The Journal of Immunology

RESEARCH ARTICLE | JUNE 01 2013

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

Jennifer Benichou; ... et. al

J Immunol (2013) 190 (11): 5567–5577.
https://doi.org/10.4049/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,*[,1] Jacob Glanville,[†,1] Eline T. Luning Prak,[‡] Roy Azran,[§,¶] Tracy C. Kuo,[†] Jaume Pons,[†] Cindy Desmarais,[‖] 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, when 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_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 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

*Mina and Everard Goodman Faculty of Life Sciences, Bar Ilan University, Ramat Gan 52900, Israel; [†]Protein Engineering and Applied Quantitative Genotherapeutics, Rinat-Pfizer, South San Francisco, CA 94080; [‡]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [§]Department of Mathematics, Bar-Ilan University, Ramat Gan 52900, Israel; [¶]Gonda Brain Research Center, Bar Ilan University, Ramat Gan, 52900, Israel; and [‖]Adaptive Biotechnologies, Seattle, WA 98102

[1]J.B. and J.G. contributed equally to the generation of the results and preparation of the manuscript.

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.P.).

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.

codon or charged residues (7–11). Other mechanisms, such as a Dμ-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 D_H 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 V_H, D_H, and J_H 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 D_H gene sequences in humans has been limited to relatively few sequences. Because selection of Abs is influenced by the V_H gene used (25, 26), earlier studies in which only a few dozen rearrangements from different V_H were analyzed were insufficient to rigorously evaluate D_H selection (which requires controlling for the V_H gene used). A second difficulty is to precisely identify D_H segments in rearranged Ab genes, which are often truncated due to exonucleolytic nibbling by nonhomologous end-joining machinery that can ultimately remove practically the entire D_H gene. Sometimes it is difficult to discern which nucleotides in the sequence are derived from the D_H 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 D_H segment identification in mature B cell populations. Some D_H genes are more similar to each other than others, which means that D_H identification fidelity depends not only upon the length of the D_H germline sequence (11–37 bp), but also upon its degree of similarity to the other germline D_H gene alleles. Finally, some D_H genes are much more frequently used than others [see for example (29)]. Thus, in a limited number of sequences, only the most abundantly used D_H genes will be sampled, as the precise distribution in relatively rarely used D_H 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 D_H with a limited number of SHMs, even those with short D_H gene sequences. In this study, we present the DRF distribution from 12 human subjects. We describe a new computational method for defining D_H segments and provide information on which D_H segments are most versus least reliably identified, which is of interest to those using high-throughput sequencing to analyze the Ab repertoire. In agreement with previous publications, we find that inverted DRF are practically 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.

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 Illumina platform. Informed consent was obtained, and subjects were asked 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 informatic screening, gel analysis, and high-throughput sequencing 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 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 titanium. Sequencing quality was assessed by keypass control. Sample quality control was confirmed by demultiplexing and V_H segment genotype. Sequencing depth was determined by diversity estimate rarefaction and simulations of germline-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 sequences at the genomic level, using seven V_H segment primers (one specific to each V_H 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 V_H 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 V_H gene segment and ending with the conserved phenylalanine encoded by the 5′ portion of the J_H gene segment. The resulting sequences were analyzed for D_H usage, orientation, and RF, as described in the section on D_H detection. A total of 12,948,437 sequence reads representing 47,358 unique sequences, of which 82% 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 V_H, D_H, and J_H genes of B cells were extracted from the IMGT database (33). In our study, we used 34 germline sequences of D_H, categorized under functional genes that included open RF, 13 sequences of J_H genes, and 188 sequences of V_H 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 V_H and J_H of each sequence were determined by aligning all possible germline V_H and J_H (based on the IMGT germline library) (33) 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 V_H and J_H 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 V_H and J_H usage as well as the distance between V_H and J_H, 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 V_H and J_H. We then took all of the sequences with the same V_H, J_H, and distance between V_H and J_H and grouped them using a phylogenic approach. The distance between V_H and J_H was computed by positioning the IMGT germline V_H and J_H genes on the observed sequence and determining the distance between the last nucleotide of V_H and the first nucleotide of J_H. All the sequences with equal V_H, J_H, 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 one to each other, up to a distance of four mutations.

