The Journal of Immunology

Characterization of the Human Ig Heavy Chain Antigen Binding Complementarity Determining Region 3 Using a Newly Developed Software Algorithm, JOINSOLVER

M. Margarida Souto-Carneiro, Nancy S. Longo, Daniel E. Russ, Hong-wei Sun and Peter E. Lipsky

J Immunol 2004; 172:6790-6802; ;
doi: 10.4049/jimmunol.172.11.6790
http://www.jimmunol.org/content/172/11/6790

The Journal of Immunology is published twice each month by
The American Association of Immunologists, Inc.,
1451 Rockville Pike, Suite 650, Rockville, MD 20852

Print ISSN: 0022-1767 Online ISSN: 1550-6606.

Characterization of the Human Ig Heavy Chain Antigen Binding Complementarity Determining Region 3 Using a Newly Developed Software Algorithm, JOINSOLVER

M. Margarida Souto-Carneiro,[1]* Nancy S. Longo,[1]* Daniel E. Russ,[1†] Hong-wei Sun,[‡] and Peter E. Lipsky[2]*

We analyzed 77 nonproductive and 574 productive human V_HDJ_H rearrangements with a newly developed program, JOINSOLVER. In the productive repertoire, the H chain complementarity determining region 3 ($CDR3_H$) was significantly shorter (46.7 ± 0.5 nucleotides) than in the nonproductive repertoire (53.8 ± 1.9 nucleotides) because of the tendency to select rearrangements with less TdT activity and shorter D segments. Using criteria established by Monte Carlo simulations, D segments could be identified in 71.4% of nonproductive and 64.4% of productive rearrangements, with a mean of 17.6 ± 0.7 and 14.6 ± 0.2 retained germline nucleotides, respectively. Eight of 27 D segments were used more frequently than expected in the nonproductive repertoire, whereas 3 D segments were positively selected and 3 were negatively selected, indicating that both molecular mechanisms and selection biased the D segment usage. There was no bias for D segment reading frame (RF) use in the nonproductive repertoire, whereas negative selection of the RFs encoding stop codons and positive selection of RF2 that frequently encodes hydrophilic amino acids were noted in the productive repertoire. Except for serine, there was no consistent selection or expression of hydrophilic amino acids. A bias toward the pairing of 5′ D segments with 3′ J_H segments was observed in the nonproductive but not the productive repertoire, whereas V_H usage was random. Rearrangements using inverted D segments, DIR family segments, chromosome 15 D segments and multiple D segments were found infrequently. Analysis of the human $CDR3_H$ with JOINSOLVER has provided comprehensive information on the influences that shape this important Ag binding region of V_H chains. ***The Journal of Immunology,*** **2004, 172: 6790–6802.**

Diversity in the Ab repertoire of Ig H chains is mainly achieved by random recombination of V_H, D, and J_H segments and enzymatic modification of the V_HDJ_H junctions. Located at the joining of the V_H, D, and J_H segments, the H chain complementarity determining region 3 ($CDR3_H$)[3] is the most diverse region in the Ig molecule. Structurally, the $CDR3_H$ is in the center of the Ag binding site, interacting directly with the other CDRs and framework regions both from H and L chains, as well as with the Ag itself (1–5). Changes in the $CDR3_H$ amino acid composition directly affect the charge, hydrophobicity, size, and shape of the Ag binding site (4, 5), and therefore, the ability of the Ab molecule to bind Ag.

Despite its crucial role in determining the nature and specificity of the Ag binding capability of the Ab molecule, the human $CDR3_H$ has not yet been fully characterized. One reason for this lack of information about the $CDR3_H$ relates to the difficulty in analyzing the sequences of this highly diverse region. Specific problems in identifying the D segment used in the components of the $CDR3_H$ relate to the overall similarity of the germline D segment sequences and the extensive exonuclease and TdT modification of the D segment. This has made precise identification of the components of the $CDR3_H$ difficult. Adding to the difficulty of analyzing the composition of the $CDR3_H$ have been the underlying assumptions of the analytical instruments used. Available software tools for Ig gene analysis such as, DNAPLOT (Centre for Protein Engineering, http://www.mrc-cpe.cam.ac.uk) and the Immunogenetic database (IMGT; 6), which assess the germline genes that most closely match the given gene sequence, use an alignment scoring system (7). This method is usually straightforward in the V_H and J_H regions, where there are large regions of sequence similarity. However, in the shorter D region, where mutations and terminal processing is common, this method is less successful. To assess the D segment more accurately, we have used a more intuitive scoring system to match D segments based upon consecutive nucleotide matches. This consecutive match scoring approach assigns a higher score for longer matches, and searches for a D segment core around which mutation or terminal processing occurs. This approach, along with an automated analytical instrument, JOINSOLVER, has made it possible to analyze a large number of human $CDR3_H$ sequences and to begin to understand the influences that shape this important Ag-binding region of Ab molecules.

It is notable that previous attempts to describe the $CDR3_H$ have often yielded conflicting results. For example, some reports claim the existence of D segment fusion in both human (8–11) and murine (12–14) V_HDJ_H sequences, the usage of DIR segments (8, 10,

*Repertoire Analysis Group, Autoimmunity Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, †Division of Computational Bioscience, Center for Information Technology, and ‡Biodata Mining and Discovery Section, Office of Science and Technology, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892

Received for publication October 16, 2003. Accepted for publication March 18, 2004.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[1] M.M.S.-C., N.S.L., and D.E.R. contributed equally to this work.

[2] Address correspondence and reprint requests to Dr. Peter E. Lipsky, Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, 9000 Rockville Pike, Building 10, Room 9N228, Bethesda, MD 20892-1820. E-mail address: LipskyP@mail.nih.gov

[3] Abbreviations used in this paper: CDR3H, H chain complementarity determining region 3; RF, reading frame; AR, amino acid residue; BLAST, basic local alignment search tool; Pr, palindromic nucleotide; RSS, recombination signal sequence; RAG, recombination-activating gene.

0022-1767/04/$02.00

15–17), and inverted D segments (10, 18, 19), whereas other analytic approaches have come to the conclusion that these are rare events in human sequences (20) and that D-D fusion is rare in mouse sequences (21). This is particularly important because it has been claimed in the mouse that the use of inverted D segments and DD fusions predispose to autoantibody formation (9, 13). Indeed, even the definition of the D segment is controversial, with some analyses using shorter consecutive nucleotide matches (8, 22–26), or allowing one mismatch (8, 25, 26), and others defining the D segment only when a match of 10 consecutive nucleotides is present (20).

In an effort to resolve some of these issues, we developed a new computer algorithm, JOINSOLVER, and used it to analyze a set of 650 V_HDJ_H sequences amplified from normal adults. Monte Carlo simulations were used to establish the required length of a D segment match to establish identity. The use of these approaches has permitted a better understanding of the variability inherent in the human $CDR3_H$.

Materials and Methods

Compilation of the database

The following Ig H chain sequences were analyzed: 1) a set of 400 sequences from genomic DNA obtained by single cell sorting of normal human adult peripheral $CD5^+/IgM^+$ and $CD5^-/IgM^+$ B cells (GenBank accession numbers Z80363-770); 2) a set of 116 sequences from cDNA obtained by single cell sorting of human tonsilar IgM^+ and IgA^+ $CD3^-/CD19^-/CD38^{+++}$ plasma cells (GenBank accession numbers AY003749-869); 3) a set of 135 sequences from cloned cDNA of human adult peripheral IgM^+ and IgG^+ B cells (GenBank accession numbers Z68345-487).

Algorithm used by JOINSOLVER for sequence analysis

A software tool, JOINSOLVER, was developed specifically to analyze the $CDR3_H$ region of the Ig genes expressed by human B cells and is available at http://joinsolver.niams.nih.gov. The strategy of JOINSOLVER is to search for D germline sequences flanking V_H and J_H germline genes. Additionally, it searches for P and N nucleotide additions in the V_HD and D_HJ junctions. The database of human D germline genes used includes all D segments from the IMGT databank (6) as well as the reverse and DIR germline genes.

JOINSOLVER initially interrogates the sequence to find the beginning of the $CDR3_H$ region that is defined as codon 93. This codon was used to define the beginning of the CDR3 based on the results of structural analyses of V_HDJ_H rearrangements (27, 28) as recommended (2–4). To identify this codon, JOINSOLVER searches for the sequence, "TAT TAC TGT", which comprises codons 90–92 of the V_H region (after Kabat et al. (1)) and is a conserved motif in most of the human V_H germline genes. If a "TAT TAC TGT" motif is not found, the search is reinitiated with 1-bp change allowed in the sequence. If a "TAT TAC TGT" with one nucleotide change is not found, then homologies with the germline genes are used to find the most likely start of the $CDR3_H$ region. If the start of the $CDR3_H$ region is not yet identified, JOINSOLVER marks the $CDR3_H$ as not found and defers finding the $CDR3_H$ region until after V and J matching.

After the V_H end of the $CDR3_H$ is defined, JOINSOLVER screens for the J_H border of the $CDR3_H$. A "C TGG GG" motif demarks the 3′ end of the $CDR3_H$ region and is conserved in all J_H sequences. A similar algorithm is used to find the "C TGG GG" at the 3′ end of the $CDR3_H$.

Once the $CDR3_H$ region is identified, V_H, J_H, and D assignment is conducted. The V region is matched to a database of germline genes from the "TAT TAC TGT" back 3′→ 5′ toward the beginning of the sequence, and forward in the 5′→ 3′ direction to the end of the germline gene. The J_H region is matched from the "C TGG GG" back to the beginning of the germline gene and forward until the end of the sequence or the end of the germline gene is identified. The V_H and J_H regions are scored with an alignment score that assigns a +5 to a nucleotide match and −4 for a mismatch between the unknown sequence and the germline (7).

The end of the V_H region is identified when the given unknown sequence matches the complete V_H germline gene or has a mismatch after the "TAT TAC TGT" with the highest scoring V_H germline. The beginning of the J_H region is defined when the unknown sequence has one mismatch before the "C TGG GG" with the highest scoring J_H region or the sequence matches the complete J_H germline gene.

In the event that the $CDR3_H$ was initially not found, JOINSOLVER looks for matches between the V and J germline databases and the unknown sequence. The unknown sequence is aligned to the highest scoring germline genes. The $CDR3_H$ region is defined as the region from codon 93 and the "C" of the "C TGG GG" motif. The V_H end and J_H start are defined the same way as if the $CDR3_H$ region had been found first.

