THE SIZE OF A MURINE HEAVY CHAIN VARIABLE REGION GENE FAMILY: IMPLICATIONS FOR THE MAGNITUDE AND EVOLUTION OF THE V_H LOCUS IN MOUSE

Thesis by

Donna L. Livant

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1985

(Submitted May 1, 1985)

To my parents and to John
with gratitude

Acknowledgements

I am indebted to many people for their help and support during the course of my graduate studies. I thank my thesis advisor, Lee Hood, for his support of this project and for providing a very stimulating environment in which to work. I thank Eric Davidson for several helpful discussions regarding the experimental measurements necessary for my thesis work, and especially for bringing the probe excess titration technique to my attention. I thank Melvin Simon for suggesting the mouse recombinant inbred line experiment to me and Cila Blatt for enabling me to do the experiment. I thank Norman Davidson for helpful discussion of my experimental results. I am grateful to Ray Owen not only for his many helpful suggestions, but also for being an unfailing source of support and encouragement. I am especially grateful to Barbara Hough-Evans for our numerous discussions of experimental techniques and experimental results without which this work would have been far more difficult. I am very much indebted to Iwona Stroynowski for our long discussions of experimental results and for her patient, detailed criticism of the presentation of my thesis work. I gratefully acknowledge John Spalding for his many suggestions regarding my data analysis. Finally, I especially thank Connie Katz and Stephanie Canada for their skillful and speedy preparation of this manuscript.

Abstract

The problem of how much antibody diversity is encoded in the germline as variable region genes has long been of interest to immunologists. We have measured the size of the J558 V_H family in the BALB/c mouse by a probe excess titration method, and found that the family contains approximately 1000 members. As a control for systematic error, we used the same method to measure the number of class I MHC genes in BALB/c. We found that the third domain of the class I D^d gene detects 36-40 class I genes. Dot blots and genome blots with copy number controls give results consistent with a J558 family size of 500-1000 V_H genes. We note that each band evident on genomic blots of DNA from several mouse strains contains multiple V_H genes, and that a significant fraction of these bands are polymorphic among the mouse strains tested. We discuss the implications of this result for both the size and evolution of the V_H locus in mouse.

Table of Contents

INTRODUCTION

A hallmark of the vertebrate immune system is its ability to recognize and distinguish between a very large number of specific molecular patterns of antigens. The antibodies produced by B lymphocytes mediate this specific recognition. Since there are a large number of antigenic determinants whose structures are very different from one another, the repertoire of antibodies produced by vertebrates to bind these determinants specifically must be correspondingly large.

In adult inbred mice, in fact, both the number of different antibodies specific for a given antigen, as well as the total number of different antibodies in the murine repertoire have been measured by several investigators. Kreth and Williamson (1) have estimated, for example, that CBA/H mice make approximately 8000 different antibodies to the NIP (4-hydroxy-3-iodo-5-nitrophenylacetyl) hapten. Similarly, Pink and Asconas (2) have estimated that C3H and CBA mice have a DNP (dinitrophenyl) repertoire size of at least 1600 unique antibodies. Since one B cell is committed to make only one antibody (reviewed in 3), given the overall frequency of B cells making a hapten-specific antibody and the number of unique antibodies (or clonotypes) specific for a given hapten, one can estimate the size of the entire antibody repertoire expressed in the B cells of the mouse. Press and Klinman (4) and Nossal *et al.* (5) have estimated, for example, that between 1/7000 (4) and 1/15,000 (5) B cells express an antibody specific for NIP. Klinman (6) also has estimated that 1/5000 B cells express an antibody specific for DNP. Assuming that the frequency and heterogeneity of the NIP and DNP clonotypes are representative of the repertoire as a whole, Klinman (3) calculates that the mouse has an average of 2.5-7.5 x 10^7 unique clonotypes in its repertoire. Köhler (7) also has arrived at a similar estimate of the murine B cell antibody repertoire size by analyzing the frequency of B cells expressing antibodies specific for β-galactosidase. More recent data from Owen *et al.* (72) measuring the frequency and heterogeneity of clonotypes present in

the IgM anti-phosphorylcholine response of BALB/c mice confirm these findings. Assuming that clonotypes are represented in the mouse by similar numbers of B cells, Klinman (3) calculates that in the mouse lymphoid system containing 2-3 x 10^8 B cells, each clonotype is represented by 3-12 B cells.

One of the primary questions in immunogenetics during the past 20 years has been how the genetic information necessary for making this large number of distinct, but nevertheless, closely related antibodies is stored in the genome. The antibody, or immunoglobulin molecule, is a tetramer, consisting of two identical heavy chains and two identical light chains. Each heavy and light chain has an amino-terminal variable (58) region referred to as V_H or V_L, respectively, and a carboxy-terminal constant region, referred to as C_H or C_L. The variable regions are responsible for antigenic recognition and binding. Each variable region has three subregions demonstrated by X-ray crystallography (63) to be antigen-contacting. These subregions known as hypervariable (59-61) or complementarity-determining (8, 62) regions are flanked by less variable framework regions. While the variable regions of immunoglobulins are numerous and diverse, there are only a few classes of constant regions in a given species. The function of the immunoglobulin constant region is to initiate one of a number of effector functions, for example complement-fixation, when the variable end of the immunoglobulin binds antigen.

The notion that V and C regions of a particular immunoglobulin polypeptide chain are encoded by two separate loci in germline DNA and undergo a joining rearrangement during lymphocyte development was first suggested by Dreyer and Bennett (9) in 1965. Since that time, many studies have demonstrated that V and C regions are encoded separately in multiple gene segments on germline DNA (reviewed in 10). One germline V gene segment of many must be joined to one constant region gene in order to form the complete immunoglobulin gene transcribed in B lymphocytes (11-14). Furthermore, many studies suggest that immunoglobulin

genes for both heavy (15-19) and light (20-22) chains undergo somatic mutation at a high rate (15). The following is a brief review of the organization of immunoglobulin genes in mice and of the mechanisms that further diversify the genetic information in the germline genome. The same principles hold for other mammalian systems as well (24-27).

Mouse immunoglobulin genes reside in three unlinked families, κ, λ and H (heavy), which are located on chromosomes 6 (28, 29), 16 (30), and 12 (28, 31, 32), respectively. In the mouse haploid genome, there are an unknown number of V_H and V_κ segments, four V_λ segments, four J_H segments, five J_κ segments, four J_λ segments, and at least 12 D_H segments (references 10 and 33 are recent reviews). A complete immunoglobulin V region is assembled from a V gene segment and a J (joining) segment in the case of the light chain families κ and λ, and from a V gene segment, a D segment, and a J segment in the case of heavy chains. Figure 1a depicts the organization of the mouse immunoglobulin gene segments in germline DNA. Figure 1b shows an example of a joined heavy chain gene of the IgM class. Complete heavy chain genes encoding immunoglobulins of other classes as well as complete light chain genes have similar structures.

The mouse uses four mechanisms to diversify further the information present in its germline genome. These are combinatorial joining, junctional diversity, junctional insertion, and somatic mutation. Combinatorial joining means that any V_κ and any J_κ can join to create a complete V_κ gene; likewise, any V_H and any D and any J_H can join to create a complete V_H gene. There appear to be some limits on which V_λ can join with J_λ (10). The mechanism of the joining is unknown but has been proposed to occur by looping out and deletion (34), by inversion (35, 36), and by sister-chromatid exchange (37, 38). Combinatorial joining of V_Hs and J_Hs has been shown to occur by Schilling *et al.* (39). Combinatorial joining of the immunoglobulin gene segments to form complete V genes is a diversifier of germline information

because the number of unique V_H genes potentially formed is the product of the number of V_H segments, J_H segments, and D segments. Similarly, the number of potential, unique V_κ genes is the product of the number of V_κ segments and the number of J_κ segments. It is worthwhile to note, however, that since Ds and J_Hs or J_κs form only the third hypervariable region of the complete V gene, combinatorial joining and the mechanismns described below which serve to diversity the V-D, D-J, and V-J junctions can affect the structure of only one of three antibody combining sites. The first and second hypervariable regions encoded in the germline V_H and V_κ gene segments are not diversified by these processes.

Imprecise joining or junctional diversity and junctional insertion refer to mechanisms varying still further the resulting V region amino acid sequence for one or two residues around the V_H-D, D-J_H, and V_κ-J_κ junctions. Junctional diversity in V_κ chains, for example, makes amino acid 96 a hot spot (14). In many κ chains codon 96 comes from one of the four junctional germline J_κ segments, but in several instances, codon 96 derives from one or two nucleotides at the 3' end of the V_κ gene segment in addition to the J_κ nucleotides (12, 41, 42), or may even be deleted altogether, suggesting that the V_κ to J_κ joining mechanism is imprecise. Similar observations have been made for J_H segments (43, 44) and D segments (40). A related phenomenon is the junctional insertion of one to four nucleotides, apparently without a template, at the D and J_H junction (40). All of the above experimental facts have been recently incorporated into the model for D-J_H joining proposed by Alt and Baltimore (45).

The least understood diversifier of antibody genes is somatic mutation (46, 47). Currently, most immunologists view somatic mutation as a system for the hypermutation of bases in and around V genes (10, 17, 23) although mechanisms such as reciprocal recombination between homologous V genes (48, 49) and gene conversion (50-52) have been proposed. Studies of antibody diversity (53) and of the diversity of

germline V region genes (17, 18) of antibodies binding phosphorylcholine in BALB/c mice constitute the most direct evidence for somatic mutation at the heavy chain locus. Similar observations have been made in the NP [(4-hydroxy-3-nitrophenyl)-acetyl] system (54, 77), but the large numbers of V_H genes cross hybridizing with the NP probe complicate the data. Somatic mutation also has been observed in V_κ (21, 55) and V_λ (10) genes as well.

Little is known about how somatic mutation works, what makes it specific for V genes, and when during B cell maturation it happens. Crews *et al.* (18), Gearhart *et al.* (53), and Bothwell *et al.* (54) have suggested that in heavy chains, V-D-J joining precedes somatic mutation, and that somatic mutation might be both temporally and mechanistically linked to the immunoglobulin heavy chain class switch (74). Their evidence was that germline V_Hs were associated with the μ constant region, whereas somatically mutated V_Hs were associated with γ or α constant regions. Problems with this hypothesis arise because both μ chains with somatically mutated V_H (23) and α chains with germline V_H (53) have been found. One could equally well propose that somatic mutation coincides with V-D-J joining, or even that it is a completely separate event. These difficulties in characterizing somatic mutation arise because in order for a particular antibody to appear in the immune response or even in a myeloma where we can study it, the clone of B cells secreting that antibody must be selected for expansion by antigen or antiidiotype. Specific T helper cells, and therefore the network of T cell regulatory mechanisms, also intervene in this process. Studies on the phosphorylcholine response (72) and on the arsonate response (73), for example, show that the repertoire of precursor B cells specific for each of these antigens is far more diverse than the antibody in the serum of mice immunized with these antigens. Hence, we know very little about the role selection plays in clonal expansion and little about the effect of both of these on the representation of germline or somatically mutated antibodies in the serum of the immune response or

in myelomas. We know even less about the mechanism of somatic mutation itself. Brenner and Milstein (56) have proposed that the point mutations occur by repair synthesis after excision using an error-prone polymerase.

In summary, the four mechanisms diversifying germline-encoded information specifying immunoglobulin variable region genes can potentially make a large number of different antibodies or clonotypes from a relatively small number of germline genes. For example, based on combinatorial diversity, a mouse with 300 V_κ genes and four functional J_κ gene segments could make a maximum of 1200 unique κ light chains. Similarly, the same mouse with 200 V_H genes, four J_H gene segments and 12 D segments has the potential to make 9600 unique heavy chains. Assuming that any heavy chain can associate with any light chain, this mouse could make 1.15×10^7 unique antibodies. Furthermore, the mechanisms of somatic mutation, junctional diversity, and junctional insertion contribute an unknown amount of additional diversity to V genes. Using similar calculations, many immunologists (10, 14, 57) have concluded that combinatorial diversity superimposed upon relatively small numbers of germline V genes and bolstered by somatic mutation and junctional diversity can potentially create enough different antibodies ($>10^7$) to satisfy the predictions of the repertoire studies (3).

A major point implicit in the foregoing discussion is the degree to which the genetic information necessary for antibody diversity is encoded directly in the genome as germline V genes. Since at least a minimum estimate of germline diversity could be made directly, as soon as cloned V genes or even purified immunoglublin message was available, many measurements of germline V gene diversity have been made during the past 10 years. Although a few earlier studies indicated that the mouse V_H (69) and V_κ (68, 71) loci might be relatively large, the overall conclusions reached by these authors, and by immunologists in general, are that the mouse has 200 V_κ gene segments (78) and 100-200 V_H gene segments (66, 78). Similarly, humans have 50 V_κ gene segments and 100 V_H gene segments (78).

However, all measurements have underlying assumptions in common. The conclusions reached by each study depend on an experimental measurement of the number of members of a particular V gene family or subgroup followed, in most of the studies, by an implicit or explicit extrapolation to the total number of germline V genes based on statistical arguments as to the number of different families or subgroups and the average number of V genes in each. The statistical arguments leading to these extrapolations depend upon the assumption that the protein sequences of secreted immunoglobulins from myelomas or joined V gene probes derived from myeloma DNA are a random, representative distribution of the number of different V gene families or subgroups found in the germline DNA. These studies further depend upon the assumption that the family in question has an average number of V genes with respect to all other families, and finally, that the number of these V genes has been measured in a controlled fashion without experimental bias or misinterpretation. We will discuss each of these assumptions in turn.

The first assumption is that the joined V genes, and therefore the V regions of secreted immunoglobulins of myelomas are a random representative sample of the members from all V gene families in germline DNA. (There are V gene probes, for example, Vh36-60 and VhJ606, which were cloned as germline genes from genomic libraries [66], but all such genes were isolated by virtue of their hybridization to one of the myeloma-derived probes.) There are several reasons why the myeloma-derived probes and protein sequences available to us may not be a representative sample of germline diversity. The most obvious reason is that, of the approximately 1500 myelomas screened, only about 10% bound a known antigen (79). The antigens used to screen myelomas were for the most part bacterial starches such as α1-3 dextran and α1-6 dextran, heat-killed bacteria, other bacterial antigens like phosphorylcholine, and planar ring compounds such as the nitrophenyl derivatives (79). The remaining 90% of myelomas have not been studied. Most myelomas in the

two susceptible mouse strains, BALB/c and NZB, were induced by repeated mineral oil injection. Even if the entire set of mineral oil-induced myelomas of BALB/c and NZB mice had been characterized, it is likely that they would still be a skewed representation of the germline repertoire. The mineral oil induction itself might make some B cell clones proliferate more than others. Prior exposure of the mouse to antigens would also result in expanded clones of certain B cells.

Isoelectric focusing of serum antibody in the primary response of BALB/c mice, to α1-3 dextran, a bacterial starch, for example, shows that both the class of the antibody and the variable regions present in the antibodies vary greatly, depending upon whether or not the mice were germ free (67). This is presumably because antibodies made against some of the bacterial antigens encountered previously by the mouse can crossreact with the α1-3 dextran antigen. Nonrepresentativeness of V region sequences found in myelomas could also be the result of bias inherent in the transformation event which makes a myeloma out of a B cell. Although we do not know whether the transformation event yielding a myeloma is selective, one very plausible mechanism for generating bias in the B cells represented as myelomas is positive selection based on the cell cycle of the B cell involved. Resting B cells might be less likely to be transformed than those actively dividing. Since immunoglobulin V gene probes correspond solely to those antibodies whose sequence is known, what we know about the number of different germline V_H genes derives at best from the V_H genes represented by approximately 10% of all mineral oil-induced myelomas regardless of how this number is measured.

Difficulties with the second assumption that the size of a particular V gene family or V region subgroup is likely to represent a size typical of most V region families, arise for the same reasons as those given above. Even in our limited sample, numbers of germline genes present in a given family range from four for phosphorylcholine (18) to approximately 1000 for α1-3 dextran. Obviously,

extrapolations based on numbers of genes in either of these families are likely to be wrong.

The final difficulty in the previous estimates of the number of germline variable region genes has to do with the accuracy of measurement of the size of a particular V gene family or V region subgroup measured. Difficulties in measuring the germline contribution to the size and diversity of a group or subgroup of V region protein sequences arise because it is uncertain which V regions found in a group of myeloma or hybridoma V region sequences represent germline genes. All V regions represented this way have undergone selection by antigen, T-cell regulatory mechanisms and transformation events. The pool of sequences represented, therefore, reflects the variable regions which fit the antigen best, were from a clonally expanded population to begin with or were rapidly dividing and perhaps more likely to be transformed. Although the phosphorylcholine germline V_H gene sequence (18) was correctly identified by protein sequencing of the group of V_H genes from myelomas binding phosphorylcholine (53), the number of occurrences of a given V region sequence in a group may not always be directly related to the likelihood that the V region sequence is present in the germline. All V regions represented in the serum in myelomas, or in hybridomas, have had the chance to undergo somatic mutation. Since somatic mutation seems to be involved in affinity maturation (53), B cells expressing antibodies whose V regions have undergone somatic mutation may well be present more abundantly by virtue of their improved antigen binding, and hence more efficient antigen selection. Thus, the number of germline sequences identified in studies like these is probably an underestimate.

The most persistent experimental difficulty with current estimates of V gene diversity, however, is that the presence of a band on a genomic blot is equated with one V gene. Controls establishing this point have been done only for mouse phosphorylcholine-binding V_H genes (18) and human $V_{\kappa I}$ genes (65). When there are a

large number (>20) of bands present, as is true for the V_H family identified by the J558 (Dex) probe, this assumption leads to an underestimation of family size of greater than tenfold. Other than estimates of family size based on the number of restriction fragments evident on genomic blots, no recent attempts to determine the size of a large variable region gene family have been made.

In summary, almost all of the current experimental methods for estimating V region diversity in germline DNA probably seriously underestimate the actual diversity present in the locus. Hence, we have estimated the size of the largest known variable region gene family, the J558 V_H family, directly in BALB/c germline DNA. We have arrived at the conclusion that the J558 V_H family contains approximately 1000 members by a probe excess titration experiment which is independent of hybridization rate. We have also demonstrated that this method of measurement yields a size for the family of histocompatibility class I genes which is consistent with published results. We further show that results consistent with the probe excess titration measurement are obtained from both dot blots and genome blots with copy number controls. We also estimate how related the members of the J558 V_H family are to the J558 V_H sequence. Finally, we examine the implications of these measurements for the size and evolution of the mouse V_H locus.

The studies we report on the J558 V_H family of the BALB/c mouse are subject to the difficulties in representativeness of the family with respect to size discussed for myeloma—derived probes in general. The J558 family is certainly the largest known V_H gene family (66), but it is unknown whether other V_H families of this size are present in the mouse germline. Therefore, we can make only a minimum estimate of V_H locus size based on our measurements. The J558 family is large enough, however, that it is a useful minimum estimate of V_H locus size. It alone is 10 times as large as the most current estimate for the size of the entire V_H locus (66). Of course, neither our experiments nor those of any other except indirectly

those of Bentley on the human V_κ locus (64) attempt to address the question of how much of the detectable germline V gene diversity is functional diversity represented in the B cell repertoire. Estimates of the fraction of pseudogenes present in the V_H locus range from 25% (18) to 40% (70) or 50% (54, 77). Furthermore, although a recent study by Manser *et al.* (80) has suggested that the V_H gene repertoire is formed stochiastically through random joining of germline V_H gene segments and random association with light chains, other recent studies (75, 76, 81) have suggested that certain V_H gene segments may be rearranged in a nonrandom manner. Much more work needs to be done in this area before we draw any conclusion about the likelihood of a particular germline V regions being joined and expressed in the repertoire. It is possible that many V genes are rarely, if ever, used even when we consider the entire species, and therefore, what we mean when we ask how many germline V genes are functional must be carefully defined.

References

1. Kreth, H. W. and Williamson, A. R. (1973) Eur. J. Immunol. **3**, 141.
2. Pink, J. R. L. and Asconas, B. (1974) Eur. J. Immunol. **4**, 426.
3. Sigal, N. H. and Klinman, N. R. (1978) B Cell Clonotype Repertoire.
4. Press, J. L. and Klinman, N. R. (1974) Eur. J. Immunol. **4**, 155.
5. Nossal, G. J. V., Stocker, J. W., Pike, B. and Goding, J. W. (1977) Cold Spring Harbor Symp. Quant. Biol. **41**, 237.
6. Klinman, N. R. (1972) J. Exp. Med. **136**, 241.
7. Köhler, G. (1976) Eur. J. Immunol. **6**, 340.
8. Davies, D. R., Padlan, E. A. and Segal, D. A. (1975) Ann. Rev. Biochem. **44**, 639.
9. Dreyer, W.J. and Bennett, J. C. (1965) Proc. Natl. Acad. Sci. USA **54**, 864.
10. Tonegawa, S. (1983) Nature **302**, 575.
11. Brack, C., Hirama, M., Lenhard-Schuller, R. and Tonegawa, S. (1978) Cell **15**, 1.
12. Early, P., Huang, H., Davis, M., Calame, J. and Hood, L. (1980) Cell **19**, 981.
13. Seidman, J. G. and Leder, P. (1978) Nature **276**, 790.
14. Weigert, M., Gatmaitan, L., Loh, E., Schilling, J. and Hood, L. (1978) Nature **276**, 785.
15. Teillaud, J.-L., Desaymard, C., Guisti, A. M., Haseltine, B., Pollock, R. R., Yelton, D. E., Zack, D. J. and Scharff, M. D. (1983) Science **222**, 721.
16. Rudikoff, S., Pawlita, M., Pumphrey, J. and Heller, M. (1984) Proc. Natl. Acad. Sci. USA **81**, 2162.
17. Kim, S., Davis, M., Sinn, E., Patten, P. and Hood, L. (1981) Cell **27**, 573.
18. Crews, S., Griffin, J., Huang, H., Calame, K. and Hood, L. (1981) Cell **25**, 59.
19. Sims, J., Rabbitts, T. H., Esters, P., Slaughter, C., Tucker, P. and Capra, J. D. (1982) Science **216**, 309.
20. Pech, M., Höchtl, J., Schuell, H. and Zachau, H. G. (1981) Nature **291**, 668.

21. Selsing, E. and Storb, U. (1981) Cell **25**, 47.

22. Bernard, O., Hozumi, N. and Tonegawa, S. (1978) Cell **15**, 1133.

23. Rudikoff, S., Pawilita, M., Pumphrey, J. and Heller, M. (1984) Proc. Natl. Acad. Sci. USA **81**, 2162.

24. Rabbitts, T. H., Matthyssens, G. and Hamlyn, P. H. (1980) Nature **284**, 238.

25. Hieter, P. A., Korsmeyer, S. J., Waldman, T. A. and Leder, P. (1981) Nature **290**, 369.

26. Sheppard, H. W. and Gutman, G. A. (1982) Cell **29**, 121.

27. Matthyssens, G. and Rabbitts, T. H. (1980) Proc. Natl. Acad. Sci. USA **77**, 6561.

28. Hengartner, H., Meo, T. and Muller, E. (1978) Proc. Natl. Acad. Sci. USA **75**, 4495.

29. Swan, D., D'Eustachio, P., Leinwand, L., Seidman, J., Keithley, D. and Ruddle, F. H. (1979) Proc. Natl. Acad. Sci. USA **76**, 2735.

30. D'Eustachio, P., Bothwell, A. L. M., Takaro, T. K., Baltimore, D. and Ruddle, F. H. (1981) J. Exp. Med. **153**, 793.

31. Meo, T., Johnson, J., Beechey, C. V., Andrews, S. J., Peters, J., Searle, A. G. (1980) Proc. Natl. Acad. Sci. USA **77**, 550.

32. D'Eustachio, P., Pravtcheva, D., Marcia, K. and Ruddle, F. H. (1980) J. Exp. Med. **151**, 1545.

33. Shimizu, A., Takahashi, N., Yaoita, Y. and Honjo, T. (1982) Cell **28**, 499.

34. Sakano, H., Hüppi, K., Heinrich, G. and Tonegawa, S. (1979) Nature **280**, 288.

35. Lewis, S., Gifford, A. and Baltimore, D. (1984) Nature **308**, 425.

36. Höchtl, J. and Zachau, H. G. (1983) Nature **302**, 260.

37. Höchtl, J., Müller, C. R. and Zachau, H. (1982) Proc. Natl. Acad. Sci. USA **79**, 1383.

38. Van Ness, B. G., Coleclough, C., Perry, R. P. and Weigert, M. (1982) Proc. Natl. Acad. Sci. USA **79**, 262.

39. Schilling, J., Clevinger, B., Davie, J. M. and Hood, L. (1980) Nature **283**, 35.

40. Kurosawa, Y. and Tonegawa, S. (1982) J. Exp. Med. **155**, 201.

41. Sakano, H., Max, E. E., Seidman, J. G. and Leder, P. (1979) Proc. Natl. Acad. Sci. USA **76**, 3450.

42. Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J. and Hood, L. (1980) Nature **283**, 497.

43. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. and Tonegawa, S. (1980) Nature **286**, 676.

44. Kurosawa, Y., von Boehmer, H., Haas, W., Sakano, H., Trauneker, A. and Tonegawa, S. (1981) Nature **290**, 565.

45. Alt, F. and Baltimore, D. (1982) Proc. Natl. Acad. Sci. USA **79**, 4118.

46. Burnet, M. (1966) Nature **210**, 1308.

47. Lederberg, J. (1959) Science **129**, 1649.

48. Gally, J. A. and Edelman, G. M. (1970) Nature **227**, 341.

49. Whitehouse, H. L. K. (1967) Nature **227**, 371.

50. Baltimore, D. (1981) Cell **24**, 592.

51. Dildrop, R., Brüggemann, M., Radbruch, A., Rajewsky, K. and Beyreuther, K. (1982) EMBO J. **5**, 635.

52. Schreier, P. H., Bothwell, A. L. M., Müller-Hill, B. and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA **78**, 4495.

53. Gearhart, P. J., Johnson, N. D., Douglas, R. and Hood, L. (1981) Nature **291**, 29.

54. Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K. and Baltimore, D. (1981) Cell **24**, 625.

55. Gershenfeld, H. K., Tsukamoto, A., Weissman, I. L. and Joho, R. (1982) Proc. Natl. Acad. Sci. USA **78**, 7674.

56. Brenner, S. and Milstein, C. (1966) Nature **211**, 242.

57. Leder, P., Max, E. E., Seidman, J. G., Kwan, S. P., Scharff, M., Nan, M. and Norman, B. (1980) Cold Spring Harbor Symp. Quant. Biol. **XLV**, 859-866.

58. Hilschmann, N. and Craig, L. C. (1965) Proc. Natl. Acad. Sci. USA **53**, 1403.

59. Kehoe, J. M. and Capra, J. D. (1971) Proc. Natl. Acad. Sci. USA **68**, 2019.

60. Capra, J. D. and Kehoe, J. M. (1974) Proc. Natl. Acad. Sci. USA **71**, 845.

61. Capra, J. D. and Kehoe, J. M. (1975) Adv. Immunol. **20**, 1.

62. Kabat, E. A. and Wu, T. T. (1971) Ann. N.Y. Acad. Sci. **190**, 382.

63. Segal, D. M., Padlan, E. A., Cohen, G. H., Rudikoff, S., Potter, M. and Davies, D. R. (1974) Proc. Natl. Acad. Sci. USA **71**, 4298.

64. Bentley, D. (1984) Nature **307**, 77.

65. Bentley, D. and Rabbitts, T. H. (1981) Cell **24**, 613.

66. Brodeur, P. and Riblet, R. (1984) Eur. J. Immunol. **14**, 922.

67. Schuler, W., Lehle, G., Weiler, E. and Kölsch, E. (1982) Eur. J. Immunol. **12**, 120.

68. Storb, U. (1974) Biochem. Biophys. Res. Commun. **57**, 31.

69. Premkumar, E., Shoyab, M. and Williamson, A. R. (1974) Proc. Natl. Acad. Sci. USA **71**, 99.

70. Rechavi, G., Ram, D., Glazer, L., Zakut, R. and Givol, D. M. (1983) Proc. Natl. Acad. Sci. USA **80**, 855.

71. Hood, L., Barstad, P., Loh, E. and Nottenburg, C. N. (1974) In *The Immune System: Genes, Receptors and Signals* (E. E. Sercarz, A. R. Williamson, and C. F. Fox, eds.), p. 119. Academic Press, New York.

72. Owen, J. A., Sigal, W. H. and Klinman, N. R. (1982) Nature **295**, 347.

73. Eichmann, K., Coutinho, A. and Melchers, F. (1977) J. Exp. Med. **146**, 1436.

74. Shimizu, A. and Honjo, T. (1984) Cell **30**, 801.

75. Perlmutter, R. M., Kearney, J., Chang, S. P. and Hood, L. (1985) Science **227**, 1597-1601.

76. Klinman, N. R. and Stone, M. R. (1983) J. Exp. Med. **158,** 1948.

77. Loh, D. Y., Bothwell, A. L. M., White-Scharf, M. E., Imanishi-Kari, T. and Baltimore, D. (1983) Cell **32,** 85.

78. Kindt, T. and Capra, J. D. (1984) *The Antibody Enigma*, Plenum Press, New York, p. 232.

79. Potter, M. (1977) Adv. Immunol. **25,** 141.

80. Manser, T., Huang, S.-Y. and Gefter, M. (1984) Science **226,** 1283.

81. Yancopoulos, G., Desiderio, S. V., Paskind, M., Kearney, J., Baltimore, D. and Alt, F. (1984) Nature **311,** 727.

MATERIALS AND METHODS

Preparation of DNAs

Genomic DNA was prepared from the livers of BALB/c J, BALB/c CRGL, BALB/c By mice and from PVG rat liver by the method of Blin and Stafford (1). Other mouse DNAs were obtained from Jackson Laboratories. Chimpanzee and orangutan DNAs were the gifts of H. Hu. Cerebratulus and Thyone DNAs were the gift of B. Evans. BALB/c J liver DNA was mechanically sheared to an average size of 500-600 ntp as previously described (2). J558 V_H was subcloned into the SmaI and PstI sites of mp8 (3) as a 267 bp PstI-PvuII insert isolated from the joined J558 gene (4). The joined J558 gene cloned into λgtWes was the generous gift of P. Brodeur and R. Riblet. $D^d\alpha 3$ is an AluI partial cloned into the mp8 SmaI site during the sequencing of the H-$2D^d$ gene (5). Preparations of both mp8 clones were grown in JM103 as previously described (6). J558 V_H and $D^d\alpha 3$ single-stranded DNA preparations used for all of the experiments described here were verified by sequence after growth and isolation. Sequences of J558 V_H and $D^d\alpha 3$ are shown in Figure 2. All sequencing was done by the chain termination method of Sanger (7).

Synthesis of single-stranded probes for titration and melting

J558 V_H and $D^d\alpha 3$ probes were prepared identically and in parallel. Approximately 2.5 μg of the single-stranded clone in m13mp8 was preannealed to 20 ng of primer fragment (8) obtained from Amersham. This reaction was carried out in a 10 μL total volume in 75 mM Tris, pH 7.5, 7.5 mM $MgCl_2$ for 15 minutes at 55°C. After preannealing, each deoxynucleotide was added to a final concentration of 125 μM. 40 μCi of $\alpha[^{32}P]$dATP (400 Ci/mmole, Amersham), and 5U of Klenow enzyme from Boeringer-Mannheim was also added. The final reaction volume was 18 μL. The synthesis reaction proceeded for 1 hr at 37°C. At this point, NaCl was added to a final concentration of 100 mM and the completed double-stranded molecules were

digested with 40 U of BRL EcoRI for 1-2 hr at 37°C. At the completion of the digestion, the reaction was made 0.5 M with respect to NaOH and 50 mM with respect to EDTA in order to denature the DNA. The reaction was then loaded onto a 2% agarose gel and electrophoresed at 30V for 16 hr. After the run, the band containing the labeled insert was electrophoresed into DE81 paper (Whatman) and eluted with 300 µL of 1.5 M NaCl/0.2 M NaOH. The solution containing the DNA was neutralized with 3 M NaOAc, pH 4.8, and the DNA precipitated with 2.5 volumes of 100% EtOH in the presence of 25 µg tRNA. The DNA was then dissolved in 100 µL 0.12 M PB/0.1% SDS and reacted at 60°C to approximately 50 x $Cot_{\frac{1}{2}}$. The probe was then pased over HAP and the single-stranded probe collected, concentrated by n-butanol extraction, and desalted over Sephadex G-50 as previously described (2, 9). This procedure resulted in a single-stranded probe of a specific activity of about 5 x 10^7 cpm/µg. Probes of higher specific activity undergo significant radiolysis during the week required to run and assay a given experiment.