D_H detection

We have computed the maximal similarity between each D_H 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 V_H gene to the end of the J_H gene. The similarity of a given sequence to a germline D_H has been defined as the fraction of nucleotides equal to the one of the germline D_H. 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_H gene). In this dataset, the D_H 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_H gene, and 2) replacement of 1–3 nt in random positions in the D_H gene. In both methods, purines and pyrimidines had a two-thirds probability of being mutated into a base in the same group (transition) and a one-third probability to be mutated to the opposite group (transversion). We then computed the best fit D_H 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_H gene length: as expected, longer D_H genes were properly classified more often.

Results

Whereas the RF usage of V_H and J_H is constrained, the RF of D_H is flexible because nucleotide additions or deletions are present at both ends of the rearranged D_H gene segment. Furthermore, D_H genes can potentially undergo inversion. Thus, D_H 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_H gene actually uses a limited number of RFs.

The analysis of D_H gene sequences in the Ab repertoire is not easy because some D_H genes are similar to each other, and nibbling of the D_H genes by the nonhomologous end-joining machinery can result in very short D_H gene sequences. Thus, the robust identification of D_H 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_H 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_H gene determination using computer simulations of D_H rearrangements and found the results to be highly precise for long D_H 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_H 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 V_H and J_H 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_H 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 $CD20^+$ 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 V_H genes at all possible positions. Sequences were discarded from the analysis if a V_H gene could not be detected with a high enough accuracy (at least 70% overlap with germline V_H and at least 120 nt of V_H sequence length were required). Nearly all sequences had a much better fit than 70%. V_H genes that were highly related (e.g., alleles of the same V_H gene) were grouped into a single gene. We then compared the sequences downstream of the identified V_H with all possible germline J_H genes, and assumed that the J_H gene could not start >50 nt before the 3′ end of the V_H gene. We found that practically all sequences for which a V_H gene could be identified, the J_H gene could also be detected with a high enough accuracy (>70% similarity to germline V_H and J_H). Note that even if the precise V_H gene could not be identified, its final position could almost always be known precisely, because similar V_H genes have similar lengths. Given the position of V_H and J_H, we extracted the sequence regions that included the D_H gene segment(s), starting with the sequence that was 30 nt before the end of the V_H germline gene and ending with the 3′ end of the germline J_H sequence. We looked for the best fit to all germline D_H genes in the IMGT database. The optimal fitting D_H gene was given a score representing the fraction of the germline D_H gene nucleotides fitting the observed sequence (see Fig. 1). The position of the D_H gene was then determined relative

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
Illumina	1	F	White	49
	2	F	White	22

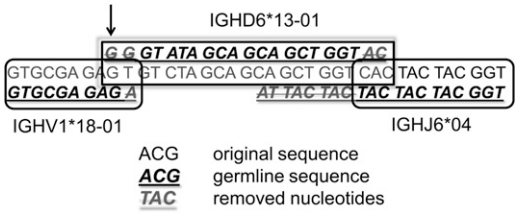

FIGURE 1. Computation of the DRF. We first compute the position of V_H (from ACGA to GAGA in the first row and the beginning of the second row). The upper line is the full sequence, and the lower line is the germline gene (V_H in this case). We then compute how the codons are aligned with respect to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

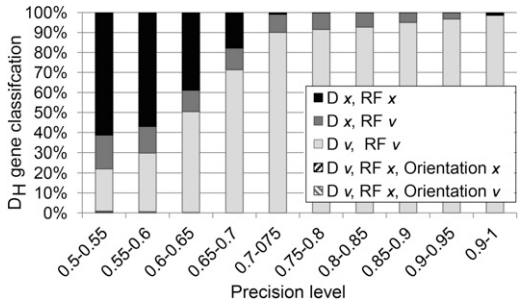

FIGURE 2. Effect of mutation frequency on D_H and DRF-identification accuracy. The *x*-axis is the fraction of nucleotides in the sampled sequence that overlaps with the full D_H germline sequence. The *y*-axis is the fraction of properly classified D_H genes in simulated rearrangements. The errors were divided into multiple categories, including errors in D_H identification or in the DRF classification or both. A detailed analysis of the effect of the D_H 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_H identification error is <10%. Most of those occur in highly similar D_H genes. RFs with a negative number are inverted RFs. Those are observed only in very short D_H genes.

to the end of the region where the sequence was identical to the corresponding germline sequence and relative to the beginning of the germline J_H sequence. The position determines the DRF relative to the beginning of the V_H germline sequence.

