The Journal of Immunology

RESEARCH ARTICLE | JANUARY 01 2004

Partitioning of Rearranged Ig Genes by Mutation Analysis Demonstrates D-D Fusion and V Gene Replacement in the Expressed Human Repertoire[1] FREE

Andrew M. Collins; ... et. al

J Immunol (2004) 172 (1): 340–348.
https://doi.org/10.4049/jimmunol.172.1.340

Partitioning of Rearranged Ig Genes by Mutation Analysis Demonstrates D-D Fusion and V Gene Replacement in the Expressed Human Repertoire[1]

Andrew M. Collins,[2]* Masashi Ikutani,* Daniela Puiu,† Gregory A. Buck,† Aradhita Nadkarni,* and Bruno Gaeta*

The accurate partitioning of Ig H chain V_HDJ_H junctions and L chain V_LJ_L junctions is problematic. We have developed a statistical approach for the partitioning of such sequences, by analyzing the distribution of point mutations between a determined V gene segment and putative Ig regions. The establishment of objective criteria for the partitioning of sequences between V_H, D, and J_H gene segments has allowed us to more carefully analyze intervening putative nontemplated (N) nucleotides. An analysis of 225 IgM H chain sequences, with five or fewer V mutations, led to the alignment of 199 sequences. Only 5.0% of sequences lacked N nucleotides at the V_HD junction (N1), and 10.6% at the DJ_H junction (N2). Long N regions (>9 nt) were seen in 20.6% of N1 regions and 17.1% of N2 regions. Using a statistical analysis based upon known features of N addition, and mutation analysis, two of these N regions aligned with D gene segments, and a third aligned with an inverted D gene segment. Nine additional sequences included possible alignments with a second D segment. Four of the remaining 40 long N1 regions included 5′ sequences having six or more matches to V gene end motifs, which may be the result of V gene replacement. Such sequences were not seen in long N2 regions. The long N regions frequently seen in the expressed repertoire of human Ig gene rearrangements can therefore only partly be explained by V gene replacement and D-D fusion. ***The Journal of Immunology*****, 2004, 172: 340–348.**

All Ig H chain V regions are encoded by gene rearrangements that take place early in B cell ontogeny. The rearranged H chain V gene is the outcome of the recombination of gene segments selected from each of three sets of germline gene segments: V_H, D, and J_H (1). Each rearranged H chain V gene encodes a polypeptide composed of supporting β-pleated sheet scaffold or framework regions (FR)[3] and three hypervariable loops, termed complementarity determining regions (CDRs). These CDRs are the principal points of interaction between the H chain and its specific Ag. The CDR3 of the H chain is encoded by the V_H-D-J_H junction, and is the most diverse of all the hypervariable loops (2). However, this diversity is not generated solely by gene segment rearrangements. Diversity is enhanced by processing of the ends of the recombining elements through exonuclease activity and by nontemplated (N) (3) and palindromic (P) (4) nucleotide addition. Following clonal selection, point mutations introduced by the process of somatic hypermutation within the germinal center reaction introduce an additional level of diversity (5).

Sequencing of junction regions, and the identification of their embedded D genes, has been critical to the discovery of the processes contributing to the diversity of the CDR3 regions. In addition to the well-known processes of exonuclease activity, and the addition of N and P nucleotides, other processes have been reported, including the use of inverted D gene segments and gene conversion (6), the insertion and deletion of trinucleotides (7), the use of D genes with irregular recombination sequences (8), and the use of more than one D gene segment in a single rearrangement (9). However, the veracity of some of these processes remains in dispute. For example, although some studies have presented evidence against the possibility of D-D recombination (10), other studies have reported such recombination events to be relatively common (11, 12). The failure to resolve this and other controversies can in part be explained by the frequent difficulties involved in arriving at an unambiguous identification of the various elements within junction regions. These alignment problems have seemed to be particularly difficult because of the effects of somatic point mutations.

Following a recent report that the extent of somatic hypermutation decays exponentially, downstream of the promoter region (13), the distribution of mutations has become amenable to new kinds of analysis, which can be applied to improve the partitioning of rearranged genes. The likelihood of a mutation occurring in any part of a sequence can now be calculated from its position within the sequence, as well as by reference to known mutational hot spots and cold spots within the sequence (14).

We have developed a statistical analysis based upon elements of the hypermutation process, to objectively evaluate any proposed partitioning of an Ig gene sequence between two elements. The analysis uses trinucleotide mutability scores that we have determined after analyzing the frequency and 5′ to 3′ distribution, within germline sequences, of each trinucleotide. The mutability scores for the different trinucleotides can be used to determine the mutability of Ig sequences, and of parts of sequences. These sequence mutability scores can then be used to predict the likely

*School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia; and †Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284

Received for publication May 29, 2003. Accepted for publication October 24, 2003.

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] This study was supported by a grant from the National Health and Medical Research Council of Australia.

[2] Address correspondence and reprint requests to Dr. Andrew Collins, School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia. E-mail address: a.collins@unsw.edu.au

[3] Abbreviations used in this paper: FR, framework region; CDR, complementarity determining region; N, nontemplated; P, palindromic.

0022-1767/04/$02.00

distribution of somatic point mutations within V segments, and between V segments and putative N regions and D segments.

The predictability of mutations within different regions of rearranged V genes makes it possible to develop new, objective criteria with which to identify D gene segments within the V_HDJ_H junction. These criteria exclude matches where apparent identity with D regions is likely to occur as a consequence of the known nucleotide preferences of TdT (15).

Such an analysis was applied to a set of relatively unmutated IgM H chain sequences. Improved identification of the ends of rearranged gene segments led to the development of a dataset of N nucleotides that could be identified with little ambiguity. The study of this N nucleotide dataset showed that human Ig junction regions frequently include long N regions. These regions were investigated for the presence of additional D gene segments. The reality of D-D fusions involving both orientations of D gene segments was confirmed, and evidence of V gene replacement was also seen. Analysis of N nucleotides also revealed nucleotide patterns that cannot presently be explained, which may represent the molecular signatures of presently unknown aspects of Ig gene rearrangement.

Materials and Methods

Trinucleotide mutability scores

Forty-nine nonproductively rearranged Ig H chain V gene sequences, obtained from GenBank database, were aligned using the VQUEST software (16), and somatic point mutations in each V segment sequence were noted by reference to the ImMunoGeneTics (IMGT) Ig gene database (16). Trinucleotide mutability scores were determined by first counting the number of occurrences of each trinucleotide in the germline V gene segments from which each of the 49 sequences were derived. For each overlapping trinucleotide, the sequence position (in the germline sequence) of the first nucleotide of the trinucleotide was recorded. For each trinucleotide, the trinucleotide counts were then adjusted to account for the position of each trinucleotide within the sequence, given the exponential decay of the mutation rate along the V gene sequences, as determined by Rada and Milstein (13): an adjusted trinucleotide count $C'_{NNN,S}$ was calculated, for each trinucleotide *NNN* in the sequence *S*, using the formula:

$$C'_{(NNN,S)} = \sum_{i \in P_{(NNN,S)}} e^{-0.0024i} \tag{1}$$

where *N* is any nucleotide A, G, T, or C, and $P_{(NNN,S)}$ is the set of sequence positions where the trinucleotide *NNN* is observed in sequence *S*.

