J. Mol. Biol. (1996) **256**, 813–817

JMB

COMMUNICATION

The Imprint of Somatic Hypermutation on the Repertoire of Human Germline V Genes

Ian M. Tomlinson[1], Gerald Walter[1,2], Peter T. Jones[1,3], Paul H. Dear[3]
Erik L. L. Sonnhammer[4] and Greg Winter[1,3]*

[1]*MRC Centre for Protein Engineering, Hills Road Cambridge CB2 2QH, UK*

[2]*Cambridge Antibody Technology Ltd. The Science Park, Melbourn Cambs SG8 6JJ, UK*

[3]*MRC Laboratory of Molecular Biology Hills Road, Cambridge CB2 2QH, UK*

[4]*Sanger Centre, Hinxton Hall Hinxton, Cambridge CB10 1RQ, UK*

**Corresponding author*

In the human immune system, antibodies with high affinities for antigen are created in two stages. A diverse primary repertoire of antibody structures is produced by the combinatorial rearrangement of germline V gene segments and antibodies are selected from this repertoire by binding to the antigen. Their affinities are then improved by somatic hypermutation and further rounds of selection. We have dissected the sequence diversity created at each stage in response to a wide range of antigens. In the primary repertoire, diversity is focused at the centre of the binding site. With somatic hypermutation, diversity spreads to regions at the periphery of the binding site that are highly conserved in the primary repertoire. We propose that evolution has favoured this complementarity as an efficient strategy for searching sequence space and that the germline V gene families evolved to exploit the diversity created by somatic hypermutation.

Keywords: antibody; diversity; evolution; germline; hypermutation

Most of the sequence diversity in antibodies is located in the complementarity-determining regions, or CDRs (Wu & Kabat, 1970; Kabat & Wu, 1971). This diversity comprises germline diversity (the choice of different variable (V), diversity (D) and joining (J) segments) and junctional diversity in the primary repertoire (Tonegawa, 1983), and somatic hypermutation introduced during the process of affinity maturation (Berek & Milstein, 1987). It has been noted that, in mice, the germline V gene segments involved in a restricted response to two haptens differ most in the CDR2 regions, whereas somatic hypermutation is biased towards ‘‘hot-spots’’ in CDR1 (reviewed by Neuberger & Milstein, 1995). We wondered whether such complementarity was a general feature of antibody repertoires.

Since all the human germline V_H and V_κ segments have now been mapped (Matsuda *et al*., 1993; Zachau, 1993; Cook *et al*., 1994; Nagoaka *et al*., 1994; Tomlinson *et al*., 1994) and sequenced (Tomlinson *et al*., 1992; Schäble & Zachau, 1993; Cox *et al*., 1994), we decided to analyse the entire human V_H and V_κ repertoires and thereby the response to a wide range of antigens. We compiled a database of 1181 rearranged V_H and 736 rearranged V_κ sequences (Figure 1), and identified the location of somatic mutations in each sequence (with the exception of the V_H CDR3 and the end of the V_κ CDR3, where somatic hypermutation cannot be distinguished from junctional diversity). In general, the number of amino acid differences introduced by somatic hypermutation is less than the number of amino acid differences between germline V gene segments (Figure 1). In addition, we note that somatic hypermutation is primarily a point mutation process and rarely results in codon insertions or deletions, whereas the CDR lengths do differ between germline V gene segments (Chothia *et al*., 1992; Tomlinson *et al*., 1995).

To compare the sequence diversity introduced by somatic hypermutation with that present in the same repertoire of germline sequences, we compiled a second database by substituting each rearranged sequence in the first database with its corresponding germline V gene segment. Although this does not strictly correspond to the primary repertoire, since the rearranged genes have been selected by antigen, it can be used to calculate the germline diversity prior to somatic hypermutation.

Present address: G. Walter, Max-Planck-Institut für Molekulare Genetik, Ihnestr. 73, D-14195 Berlin-Dahlem, Germany.