Synthesis of single-stranded probes for dot blots and genome blots

These syntheses were done as for the low specific activity probe except that the synthesis step proceeded for 30 min at 30°C in the presence of 40-80 µCi of each $\alpha[^{32}P]$ deoxynucleotide. After 30 min, each cold deoxynucleotide was added to a final concentration of 125 µM and the chase continued for 30 min at 30°C. Under these synthesis conditions, the reactions proceed for at least 500 base pairs before the chase. Therefore, the entire insert is labeled during the synthesis (data not shown). These probes were then gel-purified as before, neutralized, and used directly.

DNA solution hybridizations

All hybridization reactions were carried out in 0.12 M PB at 60°C or in 0.41 M PB at 65°C. All Cot values reported here are equivalent Cots; that is, they are corrected for the relative increase in rate due to salt concentrations above 0.18 M Na^+ (0.12 M PB). Reactions also contained 0.1% SDS. Titration reaction mixtures

contained from 3.26×10^{-5} μg to 1.16×10^{-4} μg of ^{32}P-labeled single-stranded probe at 5×10^{7} cpm/μg, and from 7.3×10^{-3} μg to 1.04 μg of sheared BALB/c DNA. In the reactions of any one set, R, the ratio of the genomic DNA mass to the probe DNA mass, was increased by adding increasing amounts of genomic DNA and keeping the mass of the probe constant. Reaction volumes were uniformly 10 μL and were sealed in 20 μL siliconized capillaries and boiled 2 min before reaction. Always, each reaction set had a sample without genomic DNA as a zero control. Reactions run without genomic DNA typically had 0-4% of the total counts in duplex; therefore, this amount was accordingly subtracted when determining the percentage of double-stranded counts in the other reactions. All sets of reactions also had a sample driven to completion with genomic DNA in order to measure the total reactivity of the probe. Most reactivity values ranged from 75% to 85%. Reactions driven to completion with excess genomic DNA contained from 3.26×10^{-5} to 1.16×10^{-4} μg of single-stranded probe and 50 μg of sheared genomic DNA. These reactions proceeded to a genomic Cot of 30,000. Reactions driven to completion with excess J558 V_H template as 100% homology controls contained 500 μg of the corresponding parent m13 clone. Chromatography by hydroxyapatite (HAP) was done as previously described (2).

Southern blotting and hybridization

DNA was transferred from 0.8% agarose (Seakem) gels (9) to nitrocellulose (Schleicher and Schuell) by the method of Southern (10). Dot blots were set up using the Schleicher and Schuell matrix in order to keep all spot sizes uniform. All blots were hybridized in 0.8 M Na^+ (5X SET, 10X Denhardt's, 0.1% sodium pyrophosphate) at a single-stranded probe concentration of 1-2 ng/ml for 36-40 hr at 68°C (9). All blots were washed down twice in 0.6 M Na^+ (4X SET, 0.2% SDS) and twice in 0.3 M Na^+ (2X SET, 0.2% SDS). Exposure times for particular blots are noted in the figure legends.

Mathematical relations used in titration experiments

The equation which describes the complete titration curve is

$$\frac{t}{t_0} = \frac{1}{1 + \frac{a}{R}}$$

where t is the number of double-stranded counts, t_0 is the total number of counts, 1/a is the mass fraction of the genome which can hybridize to the probe under the given conditions, and R is the ratio of the mass of genomic DNA to the mass of the probe DNA (11). This expression is equivalent to

$$\frac{t}{t_0} = \frac{R}{a}\left(\frac{1}{\frac{R}{a} + 1}\right),$$

when $(R/a) << 1$, then $(R/a) + 1$ approaches 1 and (t/t_0) approaches $(1/a)R$. The point with the largest R value fit to a straight line was $t/t_0 = 0.304$, $R = 2800$. At $R = 2800$, $\frac{1}{(R/a+1)} = 0.79$. Therefore, at this point, the actual value of t/t_0 should have been 0.384 instead of 0.304. Most values of R ranged around 700. Here $R/a = 6.59 \times 10^{-2}$ and $\frac{1}{(R/a+1)} = 0.94$. In this range, a t/t_0 value of 0.052 would have been closer to 0.055, an insignificant difference. The expression,

$$\frac{t}{t_0} = \frac{1}{1 + \frac{a}{R}}$$

can be derived from two initial assumptions,

$$\left(\frac{1}{a}\right) G = x + y$$

and

$$\frac{x}{y} = \left(\frac{1}{a}\right) G \times \left(\frac{1}{m}\right)$$

where G is the total mass of genomic DNA present in the reaction, m is the total mass of single-stranded probe present, and 1/a is the mass fraction of the genome reactable with the probe. G/a is then the mass of the genome reactable with the probe. X is the mass of the genome reactable with the probe that reacts with its own opposite strand, and y is the mass of the genomc reactable with the probe that does, in fact, hybridize to the probe. Then

$$x + y = \left(\frac{1}{a}\right) G$$

is just a statement of mass conservation. The second statement,

$$\frac{x}{y} = \frac{G}{a} \times \frac{1}{m}$$

amounts to the statement that the degree to which the genomic mass reactable with the probe reacts with itself (x) or with the probe (y) depends solely on the relative number of reactable sequences present in the probe or in the genome. In order to write this statement in terms of mass, it must be true that the relationship between the number of reactable sequences and the mass of reactable DNA is the same for both the probe and the genomic DNA. For both J558 V_H and $D^d\alpha 3$ this is the case. Realizing that G/m = R and that therefore x = (1/a)Ry,

$$\frac{G}{a} = \frac{mR}{a} = \frac{1}{a} Ry + y$$

and

$$y = \frac{\frac{1}{a}\, mR}{\frac{1}{a}\, R+1} .$$

Multiplying by

$$\frac{\frac{a}{mR}}{\frac{a}{mR}}$$

gives

$$y = \frac{m}{1 + \frac{a}{R}} .$$

Dividing by m and realizing that the mass of genomic DNA which does react with the probe equals the mass of the probe reacting with the genomic DNA, we get

$$\frac{y}{m} = \frac{t}{t_0} = \frac{1}{1 + \frac{a}{R}} .$$

All titration data in the linear range was fit using the method of least-squares and keeping the origin fixed.

References

1. Blin, N. and Stafford, D. W. (1976) Nucleic Acids Res. **3**, 2303.
2. Britten, R. J., Graham, D. E. and Neufeld, B. R. (1974) *Methods In Enzymology*, **XXIX**, 363-418.
3. Messing, J. and Vieira, J. (1982) Gene **19**, 269-276.
4. Brodeur, P. and Riblet, R. (1984) Eur. J. Immunol. **14**, 922.
5. Sher, B. T., Nairn, R., Coligan, J. E. and Hood, L. E. (1985) Proc. Natl. Acad. Sci. USA **82**, 1175-1179.
6. Moore, K. W., Sher, B. T., Sun, Y.-H., Eakle, K. and Hood, L. (1982) Science **215**, 679-682.
7. Sanger, F. (1981) Science **214**, 1205-1210.
8. Schreir, P. H. and Cortese, R. (1979) J. Mol. Biol. **129**, 169-172.
9. *Molecular Cloning, A Laboratory Manual*, (1982) T. Maniatis, E. F. Fritsch, J. Sambrook. Cold Spring Harbor Laboratory.
10. Southern, E. (1975) J. Mol. Biol. **98**, 503.
11. Scheller, R. H., Constantini, F. D., Kozlowski, M. R., Britten, R. J. and Davidson, E. H. (1978) Cell **15**, 189-203.

RESULTS

Measurement of the size of the family of J558-like V_H genes by probe-excess titration of BALB/c germline DNA

Figure 3 shows the data obtained from titrating the J558 V_H coding region probe (J558 V_H) with BALB/c J liver DNA randomly sheared to an average size of 500-600 base pairs. This probe was synthesized and made single stranded as described in *Methods*. All of the titrations were run at criteria equivalent to 0.18 M Na^+, 60°C. Figure 4 shows the data obtained from titrating the BALB/c H-2D^d third domain probe ($D^d\alpha 3$) with the same preparation of randomly sheared BALB/c DNA as used for J558V_H. Both probes were made in parallel and used immediately to run the titrations also in parallel. Hence, variables such as the exact age and specific activity of a particular batch of ^{32}P-labeled nucleotides, the effects due to radiolysis, the individual variations in the temperature of reaction and the various points of reaction termination are the same for both probes. Like symbols in Figures 3 and 4, therefore, represent reactions run in parallel and can be compared directly. It is important to realize that these are not rate experiments. Every point represents the kinetic termination of a probe-driven reaction with genomic DNA. As described in the figure legends of Figures 3 and 4, we ran some sets of reactions to kinetic termination, that is, to 10X $Cot_{\frac{1}{2}}$ of the probe. Others, however, we ran still further to either 30X $Cot_{\frac{1}{2}}$ or 100X $Cot_{\frac{1}{2}}$. This was to insure that the apparent size of the family of sequences reacting with either J558 V_H or $D^d\alpha 3$ was not rate-limited. All reactions were assayed by binding to hydroxyapatite (HAP) as described in *Methods*. The ordinate in Figure 3 and Figure 4 is t/t_0 or the fraction of the total counts that binds to HAP and is, therefore, the fraction of counts which is in duplex. The abcissa is the ratio of genomic DNA mass to probe mass at constant probe mass (R). The expression

$$\frac{t}{t_0} = \frac{1}{1+\frac{a}{R}}$$

where $\frac{1}{a}$ is the mass fraction of genomic DNA which will hybridize to the probe under the particular set of hybridization conditions used describes the complete curve generated by the data in these experiments. *Methods* has a discussion of why this is so. Figures 3 and 4 show the titration curve where the mass of genomic DNA present is near enough to zero that this expression reduces to

$$\frac{t}{t_0} = (\frac{1}{a})\, R.$$

As discussed in *Methods*, this expression is simply a statement that the total mass of genomic DNA hybridized is equal to the total mass of probe hybridized. At higher masses of genomic DNA, the mass of the cold competing strand of the genomic DNA becomes significant, and hence, the mass of genomic DNA hybridized is greater than the mass of the probe hybridized as measured by the number of double-stranded ^{32}P counts. In the linear range, $\frac{1}{a}$, the mass fraction of the genome which hybridizes to the probe is just the slope of the line obtained when measured values for t/t_0 are plotted against the known values of R. Figures 3 and 4 show the least-squares fit of our data. The mass fraction of the BALB/c genome reacting with V_H J558 is 9.4×10^{-5} and the mass fraction reacting with $D^d\alpha 3$ is 5×10^{-6}. Figure 5 is a replotting of both sets of data on the same graph. The mass fraction of the genomic DNA reacting with the probe is related to the number of genes in the following way:

$$N = \frac{\frac{1}{a} \times L_G}{L_P}$$

where N is the number of genes, $\frac{1}{a}$ is the mass fraction, L_G is the haploid length of the mouse genome, and L_P is the length of the probe. From the probe sequences shown in Figure 2, we know that L_P for J558 V_H is 267 ntp and L_P for $D^d\alpha 3$ is 439 ntp. We take the value for L_G to be 3×10^9 ntp. Substituting our experimentally determined values for $\frac{1}{a}$ and the known values listed above, we arrive at the result

that N = 1057 ± 33 genes standard error for the J558 V_H-like family and N = 38 ± 2 genes standard error for the class I family. The value of 38 ± 2 genes arrived at for the class I family corresponds well with the minimum value of 33-36 (3, 7, 8) obtained from the cloning experiments done by others.

A significant source of error in the determination of the size of the J558 V_H family is breakage of the genomic DNA during shearing within the sequence of interest in such a way as to generate two detectable copies instead of one. This gives an overestimation in the measurement of family size. The chances of this happening depend strongly on the length of the probe used to detect the family members (1). We believe this source of error to be negligible in our experiment because the J558 V_H probe is relatively small (267 nt) compared to the size of the genomic DNA (500-600 nt). At this size, particularly considering that the average homology of duplexes involving J558 V_H is 76% (see Figure 7), it is as likely that breaks in the genomic DNA will lead to less signal for the hybridizing copy or even no signal at all. More important, however, is the fact that our measurement for the class I family is not significantly elevated even though the class I probe, $D^d\alpha 3$, is more than 50% longer than J558 V_H. Furthermore, the average homology of the four class I third domains sequenced to date, K^d, L^d, 27.1, and 17.3A, as compared to the D^d third domain is 94% (2-6). Therefore, fragments of genomic sequences homologous to $D^d\alpha 3$ would be much more likely, on the average, to be counted as full sequences than fragments of genomic sequences homologous to J558 V_H.

A second significant source of error is the possible underestimation of the J558 V_H family size either because the duplexes are so mismatched that they hybridize very slowly and are not at kinetic termination when the reactions are stopped or because their melting temperature is near enough to the reaction temperature itself that many other related family members were not detected. We allowed many reactions to continue to 30X $Cot_{\frac{1}{2}}$ or even 100X $Cot_{\frac{1}{2}}$ and showed that

the data obtained fit the same line as did data obtained at 10X $Cot_{\frac{1}{2}}$. Because this result shows that all reactions did indeed proceed to completion, we expect no significant underestimation in the J558 V_H family size due to slow reaction rate. We will discuss the measurement of the average thermal stability of hybrids involving J558 V_H with reference to underestimation of J558 family size later. In summary, the data presented in Figures 3, 4 and 5 demonstrate that while the family size of class I is 38 ± 2 genes, that of J558 is 1057 ± 33 genes. These results are very unlikely to be overestimated because of the shearing of the genomic DNA or underestimated because of slow reaction rate.

Figure 6 is a dot blot titration of excess J558 V_H probe with BALB/c liver DNA compared with a dot blot titration of excess $D^d\alpha 3$ probe. Each spot on each array contains a total of 2048 ng of genomic DNA. From left to right, the spots have an increasing amount of BALB/c liver DNA ascending in powers of two from 1 ng in the upper left corner. The balance of the DNA in each spot is made up with salmon sperm DNA. The single-stranded probes were made in parallel as described in *Methods*. J558 V_H was present in fivefold sequence excess over the J558-like copies in the BALB/c genomic DNA. $D^d\alpha 3$ was present in 300 fold sequence excess. Both blots were hybridized in parallel at the same criterion used for the probe excess titration experiments to approximately 6X $Cot_{\frac{1}{2}}$ of their respective probe drivers. The last three spots on the lower right of each array contained only 2048 ng salmon sperm DNA, and hence are the background controls. Since the probes were made to the same specific activity, and the two arrays were exposed on the same piece of film, we can compare their relative intensities directly. The J558 V_H array has 8-16 times more signal than the $D^d\alpha 3$ array. It is also important to note that the $D^d\alpha 3$ probe is 1.6 times as long as the J558 V_H probe. This experiment shows that the family seen by the J558 V_H probe is approximately 13-26 times as large as that seen by $D^d\alpha 3$. Knowing that $D^d\alpha 3$ hybridizes to 38 ± 2 members from the probe excess

titration data shown in Figure 4, we estimate that J558 V_H is hybridizing to a family of genes whose size is approximately 500-1000. These data are less quantitative than those obtained from the probe excess titration experiment, but they represent an independent demonstration of the same result.

Thermal stability of duplexes formed between J558 V_H and genomic sequences

Figure 7 shows the thermal stability profiles of duplexes formed between J558 V_H and genomic sequences under conditions of probe excess at 10X $Cot_{\frac{1}{2}}$ and under conditions of large genomic DNA excess at a genomic Cot of 30,000. The probe excess reaction represents the point (Δ) at R = 2800 on Figure 3 set up in duplicate and run in parallel. In both cases, a 100% homologous control consisting of single-stranded J558 V_H probe reacting with its parent template, the single-stranded M13 clone was run and melted in parallel. We found the T_m of both 100% homologous control duplexes to be 90°C. The T_m of duplexes formed by J558 V_H driven by genomic DNA was 76°C and the T_m of duplexes formed by J558 V_H at R = 2800 on the titration curve was 66°C. Using the approximation that there is a 1° drop in thermal stability for each percent mismatch in the duplexes melted (9), it is apparent that the average homology of sequences reacting with J558 V_H at a genomic Cot of 30,000 is 86%. Similarly, sequences reacting with J558 V_H at R = 2800 under conditions of threefold probe excess have an average homology to J558 of 76%. From these data, we conclude that the approximately 1000 members of the J558 family have an average homology to the J558 sequence of 76%. Since the homology of the 1000 members of the J558 family is low with respect to the J558 sequence, it is likely that, had we measured the number of family members at a criteria lower by 10°, we would have found significantly more members. It is quite possible, therefore, that our measurement of 1000 sequences related to the J558 sequence is an underestimate of the total family size and represents an artificial cut-off due to the stringency of the criterion used. We note that the sequences hybridizing to the

J558 V_H probe under conditions of large genome excess are, on the average, 10% more homologous to the J558 V_H sequence. We will return to this observation later. We do not know how many members of the J558 family these more homologous members represent.

Displacement of hybridized J558 V_H by branch migration

Figure 8 is the data obtained at R values where the concentration of genomic DNA is no longer negligible. All symbols correspond to those shown in Figure 3. In fact, the linear region of positive slope at lower R values is the same data as shown in Figure 3. At similarly high R values, we find that the curve generated by the $D^d\alpha 3$ probe is that described by the equation $\frac{t}{t_0} = \frac{1}{1+(a/R)}$ (data not shown). However, reactions involving J558 V_H at high R values show sudden, large loss of signal always near an approximately equal value of genomic DNA Cot. Although we cannot assay this genomic reaction directly, we find these results consistent with the notion that the unlabeled strand of genomic DNA corresponding to that of the probe can displace by branch migration the labeled probe strand from the duplex, thus resulting in the lowering of ^{32}P-label in duplex. We note that the duplexes formed by the probe at low R values are highly mismatched—24% by our T_m data. On the other hand, the duplexes formed with the probe under conditions of high genomic DNA excess are on the average only 14% mismatched and therefore represent probe hybridization to a largely different population of J558 family members. These are probably the subpopulation of the large J558 family detected initially at low R values which is much more homologous to the J558 V_H gene. These observations imply that the members of the large J558 family detected at low R values are on average more homologous to each other than to the J558 V_H probe. In other words, the J558 sequence is not close to the consensus sequence for this large family. It is possible that, excluding the portion of the family which is on average 86% homologous to J558, the other family members are organized as a single large family whose

members are all greater than 76% homologous to each other. It is also possible that the other members are organized as several subfamilies whose members are highly homologous to each other with considerably less homology between members of different subfamilies. For reasons discussed later, we find the second alternative more consistent with our observations. In summary, we estimate the size of the J558 V_H gene family to be approximately 1000. These genes have an average homology to the J558 V_H gene of 76%, but have a greater average homology to each other.

Estimate of the size of the J558 gene family in the a, b, c, and e allotypes by genome blotting experiments

Figure 9 is an example of the data we obtained from hybridizing EcoRI or HindIII digests of the DNAs from a number of different mouse strains and substrains to single stranded J558 V_H probe. The last four lanes on the right are 100% homologous copy number controls; that is, they are known amounts of the J558 V_H gene. Each of the five bands in a given lane corresponds to the same number of J558 genes. Thus, the one copy number lane contains 0.6 pg of J558 V_H sequence. This is the mass of the DNA in one 300 base pair sequence in a genome of length 3×10^9 bp when 6 µg of the genome is loaded on the gel. Similarly, the 3 copy lane contains 1.8 pg of J558 sequence in each band, the 10 copy lane contains 6 pg of J558 sequence in each band and the 30 copy lane contains 18 pg of J558 sequence in each band. In order to insure that these copy number controls transferred to the blot with the same efficiency as a given gene in the mouse genome, these lanes also contained 6 µg of EcoRI digested salmon sperm DNA. Figure 10 shows the densitometric traces for the BALB/c By, C57L/J, A/J, DBA/2, and SJL/J lanes. Traces of these lanes as well as those for the 3 and 10 copy number control lanes were digitized and the total area under all the peaks on the trace calculated. The copy number control lanes were used to arrive at an average area per 100% homologous gene present in the

genome. The total area in each trace was then divided by this average number. The results of these calculations are listed above each trace as the number of J558 genes. Figure 11 shows the effect of homology of the duplexes formed on the signal. The homology of $\mu 2V_H$ to J558 V_H is known to be 80% by sequence (10). By digitizing the densitometric traces of these two bands, we estimate that the signal of the 80% homologous gene is approximately twofold less than that of the 100% homologous gene. Since we know that the average homology of the genes hybridizing to J558 V_H is 76%, we are justified in multiplying the apparent number of genes based on the signal intensities of the copy number controls by at least a factor of two. These numbers are given in the upper right corner of each trace in Figure 10 as the minimum total number of V_H genes in each lane.

A source of error in this experiment is the difficulty in deriving peak sizes from the densitometric traces because many peaks lie close together. We resolved the trace into its component peaks by finding the midpoint in the signal level between two adjacent peaks and drawing in the sides of the peaks extending through that point to the base line. Obviously, this is only an approximate method. Another source of error in this experiment has to do with how the rate of hybridization to a sequence bound to nitrocellulose varies with its degree of mismatch to the probe. When we did the probe excess titration experiment, we could control for the limitation of apparent family size by comparing reactions that terminated at 100X probe $Cot_{\frac{1}{2}}$ with those terminated at 10X probe $Cot_{\frac{1}{2}}$. In the experiments involving nitrocellulose, we have only a very approximate idea of the rate of hybridization for perfectly matched sequences and no clear idea of how this rate varies with mismatch. We therefore attempted to calibrate our signal with a gene of close to average homology to the J558 sequence. This is only an approximate comparison, and we expect any remaining error in our estimate to be on the side of underestimation of family size.

In summary, there are 350-550 V_H genes evident on genomic blots of DNAs from mice of the a, b, c, and e allotypes. These genes fall into approximately 35 bands when the DNA is digested with either EcoRI or HindIII. Hence, there are multiple V_H genes in each band. A similar result appears in an experiment designed to be a probe cross reaction control and shown in Figure 3 of a recent report by Brodeur and Riblet (11). The authors added known amounts of each of several cloned V_H genes to genome blots at the 5-10 copy number level. These blots were then hybridized to each one of the V_H probes corresponding to the clones present on the blots. Two of the probes used are of interest to this discussion. One is the heavy chain cDNA from the S107 myeloma whose V_H gene is a member of the phosphoryl-choline (PC) family of V_H genes. Crews et al. (12) demonstrated that the PC family has four members. The other probe of interest is Vdx11, a subclone containing the $\mu 2V_H$ gene discussed earlier (10). Both our genomic blots (data not shown) and those of Broder and Riblet (11) as well as library screens done with this probe (data not shown) indicate that, as one expects from its 80% sequence homology to the J558 gene, the $\mu 2V_H$ sequence detects a V_H family of approximately the same size as that detected by the J558 sequence. On the S107 blot, the relative intensity of the signal of the 100% homologous clone to that of the strongest band in the genomic DNA appears to be about 10 to 1, whereas on the $\mu 2V_H$ (Vdx11) blot the relative intensity of the signal from the cloned V_H gene sequence to each of the major bands in the genomic DNA appears to be approximately equal. Furthermore, the full number of bands (approximately 35) appears to be present. In other words, the relative intensities of what were, in effect, their 5-10 copy number controls compared to genomic DNA appear to be completely consistent with our results.

The observation that the J558 family is made up of multiple V_H genes surrounded by identical restriction sites makes strong predictions about the sequence composition and organization of these genes. If the V_H genes and their flanking

regions in a given band are 75-80% homologous to each other and if the positions of their differences vary randomly, then the chance of 10 V_H genes occurring on the identically sized restriction fragment is vanishingly small. Hence, either the sequences sampled by the EcoRI enzyme vary nonrandomly or the homology of the genes and their flanking regions in a given band is very high. We have repeated this result with several other restriction enzymes, including PstI, which tends to hit inside the coding region of genes homologous to the J558 V_H gene (data not shown). This means that it is unlikely that the coincidence of sites is a consequence of nonrandom variation in the sequences. It seems much more probable that it is due to a high level of homology between the flanking seqeunces surrounding the V_H genes in a given band. Since we know of no reason why the flanking regions of the genes in a given band should be much more highly conserved than the genes themselves, we must conclude that the V_H genes found in a given band have a very high homology to each other extending through their flanking regions. Based on the likelihood of a high degree of homology among the V_H genes of a given band, we suggest that each band corresponds to a small, very closely related subfamily of V_H genes. We further suggest that it is likely that most members of the J558 V_H family are organized as 35-45 subfamilies containing 3-20 members. The homologies between members of the same subfamily must be near 100%, with an unknown degree of homology among subfamilies. The average homology of all of the members of each subfamily to those of all other subfamilies is probably at least 76%.

The blot shown in Figure 9 also demonstrates numerous polymorphisms involving multiple genes among the four allotypes tested. We compared all combinations of the EcoRI band patterns of four allotypes a, b, c, and e pairwise. Of a total of 39 positions at which there was at least one band, the comparison between C57L/J(a) and A/J(e) showed 18 polymorphic positions. Similarly the ratio of polymorphic band positions to total band positions was 21/39 for the comparison

between C57L/J(a) and SJL/J(b). For C57L/J(a) vs. DBA/2(c), the fraction of polymorphic band positions was 25/44. For A/J(e) vs. DBA/2(c), the fraction was 12/39. For A/J(e) vs. SJL/J(b) the fraction was 20/40. And, for DBA/2(c) and SJL/J(b) the fraction was 36/46. We will return to the significance of this point in the discussion.

A large family of J558-like V_H genes is present in the rat and hamster genomes

Figure 12 shows two autoradiograms. Lanes A, B, C, and D are from one autoradiogram overexposed to show faint bands in orangutan and chimpanzee DNA. Lane A is the overexposed EcoRI digest of BALB/c J liver DNA, lane B is EcoRI digested rat liver DNA, lane C is EcoRI digested orangutan DNA, and lane D is EcoRI digested chimpanzee DNA. This blot was hybridized at the same criterion as the blot involving only mouse DNA shown in Figure 9. All lanes had 3 μg of DNA. Lanes E and F are from a different autoradiogram. Lane E is 2 μg EcoRI digested hamster DNA and lane F is 3 μg EcoRI digested BALB/c DNA. These autoradiograms show that both rat and hamster DNA contain many sequences that hybridize to J558 V_H. We also find faint bands in orangutan and chimpanzee DNA. Under these same conditions, we see faint bands in human, cat, and cow DNA as well. Negative lanes on these blots were salmon sperm DNA, cerebratulus lacteus (marine worm) DNA, and thyone (sea cucumber) DNA (data not shown). From these results we conclude that rodent genomes have many sequences related to the J558 V_H sequence, and that primates also have a number of more distantly related sequences. Since we also see bands in cow and cat DNA, it is likely that at least several distantly related members of this family are present in all mammals. It would be interesting to see, with the appropriate probes, how large these families really are in rodents other than mouse, and in primates.

The genes visible on genome blots map to the heavy chain variable region locus of chromosome 12

Figures 13, 14, and 15 show EcoRI digested mouse DNAs from 25 of the BXD (C57BL6 x DBA/2) Bailey (13) recombinant-inbred lines. These recombinant inbred lines were derived from crossing the two unrelated, highly inbred parental mouse strains, C57BL6 and DBA/2. After the F_2 generation, each line was maintained independently under a regimen of strict inbreeding, thus fixing the chance recombination events occurring in all of the generations following the F_1 as full homozygosity was approached in each line. A total of 10 easily recognizable polymorphisms were scored as either B type or D type in the EcoRI restriction pattern of each DNA. All lines tested showed patterns of bands identical to either the B or D parental type except for BXD19. This line showed mostly a B-type pattern of bands, but two B bands had large intensity differences with the parental B-type pattern. This line also showed at least one D-type band. On the basis of this D band, we broke the locus corresponding to the J558 family (VDX) into two loci, denoted VDX-1 and VDX-2. VDX-1 was scored as D type for BXD19 and VDX-2 was scored as B type. Table 1 shows the results of the gene linkage analysis for several markers on chromosome 12 with respect to VDX-1. NPB, NP, NPID, NP-A, GTE, SA4, and SA2 are all known V_H markers. The results of this experiment place VDX-1 and VDX-2 1 cm ±1 apart and very close to the NP and NPB V_H markers on chromosome 12.

The genes detected by J558 V_H are V_H genes

Our data strongly imply that the genes detected by the J558 V_H probe in the titration experiments and the genome blotting experiments are immunoglobulin V_H genes. The average homology of the 1000 members of the J558 family to the J558 sequence is 76%. Genetic mapping of the J558 family by recombinant-inbred mouse

lines demonstrates that this family maps to the same area of chromosome 12 as other known V_H markers. Sequences of 10 J558 family members obtained with the J558 V_H probe indicate that each of these 10 is a V_H gene (C. Readhead and D. Livant, data not shown).

References

1. Moore, G. P., Scheller, R. H., Davidson, E. H. and Britten, R. J. (1978) Cell **15**, 649-660.
2. Sher, B. T., Nairn, R., Coligan, J. E. and Hood, L. E. (1985) Proc. Natl. Acad. Sci. USA **82**, 1175-1179.
3. Fisher, D. A., Hunt, S. W. and Hood, L. (1985) J. Exp. Med., in press.
4. Moore, K. W., Sher, B. T., Sun, Y.-H., Eakle, K. A. and Hood, L. E. (1982) Science **215**, 679-682.
5. Steinmetz, M., Moore, K. W., Frelinger, J. G., Sher, B. T., Shen, F.W., Boyse, E. A. and Hood, L. E. (1981) Cell **25**, 683-692.
6. Kvist, S., Roberts, L. and Dobberstein, B. (1983) EMBO J. **2**, 245-254.
7. Winoto, A., Steinmetz, M. and Hood, L. (1983) Proc. Natl. Acad. Sci. USA **80**, 3425-3429.
8. Steinmetz, M., Winoto, A., Minard, K. and Hood, L. (1982) Cell **28**, 489-498.
9. Bonner, T. I., Brenner, D. J., Neufeld, B. R. and Britten, R. J. (1973) J. Mol. Biol. **81**, 123-136.
10. Early, P. W., Nottenburg, C., Weissman, I. and Hood, L. (1982) Mol. Cell. Biol. **2**, 829-836.
11. Brodeur, P. H. and Riblet, R. (1984) Eur. J. Immunol. **14**, 922-930.
12. Crews, S., Griffen, J., Huang, H., Calame, K. and Hood, L. (1981) Cell **25**, 59.
13. Bailey, D. (1971) Transplantation **11**, 325.

DISCUSSION

The results of the experiments reported here demonstrate that the murine V_H locus contains at least 1000 V_H genes. Previous estimates of its size (1) ranged from 100 to 400 genes. The most recent estimate of V_H locus size, that of Brodeur and Riblet (2), placed the number of V_H genes at 100 by comparing the restriction site patterns of sequences around the V_H genes detected by several non-cross hybridizing V_H probes. In interpreting their experimental results, the authors made the assumption that each band in any of these patterns corresponded to one V_H gene. We have quantified the number of genes in each band using known amounts of the J558 V_H gene. Our finding is that one band in the restriction site pattern detected by the J558 V_H probe can contain as many as 10 V_H genes 100% homologous to the J558 V_H probe. Since we know that the average homology of the V_H genes hybridizing to the J558 V_H probe is 76% and that a gene 80% homologous to the J558 gene gives two-fold less signal than the J558 V_H gene itself, we conclude that some bands contain as many as 20 V_H genes. We can interpret the observed intensities of the genomic bands hybridized to μ2, a V_H gene 80% homologous to J558, relative to the intensity of the hybridizing band from a known amount of μ2 sequence shown in a recent report by Brodeur and Riblet (2) as consistent with our results. In the titration experiments reported here, we find that approximately 1000 V_H genes hybridize to the J558 V_H probe. This number represents the minimum size for the V_H locus. Measurement of the thermal stability of duplexes formed between the members of the J558 family and the J558 V_H probe shows that these genes have an average homology to the J558 V_H gene of 76%. The finding that genomic members of the J558 family can displace the J558 probe from duplexes with other genomic members of the family suggests that members of the J558 V_H family may have a higher average homology to each other than to J558.