This adjusted trinucleotide count was used to determine the expected number of mutations of each trinucleotide, in each of the nonproductively rearranged sequences, assuming that all mutations were randomly distributed through the sequences. The expected mutation frequency for trinucleotide *NNN* in sequence *S* ($Fe_{(NNN,S)}$) allows for the fact that most observed mutations will affect three overlapping trinucleotides, and was calculated as follows:

$$Fe_{(NNN,S)} = \frac{T_{(NNN,S)} \times C'_{(NNN,S)}}{\sum_{i \in A} C'_{(i,S)}} \tag{2}$$

where $T_{(NNN,S)}$ is the total number of mutated *NNN* trinucleotides in *S*, and *A* is the set of all possible trinucleotides *nnn*, where *n* is A, G, C, or T.

The number of times each trinucleotide was expected to mutate ($Fe_{(NNN,S)}$) and the number of observed mutations ($Fo_{(NNN,S)}$) were determined, and summed for each of the 49 sequences, and the trinucleotide mutability scores (M_{NNN}) were finally calculated as follows:

$$M_{NNN} = \frac{\sum_{i \in D} Fo_{(NNN,i)}}{\sum_{i \in D} Fe_{(NNN,i)}} \tag{3}$$

where *D* is the set of 49 sequences.

V gene segment mutability scores

A representative allele from each of the Ig H chain V gene segments in the IMGT database (17) was analyzed to calculate their relative tendencies to mutate, as a consequence of the frequencies of hot spots and cold spots within their sequences. The mutability score (M_V) of any V gene sequence was calculated as the sum of the trinucleotide mutability scores (M_{NNN}) calculated for the overlapping trinucleotides that make up the sequence, after adjustment for the position ($P(t)$) of each of those trinucleotides within the sequence, as follows:

$$M_V = \sum_{t \in B(V)} e^{-0.0024P(t)} M_t \tag{4}$$

where $B(V)$ is the set of all overlapping trinucleotides in sequence *V*.

Mutability scores for sequence regions including CDRs (M_{CDR}) and FRs (M_{FR}) were also calculated using Equation 4.

Ag selection

The expected numbers of mutations in the CDR1 and CDR2 regions of 69 mutated H chain sequences (18, 19), obtained from GenBank database, were determined by reference to the mutability scores (M_{CDR} and M_{FR}) of the combined CDR1 and CDR2 sequences and the combined FR1, FR2, and FR3 sequences. These two mutability scores were then used to determine the probability (*pMut*) that any mutation will occur in either the FR or CDR regions:

$$pMut_{CDR} = M_{CDR}/(M_{FR} + M_{CDR}) \tag{5}$$

$$pMut_{FR} = M_{FR}/(M_{FR} + M_{CDR}) = 1 - pMut_{CDR} \tag{6}$$

The probabilities of the distributions of mutations in the FR and CDR regions of each of the 69 gene sequences, containing known numbers of mutations, were then calculated using the binomial distribution. The Ag selection factor was finally calculated as the mean ratio of observed V segment CDR mutations over expected CDR mutations, for the 69 sequences:

$$R = \frac{N_{CDR}}{pMut_{CDR}(N_{FR} + N_{CDR})} \tag{7}$$

where N_S is the number of mutations observed in sequence *S*.

This selection factor *R* was then used to calculate an adjusted mutability score (M'_S) for any sequence *S* containing framework sequences and CDR sequences by the following modification of Equation 4:

$$M'_S = \sum_{t \in B(FR)} e^{-0.0024P(t)} M_t + R \sum_{t \in B(CDR)} e^{-0.0024P(t)} M_t \tag{8}$$

where $B(FR)$ is the set of all overlapping trinucleotides in framework sequences and $B(CDR)$ the set of all overlapping trinucleotides in CDR sequences. Note that $P(t)$ is the position of the trinucleotide relative to the first base of the whole Ig sequence.

Partitioning of rearranged H chain V genes

Two hundred twenty-five full-length IgM sequences, with five or fewer mutations within their V gene segments, were identified among the 729 IgM sequences in the IMGT database (17). The sequences were carefully aligned, with particular focus upon each boundary within the CDR3. To partition any junction sequence, the putative V_H, D, and J_H gene segments were first determined using the VQUEST program (17). Partitioning of any sequence (*S*) was then arrived at by determining the mutability (M_S) of the germline sequence from which it was derived, and determining the probability that any number of somatic point mutations could occur in that sequence, by reference to the mutability score of its associated germline V gene segment (M_V), the number of mutations seen in that V sequence, and the binomial distribution. Consideration of the distribution of mutations within the ends of V, D, and J gene segments, in the light of these probabilities, clarified whether or not some apparent mutations were more likely to be evidence of exonuclease activity and N nucleotide addition.

In this way, the most likely 3′ ends of the V gene segments, and then the 5′ ends of the putative D gene segments, were determined. The intervening sequences were designated the putative N1 regions of N nucleotide addition. The 5′ ends of the J gene segments and the 3′ ends of the D gene segments were then similarly determined, and the intervening N2 regions were identified. Finally, where there was no evidence of exonuclease activity at a gene segment end, putative N1 and N2 regions were examined for palindromic sequences, and such nucleotides were designated as P nucleotides. Alignment of a first D gene segment was only accepted if a minimum of 8 consecutive matching nucleotides were seen, or 9 matches in a 10-nt sequence.

The acceptance criteria for a second D gene alignment depended upon the length of the putative N(D2)N region under investigation. The probabilities that N nucleotide addition could give rise to apparent D segments

were determined by first identifying all unique 6-, 7-, 8-, 9-, 10-, 11-, and 12-nt sequences that can be produced from known D gene segments, and determining their probabilities using probabilities of N addition based upon the known TdT nucleotide preferences (15), as follows: $p(A) = 0.15$, $p(T) = 0.15$, $p(G) = 0.6$, and $p(C) = 0.1$. Calculations, which are more fully described below, included the equal likelihood that G nucleotides in N regions could arise from G addition to the sense strand, or from C addition to the antisense strand (20). The analysis was performed for complete matches, as well as for sequences with one or two mismatches. The resulting probabilities were used to determine the likelihood that matches of varying lengths would be seen, given the length of the N(D2)N region under investigation. For example, if the probability that 8 N nucleotides will align perfectly with a D segment is 0.005, then the probability that such an alignment might be seen in an N(D2)N region of 18 nt is 11×0.005, or 0.055, for 11 overlapping sequences of 8 nt are found in an 18-nt sequence. These calculations were used to determine the acceptance criteria for second D gene segments in each N region, by setting 95% confidence limits based on the cumulative distribution function. These confidence limits describe the least likely sequences, which individually were highly improbable, and which together might be seen with a probability of 0.05. Identical acceptance criteria were developed for inverted D gene sequences. In practice, most long N regions were between 10 and 16 nt in length, and required matches of 8 consecutive nt, or 1 mismatch in 10 or 11 nt. Alignment including two or more mismatches could be excluded on the basis of the mutation analysis. Identification of a second D gene in the N1 region was accepted only if that D gene segment was located 5′ of the primary D gene segment, within the germline. Similarly, a second D gene in the N2 region was only accepted if that D gene segment was located 3′ of the primary D gene segment, within the germline.

