Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing

Kevin Larimore, Michael W. McCormick, Harlan S. Robins and Philip D. Greenberg

*J Immunol* published online 3 August 2012
http://www.jimmunol.org/content/early/2012/08/03/jimmunol.1201303

Supplementary Material  http://www.jimmunol.org/content/suppl/2012/08/03/jimmunol.1201303.3.DC1

Subscription Information about subscribing to *The Journal of Immunology* is online at: http://jimmunol.org/subscription

Permissions Submit copyright permission requests at: http://www.aai.org/About/Publications/JI/copyright.html

Email Alerts Receive free email-alerts when new articles cite this article. Sign up at: http://jimmunol.org/alerts

Errata An erratum has been published regarding this article. Please see next page or: /content/192/1/534.full.pdf
Shaping of Human Germline IgH Repertoires Revealed by Deep Sequencing

Kevin Larimore,* Michael W. McCormick,† Harlan S. Robins,‡ and Philip D. Greenberg*§¶

To understand better how selection processes balance the benefits of Ig repertoire diversity with the risks of autoreactivity and nonfunctionality of highly variable IgH CDR3s, we collected millions of rearranged germline IgH CDR3 sequences by deep sequencing of DNA from mature human naive B cells purified from four individuals and analyzed the data with computational methods. Long HCDR3 regions, often components of HIV-neutralizing Abs, appear to derive not only from incorporation of long D genes and insertion of large N regions but also by usage of multiple D gene segments in tandem. However, comparison of productive and out-of-frame IgH rearrangements revealed a selection bias against long HCDR3 loops, suggesting these may be disproportionately either poorly functional or autoreactive. Our data suggest that developmental selection removes HCDR3 loops containing patches of hydrophobicity, which are commonly found in some auto-antibodies, and at least 69% of the initial productive IgH rearrangements are removed from the repertoire during B cell development. Additionally, we have demonstrated the potential utility of this new technology for vaccine development with the identification in all four individuals of related candidate germline IgH precursors of the HIV-neutralizing Ab 4E10. The Journal of Immunology, 2012, 189: 000–000.

Recognition of a wide variety of foreign molecular shapes by human B cells is made possible by the assembly of unique Ig Ag binding domains in developing B cells from an array of variable (V), diversity (D), and joining (J) gene segments found in the Ig H and L chain loci through a process of genomic rearrangement termed VDJ recombination (1). The structural stability of each Ig Ag binding domain is maintained by a sandwich of seven β sheets known as the Ig fold, encoded by conserved framework region sequences, and Ag recognition is largely determined by the highly variable CDR sequences that form loops between the conserved β sheets of the Ig fold. There are six CDR loops, three in the H chain and three in the L chain, but the H chain CDR3, formed by the junction of trimmed or full-length V, D, and J gene segments plus a variable number of non-templated N nucleotides inserted between the segments, is the longest, most variable CDR loop. It forms the center of the Ag binding site and typically makes the most Ag contacts (2), thus contributing significantly to the unique Ag recognition properties of each Ab. The first and second CDR loops of Ig H chains are entirely genomically encoded within IgHV genes, with evolutionary selection of IgHV gene sequences resulting in a limited set of canonical loop conformations (3, 4). In contrast, the IgH CDR3 loop is much more variable in sequence, length, and structure (5), with the genomically encoded central D gene segment that typically forms the apex of the loop flanked on both sides by N nucleotides encoding residues proximal to the base of the loop that have been added randomly between V–D and D–J gene segments during IgH rearrangement, and therefore have not been subjected to evolutionary selection.

IgH CDR3 loops can vary in both length and sequence, allowing for the ability to recognize diverse Ags (6), but such variation must also be constrained to prevent the accumulation of poorly functional or autoreactive Igs. The progressive removal of such auto-reactive receptors from the human repertoire during B cell development has been demonstrated by sorting B cell progenitors from bone marrow along with mature peripheral B cells from the same individuals and subsequently cloning, expressing, and characterizing the autoreactivity of Igs expressed at different stages of maturation (7). These experiments revealed that the majority of Ig rearrangements in early immature B cells are autoreactive and are subsequently removed from the repertoire. We hypothesized that out-of-frame IgH rearrangements in pro-B cells, which are not expressed as functional IgH proteins but are carried in peripheral B cells that then have productively rearranged the alternate IgH locus, could also be used to gain insights into the original population of IgH rearrangements before selection. Thus, comparison of productive IgH rearrangements that have progressed through B cell development and selection with out-of-frame IgH rearrangements that were not subjected to selection pressure should elucidate aspects of the selection processes that shape human germline Ig repertoires. Because both productive and out-of-frame IgH rearrangements can be found in peripheral B cells, this approach facilitates efficient and extensive IgH repertoire sampling from multiple individuals and circumvents the need for collection of bone marrow samples, which are more difficult to obtain and typically contain fewer cells for analysis. We have used deep sequencing of the IgH rearrangements found in naive B cells to

*Department of Immunology, University of Washington, Seattle, WA 98195; †Adaptive Biotechnologies Corp., Seattle, WA 98109; ‡Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109; §Department of Medicine, University of Washington, Seattle, WA 98195; and ¶Program in Immunology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Received for publication May 8, 2012. Accepted for publication July 16, 2012.

This work was supported by the Bill and Melinda Gates Foundation (Collaboration for AIDS Vaccine Discovery Grant OPP37902).

K.L. isolated DNA from purified naïve B cells for sequencing and drafted the manuscript and figures. M.W.M. performed computational analysis of the deep sequencing data and contributed to the drafting of figures. H.S.R. oversaw the design of the IgH sequencing assay and collection of IgH sequences at Adaptive Biotechnologies Corp. and contributed to data analysis and interpretation. P.D.G. contributed to the study design and data interpretation and edited the manuscript.

