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* 2013; 190:5567-5577; Prepublished online 29 April 2013; doi: 10.4049/jimmunol.1201929
http://www.jimmunol.org/content/190/11/5567

**Supplementary Material**
http://www.jimmunol.org/content/suppl/2013/04/29/jimmunol.1201929.DC1

**References**
This article cites 44 articles, 25 of which you can access for free at:
http://www.jimmunol.org/content/190/11/5567.full#ref-list-1

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

Jennifer Benichou,*1 Jacob Glanville,†1 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, which inverted $D_H$ genes are usually avoided. The second checkpoint occurs after rearrangement, once the BCR is expressed. The second checkpoint 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: 5567–5577.

Antibodies are proteins produced by B cells. Abs consist of two H chains and two L chains. Ab 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 Ab 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(D)J recombination], junctional modifications between the rearranged gene segments, pairing of different H and L chains, and somatic hypermutation (SHM) [reviewed in (1, 2)]. V(D)J recombination occurs in an ordered and stage-specific fashion in the bone marrow, with H chain rearrangement producing a V region that contains juxtaposed $V_{H1}$, $D_{H}$, 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 sequence because it encompasses the junctions between the recombining $V_{H1}$, $D_{H}$, 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 autoimmunity (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
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 naive (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 Illu-
mina platform. Informed consent was obtained for subjects 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 repertoires 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 in silico simulation and analysis, and high-throughput screening of recovered products. The degree of germline-dependent amplification bias was assessed by comparing amplified products of stimulated naive 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 identifier (barcodes) to allow unambiguous identification of all products by sequence analysis in subsequent proce-

sures. 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 tita
nium. Sequencing quality was assessed by k-mer control. Sample quality control was confirmed by demultiplexing and VH segment gene

otype. Sequencing depth was determined by diversity estimate rare-
iﬁcation and simulations of the repertoire 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 Biotec). 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 rearranged IgH CDR3 sequences at the genomic level, using seven V H segment primers (one speciﬁc to each VH segment family) and six J H segment primers (one for each functional J H segment). Reads of 110 bp extending from the J H segment, across the NDN junction, and into the VH segment were obtained using the Illumina HiSeq sequencing platform. The immunoSEQ assay was used for the sequencing. The BCR CDR3 region was deﬁned ac-

A total of 12,948,437 sequence reads representing 47,358 unique sequences, of which 34% were productive, was analyzed for subject 1, and 12,851,153 sequence reads representing 20,374 unique sequences, of which 84% were productive, were analyzed for subject 2.

Germline V H, D H, and J H genes used for sequence composition

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

codon or charged residues (7–11). Other mechanisms, such as a D H-mediated suppression, have also been proposed to explain the preference for a single RF in mice (10–12).

In addition to the usage of up to three different forward RFs, D H rearrangement can occur by deletion or by inversion. Inversional rearrangements increase the number of potential RFs to six (three forward-facing RFs and three reverse). Rearrangements to all six possible DH gene RFs (DRF) have been reported in mice, and the use of inverted DRF has been described in autoimmune-prone strains of mice (13–16).

Whereas the joining probability of VH, D H, and JH may be affected by the structure of the locus and properties of the RAG complex, as well as by the dynamics of the rearrangement mechanism [e.g., editing (3, 17–19)], the DRF probability distribution appears to be purely driven by selection, as we will show further in this work. The usage of forward and inverted DRF, in contrast, can be based on selection, but also on possible limitations imposed by the recombination signal (20). In humans, previous reports have claimed that essentially only forward RFs are used (21) and that there are practically no D-D fusions, findings that we will confirm more definitively in this work (22–24).

