The Restricted D_H Gene Reading Frame Usage in the Expressed Human Antibody Repertoire Is Selected Based upon its Amino Acid Content

Jennifer Benichou, Jacob Glanville, Eline T. Luning Prak, Roy Azran, Tracy C. Kuo, Jaume Pons, Cindy Desmarais, Lea Tsaban and Yoram Louzoun

*J Immunol* published online 29 April 2013
http://www.jimmunol.org/content/early/2013/04/28/jimmunol.1201929
The Restricted $D_H$ Gene Reading Frame Usage in the Expressed Human Antibody Repertoire Is Selected Based upon its Amino Acid Content

Jennifer Benichou,*† Jacob Glanville,+† Eline T. Luning Prak,‡ Roy Azran,§,* Lea Tsaban,∥ and Yoram Louzoun∥,*

The Ab repertoire is not uniform. Some variable, diversity, and joining genes are used more frequently than others. Nonuniform usage can result from the rearrangement process, or from selection. To study how the Ab repertoire is selected, we analyzed one part of diversity generation that cannot be driven by the rearrangement mechanism: the reading frame usage of $D_H$ genes. We have used two high-throughput sequencing methodologies, multiple subjects and advanced algorithms to measure the $D_H$ reading frame usage in the human Ab repertoire. In most $D_H$ genes, a single reading frame is used predominantly, and inverted reading frames are practically never observed. The choice of a single $D_H$ reading frame is not limited to a single position of the $D_H$ gene. Rather, each $D_H$ gene participates in rearrangements of differing $CDR3$ lengths, restricted to multiples of three. In nonproductive rearrangements, there is practically no reading frame bias, but there is still a striking absence of inversions. Biases in $D_H$ reading frame usage are more pronounced, but also exhibit greater interindividual variation, in IgG + and IgA + than in IgM + B cells. These results suggest that there are two developmental checkpoints of $D_H$ reading frame selection. The first occurs during $VDJ$ recombination, and the second implies that $D_H$ reading frames are subjected to differential selection. Following these checkpoints, clonal selection induces a host-specific $D_H$ reading frame usage bias. The Journal of Immunology, 2013, 190: 000–000.

Antibodies are proteins produced by B cells. Abs consist of two $H$ chains and two $L$ chains. $A_b$ $H$ chains and $L$ chains consist of $C$ and $V$ regions, which are so named because they vary in their level of sequence diversity when different Ab molecules are compared. The $V$ regions that encode the greatest diversity correspond to the regions in the $A_b$ that are important for binding to a vast array of Ags. Ab diversity is achieved through a series of somatic mechanisms that produce many different sequences, but conserve genetic space. These diversification mechanisms include recombination of $V$, $D$, and $J$ gene segments [$V(DJ)$ recombination], junctional modifications between the rearranged gene segments, pairing of different $H$ and $L$ chains, and somatic hypermutation (SHM) [reviewed in (1, 2)]. $V(DJ)$ recombination occurs in an ordered and stage-specific fashion in the bone marrow, with $H$ chain rearrangement producing a $V$ region that contains juxtaposed $V_H$, $D_H$, and $J_H$ gene segments.

The fate of the B cell depends in large part on the specificity of its BCR. Self-reactive B cells must be edited, killed, or inactivated to maintain self-tolerance (3, 4). B cells that respond to pathogens are, instead, activated, clonally expanded, and undergo differentiation into specialized effector cells. During an immune response, Abs can undergo further sequence modification to optimize their effector functions (isotype switching from IgM to a different $H$ chain C region such as IgG or IgA, while keeping the same $V_H$ region). Abs of mature B cells can also undergo SHM. SHM, coupled with selection for the B cells that bind to a particular Ag with the highest affinity, results over time in the successive improvement in affinity for the Ag, a process termed affinity maturation.

Within the Ab $V$ region, there are more conserved and more variable sequences referred to as framework regions and CDRs, respectively (5). The CDRs form loops that are important for Ag binding. Among the CDRs, CDR3 is the most hypervariable in sequence because it encompasses the junctions between the recombining $V_H$, $D_H$, and $J_H$ gene segments. The position of $D_H$ in the sequence often brings it to the center of the Ab combining site. $D_H$ gene usage has been proposed to be different in autoimmune (6). The addition and deletion of nontemplated nucleotides at the junctions between the recombining gene segments allow the $D_H$ segment (which is flanked on one side by the $V_H$ gene segment and on the other by the $J_H$ gene segment) to be read in one of three forward-facing reading frames (RFs). As long as the junctional modifications at the $D$-$J$ side of the rearrangement return to the +1 RF in the $J_H$ gene segment, the rearrangement is potentially functional. Despite the availability of three forward-facing RFs, the usage of RFs is biased. For example, studies in mice (7, 8) have shown that there is selection against RFs containing a stop

---

© 2013 by The American Association of Immunologists, Inc. 0022-1767/13/$16.00

www.jimmunol.org/cgi/doi/10.4049/jimmunol.1201929

Received for publication August 2, 2012. Accepted for publication March 25, 2013.

This work was supported by the Lupus Research Institute and National Institutes of Health Grant R56-AI-090842 (to E.T.L.).

Address correspondence and reprint requests to Prof. Yoram Louzoun, Bar Ilan University Campus, Ramat Gan, 52900, Israel. E-mail address: louzouy@math.biu.ac.il

The online version of this article contains supplemental material.