Address correspondence and reprint requests to Dr. Philip D. Greenberg, University of Washington, Box 356425, 1959 NE Pacific Street, Seattle, WA 98195. E-mail address: ggreen@u.washington.edu

The online version of this article contains supplemental material.

Abbreviations used in this article: bnAb, broadly neutralizing Ab; GRAVY, grand average of hydropathicity; IMGT, Immunogenetics; MPER, membrane-proximal external region; RSS, recombination signal sequence.

Copyright © 2012 by The American Association of Immunologists, Inc. 0022-1767/12/$16.00
define better what sequences are expressed in human germline IgH repertoires and what sequences are consistently removed during B cell development. Identification of sequences consistently removed by selection processes can provide insights into the nature of Abs that might be associated with B cell autoimmunity. For example, it should be possible to determine if there is a selection bias against hydrophobic HCDR3 loops, a property commonly found in some characterized auto-antibodies (8–11).

The large number of unique IgH rearrangements in humans makes comprehensive analysis using conventional Sanger method-based capillary sequencing a technical challenge, as it requires manual cloning and sequencing of small pools of rearrangements, such that rare sequences in the repertoire are lost. The 454 high-throughput pyrosequencing technology has now made it possible to sample the Ig repertoire at depths of $\sim 10^3$ sequences per sample (12) generating reads of $\sim 300$ nt in length and permitting sequencing of entire Ig variable regions, and it is the technology of choice for analysis of mutated IgH sequences from memory B cells, which requires full-length reads including all three somatically mutated IgH CDRs. In contrast, Illumina deep sequencing by synthesis technology produces shorter sequence reads ($\sim 100$ bp) but enables much greater sequencing depth (millions of reads per sample). Moreover, by appropriately designing amplification primers, Illumina technology can permit acquisition of the entire IgH CDR3 region with flanking sequences sufficient for identification of the IgHV and J genes used as well as characterization of the D gene. Thus, for the unmutated IgH rearrangements expressed in naive B cells before encounter with Ag, it should be possible to deduce the full-length sequence of each rearrangement from these relatively short internal sequence reads at 10-fold greater sampling depth than is possible with 454 technology. Therefore, we have developed Illumina technology for deep sequencing of the unmutated IgH rearrangements in naive B cells as a means of maximizing the sampling depth of individual human germline IgH repertoires.

A major challenge for vaccine development in the 21st century has been to identify immunogens that can elicit Abs capable of neutralizing diverse strains of HIV (13–15). Such broadly neutralizing Abs (bnAbs) have been identified in some infected patients, but attempts to elicit these specificities by immunization have been unsuccessful. This in part reflects the facts that all known bnAbs contain extensive somatic mutations, whereas the computationally inferred unmutated ancestors bind HIV only weakly or not at all (16, 17). Thus, one potential limitation of current vaccines may be the inability of priming immunogens to engage and activate naive B cells expressing germline Abs that have the potential to become neutralizing after acquisition of somatic mutations through the affinity maturation process. For this reason, attempts are now being made to develop priming immunogens that can more efficiently engage the unmutated germline precursors of bnAbs predicted to be expressed by human naive B cells (18–20). However, these efforts currently rely upon the ability to infer computationally the most likely unmutated ancestors of bnAbs of interest. The computational prediction process is made difficult and imprecise due to the extreme number of mutations found in the HCDR3 region of most bnAbs that often obscures the identity of the genomic D gene used. Also, as N nucleotides are added randomly during rearrangement and are often subsequently mutated, it is not possible to infer accurately the original germline sequence. Furthermore, it cannot be determined if the computationally predicted germline ancestors of bnAbs can succeed through the selection process and actually will be found in human germline Ig repertoires where they could be engaged by priming immunogens. Extensive sampling of multiple human germline Ig repertoires with deep sequencing technology and assembly of the available repertoire provides an alternative to computational prediction of such unmutated Ab precursors. This new technology now allows the direct identification of rearranged germline Ig sequences within human naive repertoires that have the potential to match the sequence of known neutralizing Abs through the acquisition of somatic mutations. Sampling multiple individuals to identify the closest unmutated precursors of bnAbs that are commonly found in the population should facilitate the design and testing of priming immunogens that can specifically engage such precursors and increase the likelihood of being able to generate the desired high-affinity bnAb responses with subsequent boosting immunogens.

Our data suggest the majority of initial IgH rearrangements are removed from the repertoire during B cell development by selection processes that optimize the length of IgH CDR3 loops and selectively remove HCDR3 loops containing patches of hydrophobicity, which can be found in some auto-antibodies. Cationic residues were rarely encoded by unarranged IgHD genes, suggesting such codons had already been largely eliminated by evolutionary selection. Our results were consistent among multiple individuals, suggesting these observations likely reflect fundamental biases in the IgH rearrangement and selection processes. Additionally, to model the utility of deep sequencing for rational vaccine design efforts, we have used our data set to identify and compare related germline IgH precursors of the HIV-neutralizing 4E10 Ab acquired from four individuals by deep sequencing.

Materials and Methods

Isolation of genomic DNA from naive B cells

Whole blood samples (200 ml) were collected from normal human donors with approval from the Fred Hutchinson Cancer Research Center Institutional Review Board and with written informed consent. Samples were underlayed with Histopaque 1077, centrifuged, and PBMCs were collected from the interface. Naive B cells were purified from PBMC samples by magnetic separation using the Naive B Cell Isolation Kit II (Miltenyi). Cells were analyzed for purity on a FACSCanto II analyzer (BD Biosciences) after staining with fluorochrome-conjugated Abs recognizing the following human cell surface markers: IgM–allophycocyanin (BD Pharmingen), IgD–PE (BD Pharmingen), CD19–allophycocyanin-eFluor780 (eBioscience), CD2–PECy5 (eBioscience), and CD27–PECy7 (eBioscience). Genomic DNA (20–100 μg) was purified from each sorted naive B cell sample using the QIAamp DNA Blood Mini Kit (Qiagen).