When we hybridize restriction digests of mouse genomic DNA with J558 V_H, we observe at least 350-550 V_H genes depending on the strain of mouse. Furthermore, these genes fall into 35-45 bands of varying intensity. Thus, each band has multiple V_H genes. The probability of several V_H genes each occurring on the identical sized restriction fragment is very small, unless the sequences sampled by the restriction enzymes are varying nonrandomly or the homology of these sequences to each other is very high. We have repeated these results with several restriction enzymes, some of which tend to occur within the V_H coding regions themselves (data not shown). Therefore, it is likely that the clustering of many V_H genes into a relatively few restriction fragments is due to the high degree of homology among the V_H genes and their flanking regions within each discrete fragment size. We have suggested, therefore, that the several V_H genes found in a specific restriction fragment may constitute a very closely related subfamily of V_H sequences. This observation suggests that the sequence structure of the J558 V_H family consists of 35-45 V_H subfamilies having approximately 3 to 20 closely related members with unknown but substantially less sequence homology between subfamilies.

We also observe numerous polymorphisms among the restriction patterns of the a, b, c and e allotypes of mouse. When we compare these patterns pairwise, we find that from 20% to 50% of the time a given position in a pattern is polymorphic. Although the most simple explanation for the differences between allotypes is that each band contains a single V_H gene, we propose that there are multiple V_H genes per band based on our titration results, dot blot results and on our results involving genome blots with copy number controls. Thus, the same polymorphism appears in multiple V_H genes at once. We suggest that these polymorphisms result from recent duplication and deletion events involving several V_H genes at once. The observation that 20-50% of the bands containing multiple V_H genes are polymorphic when we compare the a, b, c and e allotypes pairwise implies that the closely related V_H

genes of the subfamilies contained in these bands are very close or adjacent to one another. We cannot say whether or not this is the case for those subfamilies contained in the nonpolymorphic bands. Hence, we cannot account for the multiplicity of V_H genes involved in each discrete polymorphism solely by random point mutation events in and around these V_H genes.

Instead, the multiplicity of V_H genes in each band may result from recent duplication events in this V_H family. Both amplification during DNA replication or unequal sister chromatid exchange during mitosis or meiosis could give rise to extensive duplication events involving many V_H genes and their flanking regions at once. These events must occur in the germ cells; therefore, somatic gene amplification events such as those involving the dihydrofolate reductase gene (3) or the *Drosophila* chorion genes (4) may not constitute a model system for the evolution of the genes in the V_H locus. Recent observations involving partial deficiencies at the rDNA loci of *Drosophila*, however, have revealed a high rate (7×10^{-3}/locus-generation) of large-scale loss or gain of rDNA genes occurring by unequal sister chromatid exchange during the meiotic stage of spermatogenesis (5). Apparently, in a tandemly repeated family with approximately 250 copies, such events can occur during the generation of germ cells with a high frequency. We propose that similar events could involve the J558-like V_H family relatively often and with large-scale results because of its size and the similarity of its members to one another. These events would result in both reductions and amplifications of the size of the J558 family. We realize that work by Huang (6) has shown that the frequency of recombination events between two homologous sequences depends strongly on their relative homologies. Those measurements were made on V_H sequences in *E. coli*. We do not know what the corresponding results would be in mouse germ cells. Furthermore, we must consider that we do not know what sequences in the V_H locus mediate these putative recombination events. Short, highly homologous sequences interspersed

among the V_H genes, for example, could be the targets for these events rather than the V_H genes themselves.

It is significant that the J558 family seems to be approximately the same size in all four mouse allotypes examined by genome blotting with copy number controls. Many members related to the J558 V_H gene are also evident in hamster and rat. In fact, even at the relatively elevated criterion used to detect closely related sequences in rodents, we find at least several J558-related sequences in primate DNA, as well as cat and cow DNA (data not shown). We do not know how many of these V_H genes are of functional signficance; nevertheless, it seems that the J558 V_H family is evolutionarily conserved among mammals.

In summary, we can account for both the multiplicity of V_H genes in a given restriction fragment and the involvement of many V_H genes in each restriction fragment polymorphism evident among mice of the a, b, c, and e allotypes by supposing that events magnifying or reducing large areas of the J558 V_H family have occurred at least several times in the evolutionary history of the mouse. It is possible that there is a strong selection on mice to have a V_H family of this size because losses of large numbers of family members seem to be followed by an amplification of the remaining members. This results in the skewing of the specific V_H sequences represented in each strain of mouse and coincides with the observation that each mouse strain has its own set of hybridizing restriction fragments. A large body of evidence from both comparison of the set of antibody sequences represented by BALB/c and NZB myelomas (7) and from the strain specificity of most of the known idiotypes (8) is consistent with the idea that many of the V_H genes represented in each mouse strain are not represented among all the strains.

The notion that mice of different allotypes survive with similar although not identical sets of homologous, germline V_H genes suggests that there is a great deal of degeneracy in how the information required for antigen recognition is encoded.

The most obvious source of this degeneracy is the well-known cross-reaction of a given antibody to many different antigens. The observation that this particular set of germline V_H genes is always large suggests that it might be the size of the locus and the overall similarity of its members which are important, not the presence or absence of a particular member.

Perhaps a significant fraction of the antigens encountered naturally by the mouse are bound by antibodies whose heavy chains are of overall similar shape but have a relatively random degree of variation primarily in regions involving antigen contact. In these regions, small numbers of changes make a big difference in how well a particular antibody binds a given antigen. More importantly, we suggest that the large size of the V_H locus and, at a limited number of positions, the random variations among its members mirror the large number and random variations in the shapes of the environmental antigens encountered by the mouse. Many pathenogenic antigens may evolve too rapidly to be of use in selecting specific V_H genes during evolution. It may therefore be necessary to maintain a large library of similar but randomly variable germline V region sequences in order to insure that each time a new antigen is encountered, antibodies capable of binding the antigen can be made. In this model, the advantage is in having a minimum number of germline sequences sufficiently close in structure yet with enough random variation in critical regions to make it likely that at least a few clones of B cells can be expanded when the antigen is encountered. Only when an important environmental antigen does not vary rapidly, for example, phosphorylcholine, can there be a selection operating at the level of maintenance of a particular V_H gene.

We note that the apparent size of the J558 V_H family in BALB/c germline DNA does not necessarily diminish the importance of mechanisms such as somatic mutation, junctional diversity, and combinatorial joining in creating additional diversity from the germline. We rather favor the interpretation that the existence

of these mechanisms indicates that more diversity than we previously suspected is necessary to make sure that antigen recognition always works, and that the cost of maintaining this additional information in the germline is too high to be beneficial. The size of the germline library of V_H sequences is likely to be governed by the balance of two factors: the cost of maintaining the library and the advantage of maximizing the number of opportunities to mobilize the immune system against an assault by the environment. The known somatic mechanisms producing further variation in antibody structure are thus ways the system has evolved to keep the cost of this maintenance as small as possible.

References

1. *The Antibody Enigma.* Kindt, T. and Capra, J. D. (1984) Plenum Press, New York and London, pp. 223-261.

2. Brodeur, P. H. and Riblet, R. (1984) Eur. J. Immunol. **14**, 922-930.

3. Brown, P. C., Kaufman, R. J., Haber, D. and Schimke, R. T. (1982) In: *Gene Amplification*, R. T. Schimke, ed., Cold Spring Harbor, pp. 9-15.

4. Spradling, A. C. (1982) In: *Gene Amplification*, R. T. Schimke, ed., Cold Spring Harbor, pp. 121-129.

5. Hawley, R. S. and Tartof, K. D. (1985) Genetics **109**, 691-700.

6. Huang, H., personal communication.

7. Loh, E., Hood, J. M., Riblet, R., Weigert, M. and Hood, L. (1979) J. Immunol. **122**, 44-48.

8. Rajewsky, K. and Takemori, T. (1983) Ann. Rev. Immunol. **1**, 569-607.

SUMMARY

We have demonstrated that the V_H locus in mice encompasses at least 1000 V_H genes by measuring the size of the J558 V_H gene family. We estimated the size of the J558 family by titrating the single-stranded J558 V_H coding region probe with increasing amounts of sheared liver DNA from BALB/c J mice. This experiment showed that 1057 ± 33 V_H genes hybridized to J558 V_H at a criterion equivalent to 0.12 M PB, 60°C. As a control for systematic error in our measurement, we titrated the single-stranded D^d third domain probe by the identical method in parallel. We found that 38 ± 2 genes hybridized to $D^d\alpha 3$ in agreement with previously published results (1-3). We then showed that the intensity of signal of a dot blot having varying amounts of BALB/c J DNA hybridized to J558 V_H as compared to the signal intensity of an identical dot blot hybridized to $D^d\alpha 3$ was consistent with a J558 family size of 500-1000 members. Finally, we demonstrated that approximately 500-600 V_H genes were hybridizing to genome blots of EcoRI-digested BALB/c J DNA based on J558 V_H copy number controls. We estimated the average homology of the members of the J558 family to the J558 V_H sequence to be 76% by measuring the thermal stability of duplexes formed between the J558 V_H gene and members of the J558 family.

We observed that each band on genomic blots of DNA from the BALB/c, C57L/J, DBA/2, A/J, SJL/J, and C57BL10/J mouse strains contained multiple V_H genes, and that many polymorphic bands containing multiple V_H genes were evident among the four allotypes tested. Noting these observations, we predicted that V_H genes from a given band have a very high degree of homology to each other. The high degree of homology among V_H genes from a given band and the observation that the many polymorphisms among the mouse strains tested involve multiple V_H genes at once imply that at least several large-scale duplication and deletion events have occurred during the existence of the mouse as a species. We speculate that

maintenance of a large V_H family may be advantageous because many pathogeneic antigens evolve too rapidly to be of use in selecting specific V_H genes during evolution, and that selection is therefore chiefly for the maintenance of a large set of somewhat homologous V_H genes and not generally for the maintenance of a particular V_H gene.

References

1. Fisher, D. A., Hunt, S. W. and Hood, L. (1985) J. Exp. Med., in press.

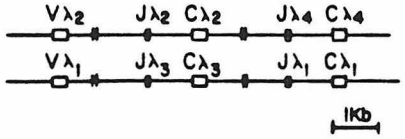
2. Winoto, A., Steinmetz, M. and Hood, L. (1983) Proc. Natl. Acad. Sci. USA **80**, 3425-3429.

3. Steinmetz, M., Winoto, A., Minard, K. and Hood, L. (1982) Cell **28**, 489-498.

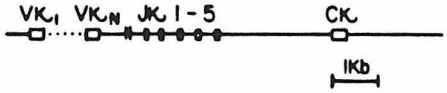
Figure 1. Structure of the mouse germline λ, κ, and H immunoglobulin families and of a joined IgM gene.

Figure 1a

λ Light Chain



κ Light Chain



Heavy Chain

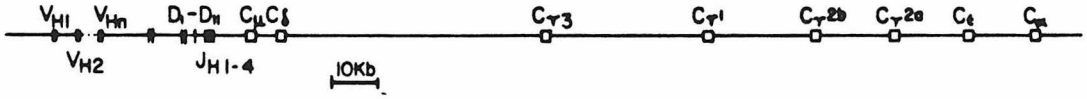


Figure 1b

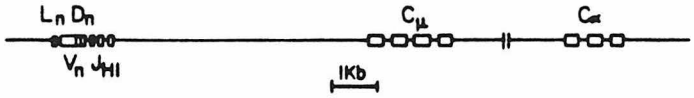


Figure 2. Sequences of the J558 V_H probe containing the J558 coding region and the $D^d\alpha 3$ probe containing the D^d third domain.

Figure 2. Sequences of J558 V_H and $D^d\alpha3$ Probes

J558 V_H

Leu	Gln	Gln	Ser	Gly	Pro	Glu	Leu	Val	Lys	Pro	Gly	Ala	Ser	Val
CTG	CAA	CAA	TCT	GGA	CCT	GAG	CTG	GTG	AAG	CCT	GGG	GCT	TCA	TGT
Lys	Met	Ser	Cys	Lys	Ala	Ser	Gly	Tyr	Thr	Phe	Thr	Asp	Tyr	Tyr
AAG	ATG	TCC	TGT	AAG	GCT	TCT	GGA	TAC	ACA	TTC	ACT	GAC	TAC	TAC
Met	Lys	Trp	Val	Lys	Gln	Ser	His	Gly	Lys	Ser	Leu	Glu	Trp	Ile
ATG	AAG	TGG	GTG	AAG	CAG	AGT	CAT	GGA	AAG	AGC	CTT	GAG	TGG	ATT
Gly	Asp	Ile	Asn	Pro	Asn	Asn	Gly	Gly	Thr	Ser	Tyr	Asn	Gln	Lys
GGA	GAT	ATT	AAT	CCT	AAC	AAT	GGT	GGT	ACT	AGC	TAC	AAC	CAG	AAG
Phe	Lys	Gly	Lys	Ala	Thr	Leu	Thr	Val	Asp	Lys	Ser	Ser	Ser	Thr
TTC	AAG	GGC	AAG	GCC	ACA	TTG	ACT	GTA	GAC	AAA	TCC	TCC	AGC	ACA
Ala	Tyr	Met	Gln	Leu	Asn	Ser	Leu	Thr	Ser	Glu	Asp	Ser	Ala	
GCC	TAC	ATG	CAG	CTC	AAC	AGC	CTG	ACA	TCT	GAG	GAC	TCT	GCA	

$D^d\alpha3$

CTCTGCTTTTGGTCACTAGTGCAATGACAGTTGAAGCGTCAAACAGACACAGAGTTCACTGTCATCATTG

ATTTAACTGAGTCTTGTGTAGATTTCAGTTTGTCTTGTTAATTGTGGAATTTCTTAAATCTTCCACACAG

sp	Pro	Pro	Lys	Ala	His	Val	Thr	His	His	Arg	Arg	Pro	Glu	Gly
AT	CCC	CCA	AAG	GCC	CAT	GTG	ACC	CAT	CAC	CGC	AGA	CCT	GAA	GGT
Asp	Val	Thr	Leu	Arg	Cys	Trp	Ala	Leu	Gly	Phe	Tyr	Pro	Ala	Asp
GAT	GTC	ACC	CTG	AGG	TGC	TGG	GCC	CTG	GGC	TTC	TAC	CCT	GCT	GAC
Ile	Thr	Leu	Thr	Trp	Gln	Leu	Asn	Gly	Glu	Glu	Leu	Thr	Gln	Glu
ATC	ACC	CTG	ACC	TGG	CAG	TTG	AAT	GGG	GAG	GAG	CTG	ACC	CAG	GAA
Met	Glu	Leu	Val	Glu	Thr	Arg	Pro	Ala	Gly	Asp	Gly	Thr	Phe	Gln
ATG	GAG	CTT	GTG	GAG	ACC	AGG	CCT	GCA	GGG	GAT	GGA	ACC	TTC	CAG
Lys	Trp	Ala	Ser	Val	Val	Val	Pro	Leu	Gly	Lys	Glu	Gln	Lys	Tyr
AAG	TGG	GCA	TCT	GTG	GTG	GTG	CCT	CTT	GGG	AAG	GAG	CAG	AAG	TAC
Thr	Cys	His	Val	Glu	His	Glu	Gly	Leu	Pro	Glu	Pro	Leu	Thr	Leu
ACA	TGC	CAT	GTG	GAA	CAT	GAG	GGG	CTG	CCT	GAG	CCC	CTC	ACC	CTG
Arg	Trp	Gly	Lys	Glu	G									
AGA	TGG	GGC	AAG	GAG	G GTGAGGCTGCAGAG									

Figure 3. Probe excess titration of J558 V_H with increasing amounts of genomic DNA. Closed and open circles represent two separate experiments reacted to 30X $Cot_{\frac{1}{2}}$ of the J558 V_H driver. Open triangles and closed triangles represent two separate experiments reacted to 10X $Cot_{\frac{1}{2}}$. Open squares represent an experiment run to 100X $Cot_{\frac{1}{2}}$. In the two experiments depicted by the closed and open circles, the mass of genomic DNA present varied from 2.03×10^{-3} µg to 4.06×10^{-2} µg. In the two experiments depicted by the open triangles and open squares, the mass of genomic DNA present varied from 1.3×10^{-2} µg to 7.8×10^{-2} µg. In the experiment depicted by the closed triangles, the mass of genomic DNA present varied from 7.3×10^{-3} µg to 9.13×10^{-2} µg. These experiments were run in 0.41 M PB/0.1% SDS at 67°C to the equivalent Cots indicated. The reaction times were 3.6 hr for the open triangles, 35.8 hr for the open squares, 31 hr for the closed and open circles, and 25 hr for the closed triangles. The ratios of counts bound are expressed on the ordinate as percents. R denotes the ratio of the total genomic mass over the total probe mass. The mass of the probe was constant for all points in a given run. Between runs, the mass of the probe present was adjusted so that the R values of points from different experiments were near each other.

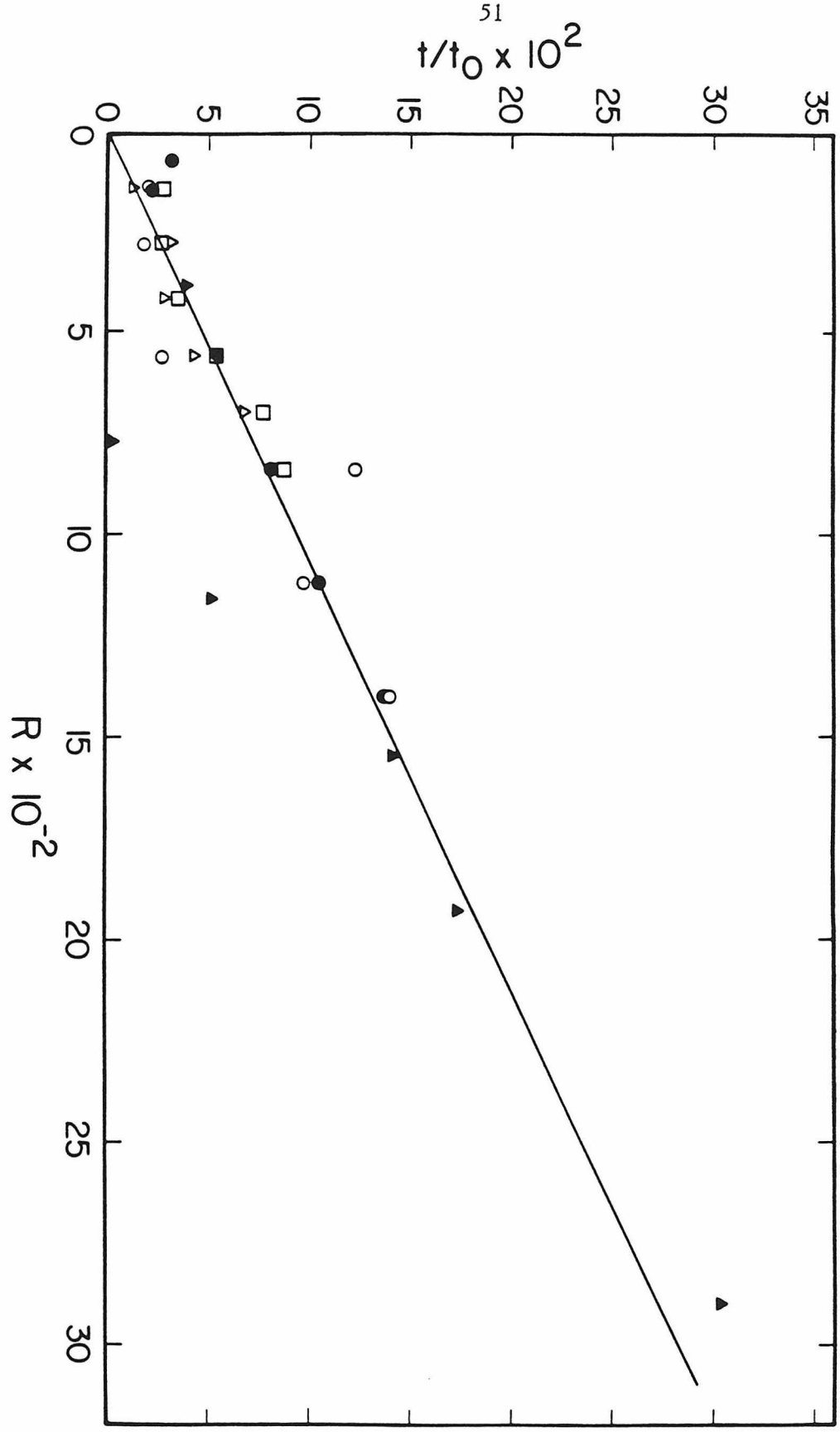

$t/t_0 \times 10^2$
0
5
10
15
20
25
30
35
$R \times 10^{-2}$
0
5
10
15
20
25
30

Figure 4. Probe excess titration of $D^d\alpha 3$ with increasing amounts of genomic DNA. Symbols correspond to those of Figure 3. Open and closed circles represent two separate experiments driven to 30X $Cot_{\frac{1}{2}}$ of the $D^d\alpha 3$ probe. Open triangles correspond to an experiment terminated at 10X $Cot_{\frac{1}{2}}$, and open squares correspond to an experiment terminated at 100X $Cot_{\frac{1}{2}}$. For the two experiments corresponding to the open triangles and open squares, the mass of genomic DNA present varied from 2.18×10^{-1} µg to 17.4 µg. For the experiments represented by the closed and open circles, the genomic mass varied from 2.43×10^{-1} µg to 4.85×10^{-1} µg. Reaction times were 10.8 hr for the open triangles, 54 hr for the open squares, and 31 hr for the closed and open circles. R denotes the ratio of the total mass of genomic DNA over the total mass of probe DNA. The fraction of counts bound is expressed as a percent on the ordinate. In all experimental points of each run, for example, all open triangles, the mass of the probe remained constant while the mass of genomic DNA varied.

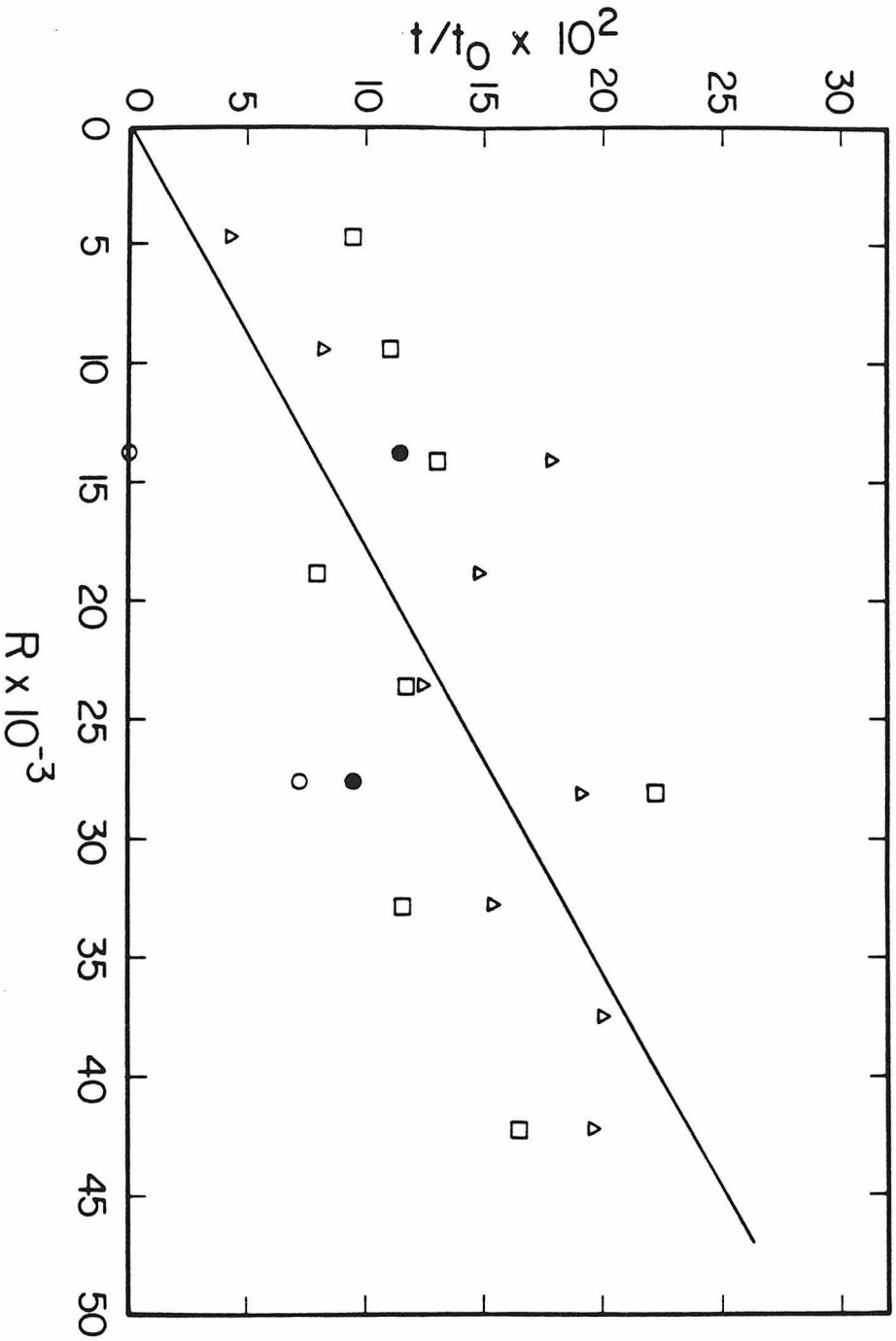

t/t0 x 10^2
0
5
10
15
20
25
30
R x 10^-3
0
5
10
15
20
25
30
35
40
45
50

Figure 5. Results of probe excess titration of J558 V_H as compared to those of $D^d\alpha 3$. Closed symbols correspond to the J558 V_H probe and open symbols to $D^d\alpha 3$. Closed circles represent the data for 30X $Cot_{\frac{1}{2}}$ reactions as described in Figure 3. Closed triangles and closed squares correspond to 10X $Cot_{\frac{1}{2}}$ reactions and 100X $Cot_{\frac{1}{2}}$ reactions, respectively, as described in Figure 3. Closed diamonds correspond to the closed triangles at 10X $Cot_{\frac{1}{2}}$ as described in Figure 3. Open triangles and squares correspond to 10X $Cot_{\frac{1}{2}}$ and 100X $Cot_{\frac{1}{2}}$ reactions, respectively, as described in Figure 4. Open circles correspond to 30X $Cot_{\frac{1}{2}}$ reactions as described in Figure 3. R denotes the mass of genomic DNA over the mass of probe DNA. The abscissa is the percent of total counts bound by HAP as previously described.

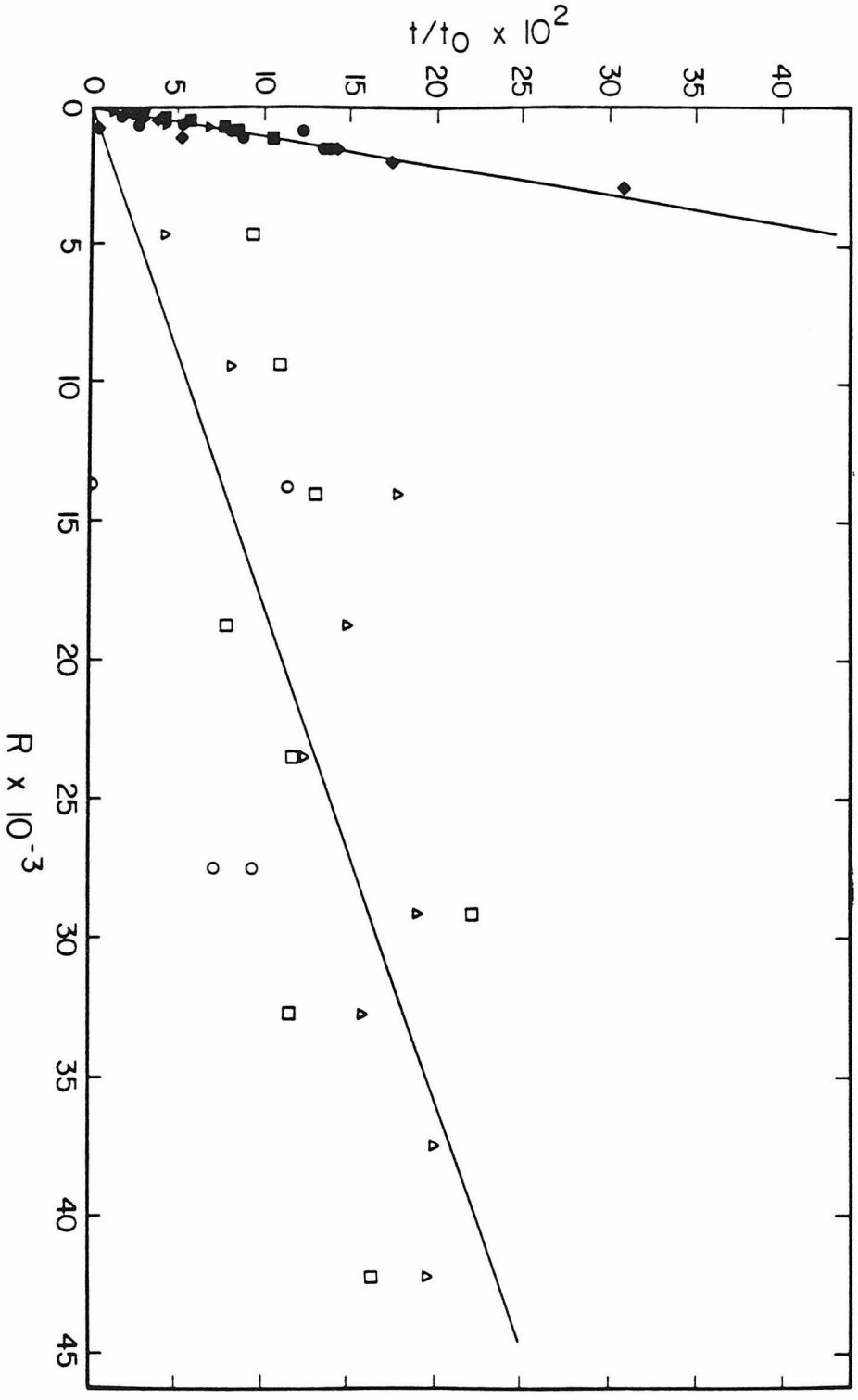

t/t$_0$ x 10^2
0
5
10
15
20
25
30
35
40
R x 10^{-3}
0
5
10
15
20
25
30
35
40
45

Figure 6. Dot blot titration of J558 V_H and $D^d\alpha 3$ probes with increasing amounts of BALB/c J liver DNA. Each array contains 15 spots arranged in order of increasing amounts of genomic DNA from left to right, top to bottom. The amounts of genomic DNA increase in factors of two from 1 ng to 2048 ng. The total amount of DNA in each spot is kept constant at 2048 ng by the addition of salmon sperm DNA. The last three spots on the lower right of each array are salmon sperm DNA only. The two probes were synthesized simultaneously to a specific activity of 2×10^9 counts/μg and made single stranded as described in ***Methods.*** Each hybridization was driven to approximately 5X $Cot_{\frac{1}{2}}$ taking into account the retarded rate of hybridization on nitrocellulose filters. The J558 V_H probe was present to approximately sixfold sequence excess over the hybridizing genomic sequences. The $D^d\alpha 3$ probe was present to approximately 300 fold sequence excess. Exposure was for 24 hr with an intensifying screen at -70°C.