Abbreviations used in this article: DRF, $D_H$ gene reading frame; IF, in-frame; OF, out-of-frame; RF, reading frame; SHM, somatic hypermutation.

Copyright © 2013 by The American Association of Immunologists, Inc. 0022-1767/13/$16.00
Materials and Methods

Study subjects

Twelve apparently healthy adult subjects (see Table I for demographic characteristics) were recruited for high-throughput sequencing using the 454 platform. Two 45-ml blood draws were collected in heparin tubes from each subject at a single time point. Mononuclear cells were isolated using Ficoll-Paque Plus (GE Healthcare), and then sorted by flow cytometry into naïve (CD20+, CD27−) and memory (CD20+ CD27+) populations. Informed consent was obtained from all donors. This work was performed in accordance with an Institutional Review Board–approved protocol at Pfizer.

Two apparently healthy adult subjects were also recruited for high-throughput sequencing using the Illumina platform at a single time point (Table I). A total of 25 cc venous peripheral blood was drawn in sodium EDTA tubes from each of the two subjects and sequenced using the Illumina platform. Informed consent was obtained for each subject and used to fill out a medical questionnaire. This work was performed in accordance with an Institutional Review Board–approved protocol at the University of Pennsylvania.

Target amplification and 454 sequencing

Unbiased amplification of repertoire was performed by 25 cycles of 5'RACE, using individual isotype-specific reverse primers. Primers were optimized for efficiency, fidelity, and completeness of repertoire recovery by informatic screening and analysis, and high-throughput sequencing of recovered products. The degree of germline-dependent amplification bias was assessed by comparing amplified products of stimulated naïve B cell pools to direct sequencing of the same pools. Cycle-dependent effects on diversity estimates were evaluated by high-throughput sequencing. All products received multiplex identifiers (barcodes) to allow unambiguous identification of all products by sequence analysis in subsequent processing steps. Multiplex identifiers differed by at least 3 bp from any other multiplex identifier sequence, and only reads with exact matches were included in the analysis. Products were sequenced with 454 titania sequencing. Quality sequencing was assessed by keypass control. Sample quality control was confirmed by demultiplexing and VnH segment genotype. Sequencing depth was determined by diversity estimate rarefaction, population simulations of uniform profile stabilization as a function of sequencing depth. A detailed discussion of the sequencing methodology has been described previously (30).

Ab DNA CDR3 analysis by Illumina sequencing

PBLs were enriched over Ficoll-Hypaque, and CD19+ cells were isolated by magnetic bead separation (Miltenyi Biotech). Genomic DNA was extracted using a Puregene Kit (Qiagen, Valencia, CA), and 800 ng was used for IgH CDR3 amplification and library construction. IgH CDR3 V regions were amplified and sequenced, as described in Larimore et al. (31). Briefly, a multiplexed PCR method was employed to amplify rearrangements of IgH sequences at the genomic level, using seven VnH segment primers (one specific to each VnH segment family) and six JnH segment primers (one for each functional JnH segment). Reads of 110 bp extending from the JnH segment, across the NDN junction, and into the VnH segment were obtained using the Illumina HiSeq sequencing platform. The ImmunoSEQ assay was used for the sequencing. The BCR CDR3 region was defined according to the IMGT convention (32), beginning with the second conserved cysteine encoded by the 3' portion of the VnH gene segment and ending with the conserved phenylalanine encoded by the 5' portion of the JnH gene segment. The resulting sequences were analyzed for DnH usage, orientation, and RF, as described in the section on DnH detection. A total of 12,948,437 sequence reads representing 47,358 unique sequences, of which 84% were productive, was analyzed for subject 2, and 12,851,153 sequence reads representing 20,374 unique sequences, of which 84% were productive, were analyzed for subject 3.

Germline VnH, DnH, and JnH genes used for sequence composition

Human Ig germline sequences for VnH, DnH, and JnH genes of B cells were extracted from the IMGT database (33). In our study, we used 34 germline sequences of DnH, categorized under functional genes that included open RF, 13 sequences of JnH genes, and 188 sequences of VnH genes.

The 454 sequence analysis

We used two different approaches to analyze the DRF usage. In the first approach, we used all unique sequences. In the second approach, we attempted to detect clones by clustering together sequences with similar CDR3 sequences, to minimize the effect of potential biases in the sequence copy numbers. Both approaches led to similar results.
Sequences were grouped into clones using a two-step approach. First, the germline VH and JH of each sequence were determined by aligning all possible germline VH and JH (based on the IMGT germline library) against the sequence, finding the highest number of overlapping nucleotides, and assuming that no deletions or insertions occurred. A full alignment using BLAST produced similar VH and JH assignments for all tested sequences that were classified clearly enough in the alignment.