Multiplex PCR amplification and sequencing of IgH CDR3 sequences

To generate the template library for the Illumina HiSeq machine, a multiplex PCR system was designed to amplify rearranged IgH loci from genomic DNA using 84 forward primers, each specific to a functional IgH V gene. The HiSeq cluster station solid-phase PCR. The Illumina HiSeq System generates reads of length 110 bp, which covers the entire range of CDR3 lengths, sequencing from the J to the V region. The six reverse J primers are positioned upstream of a unique tag for each J, allowing simple J gene segment identification. The 110 nt length includes 25–35 nt of J gene sequence, the full CDR3 region, and typically ~45 nt of V gene sequence, depending on the length of the CDR3 region in each rearrangement. This is sufficient to uniquely identify the specific V gene used in ~70% of sequences and to identify the V gene family and unambiguously define the 3’ end of the V gene in all acquired sequences. The amplification and sequencing used the same protocol as described for the similar TCRβ assay (21). Subsequent to the primary sequencing run, an additional short run was performed with a second primer designed to acquire the barcode sequence associated with each PCR product (from the 5’ end of the J primers), allowing acquisition of sequences from multiple independently barcoded samples in the same run. All sequencing was performed at Adaptive Biotechnologies Corp.
Preprocessing of genome analyzer sequence data

Raw HiSeq sequence data were preprocessed to remove errors in the primary sequence of each read and to compress the data. Approximately 20% of the sequences from the HiSeq were technical failures and were removed by a complexity filter. Each sequence was required to have a minimum of a 6-nt match to one of the 84 VH gene segments and one of the 6 JH gene segments. A nearest neighbor algorithm was used to collapse the data into unique sequences by merging closely related sequences to remove both PCR error and GA sequencing errors (21).

Data analysis

Data sets of nonredundant rearranged IgH sequences were analyzed using algorithms written in Java. Genomic Ig gene segments were identified by contiguous sequences exactly matching functional human IgH V, D, and J genes in the Immunogenetics (IMGT) database (22). IgH CDR3 regions were defined by the nucleotides found between conserved Cys and Trp codons. Bar graphs, histograms, and statistical values were generated with Microsoft Excel and the Apache statistics library (http://commons.apache.org/math/). The p values were determined by a two-tailed, equal variance Student t test for statistical significance. Error bars represent one SE of the population mean. IgH CDR3 grand average of hydrophobicity (GRAVY) scores were calculated using a freely available online tool at http://www.gravy-calculator.de/index.php. CDR3 amino acid alignments were performed using the National Center for Biotechnology Information online BLASTp tool at http://blast.ncbi.nlm.nih.gov/.

Results

Sampling of human germline IgH repertoires with deep sequencing

We adapted technology that we previously developed for TCRβ deep sequencing (21, 23, 24) to collect rearranged human germline IgH sequences from the genomic DNA of naive B cells that had been purified from normal human peripheral blood samples. Naive B cells were purified from PBMCs with Ab-coated magnetic beads. FACS analysis of the purified cells demonstrated that >97% were CD19lgD lgCD27+ (23), consistent with the naive B cell phenotype (Supplemental Fig. 1), and contained <0.9% contaminating CD19igD lgCD27 memory B cells, suggesting that a maximum of ∼1% of sequences could potentially contain somatic mutations. Primers containing linkers compatible with Illumina sequencing technology were designed to anneal to all human IgHV and IgHJ genes such that they specifically amplified, in a multiplex PCR reaction, the complete rearranged IgH VDJ junction, as well as unique portions of the 3’ end of each V gene and the 5’ end of each J gene to permit inference of their full germline sequences (Fig. 1). IgH V and J primers were designed with equal melting temperatures to produce semiquantitative amplification of all V-J pairs. Sequences containing all known human IgHV and IgHJ genes were detected, demonstrating that the PCR reaction has the potential to amplify all possible IgH rearrangements. As a result of amplification and extremely deep sampling, each non-redundant sequence was acquired an average of 20 times in each sample (range, 2–2973 copies per sequence), and, to avoid bias from potentially unequal amplification of primer pairs, we analyzed the set of unique (nonredundant) sequences from each sample. Because some V-J pairs may be amplified more efficiently than others, none of the results presented in this study rely on unbiased amplification of V-J primer pairs. PCR products were sequenced using an Illumina Genome Analyzer, collecting an average of 1.76 million IgH sequences from each of four individuals, which contained an average of 85,800 unique sequences per sample after filtering of the redundant identical sequences within each sample.

An algorithm was designed to identify the human IgHV and IgHJ genes used in each rearrangement based on matches with sequences in the IMGT database (22). The V and J gene segment boundaries were identified at the ends of contiguous sequence matches to known genes, after which a search was performed on intervening sequence for matches to known IgHD genes. Human IgHD genes range from 9 to 36 nt in length but can be found full-length or trimmed in IgH rearrangements. Therefore, for assignment of IgHD gene segments, the longest possible contiguous match to a known IgHD gene without mutations was chosen, as the probability of erroneously finding a match to any sequence by chance decreases with sequence length. For example, a random 4-nt sequence is expected to occur by chance in 1/44 = 0.3% of sequences, whereas a 5-nt match would occur by chance in 1/45 = 0.1% of sequences, resulting in uncertainty associated with these shorter matches. In our data set, 4- or 5-nt D gene segments were assigned to 2% and 4% of sequences, respectively, with the majority of D gene segment matches being longer and thus associated with lower probabilities of false assignment due to chance. Matches to known human IgHD gene segments of at least 4 consecutive nucleotides were found in >99% of sequences. Thus, the germline IgH V, D, and J genes could be identified within virtually all acquired IgH rearrangements, and the N nucleotide additions could then be defined between contiguous V, D, and J gene segment matches.

