MOUSE IMMUNOGLOBULIN HEAVY CHAIN GENE ORGANIZATION AND REARRANGEMENT: GENETIC BASES FOR ANTIBODY DIVERSITY AND REGULATED EXPRESSION

Thesis by

Philip Warren Early

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1980

(Submitted March 28, 1980)

To my parents and my teachers

Acknowledgements

I am particularly indebted to Norman Davidson and Lee Hood, who first encouraged me to work on mouse immunoglobulin genes. In my first years at Caltech, I learned many of the techniques of eukaryotic molecular biology in Norman Davidson's laboratory. I also benefitted from advice or materials given by many other people at Caltech and elsewhere, whose assistance is acknowledged in each of the following chapters.

Most of the work reported in my thesis was done in collaboration with other investigators. The paper in Chapter 2 was mainly a collaboration with Mark Davis, who prepared the phage libraries described in this and the following papers, and who did much of the characterization of the rearranged $V_H C_\alpha$ clone. The electron microscopy, which led to the discovery of introns between domains, was performed by David Kaback.

The symposium paper in Chapter 3 summarizes work done by both Mark Davis and myself in isolating and characterizing germline and rearranged immunoglobulin genes. Electron microscopy was the work of Kathryn Calame.

The results and conclusions of the paper in Chapter 4 are largely my own. Mark Davis and Kathryn Calame isolated and mapped the germline J_H clone, and Henry Huang sequenced the rearranged M603 V_H gene. The information about heavy chain switching referred to in this paper is derived from work mainly done by Mark Davis and Kathryn Calame [Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L., and Hood, L. (1980). Nature 283, 733-739.].

Chapters 5 and 6 of my thesis are the results of a collaboration with John Rogers in Randy Wall's laboratory at UCLA. The existence of a "fifth exon" in the C_μ gene was suggested by the R-loop results of Kathryn Calame. I made the observation that two of my μ cDNA clones had different 3' ends, which hybridized to separate exons in the C_μ gene. John Rogers sequenced these two cDNA clones, while I sequenced the membrane μ exons. The RNA blots were done by John Rogers and Randy Wall, using cells and RNA provided by Martha Bond. Mark Davis isolated the C_μ genomic clone.

I thank Connie Katz and Vickie Oldenburg for prompt typing, and the Jean Weigle Memorial Fund for support in the preparation of this thesis.

Abstract

Immunoglobulin heavy chains each display one of a wide range of diverse antigen-binding variable regions. At least one class of immunoglobulin, IgM, contains heavy chains which exist as two forms, either bound to the outside of a cell membrane or linked by disulfide bonds in secreted antibodies. I have used recombinant DNA techniques to isolate and determine the nucleotide sequences of genes encoding mouse immunoglobulin heavy chains. This has enabled me to examine genetic bases for the diversity of heavy chain variable regions and for the synthesis of membrane-bound and secreted forms of IgM heavy chains.

I found that genes encoding heavy chain variable regions are created somatically by joining three segments of DNA: V_H, D, and J_H. The J_H gene segments are closely linked to the IgM heavy chain constant region gene in germline DNA, where they are widely separated from V_H gene segments. In an immunoglobulin-producing cell, one V_H and one J_H gene segment are joined, together with a D sequence which is probably also a germline gene segment, to form the expressed heavy chain variable region gene. Both combinatorial association of gene segments and variations in the exact sites of DNA joining between gene segments can contribute to heavy chain variable region diversity. Based on observations of certain conserved nucleotides and spacer sequences adjacent to unrearranged immunoglobulin gene segments, I propose a mechanism for variable region gene rearrangement during differentiation.

Secreted and membrane-bound forms of IgM heavy chains were found to be encoded by separate mRNAs transcribed from the same gene. These mRNAs differ only at their 3' ends, where one encodes a 20 amino acid secretory C-terminal segment, and the other encodes a 41 amino acid transmembrane C-terminal segment. Synthesis of the two forms of IgM heavy chain mRNA appears to be developmentally regulated by controlling the site of 3' terminal polyadenylation. The site of polyadenylation defines the lengths of RNA transcripts and thereby determines which of two alternative RNA

splicing patterns will be followed, leading to mRNAs encoding either the secreted or membrane-bound forms of IgM heavy chains.

CONTENTS

Introduction

When I began working on mouse immunoglobulin genes in early 1975, it was with two major questions in mind. One was whether somatic DNA rearrangement was necessary for the expression of immunoglobulin genes, and the other was what the genetic basis was for the vast repertoire of antibody diversity. As we now know, these questions are largely the same. The DNA rearrangements which are required to form a complete immunoglobulin gene also play a key role in generating the diversity of immunoglobulin variable regions. This understanding was first reached for immunoglobulin light chains[1-3], and the work included in Chapters 2-4 of my thesis has extended it to the more complex case of immunoglobulin heavy chains.

Chapter 2 of my thesis contains the first report of functional DNA rearrangements in heavy chain genes. This paper showed that in a differentiated IgA-producing myeloma cell, the (presumptive) expressed V_H gene was only 6.8 kb from the C_α gene, rather than the much greater distance which separated these genes in the mouse germline. We also showed that, within the limits of resolution by electron microscopy, each structural domain of the α constant region is encoded by a separate exon in the C_α gene. Other investigators subsequently demonstrated this to be the case for C_γ and C_μ genes[4-6], supporting a role for exon duplication in the evolution of compound C_H genes.

The third chapter presents an overview of our laboratory's initial comparison of germline and differentiated heavy chain genes. The paper in Chapter 3 includes my genomic Southern blot showing that mouse embryo DNA contains 8-9 V_H genes homologous to a cloned cDNA probe derived from S107 heavy chain mRNA. This result demonstrated that multiple related V_H genes do exist in the undifferentiated genome, contributing to antibody diversity[7].

The manuscript in Chapter 4 of my thesis defines the germline origins of V_H genes. I isolated and sequenced two types of gene segments which join to form rearranged V_H genes: a V_H gene segment (codons 1-101) and a J_H gene segment (codons 107-123).

By comparing these germline gene segments to sequences from rearranged V_H genes, I determined that the germline genome probably contains a third type of gene segment, D, which includes codons 102-106 of the V_H gene. Gene segments of this type have not yet been isolated from germline DNA, and this remains an active area of my research. Combinatorial assortment of moderate numbers of V_H, D, and J_H gene segments during DNA joining can generate large numbers of rearranged V_H genes. In conjunction with variations in the precise points of DNA junctions between gene segments, this process probably accounts for most of the diversity of heavy chain variable regions.

The heavy chain gene segments I sequenced and light chain gene segments characterized by other investigators[1-3] are all flanked by the same conserved noncoding nucleotides. As explained in Chapter 4, these are probably the recognition sequences for enzymes which join immunoglobulin gene segments. I observed that there are actually two types of recognition sequences, in each of which the steric relationships between nucleotides presumably involved in enzyme recognition are strongly conserved. It appears that DNA joining always occurs between gene segments bordered by recognition sequences of different types. On this basis, I have proposed a model for the rearrangement of variable region gene segments, including the postulated heavy chain D gene segment.

The fifth and sixth chapters of my thesis are manuscripts dealing with the control of immunoglobulin gene expression by means other than DNA rearrangement. Immunologists had debated whether there are two forms of the μ constant region[8, 9], since IgM molecules can be found either as monomers bound to the surface of B-cells, or as secreted pentamers held together by disulfide bonds to a J chain (no relation to J gene segments). Among μ chain cDNA clones I made from M104E mRNA, I discovered two types with different 3' ends. In collaboration with J. Rogers and R. Wall at UCLA, our laboratory showed that these two μ cDNA clones are derived from two species of μ mRNA, one encoding secreted μ chains and the other encoding membrane-bound μ chains. The two forms of μ chain differ only at their C-termini, where secreted μ chains have a 20 amino

acid secretory segment and membrane-bound μ chains have a 41 amino acid segment with the characteristics of a transmembrane peptide.

By DNA sequencing, I determined that the membrane-specific C-terminus was encoded by two exons located about 2 kb 3' to the other exons of the C_μ gene. Both secreted and membrane μ mRNAs are derived from transcripts of the same μ gene. Depending on the length of the transcript, the resulting mRNA may either encode the secreted μ C-terminus or undergo RNA splicing to replace the secreted C-terminus with the membrane-bound C-terminus. The length of μ transcripts evidently is controlled at one of two possible points for 3' terminal polyadenylation. The C_μ gene thus seems to represent the first known example of regulated RNA processing altering the expression of a eukaryotic gene at different developmental stages.

1. Bernard, O., Hozumi, N., and Tonegawa, S. (1978). Cell 15, 1133-1144.

2. Max, E. E., Seidman, J. G., and Leder, P. (1979). Proc. Nat. Acad. Sci. USA 76, 3450-3454.

3. Sakano, H., Hüppi, K., Heinrich, G., and Tonegawa, S. (1979). Nature 280, 288-294.

4. Sakano, H., Rogers, J. H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R., and Tonegawa, S. (1979). Nature 277, 527-533.

5. Gough, N. M., Kemp, D. J., Tyler, B. M., Adams, J. M., and Cory, S. (1980). Proc. Nat. Acad. Sci. USA 77, 554-558.

6. Calame, K., Rogers, J., Early, P., Davis, M., Livant, D., Wall, R., and Hood, L. (1980). Nature, in press.

7. 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, 817-836.

8. Williams, P. B., Kubo, R. T., and Grey, H. M. (1978). J. Immunol. 121, 2435-2439.

9. McIlhinney, R. A. T., Richardson, N. E., and Feinstein, A. (1978). Nature 272, 555.

Chapter 1: Technical Background

My first goal in the investigation of immunoglobulin genes was not the genes themselves, but rather the mRNAs encoding certain well characterized immunoglobulin chains. Very extensive protein sequence studies, largely done in Lee Hood's laboratory, had defined a group of related myeloma proteins which all bind the phosphorylcholine moiety[1]. The heavy chains from these myeloma proteins are generally quite similar to one another, but do have a range of amino acid differences, particularly concentrated in the third hypervariable region. This seemed an ideal system in which to study the genetic origins of antibody diversity. An mRNA encoding one of these immunoglobulin heavy chains was expected to hybridize with all genes containing similar sequences. In this manner, immunoglobulin mRNAs, or cDNAs derived from them, could be used to isolate and characterize genes encoding any portion of an immunoglobulin chain.

My efforts to isolate heavy chain mRNAs were aided considerably by a procedure developed in R. Perry's laboratory, which was communicated to me by O. Valbuena and K. Marcu in advance of publication[2]. I used myeloma tumors grown subcutaneously in Balb/c or CDF1 mice to prepare microsomal RNA by this procedure. Poly(A)-containing RNA was selected by an oligo(dT)-cellulose column, and then fractionated by size on a 10-30% isokinetic sucrose gradient. Absorbance profiles of two such gradients are shown in Figure 1, for RNA from 5107 (κ, α) and J606 (κ, γ_3) myeloma tumors. Two clear peaks can be seen in both cases, with sedimentation coefficients of approximately 13S and 16S. Pooled fractions from the peaks of these gradients were analyzed by electrophoresis in methylmercury-agarose gels, using E. coli and mouse ribosomal RNAs as markers[3]. In the case of S107 and J606, the peaks contained predominant RNA species about 1200 and 1800 nucleotides in length (Figure 2). Similar RNA preparations were made from S107, M603, M167, and W3207 myeloma tumors, all of which synthesize phosphorylcholine-binding immunoglobulins containing κ and α chains. RNA was also prepared from M315 (λ_{II}, α), J606 (κ, γ_3), and M104E (λ_I, μ) myeloma tumors, which synthesize immunoglobulins not binding phosphorylcholine.

In order to verify that my purified RNAs encoded immunoglobulin chains, I translated them into protein using the rabbit reticulocyte in vitro system[4]. Translation products of the large and small RNAs from W3207 tumors are shown in Figure 3, electrophoresed in parallel with authentic κ and α chains. As expected, the 1200 nucleotide RNA produced a κ chain precursor containing a signal peptide[5], while the 1800 nucleotide RNA produced a polypeptide which migrated slightly faster than the standard α chain. The latter observation probably is the result of a signal peptide in the in vitro product almost exactly compensating for lack of the carbohydrate which is present on the in vivo synthesized α chain.

My next step was to produce pure probes for immunoglobulin genes from these mRNAs, by using recombinant DNA techniques to clone cDNA copies. I synthesized double-stranded cDNAs from immunoglobulin mRNAs by sequential reactions with avian myeloblastosis virus reverse transcriptase (primed with oligo-dT) and E. coli DNA polymerase I. The procedures I used for these steps were modifications of methods developed at Stanford by D. Kemp (personal communication) and M. Wickens[6]. I obtained double stranded cDNAs of discrete size which appeared to be full-length copies of the immunoglobulin mRNAs, as illustrated in Figure 4. Due to the short oligo-dT primer, poly-A sequences were not completely copied. Consequently, "full-length" α chain cDNA was actually 1550 nucleotides long, compared to about 1800 nucleotides for the polyadenylated mRNA. I was able to use these cDNAs to map restriction enzyme sites, as shown for M603 α chain cDNA in Figure 5. These maps were later useful in analyzing plasmid clones derived from immunoglobulin cDNAs.

I used three different procedures to clone these double stranded cDNAs. In each case, I first removed the 5' terminal "hairpin" by S1 nuclease to generate open-ended DNA duplexes (Figure 4). Initially, I used terminal transferase to generate poly-T tails on the double stranded cDNAs, which were then annealed to pMB9 plasmid DNA tailed with poly-dA[7]. Calcium-manganese shock was employed to transform the enfeebled

E. coli strain χ1776 with the recombinant plasmids[8]. Colonies of bacteria containing plasmids were tested for the presence of immunoglobulin cDNA sequences by replica-filter hybridization[9] to immunoglobulin mRNAs labeled with ^{32}P by polynucleotide kinase (Figure 6). An α chain cDNA clone (p603α1) described in the second chapter of my thesis was derived by this dA·T tailing procedure. I also constructed a clone containing γ_3 chain sequences from J606 mRNA.

Subsequently, I produced cDNA clones by ligating synthetic oligonucleotides ("linkers") containing Eco RI restriction enzyme sites to the ends of double stranded cDNAs[10]. When cut with Eco RI, these cDNAs had "sticky ends" which permitted annealing and ligation to Eco RI-cut pMB9 plasmid DNA. I first used this procedure to make a cDNA clone containing nearly the complete sequence of M167 κ mRNA[11]. Later, I constructed clones from S107 α chain mRNA which are described in Chapters 2 and 4 of my thesis. I was able to achieve a moderately high efficiency for this procedure (200-400 clones per μg of cDNA) with α chain double stranded cDNA, since it contains a natural Eco RI site in the constant region sequence. This permits "half molecules" of cDNA with only a single synthetic Eco RI linker to be cloned.

The advantage of the linker method of cloning was that it yielded recombinant plasmids with intact Eco Rl sites at both ends of the inserted cDNA, allowing easy separation of cDNA sequences from the plasmid. A simpler cloning method which achieves the same purpose was developed by W. Rowekamp and R. Firtel at U. C. San Diego (personal communication). This involves tailing double stranded cDNA with oligo-dC, and annealing it to Pst I-cut pBR322 plasmid DNA tailed with oligo-dG. Since the oligo-dG tailing regenerates single stranded Pst I sites in the plasmid DNA, the inserted cDNA will be flanked by short dG·dC tails and Pst I sites in the resultant plasmid clone. I used this procedure to clone cDNAs from M104E μ chain mRNAs (Chapter 5).

The M104E heavy chain mRNA which I had prepared contained, in addition to RNA comigrating with 18S rRNA, a major species of RNA about 2200 nucleotides in

length and at least two minor species: one the same size as α and γ_3 mRNAs, and one somewhat larger than the 2200 nucleotide species (Figure 2). J. Rogers at UCLA has subsequently shown that the major species (his estimate of size is 2400 nucleotides) and the larger of the two minor species of M104E mRNA encode two forms of μ chains (Chapter 5). The identity of the M104E mRNA which comigrates with α and γ_3 mRNAs has not been determined. I isolated 15 cDNA clones which hybridized to M104E heavy chain mRNA. Thirteen of these clones apparently encode portions of the secreted M104E μ chain, and are derived from the major species in the M104E heavy chain mRNA preparation. However, I showed that two cDNA clones, p104μ6 and p104μ4, contain a different sequence which we have determined encodes membrane-bound μ chains. These clones are derived from the large minor species of M104E heavy chain mRNA (Chapter 5).

With immunoglobulin cDNA clones as hybridization probes, it was possible to turn to the immunoglobulin genes themselves to answer questions about gene expression and rearrangement and the origins of antibody diversity. Immunoglobulin genes were isolated by screening "libraries" of recombinant Charon 4A phage containing 12-20 kb fragments of mouse genomic DNAs, a technique pioneered in the laboratories of T. Maniatis and F. Blattner[12, 13]. Immunoglobulin cDNA clones labeled with ^{32}P were used to detect recombinant phage containing immunoglobulin genes. The phage libraries described in the following chapters of my thesis were prepared by another graduate student in this laboratory, Mark Davis, with whom I have collaborated in the isolation and characterization of many of our immunoglobulin genomic clones.

An example of phage library screening is shown in Figure 7. The procedure required two cycles to obtain a clone pure enough for analysis. In the first cycle, approximately 20,000 phage plaques from the library were grown on each of fifty 150 mm petri plates. Such large numbers of plaques are necessary to obtain adequate statistical representation of the 3×10^6 kb mouse genome, only 12-20 kb of which are present in each recombinant phage plaque. A replica nitrocellulose filter from one of these plates is

shown in Figure 7 after hybridization to a ^{32}P labeled immunoglobulin cDNA plasmid. The spots detected by autoradiography correspond to the positions on the original plate of phage plaques (and phage DNA) containing an immunoglobulin gene. A plug including one of these positive plaques was picked from the first plate and used to grow about 1000 plaques on a new plate. A replica filter from this plate was again hybridized to the labeled cDNA plasmid. A much higher percentage of positive plaques were seen in the second cycle (Figure 7), all derived from the single positive plaque in the plug from the first plate. Some of the positive second cycle plaques were sufficiently separated from contaminating plaques so that they could be picked for further growth and analysis of this immunoglobulin genomic clone.

The work reported in the following chapters of my thesis involved characterizing many such immunoglobulin genomic clones. The techniques used in these analyses are described in more detail in each chapter. Basically, they included restriction enzyme digestion and gel electrophoresis of cloned DNAs, hybridizing nitrocellulose filter replicas of gels to labeled probes ("Southern blots"), electron microscopy (by D. Kaback and K. Calame) and DNA sequencing. The most detailed analysis required nucleotide sequencing, which I usually performed by the Maxam-Gilbert technique[14]. Figure 8 shows a sequencing ladder including the first identified heavy chain J gene segment, J_{H107}, described in Chapter 4 of my thesis. I also used the dideoxynucleotide method[15] to determine the partial sequences of certain immunoglobulin mRNAs described in Chapters 2 and 4 (Figure 9).

1. 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, 817-836.

2. Marcu, K. B., Valbuena, O., and Perry, R. P. (1978). Biochemistry 17, 1723-1733.

3. Bailey, J. M. and Davidson, N. (1976). Anal. Biochemistry 70, 75-85.

4. Pelham, H. R. B. and Jackson, R. J. (1976). Eur. J. Biochem. 67, 247-256.

5. Milstein, C., Brownlee, G. G., Harrison, T. M., and Mathews, M. B. (1972). Nature New Biology 239, 117-120.

6. Wickens, M. P., Buell, G. N., and Schimke, R. T. (1978). J. Biol. Chem. 253, 2483-2495.

7. Wensink, P. C., Finnegan, D. J., Donelson, J. E., and Hogness, D. S. (1974). Cell 3, 315-325.

8. Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S. P., Chick, W. L., and Gilbert, W. (1978). Proc. Nat. Acad. Sci. USA 75, 3727-3731.

9. Grunstein, M. and Hogness, D. S. (1975). Proc. Nat. Acad. Sci. USA 72, 3961-3965.

10. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D., and Goodman, H. M. (1977). Nature 270, 486-494.

11. Joho, R., Weissman, I. L., Early, P., Cole, J., and Hood, L. (1980). Proc. Nat. Acad. Sci. USA 77, 1106-1110.

12. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L., and Smithies, O. (1977). Science 196, 161-169.

13. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978). Cell 15, 687-701.

14. Maxam, A. M. and Gilbert, W. (1977). Proc. Nat. Acad. Sci. USA 74, 560-564.

15. Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Proc. Nat. Acad. Sci. USA 74, 5463-5467.

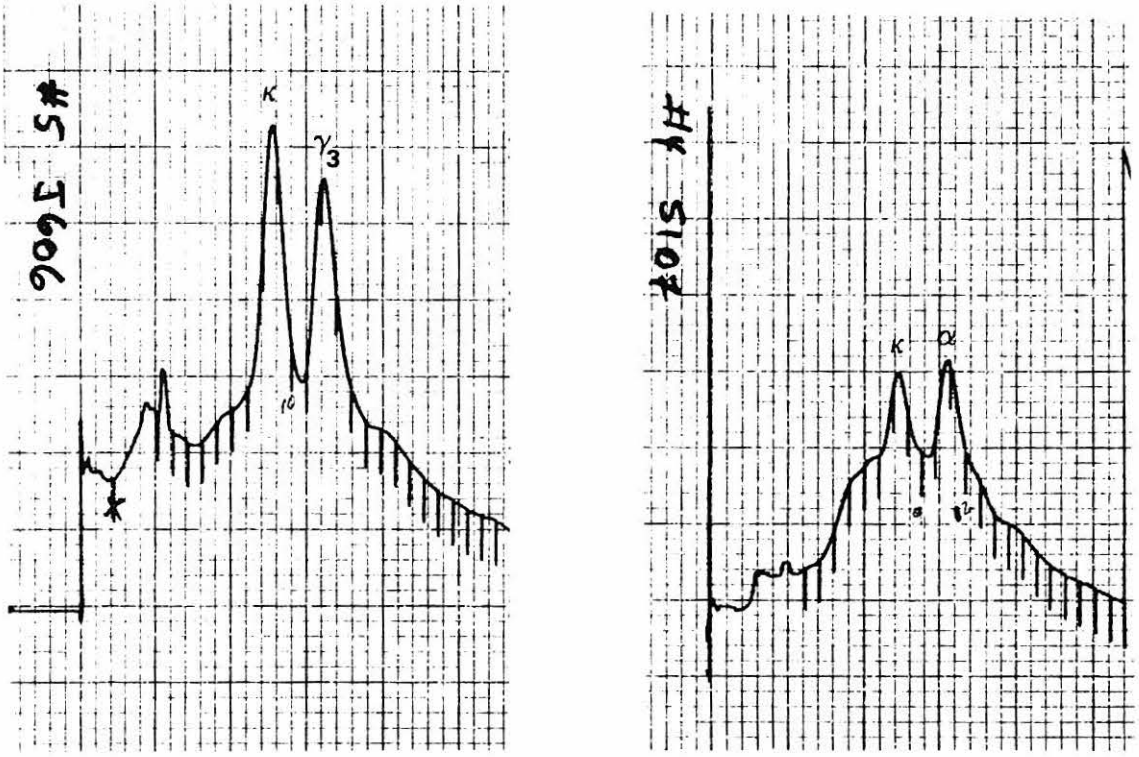


Figure 1. Sucrose gradient fractionation of poly-A+ RNAs from J606 (κ, γ_3) and S107 (κ, α) myeloma tumors. Absorbance profiles at 260 nm are shown. Isokinetic sucrose gradients [H. Noll (1967). Nature 215, 360-363] were formed in SW41 tubes: V_{mix} = 8.4 ml/tube, mixing chamber 10.4% (w/v) sucrose, reservoir 37.8% (w/v) sucrose. Buffer was 100 mM NaCl, 10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.2% Sarkosyl. Centrifugation was 37K for 12 h at 6°.