J558 V_H Class I α 3

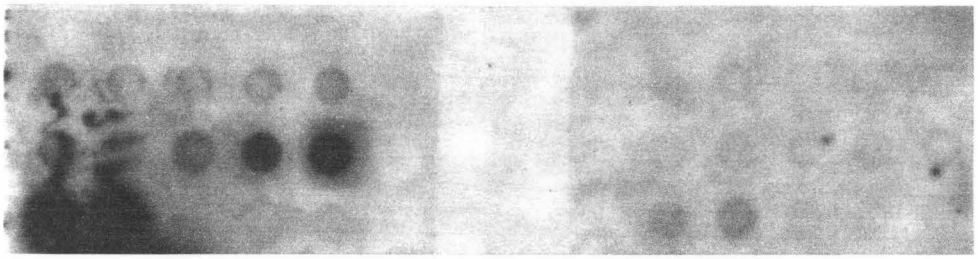


Figure 7. Melting of duplexes involving J558 V_H and genomic sequences by increasing temperature. Open circles represent the melting of a duplicate sample corresponding to the closed triangle at R = 2800 in Figure 3. Closed circles represent the parallel experiment done using the M13 clone containing the template J558 V_H sequence. Hence, this experiment is a calibration using 100% homologous sequence. Open triangles represent the melting of a sample containing a large excess (50 ug) of genomic DNA and driven to a genomic Cot of 30,000. Closed triangles are the corresponding 100% homology calibration run using the J558 V_H M13 template.

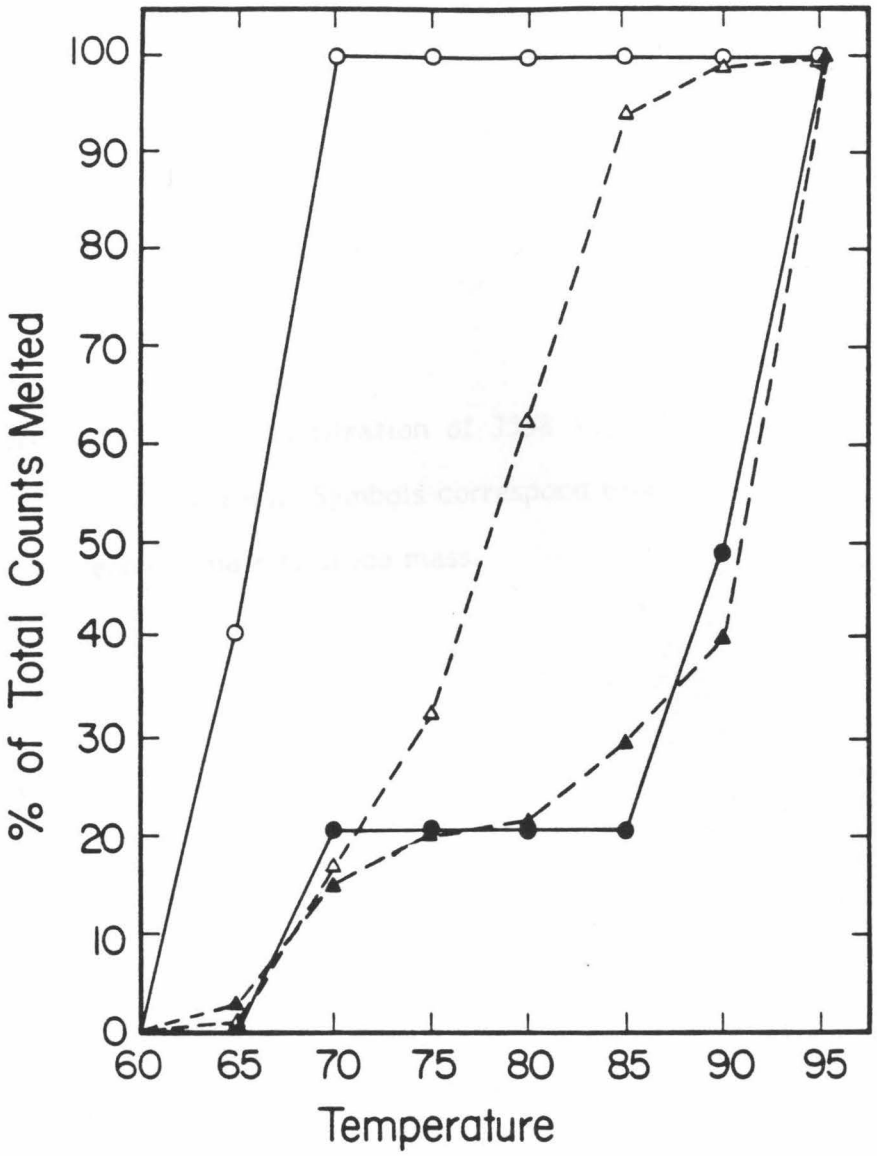

% of Total Counts Melted
100
90
80
70
60
50
40
30
20
10
0
60
65
70
75
80
85
90
95
Temperature

Figure 8. Probe excess titration of J558 V_H with increasing amounts of genomic DNA at high R values. Symbols correspond exactly to those in Figure 3. R is the ratio of genomic mass to probe mass.

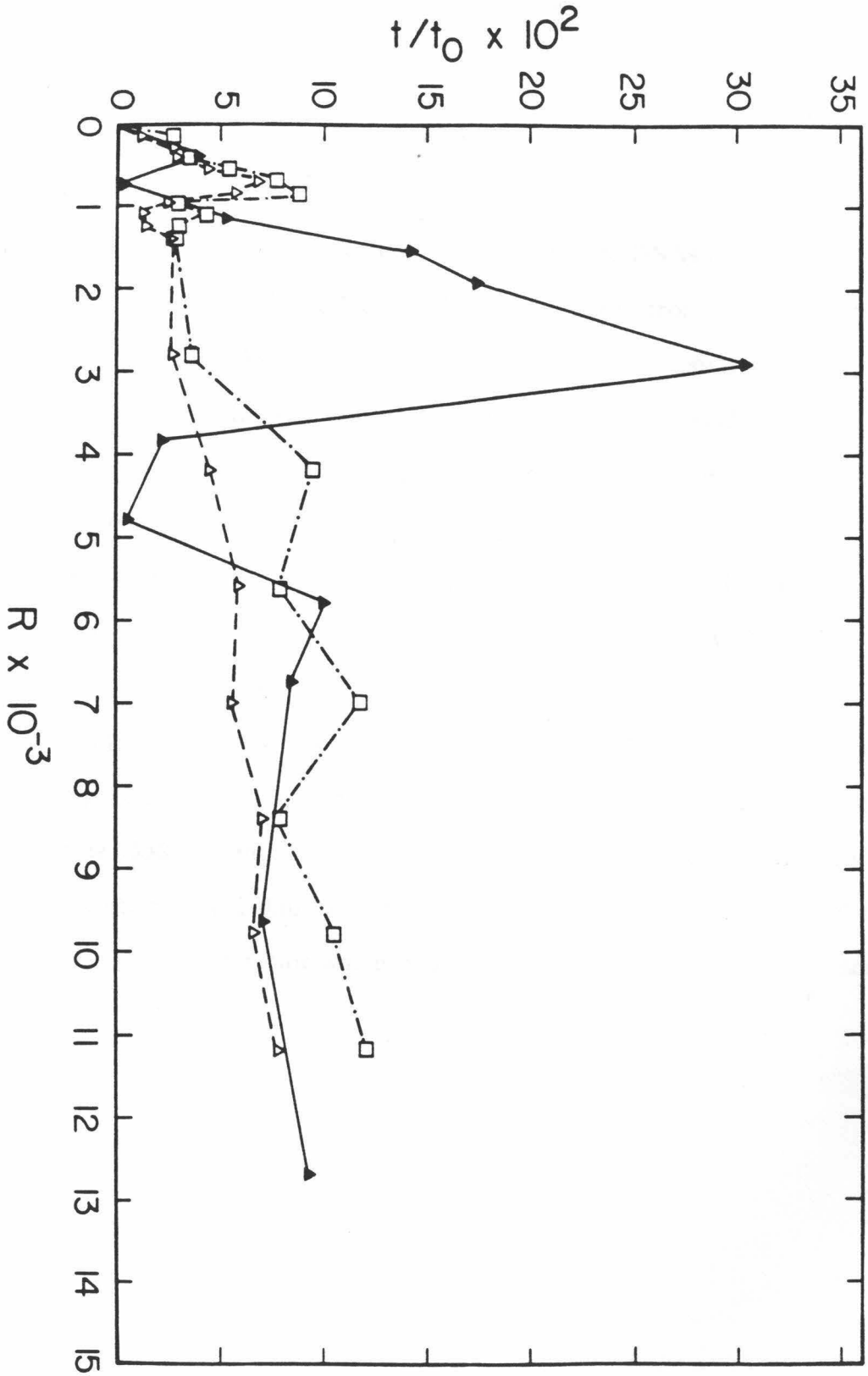

t/t0 x 10^2
0
5
10
15
20
25
30
35
R x 10^-3
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Figure 9. Genomic blot of 6 μg of EcoRI digested mouse DNAs of various strains and substrains. BALB/c By and BALB/c CRGL were prepared from liver as described in *Methods*. All other DNAs were prepared from spleen and obtained from Jackson Laboratories. In parentheses after each strain is the corresponding allotype. The numbers above the J558 copy number control lanes refer to the equivalent number of J558 genes in each band. In the 1 copy number lane, for example, there is 0.6 pg of J558 V_H sequence in each band. In order to make these control lanes, 1 μg of the λgtWes clone containing the joined J558 gene was digested in a 20 μL volume with either EcoRI, BglII, EcoRI + BglII, Hind III or EcoRI + BamHI. 1 μL of each digest was diluted to 1 ml or 100 μL with gel buffer. 2 μL of a 1:1000 dilution corresponds to 100 pg of the digested J558 V_H-containing λ clone. This is equivalent to 0.6 pg of the 300 nucleotide J558 sequence. Similarly, we used 6λ of each 1:1000 dilution for the 3 copy lane, 2λ of the 1:100 dilution for the 10 copy lane, and 6λ of the 1:100 dilution for the 30 copy lane. This blot was exposed for 20 hr with an intensifying screen at -70°C.

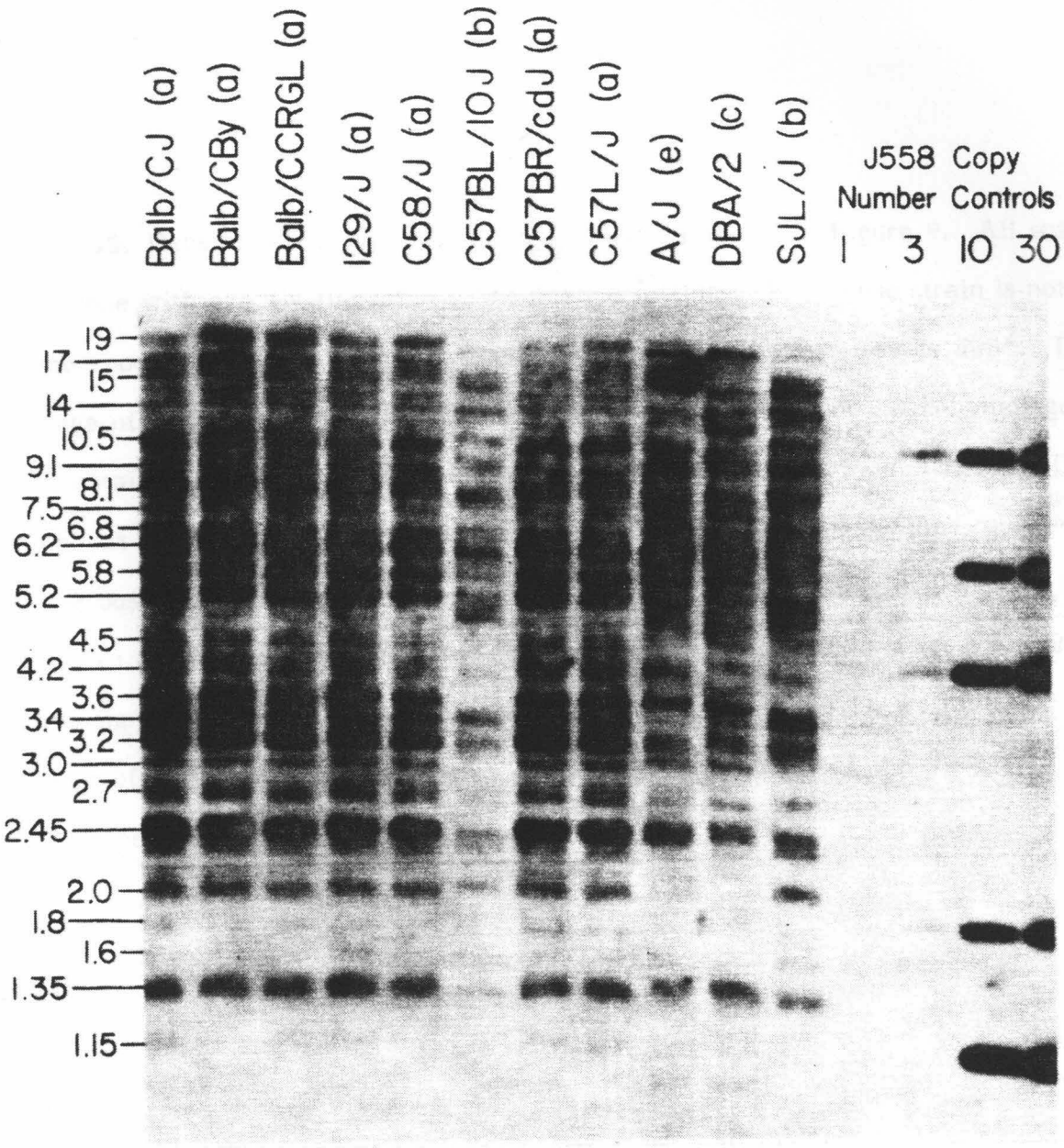
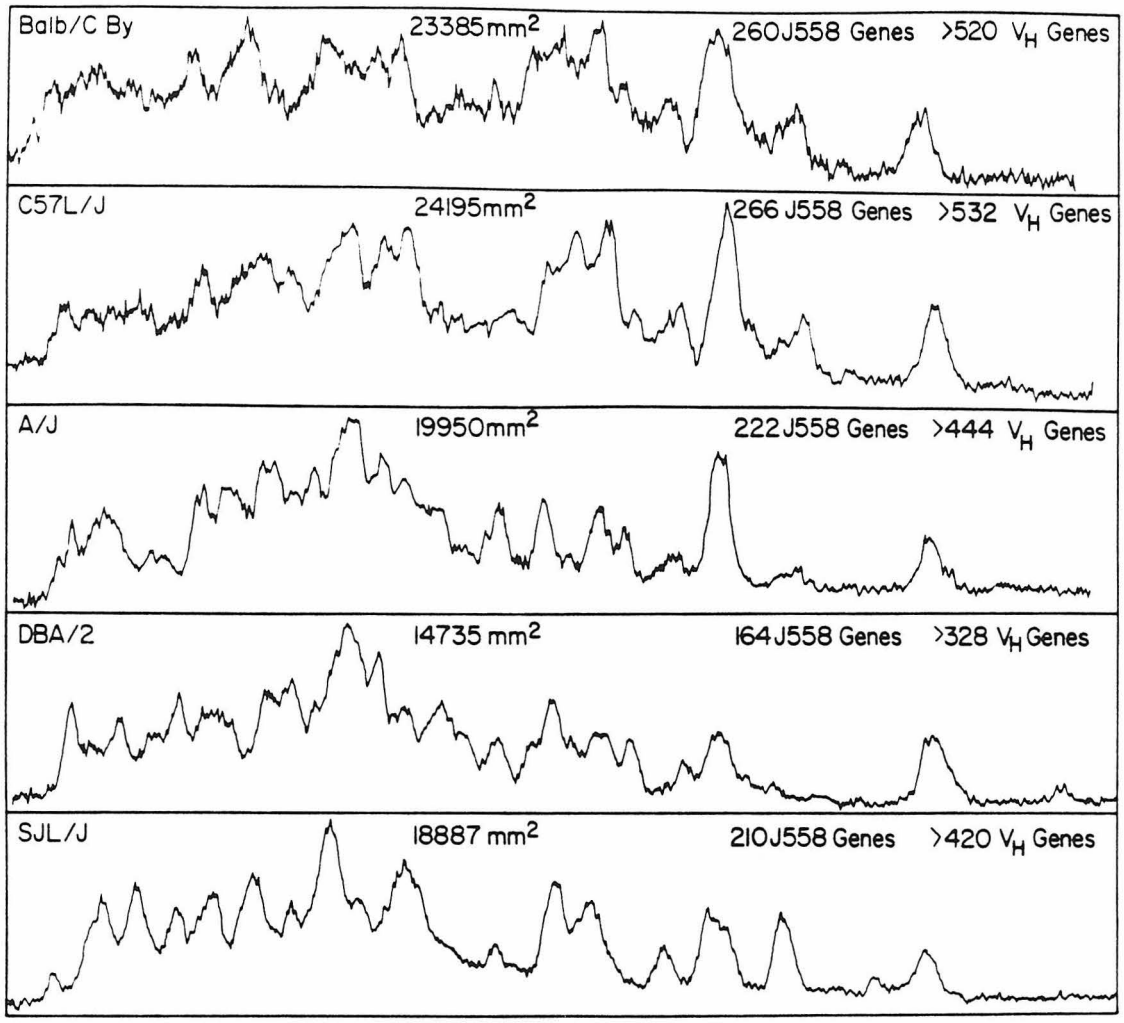

Balb/CJ (a)
Balb/CBy (a)
Balb/CCRGL (a)
129/J (a)
C58/J (a)
C57BL/10J (b)
C57BR/cdJ (a)
C57L/J (a)
A/J (e)
DBA/2 (c)
SJL/J (b)
J558 Copy Number Controls
1 3 10 30
19
17
15
14
10.5
9.1
8.1
7.5
6.8
6.2
5.8
5.2
4.5
4.2
3.6
3.4
3.2
3.0
2.7
2.45
2.0
1.8
1.6
1.35
1.15

Figure 10. Densitometric scans of the genomic blot shown in Figure 9. All scans were done with exactly the same densitometric settings. The mouse strain is noted in the upper left corner. The total area under the profile is noted as mm^2. The average of the two values obtained for the area equivalent to one 100% homologous gene obtained from the 3 and 10 copy number control lanes is 90 mm^2/gene. This value was used to arrive at the number of 100% homologous (J558) genes equivalent to the observed signal in each lane. Since the average homology of genes in this family to the J558 probe is 76%, and since an 80% homologous gene gives a factor of 2 less signal than the 100% homologous gene does (see Figure 1), we multiply the number of J558 genes by 2 to arrive at a minimum estimate of the number of V_H genes appearing on these blots.


Balb/C By
23385mm²
260J558 Genes
>520 V_H Genes
C57L/J
24195mm²
266 J558 Genes
>532 V_H Genes
A/J
19950mm²
222J558 Genes
>444 V_H Genes
DBA/2
14735 mm²
164J558 Genes
>328 V_H Genes
SJL/J
18887 mm²
210J558 Genes
>420 V_H Genes

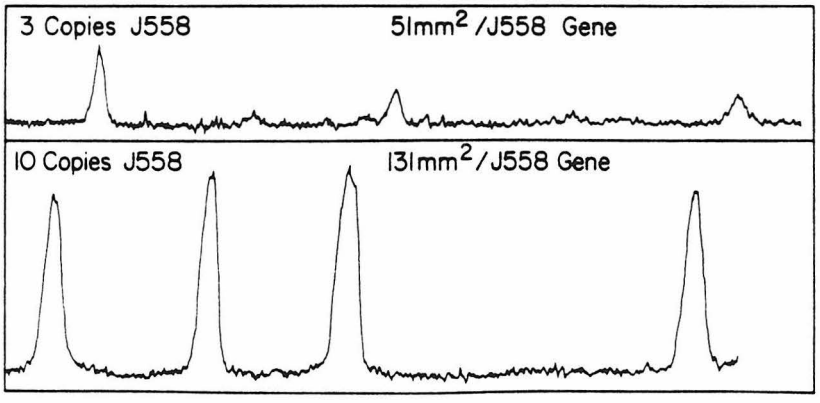

3 Copies J558
51mm²/J558 Gene
10 Copies J558
131mm²/J558 Gene

Figure 11. Densitometric trace of 5 or 10 copies of μ2, a V_H gene 80% homologous to the J558 gene and five or ten copies of the J558 gene itself. 1, 5, 10, and 20 pg of the μ2 and the J558 V_H genes were transferred from a 0.8% agarose gel onto nitrocellulose and hybridized to the J558 V_H probe. The difference in the area under the J558 peak as compared with the μ2 peak is twofold.

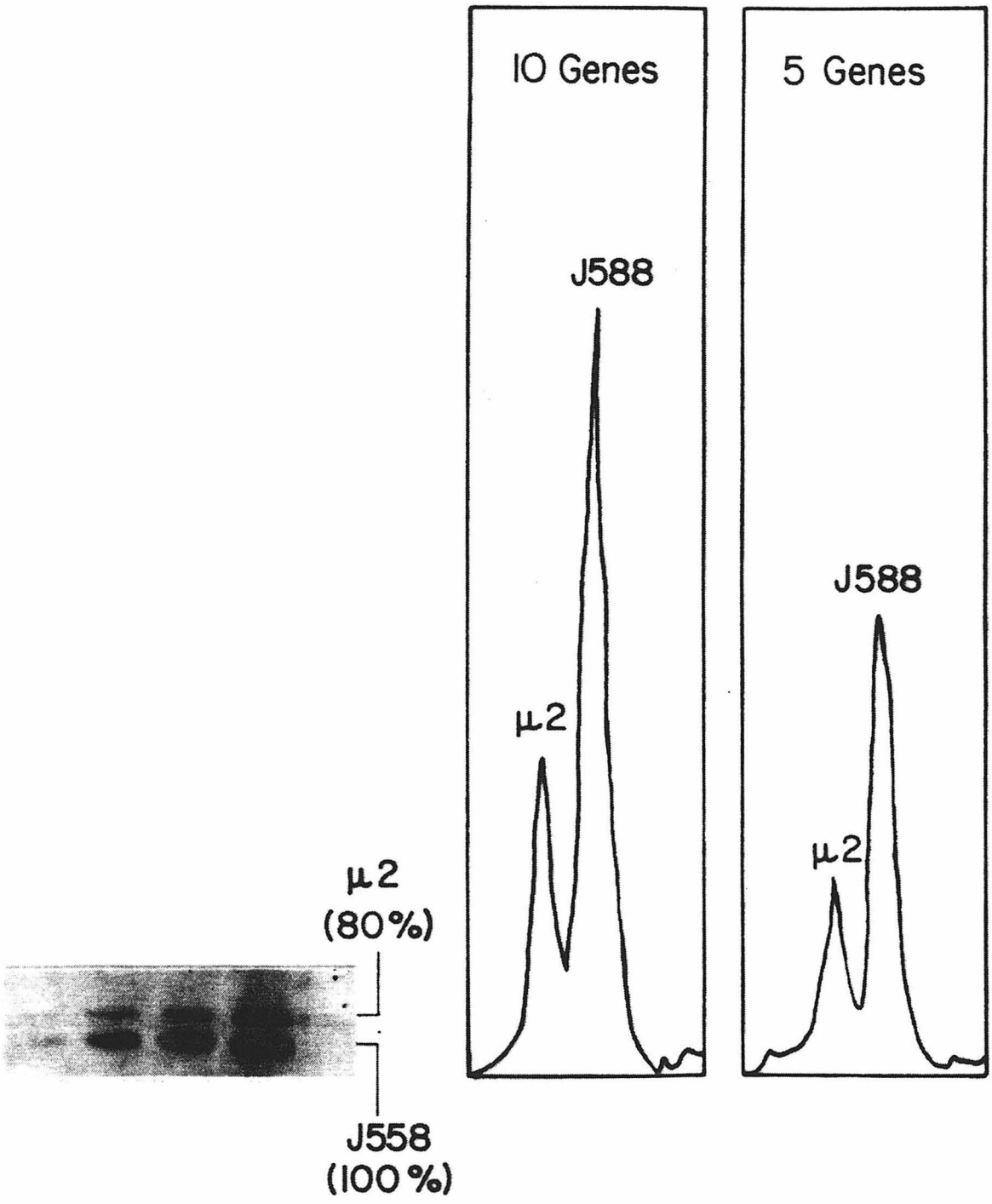

10 Genes
J588
μ2
5 Genes
J588
μ2
μ2
(80%)
J558
(100%)

Figure 12. Lanes A, B, C and D were electrophoresed on the same 0.8% agarose gel and hybridized on the same blot with single-stranded J558 V_H probe. Lane A is 3 µg of EcoRI-digested BALB/c liver DNA, lane B is 3 µg of EcoRI-digested PVG rat liver DNA, lane C is 3 µg of EcoRI-digested orangutan DNA and lane D is 3 µg of EcoRI-digested chimpanzee DNA. This blot was exposed 21 days with an intensifying screen at -70°C. Lanes E and F were electrophoresed on the same 0.8% agarose gel and hybridized on the same blot with single-stranded J558 V_H probe. Lane E is 2 µg of EcoRI-digested hamster DNA and lane F is 3 µg EcoRI-digested BALB/c liver DNA. This blot was exposed for18 hr at -70°C with an intensifying screen.

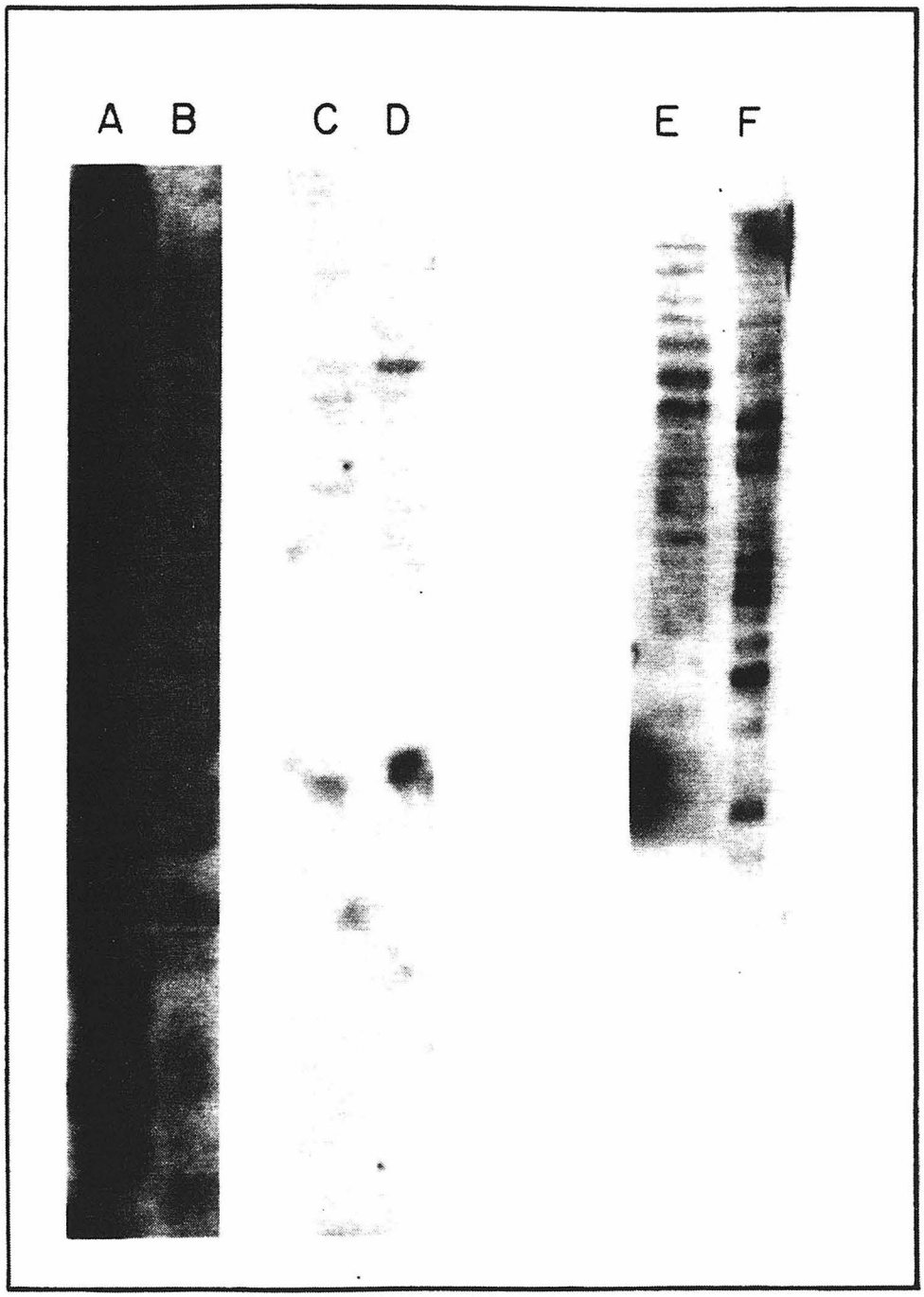

A B
C D
E F

Figure 13. An EcoRI digest of C57BL/6, DBA/2, and the BxD recombinant inbred mouse lines 1-13. 6 μg of DNA were present in each digest. These digests were resolved on a 0.8% agarose gel, transferred to nitrocellulose and hybridized to single-stranded J558 V_H probe.

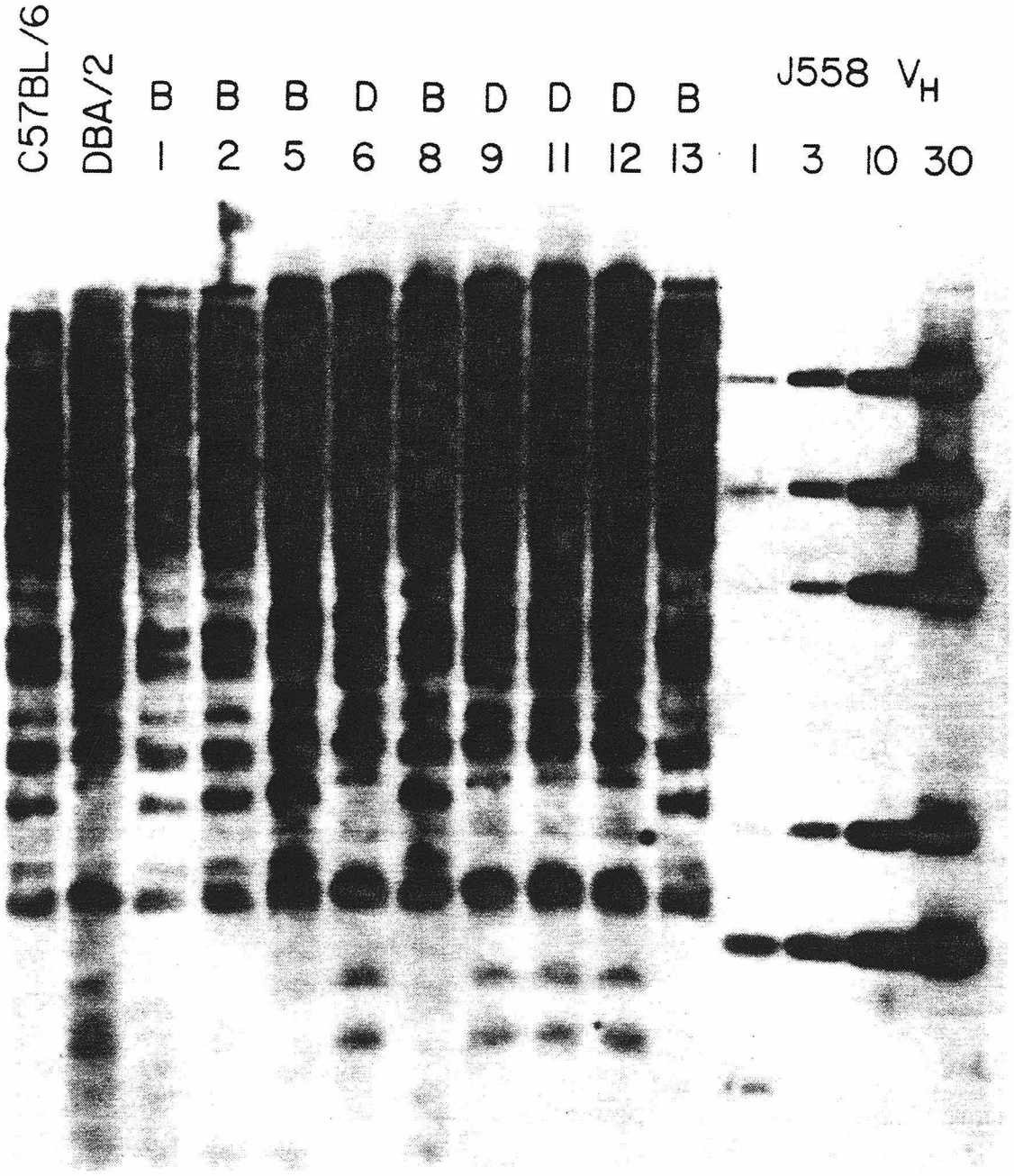

C57BL/6
DBA/2
B 1
B 2
B 5
D 6
B 8
D 9
D 11
D 12
B 13
J558 V_H 1
3
10
30

Figure 14. An EcoRI digest of C57BL/6, DBA/2 and the BxD recombinant inbred mouse lines 14-23 treated as described in Figure 13.