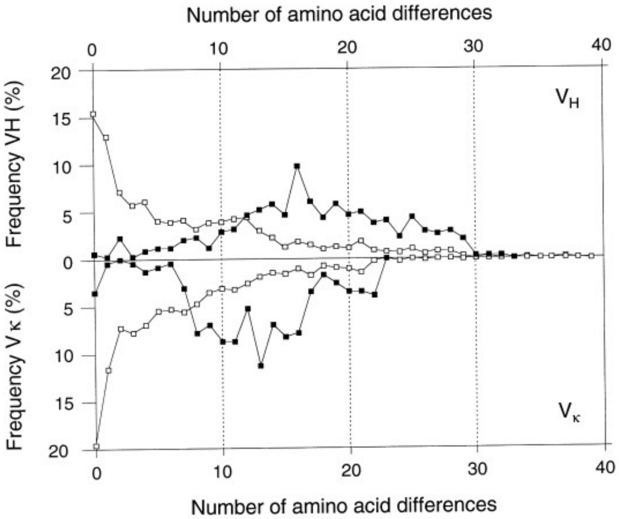

Figure 1. The contributions of germline diversity and somatic hypermutation. For germline diversity (filled squares), the number of amino acid differences between all pairs of sequences within the germline V_H families and V_κ subgroups were calculated using the 51 functional V_H (Tomlinson *et al*., 1992; Cook *et al*., 1994) and 40 functional V_κ segments (Schäble & Zachau, 1993; Tomlinson *et al*., 1995), with insertions or deletions counting as one amino acid difference (average for V_H is 17.0; average for V_κ is 12.6). (Members of distinct families/subgroups differ by at least 24 amino acids.) For somatic hypermutation (open squares), the frequencies of the number of somatic mutations per rearranged sequence were compiled from 1181 rearranged V_H and 736 rearranged V_κ genes (average for V_H is 7.3; average for V_κ is 5.8). This includes 143 rearranged V_H genes from the peripheral lymphocytes of a single individual (DP). These were either C_μ or C_γ linked, with averages of 3.7 and 10.2 mutations per rearranged sequence, respectively.

Compilation of the rearranged database was performed as follows: 46 C_μ and 97 C_γ linked V_H sequences (EMBL data library accession numbers Z68345-Z68487) were amplified, cloned and sequenced from cDNA of the donor DP essentially as described by Marks *et al*. (1991) (primer sequences available on request). Sequences were assigned to their germline counterparts in the V BASE sequence directory (available from the authors) using the MacVector sequence analysis package (IBI Kodak). In addition, 983 rearranged V_H and 670 rearranged V_κ sequences were extracted from the EMBL data library (Release 39) using the BLAST algorithm (Altschul *et al*., 1990) and automatically aligned to their germline counterparts in the V BASE directory (E.L.L.S., unpublished). A further 55 rearranged V_H and 66 rearranged V_κ sequences were entered from the literature and assigned germline counterparts using MacVector (IBI Kodak). All rearranged sequences from the EMBL data library and the literature were rearranged *in vivo*. Somatic mutations were scored by comparing each rearranged gene with its germline counterpart. Statistical analyses were performed using Apple Macintosh-based software (P.H.D., unpublished).