Up until now, the analysis of DH gene sequences in humans has been limited to relatively few sequences. Because selection of Abs is inﬂuenced by the VH gene used (25, 26), earlier studies in which only a few dozen rearrangements from different VH were analyzed were insufﬁcient to rigorously evaluate DH selection (which requires controlling for the VH gene used). A second difﬁculty is to precisely identify DH segments in rearranged Ab genes, which are often truncated due to exonucleolytic nibbling by nonhomol-

ogous end-joining machinery that can ultimately remove practi-

cally the entire DH gene. Sometimes it is difﬁcult to discern which nucleotides in the sequence are derived from the DH segment as opposed to those that result from junctional modiﬁcations or SHMs that target the CDR3 [e.g., (27, 28)]. The confounding effects introduced by SHMs may reduce the reliability of DH segment identiﬁcation in mature B cell populations. Some DH genes are more similar to each other than others, which means that DH identification fidelity depends not only upon the length of the DH germline sequence (11–37 bp), but also upon its degree of similarity to the other germline DH gene alleles. Finally, some DH genes are much more frequently used than others [see for example (29)]. Thus, in a limited number of sequences, only the most abundantly used DH genes will be sampled, as the precise distribu-

tion in relatively rarely used DH genes is hard to measure.

With the advent of high-throughput sequencing, it has become possible to sample a large enough number of naive BCR sequences to capture a large enough number of DH with a limited number of SHMs, even those with short DH gene sequences. In this study, we present the DRF distribution from 12 human subjects. We describe a new computational method for deﬁning DH segments and provide information on which DH segments are most versus least reliably identiﬁed, which is of interest to those using high-

throughput sequencing to analyze the Ab repertoire. In agree-

ment with previous publications, we ﬁnd that inverted DRF are prac-

tically never used (22, 24). However, even in the forward RF, we show that there is very strong selection of DRFs that eventually leads to the selection of one or two dominant RFs of the three possible forward RFs in the expressed Ab repertoire. These results have implications for the overall level of IgH diversity and the timing of H chain selection checkpoints in humans. An understanding of these selection checkpoints may be useful in future studies of how B cell selection is altered in disease.
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. We then took all of the sequences with the same VH, JH, and distance between VH and JH and grouped them using a phylogenic 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 phylogenetic 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 to one another, up to a distance of four mutations.

**D<sub>H</sub> detection**

We have computed the maximal similarity between each D<sub>H</sub> 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 D<sub>H</sub> gene. The similarity of a given sequence to a germline D<sub>H</sub> has been defined as the fraction of nucleotides equal to the one of the germline D<sub>H</sub>. We only further analyzed sequences with a similarity of at least 0.75.

**Validation set production**

In silico, we generated a set of 10,000 sequences by random V-D-J joining (with equal probability for each D<sub>H</sub> gene). In this dataset, the D<sub>H</sub> 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 D<sub>H</sub> gene, and 2) replacement of 1–3 nt in random positions in the D<sub>H</sub> gene. In both methods, purines and pyrimidines had a two-thirds probability of being mutated into a base in the same group (transversion) and a one-third probability to be mutated to the opposite group (transversion). We then computed the best fit D<sub>H</sub> and its orientation for each of these known rearrangements, using the same methodology as was used for the real (unknown) sequences. The fraction of high precision fits was affected by the D<sub>H</sub> gene length: as expected, longer D<sub>H</sub> genes were properly classified more often.

**Results**

Whereas the RF usage of VH and JH is constrained, the RF of D<sub>H</sub> is flexible because nucleotide additions or deletions are present at both ends of the rearranged D<sub>H</sub> gene segment. Furthermore, D<sub>H</sub> genes can potentially undergo inversion. Thus, D<sub>H</sub> 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 D<sub>H</sub> gene actually uses a limited number of RFs.