Optimization of human IgH CDR3 length during B cell development

The length of the IgH CDR3 loop is an important determinant of diversity in the naive B cell repertoire, as longer loops not only have greater potential for sequence variation but also can potentially reach into narrow antigenic pockets (25, 26). Experiments in mice have shown that the average length of the IgH CDR3 loop is increased during murine B cell development (27). However, long HCDR3 loops have previously been associated with Ab auto-reactivity and polyreactivity that is removed from the human repertoire during B cell development (7, 8, 28, 29). To investigate how IgH repertoire shaping during human B cell development influences the size of CDR3 regions expressed in mature naive B cells, we compared the distribution of CDR3 lengths in productive rearrangements that had survived B cell development and selection (i.e., the distribution of CDR3 lengths expressed in mature naive B cells) with the CDR3 length distribution found in out-of-frame IgH rearrangements, which could not be expressed as functional IgH proteins. Because out-of-frame IgH rearrangements are not expressed on the cell surface, they are not subjected to selection pressure and can be used to approximate the CDR3 length distribution found in IgH rearrangements before selection. Out-of-frame IgH rearrangements contained on average longer CDR3 loops than productive rearrangements (Fig. 2A), with mean
The difference in average IgH CDR3 loop length observed between unselected out-of-frame rearrangements and postselection productive rearrangements implies that the expressed human germline IgH repertoire is derived from a much larger population of initial productive rearrangements in pro-B cells that on average have longer CDR3 loops. This preselection productive IgH population must have contained not only the entire postselection repertoire but also additional rearrangements with longer CDR3 loops that did not survive the selection process. To calculate the percentage of initial productive IgH rearrangements that are removed from the repertoire by selection during B cell development, we first determined the minimum relative size of the preselection IgH population with the same mean and shape as the out-of-frame IgH CDR3 distribution that could contain within it the entire IgH population, assuming that the preselection HCDR3 distribution had the same mean and shape as the distribution observed in out-of-frame rearrangements, and required that the preselection histogram be scaled by a factor of 3.2 relative to the observed productive IgH population. The percentage of initial productive IgH rearrangements that ultimately survived selection was then calculated as the ratio of selected productive IgH rearrangements/preselection productive IgH rearrangements = 1/3.2 = 0.31, implying that at least 69% of initial productive IgH rearrangements in pro-B cells were removed during B cell development.

Tandem D gene segments in human IgH rearrangements

Although the vast majority of IgH rearrangements have been shown to have a canonical V–D–J structure, some studies have reported the potential use of tandem D gene segments giving rise to functional V–D–D–J rearrangements in mice and humans (28, 31–33), whereas others concluded there is no evidence of tandem D gene usage in humans (34, 35). We therefore examined our more extensive database sampling the naive B cell repertoire for the use of tandem IgHD genes as a potential source of diversity in human IgH rearrangements. To identify such events, we modified our algorithm to search for contiguous sequence matches to segments of known human IgHD genes within computationally assigned N regions of sequences in which one D gene match had already been identified. Because human IgHD genes vary in length and can be trimmed from the ends during rearrangement, true second D gene segments can be 1–36 nt in length. However, it was necessary to define a threshold for the length of contiguous matching sequence used for D gene segment identification to reduce the error introduced by short sequence matches that might have occurred by chance. For each N region interrogated, we generated a random control sequence of the same length and performed the same D gene segment search to determine the level of error expected from chance matches. We then compared this potential “noise” with the percentage of actual N region sequences containing contiguous matches to IgHD gene segments that were at or above the threshold length.

Using a 10-consecutive-nucleotide match as a minimum requirement for identifying second IgHD gene segments, we found matches in 1.7–3.3% of out-of-frame rearrangements from the four individuals, whereas the search was successful in only 0.05–0.14% of corresponding randomly generated control sequences of the same lengths (Fig. 3A; \( p < 0.0002 \)), providing confidence that rearrangements incorporating tandem D genes do occur in human pro-B cells before B cell maturation and selection. Using the same search criteria, tandem D genes were identified in only 0.28–0.54% of productive IgH rearrangements, indicating that the majority of preselection IgH rearrangements incorporating tandem D genes did not successfully progress through B cell development and selection, as this was significantly lower than in unselected out-of-frame rearrangements (\( p < 0.0008 \)). However, tandem D gene usage apparently does contribute to the diversity of the expressed germline IgH repertoire, as significantly more second D matches were found in productive IgH sequences than in corresponding randomly generated control sequences of the same lengths (0.04–0.07%; \( p < 0.0008 \)).

For the above analysis, we chose a minimum length of 10 consecutive matching nucleotides for identification of IgHD gene segments, as these relatively long contiguous matches were unlikely to occur by chance, allowing us to conclude with certainty that tandem D genes are found in productive human IgH rearrangements. However, this analysis likely underestimates the actual
The length of IgHD gene segments (0.002). Thus, the actual level of tandem D gene usage in productive and out-of-frame rearrangements (generated control sequences of the same lengths were again significant for both productive and out-of-frame rearrangements (p < 0.0001), suggesting preferential use of long D gene segments is a second mechanism contributing to the generation of elongated IgH CDR3 loops. Extensive N nucleotide addition is apparently a third mechanism that contributes to IgH CDR3 length, as productive IgH rearrangements with CDR3 loops >65 nt contained on average longer stretches of N nucleotide addition compared with the total productive IgH population (Fig. 3E, 3F; mean = 12.0 nt between D and J for long productive HCDR3 loops versus 6.1 nt for all productive HCDR3 loops, and means of 12.5 nt versus 7.3 nt at the V–D junction; p < 0.0001).

