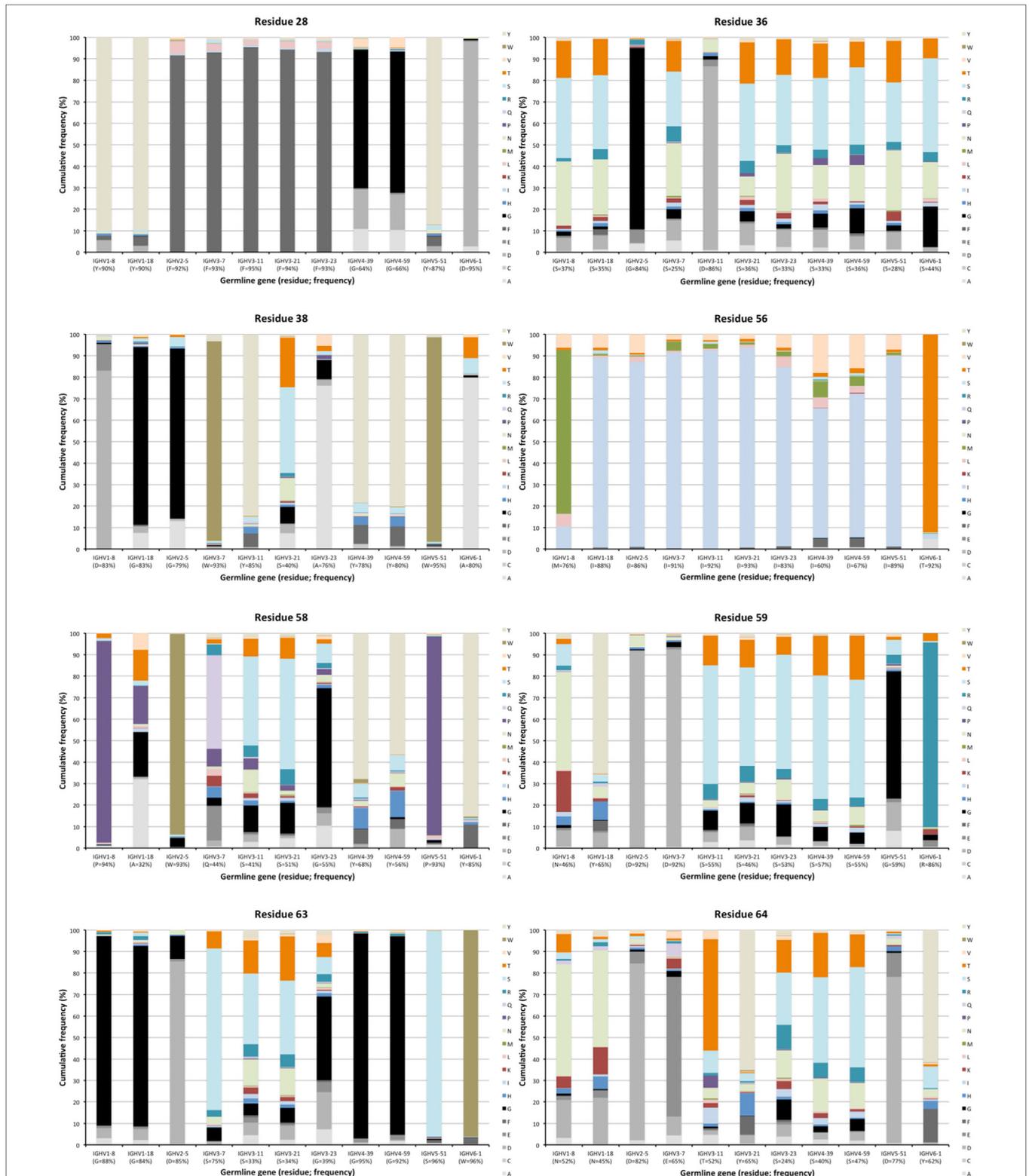
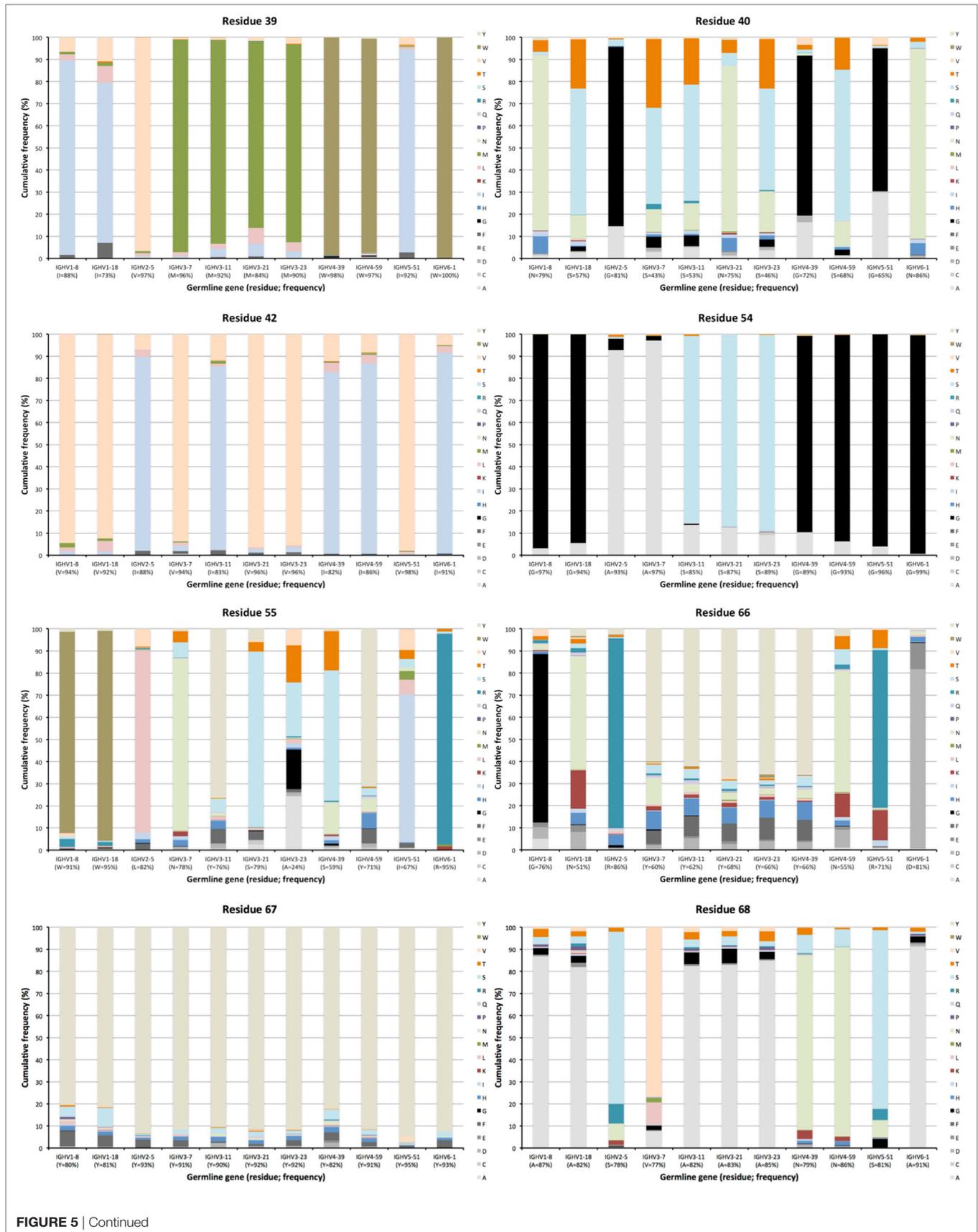


FIGURE 4 | Examples of differences in diversification of residues in complementarity determining region (CDR) encoded by transcripts of different germline gene origins. Residues shown include positions 28, 36, and 38 of CDRH1, and positions 56, 58, 59, 63, and 64 of CDRH2 of IgG. The germline-encoded residues and their frequencies in IgG-encoding transcripts are shown at the bottom of the graph. Substitutions are introduced in 5–36% (residue 28), 14–75% (residue 36), 5–60% (residue 38), 7–40% (residue 56), 6–68% (residue 58), 8–54% (residue 59), 4–67% (residue 63), and 18–76% (residue 64) of the transcripts depending on their different germline gene origins.