The analysis of D<sub>H</sub> gene sequences in the Ab repertoire is not easy because some D<sub>H</sub> genes are similar to each other, and nibbling of the D<sub>H</sub> genes by the nonhomologous end-joining machinery can result in very short D<sub>H</sub> gene sequences. Thus, the robust identification of D<sub>H</sub> 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 D<sub>H</sub> 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 D<sub>H</sub> gene determination using computer simulations of D<sub>H</sub> rearrangements and found the results to be highly precise for long D<sub>H</sub> 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 D<sub>H</sub> 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<sup>+</sup>-enriched PBLs, amplification using mixtures of VH<sub>H</sub> and J<sub>H</sub> 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 D<sub>H</sub> 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 I). For each subject, we sorted CD20<sup>+</sup> lymphocytes into naïve (CD27<sup>-</sup>) and memory (CD27<sup>+</sup>) 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<sub>H</sub> 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<sub>H</sub> with all possible germline J<sub>H</sub> genes, and assumed that the J<sub>H</sub> gene could not start >50 nt before the 3' end of the VH<sub>H</sub> gene. We found that practically all sequences for which a VH gene could be identified, the J<sub>H</sub> gene could also be detected with a high enough accuracy (>$70%$ similarity to germline VH<sub>H</sub> and J<sub>H</sub>). 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 VH<sub>H</sub> and J<sub>H</sub>, we extracted the sequence regions that included the D<sub>H</sub> gene segment(s), starting with the sequence that was 30 nt before the end of the VH<sub>H</sub> germline gene and ending with the 3' end of the germline J<sub>H</sub> sequence. We looked for the best fit to all germline D<sub>H</sub> genes in the IMGT database. The optimal fitting D<sub>H</sub> gene was given a score representing the fraction of the germline D<sub>H</sub> gene nucleotides fitting the observed sequence (see Fig. 1). The position of the D<sub>H</sub> gene was then determined relative to the germline VH<sub>H</sub> and J<sub>H</sub> genes.

| Table I. Donor demographics for 454 and Illumina sequencing |
|----------------|----------------|-------------|----------|
| Sequencing Technology | ID | Gender | Ethnographics | Age |
| 454 | | | | |
| 91 | F | White | 54 |
| 104 | F | Hispanic | 28 |
| 130 | M | Hispanic | 39 |
| 136 | M | White | 54 |
| 158 | M | Asian | 25 |
| 182 | F | White | 27 |
| 219 | F | Asian | 44 |
| 231 | F | Asian | 29 |
| 272 | M | Asian | 57 |
| 273 | M | White | 27 |
| 275 | M | Hispanic | 50 |
| 276 | F | Hispanic | 47 |
| 1 | F | White | 49 |
| 2 | F | White | 22 |
| Illumina | | | | |

[Downloaded from http://www.jimmunol.org/ by guest on April 13, 2017]
between, we compute DH (nucleotides that are added and/or removed from the DH gene). At an inverted DH. We found that the error level is a function of the overlap between the query sequence and the germline DH sequence) and the number of errors in the assignment of the DRF (compare Supplemental Fig. 1). One can clearly see that as soon as the overlap is >0.7, the fraction of sequences with a RF error is <1% and the D\textsubscript{H} identification error is <10%. Most of those occur in highly similar D\textsubscript{H} genes. RFs with a negative number are inverted RFs. Those are observed only in very short D\textsubscript{H} genes.

Methods

We have repeated the analysis separately for each D\textsubscript{H}. When averaging over each D\textsubscript{H} separately (instead of overall clones), some inverted D\textsubscript{H} genes are observed in rare short D\textsubscript{H} genes (Fig. 3A). However, in short D\textsubscript{H} genes, a small number of fitting nucleotides can be enough to ensure a 75% match. As mentioned above, for each candidate inverted D\textsubscript{H}, a reasonable candidate forward D\textsubscript{H} can be found. One can thus conclude that in very large cohorts, if inverted D\textsubscript{H} exist, they are very rare and limited to short D\textsubscript{H} 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\textsubscript{H} genes.

Note that most inverted D\textsubscript{H} 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.

FIGURE 1. Computation of the DRF. We first compute the position of V\textsubscript{H} (from ACGA to GAGA in the first row and the beginning of the second row). The upper line is the full sequence, and the lower line is the germline gene (V\textsubscript{H} in this case). We then compute how the codons are aligned with respect to the beginning of V\textsubscript{H} (see, for example, the GAG at the end of V). On the 3’ end of the sequence, we compute the position of J\textsubscript{H} (lower right-hand part of figure). Again the sequence is above the germline J\textsubscript{H} gene. In between, we compute D\textsubscript{H} (central lower part). The sequence above the text is the germline, and we compute its position versus the RF of V\textsubscript{H} in this case, it starts at the third RF (the first G of the germline D\textsubscript{H}, indicated by the arrow, is above a G in the third RF of the V\textsubscript{H} germline). Crossed-out nucleotides are germline nucleotides that were removed from either V\textsubscript{H}, D\textsubscript{H}, or J\textsubscript{H} in the join region.