Analysis of the GC content of putative N regions

To analyze whether putative N regions carried the G/C features of N regions, including high G or C content, and homogeneity of either G or C (20), the probability of the particular frequencies of G, C, and A/T nucleotides within the N regions was determined as follows:

$$p(N) = 0.5\binom{w+g+c}{w}\binom{g+c}{g}(p_W^w p_G^g p_C^c + p_W^w p_C^g p_G^c) \qquad (9)$$

where w is the number of A or T (ambiguity code W), g is the number of G, and c is the number of C nucleotides.

The probabilities of an A or T insertion (p_W) was set as 0.3, the probability of an insertion of G (p_G) as 0.6, and the probability of C insertion (p_C) as 0.1 as determined by Basu et al. (15). The equation assumes that concatenation of strands does not occur (20), and allows for the fact that, for example, C additions may appear in the sense strand through addition of G nucleotides to the antisense strand.

The possible influence of base stacking was investigated, in sets of the various putative N nucleotide regions. The expected frequency and the cumulative distribution function were determined for each dinucleotide, using probabilities calculated from the actual nucleotide frequencies seen in the N regions, and the binomial distribution. The cumulative distribution function was then used to determine the 95% confidence limits for the observed dinucleotide frequencies.

Results

Trinucleotide mutability

Trinucleotide mutability scores were determined by analysis of the distribution of mutations in V_H gene segments of 49 nonproductively rearranged H chain genes. By analyzing nonproductive rearrangements, any contribution of Ag selection to the observed patterns of mutation could be avoided. The results are presented as Table I. Mutability scores ranged from 0.28 for TTG, to 3.06 for GTA. The mean mutability of the 8 trinucleotides encompassed by the 4-nt hot spot RGYW was 1.65, and of the 8 trinucleotides in the complementary WRCY sequence was 1.37. The mean mutability of the 8 WAN trinucleotides was 1.58. Together, RGYW/WRCY and WAN trinucleotides accounted for 9 of the 10 most highly mutable trinucleotides. Biased distribution of trinucleotides within germline sequences led to significant adjustments to some scores. Nineteen mutability scores were adjusted by 10% or more, and 5′ bias in the distribution of the ACG and ATT trinucleotides led to upward adjustments of ~25%.

Table I. *Trinucleotide mutability scores determined from an analysis of 12,361 trinucleotides in 47 nonproductive human Ig H chain rearrangements*

Trinucleotide	Frequency	Observed Mutations	Expected Mutations	Unadjusted Mutability Score	Adjusted Mutability Score[a]
AAA	94	13	19.95	0.71	0.65
AAC	103	25	18.09	1.24	1.38
AAG	285	65	56.59	1.17	1.15
AAT	52	21	10.46	2.06	2.01
ACA	192	34	32.08	0.90	1.06
ACC	284	66	57.68	1.19	1.14
ACG	76	18	12.01	1.21	1.50
ACT	239	61	47.31	1.30	1.29
AGA	193	32	34.99	0.85	0.91
AGC	347	142	66.83	2.09	2.12
AGG	186	25	35.03	0.69	0.71
AGT	275	73	55.19	1.36	1.32
ATA	106	29	19.92	1.40	1.46
ATC	217	57	43.78	1.34	1.30
ATG	91	19	16.04	1.07	1.18
ATT	111	29	17.54	1.33	1.65
CAA	182	39	33.45	1.09	1.17
CAC	226	47	41.54	1.06	1.13
CAG	350	99	65.46	1.45	1.51
CAT	131	25	22.93	0.98	1.09
CCA	325	64	52.60	1.01	1.08
CCC	260	28	47.83	0.55	0.46
CCG	213	29	33.60	0.70	0.75
CCT	353	42	63.58	0.61	0.57
CGA	49	5	6.87	0.52	0.64
CGC	144	16	22.96	0.57	0.56
CGG	148	18	25.05	0.62	0.62
CGT	82	19	12.41	1.18	1.25
CTA	153	53	27.12	1.77	1.67
CTC	256	28	42.71	0.56	0.60
CTG	573	98	98.82	0.87	0.86
CTT	94	14	18.69	0.76	0.64
GAA	236	50	40.04	1.08	1.05
GAC	212	20	34.28	0.48	0.49
GAG	289	51	49.04	0.90	0.90
GAT	188	33	35.72	0.90	0.87
GCA	122	34	18.93	1.42	1.73
GCC	327	29	54.91	0.45	0.49
GCG	59	4	10.37	0.35	0.30
GCT	246	104	41.21	2.16	2.15
GGA	441	68	80.42	0.79	0.71
GGC	160	20	25.15	0.64	0.76
GGG	325	47	59.43	0.74	0.71
GGT	199	33	37.83	0.85	0.78
GTA	127	63	18.80	2.53	3.06
GTC	250	35	44.07	0.72	0.65
GTG	289	54	48.96	0.95	0.96
GTT	92	29	17.33	1.61	1.29
TAA	22	8	4.39	1.86	1.93
TAC	250	86	42.99	1.76	1.76
TAG	77	30	11.89	1.99	2.13
TAT	154	42	22.77	1.39	1.61
TCA	250	48	42.35	0.98	0.91
TCC	280	38	47.81	0.69	0.69
TCG	75	8	13.50	0.55	0.47
TCT	238	28	42.42	0.60	0.61
TGA	227	13	37.60	0.29	0.31
TGC	103	12	18.30	0.60	0.61
TGG	432	53	78.51	0.63	0.58
TGT	251	49	39.51	1.00	1.01
TTA	117	28	18.78	1.22	1.32
TTC	120	16	21.63	0.68	0.61
TTG	60	4	11.35	0.34	0.28
TTT	23	2	4.49	0.44	0.48

[a] Mutability scores were adjusted to account for the 5′ to 3′ distribution of the trinucleotides within the sequences.

Ag selection

To investigate the affect of Ag selection upon mutation patterns, productively rearranged, class-switched sequences were analyzed.

Analysis of 69 mutated, H chain sequences confirmed a tendency for mutations to accumulate in CDR1 and CDR2, at rates higher than expected solely on the basis of the nucleotide composition of the sequences, and their distance from the Ig promoter. This tendency is most evident in sequences with higher total V gene mutations, as seen in Fig. 1. The Ag selection factor (*R*), the mean ratio of observed V segment CDR mutations over expected mutations, was calculated to be 1.54.