After V_H and J_H assignment, D segment assignment is conducted using a consecutive match scoring system. All matches to the D germline genes are scored and sorted based on the V_H-J_H distance (the distance in nucleotides between the end of the V_H segment and the beginning of the J_H segment). The longest matches are aligned and returned to the user.

Monte Carlo simulation for D segment assignment

A Monte Carlo simulation was used to determine the probability of matching a randomly generated sequence of length, m_l, to the database of known human D germline genes. To accomplish this, a randomly generated set of 1×10^5 sequences of a particular length, g_l, was analyzed and searched for matches between the sequences and the D segment germline database. The value of g_l is equivalent to the V_H-J_H distance. The error in this simulation is ~1 over the square root of 100,000 or 0.00316.

Monte Carlo simulation for multiple D segment fusion assignment

A second Monte Carlo simulation was conducted using 1×10^6 randomly generated sequences for different values of g_l, m_{l1}, and m_{l2}, where m_{l1} and m_{l2} are the lengths of the first and second D matches, respectively. Because it is less likely that a longer match is random, the longer match was assigned as the first match. One million random sequences were analyzed to reduce the error to a maximum of ±0.001.

Sequence analysis

Rearrangements were considered productive if the V_HDJ_H junction maintained the reading frame (RF) into the J_H segment and contained no stop codons in the germline D segment or $CDR3_H$ junctions. When the rearrangements failed to maintain the RF into the J_H segment, or introduced stop codons during the rearrangement process, they were considered nonproductive. Junctional nucleotide additions between the V_H and D or between D and J_H segments were scored as: 1) P nucleotides, if they were inverted repeats at germline encoded ends; 2) N nucleotides, if they were nontemplated junctional additions. The junctions without N nucleotides which contained nucleotides that could not be unequivocally assigned to either coding end, were considered to be microhomologies. In cases where the nucleotide sequence between the V_H and J_H coding ends had the same number of matches with a DIR family member (17) or a D segment encoded on chromosome 15 and a conventional D segment, the latter was accepted as the D element used. Rearrangements using DD fusions, inverted, or DIR segments were excluded from the D segment RF analysis.

Basic statistics

To determine significant differences in distributions in productive and nonproductive rearrangements, the χ^2 test was used. Values of $p \leq 0.05$ were assumed to be significant. The statistical significance between observed and expected frequencies in D genes and D RFs was calculated using the χ^2 goodness-of-fit test. The Student t test was used to analyze $CDR3_H$ length; V_H-J_H distance; D segment match length; P, and N nucleotides; V_H, D, and J_H excision.

AR composition analysis of $CRD3_H$ and D-segment sequences

For this analysis, the first two amino acid residues (AR, codons 93–94 according to numbering by Kabat et al. (1)) and the last two (DX, codons 101–102) of the predicted $CDR3_H$ segment were not included to assess the nonrandom characteristics of the $CDR3_H$. The total numbers of productive $CDR3_H$, nonproductive $CDR3_H$, productive D segments, and nonproductive D segments included in the analysis are 563, 75, 390, and 60, respectively. To compare these amino acid sequences to that expected from random chance, random sequences were generated using the Genetics Computer Group SAMPLE program (Wisconsin Package version 10.2; Accelrys, San Diego, CA). Using Swiss-Prot release 38.0 and a sampling rate of 18, three sets of 523 randomly sampled human sequences were obtained consisting of a productive $CDR3_H$ length of 12, nonproductive $CDR3_H$ length of 14, and a length of 5 for both productive and nonproductive D segments. Residue compositions were calculated by using the COMPOSITION program (Genetics Computer Group, Wisconsin Package version 10.2; Accelrys). χ^2 analysis was performed with the statistics program R. Distribution change of a residue was determined to be significant

Table I. *Minimal D segment match length for a particular V_H-J_H distance, when the probability of a random match is 5 or 1%*[a]

V_H-J_H Distance (bp)	Match Length (m_{l1}) Required (bp)	
	5%	1%
8	8	
9–11	8	9
12–23	9	10
24–27	9	11
28–75	10	11
76–79	10	12

[a] The data show the condition for finding a match of length m_{l1} in a sequence with a particular V_H-J_H length (g_l) with an approximate error of ±0.3%. m_{l1} is the match length from the D segment and 5% and 1% represent the match length required for a 95% and a 99% probability that the match is not from random chance.

if its contribution was >5% of the total χ^2 sum. Where necessary, residues K, R, H or N, Q, S were combined to perform a χ^2 test properly.

Results

Length of D segment match depends on V_H-J_H distance

Based on the Monte Carlo simulation, the minimal D segment match length required for identification was found to depend on the V_H-J_H distance (g_l). Eight to 11 consecutively matching base pairs were necessary to identify a D segment with sufficiently high probability that it is unlikely to be from random chance (Table I).

Consecutive matches are an effective means to score the D segment

JOINSOLVER used a consecutive matching algorithm rather than the typical alignment scoring system to identify D segments. To compare the results obtained from these approaches, JOINSOLVER and DNAPLOT (Centre for Protein Engineering, http://www.mrc-cpe.cam.ac.uk) were used to analyze a specific sequence, Z80389. Only the sequence flanking the $CDR3_H$ is shown (Fig. 1). The DNAPLOT method (Fig. 1*B*) selected the germline gene D6-25 on the basis of having a good overall match with a basic alignment search tool (BLAST) alignment score of 63. Nucleotides in the unknown sequence were identical to 15 of 18 nucleotides in the D6-25 germline sequence. However, consecutively matching nucleotides were interrupted twice by a single mismatch and the longest consecutive match consisted of only 7 nucleotides. JOINSOLVER identified a better match by applying the consecutively matching algorithm and limiting the search to the appropriate region by excluding any putative D segment alignment in regions previously identified as V_H and J_H segments. JOINSOLVER selected germline gene D2-2 as the best match, with 13 consecutively matching nucleotides within the 51 nucleotide V_H-J_H region.

D segment matching using JOINSOLVER consecutive matching system yields different alignments than the DNAPLOT BLAST scoring system

To compare D segment alignment results from DNAPLOT and JOINSOLVER, we analyzed the D segments in 144 randomly selected unmutated and mutated V_HDJ_H rearrangements (Z80363-Z80511). Within this subset, 74% of the sequences had zero to two mutations (98–100% V_H germline homology) and 9% had more than two mutations (<97% V_H germline homology). JOINSOLVER and DNAPLOT gave comparable D segment gene matches in 50% of the rearrangements (Table II). However, JOINSOLVER performed better than DNAPLOT for 22% of the rearrangements, either by finding a D segment with a longer consecutive nucleotide match than that found by DNAPLOT or because DNAPLOT failed to find any match for the D segment. Occasionally (5.5% of the rearrangements), the highest scoring D gene identified by JOINSOLVER was also found by DNAPLOT, but it appeared with the second highest overall matching score by DNAPLOT and, therefore, was not considered to be the best match. Overall, JOINSOLVER identified the D segment in 74% of the rearrangements in this subanalysis. Furthermore, JOINSOLVER generated some (albeit not significant) D segment matching for every sequence, whereas, DNAPLOT failed to give any D segment alignment in 25% of the rearrangements. Importantly, in 47% of the rearrangements for which DNAPLOT failed to find a D segment match, JOINSOLVER was able to align the D segment.

Multiple D segment fusions are dependent on the V_H-J_H distance and the match lengths of each D segment

In some sequences, more than one possible nonoverlapping D match was found in the $CDR3_H$ region. To determine whether these can be explained by random chance, a second Monte Carlo simulation was performed to examine the conditional probability of having a second match of length m_{l2} given that a first match of length m_{l1} exists. Identifying a second D match with high probability depends on both the V_H-J_H length (g_l) and the length of the first match (m_{l1}). When the V_H-J_H distance is greater than 26 nucleotides, a match of 9 nucleotides or more is necessary to identify a second D segment with confidence if the first match is 9 or more nucleotides. When the V_H-J_H distance is 26 nucleotides or less, matches of 8 nucleotides can be used to identify a second D match with confidence when the first match is 9 or more nucleotides. Finally, when the V_H-J_H distance is 18 nucleotides or less, matches of 7 nucleotides can be used to identify a second D match, when the first match is 9 or more nucleotides. When the V_H-J_H distance is 17 nucleotides or less, a first match of 8 nucleotides and a second match of 7 nucleotides is sufficient to identify two D segments.

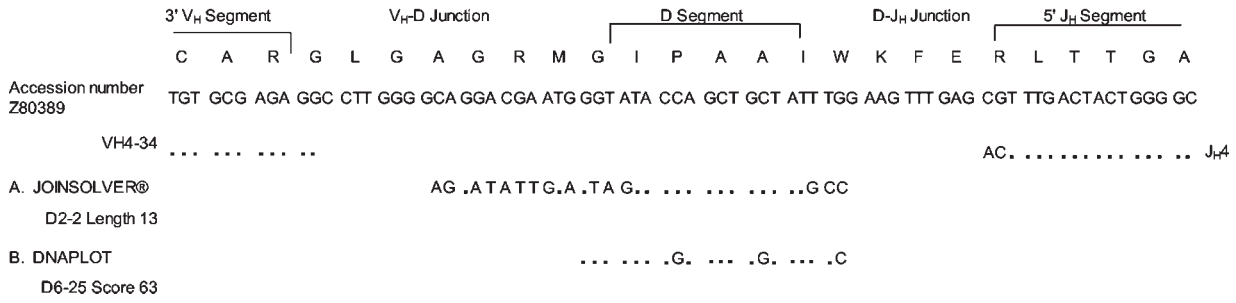

FIGURE 1. Alignment for a D segment in a nonproductive rearrangement. Comparison of a consecutive match system (*A*, JOINSOLVER) and an alignment scoring system (*B*, V-Base/DNAPLOT) to correctly identify the D segment in sequence Z80389. Although it assigns a high alignment score, the V-Base/DNAPLOT D segment choice has only seven consecutive matches, whereas, for the same rearrangement, JOINSOLVER matches 13 consecutive base pairs.