FIGURE 2. Effect of mutation frequency on D\textsubscript{H} and DRF-identification accuracy. The x-axis is the fraction of nucleotides in the sampled sequence that overlaps with the full D\textsubscript{H} germline sequence. The y-axis is the fraction of properly classified D\textsubscript{H} genes in simulated rearrangements. The errors were divided into multiple categories, including errors in D\textsubscript{H} identification or in the DRF classification or both. A detailed analysis of the effect of the D\textsubscript{Hi} gene on the precision is provided in Supplemental Fig. 1. One can clearly see that as soon as the overlap is >0.7, the fraction of sequences with a RF error is <1% and the D\textsubscript{Hi} identification error is <10%. Most of those occur in highly similar D\textsubscript{H} genes. RFs with a negative number are inverted RFs. Those are observed only in very short D\textsubscript{H} genes.

Absence of inverted DRF

We first checked the DRF usage in the total sample (without taking into account the specific D\textsubscript{H} used). We then computed how many unique clones (see Materials and Methods) used each RF for each D\textsubscript{H}. 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 D\textsubscript{H} inversions may not really contain inverted D\textsubscript{H} segments. Indeed, for each D\textsubscript{H} gene found to be in the inverted DRF, there was also a reasonable fit to a D\textsubscript{H} in a forward DRF (not always in the same D\textsubscript{H} gene). Thus, we conclude [as was previously suggested based upon smaller datasets (22)] that the use of inverted D\textsubscript{H} segments, if it occurs at all, is very infrequent in the IgH repertoire of healthy human adults.

The frequencies at which D\textsubscript{H} genes are used in our sample vary widely among D\textsubscript{H} genes. One caveat is that we are only analyzing D\textsubscript{H} genes in which the fraction of original nucleotides maintained in the rearranged sequence is >75%. This selection induces a bias toward long D\textsubscript{H} genes that can be classified properly with a better accuracy (Supplemental Fig. 1).
germline DH). The average over different DH genes. We have removed data.

Each column is a DH gene. Each shade is a DRF. We have removed
DRF rather than the other because the region of the amplification
there is no reason for the PCR amplification to overamplify one
gene. As was the case for the inverted DRF, we have repeated the
analysis separating each DH gene. The DRF usage varies among DH
genes, but it is again highly nonuniform for each DH gene (Fig.
3A). The nonuniform distribution is not the result of stop codons,
because DRFs without stop codons are practically absent. More-
over, the DRF usage pattern is highly reproducible among samples
from different individuals (Supplemental Fig. 2) and is the result
of a large number of clones for each sample and each DH gene (data not
shown). Supplemental Fig. 2 shows the DRF usage of clones and
not of total sequences, but the results for total sequences are
similar. Thus, the biased distribution cannot be the result of am-
plification errors or biases, because those would not affect the
clonal number (remember that similar sequences comprising each
clonal sequence are only counted once). For the same reason, the
biased distribution is not affected by very large clones, because again all
sequences in the same clone are only counted once. Thus, if sele-
ction occurs, it is not the result of the amplification of some clones,
but rather the selection of individual clones. Moreover, the
RACE protocol minimizes artificial skewing of VH or JH gene
usage because the primers are neither VH nor JH specific.

To test that the total DRF usage is not the result of errors (that
comprise <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 clas-
sification 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
germine gene 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 selec-
tion 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 × 18
frequency matrix (3 DRF and 18 of 34 DH genes that had >100
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-

Forward DRF usage distribution

The absence of inverted DRF may be mediated by aspects of the
rearrangement mechanism. A surprising feature of the DRF usage
that cannot be explained by the rearrangement mechanism is the
restricted usage of particular forward RFs. In the forward direction,
the DRF usage is highly skewed toward the third forward RF for
most clones, whereas a minority of the clones uses first and second
forward DRF (Fig. 3A). Again, as was the case for the inverted
DRF, most forward DRF do not contain stop codons, which would
reduce their likelihood of usage in circulating B cells. Moreover,
as will be further explained, there is no a priori advantage for one
DRF over the others, and no element of the rearrangement process
that could readily explain this preference. This unequal distribu-
tion thus seems to hint to the presence of a clear selection
mechanism for one DRF for a given DH segment.