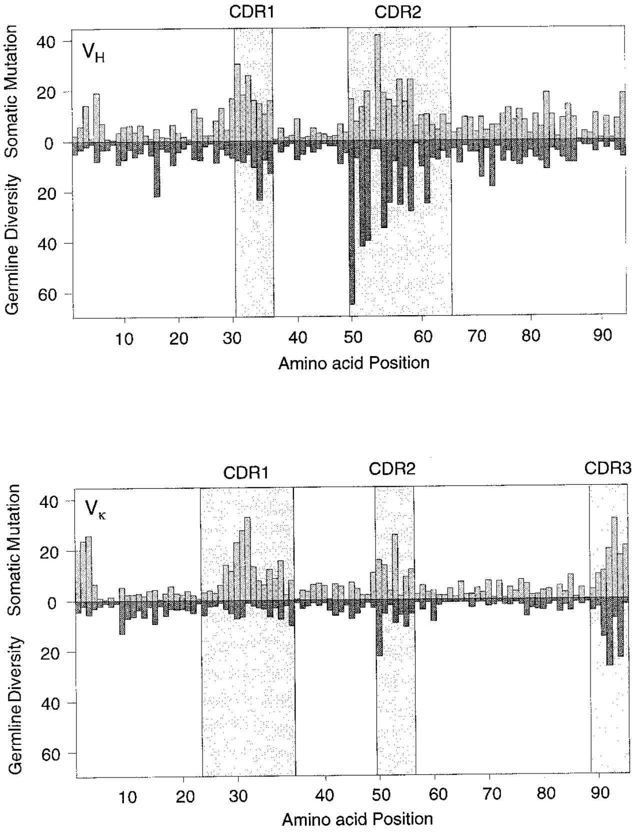

Figure 2. The patterns of germline diversity and somatic hypermutation. For each position in the V_H and V_κ regions, percentage somatic hypermutation (top of each panel in lighter tint) was calculated as the total number of differences between the rearranged sequences and their corresponding V gene segments, divided by the number of residues at that position. Mismatched PCR primers are probably responsible for some of the differences in the N-terminal residues. From the database of rearranged sequences a corresponding database of the germline counterparts was produced (see the text). This germline database was used to calculate the sequence diversity before hypermutation (bottom of each panel in darker tint) using the Kabat variability index (Wu & Kabat, 1970: variability equals the total number of different residues at each position divided by the relative frequency of the most common residue at that position). Sequence ambiguities and stop codons were excluded from the calculations. Sequence alignments, numbering and CDRs are according to Kabat *et al*. (1991), except for the CDR1 of the V_H and V_κ regions, where alignments and numbering are according to Chothia *et al*. (1992) or Tomlinson *et al*. (1995), respectively.

The patterns of sequence diversity are summarised in Figure 2. Germline diversity is greatest in the V_H CDR2 (see Tomlinson *et al*., 1992) and in specific residues in the other CDRs, for example at residues H33, L50, L91, L92 and L94. Somatic hypermutation, though more evenly distributed between the CDRs, is particularly prominent at residues H31, H31b, H52c, H56, H58, L30, L31, L31a, L53 and L93. Whilst some of these biases may be due to the presence of intrinsic hot-spots for mutation (Betz *et al*., 1993; Wagner *et al*., 1995), the pattern and conservative nature of the changes suggests that the overwhelming factor is that of selection.