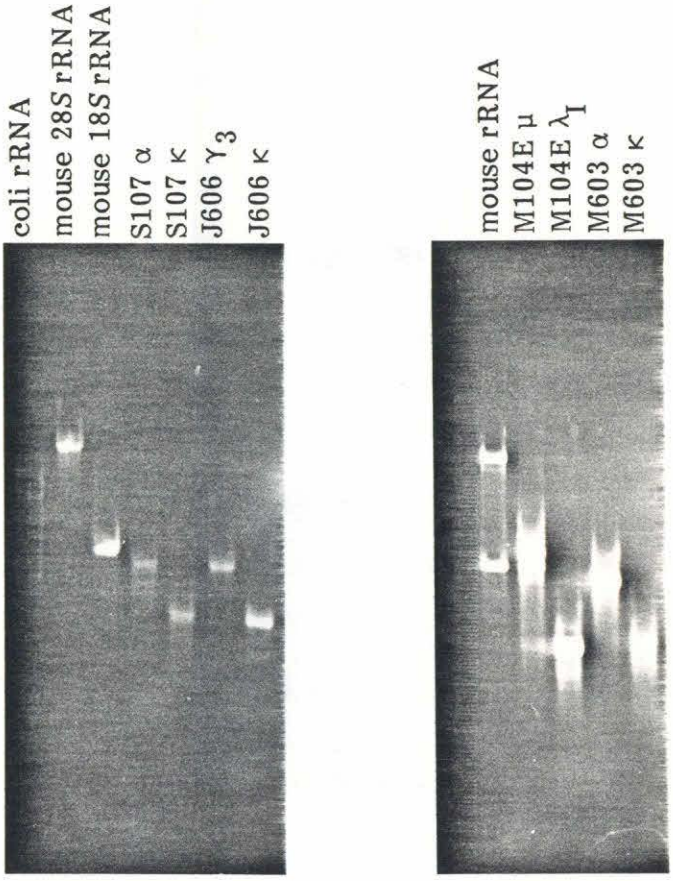


Figure 2. Methylmercury-agarose gels of immunoglobulin mRNAs, visualized by EtBr fluorescence. The 1% agarose gel includes peak fractions of S107 (κ, α) and J606 (κ, γ_3) mRNAs from the sucrose gradient in Figure 1. The 1.4% agarose gel shows M104E (λ_I, μ_s, μ_m) and M603 (κ, α) mRNAs.

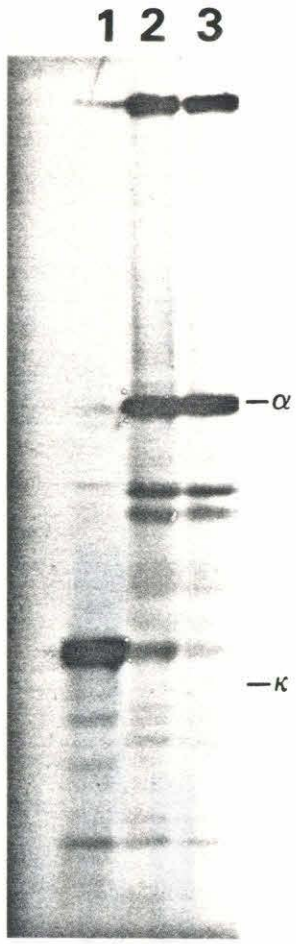


Figure 3. In vitro translation of immunoglobulin mRNAs. Sucrose gradient purified κ and α mRNAs from W3207 myeloma tumors were translated in nuclease-treated rabbit reticulocyte lysate. Reactions contained approximately 0.2 μg RNA, 40 μCi ^{3}H-leucine, and 50 μl treated lysate. Reactions were immunoprecipitated with anti-κ or anti-α rabbit antisera and staph A, and electrophoresed on a 12% Laemmli gel. Autofluorograph was performed overnight. Lane 1 shows κ mRNA translation, precipitated with anti-κ. Lane 2 is α mRNA, anti-κ (showing κ mRNA contamination in the α mRNA preparation), and lane 3 is α mRNA, anti-α. In other experiments, the α translation product was found to bind directly to staph A. Positions of marker κ and α chains are indicated.

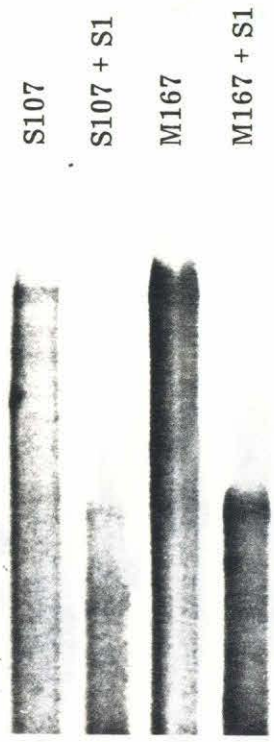


Figure 4. Autoradiograph of denaturing (alkaline) 1.5% agarose gel of double stranded cDNAs. ^{32}P labeled S107α and M167α double stranded cDNAs are shown before and after S1 digestion to remove the 5' terminal "hairpin". Both preparations contain contaminating κ cDNA.

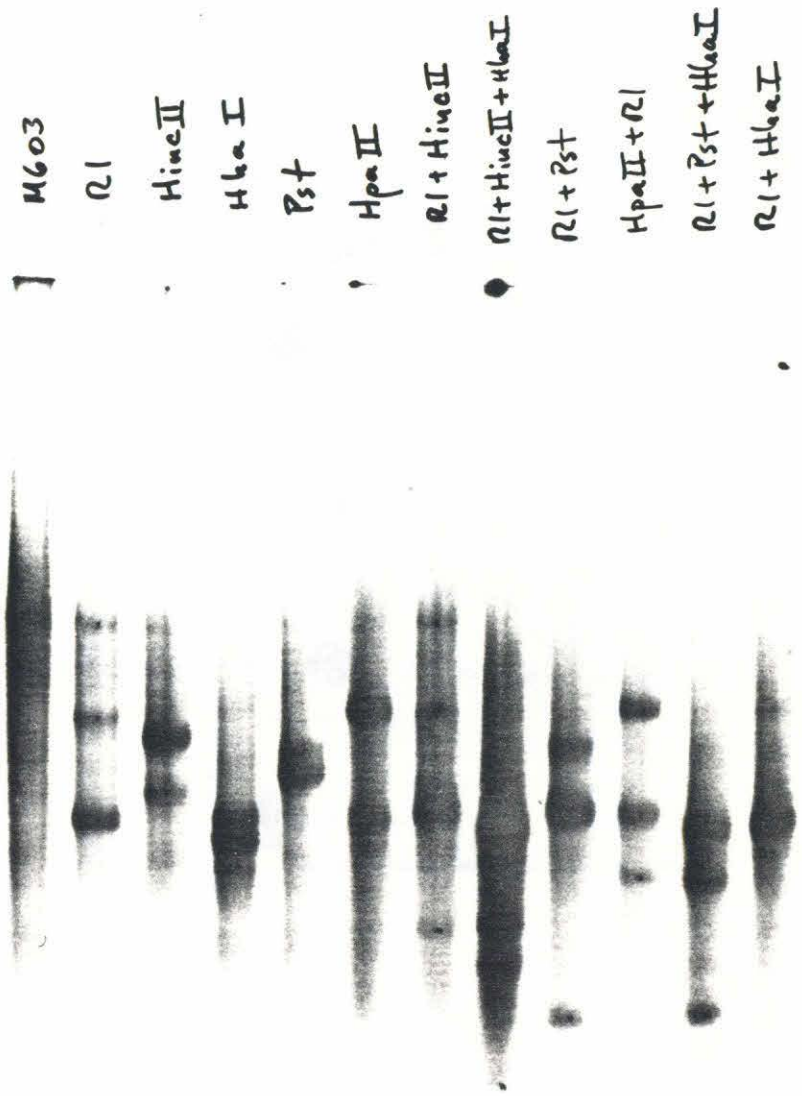


Figure 5. Autoradiograph of restriction digests of M603α double stranded cDNA. ^{32}P labeled samples were electrophoresed on a 1.4% nondenaturing agarose gel. Undigested M603α double stranded cDNA is shown in the left lane. Samples were also electrophoresed on denaturing (alkaline) agarose gels: the altered relative mobility of fragments containing a "hairpin" (see Figure 4) identified the 5' end.

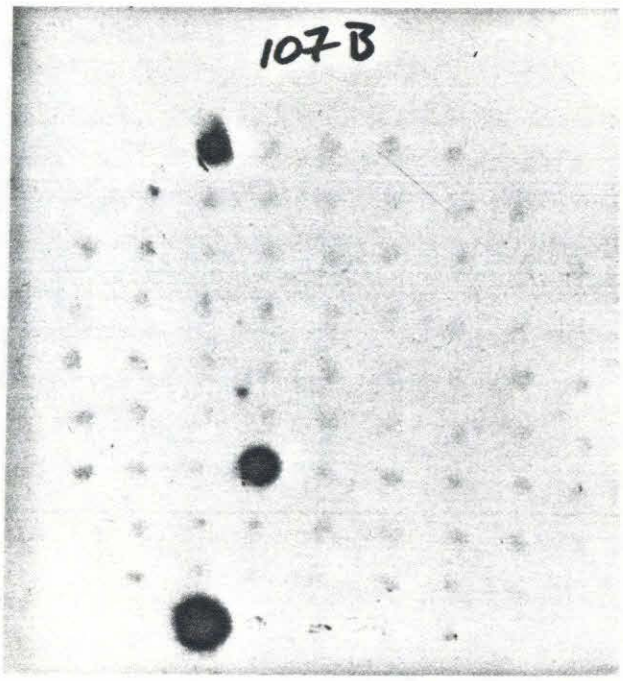


Figure 6. Detection of transformed χ1776 colonies containing S107α cDNA sequences. An autoradiograph of a nitrocellulose filter on which bacterial colonies were grown is shown after hybridization to ^{32}P kinase-labeled S107α mRNA. In this experiment, bacteria were transformed with pMB9 ligated to S107α double stranded cDNA by Eco RI linkers. Colonies which grew on tetracycline were picked, and duplicates grown on the nitrocellulose filter used for hybridization.

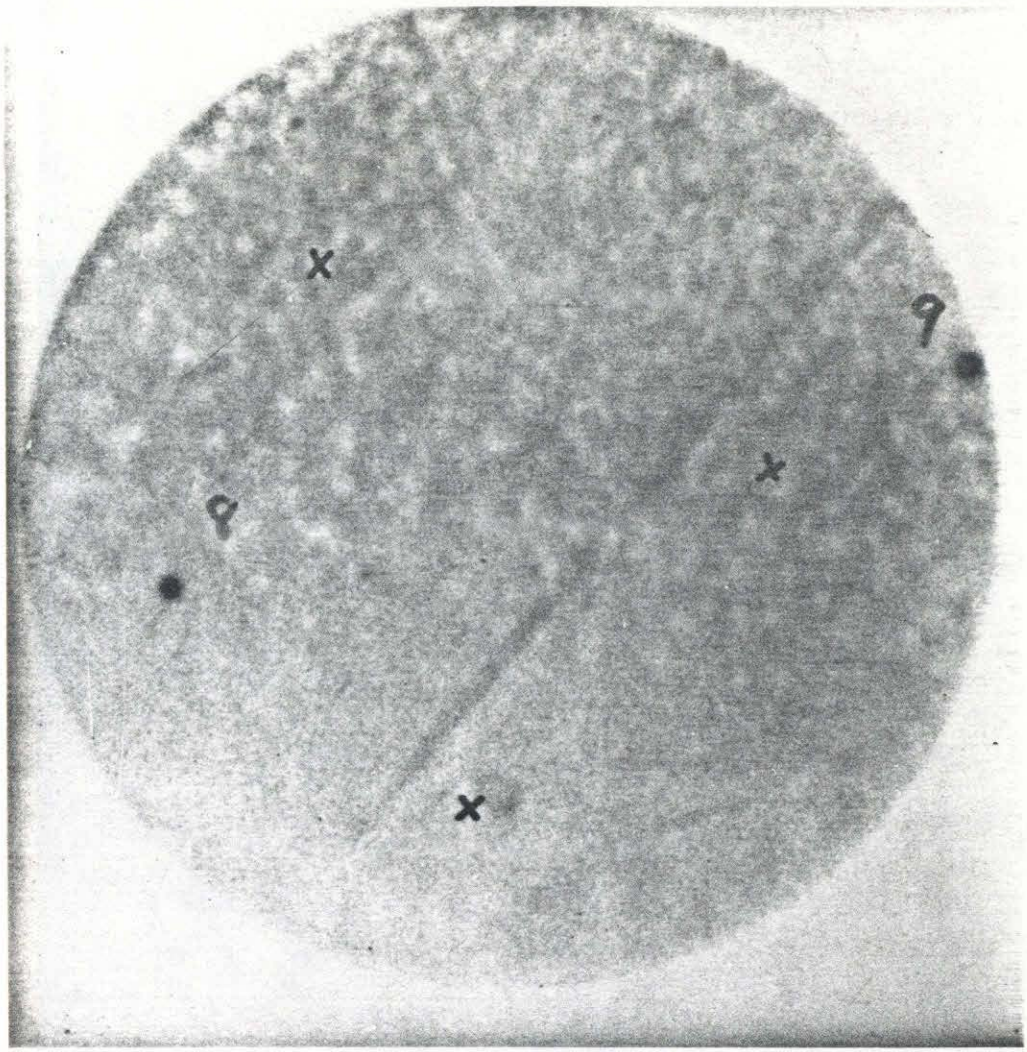


Figure 7a. First screening of phage library. An autoradiograph of a nitrocellulose replica filter is shown after hybridization to an immunoglobulin cDNA plasmid labeled with ^{32}P by nick translation. The filter contained phage DNA transferred from a petri plate covered with about 20,000 plaques from a recombinant Charon 4A phage library of mouse DNA. The two dark spots numbered 8 and 9 are plaques containing mouse immunoglobulin genes homologous to the labeled probe. Phage from these areas of the original plate were picked for further screening (Figure 7b). Crosses are alignment markings.

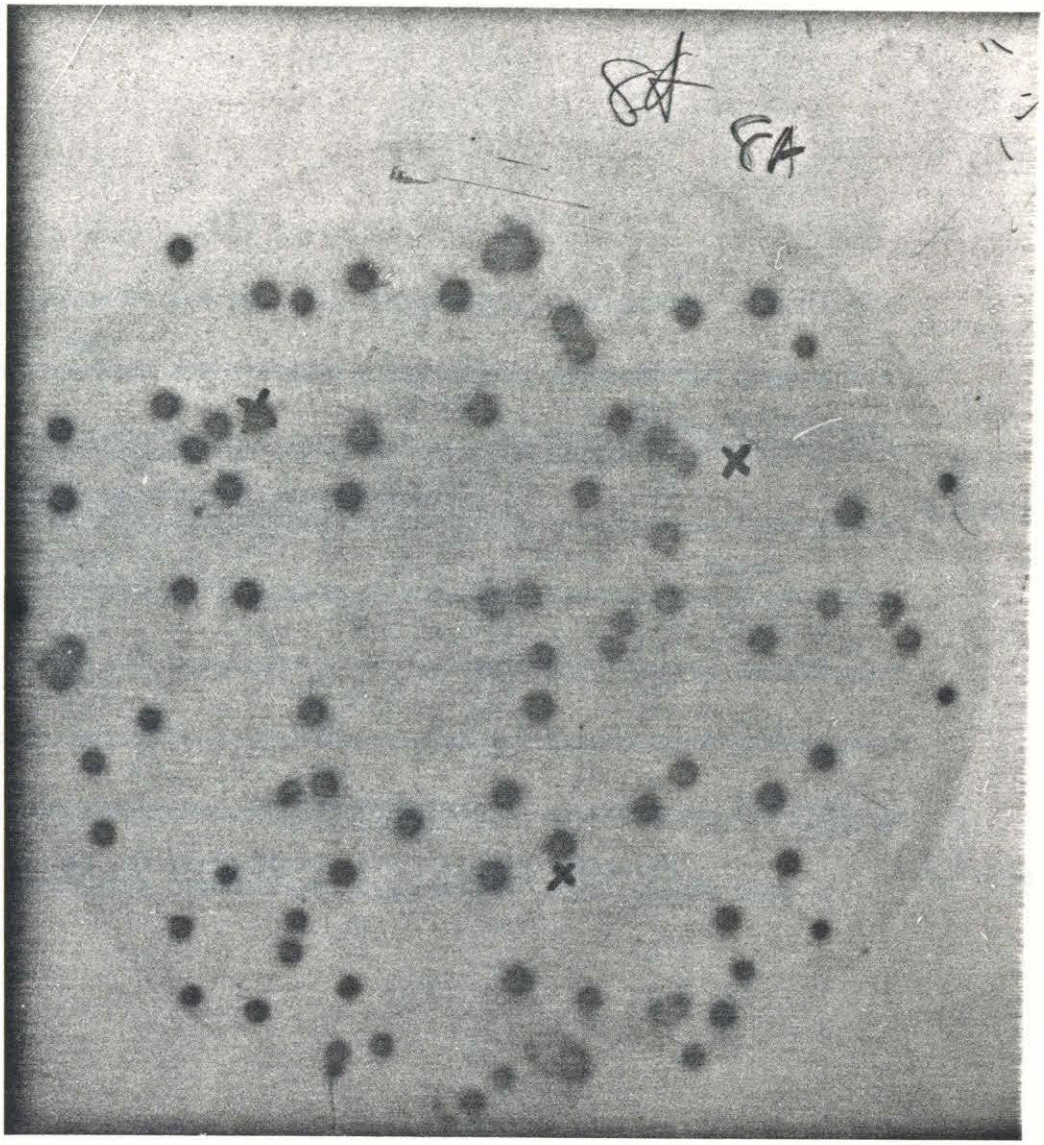


Figure 7b. Second screening of a phage library clone. About 1000 plaques from phage picked from the area of a positive plaque (8) on the first plate (Figure 7a) were grown on a second plate. An autoradiograph of a nitrocellulose replica filter from the second plate is shown after hybridization to the same labeled immunoglobulin cDNA clone used in the first cycle of screening. Phage from some of these positive plaques were picked for liquid culture growth and characterization.

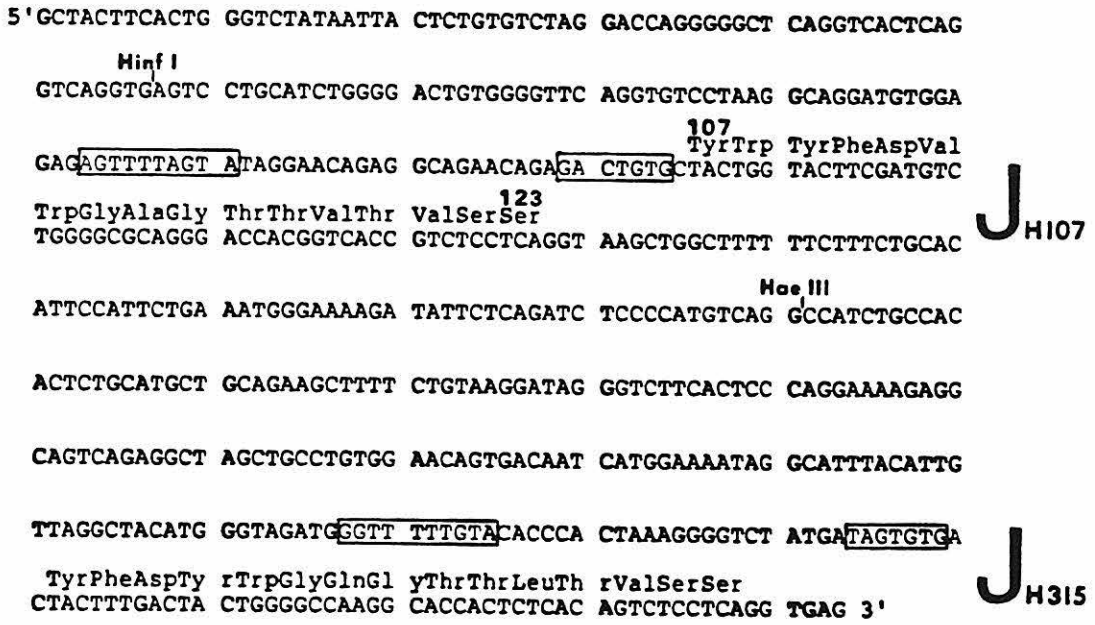
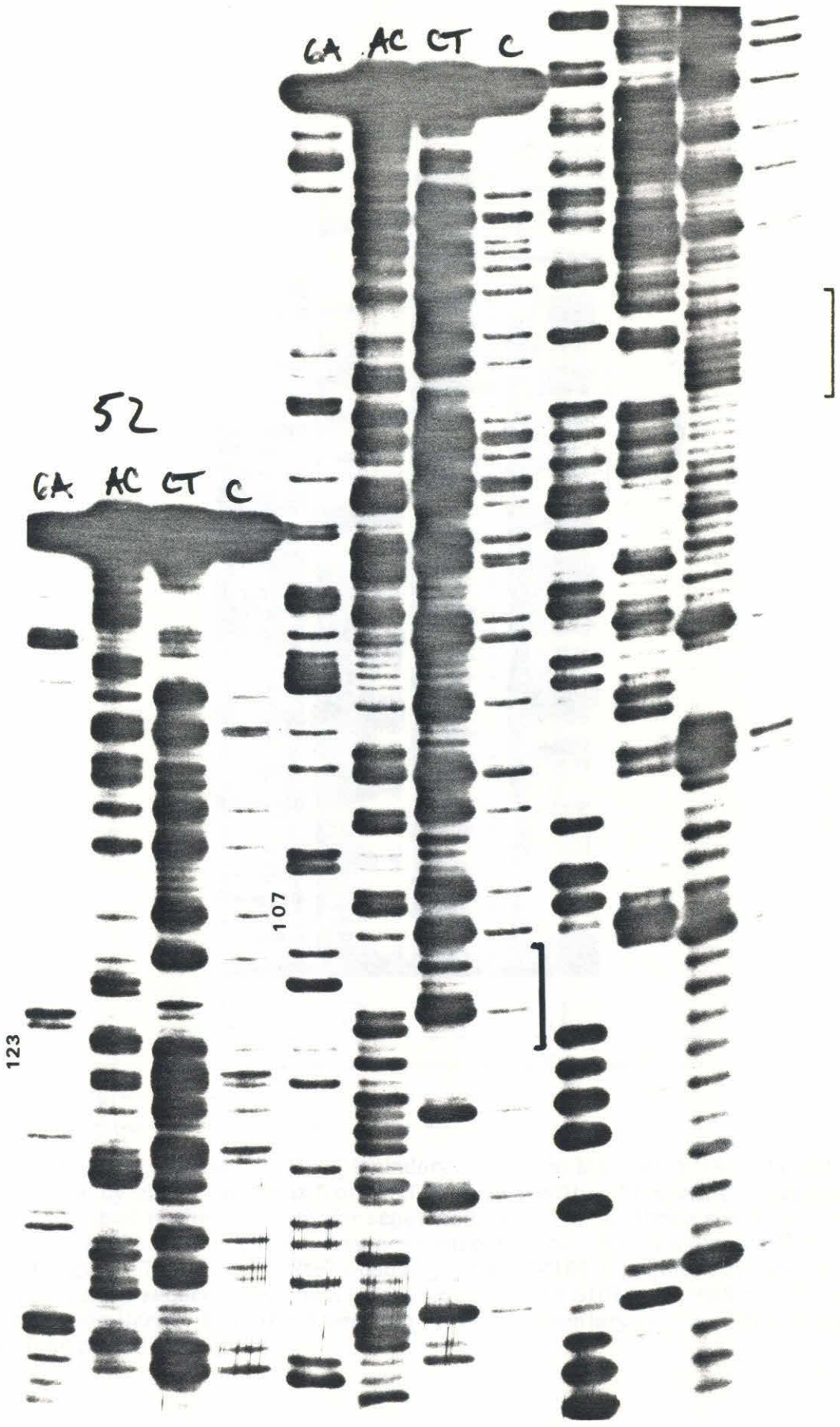


Figure 8. Sequences of two J_H germline gene segments. The following page shows a Maxam-Gilbert sequencing ladder including the J_{H107} germline gene segment. The DNA fragment was labeled at the Hinf I site shown in this figure. Sequencing reactions were G>A, A>C, C+T, and T specific. Electrophoresis was on an 8% gel, with three sample loadings. Codons at the 5' and 3' ends of the J_{H107} gene segment are numbered and the recognition sequences for DNA joining are indicated by vertical lines.