Table II. *Comparison of the accuracy of the JOINSOLVER consecutive match approach and the V-Base/ DNAPLOT alignment scoring system*[a]

Accession Number	V_H-J_H Distance (bp)	JOINSOLVER		V-Base/ DNAPLOT		
		D segment	Consecutive match length (bp)	D segment	BLAST score	Consecutive match length (bp)
Z80363	15	D3-22	9	D5-12	61	6
Z80364	27	D2-2	17	None		
Z80369	27	D6-19	10	None		
Z80372	29	D3-22	10	D6-13	24	4
Z80381	29	D3-16	10	D6-6	36	7
Z80389	51	D2-2	13	D6-25	63	7
Z80391	51	D2-2	13	None		
Z80414	23	D2-2	8	D4-17	35	6
Z80472	20	D1-26	9	D6-19	33	5
Z80478	24	D1-26	9	D2-15	29	6
Z80479	22	D1-7	11	D2-08	20	3
Z80484	27	D3-10	13	None		
Z80488	63	D2-8 and D3-9	25 and 24	None		
Z80493	33	D1-26	28	None		
Z80503	18	D3-10	9	D3-16	19	6

[a] One-hundred forty-four random sequences were selected from the 651 sequence database and analyzed for D segment alignment with JOINSOLVER and DNAPLOT.

$CDR3_H$ length, V_H-J_H distance, and D segment match tend to be longer in nonproductive rearrangements

Initially, the JOINSOLVER program was used to analyze the lengths of the $CDR3_H$, the V_H-J_H distance, and the length of the retained D segment in nonproductive and productive repertoires. As shown in Table III, the mean $CDR3_H$ length of the nonproductive rearrangements from all B cell subsets was 53.8 ± 1.9 bp, whereas the $CDR3_H$ of productive rearrangements was significantly shorter, a mean of 46.7 ± 0.5 bp, ($p < 0.01$). As with the $CDR3_H$ length, the nonproductive rearrangements have significantly ($p < 0.01$) longer V_H-J_H distances (36.2 ± 1.6 bp) than the productive ones (28.0 ± 0.4 bp). The same trend was observed when calculating the mean match length of the assigned D segments for all B cell groups (Table IV); the nonproductive repertoire had significantly ($p < 0.01$) longer consecutive matches (17.6 ± 0.7 bp) than the productive rearrangements (14.6 ± 0.2 bp).

D segment assignment relates to V_H-J_H distance

We were able to identify 71.4% of the D segments in the 77 nonproductive rearrangements and 64.4% of the D segments in the 574 productive rearrangements (Table IV). For the remaining rearrangements, no D segments were identified because the consecutive D match length was either too short (10.5% of the nonproductive; 16.9% of the productive) or had frequent point mutations (8.3% of the nonproductive; 13.6% of the productive). In addition to the 27 D segments located in the H chain locus on chromosome 14, there are 10 D segments located on chromosome 15 (29–32). When the rearrangements without a D segment match were analyzed, 3 were found to have a significant match with chromosome 15 D segments.

D segment usage

As shown in Fig. 2, the use of D segments is not random. In the nonproductive repertoire, 8 of 25 genes were used significantly more than expected from random chance. Moreover, a number of D segments were not detected in the nonproductive repertoire. Three of these (D1-14, D6-25, and D4-4) have mutations in the heptamer sequences that would be expected to limit recombination (21). In the productive repertoire, eight D segments were used more than expected from random chance. Notably, only two D segments were missing from the productive repertoire, presumably because they cannot undergo recombination effectively. Whether D4-4 is present in the productive repertoire cannot be determined because its sequence is identical to that of D4-11 (21). When the distribution of D segments in the nonproductive and productive repertoires was compared, evidence of both positive and negative selection was found. The use of two D segments was significantly greater in the productive repertoire, whereas the frequency of three D segments was significantly less, consistent with positive and negative selection of these gene segments, respectively.

Table III. *Mean $CDR3_H$ length and V_H-J_H distance for productive and nonproductive rearrangements*[a]

	Nonproductive		Productive	
	$CDR3_H$	V_H-J_H	$CDR3_H$	V_H-J_H
Peripheral B cells	57.5 ± 2.3[b]	39.5 ± 2.1[b]	46.2 ± 0.7	27.9 ± 0.5
Tonsilar plasma cells	45.8 ± 3.9	29.2 ± 3.0	47.2 ± 1.0	27.5 ± 0.8
IgG and IgM B cells	43.0 ± 3.8	26.4 ± 2.6	47.8 ± 1.0	29.0 ± 0.8
Mean	53.8 ± 1.9[b]	36.2 ± 1.6[b]	46.7 ± 0.5	28.0 ± 0.4

[a] Data shown represent mean base pairs ± SEM.
[b] Significant ($p < 0.05$) difference between nonproductive and productive rearrangements.

Table IV. *Frequency of rearrangements with unidentifiable or identifiable D segments and average D segment length*

	Nonproductive					Productive				
	D match		D length (bp)	No D match		D match		D length (bp)	No D match	
	n	(%)		*n*	(%)	*n*	(%)		*n*	(%)
Peripheral B cells	43	(78.2)	18.0 ± 0.8[a]	12	(21.8)[a]	217	(62.9)	14.5 ± 0.3[b]	128	(37.1)
Tonsilar plasma cells	6	(46.2)	17.3 ± 2.4	7	(53.8)[a]	70	(68.0)	14.8 ± 0.6	33	(32.0)
IgG and IgM B cells	6	(66.7)	15.0 ± 1.7	3	(33.3)	83	(65.9)	14.6 ± 0.5	43	(34.1)
Total	55	(71.4)	17.6 ± 0.7[a]	22	(28.6)	370	(64.4)	14.6 ± 0.2	204	(35.6)

[a] Significant ($p < 0.05$) difference between nonproductive and productive rearrangements.
[b] One rearrangement (Z80724) had a V_H segment with three mutations, a VJ length of 37, and a D region with 11 consecutive matches followed by a single mismatch and another 11 consecutive matches. Both regions of consecutive matches corresponded to IGHD3–22*01. Because the V_H segment had three mutations, the D segment could possibly have been 23 nucleotides long with a single mismatch in the middle. However, to conform with the rules established for JOINSOLVER, Z80724 was considered to have a D segment with 11 consecutive matches.

DIR and inverted D segment representation in the nonproductive and productive repertoire

DIR family members could be assigned in none of the nonproductive rearrangements and 1.1% of the productive rearrangements (Fig. 2). Notably, the frequency of usage of the DIR family members was significantly lower ($p < 0.05$) in both nonproductive and productive rearrangements than that expected from random chance. The use of inverted D segments is also absent in the nonproductive rearrangements (Fig. 2). However, inverted D segments were found significantly ($p < 0.01$) more often in the productive rearrangements (3.8%), suggesting they were positively selected.

Fusion of multiple D segments is negatively selected

Fig. 3 depicts the rearrangements with multiple D segments. This was an infrequent occurrence, but all the nonproductive rearrangements ($n = 3$) with multiple D segments were organized as $V_HD^{5'}D^{3'}J_H$, which reflects the normal sequence of recombination events. Of the productive rearrangements ($n = 5$) with putative multiple D segments, three were organized as $V_HD^{5'}D^{3'}J_H$. One of these (Z80631) had two identified D segments, which were organized $V_HD^{3'}D^{5'}J_H$. The second had three identified D segments of which the middle one originated from upstream of the 5′ D segment but also 5′ of the downstream D segment. The appearance of rearrangements with multiple D segment fusions was significantly less in the productive repertoire (1% of total rearrangements) compared with that in the nonproductive repertoire (3.9% of the total nonproductive rearrangements) ($p < 0.01$), suggesting that these rearrangements were negatively selected.

Comparison of the distribution of D segments using different levels of analytic stringency

Different levels of stringency have been applied in the identification of D segments (8, 21, 25, 26). More strict criteria may increase specificity, but at the expense of sensitivity, whereas less strict criteria may do the reverse. We were interested in determining whether the criteria used in the current analysis inappropriately included or excluded D segments. To assess this, the distribution of D segment alignments that were found with different levels of stringency (0.05 vs 0.01) was analyzed in the smaller set of 144 random sequences mentioned above. As shown in Fig. 4, only a few ($n = 8$, 5.6%) identified D segments were lost when the stringency was increased from 0.05 to 0.01. Importantly, the overall distribution of identified D segments was very similar using either cut-off. Notably, using either criterion, D7-27, the shortest germline D segment, was absent from the rearrangements assessed in this subset. These data suggest that a stringency of 0.05 is sufficient to balance the needs for sensitivity and specificity in this biologic analysis.

The infrequent appearance of the shortest D gene (D7-27) suggested the possibility that using a consecutive match approach may have biased against identifying short germline D segments. To address this possibility, alignments that failed to meet the 0.05 threshold for identification were examined. Thirty-eight D segment alignments with consecutive match lengths of only 5 to 9 nucleotides fell into this category and were considered random alignments. They represented 21 different germline D genes and every D family and varied in length from 11 to 37 nucleotides (D7-27

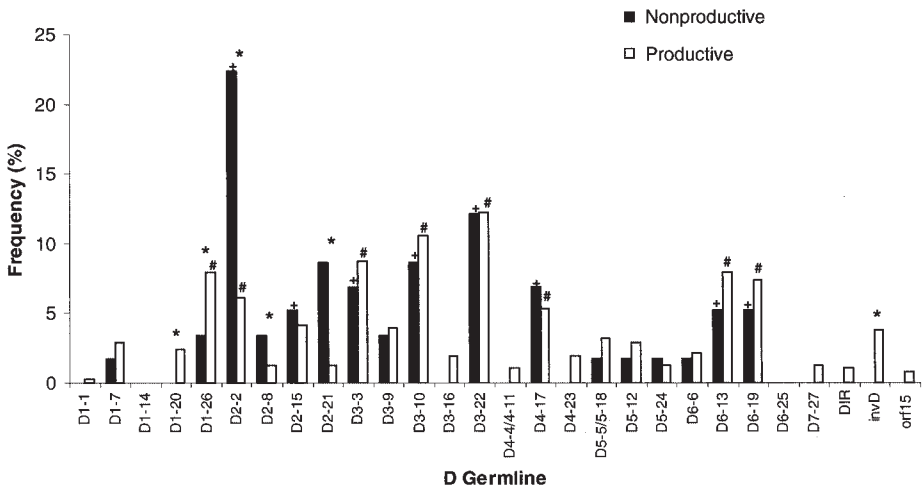

FIGURE 2. Frequency of identifiable D segments in nonproductive (■) and productive (□) rearrangements. The denominators are the total number of rearrangements with identifiable D segments in the nonproductive and productive repertoires, respectively. The # and + symbols indicate a significantly ($p < 0.05$) higher frequency when comparing productive or nonproductive rearrangements, respectively, to the expected random frequency. ∗, Significantly ($p < 0.05$) different frequencies when comparing nonproductive and productive repertoires.