V_H gene segment mutability

Mutability scores were calculated, for full-length germline V_H gene segments, from codon 1 to codon 104, the last codon of FR3, before the start of the CDR3. A single allele was selected for analysis of each gene, and the results are presented for each gene family as Table II. The mean mutability M_V of the segments was 225.8 (SD = 7.2), and scores ranged from 212.6 for IGHV2-70*01 to 240.3 for the most mutable V gene segment, IGHV6-1*01. Considerable allelic variation was also seen. For example, the 11 allelic variants of IGHV1-69 had a mean mutability of 223.9 (SD = 9.2) and ranged from 197.6 to 228.7. A few alleles with truncated FR3 regions had significantly lower mutability scores. For example, M_V for IGHV2-5*03 was 174.6.

These V gene segment mutability scores were subsequently used in the partitioning of rearranged IgM sequences, and this partitioning then allowed all the calculations that are shown below. As a guide to the process, a typical rearranged D segment with a length of ~15 nt would have a score of ~10–20, and the probability that any mutation would occur in the D segment rather than the V segment would be 0.05–0.10. Even in sequences where the V segments had five mutations, the probabilities that the D segments would include more than one mutation were almost always <0.05.

Analysis of N nucleotide addition

Two hundred twenty-four IgM H chain rearranged V genes were analyzed, and satisfactory V_HDJ_H alignments could be determined for 199 sequences (89%). Many of the sequences that could not be aligned included seven consecutive nucleotide matches to germline D sequences, but this was not considered sufficient to allow confident alignment. This introduced a systematic bias against very short D segments, but ensured the reliability of the 199 alignments that were then subjected to further analysis. The lengths of putative N1 regions, at the V_HD junctions of the 199 sequences, and of putative N2 regions at the DJ_H junction were determined, and the results are presented as Fig. 2. N nucleotide addition was absent from only 10 (5.0%) of the V_HD junctions, and from 21 (10.6%) of the DJ_H junctions. Eleven N1 sequences were identified with ≥15 nt, including 1 sequence of 20 and 1 sequence of 21 nt. Thirteen N2 sequences were identified with ≥15 nt, including 1 sequence of 22 and 2 sequences of 23 nt.

Table II. *Mutability scores (M_V) of functional human H chain V gene segments*[a]

V Gene Family	Genes	Mean M_V	SD of M_V	Range of M_V
VH1	10	224.6	5.4	213.4–232.4
VH2	3	215.0	2.3	212.6–217.2
VH3	17	227.6	7.0	214.5–240.3
VH4	10	228.2	5.1	221.8–238.3
VH5	2	216.7		213.7–219.7
VH6	1	240.3		
VH7	1	218.7		

[a] Only one allele was scored for each gene, and in all cases the allele scored was the *01 allele.

Identification of second D gene segments in long N regions

The probabilities that N region addition would give rise to apparent D segments of varying lengths and identity were calculated, and are presented as Table III. These probabilities were used to calculate the degree of identity that was required, between long N regions and D segments, in order for a second D gene segment to

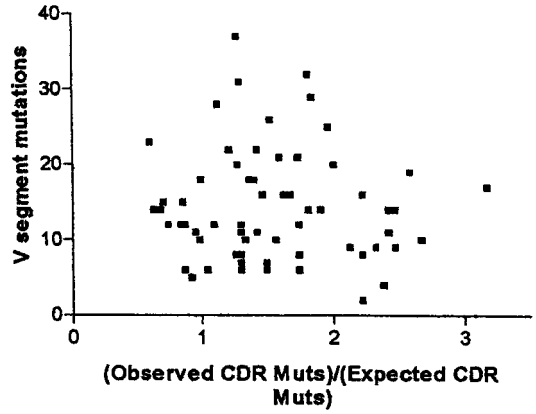

FIGURE 1. Determination of the Ag selection factor ratio (*R*). The ratio of observed/expected mutations in CDR1 and CDR2 regions of 69 highly mutated human H chain V segment sequences is plotted against total number of mutations in each V segment. *R* was calculated to be 1.54.

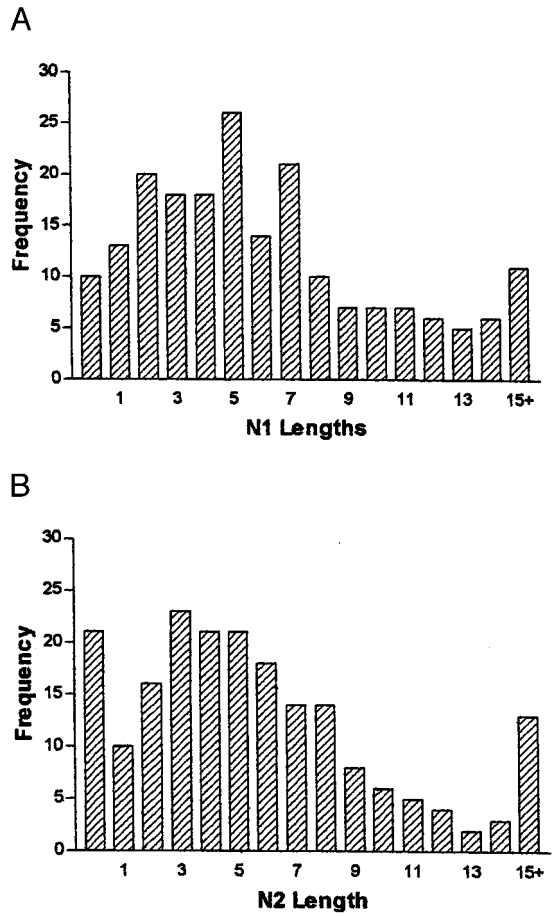

FIGURE 2. N nucleotide addition. The frequency distribution of the lengths of N1 (*a*) and N2 (*b*) regions, identified in 199 human IgM sequences.

Table III. *Probabilities that N regions of varying lengths will be identical with a germline D sequence or will mismatch by just 1 or 2 nt*

N Length	Mismatches		
	0	1	2
7	0.022	0.284	0.766
8	5.47×10^{-3}	0.103	0.505
9	1.32×10^{-3}	0.032	0.254
10	3.00×10^{-4}	8.90×10^{-3}	0.099
11	7.29×10^{-5}	2.45×10^{-3}	0.034
12	1.53×10^{-5}	5.91×10^{-3}	9.80×10^{-3}

be accepted as part of an alignment. The required matches for various N lengths are presented as Table IV, which shows, for example, that whereas 7 consecutive matches were required for identification of a D segment in an N region of 7 or 8 nt, 9 consecutive matches, or a single mismatch in 11 nt were required for any N region of ≥17 nt in length. The likelihood of alignments containing mismatches was assessed by reference to mutation analysis. Seventy-five sequences containing long putative N regions (length, >9 nt) were reanalyzed for the presence of additional D segments. Details of three sequences that appear to include second D segments are shown in Table V, including the probabilities that the degree of identity seen in the sequences could arise by random N nucleotide addition. Interestingly, the sequence X94053 includes two relatively long D segment alignments: 29 consecutive nucleotide matches to IGHD2-2*02, and 15 consecutive matches to IGHD5-12*01. An additional N2 sequence was found with 8 consecutive matches to a second D gene, but this D gene was located 3′ of the primary D gene segment, in the germline. For this reason, the alignment was not accepted as an example of D-D fusion. A number of alignments that just failed to meet the criteria for second D segments are shown as Table VI.