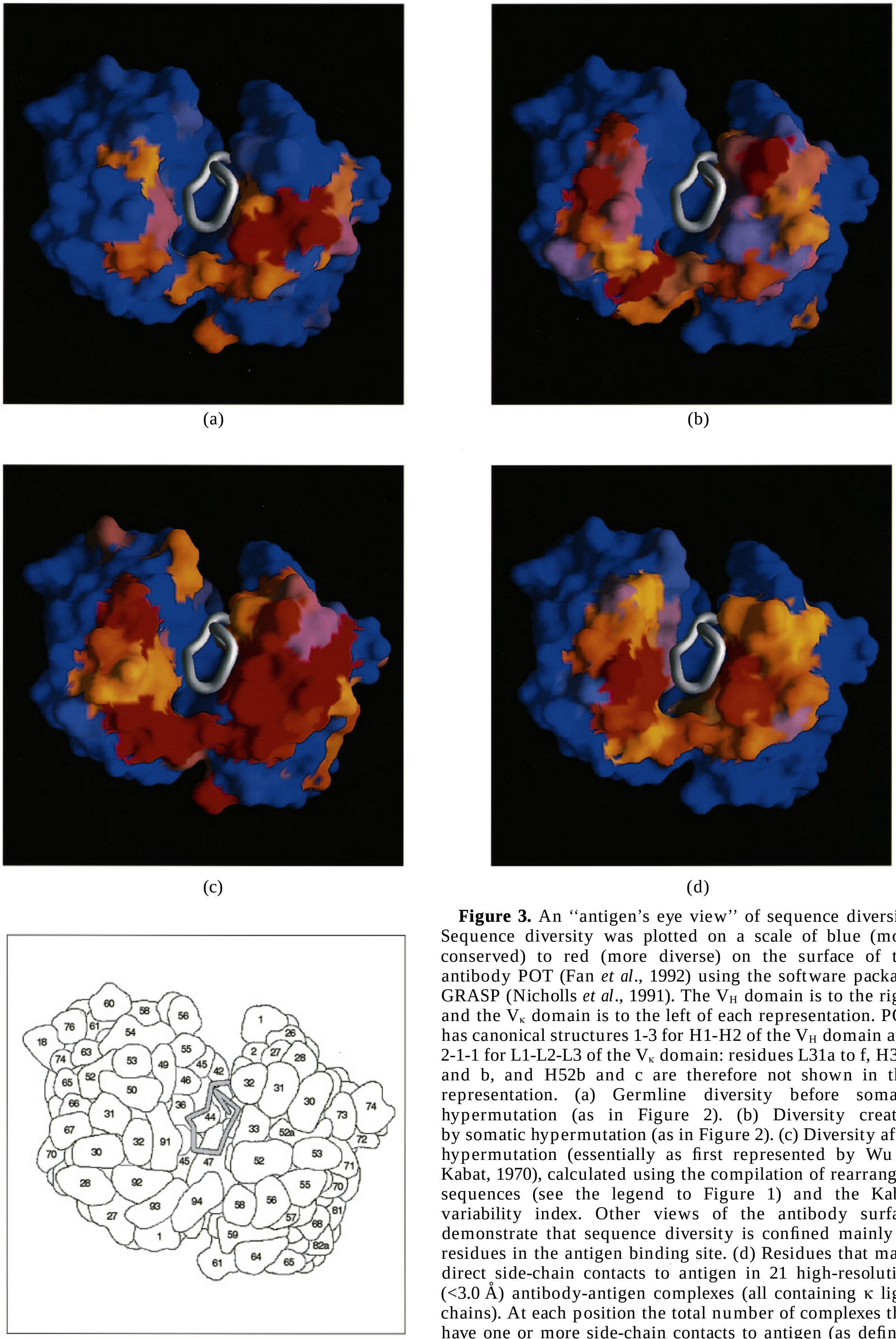

Figure 3. An ''antigen's eye view'' of sequence diversity. Sequence diversity was plotted on a scale of blue (more conserved) to red (more diverse) on the surface of the antibody POT (Fan *et al*., 1992) using the software package GRASP (Nicholls *et al*., 1991). The V_H domain is to the right and the V_κ domain is to the left of each representation. POT has canonical structures 1-3 for H1-H2 of the V_H domain and 2-1-1 for L1-L2-L3 of the V_κ domain: residues L31a to f, H31a and b, and H52b and c are therefore not shown in this representation. (a) Germline diversity before somatic hypermutation (as in Figure 2). (b) Diversity created by somatic hypermutation (as in Figure 2). (c) Diversity after hypermutation (essentially as first represented by Wu & Kabat, 1970), calculated using the compilation of rearranged sequences (see the legend to Figure 1) and the Kabat variability index. Other views of the antibody surface demonstrate that sequence diversity is confined mainly to residues in the antigen binding site. (d) Residues that make direct side-chain contacts to antigen in 21 high-resolution (<3.0 Å) antibody-antigen complexes (all containing κ light chains). At each position the total number of complexes that have one or more side-chain contacts to antigen (as defined in the original papers, see below) is plotted on a blue to red scale (blue, no structures with contacts; red, maximum of 16). Protein data bank (Bernstein *et al*., 1977) references: 1mlc, 1jhl, 1nca, 1jel, 1tet, 1ikf, 1ggi, 2igf, 1him, 1ibg, 1igj, 4fab, 1dbb, 2cgr, 1baf, 1cbv and 1mrb; other references: Tormo *et al*. (1994); Fields *et al*. (1995); Jeffrey *et al*. (1995). (e) Key for residue numbers (as in Figure 2). A photocopied transparency of the key can be used to overlay (a) to (d). The V_H CDR3 loop is shown in grey. The end of the V_κ CDR3 (also excluded from this analysis) lies at the centre of the binding site and is not visible in this representation. Residues H35, H50 and L34 are buried at the centre of the binding site.