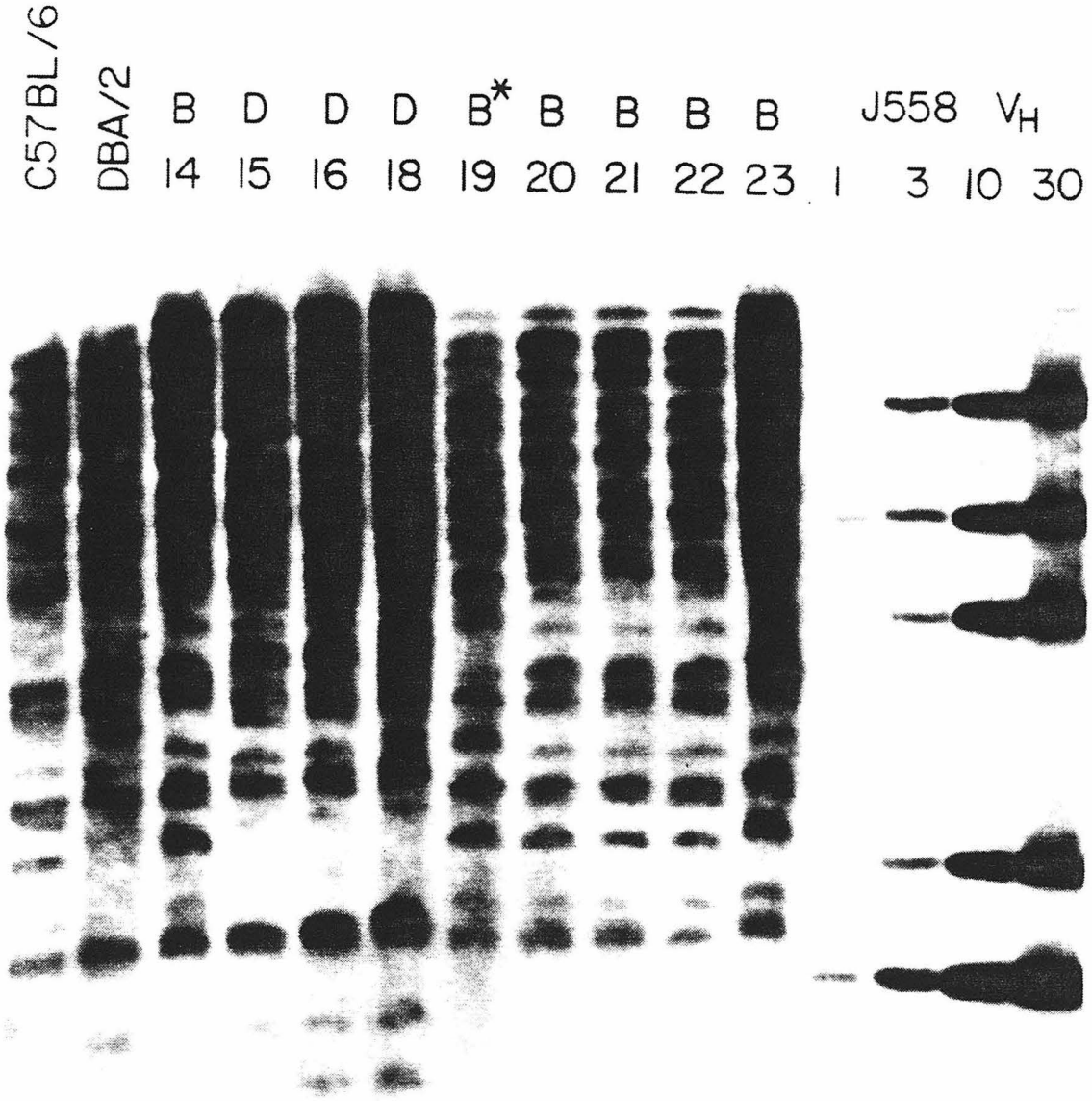

C57BL/6
DBA/2
B 14
D 15
D 16
D 18
B* 19
B 20
B 21
B 22
B 23
J558 V_H 1
3
10
30

Figure 15. An EcoRI digest of C57BL/6, DBA/2, and the BxD recombinant inbred mouse lines 24-32 treated as described in Figure 13.

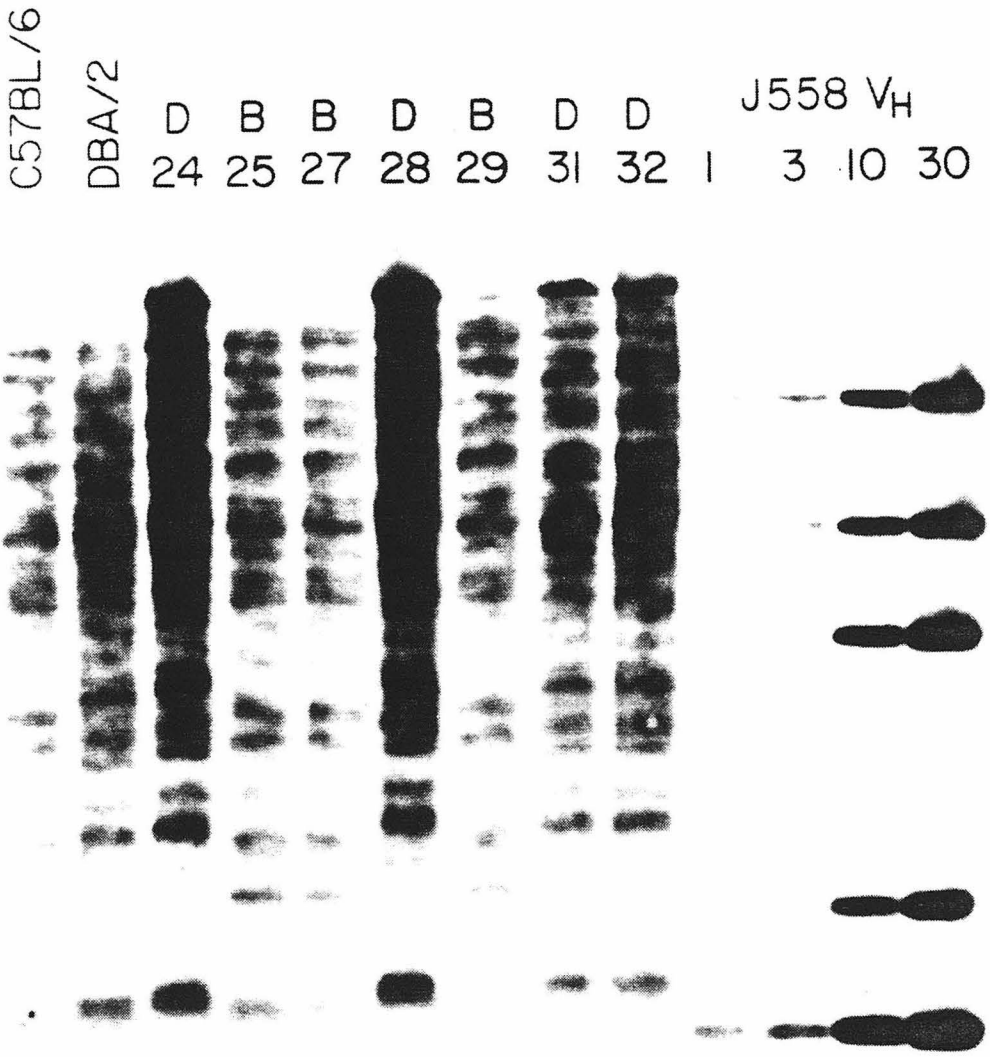

C57BL/6
DBA/2
D 24
B 25
B 27
D 28
B 29
D 31
D 32
J558 V_H
1
3
10
30

Table I shows the results of the linkage analysis of VDX-1. Matches refer to the comparison of the VDX-1 strain distribution pattern with the others shown. r is the map distance in Morgan units. The strain distribution pattern for each marker is given for the BXD lines 1, 2, 5, 6, 8, 9, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31 and 32, respectively. The 1's and 0's below denote a match or non-match, respectively, with the strain distribution pattern of VDX-1.

References

1. Stubbarao, B., Ahmed, A., Paul, W. E., Scher, I., Lieberman, R. and Mosier, D. E. (1979) J. Immunol. **122**, 2279.
2. Makela, O., Karjalainen, K., Imanishi-Kari, T. and Taylor, B. A. (1981) Immunol. Lett. **3**, 169.
3. Karjalainen, K. (1980) J. Immunol. **10**, 132.
4. Ju, S.-T. and Dorf, M. (1980) J. Immunol. **126**, 183.
5. Berek, C., Taylor, B. A. and Eichmann, K. (1976) J. Exp. Med. **144**, 1164.
6. Taylor, B. A., Bailey, D. W., Cherry, M., Riblet, R. and Weigert, M. (1975) Nature **256**, 644.
7. Tada, N., Simura, S., Binari, R., Liu, Y. and Hammerling, U. (1981) Immunogenetics **13**, 475.
8. Taylor, B. A., Bailey, D. W., Cherry, M., Riblet, R., and Weigert, M. (1975) Nature **256**, 644.

Table 1. LINKAGE ANALYSIS OF VDX-1

Chromosome No.	Gene Name	No. of Matches	No. of Possible Matches	Fraction of Non-Matches	r	Standard Deviation	Strain Distribution Pattern	Reference
12	LYB-7	5	9	.444	.333	.373	000D080008D080080880080000 000100000110000101000000000	1
12	VDX-1	25	25	.000	.000	.000	B88D8DDD88DDD08888D88D8DDD 1111111111111111111111011	
12	VDX-2	24	25	.040	.011	.011	B88D8DDD88DDD88888D88D8DDD 1111111111110111111111011	
12	NBP	22	23	.043	.012	.012	B88D8DDD88DDD88888D88D8DD0 1111111111110111111111000	2,3
12	NP	22	23	.043	.012	.012	B88D8DDD88DDD88888D88D8DD0 1111111111110111111111000	2,3
12	NPID	21	22	.045	.012	.013	B88D8DDD88DDD80888D08D8DD0 1111111111110001111011011	2,3
12	NP-A	16	17	.059	.016	.017	B8DDDDDDDDDD88808D88D8D00 1101011000111011011111000	2,3
12	GTE	24	25	.040	.011	.011	B88D8DDD88DDD88888D88D8DDD 1111111111110111111111011	4
12	SA4	20	23	.130	.041	.027	B88D8DDD88DDD8D888D08D8DD0 1111111111111001111011000	5
12	SA2	19	23	.174	.059	.036	B88D8DDD88DDD8D888DDDD8DD0 1111111111111001111011000	5
12	IGH-C	22	25	.120	.037	.024	B88D8DDD88DDD8D888D8DD8DDD 1111111111111001111011011	6
12	IGH-2	22	25	.120	.037	.024	B88D8DDD88DDD8D888D8DD8DDD 1111111111111001111011011	7
12	PRE-1	18	25	.280	.121	.067	B88DDDD888DD8DD888DDDD8D8D 1111011011110101111011001	8

APPENDIX 1

FURTHER RESEARCH

The demonstration that the V_H locus in mouse consists of more than 1,000 members immediately raises the question of how many of these members are functional. For the purposes of this discussion, a "functional" V_H gene is one which is joined to a constant region and transcribed in an early B cell. Functional V_H genes, then, are the ones found in the primary repertoire of B-cell clonotypes as they exit from the bone marrow. The kinds of experiments discussed subsequently are immediately applicable to other immunoglobulin or immunoglobulin-like loci of interest. These loci include mouse V_κ, human V_H, V_κ and V_λ, and also with modification, the mouse and human α and β loci of the T-cell receptor. We will restrict our discussion to the mouse V_H locus, realizing that we can generalize it to include these other loci as well.

The first question that comes to mind is what fraction of all functional mouse V_H genes does the J558 V_H family, as we define it, represent? If we assume that all V_H genes are joined and transcribed into functional message except those that are outright pseudogenes detectable by sequence analysis, and that all V_H genes capable of being joined are joined with a roughly equal probability, then knowing what fraction of functional V_H genes is represented by the J558 family enables us to estimate the total size of the V_H locus itself. Even if these two assumptions are invalid, we still need to know what fraction of functional germline V_H diversity we can account for with the set of V_H probes we have. Since J558 represents the great majority of known V_H sequences, this amounts to knowing what faction of functional germline V_H diversity is encompassed by the J558 family.

The best way to measure the degree to which the J558 family contributes to the total functional germline V_H diversity is to measure the mass fraction of IgM message from surface immunoglobulin negative or prereceptor B cells of the bone marrow which hybridizes to J558 V_H. Receptor B cells have undergone joining at

their heavy chain locus but not at a light chain locus. Hence, they have cytoplasmic IgM message and protein, but cannot export the protein to the cell surface because it lacks light chain. They, therefore, have not been stimulated to divide by antigen because antigen-stimulation requires binding of the antigen by cell surface immunoglobulin. The important thing is that they have joined their heavy chain locus in the milieu of the bone marrow of the intact mouse; therefore, the set of these prereceptor B cells constitutes our best approximation to the initial "read-out" of functional V_H genes from the germline. These prereceptor B cells can be purified from bone marrow B cells having surface immunoglobulin by a rosetting technique involving sheep erythrocytes coated with goat anti-mouse immunoglobulin.

It is possible that almost all V_H regions present in the IgM message of prereceptor B cells can be accounted for by the set of non-cross hybridizing V_H probes we have at present. This result would limit the size of the V_H locus to less than 1,000 functional V_H genes because the fraction of V_H pseudogenes is at least 25%. This result would also indicate that the locus is not a continuum of sequences between two extremes such as the J558 V_H and the S107 V_H whose sequence homology is 45%. Rather, we could say from this result that it is likely that almost all germline V_H sequences belong to a large family whose members are at least 76% homologous to each other on average.

On the other hand, we could find that the existing V_H probes account for only some of the total prereceptor B-cell IgM message, and that a significant fraction of it bears V_H regions for which no probe exists. After subtracting out the IgM message which will hybridize to J558 V_H and the other known V_H probes, the remaining fraction could be made into cDNA and cloned. We would then have a library of all remaining functional V_H sequences for which we now have no probe.

Because of the constraint of the maximum size of the V_H locus from genetic mapping (10 cM) and the overall size of chromosome 12, we can guess that the J558

V_H family with its 1,000 members occupies at very least 10% of the locus and perhaps almost all of it. We could not predict solely from the frequencies of sequences in the library what the rest of the V_H locus is like. The frequency of different V_H sequences in the proposed cDNA library depends not only on the relative sizes of their germline families, but also on the relative frequencies of their joining, and on their relative rates of transcription. We could, however, use single-stranded probes made from these V_H sequences to find out what the sizes of their corresponding germline V_H families are.

This approach is limited in that we still would not have access to those V_H sequences from small families which are rarely, if ever, joined. The distribution of V_H sequences in the library would, however, reflect the distribution of V_H sequences in the functional prereceptor B-cell repertoire. This repertoire, in turn, reflects the repertoire of B-cell clonotypes in the mouse which initially sees antigen and responds. It seems that, at least, the functional part of the germline V_H locus could be characterized genetically in terms of how many different families of V_H genes it contains, what size they are, and what their sequences are like.

APPENDIX 2

Ann. Immunol. (Inst. Pasteur)
1984, 135 C, 83-88

MOLECULAR GENETICS OF ANTI-CARBOHYDRATE ANTIBODIES

by **R. M. Perlmutter** (1), **S. T. Crews** (2), **J. Klotz, D. Livant** (2), **J. Siu** (2) and **L. Hood** (3)

Division of Biology, California Institute of Technology, Pasadena, California 91125 (USA)

SUMMARY

Antibodies directed against carbohydrate determinants provide useful model systems for understanding the structure and organisation of antibody genes and the generation of antibody diversity. We have used three such systems, PC, DEX and GAC, and have studied the heavy chains and V_H gene segments of each. In two of these systems, PC and GAC, much of the diversity in heavy-chain protein sequences results from somatic mutation events superimposed on expression of a single V_H gene. In the DEX system, it appears that germ-line sequence diversity may be an important contributor to the variability in heavy-chain sequence. Detailed structural analyses of this type will ultimately provide a complete picture of the mechanisms which underlie effective humoral immunity.

KEY-WORDS: Immunogenetics, Humoral immunity, Antibody diversity, Carbohydrate; Models.

INTRODUCTION.

Detailed analysis of antibody structure and genetics has been greatly aided by the discovery of well-defined antigens which elicit highly restricted antibody responses in inbred mice [1]. For more than a decade, our laboratory has pursued a molecular analysis of restricted antibody populations, particularly those directed against carbohydrate determinants found in bacterial vaccines, *e. g.* phosphorylcholine (PC) and

Manuscrit reçu le 7 septembre 1983.

(1) Supported by NIH grant AI18088.
(2) Supported by NIH predoctoral grant GM07616.
(3) Supported by NIH grant AI16913.

α-(1→3) dextran (DEX) [2, 3, 4]. Throughout these studies, our goal has been a complete understanding of the processes which generate antibody diversity. More recently, we have employed DNA cloning technology to better define the molecular genetics of anti-carbohydrate antibodies. In this report, we summarize the results of analysis of genes encoding antibodies directed against PC, DEX and group A streptococcal carbohydrate (GAC).

Structural diversity of PC-binding antibodies.

Early serologic analyses demonstrated that an idiotype associated with the BALB/c anti-PC plasmacytoma protein T15 was present on the majority of BALB/c anti-PC antibodies and was inherited as a single Mendelian trait closely linked to immunoglobulin allotype [5]. Protein-sequence determination of heavy chains from PC-binding hybridoma proteins raised in BALB/c mice further emphasized the structural similarity of these antibodies; however, IgA and IgG antibodies frequently differ from the common T15 structure by as many as 8 amino acid substitutions [2]. Three different light chains can be associated with this heavy-chain structure, each defined by a plasmacytoma prototype: T15, M603 or M167. Amino acid substitutions were frequently observed in light chains as well, and were also correlated with antibody class in that IgM antibodies appeared to have fewer substitutions than did IgA and IgG antibody light chains. These results strongly suggested that a significant portion of the diversity of anti-PC antibodies resulted from somatic processes.

Genes encoding anti-PC antibodies.

In order to examine the genetic basis of the observed structural diversity in anti-PC antibodies, we utilized a cDNA clone homologous to the T15 heavy-chain variable region to screen a BALB/c genomic library. Through this analysis, four gene segments were identified with greater than 88% homology to the cDNA probe [6]. One of these, labelled V1, encodes a protein identical to T15, thus verifying that this heavy chain is indeed represented in the germ-line. The remaining three gene segments, although closely related, do not significantly contribute to the diversity of anti-PC antibodies. One of these sequences, V3, encodes a pseudogene [7], and the remaining two members of the T15 gene family, while functional, encode proteins quite different from known anti-PC heavy chains. Thus, the observed diversity of heavy-chain pro-

---

DEX = α-(1→3) dextran.
GAC = group A streptococcal carbohydrate.
PC = phosphorylcholine.

tein sequences in the antibody response to PC results from somatic mutation processes superimposed on a single V_H gene segment. Similar analyses have revealed that a single germ-line gene likely encodes all M167-like light-chain sequences and that, here, too, observed sequence variants result from a process of somatic mutation [8, 9].

ORGANIZATION AND EVOLUTION OF THE T15 GENE FAMILY.

The four gene segments of the T15 gene family provide an ideal proving ground for the evaluation of evolutionary theories. One of these gene segments, V1, encodes a heavy chain which we suspect is important

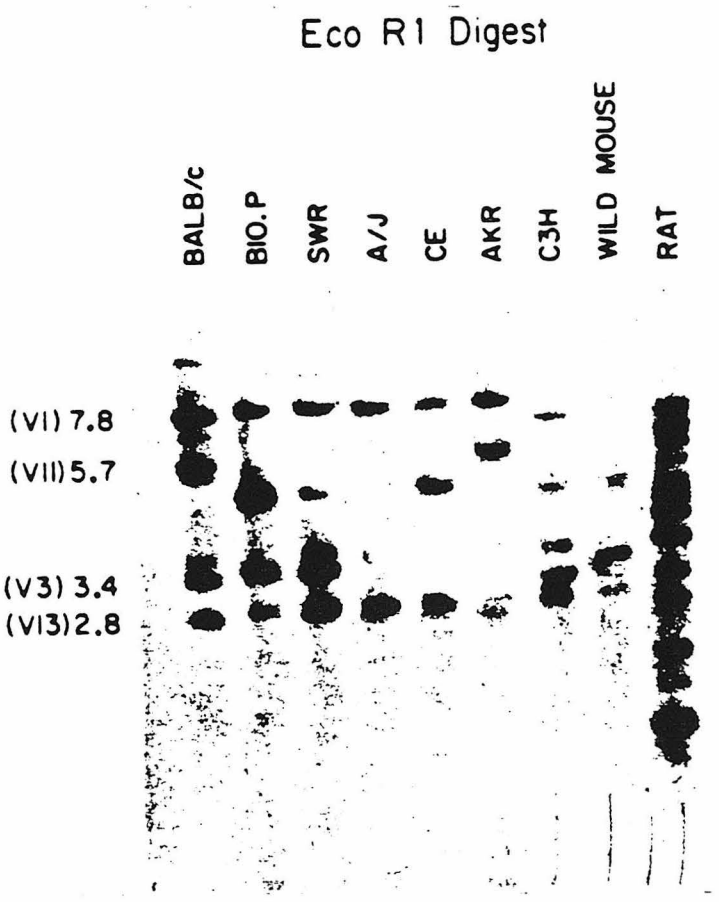


FIG. 1. — *Southern blot analysis of the T15 gene family in mice and rats.*

Ten μg of liver DNA isolated from the indicated mouse strains or from a Lewis rat were completely digested with *Eco*RI endonuclease and subjected to agarose gel electrophoresis. The fragments were transferred by blotting to nitro-cellulose paper and identified by hybridization with a ^{32}P-labelled probe homologous to the T15 V_H gene segment.

for the protection of mice against bacterial infection [10]. Thus, we would anticipate that this sequence will be strongly conserved amongst rodents. A closely related germ-line sequence, V3, is a pseudogene, and hence should not be subject to strong selective pressures. The V3 sequence is located 15 kb 5′ to V1 on chromosome 12 and probably resulted from a recent incomplete gene duplication event (J. Siu, S. Crews and R. M. Perlmutter, unpublished results). V11 and V13 have not as yet been linked to each other or to V3.

Southern blot analysis of genomic DNA from a number of mouse strains and from inbred rats shows that all strains have apparently retained both V1 and V13 elements, and that most strains also contain V11 and V3 (fig. 1). Germ-line rat DNA includes a large number of sequences homologous to the T15 probe, probably reflecting expansion of the T15 gene family during the approximately 10 million years since divergence of rats and mice. We have recently identified and completely sequenced the T15 gene family in the B10.P mouse strain (R. M. Perlmutter, J. A. Griffin, B. Berson and L. A. Hood, manuscript in preparation). From this analysis, it is clear that, although somatic mutation is responsible for much of the diversity in anti-PC antibody heavy chains, the germ-line V1 gene sequence which directs the synthesis of most IgM anti-PC heavy chains has been closely conserved during murine evolution.

Molecular genetics of anti-DEX antibodies.

The murine antibody response to DEX is also highly restricted and regulated by a gene linked to immunoglobulin allotype [11, 12]. Protein sequence analysis of anti-DEX plasmacytoma and hybridoma proteins performed in our laboratory revealed that these antibodies are all structurally quite similar in a manner analogous to that seen in anti-PC antibodies [13]. Here, it is not entirely certain that the protein sequence diversity reflects a process of somatic mutation superimposed on expression of a single germ-line gene. Idiotypic markers which define structurally different members of this antibody group identify serologically homologous molecules in different individuals from different though related mouse strains [12]. Thus, it is conceivable that multiple similar V_H gene segments which encode DEX-binding heavy chains may exist in the germ-line. We have recently initiated a study of these gene segments using a probe which contains the rearranged V_H gene from the J558 plasmacytoma which binds DEX. Here, Southern blots reveal a complex pattern of perhaps 30 different restriction fragments — a complexity which is reflected in the large number of homologous sequences which can be identified in a BALB/c germ-line library using the J558 probe (D. Livant, unpublished data). This analysis is continuing in an attempt to define the germ-line contribution to anti-DEX antibody heavy chains.

MOLECULAR GENETICS OF ANTI-CARBOHYDRATE Ig

Molecular genetics of anti-GAC antibodies.

Another restricted family of antibodies is elicited in mice by immunization with group A streptococcal vaccine. In most mouse strains, each individual animal will produce one or a few different GAC-binding antibodies after immunization; however, it is distinctly unusual for different mice to produce the same antibodies as judged by isoelectric focusing [13]. Thus, the repertoire of anti-GAC antibodies in each individual mouse is quite small, while the strain repertoire of GAC-binding antibodies appears to be quite large. In this system, as in the anti-PC antibodies, xenogeneic antiidiotypic reagents have provided evidence for an important V_H gene which dominates the anti-GAC response in both A/J and BALB/c mouse strains [14].

We have utilized a group of anti-GAC hybridomas to begin detailed analysis of GAC-binding antibodies and the genes which encode them. Here again, families of different heavy chain and light chain sequences are seen which differ by between two and five residues out of the first 60 positions. Using a rearranged V_H gene cloned from one of our GAC-binding hybridomas, we have demonstrated that here, as in the PC system, a large part of the observed protein sequence diversity appears to be generated through a process of somatic mutation acting on a single V_H gene segment (R. M. Perlmutter, J. Klotz, M. Bond, J. M. Davie and L. E. Hood, manuscript in preparation).

RÉSUMÉ

Génétique moléculaire des anticorps antiglucidiques

Les résultats obtenus au cours de l'analyse structurale de gènes V_H codant pour des anticorps dirigés contre des déterminants sucrés (DEX, PC, GAC) soulignent l'importance des mutations somatiques dans la génération de la diversité des anticorps. Cependant, le haut degré de conservation de ces gènes V_H parmi différentes souches de souris et différentes espèces implique que ces gènes sont soumis à de fortes pressions de sélection.

Mots-clés : Immunogénétique, Immunité humorale, Diversité des anticorps, Glucide ; Modèles.

REFERENCES

[1] Krause, R. M., The search for antibodies with molecular uniformity. *Advanc. Immunol.*, 1970, 12, 1-68.

[2] GEARHART, P., JOHNSON, N., DOUGLAS, R. & HOOD, L., IgG antibodies to phosphorylcholine exhibit more diversity than their IgM counterparts. *Nature* (Lond.), 1981, **291**, 29-34.

[3] HOOD, L., LOH, E., HUBERT, J., BARSTAD, P., EATON, B., EARLY, P., FUHRMAN, J., JOHNSON, N., KRONENBERG, M. & SCHILLING, J., The structure and genetics of mouse immunoglobulins: an analysis of NZB myeloma proteins and sets of BALB/c myeloma proteins binding particular haptens. *Cold Spr. Harb. Symp. quant. Biol.*, 1977, **41**, 817-836.

[4] SCHILLING, J., CLEVINGER, B., DAVIE, J. M. & HOOD, L., Structural diversity of murine antibodies binding α-(1→3) dextrans. *Nature* (Lond.), 1980, **283**, 35-40.

[5] LIEBERMAN, R., RUDIKOFF, S., HUMPHREY, W. Jr & POTTER, M., Allelic forms of anti-phosphorylcholine antibodies. *J. Immunol.*, 1981, **126**, 172-176.

[6] CREWS, S. GRIFFIN, J., HUANG, H., CALAME, K. & HOOD, L., A single V_H gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. *Cell*, 1981, **25**, 59-66.

[7] HUANG, H. CREWS, S. & HOOD, L., An immunoglobulin V_H pseudogene. *J. Mol. appl. Genet.*, 1981, **1**, 93-101.

[8] GERSHENFELD, H., TSUKAMOTO, A., WEISSMAN, I. L. & JOHO, R., Somatic diversification is required to generate the V_K genes of MOPC 511 and MOPC 167 myeloma proteins. *Proc. nat. Acad. Sci.* (Wash.), 1981, **78**, 7674-7678.

[9] SELSING, E. & STORB, U., Somatic mutation of immunoglobulin light chain variable region genes. *Cell*, 1981, **25**, 47-58.

[10] BRILES, D., FORMAN, C., HUDAK, S. & CLAFLIN, J. L., Anti-phosphorylcholine antibodies of the T15 idiotype are optimally protective against *Streptococcus pneumoniae*. *J. exp. Med.*, 1982, **156**, 1177-1185.

[11] BLOMBERG, B., GECKELER, W. R. & WEIGERT, M., Genetics of the antibody response to dextran in mice. *Science*, 1972, **177**, 178-183.

[12] HANSBURG, D., PERLMUTTER, R. M., BRILES, D. E. & DAVIE, J. M., Analysis of the diversity of murine antibodies to dextran B1355. — III. Idiotypic and spectrotypic correlations. *Europ. J. Immunol.*, 1978, **8**, 352-359.

[13] PERLMUTTER, R. M., BRILES, D. E. & DAVIE, J. M., Complete sharing of light chain spectrotypes by murine IgM and IgG antistreptococcal antibodies. *J. Immunol.*, 1977, **118**, 2161-2166.

[14] EICHMANN, K. & BEREK, C., Mendelian segregation of a mouse antibody idiotype. *Europ. J. Immunol.*, 1973, **3**, 599-605.

IMPRIMERIE BARNÉOUD — LAVAL

APPENDIX 3

Reprinted from Nature, Vol. 283, No. 5749, pp. 733-739, February 21 1980

An immunoglobulin heavy-chain gene is formed by at least two recombinational events

Mark M. Davis*, Kathryn Calame*, Philip W. Early*, Donna L. Livant*, Rolf Joho†, Irving L. Weissman† & Leroy Hood*

* Division of Biology, California Institute of Technology, Pasadena, California 91125
† Laboratory of Experimental Oncology, Department of Pathology, Stanford Medical School, Stanford, California 94305

The events of B-cell differentiation can be reconstructed in part through an analysis of the organisation of heavy-chain gene segments in differentiated B cells. A mouse immunoglobulin α heavy-chain gene is composed of at least three noncontiguous germ-line DNA segments—a V_H gene segment, a J_H gene segment associated with the C_μ gene segment, and the C_α gene segment. These gene segments are joined together by two distinct types of DNA rearrangements—a V–J joining and a C_H switch.