52
GA AC CT C
GA AC CT C
123
107

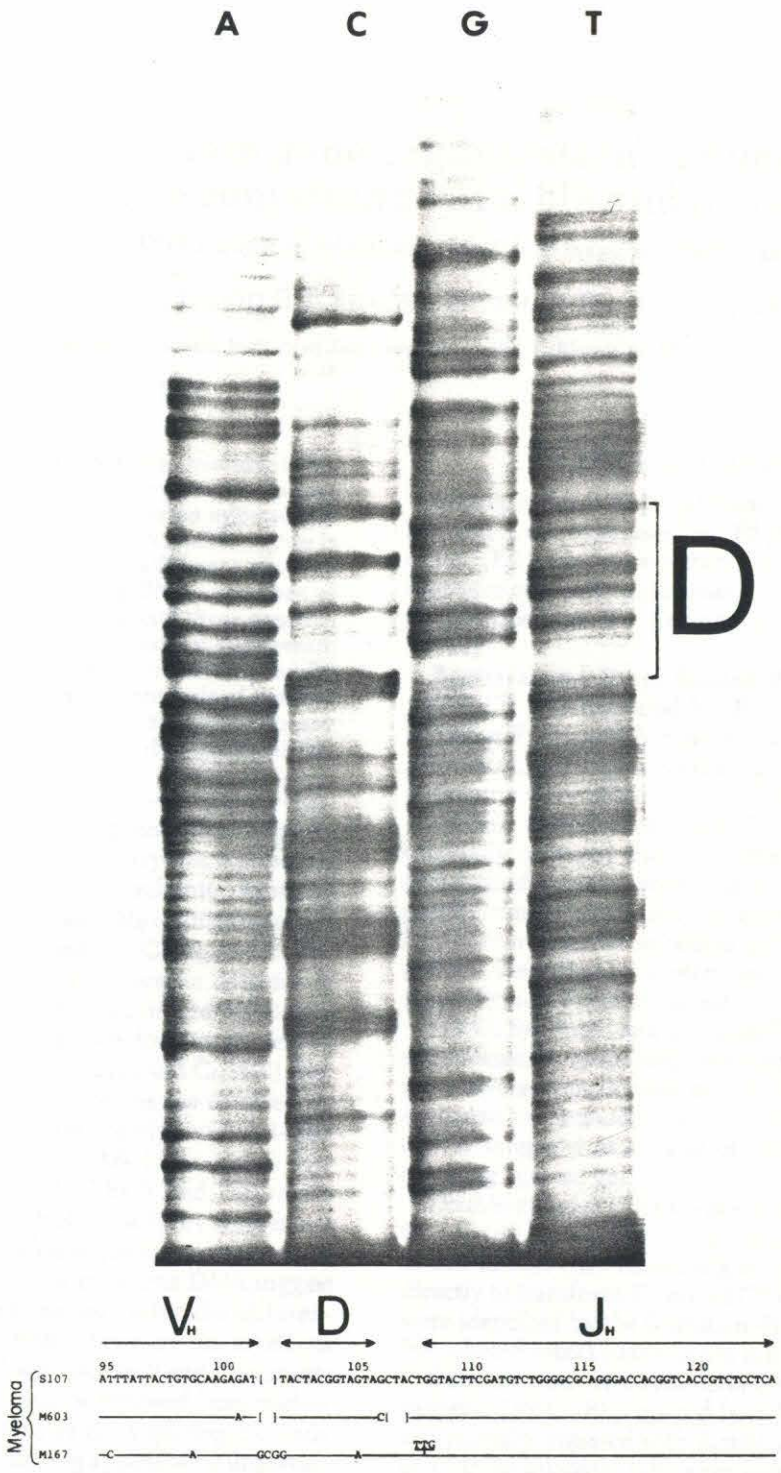


Figure 9. Sequence of the V_H-D-J_H boundary regions in M167α mRNA. The gel ladder was produced by cDNA synthesis from M167α mRNA with a different dideoxynucleotide triphosphate used in the reactions for separate lanes. The synthesis reactions were primed with a 75 nucleotide DNA fragment encoding the first 25 amino acids of the α constant region. The approximate boundaries of the M167 D segment are indicated. The M167 mRNA sequence is shown below, compared to S107 and M603 mRNA sequences determined similarly. The mRNA sequence is complementary to the cDNA sequence determined by the gel ladder.

Reprinted from
Proc. Natl. Acad. Sci. USA
Vol. 76, No. 2, pp. 857-861, February 1979
Genetics

Immunoglobulin heavy chain gene organization in mice: Analysis of a myeloma genomic clone containing variable and α constant regions

(heavy chain genomic clone/intervening DNA sequences/restriction maps/R loop mapping/V and C gene segments)

PHILIP W. EARLY*, MARK M. DAVIS*, DAVID B. KABACK†, NORMAN DAVIDSON†, AND LEROY HOOD*‡

*Division of Biology and †Department of Chemistry, California Institute of Technology, Pasadena, California 91125

Contributed by Norman Davidson, November 7, 1978

ABSTRACT We have isolated a myeloma genomic DNA clone containing the variable and constant regions of a mouse α chain. Restriction enzyme analyses and electron microscopic R loop mapping have demonstrated that the variable region is separated from the constant region by 6.8 kilobases of intervening DNA. In addition, two intervening DNA sequences of 100–200 bases separate the constant region into three approximately equal units. These intervening sequences may separate each of the segments coding for the three constant region domains of the α heavy chain. Southern blot analysis of embryo and myeloma DNA suggests that DNA rearrangement of heavy chain variable and constant regions occurs during the differentiation of antibody-producing cells.

The antibody system affords a fascinating model for the study of gene organization and expression in eukaryotes. Antibodies or immunoglobulins are composed of two subunits, light and heavy chains, each of which contains an NH_2-terminal variable (V) region and a COOH-terminal constant (C) region (1). The antibody polypeptides are divided into discrete domains or homology units, each encompassing approximately 110 residues. Accordingly, the light chain has two domains (V and C) and the α heavy chain has four (V, C_H1, C_H2, and C_H3) (2). The variable and constant regions of light chains are encoded by three distinct gene segments, V (approximately residues 1–99), J or joining (approximately residues 100–112), and C (approximately residues 113–219) (3, 4). The V and J segments encode the classical V region. Each of these DNA segments is separated by intervening nucleotide sequences in embryo or undifferentiated DNA (4). Studies of myeloma DNA suggest that these gene segments are rearranged during the differentiation of antibody-producing cells. However, in myeloma DNA, intervening sequences still separate the V and C segments (4). These intervening sequences must be removed from nuclear RNA transcripts of light chain genes by RNA splicing (5). Thus, DNA rearrangements and RNA splicing appear to be important events in the differentiation of antibody-producing cells (6).

We are interested in analyzing the genes coding for heavy chains in embryo and myeloma DNA to determine whether sequence rearrangements occur during differentiation that are comparable to those seen for light chain gene segments (4). Our initial approach has been to isolate genomic clones from a library of recombinant Charon 4A bacteriophage containing long fragments of M603 myeloma DNA. In this paper we report the characterization of one clone containing both V_H and C_α regions.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Biological and Physical Containment. Work described in this report was conducted in a P3 physical containment facility using EK2 host-vector systems, in compliance with National Institutes of Health guidelines for recombinant DNA research as published in the *Federal Register* [(1976) 41, 27902–27943].

Bacterial and Phage Strains. *Escherichia coli* K-12 strain χ1776 (7) was provided by R. Curtiss. *E. coli* K-12 strain DP50SupF (8) and Charon 4A phage (9) were provided by F. Blattner. *E. coli* strains NS428 and NS433 (10) used for *in vitro* packaging were obtained from T. Maniatis.

mRNA Preparation. BALB/c mouse myeloma tumors originally obtained from M. Potter or the Salk Institute were propagated subcutaneously. A postnuclear supernatant prepared from homogenized tissue was used to prepare poly(A)$^+$ RNA from membrane-bound polysomes (11). Heavy chain mRNA, identified by *in vitro* translation (12), was isolated by sucrose gradient fractionation.

cDNA Synthesis and Cloning. Double-stranded cDNA was synthesized by sequential reactions with avian myeloblastosis virus reverse transcriptase and *E. coli* polymerase I (13). After exclusion on Sephadex G-100, the major component of this cDNA migrated as a band of 1550 base pairs on a nondenaturing agarose gel.

Double-stranded cDNA was joined to *Eco*RI-cut pMB9 either by poly(dA), poly(dT) tailing (14), or by ligation to synthetic *Eco*RI linkers (15). Annealing or ligation mixtures were used directly to transform *E. coli* χ1776 (16). Positive transformants were identified by the Grunstein–Hogness technique (17) using ^{32}P-labeled M603 heavy chain mRNA.

Construction of M603 Library. High molecular weight genomic DNA (18) prepared from M603 subcutaneous tumors was partially digested with *Eco*RI, and fragments in the range of 12 to 20 kilobases (kb) were isolated on a sucrose gradient. Ten micrograms of M603 DNA fragments was ligated to Charon 4A arms and packaged *in vitro* to obtain a library of 3×10^6 recombinant phage (19). The library was amplified on DP50SupF as a plate lysate prior to screening.

Isolation of Clones from M603 Library. The constant region plasmid p603α1 labeled by nick translation with deoxynucleotide [^{32}P]triphosphates (20) was used to screen 400,000 clones from the M603 library plated on DP50SupF (19, 21). Duplicate nitrocellulose filters from each plate were prehybridized in 1 M NaCl/0.045 M trisodium citrate/0.2% bovine serum albumin/0.2% Ficoll/0.2% polyvinylpyrrolidone/0.1% sodium

Abbreviations: V, variable; C, constant; $NaDodSO_4$, sodium dodecyl sulfate; kb, kilobase(s); IVS, intervening sequence(s).

‡ To whom reprint requests should be addressed.

dodecyl sulfate ($NaDodSO_4$) at 68°C in a rotary water bath (D. Engel and J. Dodgson, personal communication). Denatured plasmid DNA was added to 10 ng/ml and hybridization was continued for 48 hr. Filters were washed extensively in 0.15 M NaCl/0.015 M trisodium citrate/0.1% $NaDodSO_4$/10 mM $Na_4P_2O_7$ at 68°C. Duplicated spots of hybridization were identified by autoradiography, and plaques corresponding to these locations on the filters were picked and rescreened at a low plating density to obtain pure clones.

Electron Microscopic R Loop Mapping. Duplex Ch603α6 DNA was photochemically crosslinked (one crosslink per 4 kb) in the presence of 4,5',8-trimethylpsoralen (trioxsalen). Crosslinking prevents DNA strand separation, thus permitting the R loop hybridization (22) to be carried out at a temperature close to or above the DNA strand separation temperature (unpublished data). Crosslinked DNA (5 μg/ml) was hybridized to mRNA (2 μg/ml) in 70% recrystallized formamide/0.4 M NaCl/0.1 M 1,4-piperazinediethanesulfonic acid (Pipes), pH 7.2/10 mM EDTA at 56°C for 24 hr. The R-looped DNA was either spread from 70% onto 15% formamide or treated with 1 M glyoxal at 12°C to stabilize the R loops against branch migration (unpublished observations) and then spread from 50% formamide (23).

RESULTS AND DISCUSSION

Sequence Homologies of α Heavy Chain mRNAs. We isolated heavy chain mRNA from three mouse myeloma tumors secreting IgA immunoglobulins. Both the M603 and S107 immunoglobulins bind phosphorylcholine and contain nearly identical V region protein sequences (24). The M315 V region differs from M603 at 60% of its amino acid residues. To assay the extent of nucleotide homology between these mRNA species, ^{32}P-labeled single stranded M603 cDNA was hybridized to each mRNA. The hybrid was digested with nuclease S1 (25), and the resulting cDNA cleavage products were fractionated on an alkaline agarose gel (Fig. 1). M315 mRNA protects an 1100-nucleotide piece of M603 cDNA from S1 digestion; S107 and M603 mRNA both protect the full length of 1550 nucleotides. We conclude that S107 and M603 mRNAs are closely homologous over the entire sequence and that M315 mRNA shares this homology only for the C region, as expected from the protein sequences. In subsequent experiments with M603-like genomic sequences, we used S107 mRNA as a probe for the V and C regions and M315 mRNA as a probe for the C region only.

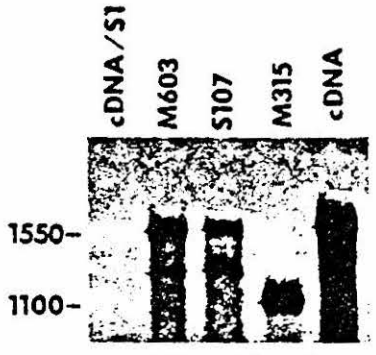


FIG. 1. Hybridization of mRNAs to M603 cDNA. Five nanograms of first-strand M603 cDNA (specific activity, 10^6 cpm/μg) was hybridized to 50 ng of the indicated heavy chain mRNA followed by digestion with nuclease S1 (25) and separation on a 2% alkaline agarose gel (26), a portion of which is shown by autoradiography. The first lane contains cDNA alone, hybridized and digested as above, and the last lane contains undigested input cDNA. The cDNA is somewhat longer than 1550 nucleotides, presumably due to an unfolded "hairpin" structure at the 5' end (27).

Characterization of cDNA Clones. Double-stranded cDNA prepared from M603 mRNA was used to construct the cDNA restriction map in Fig. 2*a*. We used poly(dA), poly(dT) tailing to obtain one recombinant plasmid, p603α1, which was shown by electron microscopy to contain an insert of approximately 600 nucleotides. Comparison of the cDNA restriction map with that of the plasmid indicates that p603α1 contains part of the C_α sequence (Fig. 2*b*). Subsequently, synthetic *Eco*RI linkers were used to clone S107 double-stranded cDNA. Regions of the cDNA included in two of these plasmids, p107αR5 and p107αR6, are indicated in Fig. 2*b*. In order to verify that the cloned sequences are derived from S107 heavy chain mRNA, a *Hin*fI restriction fragment of p107αR6 was isolated, annealed to S107 mRNA, and used to prime cDNA synthesis in the dideoxynucleotide sequencing procedure (29). The partial sequence so determined matched the known protein sequence of the S107 heavy chain between amino acids 92 and 125 (24) (data not shown).

Characterization of Genomic Clones. We screened 400,000 plaques from the M603 library with the C region plasmid p603α1, labeled with ^{32}P by nick translation. Twenty-five clones hybridized to the probe. Only three of these clones also hybridized to a nick-translated V region probe (the portion of p107αR6 shown to the left of the *Hha* I site in Fig. 2*b*). These three clones showed identical *Eco*RI restriction patterns; one, Ch603α6, was selected for further characterization. Ch603α6 contained 16.4 kb of mouse DNA with two internal *Eco*RI cleavage sites yielding fragments of 7.2, 4.8, and 4.4 kb, which are ordered in the restriction map shown in Fig. 2*c*. Filter hybridizations by the Southern blot procedure (30) localized the variable region in Ch603α6 to the 7.2-kb *Eco*RI fragment (Fig. 3). C region probes (p603α1 and the portion of p107αR6 to the right of the *Hha* I site in Fig. 2*b*) hybridized only to a 2.4-kb section bounded by *Xho* I and *Sma* I sites in the 4.8- and 4.4-kb *Eco*RI fragments. These results, displayed schematically in Fig. 2*c*, demonstrate that the V_H and C_α regions in Ch603α6 are separated by at least 4 kb of intervening DNA.

Identification of Genomic DNA Fragments Hybridizing to cDNA Clones. Hybridization of the nearly full-length cDNA probe p107αR5 to Southern blots of *Eco*RI-digested chromosomal DNA showed that the 7.2-, 4.8-, and 4.4-kb fragments of Ch603α6 all correspond to bands present in the M603 genome (Fig. 4). In embryonic DNA, the 4.8-kb middle piece of Ch603α6 was absent, although bands at 7.2 and 4.4 kb were present. These observations are consistent with a reduction in the distance between the Ch603α6 V and C regions having occurred in the derivation of the M603 genome from the embryo genome. Genetic evidence suggests that a closely related V_H region, S107, may be several hundred thousand base pairs from the C_α region in embryonic DNA (31, 32; but see ref. 33). The other changes evident from the Southern blots of embryonic and M603 DNAs cannot yet be fully explained.

Electron Microscopy Indicates That There Are Three Intervening Sequences in the Genomic Clone Ch603α6. S107 or M315 mRNA was hybridized to trioxsalen-crosslinked Ch603α6 DNA under conditions favoring R-loop formation. At least 80% of the DNA molecules examined in the electron microscope were full length and >90% of these contained R loops. Typical micrographs are shown and interpreted in Fig. 5. Data from 52 molecules were combined to generate the map shown in Fig. 2*d*. These studies indicate that Ch603α6 contains three intervening sequences (IVS). IVS 1 has a length of 6.8 kb. The R loop to the left of IVS 1 in Fig. 2*d* is due to the V_H region because it is seen in hybridizations of Ch603α6 DNA with S107 mRNA but not with M315 mRNA. The C region structures discussed below are seen in hybridizations with either S107 or

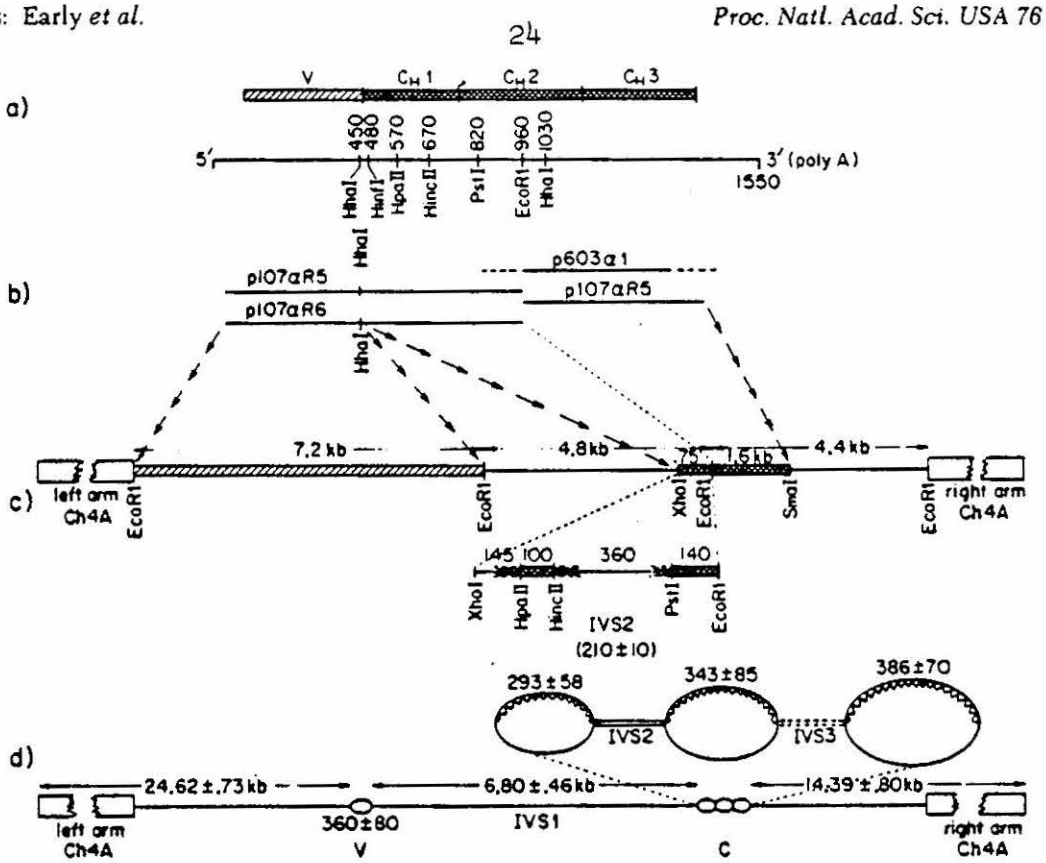


FIG. 2. Comparison of genomic and cDNA segments. (*a*) Restriction map of M603 double-stranded cDNA. The presence of a "hairpin" structure (determined by alkaline agarose gels) served as a marker for the 5' end (27). Only the *Hin*fI site identified by mRNA sequencing is shown. The structural domains depicted above the cDNA are based on a partial protein sequence of M315 and M511 heavy chains and analogy with other IgA proteins (28). The protein sequence has been aligned so that the *Hha* I site sequenced in S107 mRNA coincides with the *Hha* I site mapped in M603 cDNA. This alignment gives good correspondence between the *Hpa* II and *Hin*cII sites mapped in M603 cDNA and possible cleavage sites of these enzymes at codons for amino acids 158 and 191. (*b*) Sequences included in cDNA plasmid clones. The ends of p603α1 are uncertain, as indicated by dotted lines. The larger of the two *Eco*RI fragments in p107αR5 is present in inverted orientation relative to its position in the cDNA. This is presumably due to the independent ligation of these two fragments during the cDNA cloning procedure. The restriction fragments of Ch603α6 within which the various parts of the plasmid clones hybridize are indicated by arrows connecting the plasmid and genomic clones. The dotted line connects the internal *Eco*RI site of the cDNA plasmids with the corresponding site in Ch603α6. (*c*) Restriction map and sequence organization of Ch603α6. Restriction fragments to which V and C region probes (see arrows from plasmids) hybridize are shaded in correspondence to the protein domains. The enlarged portion of the figure shows one of the short intervening sequences (IVS 2). Compare distances between restriction sites here and those shown in *a*. The *Xho* I site does not lie within the coding sequence. (*d*) R-Loop map of Ch603α6. The right and left arms of Ch4A were taken to be 10.7 and 20.1 kb long, respectively. The enlargement of the C region R loops is drawn with the total length of the RNA·DNA duplex equal to the length of the cDNA from the 3' end to the junction with the V region shown in *a*. IVS 2 and IVS 3 are both assumed to be the same length. The *Eco*RI site shown in the enlargement of the Ch603α6 restriction map above is aligned with a point 590 nucleotides (measured on the RNA·DNA duplex regions) from the 3' end of the C region R-loop complex. This corresponds to the distance of the *Eco*RI site from the 3' end of the cDNA. The wavy line in the R loops represents RNA. The dashed line for IVS 3 indicates that its length has not been accurately measured. Errors given are SD.