N1 sequences were also analyzed for identity with the 3′ ends of germline V genes, which is considered to be evidence of V gene replacement (21) (Table VII). One sequence was found with seven consecutive nucleotide matches, 3 with six matches, 7 with five matches, and 14 with four matches. Based upon the known nucleotide preferences of TdT, and the repertoire of V ends, alignments with six or more consecutive matches are unlikely to result from TdT activity, and more likely represent V gene replacement. Although many of the shorter alignments could be expected to arise by chance in a series of long N regions, the number of such alignments is noteworthy, as is the GA-rich nature of the sequences.

Analysis of the GC content of long N regions

The GC content and G or C homogeneity was determined for the 72 remaining long N sequences. Among the 40 N1 sequences, there were 12 sequences for which the probabilities that such sequences had arisen by conventional TdT activity were <0.05. Similarly, among the 32 N2 sequences, there were 9 sequences for which the p values were <0.05, 2 sequences with p values of <0.01, and 2 sequences with p values of <0.001.

Table IV. *Identity required between putative N regions of varying lengths, and germline D segments, for a second D gene segment alignment to be accepted*

N Length	Identity/Mismatches
7, 8	7/0
9	8/0, 9/1
10	8/0, 10/1
11	8/0, 10/1, 11/2
12–14	8/0, 10/1, 12/2
15–16	8/0, 11/1, 12/2
17–30	9/0, 11/1

The G/(G+C) proportions of the 72 sequences were plotted against N length for N1 (Fig. 3*a*) and N2 (*b*) sequences. As expected, the longer N1 sequences showed a tendency toward homogeneity for either G or C. This was not the case for N2 sequences, and the difference between the homogeneity of N1 and N2 sequences, among sequences of ≥15 nt in length, was statistically significant ($p < 0.05$).

To investigate possible concatenation of N nucleotides from both strands as an explanation of long N regions, the proportion of G among GC (G/G+C) for the 5′ and 3′ ends of the sequences were determined. Surprisingly, the mean proportion of 5′ G among N2 GC nucleotides was 0.36, and of 3′ G was 0.57. The corresponding values for N1 sequences were 0.56 and 0.57, respectively. The values seen for the N2 sequences represent a significant overrepresentation of 5′ C ($p < 0.001$) and of 3′ G ($p < 0.05$). Both ends of the N1 sequences were significantly enriched for G ($p < 0.05$). Nucleotide frequencies were therefore more fully considered, by an examination of dinucleotide frequencies.

The dinucleotide frequencies in the long N regions were analyzed for evidence of base stacking, and the results are presented as Table VIII. Among long N1 sequences, there was a significant overrepresentation of the homodimers CC ($p < 0.001$) and GG ($p < 0.05$); however, there was no general overrepresentation of purine or pyrimidine homodimers that would indicate base stacking. There was also a marked underrepresentation of GC heterodimers ($p < 0.01$). GA was also overrepresented ($p < 0.05$), which may be a consequence of inclusion of GAGA and GAGG motifs within the N1 sequences, as a result of V gene replacement. The GA dinucleotides were particularly overrepresented in the 5′ portions of the N1 sequences (data not shown). In the long N2 sequences, GG was also overrepresented ($p < 0.001$), and GC was underrepresented ($p < 0.001$). Although CC was overrepresented, this result did not reach significance.

Discussion

Despite considerable recent progress in our understanding of the molecular processes at the heart of the somatic hypermutation process (22, 23), and despite our long-standing and detailed understanding of the process of Ig gene rearrangement, the analysis of rearranged Ig genes has seen little advancement since these processes were first described. As alignment algorithms inevitably involve consideration of patterns of matching and mismatching nucleotides, between a sequence under consideration and a germline sequence, an understanding of the process of somatic point mutation should assist the alignment process.

An important advance in our understanding of somatic point mutation came with the report that the probability that the hypermutation process will introduce a mutation within a sequence decays exponentially, from 5′ to 3′, downstream of the promoter region (13). As a consequence of this finding, the likelihood of a mutation occurring in any sequence can be determined by the following four factors: 1) the position of the sequence within the rearranged gene; 2) the presence of mutational hot spots within the sequence, because the mutation mechanism preferentially targets particular nucleotide sequences; 3) selection processes that may favor mutations occurring in the Ag-binding CDRs, but resist mutations in the FRs; and 4) the extent to which the hypermutation process has acted upon the sequence.

Of these factors, only the fourth factor is difficult to estimate, but the observations of Rada and Milstein (13) now provides a means

Table V. *Probable examples of second D gene segment usage, identified among 75 putative N sequences of 10 nt or more*

Accession no.	V muts[a]	N(D2)N Length	N	D[b]	N	p[c]
AF174050	0	18	TCGATT	ACTACGGgGGT[d] (IGHD4-23*01)	C	0.02
X94053	0	22	TT	TATAGTGGCTACGAT	CCTGG	<0.00001
L12190	5	14	GACA	ACTACAATAT (IGHD2-2*01 INV)		<0.001

[a] Mutations in V gene segments, between codons 10 and 104.
[b] Nucleotide shown in lowercase represents a mismatch.
[c] Probability that such an alignment, in a sequence of that length, could arise through N nucleotide addition.
[d] Probability of a single mutation occurring in this particular sequence is 0.08.

of doing so, for the number of mutations in the long V gene segments of the H and L chain are easily determined, and can be determined without ambiguity. The extent of mutation in these sequences can serve as a predictor of the level of mutation throughout a rearranged gene.

It is clear that the mutability of a particular nucleotide is influenced by upstream and/or downstream nucleotides, and a number of studies have therefore reported the mutability of dinucleotides and trinucleotides (24, 25). Other studies have examined the influence of nucleotides as many as three positions upstream or downstream of the target nucleotide (26). These studies have led to the description of major hot spots at RGYW/WRCY and WAN motifs (27–29), where R is a purine, Y is a pyrimidine, W is A or T, N is any nucleotide, and the underlined nucleotide is preferentially but not absolutely targeted. In this study, we have focused on trinucleotide mutability, acknowledging the influence of nucleotides two positions upstream and two positions downstream of the target nucleotide. We have re-examined the issue of trinucleotide mutability in the light of the report of Rada and Milstein (13), because improved mutability scores should result when due regard is paid to the 5′ to 3′ distribution of nucleotide motifs within Ig sequences.

The mutability scores derived in this study can be most directly compared with two studies, one of murine sequences (24) and one of human sequences (30). The scores of all three studies are in broad agreement, but perhaps as a consequence of the positional analysis described in this study, important differences are also seen. A measure of the ability of the trinucleotide scores to describe the mutation process is the extent to which previously reported hot-spot motifs can be predicted from the trinucleotide scores. The high scores seen for the RGYW, WRCY, and WAN trinucleotides in this study suggests that the scores are appropriate and can be usefully applied to mutation analysis.

