Comparison of the Human Germline and Rearranged V_H Repertoire Reveals Complementarity between Germline Variability and Somatic Mutation

GERALD WALTER,[a] IAN M. TOMLINSON,[b,c]
PAUL H. DEAR,[b] ERIK L. L. SONNHAMMER,[b,d]
GRAHAM P. COOK,[b,c] AND GREG WINTER[b,c]

[a]*Cambridge Antibody Technology Ltd.*
Science Park
Melbourn SG8 6EJ
United Kingdom

[b]*MRC Laboratory of Molecular Biology*
Hills Road
Cambridge CB2 2QH
United Kingdom

[c]*MRC Centre for Protein Engineering*
Hills Road
Cambridge CB2 2QH
United Kingdom

[d]*Sanger Centre*
Hinxton Hall
Cambridge CB10 1RQ
United Kingdom

We have recently characterized 88 human germline V_H segments in a single individual[1] and completed the map of the human V_H locus (1.1 Mb) on chromosome 14q32.3.[2] We have now turned our attention to the somatic processes that shape the human V_H repertoire. We have sequenced 142 rearranged genes derived from peripheral blood lymphocyte cDNA of our individual (DP) to answer the following questions:

1. Is the usage of V_H segments in rearranged heavy chain genes biased according to their positions on the map of the locus?
2. Is the frequency of somatic mutations influenced by the isotype switch from IgM to IgG?
3. Is the variability created by somatic point mutations confined to the hypervariable loops H1 and H2[3] and does it coincide with the germline variability?

By sequencing germline and rearranged V_H segments from the same individual, we have excluded germline polymorphisms from our DP data. We have also used a specifically designed software package (Sonnhammer *et al.*, manuscript in preparation) for assessing questions (1) through (3) using the EMBL DNA sequence database.

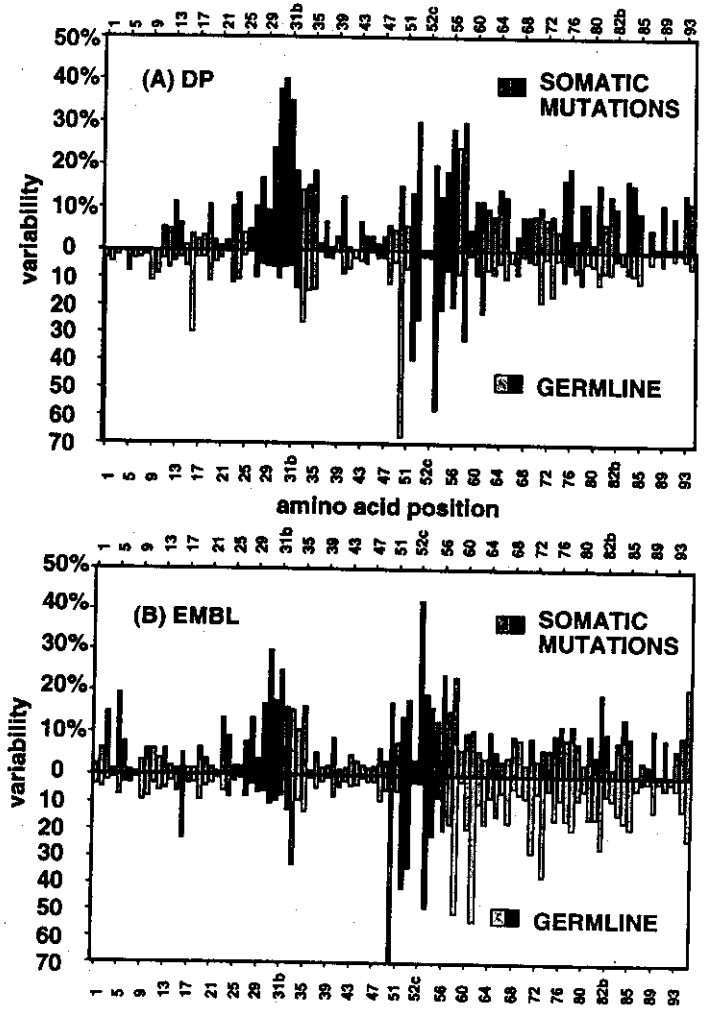

FIGURE 1. Variability of amino acid residues due to somatic point mutations compared to the variability in their germline counterparts scanned over the length of the V_H segments from our individual (DP) **(A)** and the EMBL sequence database **(B)**. The numbering is according to Kabat.[5] The dark-shaded residues form the loops H1 (position 26–32) and H2 (position 52–56),[3] which together with some neighboring residues are the regions of highest variability. *Methods*: Variability due to somatic mutations was determined for each position of the V_H segments as the number of changes per amino acid residue divided by the total number of relevant sequences at that position using purpose-written software (Dear, unpublished). For the germline variability, a score was calculated at each position as the number of different amino acids divided by the percentage frequency of occurrence of the most common amino acid.[5] The numbers of sequences scanned were 143 **(A)** and 1038 **(B)** for DP and EMBL, respectively.

CONCLUSIONS

Our data indicate the following:

1. The overall usage of V_H segments in rearranged heavy chain genes is not biased according to their positions on the map of the locus. Some V_H segments are preferentially expressed in DP (n = 143) and/or highly represented in the EMBL sequence database (n = 1038), as for example 6–1 (DP-74), 1–18 (DP-14), 3–23 (DP-47), and 1–69 (DP-10). Only one rearranged pseudogene (derived from DP-34 with one amino acid change at position 51) was found in the EMBL sequence database. V_H segments located on chromosome 15 or 16[4] were never found rearranged, nor were there any rearranged pseudogenes in DP.
2. The isotype switch from IgM to IgG has a profound influence on the overall frequency of somatic mutations as the average numbers of amino acid changes per V_H segment were calculated to be 2.7 for IgM and 10.1 for IgG.
3. The variability created by somatic point mutations clearly marks the extremities of the hypervariable loops H1 and H2 as well as some adjoining residues involved in shaping antigen-binding sites. It does coincide with the germline variability in H2 but not in H1 where the germline variability is low. That means that the pattern of somatic mutation is complementary to variability encoded by germline V_H segments, thereby maximizing the diversity required for antigen recognition (FIG. 1).

REFERENCES

1. TOMLINSON, I. M., G. WALTER, J. D. MARKS, M. B. LLEWELYN & G. WINTER. 1992. J. Mol. Biol. **227:** 776–798.
2. COOK, G. P., I. M. TOMLINSON, G. WALTER, H. RIETHMAN, N. P. CARTER, L. BULUWELA, G. WINTER & T. H. RABBITTS. 1994. Nature Genet. **7:** 162–168.
3. CHOTHIA, C., A. M. LESK, E. GHERARDI, I. M. TOMLINSON, G. WALTER, J. D. MARKS, M. B. LLEWELYN & G. WINTER. 1992. J. Mol. Biol. **227:** 799–817.
4. TOMLINSON, I. M., G. P. COOK, N. P. CARTER, R. ELASWARAPU, S. SMITH, G. WALTER, L. BULUWELA, T. H. RABBITTS & G. WINTER. 1994. Hum. Mol. Genet. **3:** 853–860.
5. KABAT, E. A., T. T. WU, H. M. PERRY, K. S. GOTTESMAN & C. FOELLER. 1991. Sequences of Proteins of Immunological Interest. 5th edit. USDHHS, NIH. Bethesda, MD.