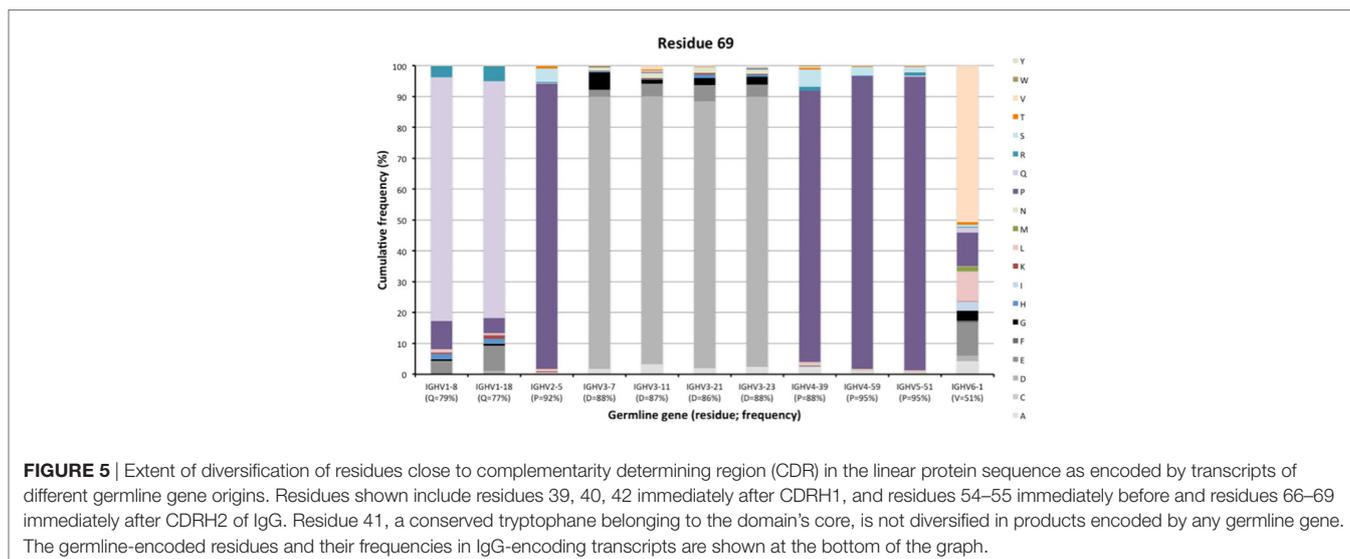


FIGURE 5 | Extent of diversification of residues close to complementarity determining region (CDR) in the linear protein sequence as encoded by transcripts of different germline gene origins. Residues shown include residues 39, 40, 42 immediately after CDRH1, and residues 54–55 immediately before and residues 66–69 immediately after CDRH2 of IgG. Residue 41, a conserved tryptophane belonging to the domain's core, is not diversified in products encoded by any germline gene. The germline-encoded residues and their frequencies in IgG-encoding transcripts are shown at the bottom of the graph.

but the diversification was largely limited in scope (such as $S \rightarrow N$ or T , or $G \rightarrow A$) (**Figure 5**). In summary, there is room for diversification of some residues often buried in the VH-VL interface, modifications that may affect the binding site or the stability of the VH-VL pair.

The upper core of antibody H chain variable domains (37) (of which residues 28, 30, 39, 78, 80, and 87 have been assessed here) is located just beneath the paratope and diversification of its residues may have profound effects on the binding site (38). Several of the residues that constitute the upper core (36) are by definition part of the sequences that comprise CDR, although their side chains are not necessarily extensively exposed on the surface of the domain. The residues play different roles, depending on their biophysical nature (37). Several residues in this core of VH, depending on its germline gene origin, are prone to accept mutations. Many germline genes encode large aromatic residues at position 28 that were rarely mutated (mostly <10%) (**Figures 2** and **4**). However, germline genes IGHV4-39 and IGHV4-59 encode a G in this position, a residue that was frequently (approximately 35%) substituted (mostly to A, D, and V) in products of IgG-encoding transcripts. Similarly, genes that encode an aromatic side chain in position 30 of VH rarely substituted it ($\leq 10\%$) while genes derived from germline genes like IGHV4-39, IGHV4-59, and IGHV6-1, which encode a hydrophobic amino acid in position 30, were more prone to substitute it, mostly for another hydrophobic residue (**Figure 9**). Residue 80 in the upper core is important for the positioning and conformation of CDR2 (38). IGHV3 germline genes incorporate R at this position, a side chain that was only very rarely ($\leq 2\%$ of reads) substituted by other residues. In contrast, R80 in antibody-encoding genes derived from IGHV1-8 underwent substitution at a high frequency. This ability for diversification is not associated with the presence of a mutational hotspot (WA/TW or RGYW/WRCY) in this codon in IGHV1-8 (Figure S3 in Supplementary Material). Other germline genes encode other residues in position 80 and these may also be substituted to

a substantial extent (**Figure 9**). Altogether, there is tolerance for diversification of many residues of the upper core in a germline-origin-dependent manner.

Antibody Evolution Provides Diversity Beyond CDR and Domain Core Structures

Hypermutation may extend to surface residues beyond CDRs, even to residues that are not located in immediate proximity to those defined to make up the CDRs. Numerous residues, in particular in FR3 carried such diversity (**Figures 2** and **10**). Sequences around residue 85 have been considered as a fourth CDR (39). This residue frequently carried diversity, a feature particularly evident in transcripts with an origin in IGHV1-8, IGHV1-18, and IGHV5-51 germline genes, in which case about 50% of the transcripts carried substitutions. This side chain is localized immediately below CDR1 in the folded domain (**Figure 10**) and it is highly conceivable that mutations may affect binding affinity and/or specificity. Certainly, antibodies derived from some germline genes show extensive evolvability in this part of the domain.

Other residues that are located at a substantial distance from CDR, were also frequently mutated. For instance, residues 90 and 92 in FR3 showed evidence of extensive diversification in transcripts derived from some germline genes, in particular those of IGHV4-39 and IGHV4-59 (**Figure 10**). Similarly, residue 101 in many VH carried a substantial level of diversification (**Figure 10**). Only in the case of IGHV3-23 was this propensity for substitution in position 101 associated with the presence of a mutational hotspot. The side chain of residue 101 is also exposed on the domain's surface near the interface with VL far away from the binding site. Some residues, although not targeted to diversification in general, may be targeted extensively in antibodies derived from some germline genes. For instance, in similarity to residue 71 of IGHV1-18 (described above; **Figure 7**), residues 46 and 81 of IGHV1-8, and to some extent residue 75 of IGHV5-51 were frequently mutated

(Figure 11). The corresponding codons of the germline genes encode T, N, and Q, respectively, while most other germline genes encode P, D, and R, respectively (Figure 1). The evolution of in particular IGHV1-8-derived H chain variable domain often introduced precisely these residues into the products.

In summary, numerous FR residues, the side chains of which are found on the surface of VH, are diversified in a germline gene-defined manner through antibodies' evolution processes *in vivo*.

Insertions and Deletions

Antibody sequences can evolve not only by hypermutation but also by insertion and deletion of entire codons (31, 32). The present dataset allows for analysis of such processes in hypermutated antibody sequences of different germline gene origins. We identified the location of such modifications (as annotated by IMGT HighV-QUEST) in transcripts derived from a number of germline genes (Figure 12; Figure S5 in Supplementary Material). Insertions were on average longer than deletions