The effect of Ag selection upon the likely frequency of mutations within CDRs remains the most uncertain of the factors influencing analysis of mutations. During an ongoing immune response, replacement mutations within the Ag-binding CDRs may be selected during rounds of replication (31). We estimated the contribution of Ag selection by calculating the mean extent to which mutations accumulate above expectations in the CDR1 and CDR2 regions of V gene segments. This Ag selection factor (*R*) was estimated to be 1.54. This is in close agreement with a previous report of CDR mutations, where the enhancement of CDR mutations was calculated to be 1.58 (32). These relatively low figures likely reflect the 5′ engagement of the mutator mechanism (13), and the consequent tendency for mutations in the CDR to be accompanied by mutations in upstream FRs. Regardless, we believe the figure is a suitable one for use in mutation analysis. There

Table VI. *Examples of possible second D gene segment usage, identified among 75 putative N sequences of 10 nt or more*

Accession no.	V muts[a]	N(D2)N Length	5′ N	D[b]	3′ N
AJ244939[c]	5	15	ATC	AGTTATG (IGHD3-16*01)	GGGAC
AJ24499[c]	1	13	ACA	ATCTCTA (IGHD5-24*01 INV)	GAG
Z1832[c]	0	14	C	GTAACTC (IGHD4-23*01)	TAGAGC
AF17404[d]	0	17	ATC	TTTGGGGcA (IGHD3-16*01)	CGTCG
AJ24498[d]	4	16		TtTTATGA (IGHD3-16*01)	CCTACAGA
U97246[d]	0	18	T	TGGgTCGGGG (IGHD3-10*01)	TTCGAGG
AJ389191[d]	0	19	CT	TAGTAGCcCC (IGHD1-26*01 INV)	TCACATG
AJ244982[d]	4	17		TAGTAGgGG (IGHD6-19*01)	AACAGGAG
AY003828[d]	0	18		TGATAGaAG (IGHD3-22*01)	GCAAGGGGC

[a] Mutations in V gene segments, between codons 10 and 104.
[b] Nucleotide shown in lowercase represents a mismatch.
[c] Sequences that just fail the alignment criteria but that are AT-rich.
[d] Sequences that fail the alignment test because of a single mutation.

Table VII. *Possible examples of V gene replacement, identified among 72 long (*n > 9*) N1 sequences*

Accession no.	Expressed *IGHV* Gene	N1 Sequence[a]	IGHV Source of V End-Like Sequence
AY0038312	*IGHV3-30*18*	CACAAAAGACTT	IGHV3-43*01
AJ245008	*IGHV4-34*01*	CAAGATCGGA	IGHV1-45*01, *02
AY003774	*IGHV3-23*01*	ACCGAGATCCCGTG	IGHV4-34*10, IGHV4-59*10
L29154	*IGHV3-15*01*	CACAGAAGGACCAG	IGHV2-5[b]

[a] Nucleotides matching the 3′ end of V gene segments are underlined.
[b] No *IGHV* gene with this motif has been reported 3′ to IGHV3-15*01.

may be circumstances in which a greater weighting should be given to CDR mutations, and Ag selection; however, such a weighting would not have substantially changed the outcome of the analysis reported here.

In this study, we have applied mutation analysis to an investigation of N nucleotide addition in human H chain sequences. Our results support early observations that human Ig sequences include long N regions (6). This is in contrast to the situation in mice, where the mean length of N regions has been reported to be 3.0 (20). Interestingly, studies using transgenic human Ig minilocus mice appear to show a more murine pattern of N addition (33).

Perhaps as a result of the different approach that we have taken to the determination of the ends of gene segments, the results of our study also challenge the high frequency of sequences that have been reported to lack N addition in adult rearrangements. Although ~30% of rearrangements in transgenic human minilocus mice appear to lack N addition (33), in our study, only 7.8% of rearrangements lacked N addition. Even this figure is likely to be an overestimate, for ~25% of rearrangements in which a single N nucleotide is added will appear to lack N addition. Inevitably, the added nucleotide will often return a germline nucleotide that had been removed by exonuclease activity. Some systematic bias in the analysis, favoring low levels of N addition, is therefore unavoidable. Only 2 of the 33 junctions that lacked N addition were fetal sequences. In contrast, many of the sequences appear to be derived from unusual B cell populations, including 14 sequences derived from subepithelial tonsillar B lymphocytes (34). Therefore, it may be that the lack of N addition in conventional B cell populations is even rarer than indicated in this study.

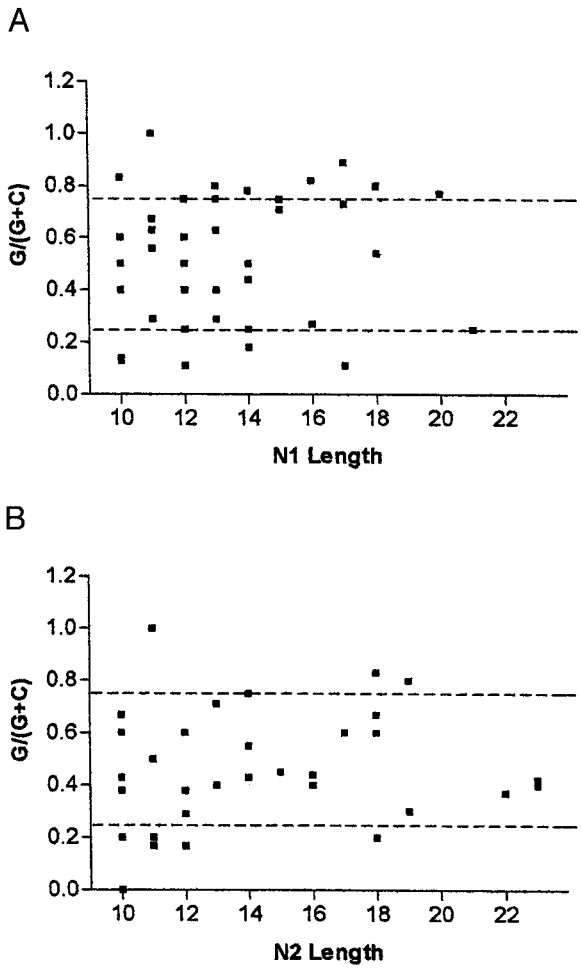

FIGURE 3. G/C homogeneity index. Homogeneity of G/C content of long N1 (*a*) and long N2 (*b*) regions identified in human IgM sequences, as shown by the proportion of G among G and C nucleotides in the sequences. Points outside the central area bounded by the dotted lines, showing proportions of 0.25 and 0.75, can be considered relatively homogenous for G or C.

Long putative N regions could arise from the presence of a second D gene segment in a rearranged V gene, but many have queried the very existence of D-D fusions since the process was first proposed (35). As part of the study that first fully described the human D segment locus, randomly generated mock sequences were aligned with D segments and used to define an alignment algorithm (10). Only 4 of 821 sequences that were then aligned met the 99% confidence limit for D-D fusion. As more than 4 sequences in 821 could be expected by chance to meet the 99% confidence limit, this was taken as evidence against the existence of D-D fusion. Nevertheless, studies have continued to report D-D fusions (33), and such fusions have sometimes been reported to be exceedingly common (11).