Note that many previous algorithms have been created for the detection of V_H, D_H, and J_H [e.g., among many others, Soda (37), IHMMune (38), and JOINSOLVER (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_H detection methodology properly identifies the D_H 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_H gene and three nucleotide changes/removed in the 3′ and 5′ ends of the D_H 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 misclassified ones, we checked the type of misclassification that occurred: either an error with respect to the RF or a misidentification of the D_H, or both. Finally, we checked whether the error was due to failure to detect an inverted D_H. We found that the error level is a function of the number of mutations in the D_H segment (nucleotide differences compared with the germline D_H sequence) and the number of nucleotides that are added and/or removed from the D_H gene. At or above a level of 75% overlap (the percentage of nucleotides overlapping between the query sequence and the germline D_H sequence), <10% of the query sequences are misclassified (Fig. 2). Note that even among these 10%, most errors are due to a misclassification of the D_H 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_H 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_H genes (Fig. 2). Therefore, we have restricted our analysis to D_H genes that can be identified with a precision of at least 0.75. Thus, this analysis excludes some of the shorter D_H 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_H used). We then computed how many unique clones (see *Materials and Methods*) used each RF for each D_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_H inversions may not really contain inverted D_H segments. Indeed, for each D_H gene found to be in the inverted DRF, there was also a reasonable fit to a D_H in a forward DRF (not always in the same D_H gene). Thus, we conclude [as was previously suggested based upon smaller datasets (22)] that the use of inverted D_H segments, if it occurs at all, is very infrequent in the IgH repertoire of healthy human adults.

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

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

Note that most inverted D_H do not contain stop codons. Thus, there is no a priori reason that these DRF should not be used. Their absence seems to highlight the presence of mechanistic differences in the inverted and forward rearrangements.

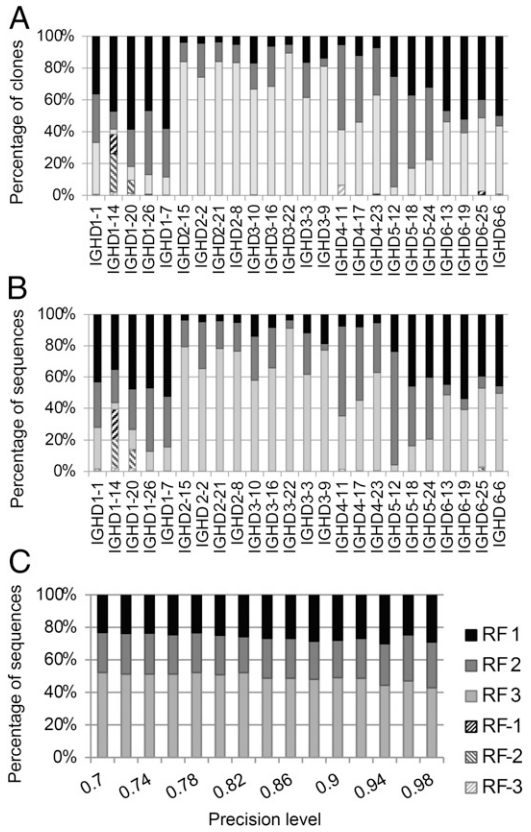

FIGURE 3. Fraction of sequences using each DRF in each D_H (454 data). (**A**) DRF usage of clone sequences. (**B**) DRF usage of all sequences. Each column is a D_H gene. Each shade is a DRF. We have removed IGHD4*04, IGHD5*05, and IGHD7*27, because we had too few sequences classified to these D_H genes. One can clearly see that the DRF usage varies significantly among D_H genes. The first D_H genes are very short and are more error prone than the others. (**C**) Effect of precision on DRF usage (average over different D_H genes). The *x*-axis is the required precision (fraction of overlapping nucleotides between a sequenced D_H with the germline D_H). The *y*-axis is the percentage of sequences using each DRF. The DRF usage is similar at different levels of precision.