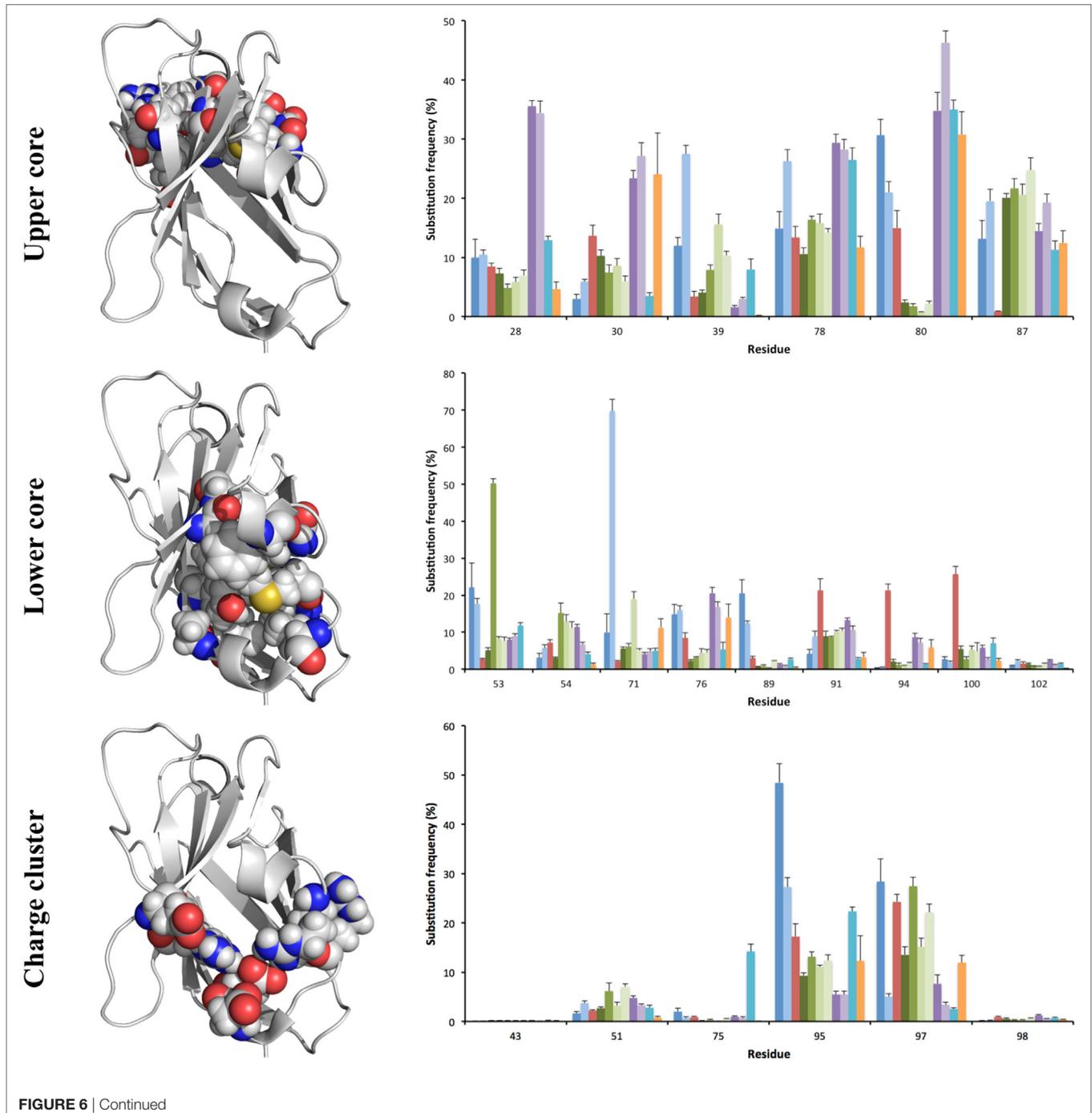


FIGURE 6 | Continued

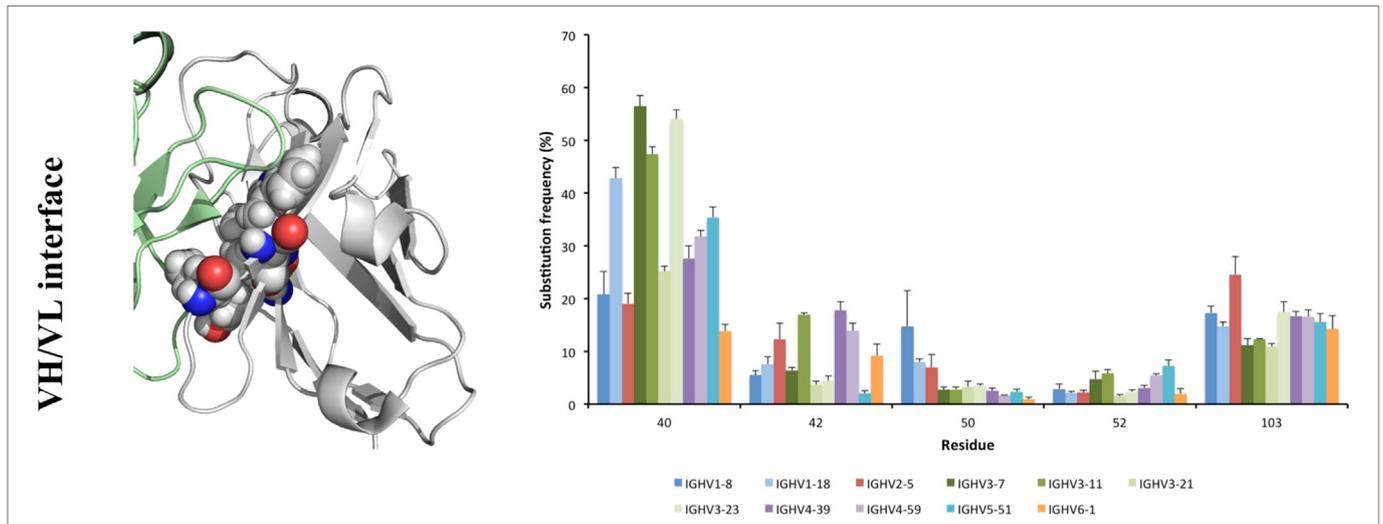


FIGURE 6 | Substitution frequency of the charge cluster, lower and upper cores, and major residues that participate in VH/VL interface of the antibody H chain. Error bars indicate SEM between repertoires of different donors. Residues belonging to these structures include those exemplified in the structure of the VH of a human antibody fragment with an origin in germline gene IGHV1-18 (PDB: 3SDY). The VH domain is shown with a gray backbone. Part of the VL domain in the illustration of the domain interface is shown in green. Achieved diversity of individual residues in the sequence populations are outlined in **Figures 7–9**, and Figure S4 in Supplementary Material.

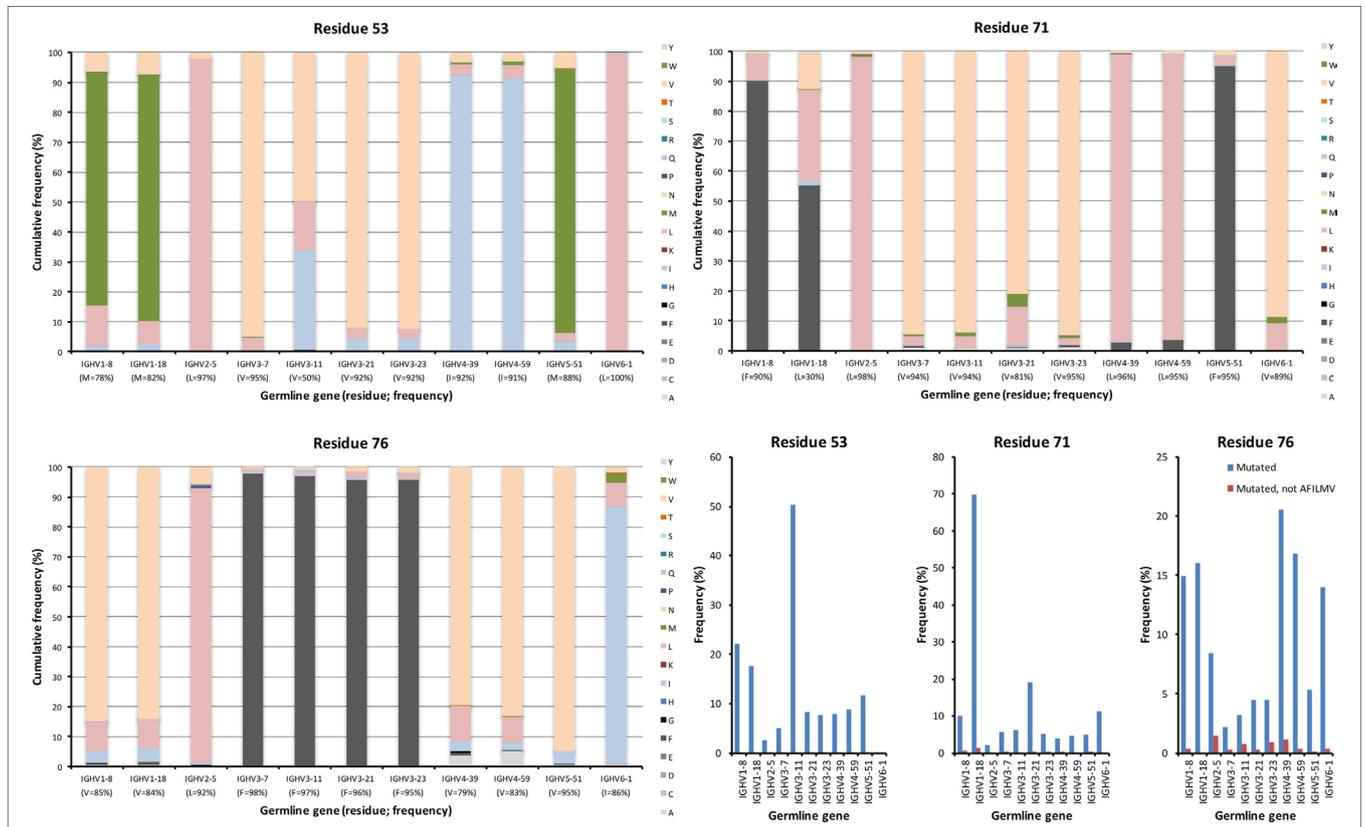


FIGURE 7 | Substitutions of lower core residues 53, 71, and 76 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes. The diversification of these mostly hydrophobic residues largely results in introduction of other hydrophobic residues (panels to the bottom right).

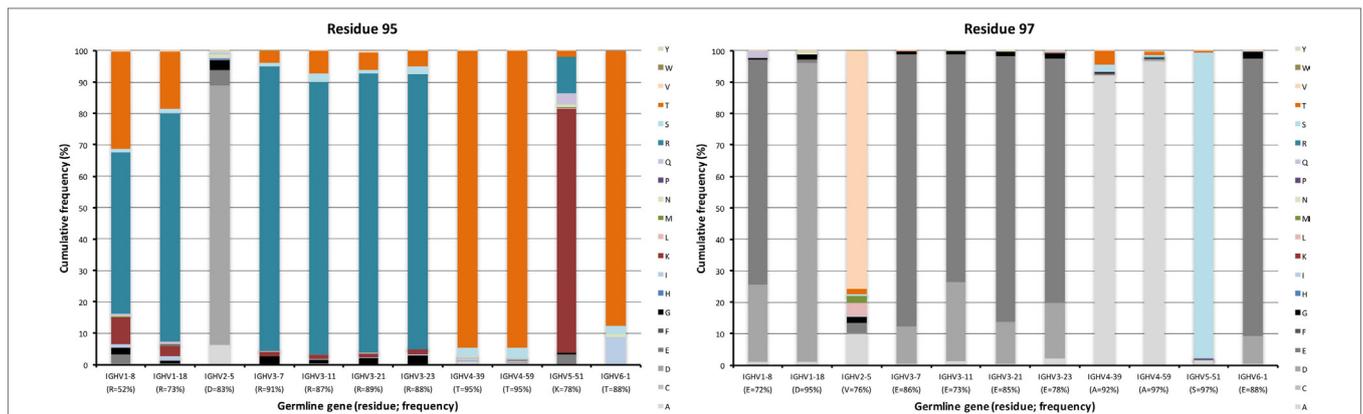


FIGURE 8 | Substitutions of charge cluster residues 95 and 97 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes.