M315 mRNAs. These results and the measured lengths of the several R loops indicate that IVS 1 occurs approximately at the junction of the V_H and C_α regions (Fig. 2*d*). Two kinds of R-loop structures were seen for IVS 1, depending on whether a single mRNA molecule (Fig. 5 *a* and *b*) hybridized to both the C and V regions of the DNA or whether these regions were hybridized to two separate mRNA molecules (Fig. 5*c*).

In addition to the 6.8-kb IVS, 58% of the molecules contained two structures that we interpret as due to two short IVS within the C_α region (IVS 2 and IVS 3 in Fig. 2*d*). In one type of short IVS structure, the single- and double-stranded arms of a C region R loop are joined at a reproducible point inside the R loop. We interpret this as due to a base-paired IVS (bpIVS in Fig. 5). Alternatively, the two arms of an R loop are not joined, but there is a small knob at a reproducible point on the double-stranded arm. We interpret this as an IVS that is not base paired to its complement on the opposite strand (ssIVS in Fig. 5). The positions on the C region R loop where bpIVS and ssIVS structures are seen are coincident. Because IVS 2 and IVS 3 are observable as knobs but are too short to be measured by electron microscopy, we believe their lengths to be in the range of 100 to 200 nucleotides.

The reproducible junction points that we attribute to base-paired IVS could be due to site-specific trioxsalen crosslinking. However, the ssIVS structures cannot be explained this way. Furthermore, these structures occurred with approximately the same frequency in R loops with uncrosslinked Ch603α6 DNA.

Both electron microscopy and restriction mapping give the same orientation of the V_H and C_α regions relative to the right and left arms of Charon 4A. The position of the 3' poly(A) end of the mRNA also has been independently determined in some R loops by an electron microscopic labeling technique (34) (Fig. 5*c*).

IVS 2 Is 210 Nucleotides in Length by Restriction Mapping. By comparing the sizes of restriction fragments produced from DNA of the genomic clone Ch603α6 with those from the cDNA clone p107αR6, we have confirmed the presence and determined the length of one of the short IVS in the C_α region (Fig. 6). Gel electrophoresis showed that the *Hin*cII and *Pst* I

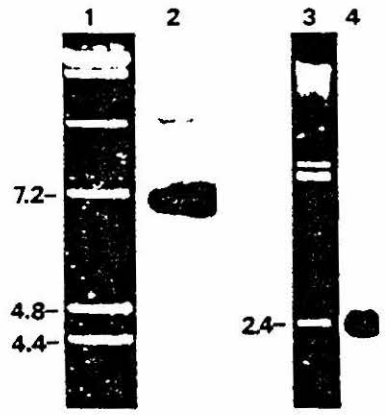


FIG. 3. Localization of the V and C regions in Ch603α6 by hybridization to Southern blots. Restriction enzyme digests of Ch603α6 DNA were electrophoresed on 1% agarose gels, transferred to nitrocellulose filters (30), and hybridized to V or C region probes prepared from gel-purified fragments of *Hha* I-digested p107αR6 (site of cleavage indicated in Fig. 2*b*). V region probe: ethidium bromide staining (lane 1) and blot (lane 2) of *Eco*RI-digested Ch603α6 DNA. Only the 7.2-kb fragment displayed strong hybridization to the V region. C region probe: ethidium bromide staining (lane 3) and blot (lane 4) of *Sma* I/*Xho* I-digested Ch603α6 DNA. Only the 2.4-kb fragment hybridized to this C region probe. The same result was obtained with probes from other parts of the C region. Unmarked bands seen by ethidium bromide fluorescence contained Ch4A vector DNA. The origins of the gels are not shown.

sites of the cDNA plasmid p107αR6 are separated by 150 nucleotides, whereas in the genomic DNA of Ch603α6 they are 360 nucleotides apart. This demonstrates the presence of an IVS of 210 nucleotides in the genomic clone which corresponds to the position assigned to IVS 2 by electron microscopy.

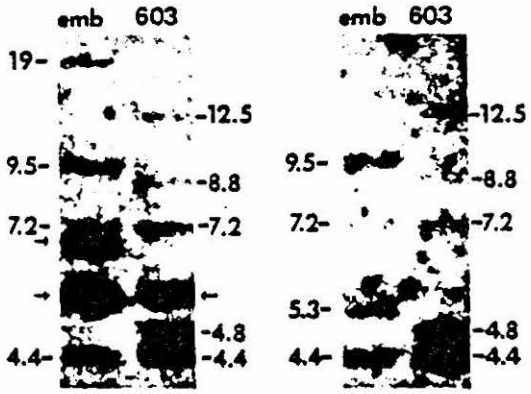


FIG. 4. Southern blots of *Eco*RI-cut genomic DNAs hybridized to p107αR5. High molecular weight genomic DNA was prepared from 12–13 day BALB/c embryos (emb) or M603 subcutaneous tumors (603) (18). Approximately 15 μg of DNA completely digested with *Eco*RI was electrophoresed on each lane of a 0.7% agarose gel. Nitrocellulose filter replicas of the gels were hybridized to ^{32}P-labeled p107αR5 (specific activity, 10^8 cpm/μg). Upper portions of the blots are not shown. In the two lanes at the left, different plasmid DNAs were added as internal standards (arrows). This obscures the 7.2- and 5.3-kb bands in the embryo DNA, which are better seen in the embryo lane at the right. The 19-kb band is not visible in the right embryo lane, probably because of poor transfer to the filter. Some clones from the M603 library contained a 4.4-kb *Eco*RI fragment hybridizing to p603α1, plus either a 12.5- or an 8.8-kb *Eco*RI fragment hybridizing to the 5′ C_α sequences in p107αR6. Thus, there may be multiple copies of the C_α gene segment, which would account for the relatively intense hybridization to the 4.4-kb band. All copies of the C_α gene in M603 DNA appeared to have undergone rearrangement or mutation from the embryo; the significance of this observation for the differentiation of antibody-producing cells is unknown.

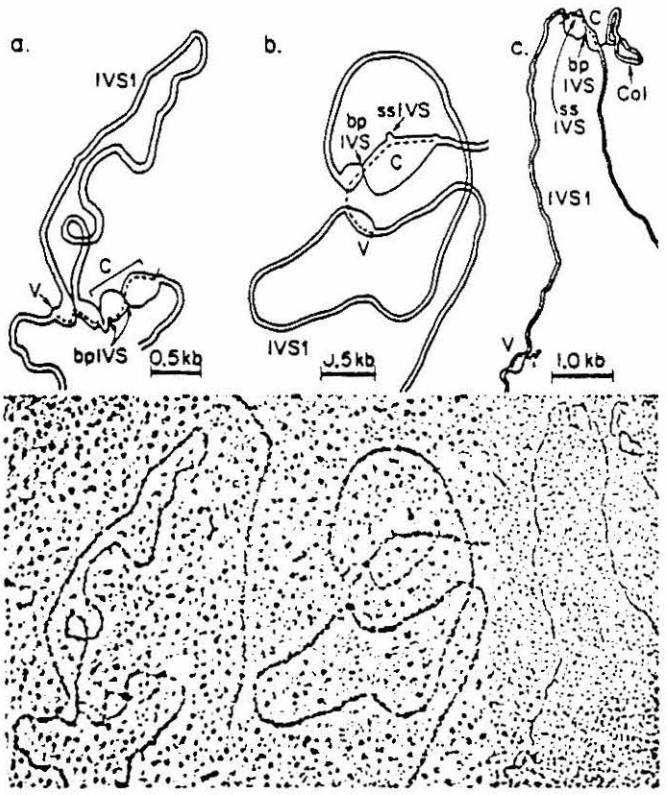


FIG. 5. Electron micrographs of R-loop structures observed with S107 mRNA hybridized to crosslinked Ch603α6 DNA. (*a*) R loop showing the large intervening sequence (IVS 1) dividing the constant (C) and variable (V) regions, and two short base-paired intervening sequences (bpIVS) in the constant region. (*b*) Similar to *a* but the C region contains a single-stranded intervening sequence (ssIVS) and a base-paired intervening sequence. (*c*) Two mRNA molecules hybridized to one Ch603α6 DNA molecule. The poly(A)$^+$ RNA at the 3′ end of the C region R loop is labeled with a poly(BrdU)-tailed circular microcolicin E1 molecule(Col). Both the V and C region R loops contain unhybridized RNA tails adjacent to IVS 1. The C region contains one bpIVS and one ssIVS. Broken lines represent RNA.

The Two Short C Region IVS May Separate the Three C_α Domains. The short IVS in the C_α gene separate the C region into three roughly equal segments. At the protein level the C_α region is divided into three roughly equal homologous structural domains: C_H1, C_H2, and C_H3 (28). Accordingly, the IVS in the DNA may separate the individual C_α domains, although this supposition will have to be verified by direct nucleic acid sequence analysis.

There are several features of immunoglobulin evolution and structure that might involve IVS separating C_H domains. First, all C regions, including C_α, contain homologous protein domains presumably derived from a common ancestral gene (2, 28, 37, 38). Second, C regions do not all contain the same number of domains, indicating that new C regions may arise by the addition or deletion of domains. Third, certain aberrant immunoglobulins (heavy chain disease proteins) often appear to involve deletions with breakpoints occurring between domains (39). Fourth, a variant myeloma cell line produced in culture shows deletion of the C_H1 domain (40). If IVS between domains facilitate unequal recombination, then the creation of new C regions with additional or deleted domains can be explained. Accordingly, the short IVS observed in the C_α region may permit the immunoglobulin domains to operate as fundamental units of evolution (41).

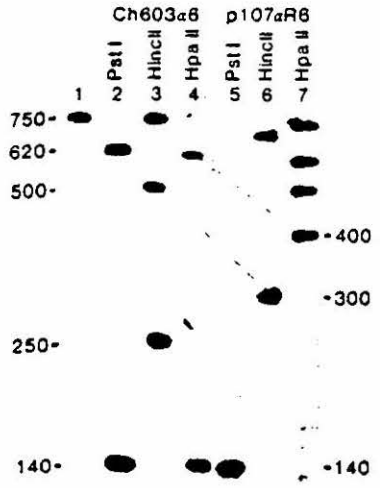


FIG. 6. Identification and measurement of IVS 2 by restriction mapping. Ch603α6 DNA digested with *Eco*RI and *Xho* I was ^{32}P end-labeled (35), and the 750-bp C region fragment (Fig. 2*c*) was eluted from a 1% agarose gel (36). An autoradiograph of part of a 5% acrylamide gel is shown comparing restriction digests of this end-labeled genomic DNA fragment and p107αR6 end-labeled after *Eco*RI digestion. Lanes: 1, Ch603α6 750-bp *Xho* I/*Eco*RI fragment; 2–4, products of digestion of this fragment with *Pst* I, *Hin*cII, and *Hpa* II, respectively (the *Hin*cII digest is incomplete); 5–7, *Pst* I, *Hin*cII, and *Hpa* II digests of p107αR6 end-labeled after *Eco*RI digestion. The *Eco*RI/*Pst* I bands of the two DNAs are the same length (140 bp), whereas the *Eco*RI/*Hin*cII and *Eco*RI/*Hpa* II bands are both 200–220 bp longer in the genomic DNA than in the cDNA (Fig. 2*c*). Comparing these results with the cDNA restriction map (Fig. 2*a*) shows that Ch603α6 contains a short intervening sequence (210 ± 10 bp) between the *Hin*cII and *Pst* I sites of the cDNA. This region includes the boundary between the C_H1 and C_H2 domains. (Other bands in lane 7 derive from *Hpa* II cleavage of the end-labeled plasmid DNA. The 790-bp fragment produced from *Hin*cII-digested p107αR6 is not shown on this part of the autoradiograph.)

We thank H. Manor for electron microscopy of p603α1, N. Johnson for providing M603 DNA, and Y.-H. Chien for electron microscopic characterization of mRNAs. *E. coli* polymerase I, DNA ligase T4, and *Eco*RI linkers were generous gifts from M. Goldberg, R. Scheller, and K. Itakura, respectively. Most of the *in vitro* packaging extracts used were the gift of T. Sargent. We thank K. Marcu, O. Valbuena, and R. Perry for advice on mRNA preparation and M. Wickens for communicating cDNA synthesis procedures prior to publication. We benefited from helpful discussions with T. Maniatis, T. Sargent, D. Goldberg, D. Engel, J. Dodgson, R. Joho, B. Klein, and D. Anderson. The work reported here was supported by National Institutes of Health Grants GM 10991 and GM 20927 and Biomedical Research Grant RRO7003A and National Science Foundation Grant PCM 71-00770. P.W.E. and M.M.D. are supported by National Institutes of Health Training Grant GM 07616; D.B.K. is a National Institutes of Health Postdoctoral Fellow.

1. Gally, J. (1973) in *The Antigens*, ed. Sela, M. (Academic, New York), Vol. 1, pp. 162–299.
2. Edelman, G. M., Cunningham, B. A., Gall, W., Gottlieb, P., Rutishauser, U. & Waxdal, M. (1969) *Proc. Natl. Acad. Sci. USA* 63, 78–85.
3. Weigert, M., Gatmaitan, L., Loh, E., Schilling, J. & Hood, L. (1978) *Nature (London)* 276, 785–790.
4. Brack, C., Hirawa, M., Lenhard-Schuller, R. & Tonegawa, S. (1978) *Cell* 15, 1–14.
5. Gilmore-Hebert, M. & Wall, R. (1978) *Proc. Natl. Acad. Sci. USA* 75, 342–345.
6. Hood, L., Huang, H. V. & Dreyer, W. J. (1977) *J. Supramol. Struct.* 7, 531–559.
7. Curtiss, R., III, Pereira, D. A., Hsu, J. C., Hull, S. C., Clarke, J. E., Maturin, L. J., Sr., Goldschmidt, R., Moody, R., Inoue, M. & Alexander, L. (1977) in *Proceedings of the 10th Miles International Symposium*, eds. Beers, R. F., Jr. & Bassett, E. G. (Raven, New York), pp. 45–56.
8. Leder, P., Tiemeier, D. & Enquist, L. (1977) *Science* 196, 175–177.
9. Blattner, F. R., Williams, B. G., Blechl, A. E., Denniston-Thompson, K., Faber, H. E., Furlong, L.-A., Grunwald, D. J., Kiefer, D. O., Moore, D. D., Schumm, J. W., Sheldon, E. L. & Smithies, O. (1977) *Science* 196, 161–169.
10. Sternberg, N., Tiemeier, D. & Enquist, L. (1977) *Gene* 1, 255–280.
11. Marcu, K. B., Valbuena, O. & Perry, R. P. (1978) *Biochemistry* 17, 1723–1733.
12. Pelham, H. R. B. & Jackson, R. J. (1976) *Eur. J. Biochem.* 67, 247–256.
13. Wickens, M. P., Buell, G. N. & Schimke, R. T. (1978) *J. Biol. Chem.* 253, 2483–2495.
14. Wensink, P. C., Finnegan, D. J., Donelson, J. E. & Hogness, D. S. (1974) *Cell* 3, 315–325.
15. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* 270, 486–494.
16. Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S. P., Chick, W. L. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* 75, 3727–3731.
17. Grunstein, M. & Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. USA* 72, 3961–3965.
18. Blin, N. & Stafford, D. W. (1976) *Nucleic Acids Res.* 3, 2303–2308.
19. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* 15, 687–701.
20. Maniatis, T., Jeffrey, A. & Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. USA* 72, 1184–1188.
21. Benton, W. D. & Davis, R. W. (1977) *Science* 196, 180–182.
22. Thomas, M., White, R. C. & Davis, R. W. (1976) *Proc. Natl. Acad. Sci. USA* 73, 2294–2298.
23. Davis, R., Simon, M. & Davidson, N. (1971) *Methods Enzymol.* 21D, 413–428.
24. Hood, L., Loh, E., Hubert, J., Barstad, P., Eaton, B., Early, P., Fuhrman, J., Johnson, N., Kronenberg, M. & Schilling, J. (1977) *Cold Spring Harbor Symp. Quant. Biol.* 41, 817–836.
25. Berk, A. J. & Sharp, P. A. (1977) *Cell* 12, 721–732.
26. McDonell, M. W., Simon, M. N. & Studier, F. W. (1977) *J. Mol. Biol.* 110, 119–146.
27. Maniatis, T., Sim, G. K., Efstratiadis, A. & Kafatos, F. C. (1976) *Cell* 8, 163–182.
28. Robinson, E. A. & Appella, E. (1977) *Proc. Natl. Acad. Sci. USA* 74, 2465–2469.
29. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
30. Southern, E. M. (1975) *J. Mol. Biol.* 98, 503–517.
31. Riblet, R. J. (1977) *ICN-UCLA Symposia on Molecular and Cellular Biology* (Academic, New York), Vol. 6, pp. 83–89.
32. Robinson, E. A. & Appella, E. (1977) *Proc. Natl. Acad. Sci. USA* 74, 2465–2469.
33. Gearhart, P. J. & Cebra, J. J. (1978) *Nature (London)* 272, 264–265.
34. Bender, W., Davidson, N., Kindle, K., Taylor, W., Silverman, M. & Firtel, R. (1978) *Cell* 15, 779–788.
35. Berkner, K. L. & Folk, W. R. (1977) *J. Biol. Chem.* 252, 3176–3184.
36. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* 74, 560–564.
37. Beale, D. & Feinstein, A. (1976) *Q. Rev. Biophys.* 9, 135–180.
38. Davies, D. R., Padlan, E. A. & Segal, D. M. (1975) *Annu. Rev. Biochem.* 44, 639–667.
39. Frangione, B., Lee, L., Haber, E. & Bloch, K. (1977) *Proc. Natl. Acad. Sci. USA* 70, 1073–1077.
40. Adetugbo, K., Milstein, C. & Secher, D. (1977) *Nature (London)* 265, 299–304.
41. Gilbert, W. (1978) *Nature (London)* 271, 501.

27

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).

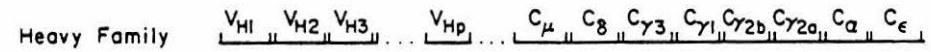


FIGURE 1. Heavy chain antibody gene family in mice. The order of C_H gene segments is uncertain, although indirect evidence supports the following alignment: $C_{\gamma 3}C_{\gamma 1}C_{\gamma 2b}C_{\gamma 2a}C_\alpha$ (20). The number of V_H gene segments is still a matter of controversy. The heavy 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.

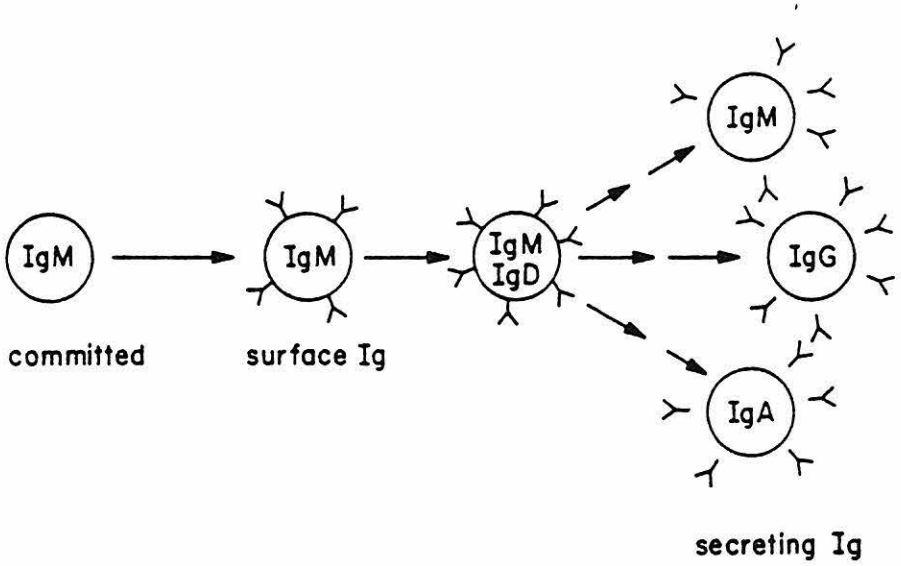


FIGURE 2. The differentiation of B cells. A B cell first becomes committed to the expression of a particular V domain (one V_L region and one V_H region) which is associated with cytoplasmic IgM molecules. Subsequently the IgM molecule is expressed on the cell surface. Later, cell-surface IgD molecules 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.