In agreement with the observed selection against IgH rearrangements containing long CDR3 loops, consistent biases against incorporation of long D gene segments and long N regions were found by comparing the unselected out-of-frame rearrangements with productive rearrangements that had progressed through B cell development and selection (Fig. 3E, 3F). Productive IgH rearrangements contained on average 6.1 N additions at the D–J junction and 7.3 N additions at the V–D junction compared with 8.9 and 9.9 N additions in out-of-frame rearrangements (Fig. 3F; p < 0.0001). Examination of D gene segment lengths found in IgH rearrangements revealed three subsets within the distribution (0–8 nt, 9–20 nt, and >20 nt), likely reflecting families of genomic D genes of similar lengths (D-1, D-7, and D-4 at 9–15 nt; D-5 and D-6 at 18–21 nt; and D-2 and D-3 at 27–36 nt). The shortest D gene segment population was increased and the longest reduced in productive IgH rearrangements after selection compared with unselected out-of-frame rearrangements (Fig. 3E, 3F), reflecting a bias against expression of longer D gene segments after selection (average D gene segment 17.4 nt for out-of-frame versus 14.3 nt for productive rearrangements; p < 0.0001). Thus, long D gene segments, long stretches of N nucleotide addition, and tandem D gene usage all contribute to the diversity of IgH CDR3 loops, but this diversity is constrained by selection against each of these features during B cell development.

**IgHD selection bias in human IgH rearrangements**

IgHD gene segments contributed on average 29% of the IgH CDR3 loop nucleotides in productive rearrangements containing a single D gene or 45% for productive rearrangements containing two tandem D gene segments, suggesting that IgHD gene segment(s) in each unique rearrangement make significant contributions to the Ag recognition properties of each resulting Ig. Thus, the distribution of D gene usage within the human germline Ig repertoire could significantly affect the identity and diversity of Ags recognized by naive B cells. Because IgHD genes can be incorporated into IgH rearrangements in all three reading frames, we investigated the relative representation of each D gene in each reading frame within our set of germline IgH rearrangements. Analysis of the distribution of D genes in out-of-frame IgH sequences from each individual, which should reveal biases imparted only by the rearrangement machinery and process, revealed biased usage of particular D genes (IgHD2-2, 3-3, and 3-22 together were found in >50% of sequences; Fig. 4A) but unbiased reading frame usage for each gene, suggesting that incorporation of D genes in each of
IgHD gene segments in each reading frame for out-of-frame (rearrangement. The number of unique IgH rearrangements containing analyses. Lines represent linear regression plots.

frames encoding stop codons were excluded from the IgH rearrangements; shown in Fig. 4C). Reading unique productive/number unique out-of-frame score (translated reading frame, the length (after rearrangement. For each germline IgHD gene individual containing segments of the same D gene reading frame (Fig. 4B) by the number of unique productive rearrangements containing hydrophobic IgHD gene segments, which is consistent with previous observations in mice (27).

FIGURE 4. Shaping of the IgHD gene usage distribution after IgH rearrangement. The number of unique IgH rearrangements containing IgHD gene segments in each reading frame for out-of-frame (A) and productive (B) rearrangements from one individual are shown. (C) The selection index for each IgHD gene reading frame was calculated by dividing the number of productive sequences (A) by the number of out-of-frame sequences (B) for each D gene reading frame from the same individual. Averages of four individuals are shown in (C); error bars represent 1 SEM.

the three possible reading frames during rearrangement was a random process. The IgHD reading frame distribution found in productive rearrangements, which reflects the cumulative biases imparted by the rearrangement machinery and selection during B cell development, by contrast did reveal clear preferences for particular reading frames of each IgHD gene in mature naive B cells after selection (Fig. 4B). Productive rearrangements were found on average 7 times more frequently than out-of-frame rearrangements in each sample, suggesting most pro-B cells that generated an out-of-frame rearrangement on one IgH allele (stochastically two of three rearrangements) were not rescued because a productive rearrangement of the alternate IgH allele either did not occur in time to promote development or could not pass the selection process.

To quantify the relative bias for expression of each possible IgHD gene reading frame after productive IgH rearrangement, we divided the number of unique productive rearrangements containing segments from each D gene reading frame (Fig. 4B) by the number of unique out-of-frame rearrangements from the same individual containing segments of the same D gene reading frame. The resulting postselection/preselection ratio for each IgHD reading frame we have termed the selection index, which describes the relative selection bias for expression of segments of each IgHD gene reading frame imparted during B cell development. The selection index for each IgHD gene reading frame differed between genes, ranging from 0 (D1-20 frame 3, which was never found in productive sequences) to 31 (D7-27 frame 2), but the overall indices were similar between individuals (Fig. 4C), suggesting that germline IgHD gene sequence attributes exist in each reading frame that consistently lead to enrichment/dimination in the human IgH repertoire during B cell maturation.

Shaping of expressed IgHD gene length and hydrophobicity
To identify features of IgHD gene reading frames associated with a positive or negative bias for expression in productive rearrangements, we looked for correlations between the selection index and specific attributes of IgHD gene reading frames. We plotted the length of each germline IgHD gene versus the selection index (from Fig. 4C), excluding frames encoding stop codons, and found germline IgHD gene length negatively correlated with selection index (Fig. 5A; linear regression $R^2 = 0.43$), which was largely anticipated based on the observed bias against expression of rearrangements using segments from intrinsically long IgHD genes.

To determine if a bias for or against expression of hydrophobic IgHD gene reading frames might exist, we plotted the calculated GRAVY score (36) versus selection index for each germline IgHD gene reading frame, again excluding reading frames encoding stop codons. Hydrophobicity negatively correlated with the selection index of IgHD gene reading frames (Fig. 5B; linear regression $R^2 = 0.40$), suggesting a bias against expression of productive rearrangements containing hydrophobic IgHD gene segments, which is consistent with previous observations in mice (27).