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 distribution thus seems to hint to the presence of a clear selection mechanism for one DRF for a given D_H segment.

To check that this result is not the peculiarity of a single individual, 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 different individuals. Thus, if specific DRF are indeed selected, this selection mechanism is 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 product that contains the DRF is internal to and nonoverlapping with the primers.

As mentioned above, the frequencies at which D_H genes are used in our sample vary widely among D_H genes. Thus, the average results could be the result of a small number of highly frequent D_H genes. As was the case for the inverted DRF, we have repeated the analysis separating each D_H. The DRF usage varies among D_H genes, but it is again highly nonuniform for each D_H gene (Fig. 3A). The nonuniform distribution is not the result of stop codons, because DRFs without stop codons are practically absent. Moreover, 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 D_H (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 amplification errors or biases, because those would not affect the clone number (remember that similar sequences comprising each clone 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 selection occurs, it is not the result of the amplification of some clones, but rather the selection of individual clones. Moreover, the RACE protocol minimizes artifactual skewing of V_H or J_H gene usage because the primers are neither V_H nor J_H specific.

To test that the total DRF usage is not the result of errors (that comprise $<10\%$ of classified sequences) in either the D_H 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 D_H 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 D_H gene, the results may be biased toward a given DRF. However, the comparison with the D_H gene is only performed with the germline and is not affected by additions around the D_H gene, and the DRF preference is also observed in CDR3 sequences where the distance between V_H and D_H, and between D_H and J_H 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 3×18 frequency matrix (3 DRF and 18 of 34 D_H 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-

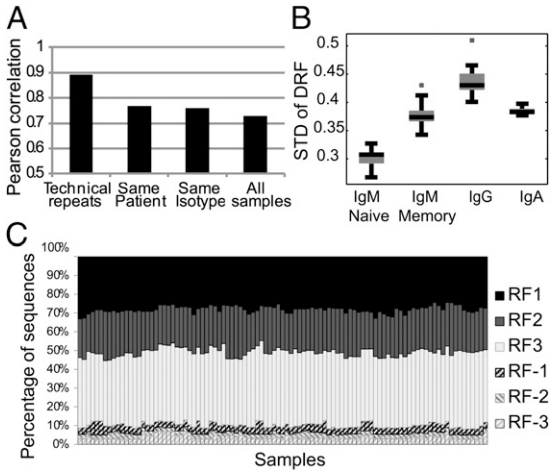

FIGURE 4. DRF usage in different donors and isotypes. (**A**) Average correlation between DRF usage within and between donors. The correlation is computed as the Pearson correlation of the frequency of each DRF in each D_H gene among samples. The highest correlation is among technical repeats, followed 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.

lation 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* 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 D_H 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 ~0.3, and it rises progressively to 0.37 in the memory IgM repertoire and to 0.43, almost the maximal variability (i.e., 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 D_H 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^+ sequences, where DRF usage is limited to practically a single DRF (*t* 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 D_H 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 J_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-J_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

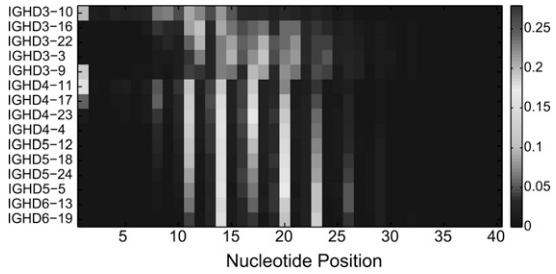

FIGURE 5. Frequency of position of end of D_H for each D_H gene. Each column is a position at the end of D_H starting from 10 positions into V_H. Each row is a D_H gene. The shades represent frequency, as shown in the grayscale bar. One can clearly see in most D_H genes jumps of three in the position, but, beyond the clear selection for a DRF, the position distribution is quite wide.

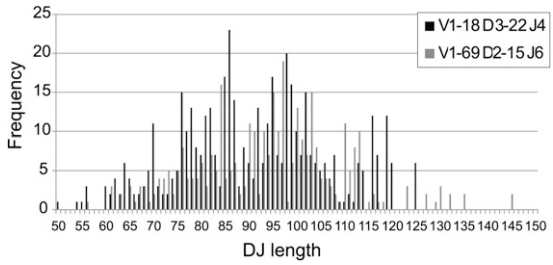

FIGURE 6. DJ region length distribution for different VDJ rearrangement in naive B cell samples. We computed in the clone sequences the length distribution of the region spanning the full germline D_H and J_H regions.