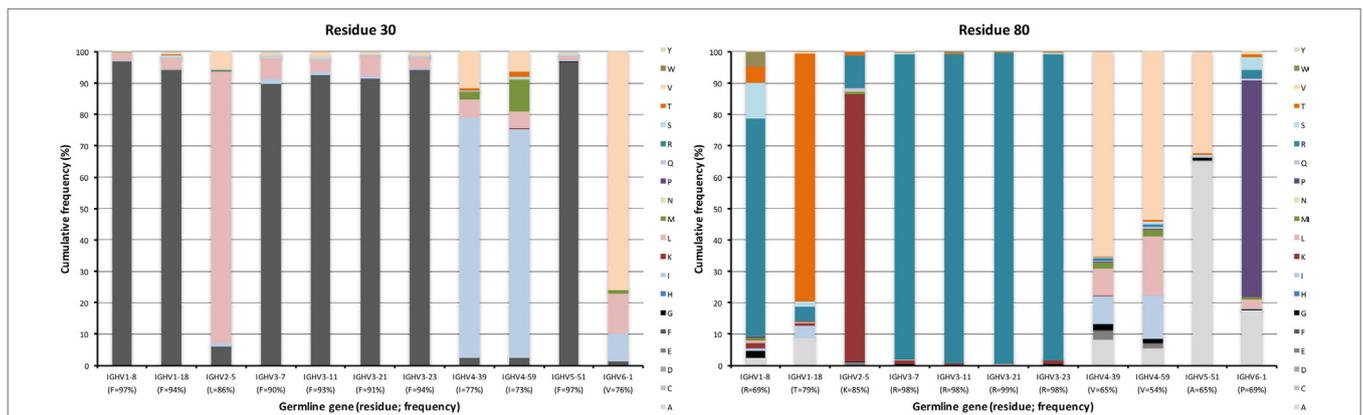


FIGURE 9 | Substitutions of upper core residues 30 and 80 as introduced during somatic hypermutation processes of IgG-encoding genes derived from different germline genes. Substitution of residue 28 is shown in **Figure 4**.

(6.7 and 4.6 bases, respectively; $p = 0.028$ using the Wilcoxon signed rank test) in in-frame transcripts with an origin in the 11 germline genes (irrespective of allelic origin) that are the focus of this study. Most such modifications occurred within/close to CDRs. Several genes (like IGHV3-23) were primarily targeted by insertions and deletions in CDRH2 while others (like IGHV2-5 and IGHV4-39) were targeted also by such modifications in CDRH1. Some genes [like IGHV3-7, IGHV5-51 (**Figure 12**), and IGHV1-69 (Figure S5 in Supplementary Material)] also extensively introduced insertions and deletions in CDRH4, i.e., in the loop situated in close proximity to other, conventional CDRs. In summary, it appears that rearranged sequences derived from different germline genes target insertions and deletions to different parts of their sequence.

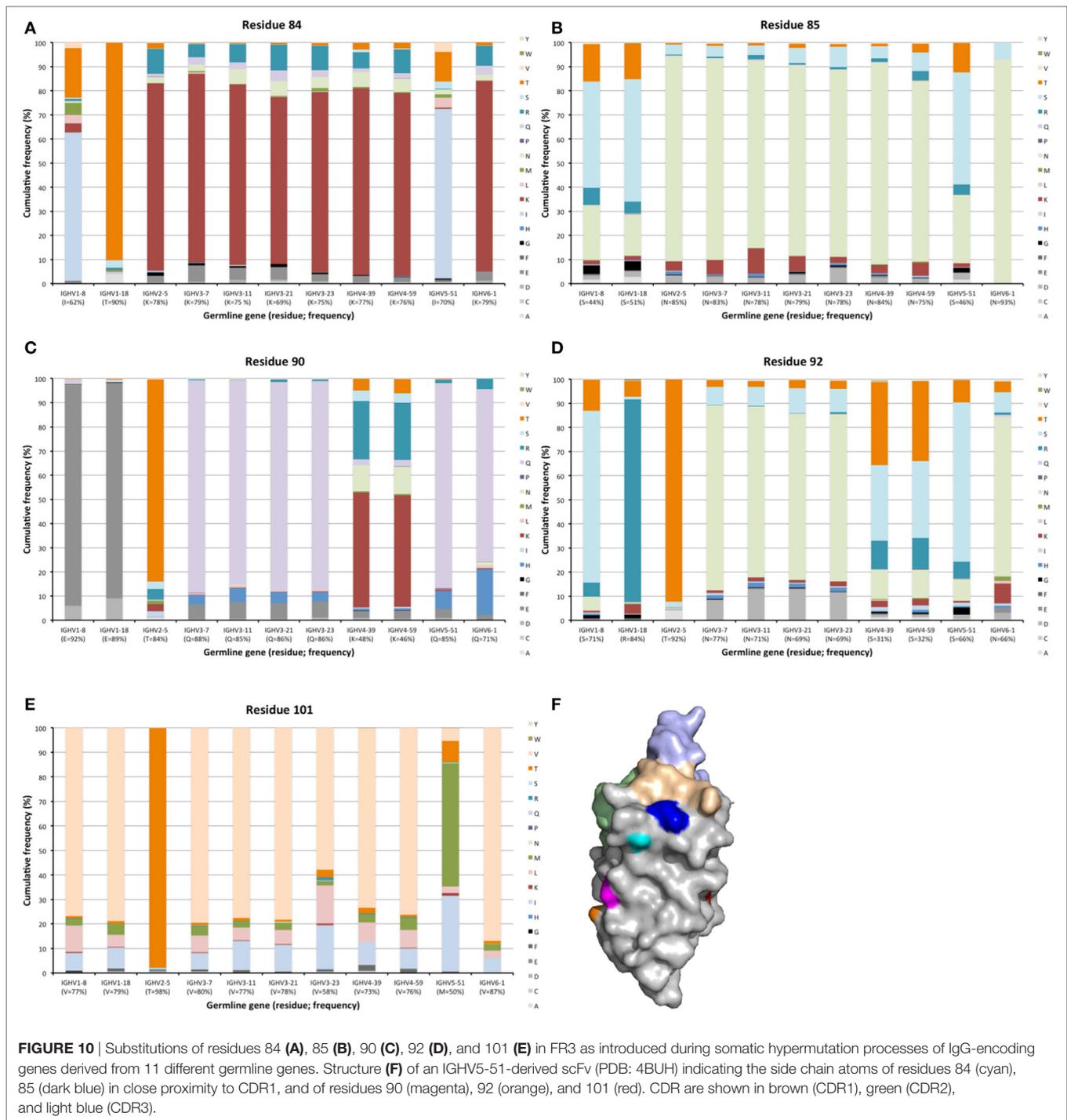
Somatic Hypermutation and Evidence of Selection

In vitro evolution through hypermutation may contain objective evidence of selection as events likely to contribute to improved binding are favored over those with no or negative influence

on antigen recognition. Such productive events are considered focused to the CDR and they may be detectable using computational approaches (33, 34, 40–42), although this possibility has also been questioned (43, 44). We investigated the evidence for such selection in only the most frequent, independent sequences of each germline in each donor to minimize the effect of random PCR and sequencing errors. Only donors that expressed a single allele of a gene were included in the analysis to minimize the risk of errors introduced by incorrect allele assignment. Such analysis demonstrated that there, despite the high degree of substitutions in FR of VH encoded by BM-derived transcripts, was a profound negative selection for mutations in FR. Although there was less selection against substitution of residues in CDR, it was not possible to identify positive selection in VH domains with an origin in any of the investigated germline genes (**Figure 13**).

DISCUSSION

Analysis of the information content of human antibody diversification holds promise for understanding antibody evolution



and affinity maturation as it occurs *in vivo*, and evolution processes as used *in vitro* to develop high affinity, highly biophysically stable antibodies. The advent of next-generation sequencing and the availability of much larger collections of antibody sequences allows for a very in-depth analysis of antibody diversity. We envisaged that such analysis would also define constraints on human repertoire development as a function of antibody germline gene origin and thus enhance the way we

in the future analyze events involving somatic diversification. Such studies and concomitant studies of antigen-antibody complex structure have been used to understand in detail how human humoral immune repertoires develop, or fail to develop efficiently. Large-scale studies, as recently reviewed, have addressed the evolution of antibodies against highly functional epitopes on viral antigens like the envelope protein of HIV-1 with the intention to enable design of immunogens that more