Identification of human germline IgH precursors with deep sequencing
For the elicitation of protective Ab responses that are highly divergent from any found in human germline repertoires, such as HIV-specific bnAbs, it may be necessary to develop priming immunogens capable of specifically engaging naive B cells that express appropriate unmutated germline Ig precursors to initiate the appropriate mutation/selection process (18, 19). The deep sequencing technology we have developed has the potential to facilitate these efforts by efficiently identifying bona fide human germline Igs as candidate precursors to any Ab of interest within multiple human repertoires. To demonstrate the potential utility of this new technology to inform rational vaccine design, we searched our germline IgH data for putative precursors of the membrane-proximal external region (MPER)-specific HIV-neutralizing Ab 4E10. We expected that identification of 4E10 precursors might be particularly challenging because the 4E10 HCDR3 loop contains multiple attributes that are preferentially removed from the germline IgH repertoire during development: it is long

FIGURE 5. Human IgH sequences using long and hydrophobic IgHD genes are reduced in frequency after rearrangement. For each germline IgHD gene translated reading frame, the length (A) and GRAVY score (B) are plotted versus the selection index (number unique productive/number unique out-of-frame IgH rearrangements; shown in Fig. 4C). Reading frames encoding stop codons were excluded from the analyses. Lines represent linear regression plots.
and contains patches of hydrophobicity. Furthermore, multiple investigators have reported reactivity of 4E10 with auto-antigens including cardiolipin and phosphatidylserine (10, 37, 38), suggesting that the most closely related precursors may often be removed from developing IgH repertoires as a result of auto-reactivity. The 4E10 Ab uses IgHV1-69, a CDR3 sequence of 20 aa between conserved C and W residues, and either IgHJ1 or IgHJ4 (39), which have identical amino acid sequences 3’ of the conserved W. We searched our germline IgH data for candidate 4E10H precursors with these attributes, which could in theory acquire point mutations during affinity maturation that would result in the 4E10H sequence. We found a total of 292 candidate precursors across four individuals from a total of 343,225 unique IgH rearrangements. These putative 4E10 IgH precursors represented 0.02–0.12% of unique IgH rearrangements from each individual. Candidate precursors were ranked by the extent to which they differed from the 4E10 sequence (requiring 16–33 nt mutations within the HCDR3 loop), and the closest precursor from each individual is shown in Fig. 6A. In addition to precursors with canonical VDJ rearrangements, VDDJ candidate precursors were also identified (data not shown) but were not as close to 4E10 as the sequences in Fig. 6A.

Although the potential 4E10 precursors identified in Fig. 6A represent the germline IgH sequences that are closest to 4E10H within each individual, they use a variety of D gene segments, and comparison of the CDR3 amino acid sequence of each closest precursor with the full set of candidate precursors from four individuals using the BLASTp algorithm did not reveal any significant homologies (data not shown), suggesting that these closest precursors were not commonly found in multiple individuals, and therefore may not be the best choice for the design and testing of 4E10 priming immunogens. However, by modifying the search criteria to allow the use of IgHJ3, which differs from J1 and J4 by one amino acid outside of the CDR3 region, we were able to identify putative 4E10H precursors within each individual that all contained IgHV1-69, IgHD3-22, IgHJ3, and a CDR3 loop of 20 aa, requiring 19–22 nt mutations within the CDR3 to match the 4E10 protein sequence (Fig. 6B). These precursors show significant amino acid homologies upon alignment, and using each of these sequences to search for similar rearrangements across individuals with the BLASTp algorithm identified a larger set of related sequences (≥80% amino acid identity upon alignment) that were shared between individuals (Fig. 6B, lower), suggesting that an appropriately designed immunogen may be capable of engaging one or more of these related precursors to initiate the necessary process of somatic hypermutation in most individuals. Our data set of human germline IgH sequences is freely available at https://clients.adaptivebiotech.com/project/home/Content/IgH%20Data%20Set/begin.view, and it can be used to identify putative germline precursors of any Ab of interest to inform rational vaccine design efforts.

**Discussion**

In this study, we have adapted Illumina deep sequencing technology to sample the human germline IgH repertoire from four normal blood donors. Although we anticipate that ultimately the database of human repertoires will be greatly expanded, the strict agreement of all of our data between the four individuals interrogated with minimal variation between individuals and the statistical significance of the comparisons suggests our conclusions have broad applicability to the construction of human B cell repertoires. Purification of naive B cells away from memory B cells allowed us to focus our analysis on the repertoire of unmutated IgH rearrangements found in mature B cells before Ag exposure, eliminating the potential bias introduced by the expansion and mutation of clones in response to foreign Ag exposures in each individual. In addition, because the sequences were unmutated, we were able unambiguously to identify genomic gene segments
within acquired IgH sequences. Our set of human germline IgH rearrangements may prove useful to other investigators as a control data set representing normal human naïve B cell repertoires, for comparison with other B cell populations, or for identification of germline precursors of known mutated human Abs of interest.

The average IgH CDR3 length naturally generated by the rearrangement machinery was found to be reduced during B cell development, consistent with suggestions from previous observations using much more limited data sets (7, 30, 32). This could result from a variety of requirements for expression of IgH chains and for the survival of developing and mature B cells, including proper and efficient protein folding, proper pairing with the surrogate L chain, subsequent pairing with a rearranged κ or λ L chain to generate a functional surface Ig, and finally the ability to overcome negative selection of autoreactive receptors. Long IgH CDR3 loops have been associated with self-reactive or polyreactive Abs (7, 8, 28, 29), and this may be a major reason for their removal from the repertoire during B cell development. However, retention of some long CDR3 loops would be expected to provide advantages to the host by increasing repertoire diversity due to the greater number of residues that can vary in sequence, potentially raising the number of foreign epitopes that can be distinctly recognized. Additionally, long CDR3 loops can facilitate binding to recessed epitopes in pathogens or the active sites of enzymes (40, 41). Our results suggest that human IgH rearrangement has evolved several mechanisms to generate longer CDR3 loops for the repertoire, including the use of long D gene segments, extended N regions, and tandem IgHD gene insertions, but only a small fraction persist with the diversity this might provide constrained by selection pressures during B cell development.