A similar strategy was used to investigate the use of inverted D gene segments (10), and the identification of just 7 sequences in 127 that met the 99% confidence limit was considered to provide little evidence for their use. Despite this study, the use of inverted D segments also continues to be reported.

In the present study, the predictability of mutations within CDR3 regions has allowed us to apply strict criteria to the investigation of D-D fusions, including D-D fusions involving inverted D gene segments, by calculating the probabilities that random N nucleotide addition would give rise to such sequences. It has been recognized for many years that D segments are G-C rich, and that apparent alignments could therefore arise from random N nucleotide addition (6). Perhaps as a consequence of the uncertainty surrounding point mutations within the CDR3, no systematic analysis has been developed before to resolve this issue.

This study clearly demonstrates the reality of D-D fusion in the expressed repertoire, with the identification of 3 sequences among 75 that met the 95% confidence limits. In 75 sequences, a small number of relatively short alignments could still be expected to arise by chance. One of the 3 sequences (AF174050) could perhaps be attributed to random N nucleotide addition. In fact, AF174050 can be aligned to a single D gene segment (IGHD3–22*01), if three central, consecutive mismatches arose from point mutation. We consider such a pattern of mutation to be improbable ($p <$

Table VIII. *Observed and expected dinucleotide frequencies in long (*n > 9*) N1 and N2 regions*

	N1				N2			
	Observed	Expected[a]	Observed/Expected	*p*	Observed	Expected	Observed/Expected	*p*
AA	19	23.5	0.8	NS	23	19.6	1.2	NS
AC	30	31.3	1.0	NS	34	28.5	1.2	NS
AG	37	35.7	1.0	NS	24	26.8	0.9	NS
AT	24	18.7	1.3	NS	14	19.0	0.7	NS
CA	24	31.3	0.8	NS	19	28.5	0.7	NS
CC	64	41.6	1.5	<0.001	51	41.5	1.2	NS
CG	35	47.4	0.7	NS	35	39.0	0.9	NS
CT	24	24.9	1.0	NS	36	27.7	1.3	NS
GA	49	35.7	1.4	<0.05	27	26.8	1.0	NS
GC	28	47.4	0.6	<0.01	19	39.0	0.5	<0.001
GG	70	54.1	1.3	<0.05	60	36.6	1.6	<0.001
GT	17	28.4	0.6	NS	17	26.0	0.7	NS
TA	20	18.7	1.1	NS	24	19.0	1.3	NS
TC	26	24.9	1.0	NS	26	27.7	0.9	NS
TG	25	28.4	0.9	NS	17	26.0	0.7	NS
TT	15	14.9	1.0	NS	24	18.4	1.3	NS

[a] Expected frequencies were calculated using probabilities based upon the actual frequency of each nucleotide in the N1 and N2 regions.

0.002). The sequence could alternatively be an example of an insertion/deletion event of the kind that has been reported in CDR1 and CDR2 regions (7, 36). Such insertion/deletion events have been associated with repetitive sequences (37), and such repetitive sequences are to be found on either side of the trinucleotide mismatch. However insertion/deletion events have been linked to the hypermutation process (36), yet there were no mutations in the V segment of this sequence.

Two alignments remain as unequivocal evidence of D-D fusions, including one alignment with an inverted *IGHD2-2* allele. It is possible that mutations within the second D segments, or unreported D segment polymorphisms could have prevented the identification of additional D-D sequences. Six sequences failed to meet the alignment criteria because of a single mismatch. Given the low level of mutation seen in the six associated V gene segments, it is most unlikely that more than one of these sequences represents a mutated D segment. Three other sequences were 1 nt short of the acceptance criteria, but the sequences were strikingly AT rich. Therefore, these sequences may be additional examples of D-D fusion. We therefore consider the number of examples of D-D fusion in the 398 V_HD and DJ_H junction sequences examined to be between 2 (0.5%) and 7 (1.8%). Therefore, D-D fusion is a rare event in the human.

Many long putative N regions with unexpected features remain unexplained in this study. As probabilities were calculated on the basis that strand concatenation does not occur during N nucleotide addition, we investigated the possibility that in fact strand concatenation is responsible for long N regions. If this is the case, the 5′ halves of long N regions should be rich in G and 3′ halves should be rich in C. Surprisingly, among long N2 regions, the opposite situation was seen. This anti-concatenation has been observed previously (20), although, in that study, the anti-concatenation did not reach statistical significance. We are unable to propose a model of N addition that can satisfactorily explain such observations.

Analysis of homodimer frequencies in N1 and N2 regions showed no significant and general overrepresentation of RR and YY dinucleotides, as has been previously reported (38). Therefore, this study does not support a role for base stacking in N nucleotide addition. However, a highly significant overrepresentation of GG and CC was seen, with an underrepresentation of GC dimers. GA heterodimers were also overrepresented in the N1 regions.

A number of long N1 regions are likely to have arisen by V gene replacement. Six of 40 long N1 regions included six or more consecutive matches with the 3′ ends of V genes. Although these alignments are short, and could reflect random TdT activity, the location of these motifs at the 5′ end of long N1 regions is striking, and such sequences were not seen at the 3′ end of N1 sequences. Only two such motifs were identified in the long N2 sequences, and both of these were located 3′ in the sequence. As many other long N1 sequences include 4- and 5-nt matches, it may be that >6 of the long N1 regions are the result of V gene replacement. Nevertheless, most long N1 regions lack the motifs, and the origin of these long sequences and the long N2 sequences cannot presently be explained.

Patterns in short nucleotide sequences can provide vital clues to processes that may contribute to the generation of Ab diversity. We have described the development of mutation analysis, which allows us to more reliably partition rearranged Ig genes. Mutation analysis also provides a guide for the interpretation of sequences between the rearranged V_H, D, and J_H segments. As a consequence, we have been able to develop objective criteria for the acceptance or rejection of putative alignments to multiple D gene segments within a V_HDJ_H rearrangement. Together, these techniques have allowed us to identify N nucleotides with greater certainty, and to develop a reliable dataset for the study of the human CDR3. We have been able to confirm the reality of D-D fusion, and we have highlighted features of the human CDR3 sequences that remain unexplained. Further investigation of patterns of nucleotides within the human CDR3 region are likely to uncover additional processes that contribute to repertoire diversity, and that are responsible for these molecular signatures, provided that such investigations are conducted using the kinds of objective partitioning algorithms that are described in this work.

Acknowledgments

We thank Dr. Dan Conrad (Medical College of Virginia, Richmond, VA) and Dr. William Sewell (Garvan Institute of Medical Research, Sydney, Australia) for their assistance with this work and for their comments on the manuscript.