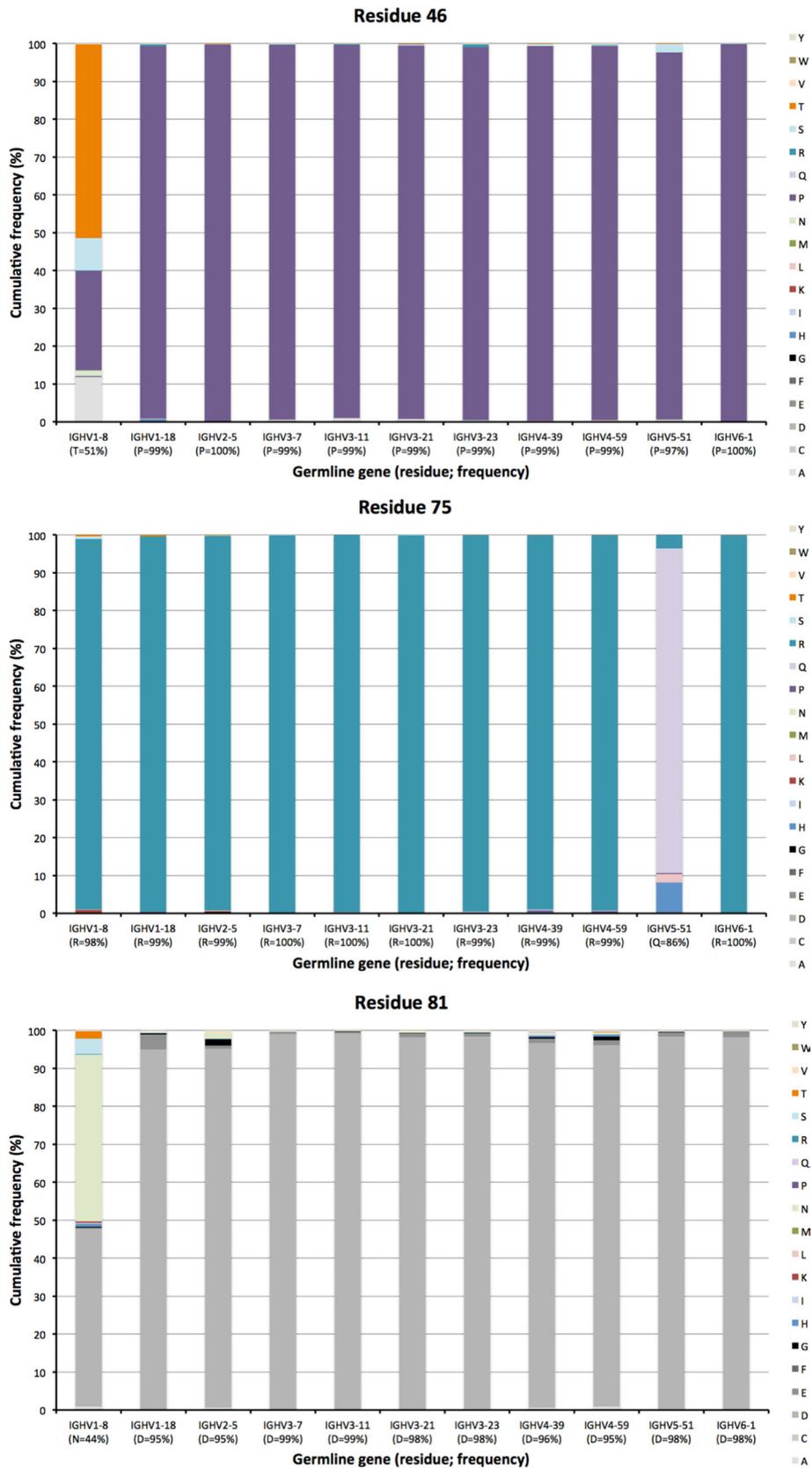
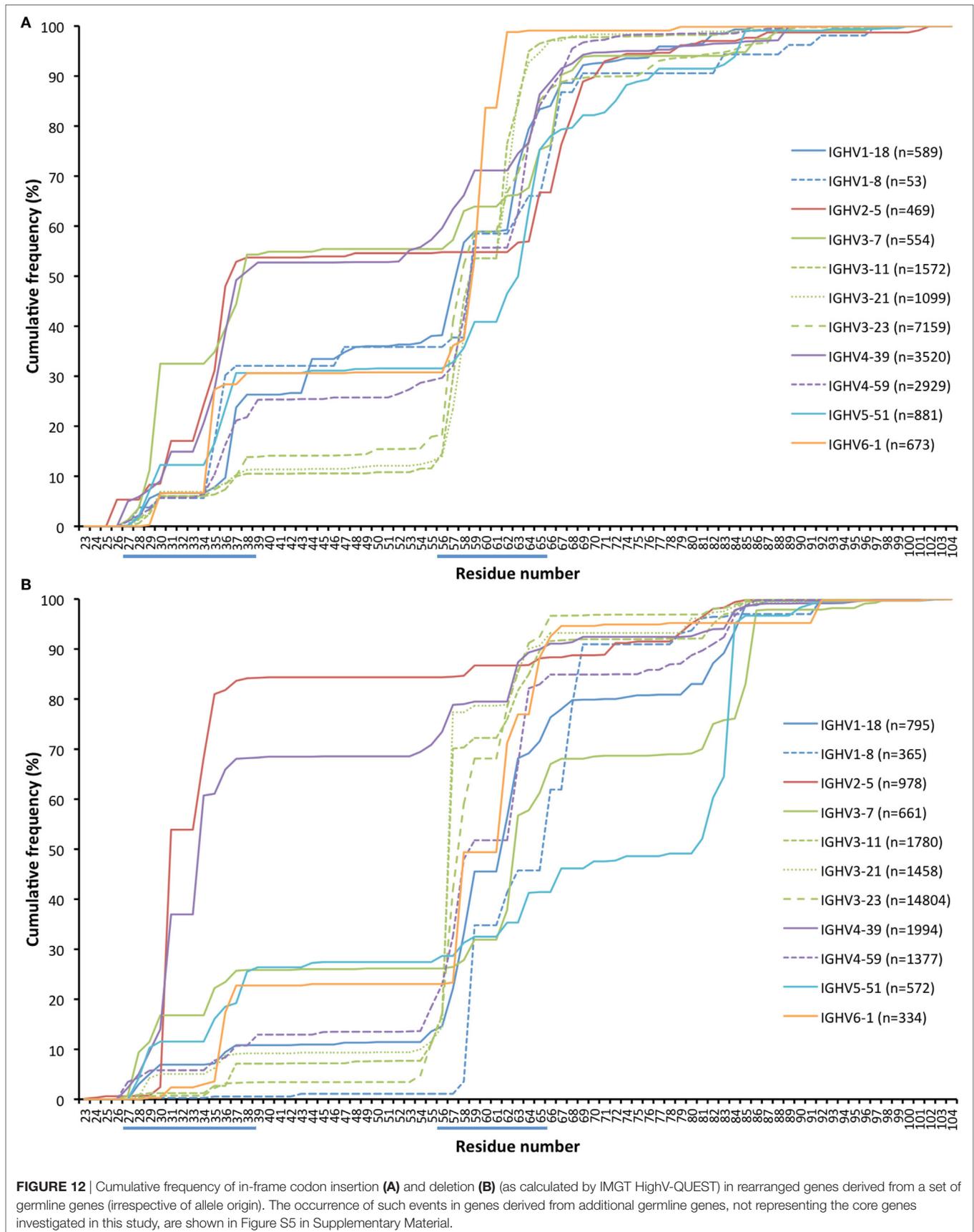


FIGURE 11 | Substitutions of residues 46, 75, and 81 as introduced during somatic hypermutation processes of IgG-encoding genes derived from 11 different germline genes.