Next, to count the clones, we grouped all sequences according to their VH and JH usage as well as the distance between VH and JH, because SHMs usually do not produce additions or deletions of nucleotides (a detailed example of clone detection can be found at: http://peptibase.cs.biu.ac.il/homepage/Lymphocyte_clone_detection.htm). Thus, every clone emerging from the same founder cell should have the same distance between VH and JH. We then took all of the sequences with the same VH, JH, and distance between VH and JH and grouped them using a phylogenetic approach. The distance between VH and JH was computed by positioning the IMGT germline VH and JH genes on the observed sequence and determining the distance between the last nucleotide of VH and the first nucleotide of JH. All the sequences with equal VH, JH, and distance were aligned together with an artificial sequence composed of the germline and gaps between them. Within each group, the sequences were aligned (using MUSCLE 3.6) (34), and a phylgetic tree was built using maximum parsimony (35) and/or neighbor-joining (36) methods (from the PHYLIP 3.69 program package). We then parsed this tree with a cutoff distance of four mutations into clones. Thus, a clone was defined as a set of sequences that are similar one to each other, up to a distance of four mutations.

**DH detection**

We have computed the maximal similarity between each DH germline gene segment in each RF and all the possible subsequences of same length in the region spanning the last 30 nt before the 3’ end of the VH gene. In both methods, purines and pyrimidines had a two-thirds probability of being mutated into a base in the same founder cell should have the same distance between VH and JH. We then took all of the sequences with the same VH, JH, and distance between VH and JH and grouped them using a phylogenetic approach. The distance between VH and JH was computed by positioning the IMGT germline VH and JH genes on the observed sequence and determining the distance between the last nucleotide of VH and the first nucleotide of JH. All the sequences with equal VH, JH, and distance were aligned together with an artificial sequence composed of the germline and gaps between them. Within each group, the sequences were aligned (using MUSCLE 3.6) (34), and a phylgetic tree was built using maximum parsimony (35) and/or neighbor-joining (36) methods (from the PHYLIP 3.69 program package). We then parsed this tree with a cutoff distance of four mutations into clones. Thus, a clone was defined as a set of sequences that are similar one to each other, up to a distance of four mutations.

**Validation set production**

In silico, we generated a set of 10,000 sequences by random V-D-J joining (with equal probability for each DH gene). In this dataset, the DH gene could be inserted in forward or backward orientation. We then added mutations between each pair of genes (V-D and D-J) to create the following: 1) replacements of 1–3 nt at each inner edge of the DH gene, and 2) replacement of 1–3 nt in random positions in the DH gene. In both methods, purines and pyrimidines had a two-thirds probability of being mutated into a base in the same founder cell should have the same distance between VH and JH. We then took all of the sequences with the same VH, JH, and distance between VH and JH and grouped them using a phylogenetic approach. The distance between VH and JH was computed by positioning the IMGT germline VH and JH genes on the observed sequence and determining the distance between the last nucleotide of VH and the first nucleotide of JH. All the sequences with equal VH, JH, and distance were aligned together with an artificial sequence composed of the germline and gaps between them. Within each group, the sequences were aligned (using MUSCLE 3.6) (34), and a phylgetic tree was built using maximum parsimony (35) and/or neighbor-joining (36) methods (from the PHYLIP 3.69 program package). We then parsed this tree with a cutoff distance of four mutations into clones. Thus, a clone was defined as a set of sequences that are similar one to each other, up to a distance of four mutations.

**Results**

Whereas the RF usage of VH and JH is constrained, the RF of DH is flexible because nucleotide additions or deletions are present at both ends of the rearranged DH gene segment. Furthermore, DH genes can potentially undergo inversion. Thus, DH segments can be read in up to six RFs. The presence of multiple possible RFs increases the variability of the CDR3. In this study, we show that the DH gene actually uses a limited number of RFs.

The analysis of DH gene sequences in the Ab repertoire is not easy because some DH genes are similar to each other, and nipping of the DH genes by the nonhomologous end-joining machinery can result in very short DH gene sequences. Thus, the robust identification of DH gene segments from CDR3 sequences involves a trade-off between sensitivity and specificity.

This analysis required the development and validation of a robust computational method for the identification of DH gene segments and their RFs from high-throughput sequencing data. We have performed several analyses to rule out certain technical artifacts in the sequencing or sampling methodology. We have further validated the precision of the DH gene determination using computer simulations of DH rearrangements and found the results to be highly precise for long DH genes. We have tested our computational methodology on artificial datasets, to ensure that the results are not due to artifacts of the computation. The code for V(D)J detection and the code to analyze and detect DH genes can be requested from the authors.

To rule out potential biases that could have been introduced as a result of the sample type, IgH amplification process, or sequencing platform, we analyzed the DRF usage using two completely different methodologies. The first method used RNA extracted from sorted B cell subsets, amplification of cDNA by 5’RACE, and pyrosequencing (by 454). The second method employed genomic DNA from CD19* -enriched PBLs, amplification using mixtures of VH and JH primers, and Illumina sequencing. The results with both approaches were very similar for the in-frame (IF) sequences. In this analysis, we have sequenced the IgH repertoire of multiple human subjects to ensure that the observed DH selection is reproducible and robust.