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 D_H genes in the forward direction. Indeed, the OF variance is not very different from what was expected due to random chance in most D_H 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.

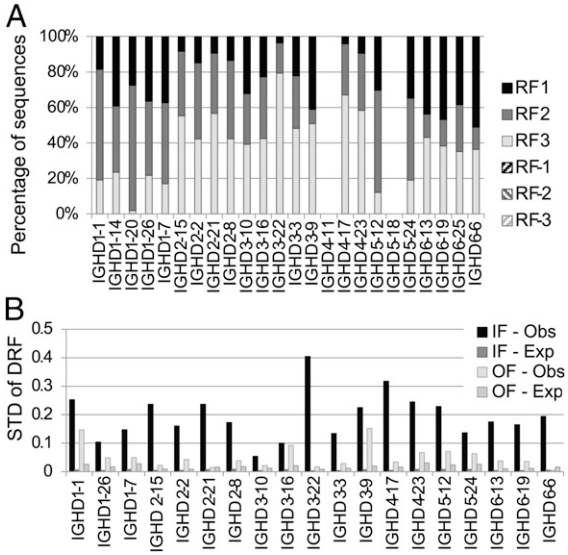

FIGURE 7. DRF usage in the Illumina data. (**A**) Fraction of sequences using each DRF in each D_H. Each column is a D_H gene. Each shade is a DRF. One can clearly see the similar DRF usage as observed in Fig. 3. Differences may arise due to the fact that different donors were recruited for each sequencing. (**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 = root of (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. D_H genes for which there were not enough OF or IF sequences (<100) were not incorporated in the analysis.

The junction between D_H and J_H exhibits a small degree of reading frame bias compared with DRF bias

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

Because the Dμ protein is generated by rearrangement between D_H and J_H 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 Dμ 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 Dμ 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 D_H gene and tend to be eliminated through nucleotide deletion and addition. DRF usage exhibited a low correlation with

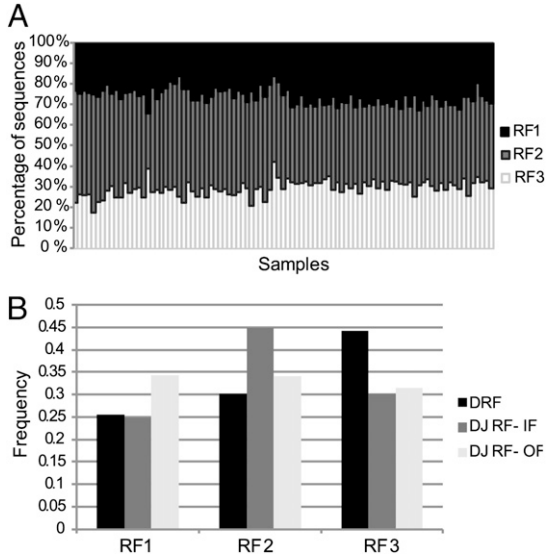

FIGURE 8. DJ rearrangement RF. (**A**) DJ rearrangement RF usage of all isotypes and donors (454 data). As in Fig. 6, we used the segment containing the D_H, N2, and J_H regions, and replaced the D_H and J_H with the full germline sequences. We then computed the RF of this segment with respect to the beginning of the C region. There is a general small bias in favor of the second RF (RF 2). (**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 D_H 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 D_H 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_H genes, represented as a single 54-position vector) and compared it with an occupancy matrix for each amino acid in each D_H in each DRF. Thus, each amino acid was assigned a (3*18) matrix representing the number of times it appears in each D_H 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.