To check that this result is not the peculiarity of a single indi-
vidual, we have compared the DRF usage among subjects and
among Ab H chain isotypes (data not shown). In all cases, the DRF
usage is skewed and the manner of skewing is very similar in
most individuals. Note also that there is no reason for the PCR amplification to overamplify one
DRF rather than the other because the region of the amplification

FIGURE 3. Fraction of sequences using each DRF in each DH (454
data). (A) DRF usage of clone sequences. (B) DRF usage of all sequences.
Each column is a DH gene. Each shade is a DRF. We have removed
IGHD4*04, IGHD5*05, and IGHD7*27, because we had too few sequences
classified to these DH genes. One can clearly see that the DRF usage varies
significantly among DH genes. The first DH genes are very short and are
more error prone than the others. (C) Effect of precision on DRF usage
(average over different DH genes). The x-axis is the required precision
(fraction of overlapping nucleotides between a sequenced DH with the
germine DH). The y-axis is the percentage of sequences using each DRF.
The DRF usage is similar at different levels of precision.
The determination of the DRF is error prone. The SD of \([1,0,0]\). The SD in the naive compartment is averaged the results for each DH gene and then averaged over DH genes. The distribution is over all samples within the same isotype. (usage among all samples (donors and isotypes). Each column is a sample, a lower variance than IgG. The differences between all groups are significant \((p < 1.e-10)\), 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.

The correlation was typically \(>0.75\) (Fig. 4A). The correlation between naive IgM repertoires was 0.8, whereas the correlations within and between memory isotypes were 0.7 and 0.77, respectively \((t\text{ test, } p < 1.e-10\) for all comparisons). The higher similarity between naive IgM sequences than between IgA and IgG suggests that there is a shared bone marrow–based selection mechanism that drives the uniform selection in IgM sequences, followed by diversification in the memory compartment.

The analysis of selection can be taken one step further by comparing the variability of DRF usage in naive versus memory B cell IgH repertoires. Completely uniform DRF usage would lead to a SD of 0 in the DRF usage (per DH gene). Conversely, in the extreme case of usage of a single DRF, the SD would be 0.57 (the SD of \([1,0,0]\)). The SD in the naive compartment is \(\sim0.3\), and it rises progressively to 0.37 in the memory IgM repertoire and to 0.43, almost the maximal variability \((i.e., \text{most limited DRF usage})\), in the IgG compartment (Fig. 4B). The rise in the DRF variance may be partially due to clonotypic enlargement. However, there is typically much more than one clone per DH gene. Thus, clonotypic enlargement is far from being the only source of the high variance.

Among the memory subsets, we found the highest degree of variance in the IgG\(^2\) sequences, where DRF usage is limited to practically a single DRF \((t\text{ test, } p < 0.01\) versus naive IgM cells). As expected, memory IgM has a variance between the naive and IgG variances. The DRF usage is less skewed in memory IgA than in IgG sequences for reasons that are not clear.

The average overall DH gene DRF distribution was practically equal among all subjects and compartments (Fig. 4C), which again favors a two- or more-stage selection model: a common selection stage among naive cells or their precursors, and further selection/diversification in the IgG/IgA pools, which is subject specific.

Distance between \(V_{H}, D_{H}\), and \(J_{H}\)