Using the out-of-frame HCDR3 length distribution to approximate the population of productive rearrangements generated before selection, we found that at least 69% of initial productive IgH rearrangements are removed from the repertoire during B cell development. A previous study demonstrated that 55–75% of Abs cloned from human early immature B cells were self-reactive based on in vitro binding assays, with the majority of this self-reactivity removed during development (7), which supports our calculation of the preselection repertoire size based on the extensive data set from peripheral B cells. Taken together, these results suggest self-reactivity, rather than nonfunctionality, may be the major reason for loss of productive IgH rearrangements during development.

There have been conflicting reports, based in large part on limitations of the available data sets, on the presence or absence of expressed human IgH rearrangements containing multiple IgHD gene segments (31–35). Our results provide highly statistically significant evidence that VDJ recombination in humans does generate rearrangements that contain tandem D gene segments, contributing to germline IgH diversity, and these sequences, although reduced in frequency during the process of B cell development, can still be identified within the mature human naïve B cell repertoire. Our ability to identify these rare rearrangements was enhanced both by the use of sorted naïve B cells bearing unmaturated IgH sequences that lack confounding somatic mutations within component gene segments and by the unprecedented number of sequences acquired from each individual for our analysis. Analyses of the VDJ recombination mechanism have suggested a “12/23 rule,” in which recombination is only allowed between IgH genes flanked by recombination signal sequences (RSS) containing 12 base spacer sequences and those containing 23 base spacer sequences. Because IgHD genes are flanked by RSS with 12 base spacers, and IgHV and J genes are flanked by RSS with 23 base spacers, this rule dictates that neither V–J nor D–D recombination should occur. However, our data indicate that recombination between RSS with 12 base spacers does occur at a low frequency in human IgH rearrangements, generating sequences that contain tandem IgHD gene segments. Comparison of D gene usage in the set of sequences containing tandem D genes with the population of standard VDJ rearrangements did not reveal any significant differences (data not shown), suggesting that D–D recombination occurs stochastically at a low level and is not due to unique sequence characteristics of particular D genes or their flanking sequences. To be certain that we identified bona fide tandem D genes within IgH rearrangements and were not misled by shorter sequences that may match D gene sequences only by chance, we restricted our analysis to include only those sequences with at least 10 nt of sequence matching a known D gene. At this level of stringency, we have a very high degree of confidence that we have identified true D–D rearrangements in ~3% of productive IgH rearrangements. However, the true rate of D–D recombination may in fact be much higher than this, as we could readily identify D gene segments shorter than 10 nt in IgH rearrangements, which were excluded from our tandem D gene analysis. Therefore, 3% is likely to be an underestimate of the overall rate of human IgH D–D rearrangement. Thus, our data confirm that the vast majority of rearrangements follow the 12/23 rule of VDJ rearrangement, but the frequency with which this “rule” is violated is greater than has been previously appreciated.

Our results show that the diversity of sequences found in the expressed germline IgH repertoire is determined in multiple stages. Positively charged residues were extremely rarely encoded by D gene sequences in all reading frames, suggesting that these had likely been eliminated from genomic IgHD coding sequences by selective pressure during evolution, and that reactivity of assembled Ig molecules with anionic auto-antigens such as DNA or cardiolipin may most often reflect residues added in N regions or acquired through somatic mutation. The rearrangement process uses particular D genes more frequently than others, with equal representation of all reading frames of each gene, as reflected by the distribution seen in out-of-frame rearrangements. Some variation in this initial IgHD gene usage existed between individuals (Supplemental Fig. 4), leading to diversity between individual IgH repertoires. These observed differences in preselection repertoires suggest variation between individuals in sequences proximal to IgHD genes (such as RSS sequences) that may affect utilization by the RAG recombination machinery, variation in the accessibility of particular IgHD genomic loci between individuals, and/or variation in the recombination machinery itself.

After rearrangement, the distribution of IgHD genes is further shaped during B cell maturation, resulting in the biased removal of rearrangements containing long and hydrophobic D genes from the repertoire. The observed selection biases were consistent between individuals, suggesting IgH CDR3 loops with these features were either not compatible with functional CDR3 loop structures or assembly with light chains, or tended to cause autoreactivity. Hydrophobic patches found in D gene reading frames may in fact often be incompatible with the function of CDR3 loops that need to bind a soluble Ag (42); however, a search of the UniProt database identified a functional human germline Ab (accession number Q9UL88) containing an untrimmed IgHD3-22 gene using the hydrophobic reading frame 2 that contains a 5-residue hydrophobic patch (IIVV1), indicating that the presence of such sequences does not preclude being functional. Of note, this unmutated Ab, which was cloned from a human fetal B cell, has characteristics that could make it both potentially protective and pathogenic, as it recognizes both foreign polysaccharides from *Pneumococcus pyogenes* and host cardiac myosin (43). Thus,
counterbalancing pressures for recognition of foreign Ag and removal of autoreactive Igs may explain why evolutionary selection has retained hydrophobic IgH gene reading frames, but selection during B cell development removes the majority of these sequences from the repertoire.

