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

**References** This article cites 53 articles, 22 of which you can access for free at: http://www.jimmunol.org/content/172/11/6790.full#ref-list-1

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

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

**Email Alerts** Receive free email-alerts when new articles cite this article. Sign up at: http://jimmunol.org/alerts
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. Lipsky2,*

We analyzed 77 nonproductive and 574 productive human V\textsubscript{H}DJ\textsubscript{H} rearrangements with a newly developed program, JOINSOLVER. In the productive repertoire, the H chain complementarity determining region 3 (CDR\textsubscript{3\textsubscript{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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 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 3 D segments were positively selected and 3 were negatively selected, indicating that both molecular mechanisms and selection biased the D segment usage.

Diversity in the Ab repertoire of Ig H chains is mainly achieved by random recombination of V\textsubscript{H}, D, and J\textsubscript{H} segments and enzymatic modification of the V\textsubscript{H}DJ\textsubscript{H} junctions. Located at the joining of the V\textsubscript{H}, D, and J\textsubscript{H} segments, the H chain complementarity determining region 3 (CDR\textsubscript{3\textsubscript{H}})\textsuperscript{3} is the most diverse region in the Ig molecule. Structurally, the CDR\textsubscript{3\textsubscript{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 CDR\textsubscript{3\textsubscript{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 CDR\textsubscript{3\textsubscript{H}} has not yet been fully characterized. One reason for this lack of information about the CDR\textsubscript{3\textsubscript{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 CDR\textsubscript{3\textsubscript{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 CDR\textsubscript{3\textsubscript{H}} difficult. Adding to the difficulty of analyzing the composition of the CDR\textsubscript{3\textsubscript{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\textsubscript{H} and J\textsubscript{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 CDR\textsubscript{3\textsubscript{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 CDR\textsubscript{3\textsubscript{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\textsubscript{H}DJ\textsubscript{H} sequences, the usage of DIR segments (8, 10,
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⁺ 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 CDR3H region of the Ig genes expressed by human B cells and is available at http://joinsolver.iamsh.nih.gov. The strategy of JOINSOLVER is to search for D germline sequences flanking VH and JH germline genes. Additionally, it searches for P and N nucleotide additions in the VH-D and DJ 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 CDR3H 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 VH-DJH 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 VH region (after Kabat et al. (1)) and is a base-peak match in most of the human VH sequenced 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 CDR3H region. If the start of the CDR3H region is not yet identified, JOINSOLVER marks the CDR3H as not found and defines the CDR3H region until after V and J matching.

After the VH end of the CDR3H is defined, JOINSOLVER screens for the JH border of the CDR3H. A “C TGG GG” motif demarks the 3’ end of the CDR3H region and is conserved in all JH sequences. A similar algorithm is used to find the “C TGG GG” at the 3’ end of the CDR3H.

Once the CDR3H region is identified, VH, JH, and D assignment is conducted. A 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 JH 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 VH and JH regions are scored with an alignment score that assigns +3 to a nucleotide match and −4 for a mismatch between the unknown sequence and the germline (7).

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

In the event that the CDR3H 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 gene. The CDR3H region is defined as the region from codon 93 and the “TGG” of the “C TGG GG” motif. The VH1 end and JH1 start are defined the same way as if the CDR3H region had been found first.

After VH and JH 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 VH-D-JH distance (the distance in nucleotides between the end of the VHH segment and the beginning of the JH region). 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, mT, to the database of known human D germline genes. To accomplish this, a randomly generated set of 1 × 10³ sequences of a particular length, mg, was analyzed and searched for matches between the sequences and the D segment germline database. The value of mg is equivalent to the VH-D-JH distance. In the error simulation it 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 × 10³ randomly generated sequences for different values of mg, mT1, and mT2, where mg and mT1 and mT2 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 VH-DJH junction maintained the reading frame (RF) into the JH segment and contained no stop codons in the germline D segment or CDR3H segments. When the rearrangements failed to maintain the RF into the JH segment, or introduced stop codons during the rearrangement process, they were considered nonproductive. Junctional nucleotide additions between the VH and D or between D and JH 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 VH and JH 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 χ² goodness-of-fit test was conducted. The Student t test was used to analyze CDR3H length; VH-D-JH distance; D segment match length; P, and N nucleotides; VH, D, and JH excision.