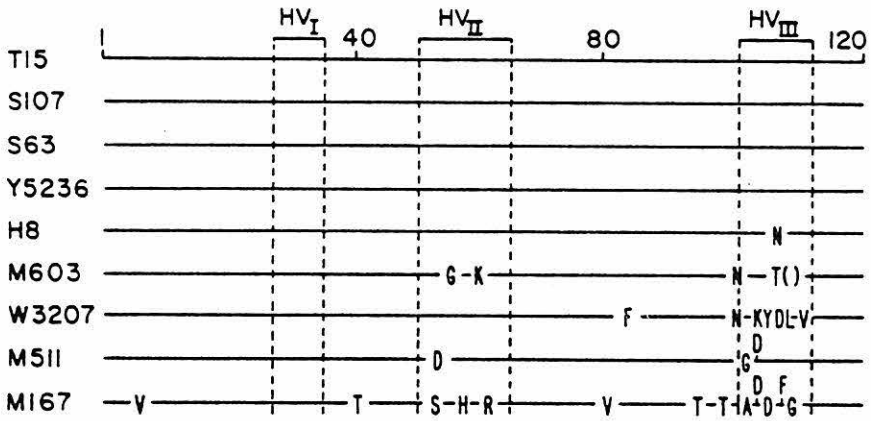


FIGURE 3. The amino acid sequences of V_H regions from immunoglobulins binding phosphorylcholine. Identities of these sequences to the V_H region of T15 are indicated by a straight line. The one letter code of Dayhoff is used to indicate amino acid substitutions (28). Deletions are indicated by brackets. 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_HC_α 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-

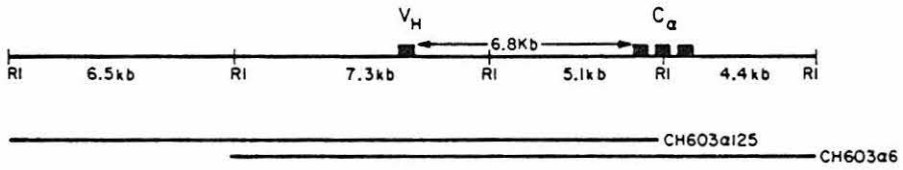


FIGURE 4. The organization of V_H and C_α gene segments from DNA derived from myeloma tumor M603. Kb denotes kilobases. R1 denotes Eco R1 cleavage sites. The distances between Eco R1 sites are indicated. CH603α125 and CH603α6 are two clones derived from the phage library of M603 DNA. The V_H gene segment 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.

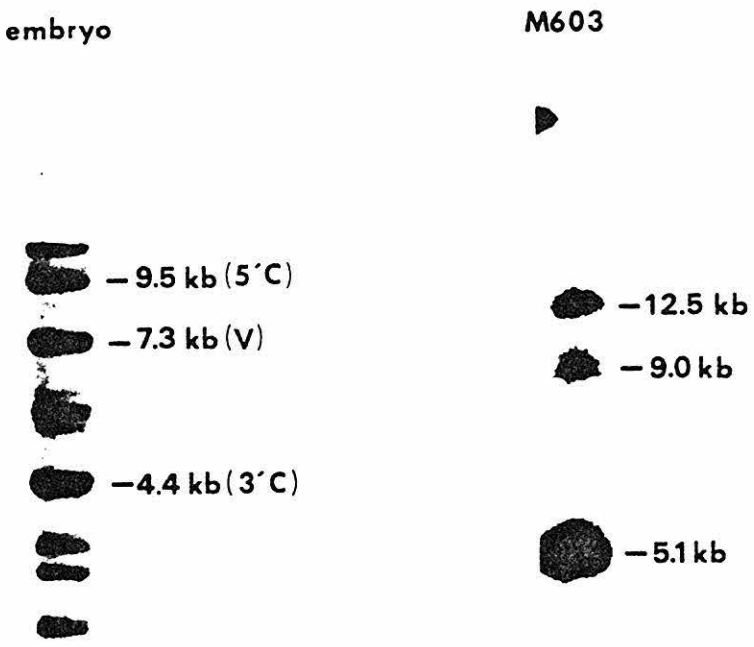


FIGURE 5. Southern blots of embryo (undifferentiated) and myeloma M603 (differentiated) DNAs. The picture on the left is a Southern blot of 13-day embryo DNA after digestion with the Eco R1 enzyme, separation of the DNA fragments on agarose, and hybridization with a cDNA probe derived from mRNA of myeloma 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).

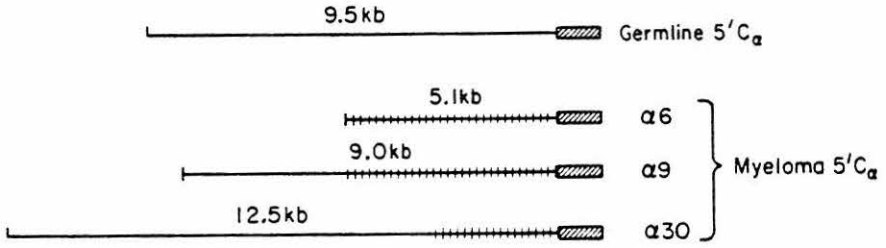


FIGURE 6. Eco R1 genomic fragments including the 5' portion of the C_α gene from myeloma M603 DNA and sperm DNA. The genomic clones α6, α9, and α30 have been derived from the M603 phage library. The structure of the germ line C_α clone comes from a Southern blot analysis of sperm or embryo DNA (Figure 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

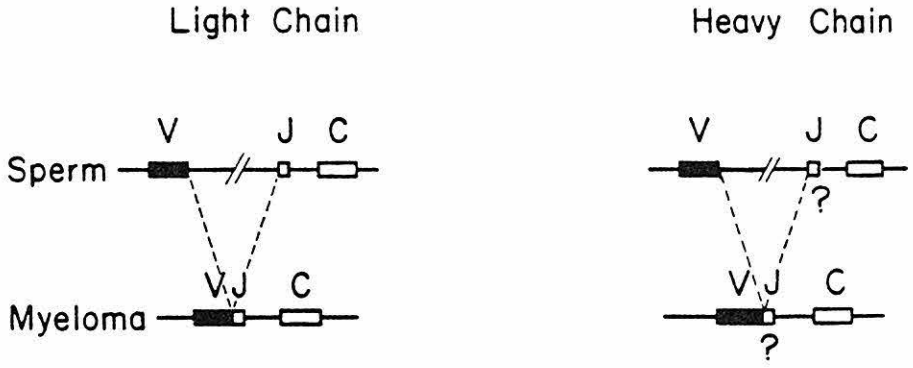


FIGURE 7. A model of the joining of light and heavy chain gene segments. An analysis of λ (4) and κ (18) light chain gene segments indicate that the 3' side of a V segment is fused to the 5' side of a J segment. The intervening DNA sequence between the J segment and the C segment remains unchanged in 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

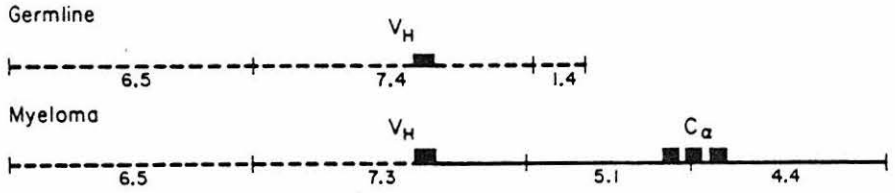


FIGURE 8. Homologies determined by heteroduplex analysis between the flanking sequences of germ line and myeloma V_H clones. The dotted lines indicate germ line sequences. Accordingly, the intervening DNA sequence between the V_H and C_α gene segments is not derived from the sperm V_H clone. The sperm 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 $\gamma 1$ 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_{κ} 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.

An Immunoglobulin Heavy Chain Variable Region Gene is Generated From Three Segments of DNA: V_H, D, and J_H

P. Early, H. Huang, M. Davis, K. Calame, and L. Hood

Division of Biology

California Institute of Technology

Pasadena, California 91125

Cell, in press April 1980

19, 981-992

Running Title: Variable Region Gene Rearrangement

Summary

We have determined the sequences of separate germline genetic elements which encode two parts of a mouse immunoglobulin heavy chain variable region. These elements, termed gene segments, are heavy chain counterparts of the variable (V) and joining (J) gene segments of immunoglobulin light chains. The V_H gene segment encodes amino acids 1-101 and the J_H gene segment amino acids 107-123 of the S107 phosphorylcholine-binding V_H region. This J_H gene segment and two other J_H gene segments are located 5' to the mu constant region gene (C_μ) in germline DNA. We also have determined the sequence of a rearranged V_H gene encoding a complete V_H region, M603, which is closely related to S107. In addition, we have partially determined the V_H coding sequences of the S107 and M167 heavy chain mRNAs. By comparing these sequences to the germline gene segments, we conclude that the germline V_H and J_H gene segments do not contain at least 13 nucleotides which are present in the rearranged V_H genes. In S107, these nucleotides encode amino acids 102-106, which form part of the third hypervariable region and consequently influence the antigen-binding specificity of the immunoglobulin molecule. This portion of the variable region may be encoded by a separate germline gene segment which can be joined to the V_H and J_H gene segments. We term this postulated genetic element the D gene segment, referring to its role in the generation of heavy chain diversity.

Essentially the same noncoding sequences are found 3' to the V_H gene segment and as inverse complements 5' to two J_H gene segments. These are the same conserved nucleotides previously found adjacent to light chain V and J gene segments. Each conserved sequence consists of blocks of seven and ten conserved nucleotides which are separated by a spacer of either 11 or 22 nonconserved nucleotides. The highly conserved spacing, corresponding to one or two turns of the DNA helix,

maintains precise spatial orientations between blocks of conserved nucleotides. Gene segments which can join to one another (V_κ and J_κ, for example) always have spacers of different lengths. Based on these observations, we propose a model for variable region gene rearrangement mediated by proteins which recognize the same conserved sequences adjacent to both light and heavy chain immunoglobulin gene segments.

Introduction

The immunoglobulin molecule is a complex entity with two major functions — recognition of foreign substances (antigen binding) and the elimination or destruction of these foreign substances (effector functions). This bipartite nature is reflected in the structure of immunoglobulins, which are composed of polypeptide chains with many alternative sequences of amino acids near their N-terminal ends (the variable or V region), but only a few possible sequences for the remainder of the chain (the constant or C region). Typical immunoglobulin molecules contain two identical heavy chains and two identical light chains. Pairs of light and heavy chains fold into discrete structural domains — the V region domain which binds antigens, and the C region domains responsible for the effector functions (Gally, 1973). V region domains consist of about 107 amino acids from a light chain and about 125 amino acids from a heavy chain. Within the V region domain, three short polypeptide loops from both the heavy and light chains form the antigen-binding pocket (Segal et al., 1974; Amzel et al., 1974). These short portions of the V region are termed the hypervariable regions, since extensive protein sequence studies of immunoglobulins have shown them to be the most frequent sites for amino acid substitutions between V regions (Wu and Kabat, 1970; Capra and Kehoe, 1974).

The heavy chains and the two types of light chains (κ and λ) which occur in vertebrate immunoglobulin molecules are each encoded by a separate multigene family (Hood et al., 1975). The functionally distinct V and C regions of immunoglobulin polypeptides are encoded by independent genetic elements within each gene family (Dreyer and Bennett, 1965). In the κ light chain gene family of mouse, which has been the most extensively studied, one germline gene encodes the C_κ region. Kappa V regions are encoded by genes which are created during differentiation by the fusion of two independent elements, V and J, which we will call gene segments.

A large number of germline V_κ gene segments encoding amino acids 1-95 can join to one of four germline J_κ gene segments encoding amino acids 96-107 of the V region (Max et al., 1979; Sakano et al., 1979). This process presumably occurs by deletion of the DNA originally separating V_κ and J_κ gene segments (Sakano et al., 1979), and results in the formation of a V_κ gene which encodes all 107 amino acids of a V_κ region as an uninterrupted nucleotide sequence. The J_κ gene segments are closely linked to the C_κ gene, so V-J joining results in a single transcription unit containing both the C_κ gene and the newly formed V_κ gene. It has been postulated that conserved blocks of noncoding nucleotides found 3' to undifferentiated V_κ gene segments and as inverse complements 5' to undifferentiated J_κ gene segments play a role in the mechanism of V-J joining (Max et al., 1979; Sakano et al., 1979).

Our studies have focused on the heavy chain gene family of mouse. In contrast to the κ gene family, the heavy chain family contains at least eight C region genes — μ, δ, γ_1, γ_{2b}, γ_{2a}, γ_3, α, ε. These C_H genes form a tightly linked cluster (Mage et al., 1973) which is linked to various V_H region markers (Riblet, 1977). The heavy chain genes also undergo DNA rearrangements which bring a V_H and a C_H gene close together to form a single transcription unit (Early et al., 1979). During B-cell differentiation, the first constant region expressed is C_μ, leading to the synthesis of an IgM molecule. Subsequently, other DNA rearrangements can produce combinations of different C_H genes with the same V_H gene (Davis et al., 1980) and thereby lead to the expression of other classes of immunoglobulins.

In this paper we examine the nature of the DNA rearrangements which result in the formation of a complete V_H gene. The V_H genes that we have chosen to study encode V_H regions from immunoglobulins which bind phosphorylcholine. The amino acid sequences of nine V_H regions derived from myeloma proteins binding phosphorylcholine have been determined (Figure 1). Four of the V_H regions are

identical to the S107 sequence. The remaining variants differ from the S107 sequence by one to 13 residues and in certain cases by insertions or deletions. Some of the variations correlate with differences in binding constants for phosphorylcholine and related antigens (Padlan et al., 1976; Goetze and Richards, 1977). Many of the amino acid substitutions and all of the deletions and insertions occur in the third hypervariable region, but amino acid substitutions also occur in the second hypervariable region and outside hypervariable regions. Since the S107 V_H region has been sequenced from four independent myeloma proteins (S107, T15, S63, Y5236), it appears to be encoded by a germline gene.

The mouse genome contains 8-9 V_H gene segments homologous to a cloned cDNA probe for phosphorylcholine-binding V_H regions (Davis et al., 1979). We have determined the nucleotide sequence of one of these germline V_H gene segments, which encodes the S107 V_H region. We also have determined the sequence of a germline J_H gene segment which is expressed in most phosphorylcholine-binding V_H regions. These gene segments are compared with a rearranged V_H gene encoding the complete M603 phosphorylcholine-binding heavy chain, and with sequences from mRNAs encoding two other phosphorylcholine-binding heavy chains. We conclude that V_H genes are created during B-cell differentiation by joining three DNA segments: V_H, D and J_H. The mechanism for DNA joining is probably the same in heavy and light chains, and involves protein recognition of blocks of conserved nucleotides adjacent to germline gene segments.

Results

Clones Containing Heavy Chain Gene Segments

Genomic clones were isolated from Charon 4A phage libraries (Maniatis et al., 1978) by the Benton-Davis (1977) filter screening procedure using ^{32}P-labeled cloned cDNAs (Early et al., 1979; Davis et al., 1980; Calame et al., 1980). Phage

in these libraries contain large (12-20 kb) inserts of either BALB/c mouse sperm DNA (germline) or DNA from M603 myeloma tumors (for details see Early et al., 1979 and Davis et al., 1980). Figure 2 depicts three genomic clones containing genes or gene segments which encode V_H regions from myeloma proteins binding phosphorylcholine (these will be termed PC genes or gene segments). Homologies between these clones were partly determined by electron microscopy of heteroduplexed DNAs and by restriction mapping (Davis et al., 1980).

The ChSpV_{PC}3 clone, isolated from a sperm DNA library, contains a germline PC V_H gene segment and will be referred to as the germline PC V_H clone. The ChSpμ27 clone, also derived from a sperm library, contains the C_μ gene. As will be shown later, ChSpμ27 includes a PC J_H gene segment located 6 kb 5' to the C_μ gene. ChSpμ27 will be called the germline J_H clone. Southern blot comparisons between sperm DNA and ChSpμ27 show that the clone has deleted approximately 2 kb of mouse DNA (Davis et al., 1980). The deletion is located about 1 kb 5' to the C_μ gene (M. Davis, unpublished). Therefore, in the mouse genome the PC J_H gene segment is approximately 8 kb 5' to the C_μ gene. The Ch603α6 clone was isolated from an M603 myeloma DNA library and encodes both the M603 V_H region and the C_α region. This clone contains the rearranged V_HC_α gene presumably expressed in the M603 myeloma tumor (Early et al., 1979). Ch603α6 will be referred to as the M603 V_H clone.

We used the method of Maxam and Gilbert (1977) to determine the sequences of V_H and J_H gene segments in these three genomic clones. The sequencing strategy employed for each clone is outlined in Figure 2.

The Germline PC V_H Gene Segment Encodes 101 Amino Acids of the V_H Region

Figure 3 depicts the sequence of the V_H gene segment in the germline PC V_H clone. This gene segment, denoted V_{H107}, encodes the first 101 amino acids of

the PC heavy chains from S107, T15, S63, Y5236 and H8 myeloma tumors (Figure 1).

The coding sequence of V_{H107} ends with the first amino acid of the third hypervariable region. The first 75 nucleotides to the 3' side of this codon do not encode any part of a phosphorylcholine-binding V_H region, and include a termination codon near the end of the V_H gene segment. The remainder of the V_H region must be encoded elsewhere in the germline genome. The noncoding sequence which follows V_{H107} includes the heptanucleotide CACAGTG and the decanucleotide GACACAAACC (Figure 3), both of which resemble conserved nucleotide sequences 3' to light chain V gene segments (Max et al., 1979; Sakano et al., 1979); see Table 1. These nucleotides may play a role in DNA rearrangements of immunoglobulin gene segments (see the Discussion).

A Hydrophobic Signal Peptide is Encoded with the V_H Region

Immunoglobulin polypeptides are translated with a hydrophobic amino terminal "signal peptide" which is subsequently removed to generate the mature immunoglobulin chain (Milstein et al., 1972). Such peptides appear to play a role in the mechanism of protein secretion (Blobel and Dobberstein, 1975). The signal peptide encoded by S107 α mRNA is shown in Figure 3. This figure depicts the nucleotide sequence for part of p107αR5, a cDNA plasmid derived from S107 mRNA (Early et al., 1979). Comparison with the germline V_{H107} sequence in Figure 3 shows that intervening DNA interrupts the signal peptide codons at an AGGT sequence 12 nucleotides from the beginning of the mature V_H region. An RNA splice site at this position also is found in light chain genes (Bernard et al., 1978; Seidman et al., 1979). We have not determined the germline location of the mRNA sequence 5' to this splice point.

The PC J_H Gene Segment is Linked to the C_μ Gene in Germline DNA

In the light chain gene families, J gene segments are closely linked to C genes

(Bernard et al., 1978; Max et al., 1979; Sakano et al., 1979). Since C_μ is the first heavy chain constant region to be expressed by lymphocytes (Raff, 1976), we reasoned that J_H gene segments would be adjacent to the C_μ gene. This gene organization would allow DNA rearrangements to form a complete V_H gene initially associated with the C_μ gene. Subsequently, other DNA rearrangements could replace the C_μ gene with different C_H genes, leading to the expression of other immunoglobulin classes with the same V_H region (Davis et al., 1980).

To locate any PC J_H gene segments in the "germline J_H" clone, which we knew contained a C_μ gene (Calame et al., 1980), restriction fragments were hybridized to a ^{32}P-labeled cDNA plasmid encoding the entire S107 heavy chain. Figure 4 shows a blot (Southern, 1975) of DNA from the germline J_H clone digested with three different restriction enzymes and separated on a polyacrylamide gel. Paired lanes contained digests either of the whole cloned DNA or of a 3.6 kb Hha I fragment (see Figure 2). Each lane shows only a single band of hybridization to the S107 cDNA plasmid. For each restriction enzyme, the band produced from the 3.6 kb Hha I fragment is smaller than the band from whole DNA. This result indicates that the region of the germline J_H clone hybridizing to the S107 cDNA plasmid includes one of the Hha I ends of the 3.6 kb restriction fragment. In other experiments (P. Early, unpublished result), the S107 cDNA plasmid only hybridized to a 6.2 kb Eco RI fragment of this clone. This Eco RI fragment includes the 5', but not the 3', end of the 3.6 kb Hha I fragment (see Figure 2).

These data allow the region of the germline J_H clone containing the PC J_H gene segment to be localized. Since only one band from each restriction digest of whole DNA hybridized to the S107 cDNA plasmid, the region of hybridization must be shared by the three different restriction fragments. Accordingly, any possible PC J_H gene segments must be limited to a sequence of 220 nucleotides

between the Hinf I and Hae III sites shown in Figure 2. This sequence is at the 5' end of a region of homology between the germline J_H clone and the M603 V_H clone (Figure 2). This portion of the germline J_H clone was sequenced to identify the germline PC J_H gene segment.

Germline J_H Gene Segments Encode Part of V_H Regions

Figure 5 shows the sequence of 530 nucleotides in the germline J_H clone, including the 220 nucleotide Hinf I/Hae III fragment which hybridizes with the S107 cDNA plasmid. There is a single PC J_H gene segment in this fragment. This germline gene segment, denoted J_{H107}, encodes amino acids 107-123 of the S107 V_H region, and terminates at the 3' end of the V_H region with an AGGT sequence associated with RNA splicing (Breathnach et al., 1978; Catterall et al., 1978). Neither the V_{H107} nor the J_{H107} gene segments encode amino acids 102-106 for any PC heavy chain. These amino acids comprise a major portion of the third hypervariable region and thus partly determine antigen-binding specificity (Padlan et al., 1976; Goetze and Richards, 1977). We will discuss the origin of the nucleotides encoding this portion of the V_H region subsequently.

A second J_H gene segment, denoted J_{H315}, is located 270 nucleotides 3' from J_{H107} (Figure 5). This J_H gene segment does not encode part of any PC V_H region and is two codons shorter than J_{H107}. These two J_H gene segments differ at 11 of 45 nucleotides. The amino acid sequence encoded by J_{H315} has been found in other heavy chain V_H regions, including M315 which binds dinitrophenol (Francis et al., 1974) and some V_H regions from myeloma proteins which bind α1,3 dextran (Schilling et al., 1980). ^{32}P-labeled heavy chain mRNA from M315 myeloma tumors hybridizes to restriction fragments of the germline J_H clone which contain J_{H315} (P. Early, unpublished results). Similarly, the J_H gene segment

variation also seen in light chain V-J joining (Max et al., 1979; Sakano et al., 1979; Weigert et al., 1978; Weigert et al., 1980).