This study demonstrates the utility of deep sequencing technology to provide insights into human Ig repertoires. Using this technology to analyze large samples of multiple repertoires has identified consistent patterns, suggesting that the aspects of germline IgH selection presented here are likely to be fundamental processes well-conserved in normal humans. The results not only provide insights into basic biologic processes governing aspects of B cell development and maturation but can also have potential practical value for vaccine development. For example, current vaccines that protect against a variety of pathogens most often use identical immunogens for priming and boosting, demonstrating that the elicited high-affinity, somatically mutated protective Abs and their unmutated germline precursors can be engaged by the same structures. However, this approach may not be effective for all Ab responses; for example, the predicted unmutated germline ancestors of extensively mutated HIV-neutralizing Abs recognize HIV only weakly or not at all (16, 44), and this may in part explain why HIV envelope glycoprotein-based vaccines fail to induce broadly protective responses. These observations suggest that some HIV-specific bnAbs contain sequence attributes required for their specificity but rarely found in human germline Ig repertoires. Efficient induction of such responses by vaccination may require the design of priming immunogens that specifically engage and promote mutation/affinity maturation of germline Ig precursors initially lacking these sequence features, followed by boosting immunogens that can further direct the affinity maturation process to generate plasma cells secreting the desired mutated neutralizing Ab.

The bnAb 4E10 (45), specific for an MPER epitope of HIV, contains an HCDR3 loop that is relatively long, includes patches of hydrophobicity, and contains a centrally located cationic amino acid: three features that we have shown are at a disadvantage for expression in human naive B cells. However, of these attributes, only the HCDR3 length must be present in germline precursors, as the other amino acid sequence features can theoretically be acquired during affinity maturation. Our data set reveals that on average, 17% of HCDR3 loops expressed in human naive B cells are at least as long as those found in 4E10 (60 nt), suggesting that the length of the CDR3 loop alone does not preclude the presence of adequate numbers of candidate germline 4E10 precursors in human naive B cell repertoires. Moreover, our results indicate that although the closest germline 4E10H precursors were not shared between individuals, more distantly related candidate germline 4E10 precursors could be identified in multiple human germline Ig repertoires, suggesting the possibility that a priming immunogen could be developed to engender reproducibly these precursors in most individuals. Ultimately, generation of the bnAb may then be accomplished by design of a boosting immunogen preferentially chosen for its ability to engage reproducibly these precursors in multiple human germline Ig repertoires that have been identified.

Disclosures

H.S.R. is a founder of Adaptive Biotechnologies Corp., a for-profit company that performed the IgH deep sequencing assays that we have analyzed. M.W.M. is an employee of Adaptive Biotechnologies Corp. The other authors have no financial conflicts of interest.

References


Corrections


It has been brought to our attention that the published sequence of the HCDR3 for 4E10 in Fig. 6 was copied incorrectly by us, resulting in two residues being misrepresented in the figure. The last four residues should read FAHW not FDYW. A corrected figure with the correct sequence is shown below. The figure legend was correct as published and is shown below for reference.

In addition, as the number of mutations in 4E10 required from germline were calculated based on this error in copying the sequence, the second to last sentence in the first paragraph on page 3227 should read, “Candidate precursors were ranked by the extent to which they differed from the 4E10 sequence (requiring 19–33 nt mutations within the HCDR3 loop). . .”

The sequence copying error does not alter any of the conclusions in or insights from the paper. We thank Dr. Prabakaran Ponraj for recognizing and alerting us to this error.

![FIGURE 6](https://www.jimmunol.org/cgi/doi/10.4049/jimmunol.1390069)

**FIGURE 6.** Candidate 4E10H germline precursors identified in multiple individuals. (A) The 4E10 HCDR3 aa sequence is shown in the *top panel*, with sequences of the closest germline IgH rearrangements using VH1-69, JH1, or JH4 and a CDR3 of 20 aa from each individual shown in the *bottom panel*. To the *right*, the number of nucleotide changes needed for each sequence to match the 4E10 AA sequence is shown. (B) Germline IgH rearrangements using VH1-69, D3-22, and JH3 with a CDR3 of 20 aa between conserved C and W residues from all four individuals are shown. The closest precursors to 4E10 of this type from each individual are shown above the consensus sequence, with additional related precursors identified by BLASTp search shown below. Letters represent amino acids, color coded according to chemical type (see the key); dashes represent gaps in the alignment; asterisks (*) represent amino acids of the same chemical type in the consensus; and X represents any amino acid.
Supplementary Materials

**Fig. S1.** Purification of naïve B cells from isolated human PBL samples. Naïve B cells were purified from normal human peripheral blood samples by depletion of CD2, CD14, CD16, CD27, CD36, CD43, and CD235a-expressing cells, using Ab-coated magnetic beads, from PBMCs that had been isolated by density gradient centrifugation. PBMC (pre-sort) and post-sort samples were stained with fluorochrome-conjugated antibodies specific for human cell surface markers: IgM-APC, IgD-PE, CD19-APCeFluor780, CD2-PECy5 and CD27-PECy7, and analyzed by flow cytometry. Representative plots are gated on events of lymphocyte size based on forward and side scatter; numbers represent the percentage of events falling within each gate.
Fig. S2. Consistent IgH CDR3 length and selection bias between individuals. The average IgH CDR3 lengths are shown for productive (black bars) and out of frame rearrangements (grey bars) in each individual.
Fig. S3. Tandem IgHD gene segments are used in human IgH rearrangements, are reduced in frequency after selection, and contribute to CDR3 length. The N regions of sequences with one human IgHD gene segment already identified were scanned for the presence of additional D gene segments based on a match of 9 or more consecutive nucleotides to known IgHD genes. Matches to random sequences of the same lengths were used to estimate the maximum error inherent to each search. The average percentage of all out of frame rearrangements, all productive rearrangements, and the subset of productive sequences with CDR3 length >65 nt containing two D gene segments (D2) from 4 individuals are shown. Error bars represent one standard error of the mean.
Fig. S4. IgHD gene usage in 4 individuals. The number of unique IgH sequences containing each D gene segment in each reading frame (frame 1, white; frame 2, striped; frame 3, black) are shown for out of frame IgH rearrangements from each of 4 individuals.