**Subjects and 454 IgH sequences**

We have sequenced the Ab H chain repertoire from 12 healthy human adult volunteers (Table 1). For each subject, we sorted CD20* lymphocytes into naive (CD27-) and memory (CD27+) subsets. We then applied a RACE protocol to separately amplify IgM, IgG, and IgA BCRs from each sample, using isotype-specific primers at the C region. The resulting sequences were compared with all germline VH genes at all possible positions. Sequences were discarded from the analysis if a VH gene could not be detected with a high enough accuracy (at least 70% overlap with germline VH and at least 120 nt of VH sequence length were required). Nearly all sequences had a much better fit than 70%. VH genes that were highly related (e.g., alleles of the same VH gene) were grouped into a single gene. We then compared the sequences downstream of the identified VH with all possible germline JH genes, and assumed that the JH gene could not start >50 nt before the 3’ end of the VH gene. We found that practically all sequences for which a VH gene could be identified, the JH gene could also be detected with a high enough accuracy (>70% similarity to germline VH and JH). Note that even if the precise VH gene could not be identified, its final position could almost always be known precisely, because similar VH genes have similar lengths. Given the position of the VH and JH, we extracted the sequence regions that included the DH gene segment(s), starting with the sequence that was 30 nt before the end of the VH germline gene and ending with the 3’ end of the germline JH sequence. We looked for the best fit to all germline DH genes in the IMGT database. The optimal fitting DH gene was given a score representing the fraction of the germline DH gene nucleotides fitting the observed sequence (see Fig. 1). The position of the DH gene was then determined relative

<table>
<thead>
<tr>
<th>ID</th>
<th>Gender</th>
<th>Ethnicographics</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>91</td>
<td>F</td>
<td>Caucasian</td>
</tr>
<tr>
<td>104</td>
<td>F</td>
<td>Hispanic</td>
<td>28</td>
</tr>
<tr>
<td>130</td>
<td>M</td>
<td>Hispanic</td>
<td>39</td>
</tr>
<tr>
<td>136</td>
<td>M</td>
<td>Caucasian</td>
<td>54</td>
</tr>
<tr>
<td>158</td>
<td>M</td>
<td>Asian</td>
<td>25</td>
</tr>
<tr>
<td>182</td>
<td>F</td>
<td>Caucasian</td>
<td>27</td>
</tr>
<tr>
<td>219</td>
<td>F</td>
<td>Asian</td>
<td>44</td>
</tr>
<tr>
<td>251</td>
<td>M</td>
<td>Caucasian</td>
<td>29</td>
</tr>
<tr>
<td>272</td>
<td>M</td>
<td>Asian</td>
<td>57</td>
</tr>
<tr>
<td>273</td>
<td>M</td>
<td>Caucasian</td>
<td>27</td>
</tr>
<tr>
<td>275</td>
<td>M</td>
<td>Hispanic</td>
<td>50</td>
</tr>
<tr>
<td>276</td>
<td>F</td>
<td>Hispanic</td>
<td>47</td>
</tr>
<tr>
<td>277</td>
<td>F</td>
<td>Caucasian</td>
<td>49</td>
</tr>
<tr>
<td>278</td>
<td>F</td>
<td>Caucasian</td>
<td>22</td>
</tr>
</tbody>
</table>
to the end of the region where the sequence was identical to the corresponding germline sequence and relative to the beginning of the germline JH sequence. The position determines the DRF relative to the beginning of the VH germline sequence.

Note that many previous algorithms have been created for the detection of VH, D, and JH [e.g., among many others, Soda (37), IHMune (38), and JOIN SOLVER (23)]. However, our algorithm allowed us to precisely check the precision on our own dataset, and to perform large-scale off-line computations.

Methodology validation

We first tested whether our D segment detection methodology properly identifies the D germline gene and the corresponding DRF. We have produced artificial V-D-J sequences and have introduced random mutations in these sequences (see Materials and Methods). We have then checked the fraction of cases in which the algorithm properly classified the results.

We have limited the analysis to contain up to three mutations in the body of the D gene and three nucleotide changes/removed in the 3’ and 5’ ends of the D gene, because sequences with more mutations than that had a too high error level and where not used, as will be further discussed. Within the mismatches, we checked the type of mismatch that occurred: either an error with respect to the RF or a misidentification of the D gene, or both. Finally, we checked whether the error was due to failure to detect an inverted D gene. We found that the error level is a function of the number of mutations in the D segment (nucleotide differences compared with the germline D sequence) and the number of nucleotides that are added and/or removed from the D gene. At or above a level of 75% overlap (the percentage of nucleotides overlapping between the query sequence and the germline D sequence), <10% of the query sequences are mismatched (Fig. 2). Note that even among these 10%, most errors are due to a misclassification of the D gene as a similar gene, and not due to errors in the assignment of the DRF (compare Supplemental Fig. 1A with 1B). This misclassification occurs because some D genes have highly similar nucleotide sequences (e.g., D-4-11 versus D-4-17). This precision is much better than currently used algorithms such as SoDA, JS, and V-Quest (50–73% precision for 2–6 mutations as used in the simulation). However, the comparison is not fair, because in this study we discard a large fraction of the sequences.

Even a low mutation frequency results in a nonnegligible error rate for highly homologous D genes (Fig. 2). Therefore, we have restricted our analysis to D genes that can be identified with a precision of at least 0.75. Thus, this analysis excludes some of the shorter D genes that practically never reach this level of precision (see Supplemental Fig. 1C, 1D).