AR composition analysis of CDR3H 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 CDR3H segment were not included to assess the nonrandom characteristics of the CDR3H. The total numbers of productive CDR3H, nonproductive CDR3H, 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 CDR3H length of 12, nonproductive CDR3H 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). χ² analysis was performed with the statistics program R. Distribution change of a residue was determined to be significant.
Table 1. Minimal D segment match length for a particular V_H-J_H distance, when the probability of a random match is 5 or 1% 

<table>
<thead>
<tr>
<th>V_H-J_H Distance (bp)</th>
<th>Match Length (m_l1) Required (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9–11</td>
<td>8</td>
</tr>
<tr>
<td>12–23</td>
<td>9</td>
</tr>
<tr>
<td>24–27</td>
<td>9</td>
</tr>
<tr>
<td>28–75</td>
<td>10</td>
</tr>
<tr>
<td>76–79</td>
<td>12</td>
</tr>
</tbody>
</table>

*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) 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 χ² sum. Where necessary, residues K, R, H or N, Q, S were combined to perform a χ² 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). 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. 1B) 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 consecutive 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.

**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) 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.
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 CDR3_H 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>
<thead>
<tr>
<th>Accession</th>
<th>V_H-J_H Distance (bp)</th>
<th>JOINSOLVER D segment</th>
<th>Consecutive match length (bp)</th>
<th>V-Base/ DNAPLOT D segment</th>
<th>BLAST score</th>
<th>Consecutive match length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z0363</td>
<td>15</td>
<td>D3-22</td>
<td>9</td>
<td>D5-12</td>
<td>61</td>
<td>6</td>
</tr>
<tr>
<td>Z0364</td>
<td>27</td>
<td>D2-2</td>
<td>17</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0369</td>
<td>27</td>
<td>D6-19</td>
<td>10</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0372</td>
<td>29</td>
<td>D3-22</td>
<td>10</td>
<td>D6-13</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Z0381</td>
<td>29</td>
<td>D3-16</td>
<td>10</td>
<td>D6-6</td>
<td>36</td>
<td>7</td>
</tr>
<tr>
<td>Z0389</td>
<td>51</td>
<td>D2-2</td>
<td>13</td>
<td>D6-25</td>
<td>63</td>
<td>7</td>
</tr>
<tr>
<td>Z0391</td>
<td>51</td>
<td>D2-2</td>
<td>13</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0414</td>
<td>23</td>
<td>D2-2</td>
<td>8</td>
<td>D4-17</td>
<td>35</td>
<td>6</td>
</tr>
<tr>
<td>Z0472</td>
<td>20</td>
<td>D1-26</td>
<td>9</td>
<td>D6-19</td>
<td>33</td>
<td>5</td>
</tr>
<tr>
<td>Z0478</td>
<td>24</td>
<td>D1-26</td>
<td>9</td>
<td>D2-15</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>Z0479</td>
<td>22</td>
<td>D1-7</td>
<td>11</td>
<td>D2-08</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Z0484</td>
<td>27</td>
<td>D3-10</td>
<td>13</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0488</td>
<td>63</td>
<td>D2-8 and D3-9</td>
<td>25 and 24</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0493</td>
<td>33</td>
<td>D1-26</td>
<td>28</td>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z0503</td>
<td>18</td>
<td>D3-10</td>
<td>9</td>
<td>D3-16</td>
<td>19</td>
<td>6</td>
</tr>
</tbody>
</table>

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

Table III. Mean CDR3_H length and V_H-J_H distance for productive and nonproductive rearrangements