References

1. Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature 302:575.*
2. Kabat, E. A., T. T. Wu, H. M. Perry, S. K. Gottesman, and C. Foeller. 1991. *Sequences of Proteins of Immunological Interest.* U.S. Department of Health and Human Services, Bethesda, MD.
3. Alt, F. W., and D. Baltimore. 1982. Joining of immunoglobulin heavy chain segments: implications from a chromosome with evidence of three D-JH fusions. *Proc. Natl. Acad. Sci. USA 79:4118.*

4. Lafaille, J. J., A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa. 1989. Junctional sequences of T cell receptor *gd* genes: implications for γδ T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell 59:859*.
5. Kim, S., M. M. Davis, E. Sinn, P. Patten, and L. Hood. 1981. Somatic hypermutation of rearranged VH genes. *Cell 27:573*.
6. Sanz, I. 1991. Multiple mechanisms participate in the generation of diversity of human H chain CDR3 regions. *J. Immunol. 147:1720*.
7. de Wildt, R. M. T., W. J. van Venrooij, G. Winter, R. M. Hoet, and I. M. Tomlinson. 1999. Somatic insertions and deletions shape the human antibody repertoire. *J. Mol. Biol. 294:701*.
8. Ichihara, Y., H. Matsuoka, and Y. Kurosawa. 1988. Organization of human immunoglobulin heavy chain diversity gene loci. *EMBO J. 13:4141*.
9. Meek, K. D., C. A. Hasemann, and J. D. Capra. 1989. Novel rearrangement at the immunoglobulin D locus. *J. Exp. Med. 170:39*.
10. Corbett, S. J., I. M. Tomlinson, E. L. 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*.
11. Dunn-Walters, D. K., M. Hackett, L. Boursier, P. J. Ciclitira, P. Morgan, S. J. Challacombe, and J. Spencer. 2000. Characteristics of human IgA and IgM genes used by plasma cells in the salivary gland resemble those used in duodenum but not those used in the spleen. *J. Immunol. 164:1595*.
12. 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*.
13. Rada, C., and C. Milstein. 2001. The intrinsic hypermutability of antibody heavy and light chain genes decays exponentially. *EMBO J. 20:4570*.
14. Jolly, C. J., S. D. Wagner, C. Rada, N. Klix, C. Milstein, and M. S. Neuberger. 1996. The targeting of somatic hypermutation. *Semin. Immunol. 8:159*.
15. Basu, M., M. V. Hegde, and M. J. Modak. 1983. Synthesis of compositionally unique DNA by terminal deoxynucleotidyl transferase. *Biochem. Biophys. Res. Commun. 111:1105*.
16. Lefranc, M.-P., V. Giudicelli, C. Ginestoux, J. Bodmer, W. Muller, R. Bontrop, M. Lemaitre, A. Malik, V. Barbie, and D. Chaume. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res. 27:209*.
17. Lefranc, M. P. 2001. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res. 29:207*.
18. Snow, R. E., C. J. Chapman, A. J. Frew, S. T. Holgate, and F. K. Stevenson. 1997. Pattern of usage and somatic hypermutation in the V_H5 gene segments of a patient with asthma: implications for IgE. *Eur. J. Immunol. 27:162*.
19. Snow, R. E., R. Djukanovic, and F. K. Stevenson. 1999. Analysis of immunoglobulin E VH transcripts in a bronchial biopsy of an asthmatic patient confirms bias towards VH5, and indicates local clonal expansion, somatic mutation, and isotype switch events. *Immunology 98:646*.
20. Kepler, T. B., M. Borrero, B. Rugerio, S. K. McCray, and S. H. Clarke. 1996. Interdependence of N nucleotide addition and recombination site choice in V(D)J rearrangement. *J. Immunol. 157:4451*.
21. Wang, Y.-H., Z. Zhang, P. D. Burrows, H. Kubagawa, S. L. J. Bridges, H. W. Findley, and M. D. Cooper. 2003. V(D)J recombinatorial repertoire diversification during intraclonal pro-B to B-cell differentiation. *Blood 101:1030*.
22. Petersen-Mahrt, S. K., R. S. Harris, and M. S. Neuberger. 2002. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature 418:99*.
23. Martin, A., P. D. Bardwell, C. J. Woo, M. Fan, M. J. Shulman, and M. D. Scharff. 2002. Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas. *Nature 415:802*.
24. Smith, D. S., G. Creadon, P. K. Jena, J. P. Portanova, B. L. Kotzin, and L. J. Wysocki. 1996. Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *J. Immunol. 156:2642*.
25. Zan, H., A. Cerutti, P. Dramitinos, A. Schaffer, Z. D. Li, and P. Casali. 1999. Induction of Ig somatic hypermutation and class switching in a human monoclonal IgM^+IgD^+ B cell line in vitro: definition of the requirements and modalities of hypermutation. *J. Immunol. 162:3437*.
26. Dunn-Walters, D. K., A. Dogan, L. Boursier, C. M. MacDonald, and J. Spencer. 1998. Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *J. Immunol. 160:2360*.
27. Rogozin, I. B., and N. A. Kolchanov. 1992. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta 1171:11*.
28. Dorner, T., S. J. Foster, N. L. Farner, and P. E. Lipsky. 1998. Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. *Eur. J. Immunol. 28:3384*.
29. Diaz, M., and P. Casali. 2002. Somatic immunoglobulin hypermutation. *Curr. Opin. Immunol. 14:235*.
30. Shapiro, G. S., K. Aviszus, D. Ikle, and L. J. Wysocki. 1999. Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition. *J. Immunol. 163:259*.
31. Chang, B., and P. Casali. 1994. The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement. *Immunol. Today 15:367*.
32. Cowell, L. G., H. J. Kim, T. Humaljoki, C. Berek, and T. B. Kepler. 1999. Enhanced evolvability in immunoglobulin V genes under somatic hypermutation. *J. Mol. Evol. 49:23*.
33. Tuaillion, N., and J. D. Capra. 2000. Evidence that terminal deoxynucleotidyltransferase expression plays a role in Ig heavy chain gene segment utilisation. *J. Immunol. 164:6387*.
34. Dono, M., S. Zupo, N. Leanza, G. Melioli, M. Fogli, A. Melagrana, N. Chiorazzi, and M. Ferrarini. 2000. Heterogeneity of tonsillar subepithelial B lymphocytes, the splenic marginal zone equivalents. *J. Immunol. 164:5596*.
35. Siebenlist, U., J. V. Ravetch, S. Korsmeyer, T. Waldman, and P. Leder. 1981. Human immunoglobulin D segments encoded in tandem multigenic families. *Nature 294:631*.
36. Wilson, P. C., O. de Bouteiller, Y. J. Liu, K. Potter, J. Banchereau, J. D. Capra, and V. Pascual. 1998. Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J. Exp. Med. 187:59*.
37. Lantto, J., and M. Ohlin. 2002. Uneven distribution of repetitive trinucleotide motifs in human immunoglobulin heavy variable genes. *J. Mol. Evol. 54:346*.
38. Gauss, G. H., and M. R. Lieber. 1996. Mechanistic constraints on diversity in human V(D)J recombination. *Mol. Cell. Biol. 16:258*.