THE antibody genes provide a unique opportunity for studying the molecular basis of one pathway of eukaryotic differentiation because the rearrangement of gene segments is correlated with the expression of antibody molecules. Antibody molecules are encoded by three unlinked gene families—λ, κ and heavy chain (H)[1]. The λ and κ families encode light (L) chains and the heavy-chain family encodes heavy chains. The light chains are encoded by three gene segments, V_L (variable), J_L (joining) and C_L (constant), which are separated in the genomes of cells undifferentiated with regard to antibody gene expression[2,3]. During differentiation of the antibody-producing or B cell, the V_L and J_L gene segments are rearranged and joined together while the intervening DNA between the J_L and C_L gene segments remains unmodified[2-4]. This process of DNA rearrangement is termed V–J joining. During the expression of the rearranged gene in the differentiated B cell, the coding regions as well as the intervening DNA between the J_L and C_L gene segments are transcribed as part of a high molecular weight nuclear transcript. The intervening region is then removed by RNA splicing to produce a light-chain mRNA with contiguous V_L, J_L and C_L coding segments[5,6]. Recently, we demonstrated that the heavy chain contains three analogous gene segments, V_H, J_H and C_H, which undergo a similar type of V–J joining during B-cell differentiation[7-9]. Each antibody-producing cell synthesises only one V_LJ_L polypeptide sequence and one V_HJ_H sequence, which together form the antigen-binding (V) domain of the antibody molecule.

The heavy-chain genes seem to have a special role in the differentiation of the antibody-producing cell as reflected in a second phenomenon known as the switching of heavy-chain constant regions or C_H switching. Antibody molecules can be divided into five different immunoglobulin classes—IgM, IgD, IgG, IgA and IgE—which are determined by one of eight distinct heavy chain genes [that is, C_μ, C_δ, ($C_{\gamma1}$, $C_{\gamma2a}$, $C_{\gamma2b}$, $C_{\gamma3}$), C_α and C_ε]. Each class of immunoglobulin seems to be associated with unique functions such as complement fixation or the release of histamine from mast cells. The process of B-cell differentiation is complex and a variety of conflicting views exists on the transitional stages. It is generally agreed, however, that IgM is the earliest immunoglobulin class that is expressed in the differentiation of a B cell (see ref. 10). The immature B lymphocyte apparently has the capacity to differentiate along a variety of discrete pathways and produce progeny which may switch from IgM expression to the expression of any one of the other classes of immunoglobulins. The terminal stage of B-cell differentiation is the plasma cell, which is committed to synthesise and secrete large quantities of a single molecular species of antibody. Several studies suggest that the specificity or V domain does not change as different immunoglobulin classes are expressed throughout this differentiation process[11-15]. During C_H switching, light-chain expression remains unaltered.

We are interested in studying the molecular mechanisms which permit a B cell (or its progeny) to express successive classes of antibody molecules. Here we demonstrate that an α heavy-chain gene derived from a terminally differentiated plasma cell is composed of three noncontiguous germ-line DNA segments. These gene segments are joined together by at least two distinct DNA rearrangements—a V–J joining and a C_H switch.

Organisation of the α heavy-chain gene is altered in differentiated as opposed to undifferentiated DNAs

We have constructed 'libraries' of recombinant phage[16] containing large (12–20 kilobases) inserts of mouse DNA in the vector Charon 4A (ref. 17). These libraries contain sufficient numbers of recombinants ($\sim 10^6$) for there to be a high probability that most single-copy sequences of a given genome are included. We have previously reported the construction of such a library from the DNA of the mouse IgA-producing myeloma tumour M603, and the subsequent isolation and characterisation of clones containing rearranged or differentiated genomic DNA. These clones, CH603α6 (α6) and CH603α125 (α125), have V_H and C_α gene segments on a single fragment of DNA[7,8]. The *Eco*RI restriction maps of these clones are shown in Fig. 1*a*. The α6 clone has three *Eco*RI fragments—one 7.3 kilobases long containing the V_H gene segment, a second of 5.1 kilobases containing the 5′ portion of the C_α gene segment, and a third of 4.4 kilobases containing the 3′ portion of the C_α gene segment. Approximately 6.8 kilobases of intervening DNA separate the V_H and C_α gene segments. The α125 clone also has three *Eco*RI fragments—one of 6.5 kilobases which is located on the 5′ side of the 7.3- and 5.1-kilobase *Eco*RI fragments also described for α6.

The genomic clones were isolated using a cloned cDNA plasmid representing the entire heavy-chain mRNA of the IgA-producing tumour S107 (denoted S107 cDNA)[7]. The S107 V_H region is about 98% homologous to the M603 V_H region at the nucleotide level[9]. Analysis by the Southern blot procedure[18,19] of *Eco*RI-digested M603 DNA hybridised to the S107 cDNA probe demonstrates the presence of three *Eco*RI fragments corresponding to those described above in the α6 clone. Thus, the rearranged α6 clone is not an artefact of the cloning or isolation procedures.

When the S107 cDNA probe is used to examine a Southern blot of *Eco*RI-digested sperm or embryo DNA, a somewhat different pattern is observed[8]. In particular, the 5.1-kilobase

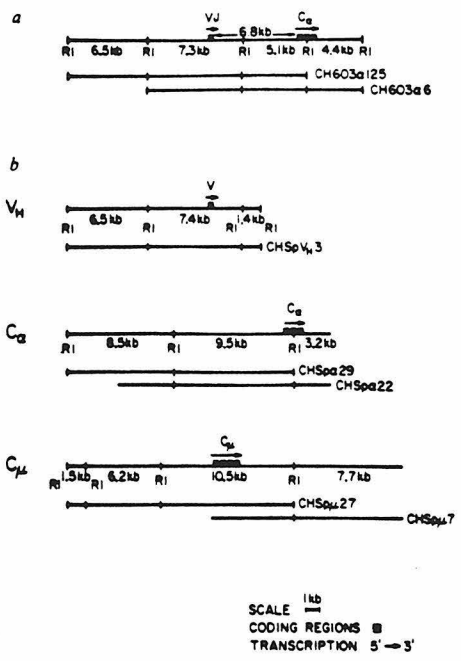


Fig. 1 *a*, α Heavy-chain clones isolated from a genomic library of myeloma M603 DNA. CH603 α6 and CH603 α125 are two overlapping clones derived from a partial *Eco*RI library of M603 DNA, constructed using the phage Charon 4A (ref. 7). The position of the respective gene segments, as well as their direction of transcription, was determined by R-loop mapping and restriction analysis[7,8]. *b*, Germ-line V_H, C_μ and C_α clones isolated from a genomic library of sperm DNA from inbred BALB/c mice. Sperm DNA[33,34] was partially digested with restriction enzymes in two ways: (1) with a mixture of *Hae*III plus *Alu*I and (2) with *Eco*RI alone. After digestion in conditions designed to maximise the yield of 12–20-kilobase fragments, these fragments were selected on sucrose gradients. The *Hae*III/*Alu*I fragments were methylated with *Eco*RI methylase and blunt end ligated with synthetic *Eco*RI cleavage sites and then cleaved with *Eco*RI essentially as described previously[16] except that *Eco*RI linker ligations were done at 18 °C to lessen the endonuclease degradation. Both types of fragments were then ligated to the isolated arms of the bacteriophage Charon 4A (ref. 16) at 4 °C. The enzymatically recombined DNAs were packaged *in vitro* using the strains of Sternberg[35] and the protocol of Hohn[36]. The efficiency of packaging was 400,000 plaque forming units (PFU) per μg inserted DNA in the case of *Eco*RI partially digested DNA and 200,000 PFU per μg for *Hae*III/*Alu*I digestions. The background of non-recombinant Charon 4A was 25% and <5%, respectively. Approximately 500,000 *Eco*RI and 1,200,000 *Hae*III/*Alu*I clones were constructed and amplified. A library of ~500,000 clones provides a 90% chance of finding a given single copy sequence and a library of 1,200,000 clones a 99% chance[37]. The use of enzymes that recognise three different sequences reduces the possibility that a particular region of interest is lost from the library because it had too many or too few restriction sites to fall within the sucrose gradient size cut. Libraries were screened[7,16] with the cDNA clone (see text) S107 or M104E C_μ cloned cDNA probes identified by DNA sequence analysis[9,38]. The C_μ clone used extends from codon 300 to the 3'-untranslated region (~1,000 base pairs).[38]. Independent overlapping clones were obtained from the C_α and C_μ gene segments. Location of the coding regions and the direction of transcription were determined by DNA sequence analyses for the CHSp PC-3 clone, by heteroduplex analyses with the α6 clone for the CHSp α29 clone, and by R-loop mapping and restriction analyses of the CHSpμ27 clone[38]. The germ-line genomic clones will be denoted μ27 (CHSpμ27), V_H3 (CHSpV_H3), and α29 (CHSpα29).

*Eco*RI fragment containing most of the large intervening sequence and the 5' portion of the C_α gene segment is not present. This observation indicates that the α6 clone is the product of one or more DNA rearrangements which presumably occurred during B-cell differentiation[8].

The α gene is composed of at least three different germ-line segments of DNA

In view of the possibility that lymphocytes earlier in the M603 lineage might first have produced IgM molecules and later IgA molecules, we decided to investigate the possible contribution of germ-line C_μ sequences as well as V_H and C_α sequences to the myeloma α gene (Fig. 1*a*). We constructed several libraries of germ-line DNA (sperm) and proceeded to isolate clones containing V_H, C_μ and C_α gene segments using cloned cDNA probes (denoted S107 V_H, C_α and C_μ) for the corresponding coding regions and the screening procedure of Benton and Davis[20]. *Eco*RI restriction maps of several germ-line clones are shown in Fig. 1*b*. We chose sperm (germ-line) DNA for our undifferentiated genomic libraries to eliminate any possibility of DNA rearrangements which may occur in somatic tissues during embryogenesis. The C_α and V_H clones seem to be representative of germ-line sequence organisation because the *Eco*RI fragments containing the corresponding coding regions (in the clones) are identical in size to those found in the Southern blot analysis of undifferentiated DNA with C_α and S107 V_H cloned cDNA probes—9.5 and 7.4 kilobases, respectively (Fig. 2*a–d*). The V_H and C_α probes were derived from restriction fragments of the S107 cloned cDNA[7] extending from approximately the 5'-untranslated region to codon 108 (V_H) and from codons 108 to 274 (C_α). Figure 2*a* shows that a Southern blot of a germ-line

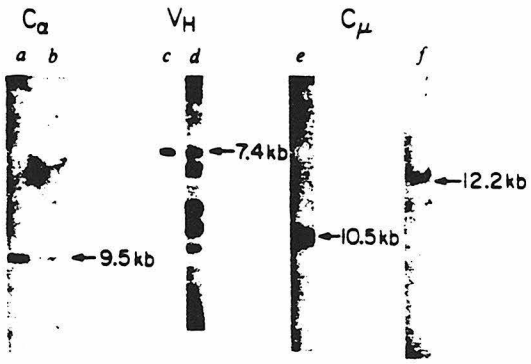


Fig. 2 Southern blot analyses of C_α, C_μ and S107 V_H coding regions in germ-line DNA. Approximately 3–20 μg of sperm DNA or 13-day embryo was digested with *Eco*RI and electrophoresed on a 0.7% neutral agarose gel for 10–12 h at 30–40 V. Gels were then blotted according to the procedures of Southern and Flavell[18,19] and hybridised with ^{32}P nick-translated cDNA[39]. Washing was for 1.5 h in 1 M NaCl, 1 M Tris *p*H 8, 0.1% SDS and 0.1% sodium pyrophosphate at 65 °C and for 1 h in 1×SSC, 0.1% SDS and 0.1% sodium pyrophosphate. Lanes *a* and *b* are α29 and embryo DNAs, respectively, hybridised to the C_α cDNA probe. Lane *c* is V_H3 DNA hybridised to the S107 cDNA probe. Lane *d* is sperm DNA hybridised to the S107 V_H cDNA probe. This probe cross-reacts with eight or nine closely related V_H gene segments[8]. Lanes *e* and *f* are μ27 and sperm DNAs, respectively, hybridised to the C_μ cDNA probe. Identical results in all cases were obtained with BALB/c 13-day embryo DNA or sperm DNA as has been found for a V_κ gene[34]. Sizes (given in kilobases) were determined by comparison with restriction fragments of PBR322 (ref. 40) or by the use of PBR322 multimers generated by *Bam*HI cleavage of PBR322 and limited ligation of the resulting monomer.

C_α clone (α29) digested with *Eco*RI and hybridised with the C_α cDNA probe yields a 9.5-kilobase fragment. A similar analysis of embryo DNA produces a band of identical size (Fig. 2*b*). Southern blots of a germ-line V_H clone (V_H3) and embryo DNA digested with *Eco*RI and hybridised to a S107 V_H cDNA probe gave, respectively, a 7.4-kilobase fragment (Fig. 2*c*) and several fragments including a 7.4-kilobase fragment (Fig. 2*d*). An important question is the relationship between the germ-line V_H3 and myeloma M603 V_H coding regions because there are at least eight V_H gene segments that hybridise to the S107 probe[8]. The germ-line V_H3 DNA sequence codes for the S107 V_H protein sequence[9] and, accordingly, differs from the M603 V_H region by a minimum of four base changes leading to three amino acid subsitutions[21]. Thus, the germ-line V_H3 and the myeloma M603 V_H gene segments may be encoded by two distinct germ-line V_H gene segments or the V_H3 gene segment may give rise to the M603 V_H gene segment by somatic mutation and selection. As we shall show subsequently, the V_H3 clone seems to be indistinguishable from the M603 clone in the V_H coding region and in more than 11-kilobases of 5′-flanking sequence by heteroduplex and restriction enzyme analyses. Therefore, our analyses of the localisation of V_H sequences in the myeloma α6 clone are valid because the germ-line V_H clone serves as a probe for the M603 V_H gene and its attendant 5′-flanking sequence.

The *Eco*RI fragment of the germ-line C_μ clone containing the C_μ coding region is smaller (10.5 kilobases; see Fig. 2*e*) than its counterpart seen on a Southern blot analysis of sperm DNA with a cloned μ cDNA probe (12.2 kilobases; see Fig. 2*f*). We believe this discrepancy arises from one (or more) deletion(s) in the DNA flanking the C_μ coding region during the propagation of the recombinant phage. In isolating μ clones from the M603 library, we obtained several 9.5–10-kilobase *Eco*RI fragments containing C_μ coding regions, whereas Southern blot analysis of M603 DNA with a cloned C_μ cDNA probe demonstrated a genomic fragment of 12.2 kilobases, as in the sperm DNA (data not shown). Attempts to isolate C_μ clones from a library of mouse liver DNA have led to similar results (N. Newell and F. Blattner, personal communication). We will present restriction enzyme data below which demonstrate that this apparent deletion in the μ clone does not affect our general conclusions.

The germ-line V_H, C_α and C_μ clones were compared with the myeloma α6 and α125 clones by heteroduplex analysis. Representative heteroduplexes from each of these comparisons show extensive homologies. The germ-line V_H clone shares approximately 11.6 kilobases of homology with the myeloma α125 clone (Fig. 3*a*). This homology extends from the 5′ end of the α125 clone up to and including the V_H coding region. The germ-line C_μ clone shows 5.0 kilobases of homology with the large intervening sequence of the myeloma α6 clone (Fig. 3*b*). Starting at its 3′ end, the germ-line C_α clone has about 6.4 kilobases of homology with the myeloma α6 clone (Fig. 3*c*). The heteroduplex measurements for these analyses are given in Table 1.

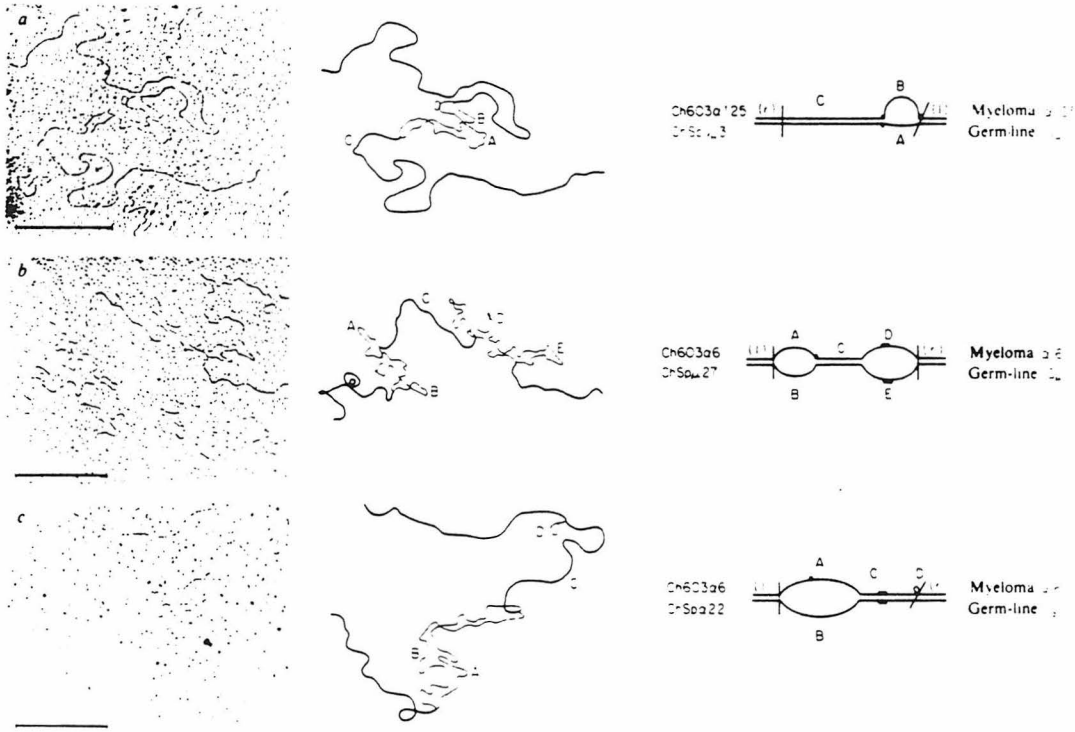


Fig. 3 Heteroduplex analyses of germ-line and somatic clones. The electron micrographs are shown on the left, tracings of these heteroduplexes in the middle and diagrammatic representations on the far right. *a*, Myeloma α125/germ-line V_H PC3. *b*, Myeloma α6/germ-line μ27. *c*, Myeloma α6/germ-line α22. Letters A–E indicate single-strand and double-strand regions for which measurements are given in Table 1. In corresponding line drawings, (r) and (l) refer to the right and left arms of the λ vector. Blocks indicate coding regions in these figures. CsCl purified phage particles were treated with 0.1 M NaOH for 10 min at 20 °C to lyse the phage and denature the DNA simultaneously. After neutralisation of the mixture, the DNA was allowed to reanneal in 50% (v/v) three times recrystallised formamide for 45 min at 20 °C before spreading for electron microscopy[41].

Table 1 Measurements of heteroduplex molecules

Heteroduplex	No. of Molecules	Distance in kilobases A	B	C	D	E
a						
Myeloma α125/ germ-line V_H3	35	4.5± 0.4	7.2± 0.7	11.6± 0.4		
b						
Myeloma α6/ germ-line μ27	30	4.8± 0.5	5.1± 0.5	5.0± 0.4	6.7± 0.4	7.5± 0.5
c						
Myeloma α6/ germ-line α22	26	9.3± 0.5	10.4± 0.9	6.4± 0.5	1.2± 0.2	

Measurements were standardised relative to two circular DNA molecules on the same grid [single-strand ΦX174 DNA (5,375 bases) and double-strand pBR322 DNA (4,365 base pairs)]. Letters refer to regions indicated in Fig. 3. The complete clone designations are given in Fig. 3. In heteroduplex *a*, C refers to the region of duplex between the germ-line V_H clone and the myeloma clone. A and B are the non-homologous single strands of these clones. Similarly, C for heteroduplexes *b* and *c* refers to the duplexes formed between myeloma and germ-line C_μ or C_α clones.

Comparative restriction analyses confirm and extend the heteroduplex data discussed above. A detailed restriction map for the M603 myeloma clones was obtained by double digestion with pairs of restriction enzymes and is shown in Fig. 4*a*. To compare the placement of these cleavage sites with those of the germ-line clones, the 5.1-kilobase restriction fragment of the α6 clone (Fig. 1*a*) that spans the region joining the germ-line C_α and C_μ sequences was subcloned. Using this fragment as a probe, detailed restriction comparisons of the myeloma α6 clone and the germ-line C_μ and C_α clones were made. Representative data are shown in Fig. 5 for these comparisons. Figure 5*a*, *b* and *c* represents *Hin*cII plus *Eco*RI digestions of the myeloma α6 5.1-kilobase RI fragment, the germ-line α29 clone and the germ-line C_μ clone, respectively. These digests were electrophoresed on an agarose gel, blotted onto a nitrocellulose filter and hybridised with the labelled 5.1-kilobase RI fragment. This enables homologous restriction fragments to be identified rapidly and precisely (arrows indicate identical fragments). Figure 5*d*, *e* and *f* shows a similar analysis using *Hin*dIII plus *Eco*RI digestions, respectively, of the myeloma 5.1-kilobase subclone, the germ-line α29 clone and the germ-line μ27 clone. These data, as well as additional restriction analyses of the germ-line V_H3 clone, demonstrate that 4 out of 4 restriction sites in the germ-line V_H3 clone, 10 out of 10 sites in germ-line α29 clone and 9 out of 10 sites in the germ-line μ27 clone corresponded exactly to those found in the myeloma α6 clone (Fig. 4). Not only do these restriction analyses independently confirm the heteroduplex results, but they also suggest that the component germ-line sequences of the α6 clone are very similar to their germ-line counterparts. Thus, the heteroduplex and restriction analyses demonstrate that the germ-line V_H gene segment and its 5′-flanking sequence, although not identical to its M603 counterpart, are very similar and may be used to analyse V_H gene segment organisation in the myeloma M603 clones.

DNA sequence analyses of the myeloma α6 clone and the germ-line μ27 clone have demonstrated that the heavy-chain gene family does have distinct V_H and J_H gene segments in the germ line[9]. Moreover, the germ-line J_H gene segment corresponding to that expressed in the myeloma α6 clone is associated with the germ-line C_μ gene[9]. This J_H gene segment contains the *Hha*I site that marks the end of the homology between the germ-line μ27 clone and the myeloma clones (Fig. 4). Accordingly, the distinct germ-line V_H and germ-line J_H gene segments are rearranged in the myeloma clones in a manner analogous to the V_L and J_L gene segments of myeloma light chain genes[2,3].

Although it seems unlikely, the M603 J_H gene segment and flanking sequences could have been fused to a germ-line C_μ clone as the result of some cloning artefact. To eliminate this possibility, we decided to demonstrate the germ-line association of the J_H flanking sequence and C_μ sequences by Southern blot analyses. On different slots of the same agarose gel, we electrophoresed mouse sperm DNA cleaved with either *Hin*cII or *Eco*RI, transferred these DNAs to a nitrocellulose filter and hybridised one lane of each digest with a C_μ probe and a second lane with the 5.1-kilobase *Eco*RI fragment from the intervening sequence of the myeloma α6 clone (Fig. 6). Figure 6*a* and *b* shows Southern blots of *Eco*RI-digested sperm DNA hybridised with the C_μ cDNA probe and the 5.1-kilobase *Eco*RI subclone from the α6 clone, respectively. In both lanes, a 12.2-kilobase band corresponding to the germ-line C_μ fragment can be identified. In a second digest (*Hin*cII) of mouse sperm DNA, a similar result is obtained with both the C_μ probe (Fig. 6*c*) and the 5.1-kilobase *Eco*RI subclone (Fig. 6*d*) hybridising to a 9.3-kilobase *Hin*cII fragment. The 9.5-kilobase *Eco*RI band in Fig. 6*b* and the 5.0-kilobase band in Fig. 6*d* correspond to C_α restriction fragments. In each case, one of the two bands from the 5.1-kilobase *Eco*RI probe co-migrated with the single C_μ DNA fragment. This analysis shows that at least part of the intervening sequence from the myeloma α6 clone is adjacent to the C_μ gene in the germ line. Thus, the apparent deletion in the germ-line μ27 clone is not a significant factor in our discussion.

A summary of these analyses is presented in Fig. 7. The myeloma α6 clone is composed of three noncontiguous germ-line gene segments: (1) a V_H gene segment and its 5′-flanking

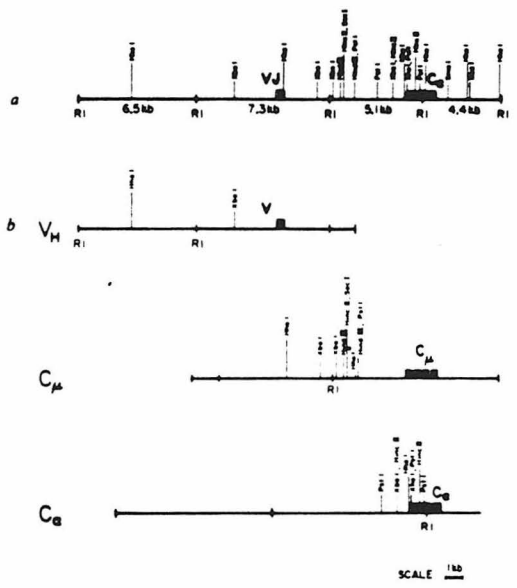


Fig. 4 *a*, Restriction map of myeloma clones α6 and α125. Specific cleavage sites were determined using double enzyme digestion and sizing by gel electrophoresis. Size standards were restriction digests of PBR322 and PBR322 multimers (see Fig. 2 legend). *Hin*dIII, *Hin*cII, *Pst*I and *Sac*I cleavage sites are only shown for the 5.1-kilobase *Eco*RI fragment. *b*, Restriction sites corresponding to those of the myeloma clones detected in the V_H, C_μ and C_α germ-line clones (Fig. 1). Restriction sites in regions corresponding to the 6.5- and 7.3-kilobase RI fragments of α6 and α125 were mapped by double digestion as above. All sites within the 5.1-kilobase RI fragment were compared by co-migration and blotting as described and illustrated in Fig. 5. The *Hha*I site marked by an asterisk in the C_μ clone was the one site found not to conform with those of the myeloma α6 clone.

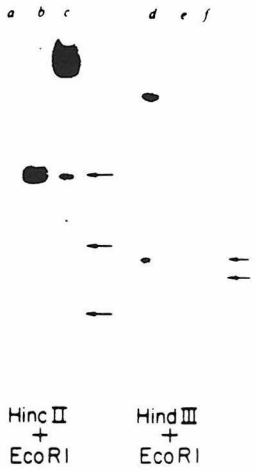


Fig. 5 Comparative restriction digests of the α6 myeloma, germ-line C_μ and C_α clones. Lanes *a–f* show parallel *Hin*cII + *Eco*RI and *Hin*dIII + *Eco*RI digests of the 5.1-kilobase *Eco*RI subclone of the α6 clone (Fig. 1*a*) and DNA from the germ-line α29 and μ27 clones (Fig. 1*b*). Samples were electrophoresed in 1% agarose, transferred to nitrocellulose and hybridised with nick-translated 5.1-kilobase α6 subclone DNA. The co-electrophoresis of restriction fragments allows a large number of different restriction sites to be compared rapidly. In this way, all 16 mapped sites in the 5.1-kilobase α6 subclone were compared with equivalent sites on the α29 and μ27 clones. Arrows show coincident bands.

sequence, (2) flanking sequences located 5′ to a germ-line C_μ gene which includes the J_H gene segment, and (3) the germ-line C_α gene segment with its flanking 3′ and 5′ sequences. Note that within the limits of the methods used here these three germ-line gene segments and their attendant flanking sequences apparently cover the entire myeloma α6 clone.

The α heavy-chain gene is formed by at least two recombinational events

Two distinct DNA rearrangements have occurred to form the M603 α gene—V–J joining and C_H switching (Fig. 8)[9]. The simplest interpretation of these observations is that the V_H gene segment is first joined to a J_H gene segment linked to the C_μ gene segment. This V–J joining, analogous to that which occurs in light chains, generates a rearranged μ gene that presumably leads to the expression of IgM molecules. V–J joining also commits an individual lymphocyte to the expression of a single V domain that remains invariant throughout subsequent steps of B-cell differentiation. Later, a C_H switch joins the V–J gene segment to the C_α gene segment to create a functional α gene and thus enables the differentiated lymphocyte to express IgA molecules. Thus, the myeloma α6 heavy-chain gene is assembled by two distinct and presumably independent DNA rearrangement events. Our heteroduplex, Southern blot and restriction mapping data show unequivocally that these two rearrangements occur at two distinct sites in the genome. These two sites of rearrangement are termed the V–J joining site and the C_H switch site, respectively.

These data are consistent with a differentiation pathway in which a B cell may switch from IgM to IgA synthesis while expressing the same V domain. We cannot establish that these two DNA rearrangement events occurred at different times, although this supposition is reasonable. We also cannot rule out the possibility of intermediate differentiation states in this B-cell lineage where IgG molecules were produced, although there is no direct evidence for such a stage.

The C_H switch may be explained by any one of several genetic models

Several mechanisms of C_H switching have been proposed[11,22–26], many of which are similar to those proposed for V–J joining[27]. These models can be categorised as involving either DNA rearrangements to replace one constant region with another (successive deletions, excision–insertions or inversions) or the differential processing of a large nuclear RNA transcript containing multiple heavy-chain constant-region genes[26].

The RNA processing model seems unlikely as a general mechanism for C_H switching at the level of antibody-secreting plasma cells. High molecular weight nuclear RNAs from three myeloma tumours hybridise only with a cDNA probe complementary to the class of immunoglobulin expressed in that tumour and not with probes from non-expressed immunoglobulin classes[28]. Moreover, cell fusion experiments hybridising two different myeloma cells demonstrate that the hybrid cells synthesise only parental V_HC_H combinations[29], a result in conflict with the simple RNA processing model. Furthermore, neither $C_{\gamma 1}$ nor C_μ DNA probes hybridise on Southern blots with the M603 α clones (data not shown), contrary to one prediction of the RNA processing model.

On the other hand, the evidence presented here strongly supports DNA rearrangement as a fundamental element in the mechanism for heavy-chain switching. As summarised in Fig. 7, a very large segment of C_μ flanking sequence has been brought adjacent to C_α and V_H gene segments in the active α gene of myeloma tumour M603. The creation of the M603 α gene apparently requires two DNA rearrangements (Fig. 8). A C_H switching

Fig. 6 Southern blots of the 5.1-kilobase *Eco*RI and C_μ fragments. Mouse sperm DNA was digested with *Eco*RI or *Hin*cII and 3 μg was loaded onto a 4 mm × 20 × 20 cm 0.7% agarose gel, electrophoresed at 40 V for 10 h and blotted as described. Probes used were either the 5.1-kilobase *Eco*RI fragment or a C_μ cDNA clone nick-translated to 4×10^8 c.p.m. per μg. Washing was as described in Fig. 3 legend except that lanes hybridised with the 5.1-kilobase *Eco*RI fragment were washed further in 10 mM NaCl, 10 mM Tris, 0.1% SDS and 0.1% NaPP, for 2 h at 68 °C to reduce the signal strength of weakly homologous repeats. Filters were exposed for 12 h with an intensifying screen at −80 °C. The faint band above the 12.2-kilobase band in *b* is a partial digestion product. All lanes shown were run on the same gel and blotted simultaneously, alignment being assisted by inclusion of pBR322 multimers as internal standards (see Fig. 2).

DNA rearrangement is not postulated in the RNA processing models[26]. The data presented here do not distinguish between the various types of DNA rearrangements proposed, but the hybridisation kinetics experiments of Honjo and Kataoka are consistent with a deletional mechanism for the C_H switch[25]. In addition, recent experiments suggest that V–J joining in mouse λ genes is accomplished by a deletional mechanism[4]. Thus, both types of DNA rearrangement, V–J joining and C_H switching, may arise through deletional mechanisms. If the deletional model is correct for either type of DNA rearrangement, the differentiation of B cells is irreversible because chromosomal information is lost with the excision of each deletional loop of chromosome.