<table>
<thead>
<tr>
<th></th>
<th>Nonproductive</th>
<th>Productive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CDR3_H bp</td>
<td>V_H-J_H bp</td>
</tr>
<tr>
<td>Peripheral B cells</td>
<td>57.5 ± 2.3</td>
<td>39.5 ± 2.1</td>
</tr>
<tr>
<td>Tonsilar plasma cells</td>
<td>45.8 ± 3.9</td>
<td>29.2 ± 3.0</td>
</tr>
<tr>
<td>IgG and IgM B cells</td>
<td>43.0 ± 3.8</td>
<td>26.4 ± 2.6</td>
</tr>
<tr>
<td>Mean</td>
<td>53.8 ± 1.9</td>
<td>36.2 ± 1.6</td>
</tr>
</tbody>
</table>

* Data shown represent mean base pairs ± SEM.
* Significant (p < 0.05) difference between nonproductive and productive rearrangements.
rearrangements were negatively selected.

Nonproductive rearrangements with multiple D segment fusions was significantly lower (p < 0.01) more often in the productive repertoire than that expected from random chance. The use of inverted D segments was 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\textsubscript{H} D\textsuperscript{5} D\textsuperscript{3} J\textsubscript{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\textsubscript{H} D\textsuperscript{5} D\textsuperscript{3} J\textsubscript{H}. One of these (Z80631) had two identified D segments, which were organized D\textsuperscript{7} D\textsuperscript{5} J\textsubscript{H}. The second had three identified D segments of which the middle one originated from upstream of the S' D segment but also S' 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 germ-line 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 germ-line 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 germ-line 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.
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...
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 D segments showed that they were frequently more enriched in hydrophilic amino acids in this RF. Although this was not a uniform finding, D segments (D6-6, D6-13, D6-19) encoded hydrophilic amino acids or glycine (D2-2, D2-15, D3-10, 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, 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 (χ² of 157 vs 885, respectively) (Fig. 5).

When the amino acid composition of the entire CDR3H 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 (χ² of 185 vs 1354, respectively, Fig. 5).

** Preferential rearrangement of 5’ D segments with 3’ JH 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 VHJH-rearrangements. From this analysis, there was...
Table VI. D segment reading frame use