One possible explanation for the skewed usage of DRFs in the IgM compartment is a rearrangement bias. Such a bias could arise if there were preferred rearrangement distances between \(V_{H}\) and \(D_{H}\) or between \(D_{H}\) and \(J_{H}\). If biased rearrangement had been the source of the skewed DRF usage, we would have expected to observe a very narrow distribution of the distances between \(V_{H}\) and \(D_{H}\) or \(D_{H}\) and \(J_{H}\). We have measured this distance for each \(D_{H}\). For most \(D_{H}\) genes, one observes a very wide distribution covering \(>30\) nt, with jumps of 3 nt (Fig. 5). Thus, for example, the frequency of a distance between \(V_{H}\) and \(D_{H}\) of 3 nt is much higher than 2 or 1 nt, but is similar to a distance of 0 nt. The same can be observed in the distance between \(D_{H}\) and \(J_{H}\) (data not shown). Although the \(J_{H}\) segment used influences the CDR3 length, the range of rearrangement lengths cannot be readily ascribed to skewed \(I_{H}\) usage for a particular \(D_{H}\) gene segment. Furthermore, the wide range of CDR3 lengths that occur at 3-nt intervals in the preferred DRF exists even at the level of specific V-D-J combinations (Fig. 6); note again that the RACE protocol is not expected to produce any length or \(V_{H}-I_{H}\) bias. Such a bias in intervals of three over a wide range of rearrangement lengths is therefore highly unlikely to be the result of biased rearrangement. One is thus left with positive or negative selection as the most likely potential explanations for the common bias in DRF usage.

**DRF Usage is Uniform in Out-of-Frame IgH Rearrangements**

To further demonstrate that the DRF bias is likely to be due to selection rather than biased recombination, we have analyzed out-of-frame (OF) rearrangements in B cell genomic DNA. If biased rearrangement contributes to the DRF bias, then this should also be observed in OF rearrangements. We thus compared the DRF bias in rearrangements that are IF with those that are OF. To obtain sufficient numbers of OF sequences, we sequenced genomic DNA using the Illumina platform (Table I, see Materials and Methods). First, we checked whether we observe the same average DRF usage using a different sequencing technology (Fig. 7A). Indeed, one can clearly see the similar biased DRF usage between the

**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 \(D_{H}\) 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 isotype 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 \(D_{H}\) 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 \((p < 1.e-10)\), 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 \(D_{H}\) gene and then averaged over \(D_{H}\) genes. Thus, rare \(D_{H}\) 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 \(D_{H}\) genes. However, in these sequences, the determination of the DRF is error prone.

**FIGURE 5.** Frequency of position of end of \(D_{H}\) for each \(D_{H}\) gene. Each column is a position at the end of \(D_{H}\) starting from 10 positions into \(V_{H}\). Each row is a \(D_{H}\) gene. The shades represent frequency, as shown in the grayscale bar. One can clearly see in most \(D_{H}\) 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 ($p < 1 \times 10^{-100}$). 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.

We wondered whether counterselection for the Dm protein could account for the DRF bias. The Dm protein is generated by DH to JH rearrangement prior to complete VDJ rearrangement in pro-B cells. The Dm protein is often encoded in RF2, and, based upon the analysis of Dm transgenic mice, Dm signaling can inhibit VH to DJ rearrangement. The Dm 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 Dm selection (40).

Because the Dm 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 Dm 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 Dm 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

FIGURE 6. DJ region length distribution for different VDJ rearrangement in naive B cell samples. We computed in the clone sequences the length distribution of the region spanning the full germline DH and JH regions.

FIGURE 7. DRF usage in the Illumina data. (A) Fraction of sequences using each DRF in each DH. Each column is a DH gene. Each shade is a DRF. One can clearly see the similar DRF usage as observed in Fig. 3. Differences may arise due to the fact that different donors were recruited for each sequencing. (B) SD of DRF usage in IF and OF sequences. Each column is the SD of the DRF usage in a given gene, as computed in Fig. 4B, using the Illumina-based sequences. The expected SD was computed as 1 over the square root of the sample size. The OF RF SD is only slightly higher than the one expected by the size of the sample, because the sample size for OF rearrangements was limited. The IF SD of DRF usage approaches the maximal possible in this case ($0.57 = \sqrt{1/3}$), and much more than expected by the sample size. One can thus conclude that most of the variance in the DRF usage is determined by selection. DH genes for which there were not enough OF or IF sequences (<100) were not incorporated in the analysis.