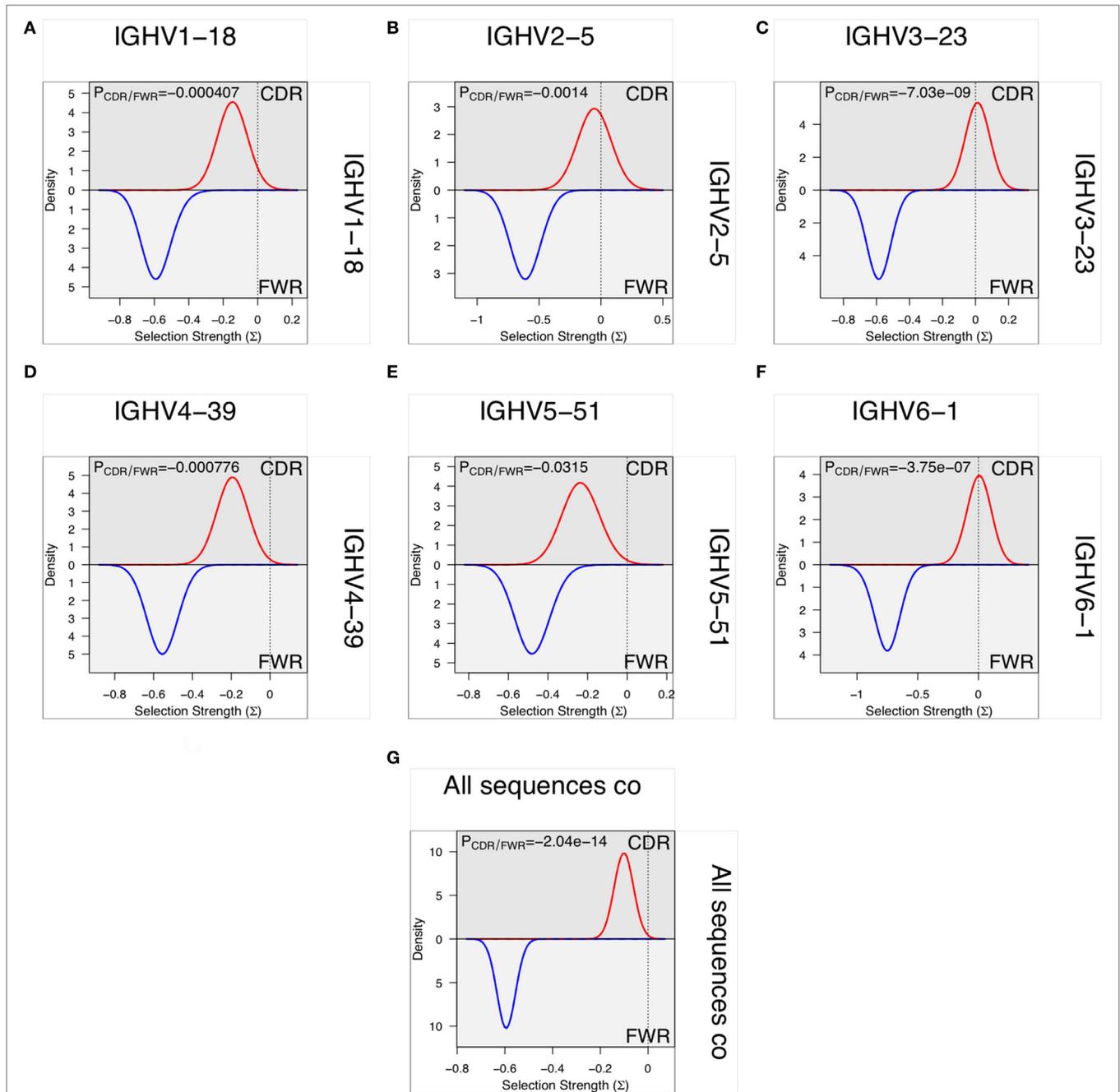


FIGURE 13 | Analysis, using BASELINE (33, 34), of selection pressure on the diversification of residues of the framework region (FR) and complementarity determining region (CDR) with an origin in six germline genes [IGHV1-18 (A), IGHV2-5 (B), IGHV3-23 (C), IGHV4-39 (D), IGHV5-51 (E), and IGHV6-1 (F)]. There is a profoundly more negative selection of diversification of residues of FR2 and FR3 than of CDR1 and CDR2 but there is no evidence for positive selection in the substitution pattern of CDR of sequences derived from any of the individual germline genes. Indeed, overall there is evidence for negative selection for substitution also in the CDR (G).

efficiently induce protective immunity in vaccinated subjects (5, 6, 45). In the present study, the aim was to deconvolute antibody diversification paths, not from a global perspective but with a focus on products of individual germline genes, to enable enhanced quality of future analysis of antibody evolution. To do so, we employed sequences that encode IgG in BM (14),

a major site of long-term, sustained antibody production. The sequences were derived from subjects diagnosed with seasonal allergic rhinitis but they were obtained out-of season of most seasonal allergens. We consequently consider them not to be biased by an on-going allergic immune response. In any event, we do not consider that major features of evolution of the IgG

response of such subjects would be dramatically different from that of non-allergic, immunocompetent, subjects.

The present study focuses on the diversity found from CDR1 until the end of FR3 of the human H chain V domain. By focusing our attention on diversification of well-defined germline genes and alleles, our analysis is minimally confounded by differences in germline gene allelic makeup between individuals or between haplotypes of an individual. In preparation of the present study, we consequently inferred the germline VH repertoires of the lymphocyte donors (19) and analyzed genes of donors with well-established germline gene allele composition. The one exception to this rule is IGHV3-23 and its, in their mature peptide-coding sequences, exactly duplicated sequence IGHV3-23D, sequences that were treated as one entity in this study. Furthermore, as some germline genes are highly similar, confounding outcomes may occur as a consequence of mutational processes rendering sequences derived from one germline gene more similar to the nucleotide sequence of other germline genes. Such highly similar germline genes (>98% nucleotide identity) were, to avoid erroneous interpretation, not investigated in the present study.

Antibodies of different germline gene origins differed substantially in terms of diversified residues, in agreement with recent findings (46). There is thus a solid basis for defined, preferred germline-centric paths of antibody evolution. Certainly, driving forces that promote higher substitution frequencies may relate to affinity maturation, stability enhancement etc. It is likely that part of the observed differences relate to the presence or absence of sequences acting as hot-spots for the mutational machinery (47). However, structural analysis have previously demonstrated that amino acids that are in mutational hot-spots are not more likely to actually undergo substitution during somatic hypermutation, suggesting that such hot-spots “are not a major driving force in determining which residues are mutated” (48). In any event, IGHV2-5, IGHV4-39, IGHV4-59, IGHV5-51, and IGHV6-1 all encode S29, a residue capable of different interactions in different antibodies (Figures 14A–D). It is, however, only in sequences derived from IGHV5-51 that this residue is targeted by extensive substitutions. This gene is also the only one among the five that carries a mutational hot-spot motif affecting this codon. If this is a selected, germline-encoded feature preventing extensive, non-functional substitution of products derived from the other four genes, or not, is currently not known. However, if this is not the case, there is a capacity for antibody evolution in antibodies derived from some genes that is not efficiently explored by the human immune system.

Beyond hot-spot involvement in the orchestration of substitution, it is also likely that structural consideration in many cases guide the ability of antibodies of different germline gene origins to tolerate or even prefer substitutions. Certainly, loops belonging to different canonical classes (49, 50), positioning side chains with identical residue numbers in entirely differently orientations and environments, may affect their ability to structurally accept substitutions. Importantly, some germline genes encode unusual residues in some positions, the side chain

of which may be suboptimal for its environment. We envisage that such unusual residue may be more commonly mutated even if they reside in FR, often attaining the more common residue following substitution, as illustrated by residues 53 of IGHV3-11 and residue 71 or IGHV1-18. Furthermore, residue 80, a residue in the upper core of the variable domain, with particular importance for the conformation of CDRH2 (38) provide interesting insight into germline-directed paths for antibody evolution. The common diversification of R80 in products with an origin in IGHV1-8 (in contrast to the lack of diversification of R80 in products derived from IGHV3 subgroup genes) (Figures 6 and 9) is not associated with the presence of a mutational hotspot in IGHV1-8 affecting codon 80 (Figure S3 in Supplementary Material). Possibly R80 is less important for maintenance of the integrity of the domain or the general architecture of the binding site in antibodies derived from this germline gene. Interestingly, reorientation of R80 has been demonstrated in one antibody (24) with a likely origin in IGHV1-8. This was also associated with a reorientation of CDRH2 as determined by X-ray crystallography (Figures 14E–H). We hypothesize that substitution of R80 in products originating from genes like IGHV1-8 may be part of an efficient route to evolve antibody functionality while still being tolerated in terms of structural stability. We hypothesize that some germline genes may even have an inherent need, or, if one so-prefer, capacity, for evolution that is not present in other germline genes. In all, the reason for such high substitution frequency *in vivo* may differ between antibodies of different germline gene origins. Future studies will have to address the difficulties encountered by, or alternatively extended opportunity of, the cells producing antibodies derived from these germline genes to gain an advantage in the race for selection through the affinity maturation processes occurring in germinal centers.

Residues beyond CDR may interact with antigen or contribute indirectly to the architecture of the binding site. There are, however, many definitions of CDRs (8–12) apart from the one used in this study (13), definitions that accommodate different viewpoints of what constitutes an antigen binding site. Indeed, substantial diversity is apparently tolerated, or even selected for, in residues in immediate proximity to CDRs as defined by IMGT (Figures 1 and 5). For instance, codon 55 of different germline genes encodes very different side chains. Is the observed difference in mutability solely a result of the presence or absence of a mutational hotspot, or is it also directed by a difference in importance of this residue for establishment of a core antigen-interacting surface (51) in products encoded by the different germline genes? It is furthermore evident that many other residues in FRs, even those belonging to the cores of the antibody fold and to the VH/VL interface, may harbor extensive diversity. We envisage that such diversity, when structurally tolerated may contribute substantially to improved biophysical properties of the encoded antibody or even to affinity maturation, for instance through stabilization of the binding site during affinity maturation (52). In particular, an area with CDR-like potential, CDR4, that resides in a loop adjacent to the