Gene organisation studies may delineate distinct pathways of B-cell differentiation

It will be interesting to determine whether all joined α genes have the same switch site. Southern blot analyses of the DNA from a second IgA-producing myeloma tumour, H8, with the 5.1-kilobase *Eco*RI probe yield a restriction fragment pattern identical to that of M603 DNA (data not shown). In particular, the 5.1-kilobase *Eco*RI fragment (Fig. 1) which contains the switch site seems the same. These data suggest that the expressed α genes in both M603 and H8 myeloma tumours have the same C_H switch site. However, Southern blot analyses of several other closely related IgA-producing tumours do not show a 5.1-kilobase *Eco*RI fragment (M.M.D., unpublished observation). Thus, there may be multiple C_H switch sites for the C_α gene segment. As it seems that B cells producing IgM or IgG may switch to the production of IgA, perhaps distinct C_H switch points reflect distinct pathways of B-cell differentiation. It will also be interesting to determine the location and number of switch sites for other immunoglobulin classes. If each C_H gene segment has a unique site or set of sites for C_H switching, one may be able to trace the distinct pathways of B-cell differentiation by studying the sequence organisation of each functional heavy-chain gene.

DNA rearrangements of antibody gene segments lead to combinatorial amplification of immunoglobulin information

V–J joining and C_H switching are mediated by DNA rearrangements which display combinatorial properties that amplify the germ-line information encoding the antibody gene families. (1) The V and J gene segments of one antibody gene family may be joined in a combinatorial manner to generate diversity in the third hypervariable regions of both κ and heavy chains[4,8,9,30,31]. For example, mice may have at least 200 V_H and 5 J_H gene segments that may be joined combinatorially to generate 1,000 different V_HJ_H coding regions. (2) One V domain may be combinatorially switched among eight or more different C_H regions to carry out a variety of different effector functions that are directed at eliminating antigen or triggering defensive mechanisms such as complement fixation. Thus, each recognition (V) domain may be switched to many different effector (C) domains. The combinatorial properties of antibody gene segments and polypeptides therefore contribute to several fundamental aspects of the vertebrate immune response—V region diversity and the combinatorial switching of antigen-

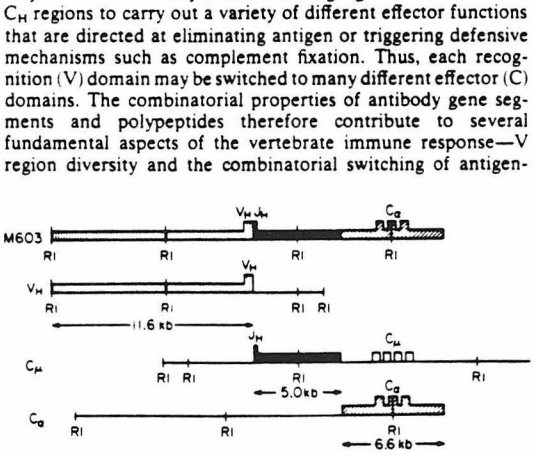


Fig. 7 Origins of the three germ-line components of the myeloma α heavy-chain gene. Various types of shading indicate homology by heteroduplex analyses and by restriction enzyme analyses.

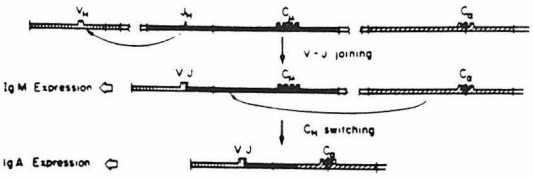


Fig. 8 Two types of DNA rearrangements leading to the creation of the myeloma α heavy-chain gene V–J joining and C_H switching. V–J joining indicates a DNA rearrangement that joins the V_H and J_H gene segments. Because the J_H segments seem to be associated with the germ-line C_μ gene[9], V–J joining permits a μ chain and IgM molecules to be expressed by the differentiating B cell. C_H switching denotes a DNA rearrangement that replaces the C_μ gene segment with a C_α gene segment. This second rearrangement presumably permits an α chain and IgA molecules to be expressed by the now fully differentiated lymphocyte.

recognition (V) domains with a variety of different effector (C) domains during B-cell differentiation. Other complex eukaryotic gene families may use similar DNA combinatorial mechanisms for information amplification[32].

This work was supported by NSF grant PCM 76–81546, ACS grant IM56 and USPHS grant AI09072. M.M.D., P.W.E. and D.L.L. are supported by NIH training grant GM 07616. K.C. is supported by NIH Fellowship GM 05442. R.J. is a fellow of the Swiss National Foundation. I.L.W. is a faculty Research Awardee of the ACS. All experiments involving recombinant organisms were conducted in accordance with the revised NIH Guidelines on recombinant DNA, using P2, EK-2 or P3, EK-2 containment. We thank David Goldberg for his gift of PBR322 multimers, Tom Sargent for packaging extracts, Keichi Itakura for synthetic RI linkers, and David Anderson, Norman Davidson, Max Delbrück, Richard Flavell, Henry Huang, Tom Maniatis and Tom Sargent for helpful discussions.

Received 1 October 1979; accepted 22 January 1980.

1. Mage, R., Lieberman, R., Potter, M. & Terry, W. in *The Antigens* (ed. Sela, M.) 299–376 (Academic, New York, 1973).
2. Brack, C., Hirowa, A., Lenhard-Schueller, R. & Tonegawa, S. *Cell* **15**, 1–14 (1978).
3. Seidman, J. G., Max, E. E. & Leder, P. *Nature* **280**, 370–375 (1979).
4. Sakano, H., Huppi, K., Heinrich, G. & Tonegawa, S. *Nature* **280**, 288–294 (1979).
5. Gilmore-Herbert, M. & Wall, R. *Proc. natn. Acad. Sci. U.S.A.* **75**, 342–345 (1978).
6. Schibler, U., Marcu, K. B. & Perry, R. P. *Cell* **15**, 1495–1509 (1978).
7. Early, P. W., Davis, M. M., Kaback, D. B., Davidson, N. & Hood, L. *Proc. natn. Acad. Sci. U.S.A.* **76**, 857–861 (1979).
8. Davis, M., Early, P., Calame, K., Livant, D. & Hood, L. in *Eukaryotic Gene Regulation* (eds Axel, R. Maniatis, T. & Fox, C. F.) 393–406 (ICN UCLA Symp., Academic, New York, 1979).
9. Early, P. W., Huang, H. V., Davis, M. M., Calame, K. & Hood, L. *Cell* (submitted).
10. Raff, M. C. *Cold Spring Harb. Symp. quant. Biol.* **41**, 159–162 (1976).
11. Sledge, C., Fain, D. S., Black, B., Kneger, R. G. & Hood, L. *Proc. natn. Acad. Sci. U.S.A.* **73**, 923–927 (1976).
12. Fudenburg, H. H., Wang, A. C., Pink, J. R. L. & Levin, A. S. *Ann. N.Y. Acad. Sci.* **190**, 501–506 (1971).
13. Wang, A. C., Geissely, J. & Fudenberg, H. H. *Biochemistry* **12**, 528–534 (1973).
14. Levin, A. S., Fudenberg, H. H., Hopper, J. E., Wilson, S. & Nisonoff, A. *Proc. natn. Acad. Sci. U.S.A.* **68**, 169–171 (1971).
15. Wang, A. C., Wilson, S. K., Hopper, J. E., Fudenberg, H. H. & Nisonoff, A. *Proc. natn. Acad. Sci. U.S.A.* **66**, 337–343 (1970).
16. Maniatis, T. *et al. Cell* **15**, 687–701 (1978).
17. Blattner, F. R. *et al. Science* **196**, 161–169 (1977).
18. Southern, E. M. *J. molec. Biol.* **98**, 503–517 (1977).
19. Jeffreys, A. J. & Flavell, R. A. *Cell* **12**, 429–439 (1977).
20. Benton, W. D. & Davis, R. W. *Science* **196**, 180–182 (1977).
21. Hood, L. *et al. Cold Spring Harb. Symp. quant. Biol.* **41**, 817–836 (1976).
22. Smithies, O. *Science* **169**, 882–883 (1970).
23. Hood, L. *Fedn Proc.* **31**, 177–178 (1972).
24. Bevan, M. J., Parkhouse, R. M. E., Williamson, A. R. & Askonas, B. A. *Prog. Biophys. molec. Biol.* **25**, 131–162 (1972).
25. Honjo, T. & Kataoka, T. *Proc. natn. Acad. Sci. U.S.A.* **75**, 2140–2144 (1978).
26. Rabbitts, T. H. *Nature* **275**, 291–296 (1978).
27. Tonegawa, S., Hozumi, N., Matthyssen, G. & Schuller, R. *Cold Spring Harb. Symp. quant. Biol.* **41**, 877–889 (1976).
28. Marcu, K. B., Schibler, U. & Perry, R. P. *Science* **204**, 1087–1088 (1979).
29. Schulman, M. J. & Kohler, G. *Nature* **274**, 917–919 (1978).
30. Max, E., Seidman, J. G. & Leder, P. *Proc. natn. Acad. Sci. U.S.A.* **76**, 3450–3454 (1979).
31. Schilling, J., Cleavinger, B., Davie, J. & Hood, L. *Nature* **283**, 35–40 (1980).
32. Hood, L., Huang, H. & Dreyer, W. J. *J. supramolec. Struct.* **7**, 531–559 (1977).
33. Witkin, S., Korngold, G. & Bendich, A. *Proc. natn. Acad. Sci. U.S.A.* **72**, 3295–3299 (1975).
34. Joho, R., Weissman, I. L., Early, P., Cole, J. & Hood, L. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
35. Sternberg, N., Tiemeier, D. & Enquist, L. *Gene* **1**, 255–280 (1977).
36. Hohn, B. & Murray, K. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3259–3263 (1977).
37. Clarke, L. & Carbon, J. *Cell* **9**, 91–99 (1976).
38. Calame, K., Rogers, J., Davis, M., Early, P., Livant, D., Wall, R. & Hood, L. *Nature* (submitted).
39. Maniatis, T., Jeffrey, A. & Kleid, D. G. *Proc. natn. Acad. Sci. U.S.A.* **72**, 1184–1188 (1975).
40. Sutcliffe, J. G. *Nucleic Acids Res.* **5**, 2721–2728 (1978).
41. Davis, R., Simon, M. & Davidson, N. *Meth. Enzym.* **21D**, 413–428 (1971).

Printed in Great Britain by Macmillan Production Ltd., Basingstoke, Hampshire

APPENDIX 4

Reprinted from Nature, Vol. 284, No. 5755, pp 452-455, April 3 1980

Mouse C_μ heavy chain immunoglobulin gene segment contains three intervening sequences separating domains

K. Calame*, J. Rogers†, P. Early*, M. Davis*, D. Livant*, R. Wall† & L. Hood*

*California Institute of Technology, Division of Biology, Pasadena, California 91125
†Molecular Biology Institute, Department of Microbiology and Immunology, UCLA School of Medicine, Los Angeles, California 90024

The IgM molecule is composed of subunits made up of two light chain and two heavy chain (μ) polypeptides. The μ chain is encoded by several gene segments—variable (V), joining (J) and constant (C_μ)[1,2]. The C_μ gene segment is of particular interest for several reasons. First, the μ chain must exist in two very different environments—as an integral membrane protein in receptor IgM molecules (μ_m) and as soluble serum protein in IgM molecules into the blood (μ_s). Second, the C_μ region in μ_s is composed of four homology units or domains ($C_\mu 1$, $C_\mu 2$, $C_\mu 3$ and $C_\mu 4$) of approximately 110 amino acid residues plus a C-terminal tail of 19 residues[3,4]. We asked two questions concerning the organisation of the C_μ gene segment. (1) Are the homology units separated by intervening DNA sequences as has been reported for α (ref. 5), γ_1 (ref. 6) and γ_{2b} (ref. 7) heavy chain genes? (2) Is the C-terminal tail separated from the $C_\mu 4$ domain by an intervening DNA sequence? If so, DNA rearrangements or RNA splicing could generate hydrophilic and hydrophobic C-terminal tails for the μ_s and μ_m polypeptides, respectively. We demonstrate here that intervening DNA sequences separate each of the four coding regions for C_μ domains, and that the coding regions for the $C_\mu 4$ domain and the C-terminal tail are directly contiguous.

A recombinant plasmid, p104Eμ12 (pμ12), containing a cDNA sequence from the heavy chain mRNA of the IgM-producing myeloma tumour, M104E, was constructed and characterised by restriction mapping and partial DNA sequence analysis (Fig. 1). A comparison of these DNA sequences with the protein sequences of the M104E myeloma μ chain[4] indicates that pμ12 contains C_μ coding sequences extending from residue 300 to the C-terminus of the μ chain at position 576. The codon for the C-terminal tyrosine is followed immediately by a stop codon (UGA).

The pμ12 probe was used to screen several genomic libraries constructed in the vector Charon 4A—a partial *Eco*RI library from the DNA of IgA-producing myeloma M603 (refs 5, 8), a partial *Eco*RI germ-line library from mouse sperm DNA[1] and a partial *Hae*III + *Alu*I germ-line library from sperm DNA[1]. Southern blot analyses of *Eco*RI-digested mouse sperm and M603 DNA using the pμ12 probe showed identical 12.2-kilobase C_μ bands[1]. This suggests that both the myeloma and germ-line libraries contain the C_μ gene segment in the germ-line or unrearranged state. Figure 2 shows the restriction enzyme patterns of three genomic clones. ChSpμ27 (μ27) from the sperm library and Ch603μ35 (μ35) from the M603 library contain *Eco*RI restriction fragments of 10.2 and 9.8 kilobases, respectively, which hybridise to the pμ12 probe. The μ27 and μ35 *Eco*RI fragments are slightly smaller than the 12.2-kilobase band observed in *Eco*RI-digested sperm and M603 DNAs. We believe that this discrepancy is caused by deletions in genomic DNA flanking the C_μ gene segement which occurred during growth or amplification of the recombinant phages[1]. Preliminary Southern blot comparisons of germ-line and μ27 DNAs localise these deletion(s) to within 1 kilobase 5′ to the C_μ coding region (M. D., unpublished results) (Fig. 2). A similar result has been obtained by others with C_μ-containing clones isolated from a mouse liver DNA library (F. Blattner and N. Newell, personal communication). However, as we show below, the C_μ coding sequences of each clone are identical within the limits of our analyses and presumably represent the true germ-line arrangement of the C_μ gene segment.

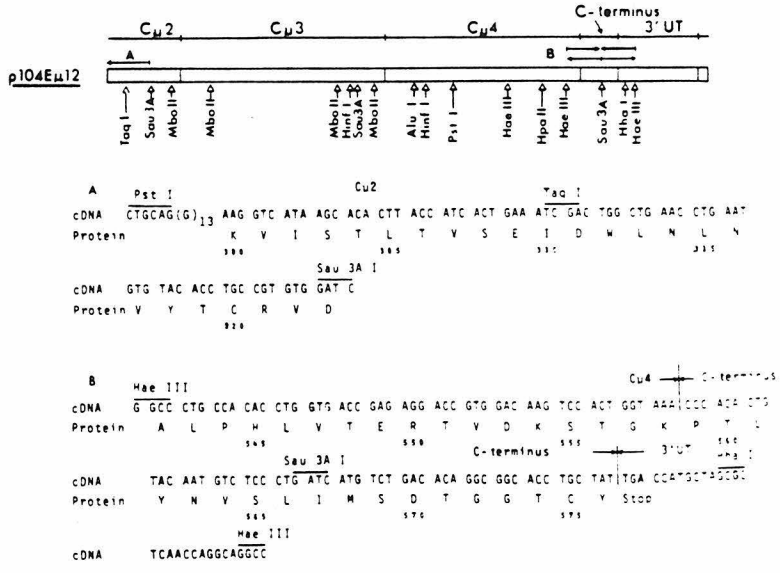


Fig. 1 Restriction map and sequences of p104Eμ12. The double-stranded cDNA segment was originally inserted into the *Pst* site of pBR322 by a dC·dG tailing procedure which reconstituted the *Pst* sites at each side[16] (W. Rowenkamp and R. Firtel, personal communication). Mapping was initially done by multiple digests of the whole plasmid. To locate *Mbo*II sites and to confirm some *Sau*3AI and *Hae*III sites, isolated restriction fragments were 5′ end-labelled, cut to separate the labelled ends, and partially digested with the appropriate restriction enzymes. DNA sequences (in regions indicated by arrows) permitted the map to be aligned with the C_μ amino-acid sequence[4]. All the restriction sites mapped are consistent with the coding sequence. There are no sites in the insert for *Eco*RI, *Bam*HI, *Kpn*I or *Xba*I. A and B show DNA sequences from p104Eμ12. Sequences were obtained by the method of Maxam and Gilbert[17], with minor modifications using the 5′ end-labelled restriction fragments indicated in the restriction map. Sequence A was determined in one direction only; nucleotides printed in lower case type were ambiguous and have been supplied from the coding requirements. Sequences B was determined in both directions and was unambiguous.

R-loop analyses using heavy chain mRNA from myeloma tumour M104E were performed and similar results were obtained for all three genomic clones. Representative electron micrographs are shown in Fig. 3. A total of 90 R-loop molecules were analysed, 62 from μ35 DNA and 28 from μ27 DNA. A small number of the molecules (6%) showed four single-stranded DNA loops of approximately equal length which we interpret to be four C_μ coding regions separated by small, base-paired intervening sequences (Fig. 3*a*, *b*). A larger fraction (74%) showed two or three loops, indicating that one or two of the intervening sequences were base-paired while the others remained single-stranded. Often a small bulge appeared at a reproducible point in the double-stranded DNA·RNA hybrid, indicating the position of the single-stranded intervening sequence (Fig. 3*d*). The remaining 20% of the R-loop molecules showed one large loop. Most of these had small bulges in the double-stranded portion at the positions of one or more single-stranded intervening sequences. The total length of the four coding regions was 1.6 ± 0.2 kilobases and the average size of the loops was 375 ± 51, 364 ± 58, 361 ± 60 and 369 ± 73 base pairs. The size of the three intervening DNA sequences could not be accurately determined by this procedure. In addition, R-loop measurements located the C_μ coding region ~ 3.6 ± 0.36 kilobases from the 3' end of the 9.8-kilobase *Eco*RI fragment of μ35. The R-loop structures that we observe indicate that the C_μ gene contains three intervening sequences which seem to separate regions coding for the four structural domains of the μ heavy chain.

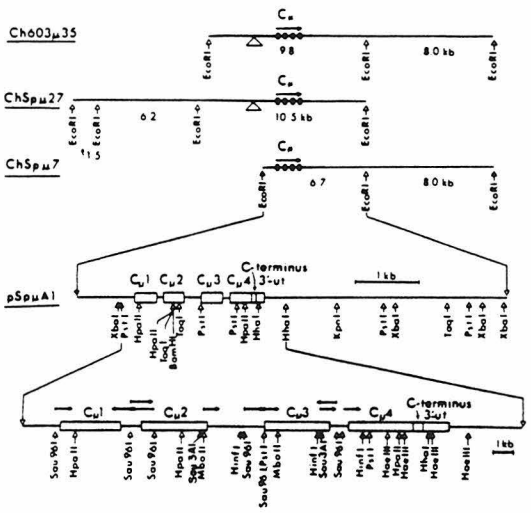


Fig. 2 Restriction maps of three C_μ genomic inserts in Charon 4A. Natural *Eco*RI sites are indicated by open arrow heads, synthetic ones by filled arrow heads. The 8-kilobase *Eco*RI fragment in μ35 may be a cloning artefact. The C_μ coding regions were located by R-loop and restriction enzyme mapping, as well as DNA sequencing. The direction of transcription, indicated by arrows above the coding regions, was determined by comparison of restriction enzyme sites in genomic clones to pμ12. Triangles indicate the approximate locations of deletions[1]. The 6.7-kilobase *Eco*RI fragment from ChSpμ7 (μ7) was recloned by ligation into the *Eco*RI site of pBR322, generating plasmid pSpμA1. In the μ7 region shown by R-looping to contain the C_μ gene, digestion with restriction enzymes determined that pμ12 could be mapped onto pSpμA1 with intervening sequences between C_μ2, C_μ3 and C_μ4. The C_μ1 domain has been positioned by R-looping mapping and DNA sequencing. The lower portion of the figure depicts the sequencing strategy used to define precisely intervening sequence boundaries (Fig. 4). All *Pst*I, *Hha*I and *Hpa*II sites in the region of the C_μ gene are shown, together with those sites for other enzymes which were used for sequencing (arrows) and for identifying intervening sequences. These include the sites bounding those fragments described in the text which span the $C_\mu 2/C_\mu 3$, $C_\mu 3/C_\mu 4$, and $C_\mu 4$/C-terminal-3' untranslated junctions.

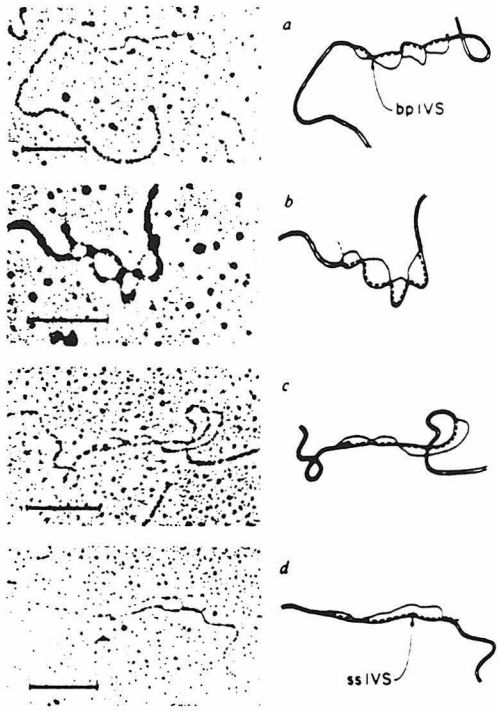


Fig. 3 Electron micrographs of R-loops formed from genomic C_μ clones and purified M104E mRNA. Cloned DNA was photochemically cross-linked with 4,5',8-trimethylpsoralen to produce an average of one cross-link every 4 kilobases[18] before incubation with M104E mRNA in R-loop conditions[19]. R-loops were either spread immediately from a hyperphase of 70% (v/v) three times recrystallised formamide or were fixed with 1 M glyoxal at 12 °C for 2 h before spreading. Both procedures gave similar results. *a*, *b*, R-loops on Ch603μ35 DNA. The arrow indicates the location of a typical base-paired intervening sequence (bpIVS). Both molecules show four C_μ coding regions interrupted by three base-paired intervening sequences. *c*, *d*, R-loops on ChSpμ27 DNA. Molecule in *c* shows two base-paired intervening sequences, that in *d* shows one base-paired intervening sequence and one single-stranded intervening sequence (ssIVS), indicated by arrow. In both molecules *c* and *d*, a third intervening sequence was not clearly observed. mRNA coding for the V region remains unhybridised and is visible in extended form in *b* and *c*. Scale bars, 0.5 μm.

The coding regions were located more precisely by restriction mapping (Fig. 2). We mapped *Hin*fI, and *Sau*3A sites in the region of the four C_μ domains in order to measure the sizes of the intervening sequences. A *Sau*3A fragment spanning the $C_\mu 2$–$C_\mu 3$ junction is 331 base pairs long in the pμ12 cDNA clone and 620 base pairs long in the germ-line clone ChSpμ7 (μ7). Therefore, an intervening DNA sequence of 289 ± 15 base pairs exists between $C_\mu 2$ and $C_\mu 3$. A *Hin*fI fragment spanning the $C_\mu 3 - C_\mu 4$ junction is 122 base pairs long in the cDNA clone and 230 base pairs long in the germ-line clone, while a *Sau*3A-*Pst*I fragment is 147 and 254 base pairs long in the two clones, respectively. Accordingly, there are 108 ± 10 base pairs of intervening sequence between $C_\mu 3$ and $C_\mu 4$. The $C_\mu 1$–$C_\mu 2$ junction was not available in a cDNA clone, but the entire

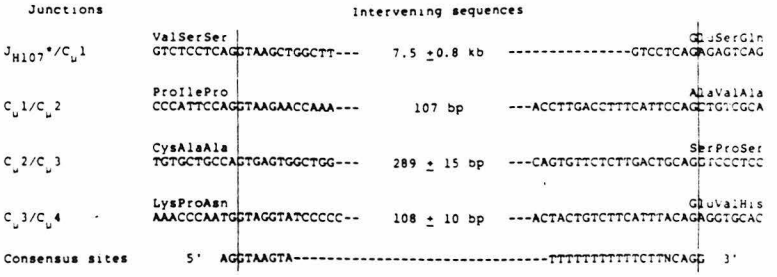


Fig. 4 Junctional sequences in the germ-line C_μ gene. *The location and sequence of J_{H107} in ChSpμ27 is from ref. 2.

intervening sequence between $C_\mu 1$ and $C_\mu 2$ from the genomic clone $\mu 7$ has been sequenced and is 107 base pairs long (J. R., unpublished results).

All of the boundaries of the coding and intervening sequences in the C_μ gene segment were sequenced using the strategy shown in Fig. 2. Figure 4 gives the sequences of the splice sites and shows that there is terminal redundancy about each of the intervening sequences. At the downstream splice sites between J_H and $C_\mu 1$, and between $C_\mu 1$ and $C_\mu 2$, the noncoding sequence is identical to the preceding coding sequence for 7 and 8 nucleotides, respectively, before the indicated splice points. The precise splice points can be designated according to the/GT $\cdots$ AG/rule[9], and the junction sequences are then found to conform generally to the 'consensus' RNA splicing sites (Fig 2)[9-12]. The intervening sequences occur in codons 127, 230, 340 and 446. These are identical to the C_μ domain boundaries as far as they could be determined from protein sequence homologies[4].

Restriction mapping also indicates that the $C_\mu 4$ coding region and the C-terminal tail are not separated by an intervening sequence but are continuously encoded in the germ-line DNA. This junction is spanned in pμ12 by a *Pst*I-*Hha*I fragment (Fig. 1), which is 297 base pairs long according to the coding requirements. The corresponding fragment from a subclone of $\mu 7$ (Fig. 2) co-migrates with the $\mu 12$ fragment to an accuracy of ± 15 base pairs. This fragment was isolated from the $\mu 7$ subclone and digested separately with *Hae*III, *Sau*3AI, and *Hpa*II. The fragments observed were identical to those predicted from the map of the $\mu 12$ cDNA clone (Fig. 1) to an accuracy of ± 6 base pairs. In additions, the $C_\mu 4$-C-terminal junction is spanned in $\mu 12$ by a completely sequenced 132-base pair *Hae*III fragment. The *Hae*III fragment from the corresponding region of $\mu 7$ co-migrates with this to within ± 2 base pairs. Thus detailed restriction analyses demonstrate that within the limits of these analyses (± 2 base pairs) there is not an intervening DNA sequence between the $C_\mu 4$ domain and the C-terminal tail.

Our current observations on the C_μ genomic clones, in conjunction with previous studies on C_α[5] and $C_{\gamma 1}$ and $C_{\gamma 2b}$ genomic clones, suggest that all immunoglobulin C_H genes will contain intervening sequences separating the regions coding for structural domains. Although the function of intervening sequences in eukaryotic genes remains unclear, their positioning precisely at the interdomain boundaries of immunoglobulin C_H genes suggests that the positions of the intervening DNA sequences may have some role in the evolution of immunoglobulin genes. As individual immunoglobulin domains probably encode distinct functions, the presence of intervening DNA sequences at the domain boundaries may facilitate the rearrangement of domain coding regions and thereby generate new combinations of heavy chain domains for selection, thereby speeding up the evolution of immunoglobulin genes[8,13]. Alternatively, it also has been proposed that intervening DNA sequences inhibit recombination and, accordingly slow the evolution of eukaryotic genes[14,15].

Our studies on the organisation and structure of the C_μ gene segment place several constraints on models for the difference between μ_m and μ_s chains. First, Southern blot analyses of embryo or germ-line DNA with the pμ12 probe show only strongly hybridising bands corresponding to a single C_μ gene segment which is present in the $\mu 35$, $\mu 27$ and $\mu 7$ clones. This is true for digests with *Eco*RI or *Hinc*II (ref. 1), as well as for *Bam*HI or *Hha*1 (M. D., unpublished results). We have isolated 10 independent genomic clones which hybridise to pμ12, and all of these seem to contain the same C_μ gene segement (M. D., K. C. and P. E., unpublished results). Thus, these restriction mapping and gene cloning results strongly suggest that there is only one C_μ gene segment in the BALB/c genome. If so, the μ_s and μ_m chains must be encoded by the same C_μ gene segment. Second, a DNA rearrangement during B-cell development to generate alternative $C_{\mu m}$ and $C_{\mu s}$ gene segments is unlikely. In B-cell development, the μ_m chain is expressed before the μ_s chain. Thus, a putative DNA rearrangement should alter the 3' structure of the expressed $C_{\mu s}$ gene segment. However, our results show that the 3' end of the C_μ gene in the pμ12 cDNA clone is identical to the 3' end of the C_μ gene in the $\mu 7$ germ-line clone, thus ruling out the possibility of a DNA rearrangement at the 3' coding region of the $C_{\mu s}$ gene segment. Finally, we can rule out a simple post-translation cleavage of a larger μ_m chain to create the μ_s chain because the μ_s coding sequence is followed immediately by a stop codon (Fig. 1).

Two models to explain the origins of μ_m and μ_s still appear plausible: (1) the μ_m chain is generated from the μ_s chain by a novel type of post-translational modification; and (2) a different COOH-terminal coding region for the μ_m chain does exist. A large nuclear RNA transcript could give rise either to μ_s mRNA or, alternatively, to μ_m mRNA by RNA termination, cleavage and/or splicing. We are currently studying the μ RNAs from a B-cell lymphoma which produces only membrane IgM to test these models.

This work was supported by NSF grant PCM 76-81546 and NIH grants CA 12800 and AI 13410. M.D., P.E. and D.L. are supported by NIH Training Grant GM 07616. K.C. is supported by NIH Fellowship GM 05442.

Received 2 November 1979; accepted 13 February 1980

1. Davis, M. M. *et al. Nature* **283**, 733 (1980).
2. Early, P. W., Huang, H. V., Davis, M. D., Calame, K. & Hood, L. *Cell* (in the press).
3. Putnam, F. W., Florent, G., Paul, C., Shinoda, T. & Shimizu, A. *Science* **182**, 287 (1973).
4. Kehry, M., Sibley, C., Fumham, J., Schilling, J. & Hood, L. *Proc. natn. Acad. Sci. U.S.A.* **76**, 2932 (1979).
5. Early, P. W., Davis, M. D., Kaback, D., Davidson, N. & Hood, L. *Proc. natn. Acad. Sci. U.S.A.* **76**, 857 (1979).
6. Sakano, H. *et al. Nature* **277**, 627 (1979).
7. Kataoka, T., Yamawaki-Kataoka, Y., Yamagishi, H. & Honjo, T. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4240 (1979).
8. Davis, M., Early, P., Calame, K., Livant, D. & Hood, L. in *Eukaryotic Gene Regulation, ICN-UCLA Symposium* (eds Axel, R., Maniatis, T. & Fox, C. F.) 393 (Academic, New York, 1979).
9. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. *Proc. natn. Acad. Sci. U.S.A.* **75**, 4853 (1978).
10. Seif, I., Khoury, G. & Dhar, R. *Nucleic Acids Res.* **6**, 3387 (1979).
11. Rogers, J. & Wall, R. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
12. Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. & Steitz, J. A. *Nature* **283**, 220–224 (1980).
13. Gilbert, W. *Nature* **271**, 501 (1978).
14. Tiemeier, D. C. *et al. Cell* **14**, 237 (1978).
15. Nishioka, Y. & Leder, P. *Cell* **18**, 875 (1979).
16. Roychoudhury, R., Jay, E. & Wu, R. *Nucleic Acids Res.* **3**, 101 (1976).
17. Maxam, A. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560 (1977).
18. Kaback, D., Angerer, L. & Davidson, N. *Nucleic Acids Res.* **6**, 2499 (1979).
19. Thomas, M., White, R. L. & Davis, R. W. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2294 (1976).