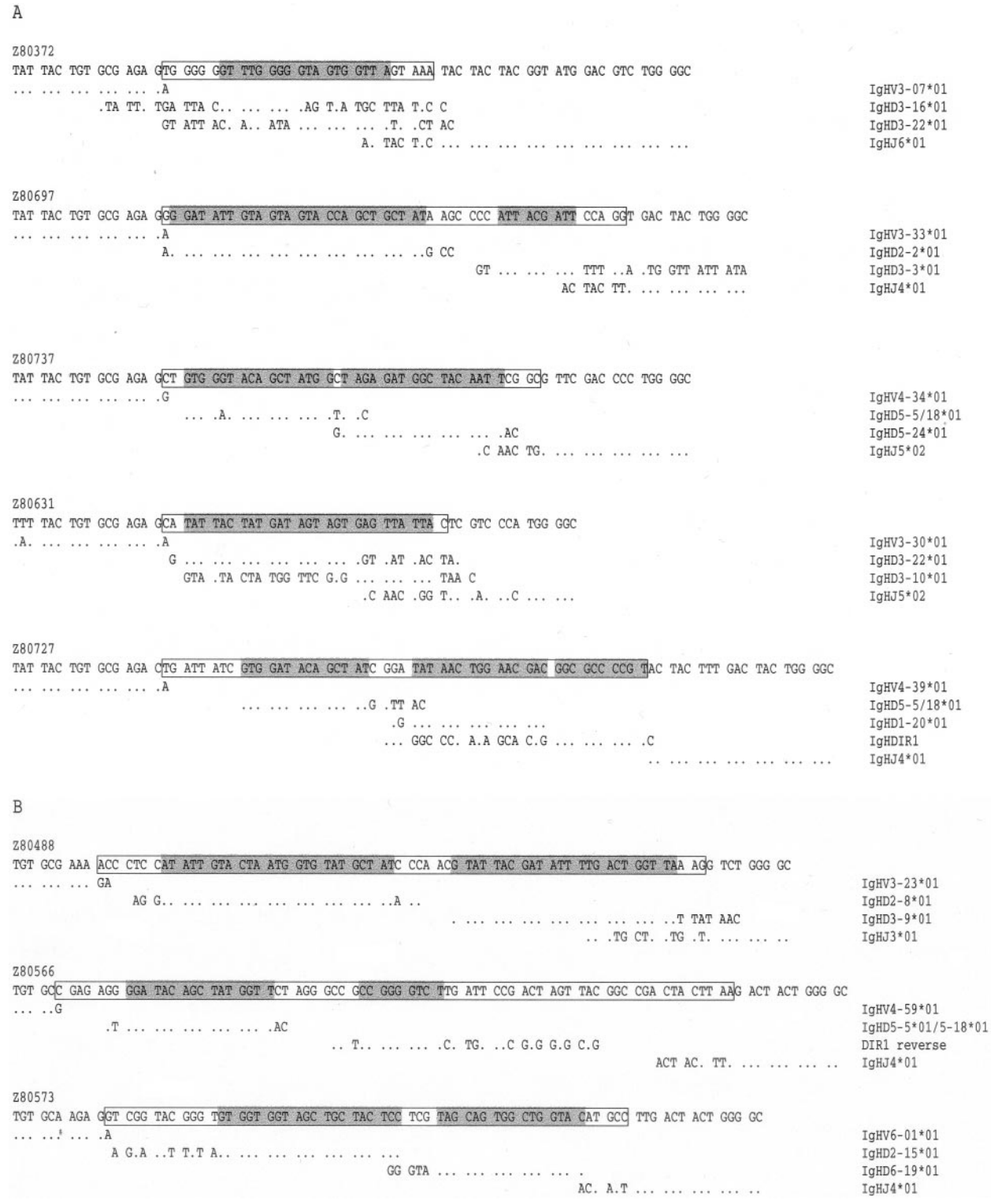

FIGURE 3. Productive (*A*) and nonproductive (*B*) rearrangements with multiple D segment fusions. The V_H-J_H distance is enclosed by a box. Gray shading within the box indicates the V_H-J_H nucleotides that match germline D segments.

and D3-16, respectively). Although there may be some bias against identification of the shortest germline genes and in favor of the longest germline genes, the failure to identify D7-27 as frequently as D segments with longer germline sequences likely reflects the decreased frequency of this segment in the adult peripheral blood repertoire, as has been suggested (33).

Distribution of D segment RFs

D segment RFs have been categorized as those containing stop codons (largely RF1), those tending to encode hydrophilic amino acids (largely RF2), and those tending to encode hydrophobic amino acids (largely RF3). Indeed, 11 of the 25 D segments have a stop codon in RF1, whereas 7 of 25 have stop codons in RF2 and

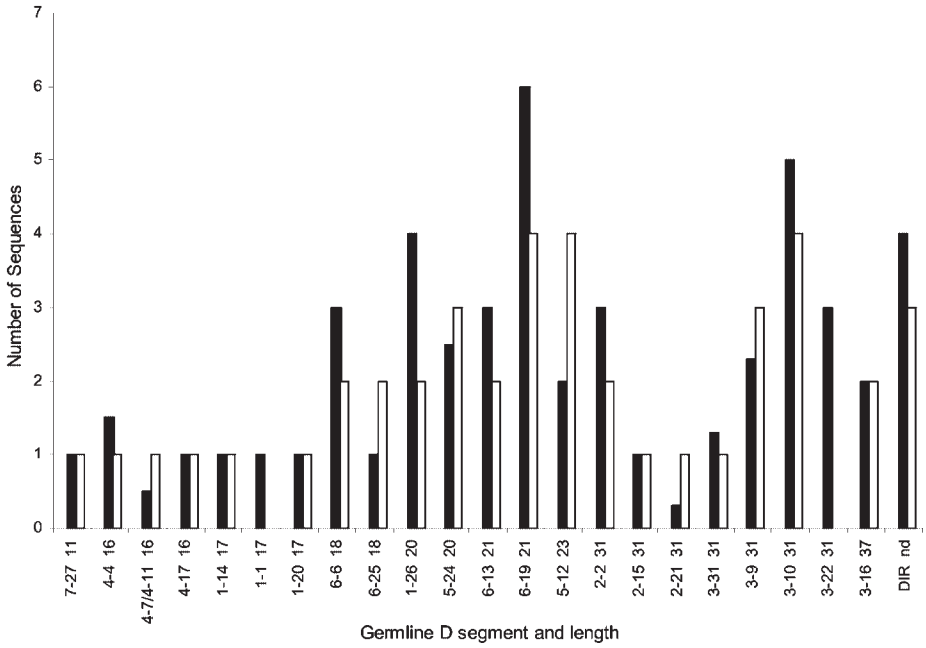

FIGURE 4. The distribution of D segments using 0.05 and 0.01 significance levels. One-hundred forty-four random sequences were analyzed by JOINSOLVER for D segment alignments. ■, Alignments excluded by using a 0.05 significance level. □, Alignments excluded using a 0.01 significance level. The D segment gene is followed by the number of nucleotides in the D germline sequence.

only 5 of 25 have stop codons in RF3. In the nonproductive repertoire, each of the D segment RFs was used at comparable frequencies (Table V). Notably, with few exceptions, D segment RFs with stop codons were used at the frequency expected in the nonproductive repertoire (Tables V and VI). In general, the use of RFs with stop codons was excluded from the productive repertoire except when the stop codon could be removed by exonucleolytic processing. When the use of RFs in the productive repertoire was analyzed, evidence for positive selection of a number of specific D gene segment RFs was noted. Thus, overall, RF2 was positively selected, whereas RF3 was not (Table VI). Specifically, RF2 was preferred in the productive repertoire by rearrangements using D2-2, D2-8, D2-15, D3-10, D3-16, and D3-22. In contrast, RF3 was preferred by rearrangements using D1-20, D1-26, D5-12 and D5-24, although the entire RF was not positively selected. Notably, no selection of rearrangements using RF1 was detected even though some D segments (D6-6, D6-13, D6-19) encoded hydrophilic amino acids in this RF. Despite this, analysis of the preferred RFs indicated that they were frequently more enriched in hydrophilic amino acids or glycine (D2-2, D2-15, D3-10, and D3-22), although this was not a uniform finding as noted above. Notably, RF1 was not positively selected even when the rearrangements lacking stop codons or enriched for hydrophilic amino acids were analyzed separately.

Amino acid analysis of the CDR3$_H$

We next analyzed the amino acid composition of the CDR3$_H$ more completely to determine whether it differs from that expected from random chance. When the amino acid distribution in the D segments in the nonproductive rearrangements was analyzed, only Y, W, and V, were used significantly ($p < 0.05$) more than random, and K, H, E, and R were used significantly ($p < 0.05$) less than random (Fig. 5). When comparing the amino acid composition of the D segments in the productive rearrangements with the random sequences, the only ARs used significantly ($p < 0.05$) more than expected from random chance were Y, W, G, and S, whereas P, L, K, Q, H, E, and R were used significantly ($p < 0.05$) less than random. Notably, S was the only residue that was both positively selected and used more than expected from random chance. Within the D segment, N was positively selected but not used more often than expected from random chance. All other residues that were used more frequently than expected from random chance were not positively selected. In general, the amino acid composition of the nonproductive rearrangements was closer to random than the productive ones (χ^2 of 157 vs 885, respectively) (Fig. 5).

When the amino acid composition of the entire CDR3$_H$ in nonproductive rearrangements was analyzed in a similar manner, only the ARs T, Y, W, and G were used significantly ($p < 0.05$) more than expected from random chance, whereas K, Q, and E were used significantly ($p < 0.05$) less than expected from random chance. In the productive rearrangements, only F, Y, W, G and D, were used significantly ($p < 0.05$) more and I, L, P, C, A, K, Q, E and R were used significantly ($p < 0.05$) less than expected from random chance. Among the ARs that were used significantly more than expected from random chance, W and D were negatively selected whereas F, Y, and G were positively selected. Notably, the nonproductive rearrangements had an amino acid composition that was more similar to random amino acid sequences than the productive rearrangements (χ^2 of 185 vs 1354, respectively, Fig. 5).

Table V. *Total distribution of D segments by RF usage*

D Segment Reading Frame	Rearrangements	
	Nonproductive	Productive
	n (% of total)	
1	22 (37.9)	85 (22.5)
2	20 (34.5)	177 (46.8)[a]
3	16 (27.6)	116 (30.7)

[a] Significant ($p < 0.05$) difference between the use of RF2 in comparison to RF1 and RF3 in the productive repertoire.