FIGURE 8. DJ rearrangement RF. (A) DJ rearrangement RF usage of all isotypes and donors (454 data). As in Fig. 6, we used the segment containing the DH, N2, and JH regions, and replaced the DH and JH with the full germline sequences. We then computed the RF of this segment with respect to the beginning of the C region. There is a general small bias in favor of the second RF (RF2). (B) DJ rearrangement RF usage in IF and OF sequences (Illumina data). We observe a uniform usage in OF sequences. However, we see again a bias toward RF 2.
the presence or absence of a stop codon in a given DRF for a given DH 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 that 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 DH 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 DH genes, represented as a single 54-position vector) and compared it with an occupancy matrix for each amino acid in each DH gene in each DRF. Thus, each amino acid was assigned a (3*18) matrix representing the number of times it appears in each DH 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 DH H gene usage on D H gene position**

A possible explanation for the DRF bias may be that each DH gene and DRF uses a different part of the DH 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 DH genes (Fig. 10). Fig. 10 shows for each position along the DH gene, for each DH gene and each RF, how often this position is used starting at the beginning of the DH in the rearranged gene. Each triplet of rows is one DH gene in three different forward RFs (DRF1, DRF2, and DRF3), and each position along the x-axis represents the nucleotide position along the DH 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 DH in that particular DRF). As one can clearly see, there are preferred positions used by each DH gene. However, these preferred positions are quite similar among the different DRF for the same gene. Still, some specific differences exist, and some DH 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 DH genes (see Table II). This preference is unlikely to be explained by frequent usage of Q and L among germline DH 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 DH gene is probably based on the rearrangement mechanism for most DH 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 DH. 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 DH genes, but tended to be similar in members of the same DH gene family. One of the most striking findings was that the skewing toward a particular DRF in the rearrangements of a given DH 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 DH gene sequence itself. Furthermore, the DRF bias was highly similar in different individuals and was most marked in naive IgM+ 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+ 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 Dµ protein. However, when we analyzed the RF skewing in the N2 junction between DH and Jµ, the effect sizes were small compared with the

**FIGURE 9.** Pearson correlation of DRF usage with the number of times each amino acid appears in each RF. The x-axis represents amino acids. The values on the y-axis are the Pearson correlation coefficients of the DRF usage and the amino acid frequency vectors. Only correlations with a $p$ value $\leq0.05$ were used. Lysine and proline are practically never used in DH genes, so they are not incorporated in the analysis. Each box contains all the samples (8 samples were analyzed for each one of the 12 subjects: IgA, IgG, IgM naive, IgM memory, each performed in duplicate). As one can clearly see, some of the amino acids are practically always positively correlated with the DRF usage, and some practically always negatively correlated. Note that the correlation is biased by the presence of stop codons. For example, leucine is often present near stop codons. We present in Supplemental Fig. 4 a similar analysis when only sequences with no stop codons are used.
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 Dμ 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 Dμ 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 that selection for particular amino acid motifs within the Dμ gene sequence is occurring. Consistent with amino acid–based selection, the usage of particular amino acids within particular DRFs is highly nonrandom.

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 cells 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 process. 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).

Table II. Specific Dμ gene starting positions and RFs that are overpresented

<table>
<thead>
<tr>
<th>Dμ Gene</th>
<th>DRF</th>
<th>Starting Position (Nucleotide)</th>
<th>Nucleotide Sequence (5’→3’)</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHD1-20</td>
<td>1</td>
<td>4</td>
<td>ATAACGTGAAACGAC</td>
<td>ITGT</td>
</tr>
<tr>
<td>IGHD1-20</td>
<td>2</td>
<td>5</td>
<td>AACCGGAAACGAC</td>
<td>NWWN</td>
</tr>
<tr>
<td>IGHD1-26</td>
<td>3</td>
<td>6</td>
<td>TGGGAGCTTACTAC</td>
<td>WLLL</td>
</tr>
<tr>
<td>IGHD3-22</td>
<td>1</td>
<td>18</td>
<td>TGTGTTATATACACT</td>
<td>QQLV</td>
</tr>
<tr>
<td>IGHD4-17</td>
<td>2</td>
<td>1</td>
<td>CTACGCGTGACTAC</td>
<td>LR^2^L</td>
</tr>
<tr>
<td>IGHD6-13</td>
<td>2</td>
<td>8</td>
<td>CACACGGCTTAC</td>
<td>QQLV</td>
</tr>
<tr>
<td>IGHD6-19</td>
<td>2</td>
<td>8</td>
<td>GATGGCTGCTACT</td>
<td>QVLW</td>
</tr>
<tr>
<td>IGHD6-25</td>
<td>2</td>
<td>8</td>
<td>CAGCGGCTAC</td>
<td>QRL</td>
</tr>
<tr>
<td>IGHD6-6</td>
<td>2</td>
<td>7</td>
<td>CAGCTCTCC</td>
<td>QLV</td>
</tr>
<tr>
<td>IGHD6-6</td>
<td>2</td>
<td>8</td>
<td>CAGCTCTCC</td>
<td>QVL</td>
</tr>
</tbody>
</table>