APPENDIX 5

EUCARYOTIC GENE REGULATION

THE ORGANIZATION AND REARRANGEMENT OF HEAVY CHAIN IMMUNOGLOBULIN GENES IN MICE[1]

M. Davis, P. Early, K. Calame, D. Livant, and L. Hood

Division of Biology, California Institute of Technology, Pasadena, California 91125

ABSTRACT A preliminary analysis of several heavy chain variable (V) and constant region (C) gene segments from sperm (undifferentiated) and myeloma (differentiated) DNA has revealed the following: 1) the V_H and C_α genes are separate in the germ line; 2) the V_H and C_α genes are rearranged during the differentiation of the antibody-producing cell; 3) multiple rearranged C_α genes are present in the DNA of a single myeloma tumor; 4) small intervening sequences may separate the domains of the α and μ constant region genes; and 5) at least 8-9 germ line V_H genes exist for antibodies binding phosphorylcholine.

INTRODUCTION

The antibody gene families have several interesting organizational features. There are three distinct gene families - two code for light (L) chains, λ and κ, and the third codes for heavy (H) chains. They are composed of three distinct coding segments which are separated from one another by intervening DNA sequences - V (variable), J (joining) and C (constant). The V and J segments together comprise the V region of the antibody polypeptide which encodes the immunoglobulin domain concerned with antigen recognition. Moreover, each antibody gene family appears to contain multiple V and J segments.

The antibody gene families present two fascinating biological problems. First, it has been estimated that mammals can synthesize 10^5 to 10^8 different antibody molecules. What genetic mechanisms are responsible for this diversity of antibody molecules? We hope to assess the relative contributions of three genetic mechanisms: multiple germ line V genes (1), somatic mutation (2), and the joining in a combinatorial fashion of multiple V and J segments (3). Second,

[1]This work was supported by NSF grant PCM 76-81546.

ISBN 0-12-068350-4

how are antibody gene segments rearranged during the differentiation of antibody-producing cells? These DNA rearrangements presumably are fundamental components of the molecular events that commit the antibody-producing cell to the synthesis of a single type of antibody molecule as well as contributing to antibody diversity in the combinatorial joining of V and J segments (3,4).

We have focused on the analysis of the heavy chain gene family because, in addition to being an excellent system for studying the phenomena mentioned above, it has intricacies not exhibited in light chains. The heavy chain gene family of the mouse is comprised of an unknown number of variable (V_H) gene segments and at least eight different constant (C_H) gene segments (5) (Figure 1).

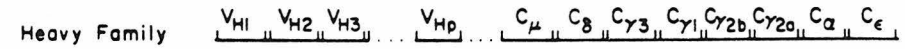


FIGURE 1. Heavy chain antibody gene family in mice. The order of C_H gene segments is uncertain, although indirect evidence supports the following alignment: $C_{\gamma 3}C_{\gamma 1}C_{\gamma 2b}C_{\gamma 2a}C_{\alpha}$ (20). The number of V_H gene segments is still a matter of controversy. The heavy chain gene family also has multiple J segments that are not depicted in this figure (see text).

The various classes and subclasses of immunoglobulins are determined by the C_H gene segments (e.g., C_{μ}-IgM, C_{γ}-IgG, C_{α}-IgA, etc.). Moreover, during the differentiation of the antibody-producing cell, distinct classes of immunoglobulins are expressed in a reproducible order (Figure 2). First IgM is expressed; later IgD and IgM are expressed; and eventually the other classes of immunoglobulins are expressed (6). In the lineage of a particular antibody-producing cell, it appears that these developmental shifts in immunoglobulin class expression occur by associating a particular V_H gene segment with different C_H gene segments while maintaining the expression of the same light chain gene segments. Therefore, a question of particular interest is the nature of the DNA rearrangements which lead to sequential and at times, simultaneous, expression of different heavy chain classes. Fortunately, tumors of antibody-producing cells exist which "freeze" this developmental pathway at many different points. Thus in time we will understand how the antibody gene organization for sperm cells (undifferentiated DNA) differs from that of tumor cell lines producing IgM, IgM + IgD and IgA (i.e., various stages of differentiation). Accordingly, our our initial efforts are focused on understanding the gene

organization in DNA at the beginning (sperm or embryo) and the end (IgA-producing myeloma) of a heavy chain differentiation pathway.

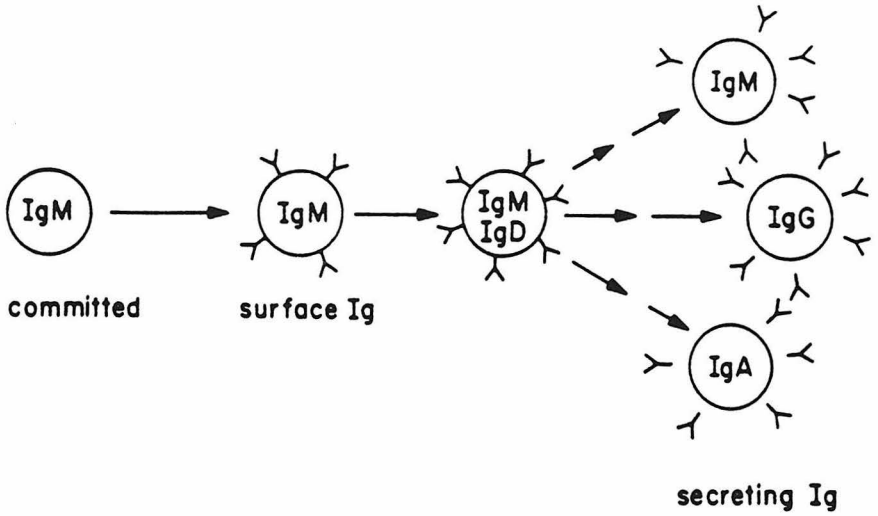


FIGURE 2. The differentiation of B cells. A B cell first becomes committed to the expression of a particular V domain (one V_L region and one V_H region) which is associated with cytoplasmic IgM molecules. Subsequently the IgM molecule is expressed on the cell surface. Later, cell-surface IgD molecules appear. Subsequent differentiation events lead to a terminally differentiated cell which specializes in the synthesis of soluble antibodies of one of a variety of immunoglobulin classes. For an individual B cell, the same V domain is associated with the various classes of immunoglobulins throughout the differentiation pathway.

THE PHOSPHORYLCHOLINE ANTIBODY SYSTEM

We have chosen to examine some of the questions posed above for a series of antibody-producing cells which synthesize immunoglobulin binding phosphorylcholine because this system allows us to analyze directly the biology of the immune response to phosphorylcholine (PC). Let us summarize the salient features of this system. First, several thousand myeloma tumors have been screened and twelve appear to

synthesize immunoglobulins binding phosphorylcholine (7). Our laboratory has determined the amino acid sequences of the V_H regions for seven of these tumors (8,9) and other laboratories have analyzed several additional sequences (10) (Figure 3). The V_H sequences from myeloma proteins binding phosphorylcholine illustrate several features of V diversity. 1) Four V_H sequences are identical. Since these identical V_H sequences were expressed independently in different mice, it appears that they are encoded by a germ line V_H gene segment designated T15. This reasoning argues that it is unlikely that four somatic variants would be identical in amino acid sequence. 2) The variant sequences differ by one to eleven amino acid substitutions and also exhibit sequence gaps. Accordingly, one can hope to determine the nature and extent of diversity generated from somatic genetic mechanisms by sequencing germ line PC V_H gene segments and comparing them with the protein diversity patterns reflected in their myeloma counterparts. Second, antisera have been raised which are specific for the V domains of several myeloma proteins binding phosphorylcholine. These antisera are termed anti-idiotypic antisera. Anti-idiotypic antisera to T15 can be used to map genetic elements which control the expression of this V_H domain. The T15 idiotype maps about 0.4 centiMorgans (cM) from the C_H gene cluster (11) and simplistic genetic calculations suggest the PC V_H and C_H gene segments are separated by hundreds of thousands or even a million nucleotides. For example, mouse chromosomes have about 25 chiasmata per meiosis (12). With a genome of 3×10^9 nucleotide pairs, 0.4 cM of DNA in the mouse would span about 10^6 nucleotide pairs, if meiotic recombination were random. Third, the T15 idiotype appears to be present on at least one type of T cells ("helper T cells") (13), implying that T-cell receptors and B-cell immunoglobulins may share the same V_H repertoire of genes. Thus an analysis of the phosphorylcholine system may provide opportunities to analyze T-cell receptors. Finally, the hybridoma system of Milstein and Köhler (14) has been employed to generate homogeneous antibodies to phosphorylcholine. In collaboration with Dr. Patricia Gearhart, we are analyzing 20 hybridomas to phosphorylcholine in order to broaden our knowledge about the phenotypic diversity patterns of the phosphorylcholine system. The importance of detailed protein sequence studies on the products of complex multigenic systems such as the antibody gene families cannot be overemphasized, for these phenotypic diversity patterns are one of the end results of heavy chain gene organization and rearrangements and any meaningful understanding of this system at the DNA level must account for the resultant diversity of its gene products. Thus we hope the phosphorylcholine system will provide

insights into antibody gene diversity and organization and permit us, in time, to begin analyzing the more complex regulatory events of this sophisticated system.

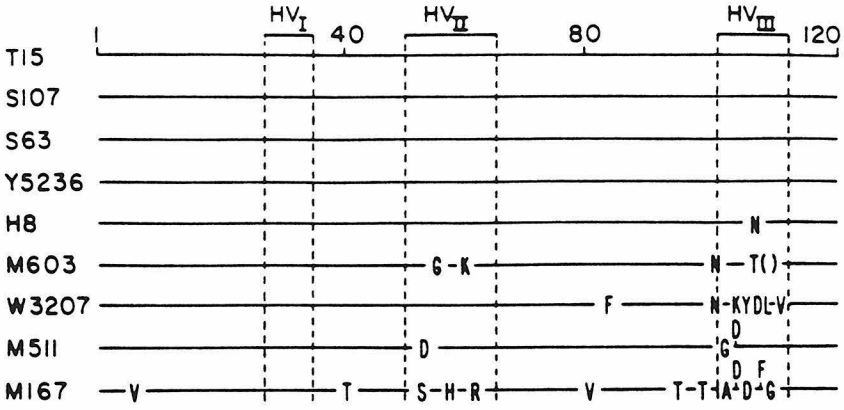


FIGURE 3. The amino acid sequences of V_H regions from immunoglobulins binding phosphorylcholine. Identities of these sequences to the V_H region of T15 are indicated by a straight line. The one letter code of Dayhoff is used to indicate amino acid substitutions (28). Deletions are indicated by brackets. Insertions are denoted by a vertical bar. The three hypervariable regions which fold in three dimensions to constitute the walls of the antigen-binding site of the V domain are designated by HV_I, HV_{II}, HV_{III} and dotted lines.

OUR APPROACH

We have constructed libraries in Charon 4A bacteriophage from partial restriction digests of sperm, embryo, and myeloma DNA (15, 16). The sperm and embryo libraries are a source of undifferentiated DNA. The myeloma library, derived from the tumor MOPC 603 which synthesizes IgA molecules binding phosphorylcholine, represents a terminal stage in the differentiation of an antibody-producing cell. We also have purified mRNA from a variety of myeloma tumors, and used these as templates for the synthesis of double-stranded DNA copies which were then inserted into plasmids (16). Our initial approach has been to compare the genomic organizations of undifferentiated (sperm or embryo) and differentiated (IgA myeloma tumor) DNAs. To this end we have isolated a number of genomic clones from both the M603 library and from a sperm library, using cDNA probes for the complete $V_H C_\alpha$ coding region of myeloma protein S107. The V_H regions of the

S107 and the M603 immunoglobulins are very closely related (Figure 3) and the corresponding mRNAs completely protect one another in S1 nuclease digestion experiments (16). Certain of these initial experiments have recently been published in a paper which describes for the first time a heavy chain genomic clone containing the V_H and C_α gene segments and the presence of intervening sequences within the C_α coding region, probably separating the coding regions for immunoglobulin α domains (16). These results as well as more recent observations are summarized below.

EXPERIMENTAL OBSERVATIONS

The Variable and Constant Regions of α Heavy Chains Appear to be Encoded by Distinct V_H and C_α Gene Segments which are Rearranged During Differentiation. We have analyzed a series of overlapping genomic clones from the M603 library which have the general structures illustrated in Figure 4. The V and the C gene segments are separated by 6.8 kilobases. Furthermore, idiotypic mapping, discussed above, suggests that these regions were separated by hundreds of thousands of nucleotides prior to differentiation of this antibody-producing cell with the concomitant DNA rearrangements. A heteroduplex comparison of a sperm V_H clone with the myeloma M603 clone, which will be discussed subsequently, also provides evidence for the rearrangement of the V_H gene segment in the myeloma DNA. Accordingly, the V_H and C_α gene segments are originally widely separated from one another. As the antibody-producing cell differentiates, DNA rearrangements of antibody V and C gene segments occur over extensive stretches of DNA.

The C_α Gene Segments from the M603 Myeloma Library are Present in Multiple Rearranged Forms. A comparison of Southern blots on sperm M603 DNA using the C_α probe demonstrates that the myeloma DNA has three forms of the C_α gene, none of which are identical to their germ line counterpart (Figure 5). These three forms have been isolated from the M603 library as Charon 4A clones (Figure 6). Restriction enzyme analyses and heteroduplex comparisons demonstrate that, although they share 2.7 or more kilobases of homology just 5' to the C_α gene, each of these three clones is distinct from the others in their more 5' regions.

These observations raise several interesting possibilities. The absence of a germ line-like C_α gene segment in the M603 DNA suggests that the C_α gene segments in both the maternal and paternal chromosomes coding for heavy chain genes have been rearranged. Immunoglobulin-producing cells exhibit allelic exclusion; that is, a particular antibody-

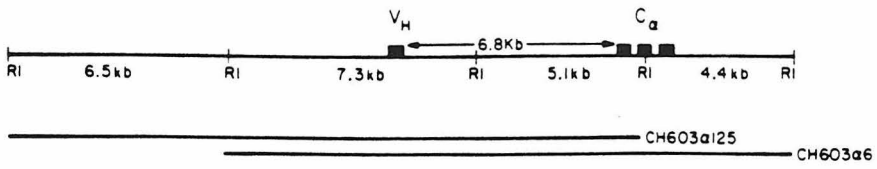


FIGURE 4. The organization of V_H and C_α gene segments from DNA derived from myeloma tumor M603. Kb denotes kilobases. R1 denotes Eco R1 cleavage sites. The distances between Eco R1 sites are indicated. CH603α125 and CH603α6 are two clones derived from the phage library of M603 DNA. The V_H gene segment is separated from the Cα gene segment by 6.8 kilobases of intervening DNA. R-loop mapping and restriction enzyme analyses demonstrate that the C_α segment is divided into three approximately equal segments, presumably coding regions for the three C_α domains, by two small intervening DNA sequences (16).

producing cell may express the maternal or paternal allele for a particular immunoglobulin family, but not both alleles. In the past the phenomenon of allelic exclusion has been explained by suggesting that either the maternal or paternal chromosome does not rearrange at the DNA level and, accordingly, cannot express an immunoglobulin polypeptide. This suggestion has come from Southern blot analyses of myeloma DNAs in which the germ line pattern of constant gene segments for light chains appears to be preserved (17). Our data on the alpha constant region genes of the M603 myeloma DNA suggests that both the maternal and paternal chromosomes undergo rearrangements, but that one of these rearrangements is abortive in the sense no gene product is expressed. It will be interesting to determine whether these abortive DNA rearrangements include V gene segments; or whether only the C gene segment is involved in the rearrangement. Moreover, it will be interesting to analyze carefully the myeloma examples that appear to have germ line C fragments to determine whether the DNA rearrangements have been missed due to technical limitations of the Southern blotting technique, or contamination with somatic DNA. It may be that all myeloma DNAs in fact rearrange both the paternal and maternal chromosomes--one in a productive and the second in an abortive fashion.

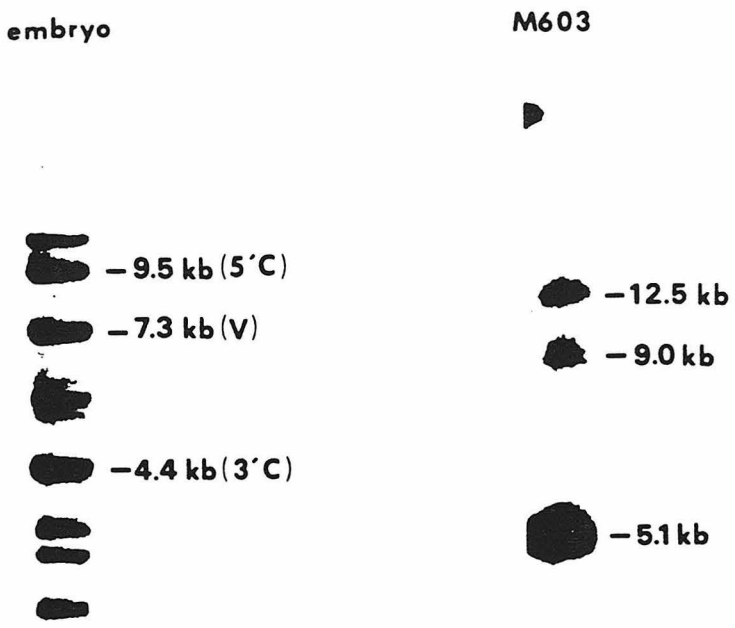


FIGURE 5. Southern blots of embryo (undifferentiated) and myeloma M603 (differentiated) DNAs. The picture on the left is a Southern blot of 13-day embryo DNA after digestion with the Eco R1 enzyme, separation of the DNA fragments on agarose, and hybridization with a cDNA probe derived from mRNA of myeloma tumor S107. This probe contains both the V_H and C_α coding regions. Assignments of the C_α fragments are based on Southern blots with separated V_H and C_α probes (data not shown). The remaining fragments must be V_H gene fragments. Thus there are at least 8-9 germ line V_H genes which cross-hybridize with the V_H probe from myeloma tumor S107. The exposure on the right is a Southern blot of tumor M603 DNA after Eco R1 digestion and hybridization to a plasmid containing the 5' half of the C_α coding region (an R1 site separates the 5' from the 3' half of the C_α gene segment; see Figure 4). The 5' C_α probe gives just one 9.5 kilobase band in the embryo DNA (data not shown) and 5.1, 9.0 and 12.5 kilobase bands in the M603 DNA. Hybridization to the 3' half of the C_α coding region gives a 4.4 kilobase band in both embryo and myeloma DNA (not shown).

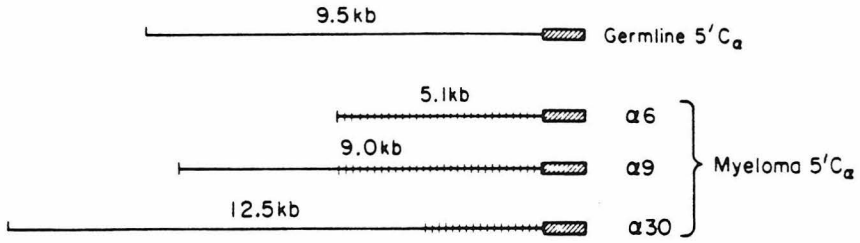


FIGURE 6. Eco R1 genomic fragments including the 5' portion of the C_α gene from myeloma M603 DNA and sperm DNA. The genomic clones α6, α9, and α30 have been derived from the M603 phage library. The structure of the germ line C_α clone comes from a Southern blot analysis of sperm or embryo DNA (Figure 5). The boxes represent the 5' potion of the C_α coding sequence (see Figure 4), whereas the hashmarks represent DNA homologies revealed by heteroduplex analyses.

One surprising observation that is difficult to explain is the presence of three distinct C_α clones in the M603 DNA. Several explanations may be offered, none really satisfactory. First, the germ line may contain two C_α genes, both the same size by Eco R1 restriction analysis. Both of these C_α genes may undergo rearrangements of several different types. Second, perhaps the abortive rearrangement is unstable and may be subject to additional DNA rearrangements. Third, perhaps there are several different M603 cell types in the uncloned tumor from which the DNA was derived. The possibility that the M603 C_α pattern is some aberration of this particular tumor line seems unlikely because at least one other phosphorylcholine binding tumor (H8) has an identical pattern on Southern blots (M. Davis and P. Early, unpublished). Thus in the case of the C_α gene segments, it appears that both the maternal and paternal chromosomes undergo DNA rearrangements, some of which are abortive (nonproductive) while others lead to the expression of one V_H-C_H pair of gene segments.

The V and C Rearrangements in Heavy Chains Resemble Those of Light Chains in Some Respects but Not Others. The V_L and C_L gene segments are rearranged by a fusion at the DNA level of V_L and J_L gene segments with the removal (or rearrangement) of the intervening DNA (Figure 7) (4, 17). Accordingly, the DNA 5' to the V_L gene segment is identical to that of the unrearranged V_L gene and the intervening DNA between the V and C gene segments is derived from the region

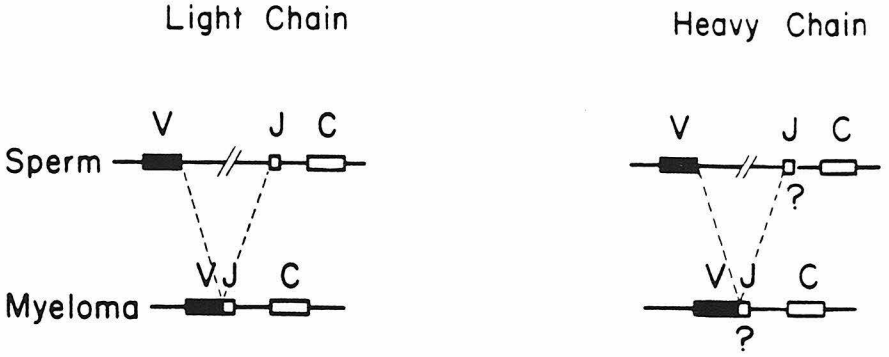


FIGURE 7. A model of the joining of light and heavy chain gene segments. An analysis of λ (4) and κ (18) light chain gene segments indicate that the 3' side of a V segment is fused to the 5' side of a J segment. The intervening DNA sequence between the J segment and the C segment remains unchanged in the DNA rearrangement process. The heavy chain gene segments appear to rearrange in a similar fashion, although the organization of the intervening DNA sequence between the J and C gene segments is altered, presumably because of multiple DNA rearrangements between one V_H gene segment and two (or more) C_H gene segments (see text).

5' to the unrearranged C_L gene. The existence of J_H segments for heavy chains is strongly implied from protein sequence data (18) and has recently been demonstrated by the DNA sequence analysis of a sperm clone containing a V_H segment (P. Early and M. Davis, unpublished observation). Comparison of a sperm V_H clone and the joined V_H and C_α myeloma clone (α6) by DNA heteroduplex analysis demonstrates that those regions 5' to the V segment are homologous and those regions 3' to the V segment are nonhomologous (Figure 8). In this respect the heavy chain variable region gene segment appears to rearrange in a manner similar to its light chain counterparts (Figure 7).

The rearrangement of V_H and C_H gene segments differs from those of the light chains in one important regard. Certain of the intervening sequences between the V_H and C_α gene segments of the α6 clone (Figure 4) are not derived from germ line DNA 5' to the C_α gene. For example, a Southern blot analysis of germ line DNA with a C_α probe shows that the closest Eco R1 site is 9.5 kilobases from the 5' side of the C_α gene segment (Figure 5). However, the α6 clone

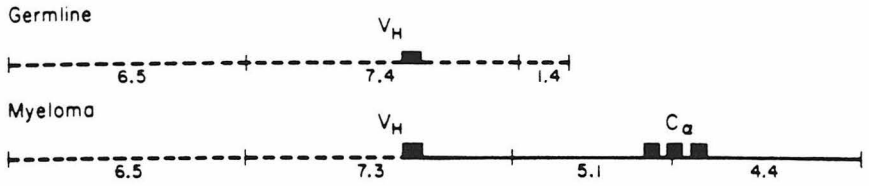


FIGURE 8. Homologies determined by heteroduplex analysis between the flanking sequences of germ line and myeloma V_H clones. The dotted lines indicate germ line sequences. Accordingly, the intervening DNA sequence between the V_H and C_α gene segments is not derived from the sperm V_H clone. The sperm library was constructed by M. Davis and R. Joho.

from the M603 DNA has an Eco R1 site 5.1 kilobases from the 5' end of the C_α gene segment. In addition, as discussed above, this α6 DNA is not homologous to DNA of the sperm V_H clone (Figure 8). Moreover, the Eco R1 site of the α6 clone in the DNA between the V and C gene segments does not seem to have been created by a spurious mutation, since Southern blots of DNA from an independently derived tumor line (H8) show the same C_α Eco R1 fragment. One explanation for the origin of the DNA sequence between V_H and C_α gene segments in the α6 clone containing this Eco R1 site is that it arises from the DNA rearrangement events of an earlier stage in differentiation, in which this V_H gene segment was formerly joined to a different C_H (or J) gene segment. Indeed, during the differentiation of antibody-producing cells, the V_H gene segment appears initially to be joined to a C_μ gene (Figure 2), so we would predict that some of the intervening DNA in the α6 clone between the V_H and C_α gene segments may be derived from the 5' side of a germ line C_μ gene segment. The subsequent joining of this V_H segment to a C_α gene segment later in development might displace or delete (19) the C_μ gene, but not all of its flanking sequences.

Intervening Sequences Appear to Separate the Domains of the C_H Genes. The C_α polypeptide is divided into three discrete molecular domains, each of which encompasses about 110 amino acid residues (20). We initially used R-loop mapping to demonstrate the existence of two small intervening sequences (IVS2, IVS3) which separate the C_α coding region into three roughly equal segments (Figure 4) (16). Subsequent restriction enzyme analyses of the M603 genomic clone (α6) places IVS2 within 30 amino acids of the domain boundary

between the $C_{\alpha}1$ and $C_{\alpha}2$ homology units (16; M. Davis, unpublished). Thus it appears likely that the two intervening sequences will separate the C_{α} gene into three distinct coding segments, one for each C_{α} domain (Figure 4). In addition, we have analyzed a μ genomic clone from the M603 library by R-loop mapping. The C_{μ} region has four domains (21) and, as expected, R-loop analysis demonstrates that the C_{μ} coding region is divided by three small intervening sequences into four roughly equivalent segments (K. Calame, P. Early, M. Davis, D. Livant, unpublished observations). The analysis of a genomic γ1 clone has established that intervening sequences separate the three $C_{\gamma 1}$ domains and the hinge region from one another precisely at the interdomain boundaries (22). Therefore it appears reasonable to conclude that intervening sequences will divide all of the immunoglobulin C genes coding into segments for structural domains (see Figure 1).

The function of intervening sequences has generated spirited controversy and discussion. Individual domains of the immunoglobulin molecule carry out discrete and independent functions (20). Accordingly, the immunoglobulin intervening sequences appear to perform the important task of breaking the coding regions into discrete units which may then rearrange independently of one another through recombination at either the DNA level or the nuclear RNA level as proposed by Gilbert (23). Several lines of evidence suggest that the domains of immunoglobulins may be discrete evolutionary units. First, C_H regions with two, three, and four domains are present in vertebrate antibodies. Second, heavy chain disease deletions (24) and spontaneous deletions in tissue culture lines (25) suggest that frequent nonhomologous crossing-over occurs at or between domain boundaries. Perhaps intervening sequences not only separate domains but facilitate recombination as well. It will certainly be interesting to determine the homology relationships, if any, of the various immunoglobulin intervening sequences to one another.

The Germ Line V Gene Segments of Mouse Heavy Chains Appear to be as Diverse as Their V_K Counterparts. The V_H regions derived from myeloma proteins binding phosphorylcholine show a limited range of heterogeneity (Figure 3). We are interested in determining whether these different V_H sequences are germ line or in part derived by somatic mutation. Southern blot analysis of embryo DNA employing the S107 cDNA probe reveals at least 8-9 restriction fragments which hybridize to the S107 V region probe (Figure 5). The PC V_H regions represent a single group of heavy chain variable regions (26). Approximately 20 other groups of

V_H regions have been defined (26). Therefore, if each group is on the average encoded by ~10 germ line genes, the heavy chain gene family may be comprised of approximately 200 V_H genes. Since the amino acid sequence analyses of mouse V_H regions are relatively limited, it appears likely that in time many additional V_H groups will be defined. By similar analyses, the V_K family of mouse appears to be encoded by 200 or more germ line V genes (3, 27). We have isolated several different PC V_H genes and are now in the process of sequencing them to determine the relative contributions of germ line diversity, somatic mutation, and combinatorial joining of V_H and J_H segments to antibody variability.

The Generality of Nucleic Acid Rearrangements. The intriguing general question posed by the studies on immunoglobulin genes is whether DNA rearrangements are a fundamental aspect of differentiation in other eukaryotic systems. An answer to this question will await more detailed analyses of other gene families, both simple and complex.

ACKNOWLEDGMENTS

The work here is supported by National Science Foundation Grant PCM 76-81546. MD, PE, and DL are supported by National Institutes of Health Training Grant GM 07616. KC is supported by National Institutes of Health Fellowship GM 05442.

REFERENCES

1. Hood, L., Campbell, J. H., and Elgin, S. C. R. (1975). Ann. Rev. Genet. 9, 305.
2. Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R., and Weigert, M. (1974). "The Immune System," ICN-UCLA Symp., p. 89. Academic Press.
3. Weigert, M., Gatmaitan, L., Loh, E., Schilling, J., and Hood, L. (1978). Nature 276, 785.
4. Brack, C., Hirawa, M., Lenhard-Schueller, R., and Tonegawa, S. (1978). Cell 15, 1.
5. Mage, R., Lieberman, R., Potter, M., and Terry, W. (1973). In "The Antigens" (M. Sela, ed.), Vol. I, p. 300. Academic Press.
6. Goding, J. W., Scott, D. W., and Layton, J. E. (1977). Immunol. Rev. 37, 152.
7. Potter, M. (1970). Physiol. Rev. 52, 631.
8. Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M., and Schilling, J. (1976). Cold Spring Harbor Symp. Quant. Biol. 41, 817.

9. Hubert, J., Johnson, N., Barstad, P., Rudikoff, S., and Hood, L. In preparation.
10. Rao, D. N., Rudikoff, S., and Potter, M. (1978). Biochemistry 17, 5555.
11. Riblet, R. J. (1977). "Molecular and Cellular Biology," ICN-UCLA Symp., Vol. 6, p. 83. Academic Press.
12. Klein, J. (1975). "The Biology of the Mouse Histocompatibility Complex." Springer-Verlag.
13. Cosenza, H., Augustin, A., and Julius, M. (1977). Cold Spring Harbor Symp. Quant. Biol. 41, 709.
14. Köhler, G., and Milstein, C. (1976). Eur. J. Immunol. 6, 511.
15. Maniatis, T., Hardison, R., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G., and Efstratiadis, A. (1978). Cell 15, 687.
16. Early, P., Davis, M., Kaback, D., Davidson, N., and Hood, L. (1979). Proc. Nat. Acad. Sci. USA 76, 857.
17. Seidman, J. G., and Leder, P. (1978). Nature 276, 790.
18. Schilling, J., Clevinger, B., Davie, J., and Hood, L. In preparation.
19. Honjo, T., and Kataoka, T. (1978). Proc. Nat. Acad. Sci. USA 75, 2140.
20. Edelman, G. M., Cunningham, B. A., Gall, W., Gottlieb, P., Rutishauser, U., and Waxdal, M. (1969). Proc. Nat. Acad. Sci. USA 63, 78.
21. Beale, D., and Feinstein, A. (1976). Quart. Rev. Biophys. 9, 135.
22. Sakano, H., Rogers, J. H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R., and Tonegawa, S. (1979). Nature 277, 627.
23. Gilbert, W. (1978). Nature 271, 501.
24. Frangione, B., Lee, L., Haber, E., and Bloch, K. (1977). Proc. Nat. Acad. Sci. USA 70, 1073.
25. Adetugbo, K., Milstein, C., and Secher, D. (1977). Nature 265, 299.
26. Barstad, P., Rudikoff, S., Potter, M., Cohn, M., Konigsberg, W., and Hood, L. (1974). Science 183, 962.
27. Seidman, J., Leder, A., Nau, M., Norman, B., and Leder, P. (1978). Science 202, 11.
28. Dayhoff, M. O. (1972). In "Atlas of Protein Sequence and Structure," Vol. 5, Biomedical Research Foundation, Washington, D.C.