Absence of inverted DRF

We first checked the DRF usage in the total sample (without taking into account the specific D gene used). We then computed how many unique clones (see Materials and Methods) used each RF for each D gene. The inverted DRF usage frequency is lower than the precision level of the methodology (<1% of the unique clones contained a possible inverted DRF, compared with an error level of 5%). Thus, sequences that are identified as having potential DH inversions may not really contain inverted D segments. Indeed, for each D gene found to be in the inverted DRF, there was also a reasonable fit to a D gene in a forward DRF (not always in the same D gene). Thus, we conclude [as was previously suggested based upon smaller datasets (22)] that the use of inverted D segments, if it occurs at all, is very infrequent in the IgH repertoire of healthy human adults.

The frequencies at which D genes are used in our sample vary widely among D genes. One caveat is that we are only analyzing D genes in which the fraction of original nucleotides maintained in the rearranged sequence is >75%. This selection induces a bias toward long D genes that can be classified properly with a better accuracy (Supplemental Fig. 1).

We have repeated the analysis separately for each D gene. When averaging over each D separately (instead of overall clones), some inverted D genes are observed in rare short D genes (Fig. 3A). However, in short D genes, a small number of fitting nucleotides can be enough to ensure a 75% match. As mentioned above, for each candidate inverted D, a reasonable candidate forward D can be found. One can thus conclude that in very large cohorts, if inverted D exist, they are very rare and limited to short D genes. We repeated the analysis using all the sequences and not only clones, to make sure that this bias is not a result of artifacts from our clone detection methodology (Fig. 3B). When using all the sequences, there are again practically no inverted D genes.

Note that most inverted D genes do not contain stop codons. Thus, there is no a priori reason that these DRFs should not be used. Their absence seems to highlight the presence of mechanistic differences in the inverted and forward rearrangements.
The relative DRF usage distribution can be treated as a $3 \times 18$ frequency matrix (3 DRF and 18 of 34 DH genes that had $\geq 10\%$ of classified sequences) in either the DH or DRF classification, we repeated the analysis and increased the precision level requested, until only sequences with a 100% precise classification were used (i.e., the full DH gene is observed in the original sequence). Note that in such a case, the error level is negligible, and a limited number of sequences are used (most of these sequences have only insertions and no deletions at the VD and DJ junctions). Even under these conditions, the average DRF usage did not change (Fig. 3C). Thus, the nonuniform DRF usage seems to be a real feature of the sequences, and not an artifact of the methodology. In contrast, one may worry that because we are looking only at sequences with a good enough classification of the DH gene, the results may be biased toward a given DRF. However, the comparison with the DH gene is only performed with the germline and is not affected by additions around the DH gene, and the DRF preference is also observed in CDR3 sequences where the distance between VH and DH, and between DH and JH is positive, where the majority of sequences are taken into account (Supplemental Fig. 3).

**Comparison of DRF in IgM*+, IgG*, and IgA* B cell subsets**

To determine whether DRF skewing could be mediated by selection in the periphery, we compared DRF usage in naive (CD27−) and memory (CD27+) sorted B cell subsets. We amplified IgM rearrangements (from cDNA) of the naive B cells, and, to further analyze the memory B cell pool, we amplified IgM, IgA, and IgG rearrangements from the CD27+ fraction (see Materials and Methods).

The relative DRF usage distribution can be treated as a $3 \times 18$ frequency matrix (3 DRF and 18 of 34 DH genes that had $\geq 10\%$ sequences in most tested donors) in which the sum of each column is 1. We rearranged this matrix as a vector and computed the correlation of this vector between all samples (12 subjects, 4 compartments: naive [IgM] and memory [CD27+] IgM, IgG, or IgA). We observed a very similar distribution among all samples, with the maximal similarity being among technical repeats of the same sample, followed by different isotypes from the same donor (corr = 0.8) (Fig. 4A). But even between individuals, the corre-


D genes typically use a single reading frame

---

**FIGURE 4.** DRF usage in different donors and isotypes. (A) Average correlation between DRF usage within and between donors. The correlation is computed as the Pearson correlation of the frequency of each DRF in each DH gene among samples. The highest correlation is among technical repeats, followed by the correlation between different isotypes in the same donor and similar isotypes among subjects. The lowest correlation is between samples of different isotypes in different subjects. Technical repeats are from the same donor and the same isotype. Same donor represents the same subject and different isotype. Same isotype represents the same genotype in different subjects. All samples represent different isotypes and donors (i.e., those are all sample combinations, except for the ones considered in the previous columns). (B) Average SD of DRF usage as a function of isotype. A uniform distribution would represent a SD of 0, whereas a single DRF used for each DH gene leads to a SD of 0.57. The SD is lowest for naive IgM, followed by memory IgM and then IgG. IgA has a lower variance than IgG. The differences between all groups are significant (t test, p < 1.e-10), except for the memory IgM and the IgA. The distribution is over all samples within the same isotype. (C) Average DRF usage among all samples (donors and isotypes). Each column is a sample, and the y-axis is the fraction of sequences using each DRF. We have first averaged the results for each DH gene and then averaged over DH genes. Thus, rare DH genes are overexpressed in this analysis. One can clearly see that the average pattern is very similar among all samples. The presence of inverted DRF is fully due to short DH genes. However, in these sequences, the determination of the DRF is error prone.

---

**FIGURE 5.** Frequency of position of end of DH for each DH gene. Each column is a DH gene. The shades represent frequency, as shown in the grayscale bar. One can clearly see in most DH genes jumps of three in the position, but, beyond the clear selection for a DRF, the position distribution is quite wide.
cDNA 454 samples and the DNA Illumina samples. The fit shows again that the DRF usage is not an artifact of the sequencing methodology.