Preferential rearrangement of 5′ D segments with 3′ J$_H$ segments

The analysis of nonproductive rearrangements is a valuable mechanism to study the molecular events before Ag selection (34). Thus, we analyzed the combinatorial preferences of the nonproductive rearrangements to determine whether there was a bias for particular V$_H$DJ$_H$-rearrangements. From this analysis, there was

Table VI. *D segment reading frame use*[a]

D Gene	RF1	Frequency of RF1 Usage NP	Frequency of RF1 Usage P	RF2	Frequency of RF2 Usage NP	Frequency of RF2 Usage P	RF3	Frequency of RF3 Usage NP	Frequency of RF3 Usage P
1-1	GTTGT	0	0	VQLER	0	0	YNWND	0	1 (100)[b,c]
1-7	GITGT	0	4.5 (53)[b,c]	V*LEL	1 (100)	0	YNWNY	0	4 (47)[c]
1-14	GITGT	0	0	V*PEP	0	0	YNRNH	0	0
1-20	GITGT	0	2.5 (26)	V*LER	0	0	YNWND	0	7 (74)[b,c]
1-26	GIVGAT	2 (100)[b,c]	10 (32)	V*WELL	0	3 (10)	YSGSYY	0	18 (58)[b,c]
2-2	RIL**YQLLC	3 (23)[c]	0	GYCSSTSCYA	5 (8)	17 (68)[b]	DIVVVPAAM	5 (38)	8 (32)
2-8	RILY*WCMLY	1 (50)	0	GYCTNGVCYT	0	4 (80)[b,c]	DIVLMVYAI	1 (50)	1 (20)
2-15	RIL*WW*LLL	1 (33)[c]	0	GYCSGGSCYS	1 (33)	13 (81)[b,c]	DIVVVVAAT	1 (33)	3 (19)
2-21	SILWW*LLF	2 (40)[c]	0	AYCGGDCYS	1 (20)	1 (25)[b,c]	HIVVVIAI	2 (40)	3 (75)[b]
3-3	VLRFLEWLLY	1 (33)	10 (29)	YYDFWSGYYT	2 (66)	14 (41)	ITIFGVVII	1 (33)	10 (29)
3-9	VLRYFDWLL*	0	5 (33)	YYDILTGYYN	0	10 (66)[b,c]	ITIF*LVII	2 (100)	0
3-10	VLLWFGELL*	2 (40)	10 (24)	YYYGSGSYYN	1 (20)	21 (51)[b,c]	ITMVRGVII	2 (40)	10 (24)
3-16	VL*LRLGELCLY	0	1 (13)	YYDYVWGSYAYT	0	7 (88)[b,c]	IMITFGGVMLI	0	0
3-22	VLL***WLLL	4 (50)	0	YYYDSSGYYY	3 (38)	40 (87)[b,c]	ITMIVVVIT	1 (13)	6 (13)
4-4/4-11	*LQ*L	0	0	DYSNY	0	2 (50)[c]	TTVT	0	2 (50)[b,c]
4-17	*LR*L	2 (50)	0	DYGDY	2 (50)[b]	11 (52)[b]	TTVT	0	10 (48)[c]
4-23	*LRW*L	0	0	DYGGNS	0	3 (60)[b,c]	TTVVT	0	2 (40)[c]
5-5/5-18	VDTAMV	0	6 (50)	WIQLWL	0	1 (9)	GYSYGY	1 (100)	5 (42)
5-12	VDIVATI	1 (50)	3 (27)	WI*WLRL	0	1 (9)	GYSGYDY	1 (50)	7 (64)[b]
5-24	VEMATI	1 (100)	3 (50)	*RWLQL	0	1 (17)	RDGYNY	0	2 (33)[b,c]
6-6	EYSSSS	1 (100)	4 (44)	SIAAR	0	5 (56)[b]	V*QLV	0	0
6-13	GYSSSWY	0	13 (42)	GIAAAG	2 (100)[b]	14 (45)	V*QQLV	0	4 (13)
6-19	GYSSGWY	1 (33)	15 (47)	GIAVAG	2 (66)[b]	12 (38)	V*QWLV	0	5 (16)
6-25	GYSSGY	0	0	GIAAA	0	0	V*QRL	0	0
7-27	LTG	0	2 (40)	*LG	0	0	NWG	0	3 (60)[b]

[a] Columns 2, 5, and 8 show the amino acid translation of each D segment for each RF followed by the number of times a nonproductive or productive rearrangement using a specific D segment uses the indicated RF. The number in parentheses indicates the percentages with which rearrangements using the specific D segment use the given reading frame. The asterisk (*) represents stop codons; hydrophilic amino acid residues are italicized; hydrophobic amino acid residues are black. Amino acids are defined as hydrophilic or hydrophobic as described (38).

[b] Significantly ($p < 0.05$) higher frequency when compared to the expected random frequency.

[c] Significantly ($p < 0.05$) higher frequency between nonproductive and productive.

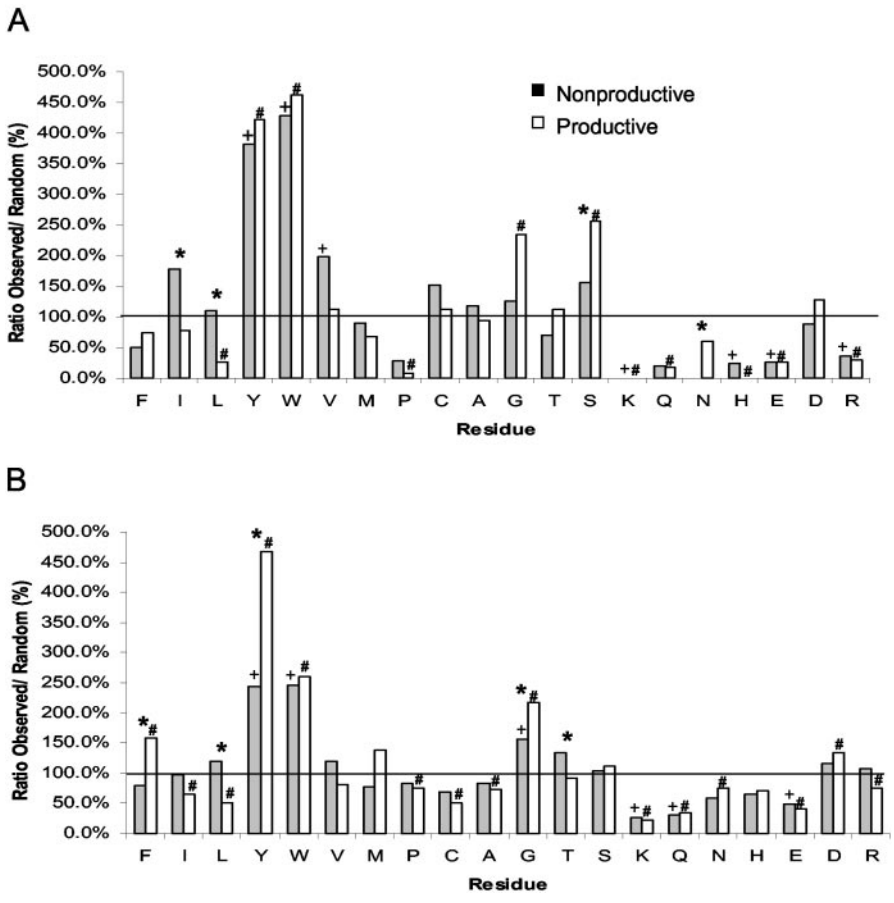

FIGURE 5. Distribution of the ARs in the D segment (*A*) and $CDR3_H$ (*B*) for both nonproductive (▨) and productive (□) rearrangements compared with randomly sampled human sequences. ARs are organized from left to right according to their hydropathy values (44). +, A significant ($p < 0.05$) difference between nonproductive rearrangements and random human sequences; #, a significant ($p < 0.05$) difference between productive rearrangements and random human sequences; ∗, a significant ($p < 0.05$) difference between nonproductive and productive rearrangements.

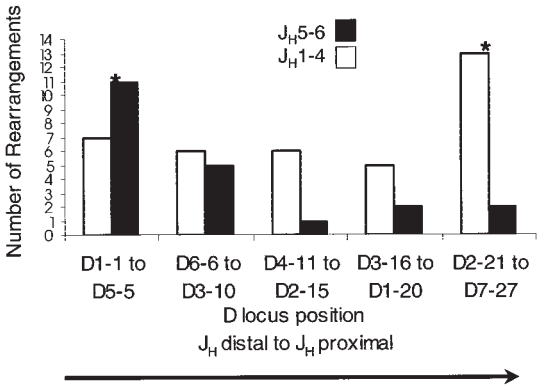

FIGURE 6. Frequency of use of D segments by nonproductive rearrangements using D proximal and D distal J_H genes. D-proximal J_H genes (J_H1, 2, 3, and 4, □); D-distal J_H genes (J_H5 and 6, ■). The D genes are divided into five groups according to their position in the locus (D1-1 to D5-5 are the most J_H distal and D2-21 to D7-27 are the most J_H proximal). *, Significant ($p < 0.01$) bias in the pairing of D and J_H segments. Only the 59 rearrangements for which a D segment could be assigned were analyzed.

no bias for a particular V_HD pairing, because 5′ or 3′ D segments were indiscriminately paired with 5′ or 3′ V_H segments (data not shown). Conversely, a significant bias ($p < 0.01$, χ^2 test) could be observed in the pairing of D and J_H segments, with 5′ (J_H distal) D segments coupled preferentially to 3′ J_H segments (Fig. 6). This bias was not found in productive rearrangements (data not shown).

TdT activity and microhomology

The number of N nucleotides inserted between the V_H and the D coding ends was similar to that inserted between the D and J_H in both productive and nonproductive rearrangements (Table VII). Nevertheless, there were significantly more N nucleotide additions in the V_HD and DJ_H junctions in the nonproductive than in the productive rearrangements. Although the presence of N nucleotides was the most common situation, there were some sequences that lacked N nucleotide additions at either the V_HD (nonproductive 3.6%, productive 3.0%) or DJ_H junction (nonproductive 7.2%, productive 7.3%), even though there were no significant differences between nonproductive and productive rearrangements. Rearrangements lacked TdT activity significantly more often in the DJ_H junction than in the V_HD junction ($p < 0.05$).