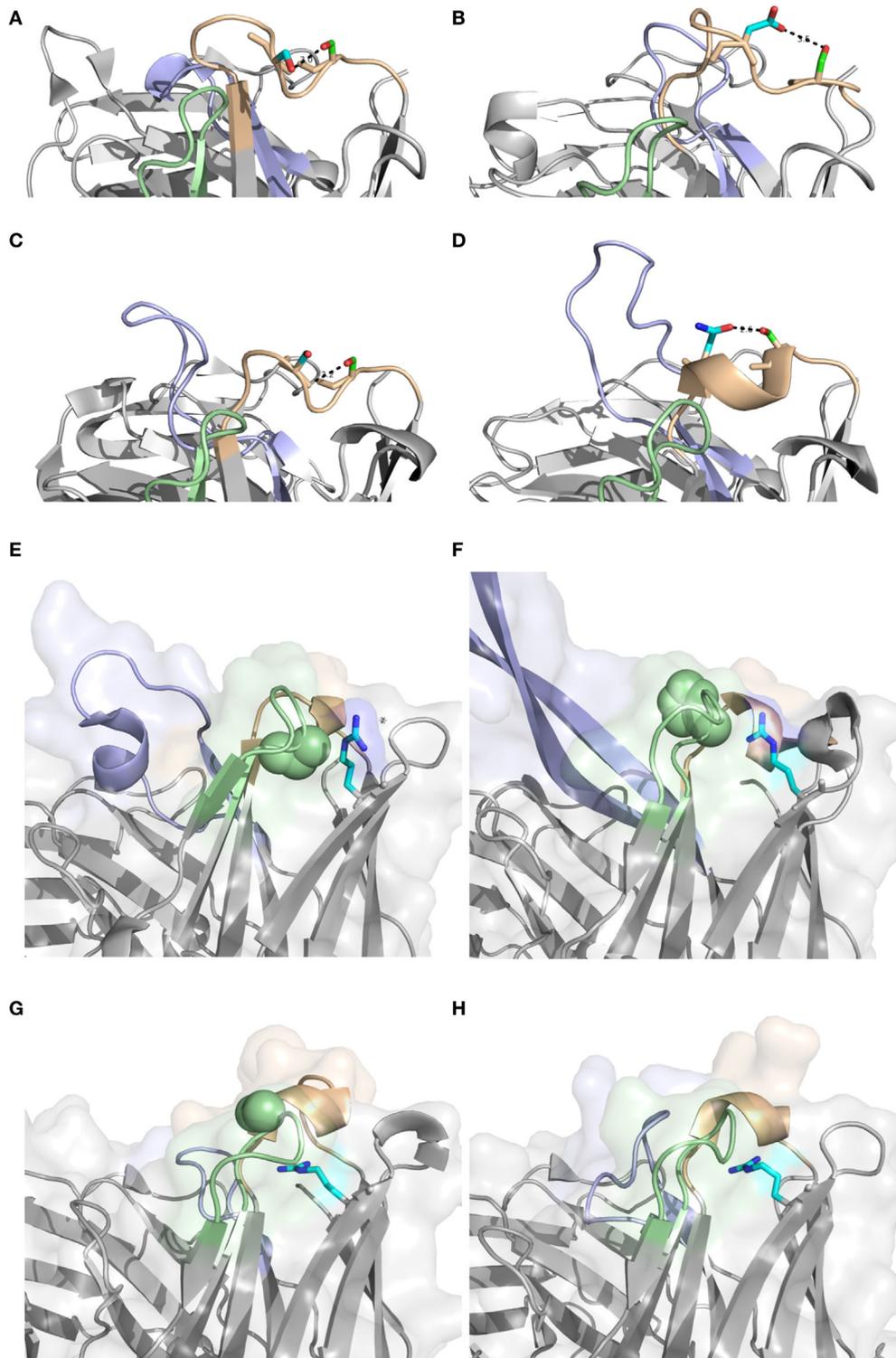


FIGURE 14 | Position 29 of rearranged sequences of different germline gene origins are to different extents targeted by mutagenesis resulting in amino acid substitution (**Figure 3**). A diversity of potential polar interactions of the side chain of S29 have been implicated such as those to O_γ of S31 (IGHV2-5; PDB: 3QRG) (**A**), O_{δ2} of D31 (IGHV4-39; PDB: 5C6T) (**B**), backbone O of S31 (IGHV4-59; PDB: 3H11) (**C**), and O_{δ1} of N36 (IGHV5-51; PDB: 4BUH) (**D**). H chain CDR1, CDR2, and CDR3, are colored in brown, green, and blue, respectively. Side chain atoms are colored in red (oxygen), blue (nitrogen), green (carbon of residue 29) and cyan (carbon of residue 31 or 36). Position of R80 (carbon of side chain shown in cyan) and residue 58 (residues of the side chain as spheres). Structures PDB: 3×3G (with P58) (**E**) and PDB: 3U1S (with H58) (**F**) both derived from IGHV1-8, and PDB: 3FZU (with S58) (**G**) and PDB: 2R56 (with G58) (**H**), both derived from the IGHV3 subgroup, are shown. H chain CDR1, CDR2, and CDR3, are colored in brown, green, and blue, respectively.

classical CDRs in the folded structure have been defined and exploited (11, 39, 48, 53). It has been suggested to be able to accommodate extensive diversity (39). We have now identified that some, but not all, germline genes introduce diversity in this loop. Such diversity may contribute to functional evolution of antibodies of such germline gene origins. In all, the preferred paths of evolution of antibody V domains extend substantially beyond conventional CDRs in ways directed by an antibody's germline gene origin.

Antibody variable domains, indeed, diversify not only by substitution but also through insertion and deletion of residues into the variable domain sequence (31, 32). In this study, we observed germline gene-inherited patterns that differently target genes with such insertions and deletions, not only in conventional CDRs but also in CDR4 (**Figure 12**; Figure S5 in Supplementary Material). In the past, we hypothesized that the presence of repetitive codons might be one feature that targets such modifications to a particular part of a gene (54). It is also conceivable that parts of VH domains of different germline gene origins are able to structurally harbor such diversity to different extents, a factor that certainly needs further assessment. Nevertheless, it is conceivable that any immune response that relies on introduction of sequence insertion and/or deletion *in vivo* will only recruit members derived from those germline genes that introduce such diversification with ease in critical parts of the sequence. Important immune responses that require such modification have been reported (5, 55) and other responses relying on such evolution, such as those targeting occluded sites, will likely be described in the future. By understanding the ability of particular germline genes to diversify by insertion and deletion, it will be possible to develop our understanding of selection of germline genes made by the immune system in the generation of these particularly difficult immune responses.

Overall, the diverse pattern of diversification even beyond conventional CDR likely complicates computational efforts to assess the involvement of selection during antibody development. We employed one such analysis technology on highly expressed IgG H chain V domain sequences encoded in BM but found no evidence of a positive selection force in the mutational pattern targeting CDR. Our findings are in line with past studies demonstrating a failure to identify evidence of positive selection in CDR, while evidence of negative selection of modifications in FR is detected (43, 44). In agreement with a recent study (46), and given the diversity of paths through which antibodies of different origins evolve, we suggest that any approach to assess selection ought to take germline gene-specific mutational patterns as found in selected and non-selected repertoires into account and not rely entirely on an analysis of mutations based on current global CDR definitions. Processes to facilitate analysis of selection in a germline gene origin-centric fashion have been initiated elsewhere (46). We foresee that such a development will be required if computational approaches are to accurately address the impact of selection on antibody repertoire development. In some situations, such as IgE responses, this aspect is a matter of

substantial biological controversy (56, 57) and certainly need further investigations, the outcomes of which will impact our understanding of fundamental biological processes associated with disease.

In summary, we identified germline gene-unique patterns of evolution that occur during hypermutation of antibodies of diverse IGHV germline gene origins. Our findings extend the findings of a recent study, published during preparation of the present manuscript, that identified gene-specific substitution profiles of antibodies of different germline gene origins (46). Collectively, we have demonstrated a diversity of paths taken by antibodies of different germline gene origins to evolve by somatic hypermutation, including not only base substitution but also processes of codon insertion and deletion. Our study forms the basis for improved understanding of molecular evolution as it proceeds in immune responses *in vivo* and establishes a foundation for future germline gene origin-centered analysis approaches.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Regionala etikprövningsnämnden (Lund). All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Regionala etikprövningsnämnden (Lund).

AUTHOR CONTRIBUTIONS

UK: bioinformatic pipeline development and bioinformatic analysis, manuscript preparation, and approved the final manuscript. HP: conceived the study, manuscript preparation, and approved the final manuscript. FL: initial bioinformatic pipeline development, manuscript preparation, and approved the final manuscript. LG: patient management, manuscript preparation, and approved the final manuscript. MO: conceived the study, bioinformatic analysis, main responsibility for manuscript preparation, and approved the final manuscript.

FUNDING

This study was supported by grants from the Swedish Research Council (grant number 2016-01720), and Lund University's Avtal om Läkarutbildning och Forskning (ALF). 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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at <http://www.frontiersin.org/article/10.3389/fimmu.2017.01433/full#supplementary-material>.