Consistent with selection, the OF DRF usage was practically uniform for all DH genes in the forward direction. Indeed, the OF variance is not very different from what was expected due to random chance in most DH genes. Conversely, in the IF rearrangements, the observed bias is similar to the one observed in the cDNA sequence analysis described using 454 sequencing (Fig. 7B), and significantly different from random for most alleles ($p < 1.e-100$). Because the number of OF sequences was much smaller than the number of IF sequences, the expected variance in the OF DRF was much larger than the expected variance in the IF DRF.

The junction between DH and JH exhibits a small degree of reading frame bias compared with DRF bias

We wondered whether counterselection for the $\Delta_{th}$ protein could account for the DRF bias. The $\Delta_{th}$ protein is generated by DH to JH rearrangement prior to complete VDJ rearrangement in pro-B cells. The $\Delta_{th}$ protein is often encoded in RF2, and, based upon the analysis of $\Delta_{th}$ transgenic mice, $\Delta_{th}$ signaling can inhibit VH to DJ rearrangement. The $\Delta_{th}$ protein can associate with the surrogate L chain as well as Igβ (11), and, intriguingly, the signaling adaptor and tumor suppressor protein called SLP-65 (39) is required for $\Delta_{th}$ selection (40).

Because the $\Delta_{th}$ protein is generated by rearrangement between DH and JH only at the N2 junction, it cannot account for skewing in the DRF of the fully rearranged VDJ unless the N1 junction is somehow constrained. To test this idea, we analyzed the RF usage of N2 in isolation (without regard for N1). If the $\Delta_{th}$ protein is counterselected and contributes significantly to the final VDJ DRF usage, then we expect to find counterselection of N2 RF2. Indeed this is the case (Fig. 8A). RF2 is underrepresented. Intriguingly, N2 RF3 is more frequently used than either N1 or N2. But the effect sizes are small compared with the skewing observed when both the N1 and N2 junctions are taken into account (Fig. 8B). These data indicate that the skewing introduced by the $\Delta_{th}$ RF is insufficient to fully account for the DRF bias.

Stop codon usage is insufficient to account for DRF bias

We next wondered whether negative selection for disfavored DRFs arose due to higher frequencies of stop codon usage in the disfavored DRFs. This possibility seems unlikely because stop codons are only found in 30% of the forward DRF (a table of human DH genes and amino acid conversion in each reading frame can be found at http://peptibase.cs.biu.ac.il/homepage/Lymphocyte_trans_aa.htm). Moreover, stop codons often reside in the extremities of the DH gene and tend to be eliminated through nucleotide deletion and addition. DRF usage exhibited a low correlation with

The Journal of Immunology 7

by guest on July 26, 2017 http://www.jimmunol.org/ Downloaded from
the presence or absence of a stop codon in a given DRF for a given D<sub>Hi</sub> gene. Thus, stop codons are unlikely to be significant drivers of DRF selection. Furthermore, there are alternative DRFs that contain no stop codon (for example, DRF 2 and 3 of IGHD1-20*01), but nonetheless display a very clear bias (0.8 versus 0.2 in the case of IGHD1-20*01).

This leaves two final scenarios that are not mutually exclusive. Either specific DRFs are selected at the amino acid level in the germline through evolution (41), or specific properties of D<sub>Hi</sub> genes are selected somatically (either positively or negatively) during the life of the B cell (9–12, 16, 42–44).

**Correlation of amino acid with DRF frequency**

To test whether skewed DRF usage correlates with the presence (or absence) of particular amino acids, we computed a DRF usage probability matrix (3 DRF *18 D<sub>Hi</sub> genes, represented as a single 54-position vector) and compared it with an occupancy matrix for each amino acid in each D<sub>Hi</sub> in each DRF. Thus, each amino acid was assigned a (3*18) matrix representing the number of times it appears in each D<sub>Hi</sub> and DRF (represented again as a single 54-position vector). We then computed for each sample and each amino acid the Pearson correlation between the DRF usage probability matrix and the amino acid occupancy matrix. The result was a correlation value for each amino acid for each sample. For most amino acids, this correlation is highly significant (p < 0.01 for 16 of 20 aa using a t test over the correlation in all samples versus 0). Moreover, the correlation is highly consistent among all samples. Whereas the overall average correlation in the DRF usage among all samples is 0.75–0.8, the list of the 10 most highly correlated amino acids (both positive and negative) is actually conserved in 95% of the samples. Strikingly, some amino acids are highly positively correlated with DRF usage, such as threonine, tyrosine, or serine, and some are highly negatively correlated (leucine, glutamine, and others; Fig. 9). Moreover, except for methionine, asparagine, and valine, all amino acids are either always positively correlated with expression or negatively correlated in the vast majority of samples.