Sequences of Three PC mRNAs Show Evidence for Two Sites of DNA Joining in V_H Genes

The partial sequences of three mRNAs encoding heavy chains from myeloma proteins binding phosphorylcholine are compared in Figure 7. The data include the codons for amino acids 95-123 of the S107 heavy chain and the homologous portions of the M603 and M167 heavy chains. These sequences include the third hypervariable regions, which differ in size for each case (Figure 1). The mRNA sequences were determined by extension of a cloned cDNA primer in the presence of dideoxynucleotide triphosphates (Sanger et al., 1977).

The deletions and insertions in these mRNAs relative to the S107 sequence fall near the boundaries of the germline V_{H107} and J_{H107} gene segments. A high degree of diversity also is generated in light chains by slight differences in the points at which the V and J gene segments are joined (Max et al., 1979; Sakano et al., 1979; Weigert et al., 1978; Weigert et al., 1980). Diversity in PC heavy chains may be generated by a similar mechanism at two DNA junctions. In S107, these junctions would occur at the codons for amino acids 102 and 106.

The scattered single nucleotide differences between these three mRNAs may reflect their distinct origins from different germline gene segments. The M167 V_H region almost certainly is derived from another germline V_H gene segment, since the remainder of this V_H region differs extensively from S107 (Figure 1). Alternatively, some of the differences between these mRNAs might have arisen from somatic mutation during B-cell differentiation prior to neoplastic transformations, or during growth and propagation of the myeloma tumors.

The J_H sequences of these three mRNAs are identical except in the third hypervariable region. All could be derived from the J_{H107} gene segment if DNA

joining occurred at slightly different positions in each case (Figure 7). The J_{H107} gene segment may be the only J_H gene segment used in the PC heavy chains analyzed to date (compare the protein sequences in Figure 1). There are no other closely related J_H gene segments in the germline J_H clone (Figures 2 and 4).

Discussion

Germline Origins of Rearranged V_H Genes

We have determined the DNA sequence of two germline gene segments, V_{H107} and J_{H107}, which together encode most of the variable region of the S107 heavy chain. The V_{H107} gene segment contains the codons for the first 101 amino acids of the mature heavy chain, including the first and second hypervariable regions. The J_{H107} gene segment encodes amino acids 107-123 of the S107 V_H region. The J_{H107} gene segment is approximately 8 kb to the 5' side of the C_μ gene in germline DNA, while the V_{H107} gene segment is not closely associated with the C_μ or C_α genes in the germline.

We also have determined the sequence of one rearranged PC V_H gene from the M603 myeloma tumor. As shown in Figures 6 and 7, the M603 V gene includes the last 16 codons of the J_{H107} gene segment. The first three hundred nucleotides to the 3' side of the J_{H107} sequence show only three substitutions between the germline J_H clone and the M603 V_H clone (Figure 6). These differences could be due to polymorphism among BALB/c mice or mutation in the myeloma tumors. The region of homology between the myeloma M603 and the germline J_H clones extends 5 kb from the 3' side of the J_{H107} gene segment (Figure 2). There are no other J_H gene segments closely related to J_{H107} in this region of homology or elsewhere in the germline J_H clone (Figure 4). Southern blot experiments also have shown that a probe from this region of homology on the 3' side of the J_H

sequence hybridizes strongly only to the C_μ band when germline DNA is digested with either of two restriction endonucleases (Davis et al., 1980). These results indicate that the J_{H107} gene segment is the unique germline J_H precursor for the rearranged M603 V_H gene. The J_{H107} gene segment also is probably the precursor for the other known PC variable regions, as suggested by mRNA (Figure 7) and protein sequences (Figure 1).

The germline V_{H107} gene segment is not completely identical to the corresponding portion of the M603 V_H gene. There are seven differences in 303 coding nucleotides, creating three amino acid substitutions (Figure 6). However, the germline V_{H107} clone and the rearranged M603 V_H clone share 11.3 kb of homology 5' to the V_H sequences (Figure 2). If another germline V_H gene segment is the precursor for the M603 gene, it must be very closely related to V_{H107}. If the V_{H107} gene segment is the germline precursor of the rearranged M603 V_H gene, it must have accumulated seven nucleotide differences (three of which are silent), either during B-cell differentiation in the mouse from which the M603 tumor was initially isolated, or during growth and propagation of the tumor as it was passaged more than 100 generations. The V_{H107} gene segment does encode the first 101 amino acids of the most frequently expressed phosphorylcholine-binding V_H region, exemplified by S107, without somatic mutation. Thus, the first and second hypervariable regions are not encoded by "minigenes" inserted into germline V_H genes during lymphocyte differentiation as postulated by Kabat and his coworkers (Wu and Kabat, 1970; Kabat et al., 1978).

All J_H Gene Segments may be 5' to the C_μ Gene in Germline DNA

In the germline, J_H gene segments probably are only associated with the C_μ gene and not with other C_H genes. Two lines of evidence support this conclusion. First, we have demonstrated that three separate J_H gene segments, J_{H107}, J_{H315},

and a J_H gene segment which hybridizes to M21 mRNA, all are 5' to the germline C_μ gene. Second, as B cells differentiate, they produce combinations of different C_H genes with the same V_H gene (Davis et al., 1980). Since the J_H gene segment forms a significant part of the V_H gene, the initial V_H-D-J_H joining event is expected to provide the rearranged V_H gene for all subsequent immunoglobulin classes derived from a particular B cell. The first heavy chain constant region expressed is always C_μ (Raff, 1976), and C_μ is apparently the constant region gene associated with V_H-D-J_H joining.

The D Gene Segment: A Third Part of the V_H Gene

Amino acids 102-106 of the S107 or M603 V_H regions are not encoded by either the V_{H107} or J_{H107} germline gene segments. Figure 7 shows that at least 13 nucleotides are added in rearranged V_H genes between the V_{H107} and J_{H107} gene segments. These nucleotides encode part of the third hypervariable region and can affect the antigen binding properties of the immunoglobulin molecule (Padlan et al., 1976; Goetze and Richards, 1977). This observation is very different from the V-J joining described for immunoglobulin light chains (Bernard et al., 1978; Max et al., 1979; Sakano et al., 1979), where V and J gene segments together account for the complete rearranged V gene. Protein sequence and serological studies also have suggested that the third hypervariable region in heavy chains may have an origin independent from the rest of the V_H region (Wu and Kabat, 1970; Capra and Kindt, 1975; Rao et al., 1979; Schilling et al., 1980).

Where do the "extra" nucleotides in the PC V_H genes come from? Other explanations are possible, but they probably originate from a third germline gene segment (D, or diversity) which joins between V_H and J_H gene segments. Note that in the V_H and J_H gene segments (Figures 3 and 5), coding sequences are separated by at most three nucleotides from the conserved heptanucleotide sequence which

is associated with the point of DNA joining in light chains (Max et al., 1979; Sakano et al., 1979). The position of this conserved sequence, which probably plays the same role in both heavy and light chain DNA rearrangements, leaves no room for the D nucleotides to arise from DNA directly contiguous to either the V_H or J_H gene segments. Our model of V_H-D-J_H joining is illustrated in Figure 8. We propose that germline D gene segments are located between the V_H and J_H gene segments so that the V_H, D and J_H gene segments can be joined in the same order as they are found in the rearranged V_H gene. The D gene segments might be associated with the C_μ gene, as are J_H gene segments, or alternatively, they might be closely linked to V_H gene segments. If the latter possibility is correct, each group of V_H gene segments might have its own set of D gene segments. The V_H regions from myeloma proteins binding levan may lack D segments (Schilling et al., 1980). Thus, some V_H gene segments might join directly to J_H gene segments in a manner similar to light chain V-J joining.

DNA Rearrangement as a Generator of Antibody Diversity

DNA joining can produce variable region diversity by at least two means: combinatorial joining of germline gene segments and codon alterations at the site of joining. Combinatorial joining suggests that any gene segment may be joined to any other. Thus, 200 V_κ and 4 J_κ gene segments could generate 800 V_κ genes. If V_H regions are encoded by three sets of germline gene segments, V_H, D, and J_H, the extent of combinatorial diversity will be greater than for light chains which are encoded only by V and J gene segments.

In light chains, additional diversity is generated by variations in the site of V-J joining (Max et al., 1979; Sakano et al., 1979; Weigert et al., 1978; Weigert et al., 1980). This junctional diversification mechanism can add or subtract single V or J codons

from the rearranged gene, or create new hybrid codons. In heavy chains, the joining of V_H, D and J_H gene segments would provide two sites for diversification by this junctional somatic mutational mechanism. A comparison of the S107 and M603 mRNAs (Figure 7) provides an example of such junctional diversity between D and J_{H107}.

The Mechanism of DNA Rearrangement in V Genes

Light and heavy chain V and J gene segments all share one important feature. Blocks of relatively conserved noncoding nucleotides occur 3' to V gene segments and 5' to J gene segments (Bernard et al., 1978; Max et al., 1979; Sakano et al., 1979). The most common nucleotides adjacent to either light or heavy chain V gene segments are CACAGTG....$^{A}_{G}$ACA$^{A}_{T}$AAACC, where the conserved hepta- and decanucleotides are separated by either 11 or 22 essentially random nucleotides (Table 1). Similarly, light and heavy chain J gene segments are adjacent to GGTTTTTGTA....CACTGTG, where again the blocks of conserved nucleotides are separated by either 11 or 22 random nucleotides (Table 1). The conserved nucleotides next to V and J gene segments are nearly inverse complements of one another, as seen in Figure 8. By contrast, the 11 or 22 nucleotides which separate blocks of conserved nucleotides show no particular homologies, except between some related V gene segments (see $V_{\kappa 41}$, $V_{\kappa 2}$, and $V_{\kappa 3}$ in Table 1).

What is remarkable, though, is the highly conserved spacing between the hepta- and decanucleotide sequences. In every case, the spacing is either 11 or 22 nucleotides, deviating from these values by at most one nucleotide (Table 1). The only exception is a J_κ gene segment which appears to be nonfunctional (Max et al., 1979; Sakano et al., 1979). V_κ gene segments have an 11 nucleotide spacer, and J_κ gene segments have a 22 nucleotide spacer. The $V_{\lambda I}$ gene segment has a 22 nucleotide spacer, and the $J_{\lambda I}$ gene segment an 11 nucleotide spacer. Heavy

chain V and J gene segments have 22 nucleotide spacers. No information is yet available for the putative heavy chain D gene segments.

One turn of the DNA helix requires about 10.4 nucleotide pairs (Wang, 1979). Thus, the conserved spacing of 11 or 22 nucleotides corresponds to either one or two turns of the helix, plus an extra nucleotide. This spacing means that the near ends of the conserved hepta- and decanucleotides maintain a precise relative orientation on the same side of the DNA helix. While separated by one or two turns of DNA, the two blocks of conserved nucleotides are in much the same relative orientation on the helix as if they formed a continuous stretch of 17 conserved nucleotides.

The foregoing points may be summarized as follows. The same noncoding nucleotides are conserved in all three immunoglobulin gene families, which diverged in evolution sometime prior to the origin of vertebrates (Marchalonis and Cone, 1973). Accordingly, these sequences have been conserved for more than 500 million years. The conserved nucleotides 3' to the V gene segments and 5' to J gene segments are approximately inverse complements of one another. The conserved nucleotides always occur as a heptanucleotide and a decanucleotide, separated by either 11 or 22 essentially random nucleotides. This spacing corresponds closely to one or two turns of the DNA helix. The data currently available indicate that gene segments of one type within a gene family all have spacers of the same size. At least for the light chains, gene segments which can join to one another have different sized spacers: V_κ spacers are "one turn", J_κ spacers are "two turns"; the $V_{\lambda I}$ spacer is "two turns", the $J_{\lambda I}$ "one turn".

These observations suggest a model for DNA rearrangement in V genes. Specific joining proteins exist in the precursors of B lymphocytes which can recognize and bind to the blocks of conserved nucleotides adjacent to V and J gene segments

(and probably D segments as well). The same joining proteins could bind in opposite orientations to both V and J gene segments, since the recognition nucleotides are nearly inverse complements of one another (Figure 8). The variations from the "prototype" recognition sequence $\frac{\text{TACAAAAACC}}{\text{ATGTTTTTGG}}$ are most pronounced for V gene segments (Table 1). This may reduce joining protein binding to individual V gene segments to compensate for their larger numbers, if optimal DNA joining requires roughly equal numbers of protein molecules bound to each type of gene segment. One form of joining protein may bind to blocks of nucleotides with a "one turn" spacer, and another form to blocks with a "two turn" spacer. These forms may be different aggregations of the same subunits. The strong conservation of spacer lengths in the sequences shown in Table 1 probably reflects a fairly rigid orientation of two binding sites in the joining proteins, one for each block of conserved nucleotides. Greater deviations of spacer length would not allow both sites to interact with the DNA simultaneously. A joining protein bound across a "two turn" spacer can interact with a joining protein bound across a "one turn" spacer to form a complex (Figure 8) in which the two gene segments are subsequently cut and ligated together. Slight differences in the points of cutting and joining can produce junctional codon variations (Max et al., 1979; Sakano et al., 1979; Weigert et al., 1978; Weigert et al., 1980).

If there is a functional difference between 11 and 22 nucleotide spacers as we suggest, gene segments of the same type would be prevented from joining to one another. Otherwise if, for example, two V_κ gene segments were to join, breakage and inversion of the chromosome could result. Our model predicts that the heavy chain D gene segments will have 11 nucleotide spacers on both sides. This would allow both V_H-D and D-J_H joining to occur by the "one turn plus two turns" mechanism. It would also prevent V_H and J_H gene segments from joining without a D gene segment, unless some V_H or J_H gene segments had 11 nucleotide spacers.

One important implication of this model is that similar or identical enzyme systems would mediate the DNA rearrangements of the V-J (and D) gene segments in all three immunoglobulin families. This supposition would account both for the conserved recognition sequences and the conserved spacing relationships.

There is one additional interesting observation. The D nucleotides in S107 mRNA form a palindrome around a central G (Figure 9). The same sort of symmetry also is seen in the D nucleotides of M603 and M167 mRNAs (Figure 7). Whether this symmetry might be important for rearrangement of the postulated D gene segments is unknown.

There are many cases in prokaryotes in which proteins bind specifically to DNA sequences with a two-fold axis of symmetry (see Lewin, 1977). The structure shown in Figure 8 differs from these prokaryotic binding sites mainly by including a large loop of DNA between the two nearly symmetric halves. A mechanism for DNA rearrangement which involves protein recognition of specific sites in duplex DNA appears more plausible than the cruciform structure postulated by others (Sakano et al., 1979; Max et al., 1979). Such a structure is not likely to form spontaneously between widely separated regions of DNA containing immunoglobulin gene segments. Furthermore, the cruciform model would not account for the conservation of "one turn" and "two turn" spacers or of the same blocks of noncoding nucleotides in widely divergent gene families. An analogy also may be drawn with the operator sequences in λ phage, where repressor binding occurs to the symmetric duplex DNA rather than an alternative cruciform structure (Maniatis and Ptashne, 1973).

The Evolution of Systems for Specific Protein-DNA Interactions

In the immunoglobulin gene segments, a system of protein-DNA interactions has apparently evolved which recognizes two specific blocks of conserved nucleotides

and the spacing between them. Probably two separate protein domains or subunits exist, one binding to $\frac{\text{CACAGTG}}{\text{GTGTCAC}}$ and one to $\frac{\text{TACAAAAACC}}{\text{ATGTTTTTGG}}$. These separate binding sites may once have been present on independent proteins, but have since been combined to produce two alternative kinds of joining proteins. We postulate that one of these kinds of joining proteins binds to blocks of specific nucleotides separated by a "one turn" spacer, and the other kind to blocks separated by a "two turn" spacer. This may represent a general strategy in eukaryotes for generating highly specific DNA-binding proteins by combining smaller sequence-specific proteins. The specificity of the combined protein is determined not only by the subunit specificities, but also by the overall protein conformation, which determines the length of any spacer between blocks of recognition nucleotides.

Mouse light and heavy chain gene families are known to be on separate chromosomes (Mage et al., 1973; Hengartner et al., 1978; Swan et al., 1979). This chromosomal separation may be a necessary consequence of the use of the same or similar recognition signals for DNA rearrangement in all these gene families. Otherwise, joining might occur between V and J (or D) gene segments from different families, leading to nonfunctional immunoglobulins.

Experimental Procedures

Cloned cDNA Probes

p107αR5 contains 1550 nucleotides of the S107α mRNA sequence, including the entire translated portion (Early et al., 1979). The insert is bordered by Eco RI synthetic linkers in the Eco RI site of pMB9. p107αR5 was subcloned to produce a plasmid which hybridizes only to PC V_H gene segments. A 445 bp V_H fragment of p107αR5 can be prepared by digestion with Eco RI + Hha I. This fragment is bounded by an Eco RI linker and a Hha I cleavage site in amino acid codons 114-115,

the middle of the J_H sequence in the S107 V_H gene. Terminal transferase (P/L Biochemicals) was used to add oligo(dC)$_{10-20}$ to Eco RI + Hha I digested p107αR5 (Roychoudhury et al., 1976). The tailed DNAs were separated on an 8% polyacrylamide gel, and the 445 bp V_H fragment eluted by the method of Maxam and Gilbert (1977), omitting SDS. Oligo(dG)$_{10-20}$ was added to pBR322 which had been linearized with Pst I (W. Rowekamp and R. Firtel, personal communication). Equimolar amounts of these two tailed DNAs were annealed and used directly to transform E. coli χ1776 (Villa-Komaroff et al., 1978). One of the resultant plasmid clones, p107V1, contains the 445 bp V_H fragment of p107αR5, flanked by oligo(dG:dC) tracts and the regenerated pBR322 Pst I sites. This plasmid was labeled with ^{32}P by nick translation (Maniatis et al., 1975) as a probe for PC V_H gene segments. It hybridizes only weakly to the J_{H107} gene segment under the conditions used (Engel and Dodgson, 1978). ^{32}P-labeled p107αR5 was used to detect PC J_H gene segments (it will also hybridize to PC V_H gene segments and the C_α gene).

Germline Clones

The germline Charon 4A phage library from which ChSpPC3 and ChSpμ27 were isolated was derived from BALB/c mouse sperm DNA partially digested with Eco RI (Davis et al., 1980). The cells used to prepare DNA were assayed by light microscopy to be greater than 90% sperm (Joho et al., 1980). We also have isolated genomic clones with gene segments homologous to V_{H107} and J_{H107} from a germline library containing greater than 99% sperm DNA (P. Early, M. Davis, K. Calame, unpublished observations). Southern blots of Eco RI digested sperm and embryo DNAs show no difference in the bands observed by hybridization to S107 PC V_H probes (Davis et al., 1979, 1980).

mRNA Sequencing

A modification of the dideoxynucleotide procedure (Sanger et al., 1977) was employed

to determine the sequences of cDNAs made from α mRNA templates. An 80 bp Hinf I fragment of p107αR5 encoding C_α amino acids 125-150 was isolated from 10% polyacrylamide gels by the procedure of Maxam and Gilbert (1977), omitting SDS. Approximately 0.3 pmole each of an mRNA and the 80 bp Hinf I fragment were mixed in 2 µl of 33 mM NaCl, 33 mM Tris-HCl, pH 7.5, and sealed in a siliconized glass capillary. This was boiled 2 min and annealed at 65°C 30 min. The capillary was broken and the contents blown into a tube containing 1 µl 66 mM $MgCl_2$, 66 mM dithiothreitol and 5 µCi α [^{32}P]dCTP (Amersham, 350 Ci/mmole). The volume was adjusted to 10 µl with H_2O and 2 µl distributed to each reaction tube. Deoxynucleotide triphosphates, dideoxynucleotide triphosphates (P/L Biochemicals), buffer and AMV RNA-dependent DNA polymerase (J. Beard) were added and incubated essentially as described (Sanger et al., 1977). Reactions were lyophilized and suspended in 3.5 µl of 80% formamide, 10 mM EDTA plus dyes, prior to electrophoresis on a 0.4 mm 8% Tris-borate urea gel (Sanger and Coulson, 1978).

Acknowledgements

This work was supported by NSF grant PMC 76-81546. P.E. and M.D. are supported by National Institutes of Health Training Grant GM 07616. K.C. is supported by National Institutes of Health Fellowship GM 05442. H.H. is supported by American Cancer Society Fellowship PF1351. We thank Tom Maniatis, Norman Davidson, Max Delbruck, Jim Posakony, Jim Schilling, and Steve Crews for helpful discussions. During the preparation of this manuscript, we were advised by N. Newell, J. Richards, and F. Blattner of their discovery of additional J_H gene segments.

References

Amzel, L. M., Poljak, R. J., Saul, F., Varga, J. M., and Richards, F. F. (1974). Proc. Nat. Acad. Sci. USA 71, 1427-1430.

Benton, W. D., and Davis, R. W. (1977). Science 196, 180-182.

Bernard, O., Hozumi, N., and Tonegawa, S. (1978). Cell 15, 1133-1144.

Blobel, G., and Dobberstein, B. (1975). J. Cell Biol. 67, 835-851.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., and Chambon, P. (1978). Proc. Nat. Acad. Sci. USA 75, 4853-4857.

Calame, K., Rogers, J., Early, P. W., Davis, M. M., Livant, D. L., Wall, R., and Hood, L. (1980). Nature, submitted.

Capra, J. D., and Kehoe, J. M. (1974). Proc. Nat. Acad. Sci. USA 71, 845-848.

Capra, J. D., and Kindt, T. J. (1975). Immunogenetics 1, 417-427.

Catterall, J. F., O'Malley, B. W., Robertson, M. A., Staden, R., Tanaka, Y., and Brownlee, G. G. (1978). Nature 275, 510-513.

Davis, M. M., Early, P. W., Calame, K., Livant, D. L., and Hood, L. (1979). In Eukaryotic Gene Regulation, R. Axel, T. Maniatis, and C. F. Fox, eds. (New York: Academic Press) in press.

Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L., and Hood, L. (1980). Nature, submitted.

Dreyer, W. J., and Bennett, J. C. (1965). Proc. Nat. Acad. Sci. USA 54, 864-868.

Early, P. W., Davis, M. M., Kaback, D. B., Davidson, N., and Hood, L. (1979). Proc. Nat. Acad. Sci. USA 76, 857-861.

Engel, J. D., and Dodgson, J. B. (1978). J. Biol. Chem. 253, 8239-8246.

Francis, S. H., Leslie, R. G. Q., Hood, L., and Eisen, H. N. (1974). Proc. Nat. Acad Sci. USA 71, 1123-1127.