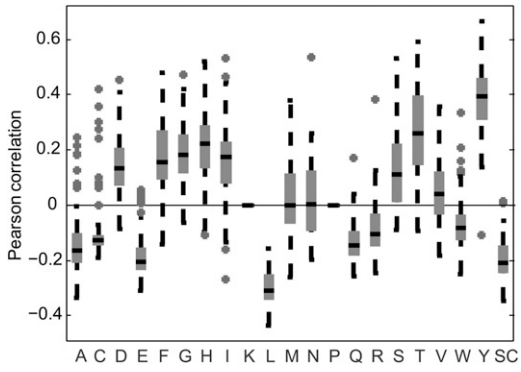

FIGURE 9. Pearson correlation of DRF usage with the number of times each amino acid appears in each RF. The *x*-axis represents amino acids. The values on the *y*-axis are the Pearson correlation coefficients of the DRF usage and the amino acid frequency vectors. Only correlations with a p value ≤ 0.05 were used. Lysine and proline are practically never used in D_H 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.

Effect of DRF and D_H gene usage on D_H gene position

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

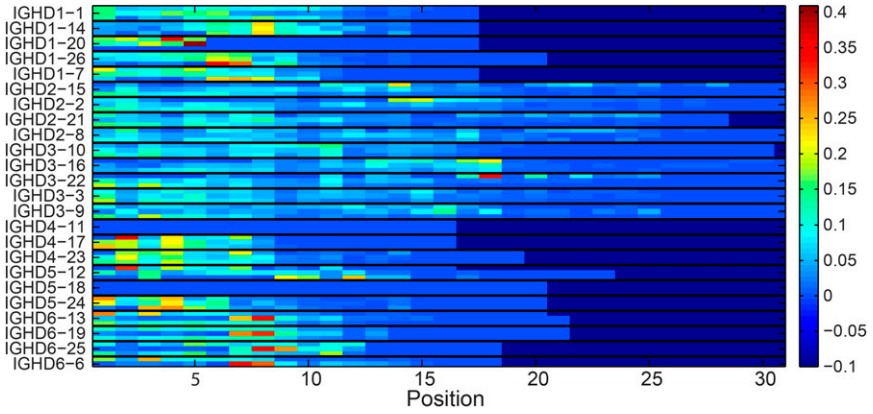

FIGURE 10. Relative frequency of the starting position in the D_H gene. Each position in the *x*-axis is a nucleotide along the D_H gene, and each triplet of rows represents a D_H gene. Within each triplet, each row is a DRF. The colors represent the frequency of this position in the DNA samples. Only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

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 D_H 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

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

D_H Gene	DRF	Starting Position (Nucleotide)	Nucleotide Sequence (5′→3′)	Amino Acid Sequence
IGHD1-20	1	4	ATAACTGGAACGAC	ITGT
IGHD1-20	2	5	AACTGGAACGAC	NWND
IGHD1-26	3	6	TGGGAGCTACTAC	WELL
IGHD3-22	1	18	TGGTTATTACTAC	WLLL
IGHD4-17	1	2	CTACGGTGACTAC	LR*L
IGHD6-13	2	8	CAGCAGCTGGTAC	QQLV
IGHD6-19	2	8	CAGTGGCTGGTAC	QWLV
IGHD6-25	2	8	CAGCGGCTAC	QRL
IGHD6-6	2	7	CAGCTCGTCC	QLV
IGHD6-6	2	8	CAGCTCGTCC	QLV

The first column is the D_H gene; the second column the DRF; the third column is the position in the D_H gene used as a beginning of the D_H gene. The last columns represent the nucleotide and amino acids of the resulting D_H 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 D_H 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 D_H is seen over a large range of rearrangement lengths. This suggests that there is an important region at the core of the D_H gene sequence that may be getting selected. We show that the Dμ 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 D_H 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 germline D_H sequence. Because D_H 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 D_H 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