<table>
<thead>
<tr>
<th>D Gene</th>
<th>Frequency of RF1 Usage</th>
<th>Frequency of RF2 Usage</th>
<th>Frequency of RF3 Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RF1</td>
<td>RF2</td>
<td>RF3</td>
</tr>
<tr>
<td></td>
<td>NP</td>
<td>P</td>
<td>NP</td>
</tr>
</tbody>
</table>
| 1-1    | GTRCT  | 0  | 0  | VQLER  | 0  | 0  | YMWMD  | 0  | 1 (100)
| 1-7    | GTRCT  | 0  | 4.5 (53) | V*LEL  | 1 (100) | 0  | YMWNY  | 0  | 4 (47)
| 1-14   | GTRCT  | 0  | 0  | V*PEP  | 0  | 0  | YNRHH  | 0  | 0
| 1-20   | GTRCT  | 0  | 2.5 (26) | V*HLL  | 0  | 0  | YMWMD  | 0  | 7 (74)
| 1-26   | GTRCT  | 2 (100) | 10 (32) | V*HLL  | 0  | 3 (10)  | VEGTY  | 0  | 18 (58)
| 2-2    | RIL**YQLLC  | 3 (23)  | 0  | YCSSTSCYA  | 5 (8)  | 17 (68)  | DIVVFAAM  | 5 (38) | 8 (32)
| 2-8    | RIL**WCMLY  | 1 (50)  | 0  | GYCTGVCYT  | 0  | 4 (80)  | DIVMVYAI  | 1 (50) | 1 (20)
| 2-15   | RIL**WW*LLL  | 1 (33)  | 0  | GCSCGSCYS  | 1 (33) | 13 (81)  | DIVVVVAAT  | 1 (33) | 3 (19)
| 2-21   | SILW**MLF  | 2 (40) | 0  | AIGCGDCYS  | 1 (20) | 1 (25)  | HVVIVIAI  | 2 (40) | 3 (75)
| 2-3    | VLRF**WLLL  | 1 (33) | 10 (29) | YDFWGGTYT  | 2 (66) | 14 (41)  | ITIPGVVII  | 1 (33) | 10 (29)
| 2-9    | VLRFD*WLLL  | 0  | 5 (33) | VDIILQYYN  | 0  | 10 (66)  | ITIPLVI  | 2 (100) | 0
| 2-10   | VLLWF*ELL  | 2 (40) | 10 (24) | YYGSGGGYN  | 1 (20) | 21 (51)  | ITIVRGVII  | 2 (40) | 10 (24)
| 2-16   | VLL**LRLGELY  | 0  | 1 (13) | YYDVNGSGYAT  | 7 (88) | 0  | IMTIFGGVMI  | 0  | 0
| 2-22   | VLL**WLLL  | 4 (50) | 0  | YYDSSGYY  | 3 (38) | 40 (87)  | ITIVVVI  | 1 (13) | 6 (13)
| 4-4/11  | *LQ**L  | 0  | 0  | DINSY  | 0  | 2 (50)  | TTNT  | 0  | 2 (50)
| 4-17   | *LQ**L  | 2 (50) | 0  | DYGDY  | 2 (50) | 11 (52)  | TTNT  | 0  | 10 (48)
| 4-23   | *LQ**L  | 0  | 0  | DYGGSNS  | 0  | 3 (60)  | TTNT  | 0  | 2 (40)
| 5-5/18  | VDTMW  | 0  | 6 (50) | NQLW  | 1 (9)  | GYSYD  | 1 (100) | 5 (42)
| 5-12   | VDIVATI  | 1 (50) | 3 (27) | WI*WRL  | 1 (9)  | GYSYD  | 1 (50) | 7 (64)
| 5-24   | VENATI  | 1 (100) | 3 (50) | *RLQL  | 0  | 1 (17)  | RDGNY  | 0  | 2 (33)
| 6-6    | ESSSSS  | 1 (100) | 4 (44) | SIAAR  | 0  | 5 (56)  | V*QLV  | 0  | 0
| 6-13   | GYSSHY  | 0  | 13 (42) | GIAAA  | 2 (100) | 14 (45)  | V*QQLV  | 0  | 4 (13)
| 6-19   | GYSSHY  | 1 (33) | 15 (47) | GIAAVG  | 2 (66) | 12 (38)  | V*QQLV  | 0  | 5 (16)
| 6-25   | GYSSGY  | 0  | 0  | GIAAA  | 0  | 0  | V*QQLV  | 0  | 0
| 7-27   | IZG**  | 0  | 2 (40) | *LQ  | 0  | 0  | NWG  | 0  | 3 (60)

* 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).

** Significantly (p < 0.05) higher frequency when compared to the expected random frequency.

*** Significantly (p < 0.05) frequency between nonproductive and productive.

** FIGURE 5. Distribution of the ARs in the D segment (A) and CDR3 (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.
The V<sub>H</sub>-D junction was more frequent in the productive than in the nonproductive rearrangements (Table VIII). D segment excision was similar in the nonproductive or productive repertoires, with more excision at the 3′ end. The I<sub>D</sub> coding end was excised to a significantly (p < 0.01) greater degree in both nonproductive and productive rearrangements than the V<sub>H</sub> and D<sub>5′</sub> coding ends. P nucleotides were significantly (p < 0.01) more abundant in the V<sub>H</sub>D junctions than in the D<sub>1</sub>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<sub>H</sub>D and D<sub>1</sub>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<sub>H</sub>. Within the CDR3<sub>H</sub>, 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<sub>H</sub>-J<sub>H</sub> 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 segments with a 95% probability strategy, we were able to identify D segments significantly (p < 0.05) higher percentage of sequences with microhomology than the V<sub>H</sub>D junction.