We then used the structure of the human antibody POT (Fan *et al*., 1992) to provide an ''antigen's eye view'' of this diversity. Before hypermutation, sequence diversity is focused towards the centre of the antigen binding site (Figure 3(a)). The V_H CDR3 and the end of the V_κ CDR3 (excluded from this analysis, see above) also are diverse in sequence and length in the primary repertoire and lie at the centre of the antigen binding site (Figure 3). Somatic hypermutation spreads diversity to regions at the periphery of the binding site that are conserved in the primary repertoire (Figure 3(b)); for example, residues H30, H31, H32, L30, L31, L32, L53 and L93. As a result, the pattern of sequence diversity of the mature repertoire is more evenly distributed across the antigen binding site (Figure 3(c)).

In general, the most diverse residues are those that make the most contacts, emphasising their importance in the antibody-antigen interaction (Figure 3(d)). However, some conserved residues do make contacts (for example tryptophan H47 and tyrosine L49) and some of the more diverse residues do not (for example, H61, H71, H73 and L60), suggesting that they may play an indirect role in antigen binding.

The pattern and extent of diversity introduced by somatic hypermutation is therefore complementary to that in the primary repertoire, antibodies using two intermeshing regions of the binding site to make and/or improve binding contacts. This poses the question of how such complementarity evolved.

Both human and mouse antibodies undergo somatic hypermutation, and the organisation of their V gene loci indicates that the duplication and diversification of their V gene families occurred independently (Brodeur & Riblet, 1984; Kofler *et al*., 1989; Matsuda *et al*., 1993; Zachau, 1993; Cook *et al*., 1994). Hence, it appears that the mechanism of somatic hypermutation predates family diversification and could have played a role in the evolution of the human V genes. Indeed, it has been proposed that diversity in the germline V gene segments arose by homologous recombination with somatically mutated rearranged V genes (Rothenfluh *et al*., 1994). Our findings indicate that the mechanism cannot be direct: the patterns of germline and somatic diversity are complementary and the number of differences between the V gene segments is larger than the number of changes created by somatic hypermutation.

Instead, we propose that evolution has favoured complementarity as an efficient strategy for searching sequence space (Kauffman, 1993). Some sites have undergone germline diversification, whilst others (including hot-spots, Betz *et al*., 1993; Wagner *et al*., 1995) have been conserved for alteration by somatic hypermutation. In this way somatic hypermutation has left an evolutionary imprint on the sequences of the human V gene segments.

Acknowledgements

We thank Jonathan Cox for help constructing the rearranged database of V_κ genes and Graham Cook for helpful comments on the manuscript.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Berek, C. & Milstein, C. (1987). Mutation drift and repertoire shift in the maturation of the immune response. *Immunol. Rev.* **96**, 23–41.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Betz, A. G., Rada, C., Pannell, R., Milstein, C. & Neuberger, M. S. (1993). Passenger transgenes reveal intrinsic specificity of the antibody hypermutation mechansim: clustering, polarity, and specific hotspots. *Proc. Natl Acad. Sci. USA*, **90**, 2385–2388.

Brodeur, P. H. & Riblet, R. (1984). The immunoglobulin heavy-chain variable region (Igh-V) locus in the mouse. I. One hundred Igh-V genes comprise seven families of homologous genes. *Eur. J. Immunol.* **14**, 922–930.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). Structural repertoire of the human V_H segments. *J. Mol. Biol.* **227**, 799–817.

Cook, G. P., Tomlinson, I. M., Walter, G., Riethman, H., Carter, N. P., Buluwela, L., Winter, G. & Rabbitts, T. H. (1994). A map of the human immunoglobulin V_H locus completed by analysis of the telomeric region of chromosome 14q. *Nature Genet.* **7**, 162–168.

Cox, J. P. L., Tomlinson, I. M. & Winter, G. (1994). A directory of human germ-line V_κ segments reveals a strong bias in their usage. *Eur. J. Immunol.* **24**, 827–836.

Fan, Z.-C., Shan, L., Guddat, L. W., He, X.-M., Gray, W. R., Raison, R. L. & Edmondson, A. B. (1992). Three-dimensional structure of an Fv from a human IgM immunoglobulin. *J. Mol. Biol.* **228**, 188–207.

Fields, B. A., Goldbaum, F. A., Ysern, X., Poljak, R. J. & Mariuzza, R. A. (1995). Molecular basis of antigen mimicry by an anti-idiotope. *Nature*, **374**, 739–742.