1. Blackwell, T. K., and F. W. Alt. 1989. Mechanism and developmental program of immunoglobulin gene rearrangement in mammals. *Annu. Rev. Genet.* 23: 605–636.
2. Fugmann, S. D., A. I. Lee, P. E. Shockett, I. J. Villey, and D. G. Schatz. 2000. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. *Annu. Rev. Immunol.* 18: 495–527.
3. Nemazee, D. 2006. Receptor editing in lymphocyte development and central tolerance. *Nat. Rev. Immunol.* 6: 728–740.
4. Nemazee, D., and M. Weigert. 2000. Revising B cell receptors. *J. Exp. Med.* 191: 1813–1817.
5. Kabat, E. A., T. T. Wu, and H. Bilofsky. 1978. Variable region genes for the immunoglobulin framework are assembled from small segments of DNA—a hypothesis. *Proc. Natl. Acad. Sci. USA* 75: 2429–2433.
6. Schroeder, H. W., Jr., M. A. Walter, M. H. Hofker, A. Ebens, K. Willems van Dijk, L. C. Liao, D. W. Cox, E. C. Milner, and R. M. Perlmutter. 1988. Physical linkage of a human immunoglobulin heavy chain variable region gene segment to diversity and joining region elements. *Proc. Natl. Acad. Sci. USA* 85: 8196–8200.
7. Volpe, J. M., and T. B. Kepler. 2008. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res.* 4: 3.
8. Louzoun, Y., P. E. Luning, S. Litwin, and M. Weigert. 2002. D is for different: differences between H and L chain rearrangement. *Semin. Immunol.* 14: 239-241.
9. Schelonka, R. L., M. Zemlin, R. Kobayashi, G. C. Ippolito, Y. Zhuang, G. L. Gartland, A. Szalai, K. Fujihashi, K. Rajewsky, and H. W. Schroeder, Jr. 2008. Preferential use of DH reading frame 2 alters B cell development and antigen-specific antibody production. *J. Immunol.* 181: 8409–8415.
10. Zemlin, M., R. L. Schelonka, G. C. Ippolito, C. Zemlin, Y. Zhuang, G. L. Gartland, L. Nitschke, J. Pelkonen, K. Rajewsky, and H. W. Schroeder, Jr. 2008. Regulation of repertoire development through genetic control of DH reading frame preference. *J. Immunol.* 181: 8416–8424.
11. Gu, H., D. Kitamura, and K. Rajewsky. 1991. B cell development regulated by gene rearrangement: arrest of maturation by membrane-bound D mu protein and selection of DH element reading frames. *Cell* 65: 47–54.
12. Reth, M. G., and F. W. Alt. 1984. Novel immunoglobulin heavy chains are produced from DJH gene segment rearrangements in lymphoid cells. *Nature* 312: 418–423.
13. Alt, F. W., and D. Baltimore. 1982. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc. Natl. Acad. Sci. USA* 79: 4118–4122.
14. Meek, K. D., C. A. Hasemann, and J. D. Capra. 1989. Novel rearrangements at the immunoglobulin D locus: inversions and fusions add to IgH somatic diversity. *J. Exp. Med.* 170: 39–57.
15. Sollbach, A. E., and G. E. Wu. 1995. Inversions produced during V(D)J rearrangement at IgH, the immunoglobulin heavy-chain locus. *Mol. Cell. Biol.* 15: 671–681.
16. Radic, M. Z., J. Mackle, J. Erikson, C. Mol, W. F. Anderson, and M. Weigert. 1993. Residues that mediate DNA binding of autoimmune antibodies. *J. Immunol.* 150: 4966–4977.
17. Nemazee, D. 2000. Receptor editing in B cells. *Adv. Immunol.* 74: 89–126.
18. Chen, C., Z. Nagy, E. L. Prak, and M. Weigert. 1995. Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing. *Immunity* 3: 747–755.
19. Prak, E. L., M. Trounstine, D. Huszar, and M. Weigert. 1994. Light chain editing in kappa-deficient animals: a potential mechanism of B cell tolerance. *J. Exp. Med.* 180: 1805–1815.
20. Gauss, G. H., and M. R. Lieber. 1992. The basis for the mechanistic bias for deletional over inversional V(D)J recombination. *Genes Dev.* 6: 1553–1561.
21. Ohm-Laursen, L., M. Nielsen, S. R. Larsen, and T. Barington. 2006. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VH replacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology* 119: 265–277.
22. Corbett, S. J., I. M. Tomlinson, E. L. Sonnhammer, D. Buck, and G. Winter. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, "minor" D segments or D-D recombination. *J. Mol. Biol.* 270: 587–597.
23. Souto-Carneiro, M. M., N. S. Longo, D. E. Russ, H. W. Sun, and P. E. Lipsky. 2004. Characterization of the human Ig heavy chain antigen binding complementarity determining region 3 using a newly developed software algorithm, JOINSOLVER. *J. Immunol.* 172: 6790–6802.
24. Briney, B. S., J. R. Willis, M. D. Hicar, J. W. Thomas, II, and J. E. Crowe, Jr. 2012. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* 137: 56–64.
25. ten Boekel, E., F. Melchers, and A. G. Rolink. 1997. Changes in the V(H) gene repertoire of developing precursor B lymphocytes in mouse bone marrow mediated by the pre-B cell receptor. *Immunity* 7: 357–368.
26. Meng, W., L. Yunk, L. S. Wang, A. Maganty, E. Xue, P. L. Cohen, R. A. Eisenberg, M. G. Weigert, S. J. Mancini, and E. T. Prak. 2011. Selection of individual VH genes occurs at the pro-B to pre-B cell transition. *J. Immunol.* 187: 1835–1844.
27. Cohen, R. M., S. H. Kleinstein, and Y. Louzoun. 2011. Somatic hypermutation targeting is influenced by location within the immunoglobulin V region. *Mol. Immunol.* 48: 1477–1483.
28. Dörner, T., H. P. Brezinschek, R. I. Brezinschek, S. J. Foster, R. Domiati-Saad, and P. E. Lipsky. 1997. Analysis of the frequency and pattern of somatic mutations within nonproductively rearranged human variable heavy chain genes. *J. Immunol.* 158: 2779–2789.
29. Xue, W., S. Luo, W. H. Adler, D. H. Schulze, and J. E. Berman. 1997. Immunoglobulin heavy chain junctional diversity in young and aged humans. *Hum. Immunol.* 57: 80–92.