Gally, J. A. (1973). In The Antigens, M. Sela, ed. (New York: Academic Press) pp. 162-298.

Goetze, A. M., and Richards, J. H. (1977). Proc. Nat. Acad. Sci. USA 74, 2109-2112.

Hengartner, H., Meo, T., Muller, E. (1978). Proc. Nat. Acad. Sci. USA 75, 4494-4498.

Hood, L., Campbell, J. Y., and Elgin, S. C. R. (1975). Ann. Rev. Genetics 9, 305-353.

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, 817-836.

Joho, R., Weissman, I. L., Early, P., Cole, J., and Hood, L. (1980). Proc. Nat. Acad. Sci. USA, in press.

Kabat, E. A. (1976). Structural Concepts in Immunology and Immunochemistry, 2nd ed. (New York: Holt, Rinehart, Winston) p. 294.

Kabat, E. A., Wu, T. T., and Bilofsky, H. (1978). Proc. Nat. Acad. Sci. USA 75, 2429-2433.

Lewin, B. (1977). Gene Expression, Vol. 3 (New York: Wiley) pp. 337-343.

Mage, R., Lieberman, R., Potter, M., and Terry, W. D.-(1973) In The Antigens, M. Sela, ed. (New York: Academic Press) pp. 299-376.

Maniatis, T., and Ptashne, M. (1973). Proc. Nat. Acad. Sci. USA 70, 1531-1535.

Maniatis, T., Jeffrey, A., and Kleid, D. G. (1975). Proc. Nat. Acad. Sci. USA 72, 1184-1188.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., and Efstratiadis, A. (1978). Cell 15, 687-701.

Marchalonis, J. J., and Cone, R. E. (1973). Aust. J. Exper. Biol. Med. Sci. 51, 461-488.

Max, E. E., Seidman, J. G., and Leder, P. (1979). Proc. Nat. Acad. Sci. USA 76, 3450-3454.

Maxam, A. M., and Gilbert, W. (1977). Proc. Nat. Acad. Sci. USA 74, 560-564.

Milstein, C., Brownlee, G. G., Harrison, T. M., and Mathews, M. B. (1972). Nature New Biol. 239, 117-120.

Padlan, E. A., Davies, D. R., Rudikoff, S., and Potter, M. (1976). Immunochemistry 13, 945-949.

Raff, M. C. (1976). Cold Spring Harbor Symp. Quant. Biol. 41, 159-162.

Rao, D. N., Rudikoff, S., Krutzsch, H., and Potter, M. (1979). Proc. Nat. Acad. Sci. USA 76, 2890-2894.

Riblet, R. J. (1977). In The Immune System: Genetics and Regulation, E. E. Sercarz, L. A. Herzenberg, and C. F. Fox, eds. (New York: Academic Press) pp. 83-89.

Roychoudhury, R., Jay, E., and Wu, R. (1976). Nucleic Acids Res. 3, 101-116.

Sakano, H., Huppi, K., Heinrich, G., and Tonegawa, S. (1979). Nature 280, 288-294.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Proc. Nat. Acad. Sci. USA 74, 5463-5467.

Sanger, F., and Coulson, A. R. (1978). FEBS Lett. 87, 107-110.

Schilling, J., Clevinger, B., Davie, J. M., and Hood, L. (1980). Nature 283, 35-40.

Segal, D. M., Padlan, E. A., Cohen, G. H., Rudikoff, S., Potter, M., and Davies, D. R. (1974). Proc. Nat. Acad. Sci. USA 71, 4298-4302.

Seidman, J. G., Max, E. E., and Leder, P. (1979). Nature 280, 370-375.

Southern, E. M. (1975). J. Mol. Biol. 98, 503-517.

Swan, D., D'Eustachio, P., Leinwand, L., Seidman, J., Keithley, D., and Ruddle, F. H. (1979). Proc. Nat. Acad. Sci. USA 76, 2735-2739.

Villa-Komaroff, L., Efstratiadis, A., Broome, S., Lomedico, P., Tizard, R., Naber, S. P., Chick, W. L., and Gilbert, W. (1978). Proc. Nat. Acad. Sci. USA 75, 3727-3731.

Wang, J. C. (1979). Proc. Nat. Acad. Sci. USA 76, 200-203.

Weigert, M., Gatmaitan, L., Loh, E., Schilling, J., and Hood, L. (1978). Nature 276, 785-790.

Weigert, M., Perry, R., Kelley, D., Hunkapiller, T., Schilling, J., Hood, L. (1980). Nature, in press.

Wu, T. T., and Kabat, E. A. (1970). J. Exp. Med. 132, 211-240.

Figure Legends

Figure 1. The Phosphorylcholine-binding (PC) Group of V_H Regions

The S107 protein sequence is indicated by a horizontal line. Differences are shown by the one-letter amino acid code; insertions are indicated by a vertical line, deletions by brackets. HV1-3 are the hypervariable regions (Kabat, 1976). Numbering is from the first amino acid of S107. Redrawn from Hood et al., 1976.

Figure 2. Genomic Clones and Sequencing Strategies

Homology between the myeloma and germline clones (indicated by shading) was determined by electron microscopy of heteroduplexes and by restriction site comparisons (Davis et al., 1980), as well as by nucleotide sequences. The raised boxes are coding sequences, with the 5' ends to the left. $ChSpV_{PC}3$ (germline V_H) and ChSpμ27 (germline J_H) were isolated from a germline (sperm) library. The map of Ch603α6/Ch603α125 is a composite of two overlapping clones from an M603 myeloma library. Ch603α6 contains the rightward three Eco RI fragments, and Ch603α125 the leftward three fragments. In the text, this composite is referred to as "Ch603α6" or the M603 V_H clone. The expanded maps under each clone show the locations of coding sequences more exactly. The direction and extent of sequence determinations are shown by arrows. The expanded map around the J_{H107} gene segment uses crosshatches to show the region of overlap between restriction fragments which hybridize to a PC J_H probe (Figure 4). The J_{H107} gene segment is the only J_H gene segment within these bounds. The numbers above the line indicate distances of the restriction sites from the Hha I site in J_{H107}.

Figure 3. Sequence of a PC V_H Gene Segment

The top lines show a partial sequence of the S107 cDNA plasmid p107αR5. The presumptive "signal peptide" is italicized. Amino acids are numbered from the beginning of the mature S107 heavy chain. The lower lines show the sequence of V_{H107}, the PC germline V_H gene segment from $ChSpV_{PC}3$. HV1 and HV2 indicate the first two hypervariable regions (Kabat, 1976). The boxes enclose conserved noncoding sequences shared with light chain V gene segments. A termination codon following the V_{H107} gene segment is underlined.

Figure 4. Southern Blot to Locate the PC J_H Gene Segment

Either whole ChSpμ27 DNA, or a 3.6 kb Hha I fragment from this clone (see Figure 2) was digested with the indicated restriction enzymes. Digests were electrophoresed on an 8% acrylamide gel, denatured, and transferred to nitrocellulose (Southern, 1975) with 1 M sodium acetate (pH 5). The blot was hybridized to p107αR5 labeled with ^{32}P by nick translation. p107αR5 is a cDNA plasmid encoding the complete S107 heavy chain, including the J_H sequence.

Figure 5. Sequences of Two Germline J_H Gene Segments in ChSpμ27

The amino acids encoded by the J_{H107} gene segment are numbered from the beginning of the mature S107 heavy chain. The Hinf I and Hae III sites mark the boundaries of the PC J_H locus determined by hybridization (Figure 2). J_{H315} has been found only in non-PC V_H regions, including M315. Boxes enclose conserved noncoding sequences shared with light chain J gene segments.

Figure 6. Sequence of the M603 V_H Gene

Amino acids are numbered from the beginning of the mature M603 heavy chain. HV1-3 denote the three hypervariable regions (Kabat, 1976). Where the V_{H107} or J_{H107} germline gene segments differ from the M603 sequence, the germline nucleotides are shown below. The D nucleotides are boxed, and the nucleotides derived from J_{H107} are underlined.

Figure 7. Comparison of PC mRNAs and Germline Gene Segments

The dideoxynucleotide procedure (Sanger et al., 1977) was used to determine partial sequences of cDNAs made to heavy chain mRNAs from S107, M603 and M167 myelomas. A Hinf I fragment of p107αR5 encoding C_α amino acids 125-150 served as primer for DNA synthesis on the mRNA templates. Sequences have been positioned with deletions or insertions to maximize homology. Identity with the S107 mRNA sequence is indicated by horizontal lines. Arrows above show the minimum extent of each gene segment in these mRNAs. The boundaries between the gene segments probably differ somewhat for each mRNA, but cannot be precisely located from these data. The sequence of the M603 gene also was determined from the M603 V_H clone. For comparison, portions of the germline V_{H107} and J_{H107} gene segments are shown beneath, with the conserved noncoding sequences in boxes.

Figure 8. Model for V_H Gene Rearrangement

The upper diagram depicts possible germline D gene segments. Relative distances of the various gene segments from one another are undetermined. The short arrows indicate conserved noncoding sequences (see Table 1) which may be involved in DNA rearrangement. In this model, DNA rearrangement joins V_H-D-J_H gene segments, or perhaps in some cases just V_H-J_H. Intervening DNA may be deleted,

Figure 8 (continued)

or could undergo other types of rearrangement. The lower diagram shows paired V_H and D (alternatively D and J_H, or V and J) gene segments. Putative DNA-joining proteins might bind to the areas enclosed by dashed lines. The gene segments are represented as colinear to emphasize the symmetry of the conserved noncoding nucleotides. The actual structure may bring the ends of the two gene segments into close proximity.

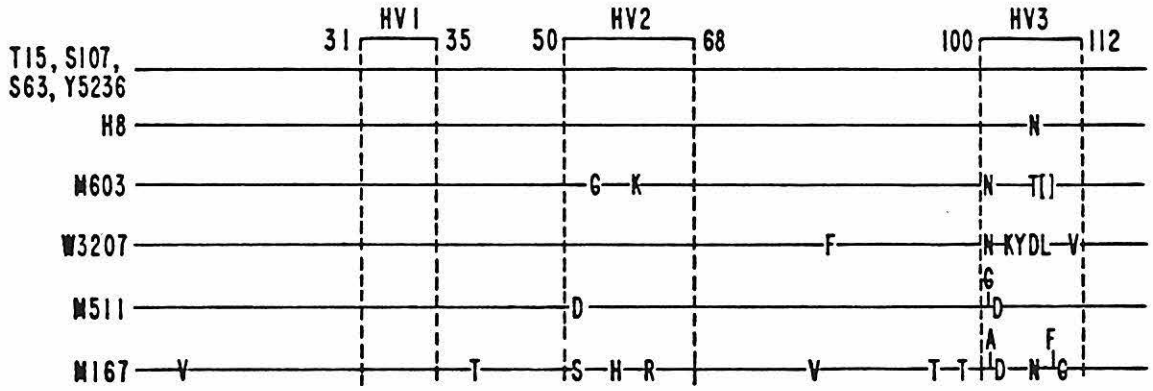
Figure 9. Symmetry of the S107 D Sequence

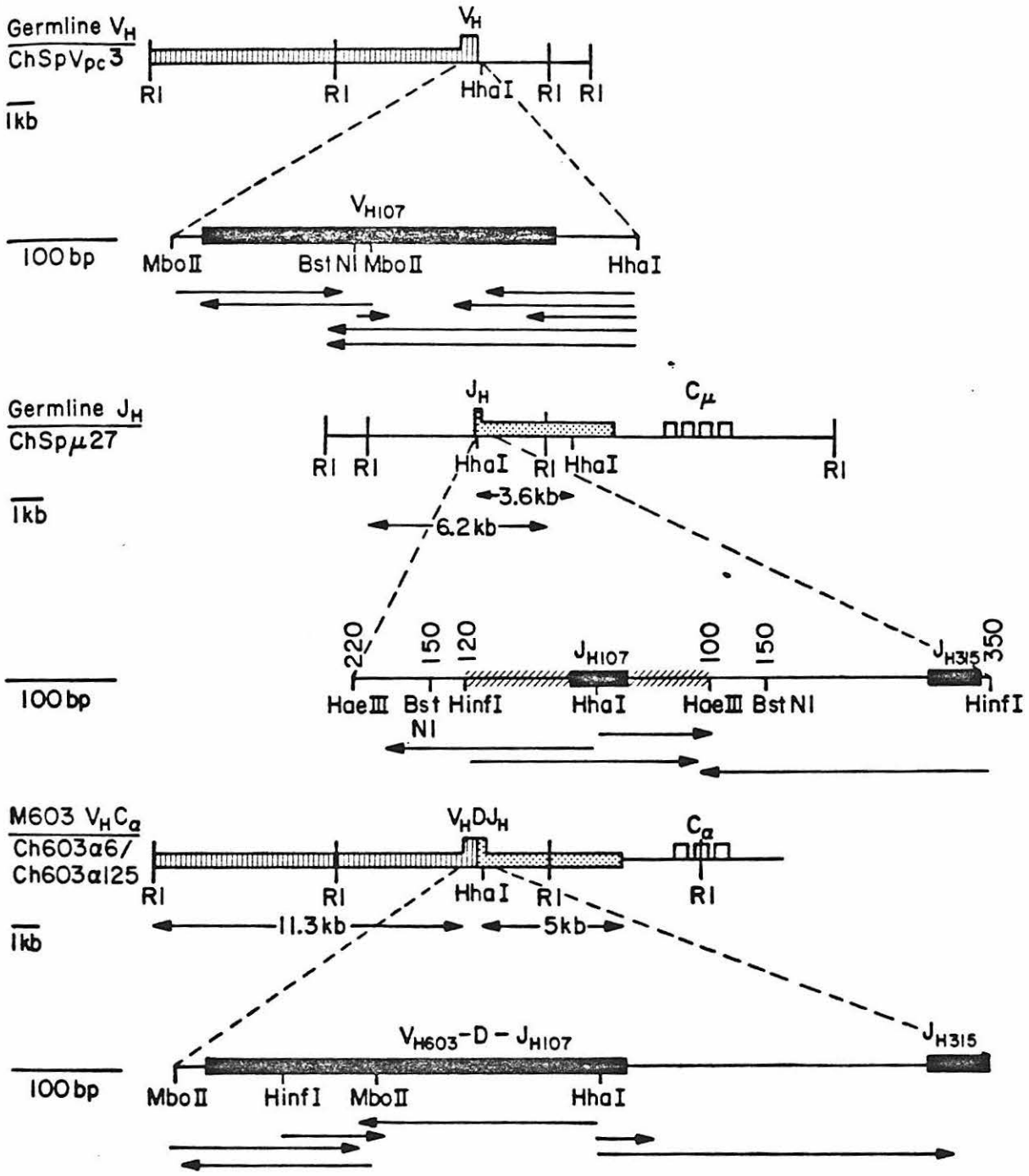
The partial sequence shown was determined from S107 mRNA. Portions of this sequence which could derive from the V_{H107} and J_{H107} germline gene segments are listed horizontally. The vertical stem (D) is self-complementary. An alternative representation of the D nucleotides would be as a palindrome in duplex DNA.

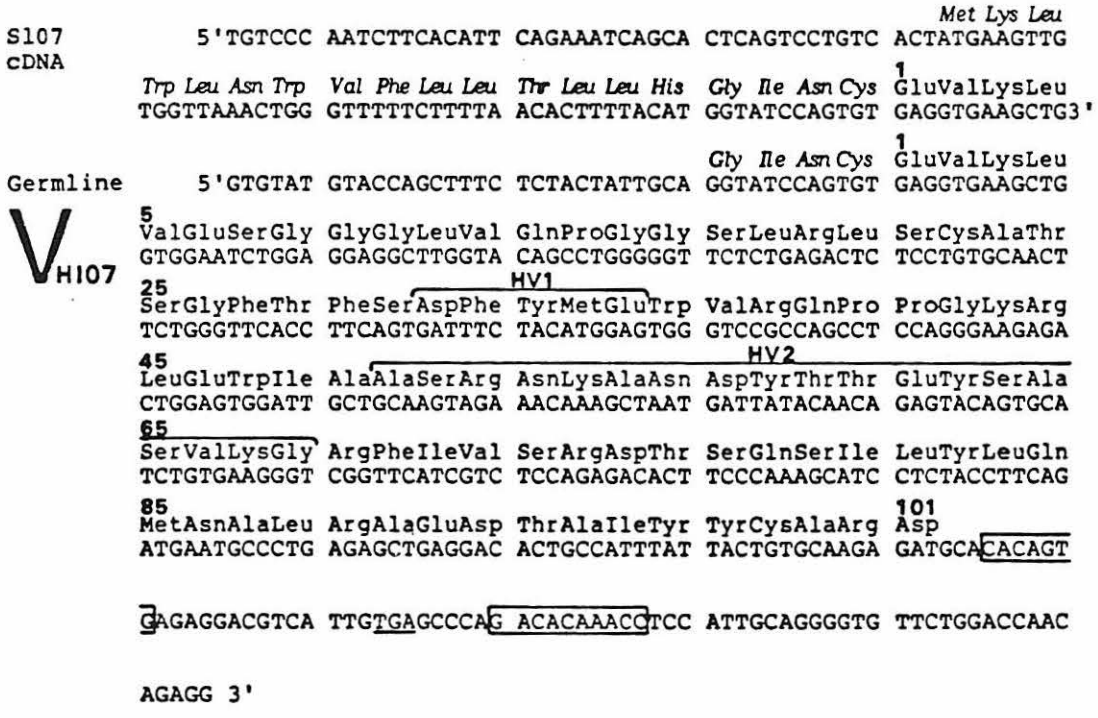
Table 1. Comparison of Noncoding Nucleotides Adjacent to V and J Gene Segments

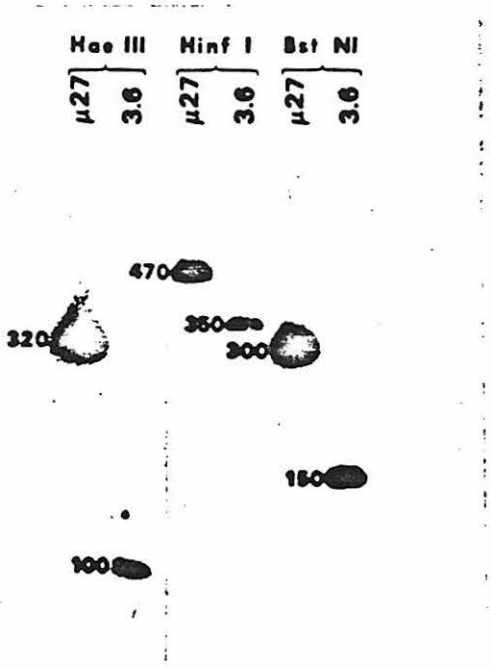
$V_{\kappa 41}$	CACAGTGATACAAATCATAACATAAACC	(11)
$V_{\kappa 2}$	CACAGTGATTCAAGCCATGACATAAACC	(11)
$V_{\kappa 3}$	CACAGTGATTCAAGCCATGACATAAACC	(11)
$V_{\kappa 21}$	CACAGTGCTCAGGGCTGAACAAAAACC	(10)
V_{H107}	CACAGTGAGAGGACGTCATTGTGAGCCCAGACACAAACC	(22)
$V_{\lambda I}$	CACAATGACATGTGTAGATGGGGAAGTAGATCAAGAACA	(22)
$J_{\kappa 1}$	GGTTTTTGTAGAGAGGGGCATGTCATAGTCCTCACTGTG	(22)
$J_{\kappa 2}$	GGTTTTTGTAAAGGGGGGCGCAGTGATATGAATCACTGTG	(23)
$*J_{\kappa 3}$	GGGTTTTGTGGAGGTAAAGTTAAAATAAATCACTGTA	(20)
$J_{\kappa 4}$	AGTTTTTGTATGGGGGTTGAGTGAAGGGACACCAGTGTG	(22)
$J_{\kappa 5}$	GGTTTTTGTACAGCCAGACAGTGGAGTACTACCACTGTG	(22)
J_{H107}	AGTTTTAGTATAGGAACAGAGGCAGAACAGAGACTGTG	(21)
J_{H315}	GGTTTTTGTACACCCACTAAAGGGGTCTATGATAGTGTG	(22)
$J_{\lambda I}$	GGTTTTTGCATGAGTCTATATCACAGTG	(11)

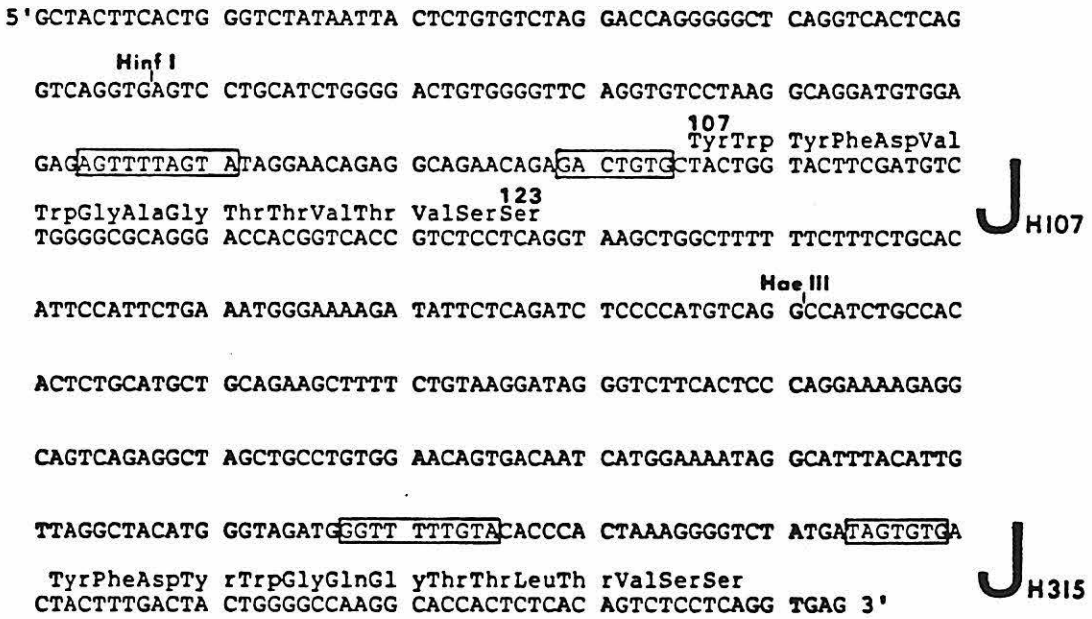
Kappa sequences are from Max et al., 1979 and Sakano et al., 1979; lambda from Bernard et al., 1978; and heavy chains from this paper. The mRNA-sense strand is shown 3' to V gene segments and 5' to J gene segments. The asterisk indicates a possibly nonfunctional J_{κ} gene segment; it has not been found in a rearranged gene. The column to the right lists the number of nucleotides between the underlined hepta- and decanucleotides.