REFERENCES

- Methot SP, Di Noia JM. Molecular mechanisms of somatic hypermutation and class switch recombination. *Adv Immunol* (2017) 133:37–87. doi:10.1016/bs.ai.2016.11.002
- Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat Biotechnol* (2014) 32:158–68. doi:10.1038/nbt.2782
- Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* (2015) 7:121. doi:10.1186/s13073-015-0243-2
- Boyd SD, Crowe JE Jr. Deep sequencing and human antibody repertoire analysis. *Curr Opin Immunol* (2016) 40:103–9. doi:10.1016/j.coi.2016.03.008
- Kepler TB, Wiehe K. Genetic and structural analyses of affinity maturation in the humoral response to HIV-1. *Immunol Rev* (2017) 275:129–44. doi:10.1111/imr.12513
- Kwong PD, Chuang GY, Dekosky BJ, Gindin T, Georgiev IS, Lemmin T, et al. Antibodyomics: bioinformatics technologies for understanding B-cell immunity to HIV-1. *Immunol Rev* (2017) 275:108–28. doi:10.1111/imr.12480
- Lefranc MP. IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harb Protoc* (2011) 2011:595–603. doi:10.1101/pdb.top115
- Chothia C, Lesk AM. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol* (1987) 196:901–17. doi:10.1016/0022-2836(87)90412-8
- Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* (1989) 342:877–83. doi:10.1038/342877a0
- Kabat EA, Wu TT, Perry HM, Gottesmann KS, Foeller C. *Sequences of Proteins of Immunological Interest*. Bethesda, MD: U.S. Department of Health and Human Services (1991).
- Honegger A, Plückthun A. Yet another numbering scheme for immunoglobulin variable domains: an automatic modeling and analysis tool. *J Mol Biol* (2001) 309:657–70. doi:10.1006/jmbi.2001.4662
- Abhinandan KR, Martin AC. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol* (2008) 45:3832–9. doi:10.1016/j.molimm.2008.05.022
- Lefranc MP. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc* (2011) 2011:633–42. doi:10.1101/pdb.ip85
- Levin M, Levander F, Palmason R, Greiff L, Ohlin M. Antibody-encoding repertoires of bone marrow and peripheral blood—a focus on IgE. *J Allergy Clin Immunol* (2017) 139:1026–30. doi:10.1016/j.jaci.2016.06.040
- Corcoran MM, Phad GE, Bernat NV, Stahl-Hennig C, Sumida N, Persson MA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun* (2016) 7:13642. doi:10.1038/ncomms13642
- Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* (2014) 30:1930–2. doi:10.1093/bioinformatics/btu138
- Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) tools for the nucleotide analysis of immunoglobulin (IG) and T cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi:10.1007/978-1-61779-842-9_32
- Kirik U, Greiff L, Levander F, Ohlin M. Data on haplotype-supplemented immunoglobulin germline gene inference. *Data Brief* (2017) 13:620–40. doi:10.1016/j.dib.2017.06.031
- Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol* (2017) 87:12–22. doi:10.1016/j.molimm.2017.03.012
- Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. *Bioinformatics* (2007) 23:2947–8. doi:10.1093/bioinformatics/btm404
- Giudicelli V, Brochet X, Lefranc MP. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T cell receptor (TR) nucleotide sequences. *Cold Spring Harb Protoc* (2011) 2011:695–715. doi:10.1101/pdb.prot5633
- Ekiert DC, Friesen RH, Bhabha G, Kwaks T, Jongeneelen M, Yu W, et al. A highly conserved neutralizing epitope on group 2 influenza A viruses. *Science* (2011) 333:843–50. doi:10.1126/science.1204839
- McLellan JS, Pancera M, Carrico C, Gorman J, Julien JP, Khayat R, et al. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* (2011) 480:336–43. doi:10.1038/nature10696
- Tamada T, Shinmi D, Ikeda M, Yonezawa Y, Kataoka S, Kuroki R, et al. TRAIL-R2 superoligomerization induced by human monoclonal agonistic antibody KMTR2. *Sci Rep* (2015) 5:17936. doi:10.1038/srep17936
- Niemi M, Jylha S, Laukkanen ML, Söderlund H, Makinen-Kiljunen S, Kallio JM, et al. Molecular interactions between a recombinant IgE antibody and the beta-lactoglobulin allergen. *Structure* (2007) 15:1413–21. doi:10.1016/j.str.2007.09.012
- Houde D, Arndt J, Domeier W, Berkowitz S, Engen JR. Characterization of IgG1 conformation and conformational dynamics by hydrogen/deuterium exchange mass spectrometry. *Anal Chem* (2009) 81:2644–51. doi:10.1021/ac802575y
- Chandramouli S, Ciferri C, Nikitin PA, Calo S, Gerrein R, Balabanis K, et al. Structure of HCMV glycoprotein B in the postfusion conformation bound to a neutralizing human antibody. *Nat Commun* (2015) 6:8176. doi:10.1038/ncomms9176
- Chen L, Kwon YD, Zhou TQ, Wu XL, O'dell S, Cavacini L, et al. Structural basis of immune evasion at the site of CD4 attachment on HIV-1 gp120. *Science* (2009) 326:1123–7. doi:10.1126/science.1175868
- Levin M, Davies AM, Liljekvist M, Carlsson F, Gould HJ, Sutton BJ, et al. Human IgE against the major allergen Bet v 1 – defining an epitope with limited cross-reactivity between different PR-10 family proteins. *Clin Exp Allergy* (2014) 44:288–99. doi:10.1111/cea.12230
- Ehrenmann F, Lefranc MP. 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) 2011:750–61. doi:10.1101/pdb.prot5637
- Ohlin M, Borrebaeck CAK. Insertions and deletions in hypervariable loops of antibody heavy chains contribute to molecular diversity. *Mol Immunol* (1998) 35:233–8. doi:10.1016/S0161-5890(98)00030-3
- Wilson PC, De Bouteiller O, Liu YJ, Potter K, Bancheau J, Capra JD, et al. Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J Exp Med* (1998) 187:59–70. doi:10.1084/jem.187.1.59
- Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res* (2011) 39:W499–504. doi:10.1093/nar/gkr413
- Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res* (2012) 40:e134. doi:10.1093/nar/gks457
- Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol* (2010) 184:6986–92. doi:10.4049/jimmunol.1000445
- Honegger A, Malebranche AD, Rothlisberger D, Plückthun A. The influence of the framework core residues on the biophysical properties of immunoglobulin heavy chain variable domains. *Protein Eng Des Sel* (2009) 22:121–34. doi:10.1093/protein/gzn077
- Ewert S, Huber T, Honegger A, Plückthun A. Biophysical properties of human antibody variable domains. *J Mol Biol* (2003) 325:531–53. doi:10.1016/S0022-2836(02)01237-8
- Tramontano A, Chothia C, Lesk AM. Framework residue-71 is a major determinant of the position and conformation of the 2nd hypervariable region in the Vh domains of immunoglobulins. *J Mol Biol* (1990) 215:175–82. doi:10.1016/S0022-2836(05)80102-0
- Bond CJ, Wiesmann C, Marsters JC Jr, Sidhu SS. A structure-based database of antibody variable domain diversity. *J Mol Biol* (2005) 348:699–709. doi:10.1016/j.jmb.2005.02.063
- Lossos IS, Tibshirani R, Narasimhan B, Levy R. The inference of antigen selection on Ig genes. *J Immunol* (2000) 165:5122–6. doi:10.4049/jimmunol.165.9.5122
- Dahlke I, Nott DJ, Ruhno J, Sewell WA, Collins AM. Antigen selection in the IgE response of allergic and nonallergic individuals. *J Allergy Clin Immunol* (2006) 117:1477–83. doi:10.1016/j.jaci.2005.12.1359
- Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol* (2008) 20:683–94. doi:10.1093/intimm/dxn026

43. MacDonald CM, Boursier L, D'cruz DP, Dunn-Walters DK, Spencer J. Mathematical analysis of antigen selection in somatically mutated immunoglobulin genes associated with autoimmunity. *Lupus* (2010) 19:1161–70. doi:10.1177/0961203310367657
44. Levin M, Ohlin M. Inconclusive evidence for or against positive antigen selection in the shaping of human immunoglobulin E repertoires: a call for new approaches. *Int Arch Allergy Immunol* (2013) 161:122–6. doi:10.1159/000345421
45. Ward AB, Wilson IA. The HIV-1 envelope glycoprotein structure: nailing down a moving target. *Immunol Rev* (2017) 275:21–32. doi:10.1111/imr.12507
46. Sheng Z, Schramm CA, Kong R, Program NCS, Mullikin JC, Mascola JR, et al. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol* (2017) 8:537. doi:10.3389/fimmu.2017.00537
47. Rogozin IB, Diaz M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J Immunol* (2004) 172:3382–4. doi:10.4049/jimmunol.172.6.3382
48. Burkovitz A, Sela-Culang I, Ofra Y. Large-scale analysis of somatic hypermutations in antibodies reveals which structural regions, positions and amino acids are modified to improve affinity. *FEBS J* (2014) 281:306–19. doi:10.1111/febs.12597
49. Al-Lazikani B, Lesk AM, Chothia C. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol* (1997) 273:927–48. doi:10.1006/jmbi.1997.1354
50. Nowak J, Baker T, Georges G, Kelm S, Klostermann S, Shi J, et al. Length-independent structural similarities enrich the antibody CDR canonical class model. *MAbs* (2016) 8:751–60. doi:10.1080/19420862.2016.1158370
51. Tomlinson IM, Walter G, Jones PT, Dear PH, Sonnhammer EL, Winter G. The imprint of somatic hypermutation on the repertoire of human germline V genes. *J Mol Biol* (1996) 256:813–7. doi:10.1006/jmbi.1996.0127
52. Wedemayer GJ, Patten PA, Wang LH, Schultz PG, Stevens RC. Structural insights into the evolution of an antibody combining site. *Science* (1997) 276:1665–9. doi:10.1126/science.276.5319.1665
53. Young NM, Watson DC, Cunningham AM, Mackenzie CR. The intrinsic cysteine and histidine residues of the anti-*Salmonella* antibody Se155-4: a model for the introduction of new functions into antibody-binding sites. *Protein Eng Des Sel* (2014) 27:383–90. doi:10.1093/protein/gzu018
54. Lantto J, Ohlin M. Uneven distribution of repetitive trinucleotide motifs in human immunoglobulin heavy variable genes. *J Mol Evol* (2002) 54:346–53. doi:10.1007/s00239-001-0049-2
55. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity* (2013) 39:245–58. doi:10.1016/j.immuni.2013.04.012
56. Davies JM, Platts-Mills TA, Aalberse RC. The enigma of IgE+ B-cell memory in human subjects. *J Allergy Clin Immunol* (2013) 131:972–6. doi:10.1016/j.jaci.2012.12.1569
57. Gadermaier E, Levin M, Flicker S, Ohlin M. The human IgE repertoire. *Int Arch Allergy Immunol* (2014) 163:77–91. doi:10.1159/000355947

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Kirik, Persson, Levander, Greiff and Ohlin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.