30. Glanville, J., W. Zhai, J. Berka, D. Telman, G. Huerta, G. R. Mehta, I. Ni, L. Mei, P. D. Sundar, G. M. R. Day, et al. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* 106: 20216–20221.
31. Larimore, K., M. W. McCormick, H. S. Robins, and P. D. Greenberg. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.* 189: 3221–3230.
32. Yousfi Monod, M., V. Giudicelli, D. Chaume, and M. P. Lefranc. 2004. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics* 20(Suppl. 1): i379–i385.
33. Lefranc, M. P., V. Giudicelli, C. Ginestoux, J. Bodmer, W. Müller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbié, and D. Chaume. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* 27: 209–212.
34. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32: 1792–1797.
35. Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980–984.
36. Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406–425.
37. Volpe, J. M., L. G. Cowell, and T. B. Kepler. 2006. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics* 22: 438–444.
38. Gaëta, B. A., H. R. Malming, K. J. L. Jackson, M. E. Bain, P. Wilson, and A. M. Collins. 2007. iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics* 23: 1580–1587.
39. Guo, B., R. M. Kato, M. Garcia-Lloret, M. I. Wahl, and D. J. Rawlings. 2000. Engagement of the human pre-B cell receptor generates a lipid raft-dependent calcium signaling complex. *Immunity* 13: 243–253.
40. Hayashi, K., M. Yamamoto, T. Nojima, R. Goitsuka, and D. Kitamura. 2003. Distinct signaling requirements for Dmu selection, IgH allelic exclusion, pre-B cell transition, and tumor suppression in B cell progenitors. *Immunity* 18: 825–836.
41. Cohn, M., and R. E. Langman. 1990. The protection: the unit of humoral immunity selected by evolution. *Immunol. Rev.* 115: 7–147.
42. Louzoun, Y., T. Friedman, E. Luning Prak, S. Litwin, and M. Weigert. 2002. Analysis of B cell receptor production and rearrangement* 1. Part I. Light chain rearrangement. *Semin Immunol* 14: 169–190.
43. Li, Y., Y. Louzoun, and M. Weigert. 2004. Editing anti-DNA B cells by Vlambdax. *J. Exp. Med.* 199: 337–346.
44. Li, H., Y. Jiang, E. L. Prak, M. Radic, and M. Weigert. 2001. Editors and editing of anti-DNA receptors. *Immunity* 15: 947–957.
45. Wardemann, H., S. Yurasov, A. Schaefer, J. W. Young, E. Meffre, and M. C. Nussenzweig. 2003. Predominant autoantibody production by early human B cell precursors. *Science* 301: 1374–1377.
46. Kabat, E. A., and T. T. Wu. 1991. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities: relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J. Immunol.* 147: 1709–1719.