Jeffrey, P. D., Bajorath, J., Chanf, C. Y., Yelton, D., Hellström, I., Hellström, K. E. & Sheriff, S. (1995). The X-ray structure of an anti-tumour antibody in complex with antigen. *Nature Struct. Biol.* **2**, 466–471.

Kabat, E. A. & Wu, T. T. (1971). Attempts to locate complementarity-determining residues in the variable positions of light and heavy-chains. *Ann. NY Acad. Sci.* **190**, 382–393.

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immunological Interest*, US Department of Health and Human Services, Bethesda, MD.

Kauffman, S. A. (1993). *The Origins of Order*, Oxford University Press, Oxford.

Kofler, R., Duchosal, M. A. & Dixon, F. J. (1989). Complexity, polymorphism and connectivity of mouse V_κ gene families. *Immunogenetics*, **29**, 65–74.

Marks, J. D., Tristrem, M., Karpas, A. & Winter, G. (1991). Oligonucleotide primers for polymerase chain reaction amplification of human immunoglobulin variable genes and design of family-specific oligonucleotide probes. *Eur. J. Immunol.* **21**, 985–991.

Matsuda, F., Shin, E. K., Nagaoka, H., Matsumara, R., Haino, M., Fukita, Y., Taka-ishi, S., Imai, T., Riley, J. H., Anand, R., Soeda, E. & Honjo, T. (1993). Structure and physical map of 64 variable segments in the 3' 0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nature Genet.* **3**, 88–94.

Nagaoka, H., Ozawa, K., Matsuda, F., Hayashida, H., Matsumura, R., Haino, M., Shin, E. K., Fukita, Y., Imai, T., Anand, R., Yokoyama, K., Eki, T., Soeda, E. & Honjo, T. (1994). Recent translocation of variable and diversity segments of the human immunoglobulin heavy-chain from chromosome 14 to chromosomes 15 and 16. *Genomics*, **22**, 189–197.

Neuberger, M. S. & Milstein, C. (1995). Somatic hypermutation. *Curr. Opin. Immunol.* **7**, 248–254.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281–296.

Rothenfluh, H. S., Gibbs, A. J., Blanden, R. V. & Steele, E. J. (1994). Analysis of the patterns of DNA sequence variation in flanking and coding regions of murine germ-line immunoglobulin heavy-chain variable genes: evolutionary implications. *Proc. Natl Acad. Sci. USA*, **91**, 12163–12167.

Schäble, K. F. & Zachau, H. G. (1993). The variable genes of the human immunoglobulin κ locus. *Biol. Chem. Hoppe-Seyler*, **374**, 1001–1022.

Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline V_H segments reveals about fifty groups of V_H segments with different hypervariable loops. *J. Mol. Biol.* **227**, 776–798.

Tomlinson, I. M., Cook, G. P., Carter, N. P., Elaswarapu, R., Smith, S., Walter, G., Buluwela, L., Rabbitts, T. H. & Winter, G. (1994). Human immunoglobulin V_H and D segments on chromosomes 15q11.2 and 16p11.2. *Hum. Mol. Genet.* **3**, 853–860.

Tomlinson, I. M., Cox, J. P. L., Gherardi, E., Lesk, A. M. & Chothia, C. (1995). The structural repertoire of the human V_κ domain. *EMBO J.* **14**, 4628–4638.

Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, **302**, 575–581.

Tormo, J., Blaas, D., Parry, N. R., Rowlands, D., Stuart, D. & Fita, I. (1994). Crystal structure of a human rhinovirus neutralizing antibody complexed with a peptide derived from viral capsid protein VP2. *EMBO J.* **13**, 2247–2256.

Wagner, S. D., Milstein, C. & Neuberger, M. S. (1995). Codon bias targets mutation. *Nature*, **376**, 732.

Wu, T. T. & Kabat, E. A. (1970). An analysis of the sequences of the variable regions of Bence-Jones proteins and myeloma light-chains and their implications for antibody complementarity. *J. Expt. Med.* **132**, 211–250.

Zachau, H. G. (1993). The immunoglobulin κ locus—or—what has been learned from looking closely at one-tenth of a percent of the human genome. *Gene*, **135**, 167–173.

Edited by J. Karn

(Received 12 October 1995; accepted 28 November 1995)