As shown in Table VII, the presence of microhomology on both the V_HD and DJ_H junction is more frequent in the productive than in the nonproductive repertoire. Moreover, the DJ_H junction had a significantly ($p < 0.05$) higher percentage of sequences with microhomology than the V_HD junction.

Exonuclease activity

The V_H coding end had significantly ($p < 0.01$) less exonucleolytic excision when compared with the D and J_H coding ends, both in the nonproductive and productive repertoires (Table VIII). D segment excision was similar in the nonproductive or productive repertoires, with more excision at the 3′ end. The J_H coding end was excised to a significantly ($p < 0.01$) greater degree in both nonproductive and productive rearrangements than the V_H and D5′ coding ends. P nucleotides were significantly ($p < 0.01$) more abundant in the V_HD junctions than in the DJ_H junction in both nonproductive and productive rearrangements (Table IX). Palindromic (Pr) nucleotides in processed coding ends that could have developed from an overhanging hairpin intermediate structure (35) were not found at a greater frequency than expected from random chance (data not shown). As a result of the various modifications, the mean lengths of the V_HD and DJ_H junctions were 10.2 ± 1.0 bp and 9.2 ± 0.9 bp in nonproductive rearrangements and 7.7 ± 0.3 bp and 7.1 ± 0.3 bp in productive rearrangements, respectively.

Discussion

We have developed a novel software algorithm, JOINSOLVER, to analyze the human $CDR3_H$. Within the $CDR3_H$, the definition of the D segment has been particularly problematic because of its short size and extensive terminal processing. Many attempts have been made to define the minimum length needed for D segment assignment (8, 21, 24–26, 36), yet there is still no consensus definition. Thus, we used novel methods to assign D segments. The first involved the use of a consecutive matching approach rather than the more standard alignment scoring system. The consecutive matching approach permitted the secure assignment of more D segments than the alignment scoring method. The second used methods to limit the search for identity to the V_H-J_H region only. Finally, a Monte Carlo simulation was used to determine the consecutive match necessary to assign a D segment. We opted to distinguish an actual D segment match from random sequence identity using a 95% probability. This level of confidence seems more appropriate for biological systems, because it balances the sensitivity and specificity of the D segment assignments. A previous alignment scoring approach used a 99% probability for D segment assignment (20), which increases the specificity, but omits a large number of apparently real D matches. Using the consecutive matching approach along with a 95% probability strategy, we were able to identify D

Table VII. *TdT activity and frequency of microhomology in the V_H-D or D-J_H junctions of nonproductive and productive rearrangements*

	Nonproductive						Productive					
	N addition (bp)		No N addition (% of total)		Microhomology (% of total)		N addition (bp)		No N addition (% of total)		Microhomology (% of total)	
	V_H-D	D-J_H	V_H-D	D-J_H	V_H-D	D-J_H	V_H-D	D-J_H	V_H-D	D-J_H	V_H-D	D-J_H
Peripheral B cells	11.0 ± 1.1[a]	9.5 ± 1.1	4.6[a]	9.2[b]	2.3[a]	4.6[b]	7.8 ± 0.3[c]	7.3 ± 0.4	1.8	8.8[b]	0.5	2.8[b]
Tonsilar plasma cells	7.0 ± 2.9	8.5 ± 1.2	0[a]	0[a]	0[a]	0[a]	6.7 ± 0.5	6.6 ± 0.6	4.3	5.7	2.8	2.8
IgG and IgM cells	6.3 ± 1.4	8.5 ± 2.0	0[a]	0[a]	0[a]	0[a]	8.0 ± 0.7	7.0 ± 0.6	4.8	4.8	3.6	4.8
Mean	10.2 ± 1.0[a]	9.2 ± 0.9[a]	3.6	7.2	1.8	3.6[b]	7.7 ± 0.3	7.1 ± 0.3	3.0	7.3[b]	1.6	3.2[b]

[a] Significant ($p < 0.05$) differences between nonproductive and productive rearrangements.
[b] Significant ($p < 0.05$) differences between the V_HD and DJ_H junctions.
[c] One rearrangement (Z80724) had a D region containing 11 5′ consecutive matches followed by one mismatch and 11 3′ consecutive matches. Both consecutive matching segments were identified as IGHD3–22*01. The 3′ 11 consecutive match was assigned as the D segment. The 5′ 11 consecutive match was considered as a match only for the purpose of identifying the V_H-D junction and analyzing junctional diversity in this region.

Table VIII. *Exonuclease activity in the V_HD or DJ_H junctions*

	Excision Site	Peripheral B Cells	Tonsilar Plasma Cells	IgG and IgM Cells	Mean
Nonproductive	V_H3'	2.2 ± 0.4	2.2 ± 0.8	1.1 ± 0.6	2.1 ± 0.3[a,b]
	D5′	4.6 ± 0.6	4.8 ± 2.5	3.2 ± 1.1	4.5 ± 0.5[a,b]
	D3′	4.7 ± 0.7	3.5 ± 1.7	8.7 ± 3.1	5.0 ± 0.6[a]
	J_H5'	6.5 ± 0.7	7.5 ± 1.2	7.6 ± 1.6	6.8 ± 0.6[a,b]
Productive	V_H3'	1.9 ± 0.1	1.5 ± 0.2	1.6 ± 0.2	1.8 ± 0.1[a,b]
	D5′	4.5 ± 0.3[c]	4.9 ± 0.4	5.3 ± 0.5	4.7 ± 0.2[a,b]
	D3′	5.5 ± 0.3	5.2 ± 0.5	5.3 ± 0.5	5.4 ± 0.2[a]
	J_H5'	6.0 ± 0.3	5.9 ± 0.4	5.6 ± 0.4	5.9 ± 0.2[a,b]

[a] Significant ($p < 0.05$) differences between the excision in the V_H coding end and the other coding ends.
[b] Significant ($p < 0.05$) differences between the excision in the D5′ coding end and the other coding ends.
[c] One rearrangement (Z80724) had a D region containing 11 5′ consecutive matches followed by one mismatch and 11 3′ consecutive matches. Both consecutive matching segments were identified as IGHD3–22*01. For analysis, the 3′ 11 consecutive match was assigned as the D segment and the 5′ 11 consecutive match was used only to identify the V_H-D junction.

segments in >68% of the analyzed rearrangements. By contrast, the alignment scoring system and higher stringency used in the previous analysis (20) resulted in only 50.5% of rearrangements having a D segment assignment. Importantly, we could detect no consistent bias in D segment assignment when the current database was analyzed with the higher stringency used in the previous approach. Moreover, there was general similarity between the D segments assigned in the current analysis and those reported previously. Thus, of the eight D segments that were found at a greater frequency in the productive repertoire than expected in the current analysis (D1-26, D2-2, D3-3, D3-10, D3-22, D4-17, D6-13, and D6-19), six were previously identified as the most frequently used D segments. Similarly, of the six most frequently used D segments in the previous analysis (D2-2, D3-3, D3-10, D3-22, D6-13, and D6-19), all were overrepresented in the current analysis of the productive repertoire. These results indicate that the current approach identifies more D segments than previous methods, but does not bias the analysis inappropriately.

The use of D segments was not random. Analysis of the nonproductive repertoire provided information concerning biased use of D segments during V_HDJ_H recombination. Eight D segments (D2-2, D2-15, D3-3, D3-10, D3-22, D4-17, and D6-19) were significantly overrepresented in the nonproductive repertoire, suggesting that they were preferentially used during V_HDJ_H recombination. The reasons for the preferential usage in the nonproductive repertoire are not clear as these segments are both long and short and scattered throughout the locus. Moreover, it is unlikely that the recombination signal sequences (RSS) play a major role as, for example D2-2 (overrepresented) and D2-8 (used at the expected frequency) have identical RSS and are the same length (20), but are used at markedly different frequencies.

Analysis of the productive repertoire indicated that a number of D segments were also overrepresented. Some of these, such as D3-3, D3-10, D3-22, D4-17, D6-13, and D6-19, were not positively selected but rather appeared frequently because of biased use during recombination with no subsequent evidence of negative selection. In contrast, D1-26 was overrepresented in the productive repertoire as a result of positive selection. Another D segment (D1-20) manifested evidence of positive selection, even though it did not appear more frequently than expected in the productive repertoire. Finally, a number of D segments (D2-2, D2-8, D2-21) were clearly negatively selected. The final distribution of D segments in the productive repertoire, therefore, results from biases introduced during V_HDJ_H recombination and subsequently from positive and negative selection. The basis of these molecular and selective events is currently unresolved, but do not appear to relate solely to D segment length, RF bias, or position in the locus.

Germline D segments vary in length from 11 nucleotides (D7-27) to 37 nucleotides (D3-16). Because the length of the $CDR3_H$ appears to be restricted in the productive repertoire (46.7 ± 0.5 bp) and regulated by selection, the use of longer D segments may be limited, unless these are exposed to extensive exonuclease cleavage during recombination. Indeed, the finding that the length of the D segment after exonuclease cleavage is only 14.6 ± 0.2 nucleotides in the productive repertoire suggests that the length of the germline D segment plays little role in biasing the repertoire. In this regard, the longest D segment (D3-16) was positively selected along with a number of shorter segments, whereas some, but not all, long segments were negatively selected. Thus, it appears that the original length of the germline D segment does not play a crucial role in the selection of particular V_HDJ_H rearrangements, which is likely related to subsequent exonucleolytic activity that reduces the size of the D segment.

It is notable that the apparent use of DIR family members and/or inverted D segments was identified in this set of rearrangements. Although both events were absent in the nonproductive repertoire implying that they were rarely used in rearrangements, they were more frequent in the productive repertoire suggesting that their use could contribute to diversity. There is controversy concerning the use of DIR segments and inverted D segments, with some studies reporting their use and others not (20, 23, 37). Notably, however,

Table IX. *Number of P nucleotides (bp)*

	Junction	Peripheral B Cells	Tonsilar Plasma Cells	IgG and IgM Cells	Mean
Nonproductive	V_HD	0.6 ± 0.1[a]	0.7 ± 0.3[a]	0.3 ± 0.3[a]	0.6 ± 0.1[a]
	DJ_H	0.2 ± 0.1	0	0.2 ± 0.2	0.1 ± 0.1
Productive	V_HD	0.4 ± 0.1[a]	0.5 ± 0.1[a]	0.7 ± 0.1[a]	0.5 ± 0.4[a]
	DJ_H	0.2 ± 0.04	0.2 ± 0.1	0.2 ± 0.1	0.2 ± 0.03

[a] Significant difference between the number of P nucleotides in V_HD vs DJ_H joins of productive and nonproductive rearrangements ($p < 0.01$).

even in the stringent analysis of Corbett et al. (20) a low frequency (0.5–1% of rearrangements) used these elements. The bulk of the data support the conclusion that DIR family members and inverted D segments are used rarely in human V_HDJ_H rearrangements as could be expected from the molecular constraints imposed on their use in recombination.