**Effect of DRF and D<sub>Hi</sub> gene usage on D<sub>Hi</sub> gene position**

A possible explanation for the DRF bias may be that each D<sub>Hi</sub> gene and DRF uses a different part of the D<sub>Hi</sub> gene. For example, for a given gene, the end of the gene may be typically used, whereas for other genes the initial part may be used, and the end deleted in the junction production. Interestingly, there are clear differences between the different D<sub>Hi</sub> genes (Fig. 10). Fig. 10 shows for each position along the D<sub>Hi</sub> gene, for each D<sub>Hi</sub> gene and each RF, how often this position is used starting at the beginning of the D<sub>Hi</sub> in the rearranged gene. Each triplet of rows is one D<sub>Hi</sub> gene in three different forward RFs (DRF1, DRF2, and DRF3), and each position along the x-axis represents the nucleotide position along the D<sub>Hi</sub> germline gene sequence. The colors represent the relative frequency at which a position is used (the number of times that that nucleotide is used divided by the number of rearrangements involving that particular D<sub>Hi</sub> in that particular DRF). As one can clearly see, there are preferred positions used by each D<sub>Hi</sub> gene. However, these preferred positions are quite similar among the different DRF for the same gene. Still, some specific differences exist, and some D<sub>Hi</sub> genes tend to use preferred positions in some specific RFs (the red boxes in Fig. 10). We have thus analyzed the amino acid sequences that are encoded by nucleotides in these positions, and there seems to be a preferred usage of a QLV sequence for some of the D<sub>Hi</sub> genes (see Table II). This preference is unlikely to be explained by frequent usage of Q and L among germline D<sub>Hi</sub> genes, because RFs individually containing Q and L individually are actually infrequently used (Fig. 9). Rather, this specific combination of amino acids seems to be favored. Thus, whereas selection for the starting point of a D<sub>Hi</sub> gene is probably based on the rearrangement mechanism for most D<sub>Hi</sub> genes, there are some specific sequences that are highly positively selected.

**Discussion**

We show in this work in human peripheral blood B cells that a single DRF predominates for most D<sub>Hi</sub>. Thus, in reality there is much less freedom in DRF usage, and the expressed IgH repertoire is therefore not as diversified by alternative DRF as is theoretically possible. The skewing toward particular DRFs mapped not only to individual D<sub>Hi</sub> genes, but tended to be similar in members of the same D<sub>Hi</sub> gene family. One of the most striking findings was that the skewing toward a particular DRF in the rearrangements of a given D<sub>Hi</sub> gene was preserved over a wide range of rearrangement lengths. This finding suggests that the DRF bias reflects a selective process that is intrinsic to the D<sub>Hi</sub> gene sequence itself. Furthermore, the DRF bias was highly similar in different individuals and was most marked in naive IgM<sup>+</sup> B cells. Collectively, these findings indicate that there is something about the DRF bias that is wired into the preimmune repertoire. But how and why does this happen?

There are two general ways in which DRF bias could arise. The first is that bias could be introduced at the time of rearrangement. The second possibility is that rearrangements are random, but that selection (operating either negatively or positively) favors certain DRFs. Biased rearrangement would predict that OF rearrangements should also exhibit biases in DRF, yet this is not observed: instead, the usage of DRF is fairly uniform among OF rearrangements. As naive IgM<sup>+</sup> B cells exhibited the most striking and consistent DRF bias, we reasoned that the major checkpoint for DRF selection must occur early during B cell development. The earliest stage in which this could occur is in pro-B cells (when H chain rearrangement is occurring) and could involve the D<sub>µ</sub> protein. However, when we analyzed the RF skewing in the N2 junction between D<sub>Hi</sub> and J<sub>µ</sub>, the effect sizes were small compared with the
level of skewing when N1 and N2 junctions were viewed in aggregate in the completed VDJ rearrangement. Furthermore, there was no specific distance between the N1 and N2 rearrangement junctions that was favored. Instead, the DRF bias existed over a wide range of rearrangement lengths. Thus, counterselection for the D_{μ} protein appears to be insufficient to account for the DRF bias. Furthermore, the frequency of stop codon usage in the different DRFs was also insufficient to account for the DRF bias. Collectively, these findings favor the remaining alternative, namely that selection for particular amino acid motifs within the D_{μ} gene sequence is occurring. Consistent with amino acid–based selection, the usage of particular amino acids within particular DRFs is highly nonrandom.

Currently, we have no clear explanation for the mechanism relating the amino acid usage with the DRF usage. The simplest explanation would be that some target Ags bind to specific amino acids. Thus, the main element that shapes the DRF bias could be simply the need to express or avoid certain specific amino acids in the CDR3 region. As such, these results may provide a map of permissive CDR3 amino acid motifs that could be helpful in the analysis of Ab and autoantibody repertoires. The DRF bias could be altered if there is relaxed selection stringency, as could occur in autoimmunity.