**Exonuclease activity**

The V<sub>H</sub> coding end had significantly (p < 0.01) less exonucleolytic excision when compared with the D and I<sub>D</sub> 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 I<sub>D</sub> coding end was excised to a significantly (p < 0.01) greater degree in both nonproductive and productive rearrangements than the V<sub>H</sub> and D<sub>5′</sub> coding ends. P nucleotides were significantly (p < 0.01) more abundant in the V<sub>H</sub>D junctions than in the D<sub>1</sub>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<sub>H</sub>D and D<sub>1</sub>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.

---

**Table VII. TdT activity and frequency of microhomology in the V<sub>H</sub>D or D-J<sub>H</sub> junctions of nonproductive and productive rearrangements**

<table>
<thead>
<tr>
<th></th>
<th>Nonproductive</th>
<th></th>
<th>Productive</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V&lt;sub&gt;H&lt;/sub&gt;D</td>
<td>D-J&lt;sub&gt;H&lt;/sub&gt;</td>
<td>V&lt;sub&gt;H&lt;/sub&gt;D</td>
<td>D-J&lt;sub&gt;H&lt;/sub&gt;</td>
</tr>
<tr>
<td>Peripheral B cells</td>
<td>11.0 ± 1.1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9.5 ± 1.1&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4.6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>9.2&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Tonsilar plasma cells</td>
<td>7.0 ± 2.9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8.5 ± 1.2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>IgG and IgM cells</td>
<td>6.3 ± 1.4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>8.5 ± 2.0&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mean</td>
<td>10.2 ± 1.0&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9.2 ± 0.9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>3.6</td>
<td>7.2</td>
</tr>
</tbody>
</table>

<sup>a</sup> Significant (p < 0.05) differences between nonproductive and productive rearrangements.

<sup>b</sup> Significant (p < 0.05) differences between the V<sub>H</sub>D and D-J<sub>H</sub> junctions.

* One rearrangement (280724) 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<sub>H</sub>D junction and analyzing junctional diversity in this region.
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, D4-17, 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\textsubscript{H}DJ\textsubscript{H} recombination. Eight D segments (D2-2, D2-8, D2-21) were significantly overrepresented in the nonproductive repertoire, suggesting that they were preferentially used during V\textsubscript{H}DJ\textsubscript{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\textsubscript{H}DJ\textsubscript{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 CDR\textsubscript{3}\textsubscript{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\textsubscript{H}DJ\textsubscript{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 VIII. Exonuclease activity in the V\textsubscript{H}D or DJ\textsubscript{H} junctions |
|---------------------------------|------------------|------------------|------------------|------------------|------------------|
| Excision Site                  | Peripheral B Cells | Tonsilar Plasma Cells | IgG and IgM Cells | Mean             |
| Nonproductive                 | V\textsubscript{H}D | 2.2 ± 0.4         | 2.2 ± 0.8        | 1.1 ± 0.6        | 2.1 ± 0.3\textsuperscript{a,b} |
|                               | DJ\textsubscript{H} | 4.6 ± 0.6         | 4.8 ± 2.5        | 3.2 ± 1.1        | 4.5 ± 0.5\textsuperscript{a} |
|                               | D3\textsuperscript{b} | 4.7 ± 0.7         | 3.5 ± 1.7        | 8.7 ± 3.1        | 5.0 ± 0.6\textsuperscript{a} |
|                               | J\textsubscript{H}\textsuperscript{a} | 6.5 ± 0.7         | 7.5 ± 1.2        | 7.6 ± 1.6        | 6.8 ± 0.6\textsuperscript{a,b} |
| Productive                    | V\textsubscript{H}D | 1.9 ± 0.1         | 1.5 ± 0.2        | 1.6 ± 0.2        | 1.8 ± 0.1\textsuperscript{a,b} |
|                               | DJ\textsubscript{H} | 4.5 ± 0.3\textsuperscript{a} | 4.9 ± 0.4        | 5.3 ± 0.5        | 4.7 ± 0.2\textsuperscript{a,b} |
|                               | D3\textsuperscript{b} | 5.5 ± 0.3         | 5.2 ± 0.5        | 5.3 ± 0.5        | 5.4 ± 0.2\textsuperscript{a} |
|                               | J\textsubscript{H}\textsuperscript{a} | 6.0 ± 0.3         | 5.9 ± 0.4        | 5.6 ± 0.4        | 5.9 ± 0.2\textsuperscript{a,b} |