The presence of multiple D segments in a single rearrangement, i.e., V_HDDJ_H, has also been a matter of controversy. The presence of such DD recombination violates the "12/23 rule", because it would disregard the strict sequential recognition by the recombination-activating gene (RAG) proteins of a 23-bp spacer associated RSS following a 12-bp spacer associated RSS (38). Whereas some studies provide evidence for the existence of DD fusions both in human (8, 16, 18, 24, 39–41) and mice (9, 12, 13), other studies conclude that such multiple D recombinations do not occur or are infrequent events (20, 25, 42). A second Monte Carlo simulation was preformed to assess the statistical probability of the existence of such multiple D segment recombinations. This analysis strongly implied that multiple D segment recombinations can occur in the human V_HDJ_H repertoire, but their frequency does not appear to be as high as suggested by some previous reports (8, 10) nor as unlikely as concluded by others (20). Of the eight sequences shown in Fig. 3, four (Z80737, Z80727, Z80488, and Z80573) unequivocally contain two D segments even using the stringent criterion of Corbett (20). Moreover, in three other rearrangements (Z80372, ZZ80697, and Z80631), the likelihood that the second D match occurred by random chance ranged between 0.9 and 2.6%. Therefore, six of the eight sequences with the putative D-D fusions are likely to be authentic (two nonproductive, four productive). These results indicate that rearrangements using two D segments are uncommon but real. Because DD fusions appear more frequently in the nonproductive compared with the productive repertoire, it is likely that such fusions producing longer $CDR3_H$ are negatively selected, possibly because they may distort the Ag binding site or encode autoantibodies (9, 13). As a result, the use of multiple D segments is unlikely to play a major role in contributing to diversity in the human V_H repertoire. It is notable that the frequent use of D-D fusions in the mouse has also been questioned (21). Moreover, in human B cells identified with an Ab to V-pre-B, the increased use of productive V_HDJ_H rearrangements with D-D fusions that meet the current criteria (8/136 = 5.9%) indicates that this event is uncommon, even in this population (11).

D segment RFs are determined by the combined effect of exonuclease and TdT that remove or add nucleotides at the $V_H \rightarrow DJ_H$ junction (43). Changes in RF impact the amino acid sequence which changes the hydropathic character of the $CDR3_H$. Evidence for the preferential use of D segment RFs that encode hydrophilic amino acids has previously been presented (21). In the current study, we found that each of the RFs was used comparably in the nonproductive repertoire, implying that there was no combinatorial bias in their usage. However, the distribution of RFs in the productive repertoire was clearly not random, with RF2 overrepresented, RF1 underrepresented, and RF3 appearing at the expected frequency. The underrepresentation of RF1 relates to the more frequent presence of stop codons that preclude the appearance of these RFs in the productive repertoire unless the stop codon is removed by exonuclease activity. RF2 appears to be overrepresented in the productive repertoire because of the frequent presence of hydrophilic amino acids that are positively selected, such as can be found in D2-2, D2-8, D2-15, D3-9, D3-10, D3-16, and D3-22. Using either the method of Black and Mould (44) in which T, S, K, Q, N, H, E, D, and R or the Kyte and Doolittle (45) analysis in which T, S, W, Y, K, Q, N, H, E, D, and R are viewed as hydrophilic, all of these D segments encode amino acids in RF2 that are more hydrophilic than those encoded by the other RFs without stop codons. The data are consistent with the conclusion that there is positive selection of D segment RFs that encode hydrophilic amino acids. It is notable that there was no evidence of positive selection of some D segment RFs (RF1 in D6-6, D6-13, D6-19) encoding hydrophilic amino acids. The explanation of this is not clear, but in two of them (D6-13 and D6-19) the Kyte and Doolittle (45) and Hopp/Woods (46) analyses identified different RFs as the most hydrophilic. Alternatively, the sparcity of glycine residues that may contribute to flexibility of the $CDR3_H$ (47) may limit the ability of these D segment RFs from being positively selected. When the RFs are grouped into those actually with stop codons, those actually encoding hydrophilic amino acids and those actually encoding hydrophobic amino acids according to the Kyte and Doolittle algorithm (45), no differences were noted in their usage by nonproductive rearrangements. However, marked enrichments in productive rearrangements using the hydrophilic RF (60.3%) compared with the hydrophobic RF (32.9%) and the RF with stop codons (6.8%) was noted. Again this result is consistent with the conclusion that there is positive selection of D segment RFs encoding hydrophilic amino acids.

Because of the evidence of positive selection of D segment RFs encoding hydrophilic amino acids, the overall hydropathicity of the $CDR3_H$ was analyzed. This indicated that of all the hydrophilic amino acids encoded by the D segment, only S and N were positively selected. The D segment contributed an overabundance of hydrophobic amino acids to the $CDR3_H$ as detected in the nonproductive repertoire, including W, some of which, such as I and L, were negatively selected. In addition, the D segment contributed an increased number of Y residues to the nonproductive repertoire, whose presence was not subsequently altered by selection. Therefore, despite the positive selection of D segment RFs encoding hydrophilic amino acids, the only hydrophilic amino acid encoded by this region that was both overrepresented and positively selected was S. Notably, this contribution was counteracted by the amino acids encoded by other portions of the $CDR3_H$ (V_H, J_H, and junctional diversity).

As previously reported (20, 42), the mean degree of TdT activity on the V_HD and DJ_H junctions is similar in both productive and nonproductive repertoires. It is notable that there is a small, but significant, difference in the percentage of DJ_H junctions with no N additions compared with the percentage of V_HD junctions with no N additions. The difference may be related to developmentally regulated levels of TdT expression, as revealed by less frequent junctional TdT activity in fetal and neonatal repertoires compared with the adult repertoire (25, 48–50). Thus, the increased number of N nucleotides in the V_HD junction, which is formed after the DJ_H junction in rearranging B cells, could be related to a higher level of TdT activity. The exonucleolytic activity was greater on the DJ_H junction than on the V_HD join. This was especially notable when the degree of exonuclease processing of the V_H and J_H segments were compared and could relate to the primary sequence of these regions. The 5′ coding ends of the J_H segments are slightly more AT-rich, potentially making them preferential substrates for exonucleolytic processing (35, 51). The differences in the processing of V_H and J_H segments was reflected by the appearance of P nucleotides that were more frequent in the V_HD junction than in the DJ_H joint. Finally, it has been suggested that the presence of Pr nucleotides might reflect a second round of RAG-mediated cleavage and "hairpining" (35). It is notable in the current study that the frequency of Pr nucleotides (35, 52) in any of the coding ends was not significantly different than the likelihood that specific consecutive nucleotides would be found by random chance. Because of this finding, we recalculated the frequency of Pr nucleotides in the

original report (35) and found that the frequency of Pr overhangs was below the frequency of consecutive nucleotides occurring by random chance, even for the longer insertions. Therefore, it is unlikely that Pr nucleotides play a role in the generation of junctional diversity. Moreover, it is unlikely that "rehairpining" of the coding ends occurs during V_HDJ_H rearrangement.

The formation of microhomology may influence the development of the $CDR3_H$ by constraining nucleolytic processing or preventing access of TdT to the coding ends (35, 53, 54). In the present analysis, a low frequency of microhomologies was observed. Notably, the frequency of microhomologies seems to be less than that found in fetal (49) or neonatal (25) arrangements.

Analysis of the nonproductive rearrangements provides insight into the molecular mechanisms occurring before selection. By analyzing the nonproductive repertoire for V_H, D, and J_H segments, we were able to detect biases in the association of these genetic elements without the superimposed influence of selection. A significant bias was noted in the tendency for 5′ D segments to rearrange with 3′ J_H segments without relation to the position of the V_H gene. The most likely explanation for this finding is that there are multiple DJ_H arrangements before $V_H \rightarrow DJ_H$ rearrangement occurs and the rearrangement process ceases. The net result would be the tendency for 5′ D segments to be found preferentially rearranged to 3′ J_H segments because the initial rearrangements would be deleted as the progressive rearrangement process proceeds. The finding that the distribution of V_H genes is random suggests that this process occurs before $V_H \rightarrow DJ_H$ rearrangement occurs. A similar process has been suggested to occur in the mouse (9, 55). Whether this process serves to increase diversity or rather merely reflects the persistent expression of RAG proteins and availability of the H chain locus during B cell development remains to be determined. The finding that the bias is lost in the productive repertoire presumably as a result of selection is more consistent with the latter interpretation.

The development of JOINSOLVER has permitted a detailed analysis of the human adult $CDR3_H$ and has facilitated the development of new insights into the molecular and selective mechanisms that underlie the generation of this Ag binding region of human Ig molecules. Importantly, this approach and the database generated should be of great value in determining abnormalities in individuals with immune disorders.