In parallel to the current analysis, DRF bias in human B cells has been studied by Larimore et al. (31), which confirms our observation on the larger DRF bias in productive than in nonproductive rearrangements, the authors also observed biases against longer CDR3 sequences with higher hydrophobicity (GRAVY) scores among productive rearrangements. Counterselection against longer CDR3 sequences has also been documented during early B cell development in humans (45) and is observed in our sequences. However, we did not observe the skewing toward hydrophobic residues. In our analysis, we computed the correlation between the DRF usage with the presence or absence of the different amino acids (Fig. 9). Therefore, we should expect a negative correlation with hydrophobic amino acids. Instead, we show that hydrophobic amino acids either positively (phenylalanine, isoleucine) or negatively (leucine, tryptophan) correlated with the DRF usage, and some did not correlate at all (methionine, valine). Our analysis shows that, in addition to major shifts during early B cell development, DRF usage appears to be further shaped by peripheral selection, as could arise through exposure and immune responses to foreign Ags produced by pathogens. We have compared the naive and memory B cell repertoires. Biases in DRF usage are most pronounced and consistent between different individuals in the naive B cell repertoire. Among the CD27^+ B cell pools there are greater interindividual differences. Among IgG^+ CD27^+ B cells the DRF bias is less pronounced, and overall the repertoire appears to contract, with a smaller and shorter range of CDR3 lengths. Additionally, there are differences in DRF usage between IgA^+ and IgG^+ subsets. Collectively, these data suggest that DRF is subjected to more than one selection checkpoint, the first occurring in the bone marrow, followed by a second occurring peripherally. We have no clear explanation why IgG selection appears to be more stringent than IgA selection. It does not appear to be due to differences in sample size and because unique sequences rather than total copies were analyzed; thus, these data are not likely to be significantly skewed by clonal expansion. The conclusion from all these observations is that the repertoire is shaped by a common and seemingly quite stringent selection

![Figure 10](http://www.jimmunol.org/)

**FIGURE 10.** Relative frequency of the starting position in the D_{υ} gene. Each position in the x-axis is a nucleotide along the D_{υ} gene, and each triplet of rows represents a D_{υ} gene. Within each triplet, each row is a DRF. The colors represent the frequency of this position in the DNA samples. Only RFs (i.e., rows) with at least 100 sequences were taken into account (otherwise, the values are zeros).

<table>
<thead>
<tr>
<th>D_{υ} Gene</th>
<th>DRF</th>
<th>Starting Position (Nucleotide)</th>
<th>Nucleotide Sequence</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHD1-20</td>
<td>1</td>
<td>4</td>
<td>ATAACTGGGACGAC</td>
<td>ITGT</td>
</tr>
<tr>
<td>IGHD1-20</td>
<td>2</td>
<td>5</td>
<td>AACTGGGACGAC</td>
<td>NWND</td>
</tr>
<tr>
<td>IGHD1-26</td>
<td>3</td>
<td>6</td>
<td>TGDAACCTACTAC</td>
<td>WELL</td>
</tr>
<tr>
<td>IGHD3-22</td>
<td>1</td>
<td>18</td>
<td>TGGTTATTACTAC</td>
<td>WLLL</td>
</tr>
<tr>
<td>IGHD4-17</td>
<td>1</td>
<td>2</td>
<td>CTACCGGACTAC</td>
<td>LR^*L</td>
</tr>
<tr>
<td>IGHD6-13</td>
<td>2</td>
<td>8</td>
<td>CAGCAAGCTGTGAC</td>
<td>QQLV</td>
</tr>
<tr>
<td>IGHD6-19</td>
<td>2</td>
<td>8</td>
<td>CAGTGCGCTGTGAC</td>
<td>QWLV</td>
</tr>
<tr>
<td>IGHD6-25</td>
<td>2</td>
<td>8</td>
<td>CAGCGCCTCAC</td>
<td>QRL</td>
</tr>
<tr>
<td>IGHD6-6</td>
<td>2</td>
<td>7</td>
<td>CAGCTGGGCCC</td>
<td>QLV</td>
</tr>
<tr>
<td>IGHD6-6</td>
<td>2</td>
<td>8</td>
<td>CAGCTGGGCCC</td>
<td>QLV</td>
</tr>
</tbody>
</table>

The first column is the D_{υ} gene; the second column the DRF; the third column is the position in the D_{υ} gene used as a beginning of the D_{υ} gene. The last columns represent the nucleotide and amino acids of the resulting D_{υ} genes. An interesting QLV motif appears in many of the sequences. The nucleotide sequences are not the full germline sequences, but the relative regions translated, taking into account the DRF and the starting position in each case.
mechanism that operates primarily in the naive pool, followed by further selection in the periphery. This interpretation is reinforced by the very clear correlation between the presence of specific amino acids in the CDR3 with the usage of a given DRF. This correlation is very similar in samples from different individuals, again highlighting a common selection mechanism. However, it is not clear how this selection operates. Selection could be influenced by self-Ags that are highly expressed in the bone marrow or in the periphery, as proposed in the context of natural Abs. Alternatively or in addition, the selection could be negative, disfavoring Abs with certain biochemical properties in their CDR3s. In mice, we and others (7–11, 46) have shown that charged amino acids induce negative selection, which constrains the DRF (8).

This study advances our understanding of B cell repertoire selection in two basic ways, as follows:

First, this study shows that D4 gene segments do not afford maximal diversity, due to the biased usage of particular DRFs. This restriction in diversity arises after V(DJ) recombination because OF rearrangements do not exhibit a similar bias. This implies that selection for particular RFs occurs after their generation. Intriguingly, the restricted usage of a subset of amino acids in a particular D4 is seen over a large range of rearrangement lengths. This suggests that there is an important region at the core of the D4 gene sequence that may be getting selected. We show that the D3 protein and the frequency of stop codons in the different DRFs are insufficient to account for the degree of DRF skewing. Finally, we show that most of the DRF skewing occurs in different DRFs are insufficient to account for the degree of DRF skewing.

Disclosures

The authors have no financial conflicts of interest.

References