\textsuperscript{a} Significant (p < 0.05) differences between the excision in the V\textsubscript{H} coding end and the other coding ends.

\textsuperscript{b} Significant (p < 0.05) differences between the excision in the DJ\textsubscript{H} coding end and the other coding ends.

| Table IX. Number of P nucleotides (bp) |
|--------------------------------------|------------------|------------------|------------------|------------------|
| Junction                             | Peripheral B Cells | Tonsilar Plasma Cells | IgG and IgM Cells | Mean             |
| Nonproductive                       | V\textsubscript{H}D | 0.6 ± 0.1\textsuperscript{a} | 0.7 ± 0.3\textsuperscript{a} | 0.3 ± 0.3\textsuperscript{a} | 0.6 ± 0.1\textsuperscript{a} |
|                                   | DJ\textsubscript{H} | 0.2 ± 0.1       | 0               | 0.2 ± 0.2        | 0.1 ± 0.1        |
| Productive                         | V\textsubscript{H}D | 0.4 ± 0.1\textsuperscript{a} | 0.5 ± 0.1\textsuperscript{a} | 0.7 ± 0.1\textsuperscript{a} | 0.5 ± 0.4\textsuperscript{a} |
|                                   | DJ\textsubscript{H} | 0.2 ± 0.04      | 0.2 ± 0.1       | 0.2 ± 0.1        | 0.2 ± 0.03       |

\textsuperscript{a} Significant difference between the number of P nucleotides in V\textsubscript{H}D vs DJ\textsubscript{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_{H}DJ_{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_{H}D_{1}D_{2}D_{3}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_{H}DJ_{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, Z80697, 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_{H}D_{1}DJ_{H} rearrangements with D-D junctions 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}→DJ_{H} junction (43). Changes in RF impact the amino acid sequence which changes the hydrophatic 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 productive and 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 Hoppe/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 hydrophaticity 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_{H}D 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_{H}D 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_{H}D 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_{H}D join. This was especially notable when the degree of exonucleolytic processing of the V_{H} and J_{H} segments were compared and could relate to the primary sequence of these regions. The S' 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_{H}D 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 “hairpinning” (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 VH-DJ rearrangement.

The formation of microhomology may influence the development of the CDR3H 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 (61). Notably, the frequency of microhomologies seems to be less than that found in fetal (49) or neonatal (25) arrangements.

In analyzing the nonproductive repertoire for VH, DJ, and JH segments, a low frequency of microhomologies was observed. The persistent expression of RAG proteins and the presence of TdT may reflect the continuous need to access the coding ends (35, 53, 54). The net result would be the tendency for DJH segments to be found preferentially rearranged with VH segments because the initial rearrangements would be deleted as the progressive rearrangement process proceeds. The finding that the distribution of VH genes is random suggests that this process occurs before VH-DJH 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 JOINSERVER has permitted a detailed analysis of the human adult CDR3H 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 human Ig molecules. Importantly, this approach and the database generated should be of great value in determining abnormalities in human Ig molecules. Importantly, this approach and the database generated should be of great value in determining abnormalities in human Ig molecules. Importantly, this approach and the database generated should be of great value in determining abnormalities in human Ig molecules.