References

1. Kabat, E. A., and National Institutes of Health. 1991. *Sequences of Proteins of Immunological Interest*, 5th Ed. Public Health Service, National Institutes of Health, Washington.
2. Chothia, C., A. M. Lesk, A. Tramontano, M. Levitt, S. J. Smith-Gill, G. Air, S. Sheriff, E. A. Padlan, D. Davies, and W. R. Tulip. 1989. Conformations of immunoglobulin hypervariable regions. *Nature 342:877.*
3. Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A. M. Lesk. 1997. Antibody structure, prediction and redesign. *Biophys. Chem. 68:9.*
4. Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A. M. Lesk. 1998. Conformations of the third hypervariable region in the V_H domain of immunoglobulins. *J. Mol. Biol. 275:269.*
5. Knappik, A., L. Ge, A. Honegger, P. Pack, M. Fischer, G. Wellnhofer, A. Hoess, J. Wolle, A. Pluckthun, and B. Virnekas. 2000. Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J. Mol. Biol. 296:57.*
6. Lefranc, M. P. 2001. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res. 29:207.*
7. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol. 215:403.*
8. Brezinschek, H. P., S. J. Foster, R. I. Brezinschek, T. Dorner, R. Domiati-Saad, and P. E. Lipsky. 1997. Analysis of the human V_H gene repertoire: differential effects of selection and somatic hypermutation on human peripheral $CD5^+/IgM^+$ and $CD5^-/IgM^+$ B cells. *J. Clin. Invest. 99:2488.*
9. Monestier, M., and M. Zouali. 2002. Receptor revision and systemic lupus. *Scand. J. Immunol. 55:425.*
10. Raaphorst, F. M., E. Timmers, M. J. Kenter, M. J. Van Tol, J. M. Vossen, and R. K. Schuurman. 1992. Restricted utilization of germ-line V_H3 genes and short diverse third complementarity-determining regions (CDR3) in human fetal B lymphocyte immunoglobulin heavy chain rearrangements. *Eur. J. Immunol. 22: 247.*
11. Meffre, E., E. Davis, C. Schiff, C. Cunningham-Rundles, L. B. Ivashkiv, L. M. Staudt, J. W. Young, and M. C. Nussenzweig. 2000. Circulating human B cells that express surrogate light chains and edited receptors. *Nat. Immunol. 1:207.*
12. Klonowski, K. D., and M. Monestier. 2000. Heavy chain revision in MRL mice: a potential mechanism for the development of autoreactive B cell precursors. *J. Immunol. 165:4487.*
13. Klonowski, K. D., L. L. Primiano, and M. Monestier. 1999. Atypical V_H-D-J_H rearrangements in newborn autoimmune MRL mice. *J. Immunol. 162:1566.*
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.*
15. Sanz, I., S. S. Wang, G. Meneses, and M. Fischbach. 1994. Molecular characterization of human Ig heavy chain DIR genes. *J. Immunol. 152:3958.*
16. Tuaillon, N., and J. D. Capra. 1998. Use of D gene segments with irregular spacers in terminal deoxynucleotidyltransferase $(TdT)^{+/+}$ and $TdT^{-/-}$ mice carrying a human Ig heavy chain transgenic minilocus. *Proc. Natl. Acad. Sci. USA 95:1703.*
17. Ichihara, Y., M. Abe, H. Yasui, H. Matsuoka, and Y. Kurosawa. 1988. At least five D_H genes of human immunoglobulin heavy chains are encoded in 9-kilobase DNA fragments. *Eur. J. Immunol. 18:649.*
18. Gellert, M. 1992. Molecular analysis of V(D)J recombination. *Annu. Rev. Genet. 26:425.*
19. Tuaillon, N., A. B. Miller, P. W. Tucker, and J. D. Capra. 1995. Analysis of direct and inverted DJ_H rearrangements in a human Ig heavy chain transgenic minilocus. *J. Immunol. 154:6453.*
20. 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.*
21. Kompfner, E., P. Oliveira, A. Montalbano, and A. J. Feeney. 2001. Unusual germline DSP2 gene accounts for all apparent V-D-D-J rearrangements in newborn, but not adult, MRL mice. *J. Immunol. 167:6933.*
22. Brezinschek, H. P., R. I. Brezinschek, and P. E. Lipsky. 1995. Analysis of the heavy chain repertoire of human peripheral B cells using single-cell polymerase chain reaction. *J. Immunol. 155:190.*
23. Brezinschek, H. P., R. I. Brezinschek, T. Dorner, and P. E. Lipsky. 1998. Similar characteristics of the CDR3 of V_H1–69/DP-10 rearrangements in normal human peripheral blood and chronic lymphocytic leukaemia B cells. *Br. J. Haematol. 102:516.*
24. Yavuz, S., A. C. Grammer, A. S. Yavuz, T. Nanki, and P. E. Lipsky. 2001. Comparative characteristics of μ chain and α chain transcripts expressed by individual tonsil plasma cells. *Mol. Immunol. 38:19.*
25. Bauer, K., M. Zemlin, M. Hummel, S. Pfeiffer, J. Karstaedt, G. Steinhauser, X. Xiao, H. Versmold, and C. Berek. 2002. Diversification of Ig heavy chain genes in human preterm neonates prematurely exposed to environmental antigens. *J. Immunol. 169:1349.*
26. Rosner, K., D. B. Winter, R. E. Tarone, G. L. Skovgaard, V. A. Bohr, and P. J. Gearhart. 2001. Third complementarity-determining region of mutated V_H immunoglobulin genes contains shorter V, D, J, P, and N components than non-mutated genes. *Immunology 103:179.*
27. Chothia, C., and A. M. Lesk. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J. Mol. Biol. 196:901.*
28. Satow, Y., G. H. Cohen, E. A. Padlan, and D. R. Davies. 1986. Phosphocholine binding immunoglobulin Fab McPC603: an x-ray diffraction study at 2.7 A. *J. Mol. Biol. 190:593.*
29. Matsuda, F., E. K. Shin, Y. Hirabayashi, H. Nagaoka, M. C. Yoshida, S. Q. Zong, and T. Honjo. 1990. Organization of variable region segments of the human immunoglobulin heavy chain: duplication of the D5 cluster within the locus and interchromosomal translocation of variable region segments. *EMBO J. 9:2501.*
30. Matsuda, F., K. H. Lee, S. Nakai, T. Sato, M. Kodaira, S. Q. Zong, H. Ohno, S. Fukuhara, and T. Honjo. 1988. Dispersed localization of D segments in the human immunoglobulin heavy-chain locus. *EMBO J. 7:1047.*
31. Zong, S. Q., S. Nakai, F. Matsuda, K. H. Lee, and T. Honjo. 1988. Human immunoglobulin D segments: isolation of a new D segment and polymorphic deletion of the D1 segment. *Immunol. Lett. 17:329.*
32. Buluwela, L., D. G. Albertson, P. Sherrington, P. H. Rabbitts, N. Spurr, and T. H. Rabbitts. 1988. The use of chromosomal translocations to study human immunoglobulin gene organization: mapping D_H segments within 35 kb of the Cμ gene and identification of a new D_H locus. *EMBO J. 7:2003.*
33. Shiokawa, S., F. Mortari, J. O. Lima, C. Nunez, F. E. Bertrand, III, P. M. Kirkham, S. Zhu, A. P. Dasanayake, and H. W. Schroeder, Jr. 1999. IgM heavy chain complementarity-determining region 3 diversity is constrained by genetic and somatic mechanisms until two months after birth. *J. Immunol. 162: 6060.*
34. Girschick, H. J., and P. E. Lipsky. 2002. The κ gene repertoire of human neonatal B cells. *Mol. Immunol. 38:1113.*
35. Gauss, G. H., and M. R. Lieber. 1996. Mechanistic constraints on diversity in human V(D)J recombination. *Mol. Cell. Biol. 16:258.*
36. Mortari, F., J. Y. Wang, and H. W. Schroeder, Jr. 1993. Human cord blood antibody repertoire: mixed population of V_H gene segments and CDR3 distribution in the expressed Cα and Cγ repertoires. *J. Immunol. 150:1348.*
37. Moore, B. B., and K. Meek. 1995. Recombination potential of the human DIR elements. *J. Immunol. 154:2175.*

38. Gellert, M. 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem. 71:101.*
39. Kurosawa, Y., and S. Tonegawa. 1982. Organization, structure, and assembly of immunoglobulin heavy chain diversity DNA segments. *J. Exp. Med. 155:201.*
40. Yamada, M., R. Wasserman, B. A. Reichard, S. Shane, A. J. Caton, and G. Rovera. 1991. Preferential utilization of specific immunoglobulin heavy chain diversity and joining segments in adult human peripheral blood B lymphocytes. *J. Exp. Med. 173:395.*
41. Clausen, B. E., S. L. Bridges, Jr., J. C. Lavelle, P. G. Fowler, S. Gay, W. J. Koopman, and H. W. Schroeder, Jr. 1998. Clonally-related immunoglobulin V_H domains and nonrandom use of D_H gene segments in rheumatoid arthritis synovium. *Mol. Med. 4:240.*
42. Collins, A. M., M. Ikutani, D. Puiu, G. A. Buck, A. Nadkarni, and B. Gaeta. 2004. Partitioning of rearranged Ig genes by mutation analysis demonstrates D-D fusion and V gene replacement in the expressed human repertoire. *J. Immunol. 172:340.*
43. Ichihara, Y., H. Matsuoka, and Y. Kurosawa. 1988. Organization of human immunoglobulin heavy chain diversity gene loci. *EMBO J. 7:4141.*
44. Black, S. D., and D. R. Mould. 1991. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal. Biochem. 193:72.*
45. Kyte, J., and R. F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol. 157:105.*
46. Hopp, T. P., and K. R. Woods. 1983. A computer program for predicting protein antigenic determinants. *Mol. Immunol. 20:483.*
47. Betts, M. J., and Russell, R. B. 2003. Amino acid properties and consequences of substitutions. In *Bioinformatics for Geneticists*, 1st Ed. M. R. Barnes and I. C. Gray, eds. Wiley, Hoboken.
48. Feeney, A. J., B. R. Lawson, D. H. Kono, and A. N. Theofilopoulos. 2001. Terminal deoxynucleotidyl transferase deficiency decreases autoimmune disease in MRL-Fas^{lpr} mice. *J. Immunol. 167:3486.*
49. Lee, J., N. L. Monson, and P. E. Lipsky. 2000. The VλJ lambda repertoire in human fetal spleen: evidence for positive selection and extensive receptor editing. *J. Immunol. 165:6322.*
50. Nadel, B., S. Tehranchi, and A. J. Feeney. 1995. Coding end processing is similar throughout ontogeny. *J. Immunol. 154:6430.*
51. Boubnov, N. V., Z. P. Wills, and D. T. Weaver. 1995. Coding sequence composition flanking either signal element alters V(D)J recombination efficiency. *Nucleic Acids Res. 23:1060.*
52. Ma, Y., U. Pannicke, K. Schwarz, and M. R. Lieber. 2002. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell 108:781.*
53. Paull, T. T., and M. Gellert. 2000. A mechanistic basis for Mre11-directed DNA joining at microhomologies. *Proc. Natl. Acad. Sci. USA 97:6409.*
54. Gauss, G. H., I. Domain, C. L. Hsieh, and M. R. Lieber. 1998. V(D)J recombination activity in human hematopoietic cells: correlation with developmental stage and genome stability. *Eur. J. Immunol. 28:351.*
55. Klonowski, K. D., and M. Monestier. 2001. Ig heavy-chain gene revision: leaping towards autoimmunity. *Trends Immunol. 22:400.*