The first column is the Dμ gene; the second column the DRF; the third column is the position in the Dμ gene used as a beginning of the Dμ gene. The last columns represent the nucleotide and amino acids of the resulting Dμ 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(D)J 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 Dp 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 naive IgM+ B cells and that DRF usage shifts slightly in the more mature (CD27+) B cell pools. Taken together, these findings indicate that there are two windows for DRF selection during B cell development. By defining the skewed DRF usage in healthy individuals, we can use this information to study how B cell selection may be altered in B cell disorders such as autoimmunity or cancer.

Second, we show that the bias in DRF varies by individual D4 gene segment and is reproducible in different individuals. This implies that the selection mechanism leading to DRF bias may be evolutionarily conserved. Analyzing the DRF distribution in different species may provide insight into the potential (self) Ags that drive this process.

There are three potential limitations to this analysis. First, the amino acid: DRF correlation is biased by the presence of stop codons. For example, leucine is often present near a stop codon. However, even when only sequences without stop codons are used, clear correlations are observed (Supplemental Fig. 4). Second, this analysis utilizes the entire germinal D4 sequence. Because D4 genes are frequently nibbled from both ends during nonhomologous end joining, the amino acid occupancy matrix may not adequately reflect the true amino acid usage. Indeed, one can, for example, observe a very limited negative correlation of the DRF frequency with the presence of stop codons. The correlation is limited, because stop codons are often near the borders of the D4 gene and are nibbled away. Note that in this work we have not analyzed the effect of somatic hypermutation on DRF. This is a fascinating issue and one that requires far more analysis in the future.

Disclosures
The authors have no financial conflicts of interest.

References


**Supplemental Figure 1. Error level distribution.** We show here the fraction of correctly classified sequences at a level of 75% overlap (fraction of nucleotides overlapping between the sequence and the germline D_H). At this overlap level, only less than 10% of the sequences are misclassified. Most errors are due to a misclassification of the D_H gene and not the DRF. A) Fraction of correct classifications of both the DRF and the D_H gene, as a function of the D_H germline (columns) and the number of mutations (rows). B) Fraction of correct classification of the DRF, as a function of the D_H germline and the number of mutations. C) Fraction of correct classification of the DRF and the D_H genes, as a function of the D_H length and the number of mutations. D) Fraction of correct classification of the DRF, as a function of the D_H length and the number of mutations.

**Supplemental Figure 2. Comparison between D_H specific DRF usage in different samples.** The following figures show the DRF usage for all samples (all isotypes together) of 4 donors. The Y-axis is the fraction of clones using each DRF. Some D_H genes are too similar to other genes and were not included in the analysis. The DRF usage pattern is highly similar among donors.

**Supplemental Figure 3. DRF usage among D_H genes for sequences without deletions.** The following figure is equivalent to Fig. 3 in the main text, but this analysis was performed only with sequences containing the full germline D_H, i.e. sequences with no deletions at both ends of the D_H gene. The same biased DRF usage is still observed.

**Supplemental Figure 4. Pearson correlation of DRF usage with the number of times each amino acid appears in each reading frame.** This figure is similar to Fig. 9 in the main text, but the analysis is only performed on sequences that do not contain SCs to avoid biases. Even when only sequences with no SCs are used, clear correlations are observed.
Fig. S3