HV1
HV2
HV3
31
35
50
68
100
112
T15, S107, S63, Y5236
H8
M603
W3207
M511
M167
N
G K
N T[]
F
N KYDL V
G
D
D
A
F
V
T
S H R
V
T T
D N G

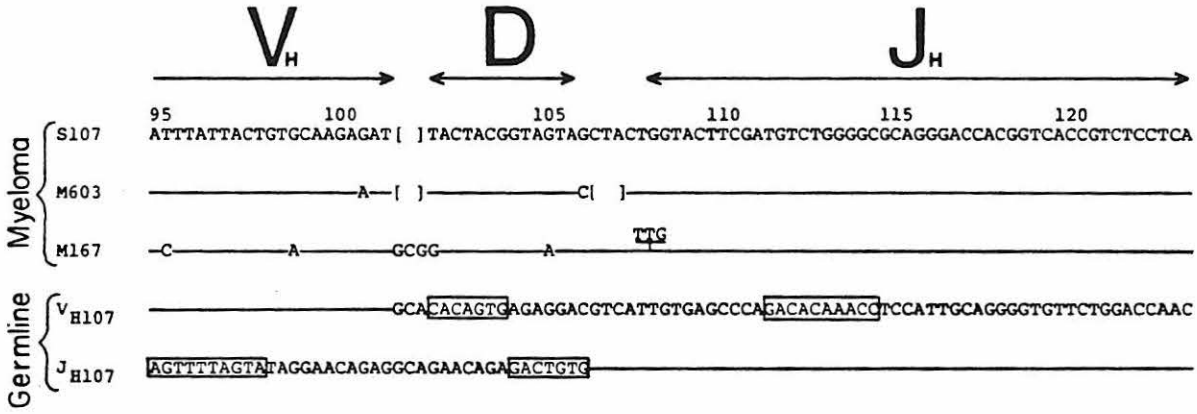

Germline V_H
ChSpV_{pc}3
V_H
RI
RI
HhaI
RI
RI
1kb
V_{H107}
100 bp
MboII
BstNI MboII
HhaI
Germline J_H
ChSpμ27
J_H
C_μ
RI RI
HhaI
RI
HhaI
RI
3.6kb
6.2kb
1kb
220
150
120
J_{H107}
100
150
J_{H315}
350
100 bp
HaeIII
Bst NI
HinfI
HhaI
HaeIII
BstNI
HinfI
M603 V_HC_α
Ch603α6/
Ch603α125
V_HDJ_H
C_α
RI
RI
HhaI
RI
RI
11.3kb
5kb
1kb
V_{H603}-D-J_{H107}
J_{H315}
100 bp
MboII
HinfI
MboII
HhaI

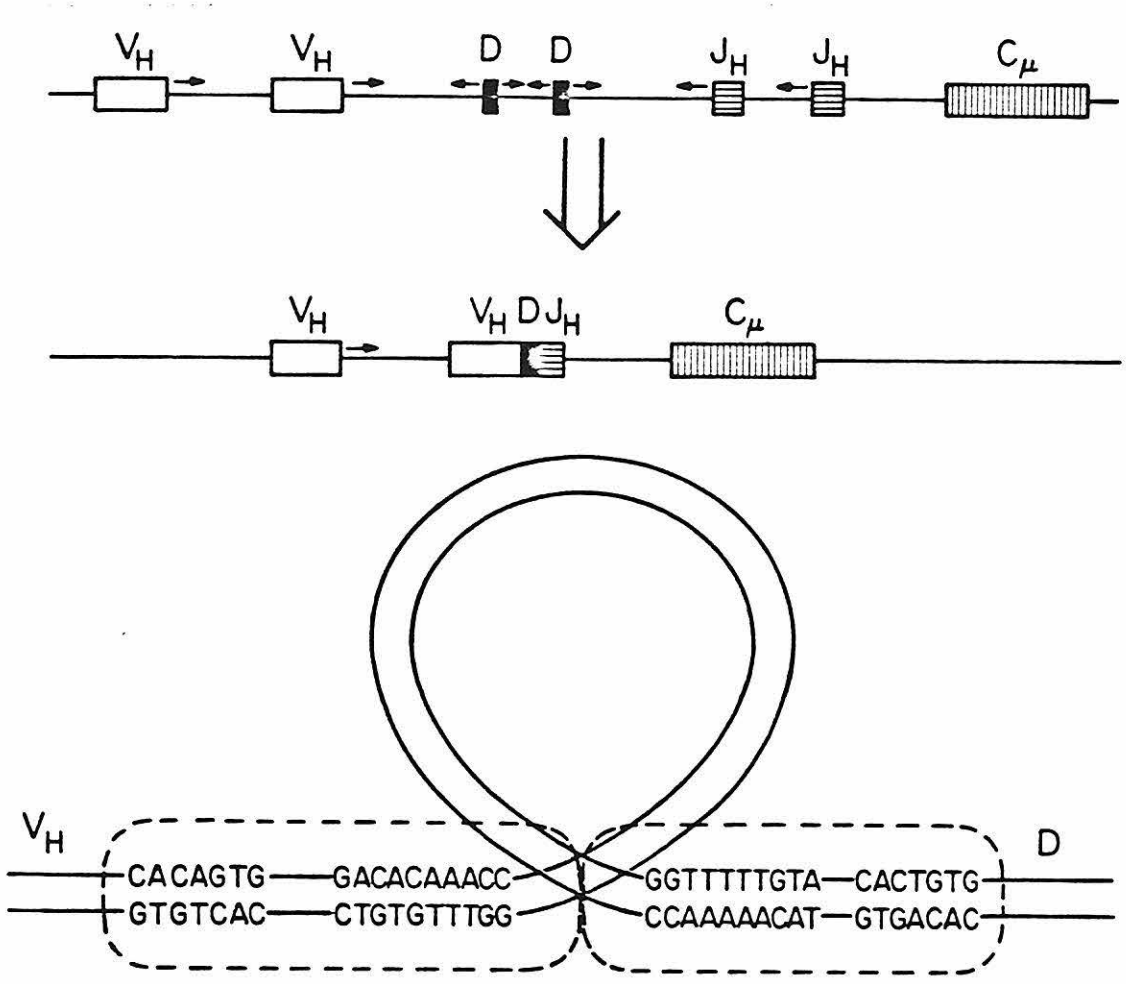

S107 cDNA
5'TGTCCC AATCTTCACATT CAGAAATCAGCA CTCAGTCCTGTC ACTATGAAGTTG
Met Lys Leu
1
Trp Leu Asn Trp Val Phe Leu Leu Thr Leu Leu His Gly Ile Asn Cys GluValLysLeu
TGGTTAAACTGG GTTTTTCTTTTA ACACTTTTACAT GGTATCCAGTGT GAGGTGAAGCTG3'
Germline
5'GTGTAT GTACCAGCTTTC TCTACTATTGCA GGTATCCAGTGT GAGGTGAAGCTG
Gly Ile Asn Cys
1
GluValLysLeu
VH107
5
ValGluSerGly GlyGlyLeuVal GlnProGlyGly SerLeuArgLeu SerCysAlaThr
GTGGAATCTGGA GGAGGCTTGGTA CAGCCTGGGGGT TCTCTGAGACTC TCCTGTGCAACT
HV1
25
SerGlyPheThr PheSerAspPhe TyrMetGluTrp ValArgGlnPro ProGlyLysArg
TCTGGGTTCACC TTCAGTGATTTC TACATGGAGTGG GTCCGCCAGCCT CCAGGGAAGAGA
HV2
45
LeuGluTrpIle AlaAlaSerArg AsnLysAlaAsn AspTyrThrThr GluTyrSerAla
CTGGAGTGGATT GCTGCAAGTAGA AACAAAGCTAAT GATTATACAACA GAGTACAGTGCA
65
SerValLysGly ArgPheIleVal SerArgAspThr SerGlnSerIle LeuTyrLeuGln
TCTGTGAAGGGT CGGTTCATCGTC TCCAGAGACACT TCCCAAAGCATC CTCTACCTTCAG
85
MetAsnAlaLeu ArgAlaGluAsp ThrAlaIleTyr TyrCysAlaArg
101
Asp
ATGAATGCCCTG AGAGCTGAGGAC ACTGCCATTTAT TACTGTGCAAGA GATGCACACAGT
GAGAGGACGTCA TTGTGAGCCCAG ACACAAACCTCC ATTGCAGGGGTG TTCTGGACCAAC
AGAGG 3'

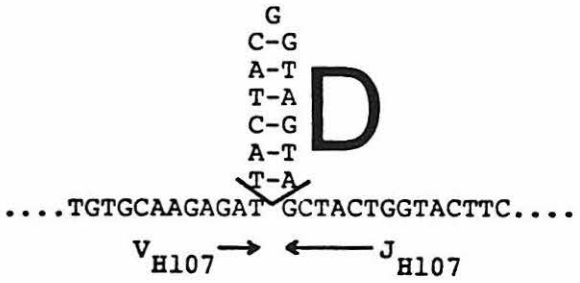

Hae III
Hinf I
Bst NI
μ27
3.6
μ27
3.6
μ27
3.6
470
320
350
300
150
100


5'GCTACTTCACTG GGTCTATAATTA CTCTGTGTCTAG GACCAGGGGGCT CAGGTCACTCAG
Hinf I
GTCAGGTGAGTC CTGCATCTGGGG ACTGTGGGGTTC AGGTGTCCTAAG GCAGGATGTGGA
107
TyrTrp TyrPheAspVal
GAGAGTTTTAGT ATAGGAACAGAG GCAGAACAGAGA CTGTGCTACTGG TACTTCGATGTC
123
TrpGlyAlaGly ThrThrValThr ValSerSer
TGGGGCGCAGGG ACCACGGTCACC GTCTCCTCAGGT AAGCTGGCTTTT TTCTTTCTGCAC
JH107
Hae III
ATTCCATTCTGA AATGGGAAAAGA TATTCTCAGATC TCCCCATGTCAG GCCATCTGCCAC
ACTCTGCATGCT GCAGAAGCTTTT CTGTAAGGATAG GGTCTTCACTCC CAGGAAAAGAGG
CAGTCAGAGGCT AGCTGCCTGTGG AACAGTGACAAT CATGGAAAATAG GCATTTACATTG
TTAGGCTACATG GGTAGATGGGTT TTTGTACACCCA CTAAAGGGGTCT ATGATAGTGTGA
JH315
TyrPheAspTy rTrpGlyGlnGl yThrThrLeuTh rValSerSer
CTACTTTGACTA CTGGGGCCAAGG CACCACTCTCAC AGTCTCCTCAGG TGAG 3'


M603
5'GTGTAT GTACCAGCTTTC TCTACTATTGCA GGTATCCAGTGT GAGGTGAAGCTG
Gly Ile Asn Cys
1
GluValLysLeu
5
ValGluSerGly GlyGlyLeuVal GlnProGlyGly SerLeuArgLeu SerCysAlaThr
GTGGAATCTGGA GGAGGCTTGGTA CAGCCTGGGGGT TCTCTGAGACTC TCCTGTGCAACT
VH
25
HV1
SerGlyPheThr PheSerAspPhe TyrMetGluTrp ValArgGlnPro ProGlyLysArg
TCTGGGTTCACC TTCAGTGATTTC TACATGGAGTGG GTCCGCCAGCCT CCAGGGAAGAGA
45
HV2
LeuGluTrpIle AlaAlaSerArg AsnLysGlyAsn LysTyrThrThr GluTyrSerAla
CTGGAGTGGATT GCTGCAAGTAGA AACAAGGGTAAT AAATATACAACA GAATACAGTGCA
A C G T G
65
SerValLysGly ArgPheIleVal SerArgAspThr SerGlnSerIle LeuTyrLeuGln
TCTGTGAAGGGT CGGTTCATCGTC TCCAGAGACACT TCCCAAAGCATC CTCTACCTTCAG
85
HV3
MetAsnAlaLeu ArgAlaGluAsp ThrAlaIleTyr TyrCysAlaArg AsnTyrTyrGly
ATGAATGCCCTG AGAGCTGAGGAC ACAGCCATTTAT TACTGTGCAAGA AATTACTACGGT
T G
105
122
SerThrTrpTyr PheAspValTrp GlyAlaGlyThr ThrValThrVal SerSer
AGTACCTGGTAC TTCGATGTCTGG GGCGCNGGGACC ACGGTCACCGTC TCCTCAGGTAAG
CTGGCTTTTTTC TTTTTGCACATT CCATTCTGAAAT GGGAAAAGATAT TCTCAGATCTCC
C
CCATGTCAGGCC ATCTGCCACACT CTGCACGCTGCA GAAGCTTTTCTG TAAGGATAGGGT
T
CTTCACTCCCAG GAAAAGAGGCAG TCAGAGGCTAGC TGCCTGTGGAAC AGTGACAATCAT
GGAAAATAGGCA TTTACATTGTTA GGCTACAAGGGT AGATGGGTTTTT GTACACCCACTA
T
Ty rPheAspTyrTr pGlyGlnGlyTh rThrLeuThrVal
AAGGGGTCTATG ATAGTGTGACTA CTTTGACTACTG GGGCCAAGGCAC CACTCTCACAGT
JH315


V H
D
J H
95
100
105
110
115
120
Myeloma
S107
ATTTATTACTGTGCAAGAGAT[ ]TACTACGGTAGTAGCTACTGGTACTTCGATGTCTGGGGCGCAGGGACCACGGTCACCGTCTCCTCA
M603
A—[ ]
C[ ]
M167
C
A
GCGG
A
TTG
Germline
V H107
GCACACAGTGAGAGGACGTCATTGTGAGCCCAGACACAAACCTCCATTGCAGGGGTGTTCTGGACCAAC
J H107
AGTTTTAGTATAGGAACAGAGGCAGAACAGAGACTGTG


V_H
V_H
D
D
J_H
J_H
C_μ
V_H
V_H D J_H
C_μ
V_H
CACAGTG
GACACAAACC
GGTTTTTGTA
CACTGTG
D
GTGTCAC
CTGTGTTTGG
CCAAAAACAT
GTGACAC


G
C-G
A-T
T-A
C-G
A-T
T-A
D
....TGTGCAAGAGAT GCTACTGGTACTTC....
V_{H107}
J_{H107}

Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin μ chain (I)

J. Rogers[+], P. Early[*], C. Carter[+], K. Calame[*], M. Bond[*], L. Hood[*], and R. Wall[+]

[+]Molecular Biology Institute and Department of Microbiology and Immunology, UCLA School of Medicine, Los Angeles, California 90024

[*]Division of Biology, California Institute of Technology, Pasadena, California 91125

Cell (June 1980), in press.

Running title: Immunoglobulin μ mRNAs

Summary

During differentiation, B lymphocytes undergo a shift from expression of membrane-bound IgM to IgM secretion. The μ chains of membrane and secreted IgM, μ_m and μ_s respectively, differ in the amino acid sequence of their carboxy-terminal regions. In this paper, we demonstrate that μ_m and μ_s heavy chains are encoded by separate mRNAs of 2.7 kilobases and 2.4 kilobases, respectively. Restriction mapping and sequence analysis of μ cDNA clones from a myeloma tumor that produces both types of μ chain indicate that the μ_m and μ_s mRNAs are identical throughout the coding region up to the 3' end of the fourth constant region ($C_\mu 4$) domain, but differ in their C-terminal coding and 3' untranslated segments. From the nucleotide sequence of the μ_m cDNA clone, we predict the amino acid sequence of the 41-residue μ_m C-terminal segment or 'M (membrane) segment'. This sequence has characteristics consistent with its being a transmembrane peptide. Thus, the μ_s chain has a 20-residue hydrophilic C-terminal segment after the $C_\mu 4$ domain, and the μ_m chain has a 41-residue C-terminal segment containing a hydrophobic sequence. We propose that comparable C-terminal segments also will be found in other membrane-bound immunoglobulin heavy chains.

Introduction

The immunoglobulin molecule can exist in two very different environments: in the cell membrane as a surface antigen receptor and in solution as a secreted antibody. Immunoglobulin molecules in both environments share the same overall structure. The immunoglobulin molecule is composed of two identical light chains and two identical heavy chains. The light and heavy chains can each be divided into an N-terminal variable (V) region and a C-terminal constant (C) region. The V regions are responsible for antigen binding, whereas the C regions embody the various effector functions of the molecule. The various classes of immunoglobulin with different functions (IgM, IgD, IgG, IgA, IgE) are distinguished by different heavy chains (μ, δ, γ, α, ε) with the differences residing in their C_H regions (C_μ, C_δ, C_γ, C_α, C_ε). Rearrangement of gene segments during B-cell differentiation (Early et al., 1979; Davis et al., 1980; Early et al., 1980a) produces an active immunoglobulin heavy chain gene in which the N-terminal signal sequence (Blobel and Dobberstein, 1975), the V_H region, and the series of homologous domains which constitute the C_H region, are all encoded by separate segments or 'exons'.

The B-cell lineage moves through an orderly series of early developmental stages which are distinguished by the placement of the μ chain. As a stem cell differentiates to a pre-B cell, μ heavy chain (but not light chain) is expressed in the cytoplasm (Burrows et al., 1979). As the pre-B cell changes to a B cell, light chain is expressed and IgM monomers (μ_2L_2) are displayed as cell-surface antigen receptors (Vitetta et al., 1971). Subsequently, the B cell can differentiate to become a plasma cell which actively secretes IgM molecules. These are secreted in pentameric form by virtue of disulphide linkages between the μ chains and a J or joining chain to generate the pentameric $(\mu_2L_2)_5J$ molecule (Koshland,

1975; Della Corte and Parkhouse, 1973). Thus, at different stages of development the B cell expresses predominantly membrane IgM or secreted IgM molecules.

There are tumors of the B-cell lineage corresponding to each of these developmental stages, which permit ready analysis of the immunoglobulin gene expression. The B cell is represented by B-cell lymphomas, and the plasma cell by plasmacytomas or myeloma tumors. Moreover, a single cell lineage can be analyzed in both states because a B-cell lymphoma carrying membrane IgM, upon cell fusion with an appropriate myeloma cell line to form a hybridoma, is induced to secrete its IgM in the pentameric form (Raschke et al., 1979; Eshhar et al., 1979; Laskov et al., 1979).

Several investigators, beginning with Melcher and Uhr (1973), have demonstrated that the μ chains of membrane IgM (μ_m) and secreted IgM (μ_s) differ in structure. The μ_m chain is larger than the μ_s chain (Bergman and Haimovich, 1978; Vassalli et al., 1979; Sibley et al., 1980; A. Williamson, personal communication) and has hydrophobic properties not exhibited by the μ_s chain (Parkhouse et al., 1979; Vassalli et al., 1979). The μ_m and μ_s chains have identical N-terminal sequences and appear by comparative peptide map analyses to be identical throughout most of their V and C_μ regions (Yuan et al., 1979; Kehry et al., 1980). Analyses with carboxypeptidase suggest that the μ_m and μ_s chains have distinct C-terminal sequences (Williams et al., 1978; Kehry et al., 1980). Kehry et al. (1980) have demonstrated that the C-terminus of the μ_m chain is distinct in amino acid sequence from its μ_s counterpart. Moreover, the μ_s-μ_m differences appear to reside in the C-terminal segment which in the μ_s chain extends for 20 residues beyond the $C_\mu 4$ domain. Taken together, these data suggest that the μ_m chain has a distinct, possibly hydrophobic, C-terminal segment longer than that of the μ_s chain.

Four possible models could be advanced to account for the differences between μ_m and μ_s chains. (i) The μ_m chain may be synthesized with a C-terminal hydrophobic

sequence which is removed by proteolytic cleavage to generate a shorter μ_s chain with a hydrophilic C-terminus. We have eliminated this model by showing that the μ_s mRNA contains a stop codon precisely at the end of the coding sequence of μ_s (Calame et al., 1980). (ii) The C_μ gene segment may have its 3' coding region rearranged during B-cell differentiation, in a manner analogous to the somatic joining of sequences which encode the V and J (joining) gene segments (Bernard et al., 1978; Seidman et al., 1979; Early et al., 1980a). This was ruled out for the μ_s chain by the demonstration that the μ_s C-terminal segment is already encoded in the germline DNA in contiguous juxtaposition to the $C_\mu 4$ domain (Calame et al., 1980). (iii) The μ_m and μ_s chains may be encoded by two different C_μ genes. This model appears unlikely because the C_μ gene segment is present in a single copy per haploid complement of germline DNA (Cory and Adams, 1980; Calame et al., 1980; Early et al., 1980b). (iv) The μ_m and μ_s chains may be generated by alternative RNA splicing events from a larger nuclear transcript. This is the correct model, as we demonstrate in this and a companion paper (Early et al., 1980b).

We demonstrate in this paper that B cells and plasma cells have two distinct μ mRNAs. One codes for the μ_s chain. The other has a different C-terminal coding sequence which appears to be generated by RNA splicing at the end of the $C_\mu 4$ domain. We present evidence that this mRNA codes for the μ_m chain. In the following paper (Early et al., 1980b), we analyze the structure of a C_μ genomic clone, including the coding region for the μ_m C-terminal segment, and propose a mechanism for developmentally regulated expression of μ_m and μ_s chains by RNA processing.

Results

Cell Lines Producing μ Chains

In this study, we analyzed μ mRNAs from mouse tumor cell lines which correspond to various stages in B-cell development. Two mouse B-cell lymphomas, WEHI 279 (W279) and WEHI 231 (W231), correspond to early B cells in that both produce predominantly membrane IgM (Warner et al., 1979). However, a majority of the cytoplasmic μ chains in these cells have the structure of μ_s chains, and a small amount of pentameric IgM is secreted (Kehry et al., 1980). Upon cell fusion of W279 or W231 with the mouse myeloma cell line MPC 11, membrane IgM production ceases and pentameric IgM molecules are secreted (Raschke et al., 1979; W. Raschke, personal communication). The parental line of MPC 11 secretes IgG2b and contains J chains, while the resulting hybridomas secrete both IgG2b and pentameric IgM with J chains. Thus, the fusion of W279 or W231 with MPC 11 appears to switch off the production of μ_m chains and stimulate the secretion of μ_s chains. Accordingly, these hybridomas correspond to terminally differentiated plasma cells.

The mouse myeloma tumor MOPC 104E (M104E) secretes pentameric IgM molecules and thus also represents the plasma cell stage. However, as with certain other IgM-secreting myeloma tumors, M104E also produces some membrane IgM molecules (Anderson et al., 1974; E. Vitetta, personal communication).

R-loop Mapping Suggests that M104E Cells Contain Two Classes of μ mRNA

Heavy chain mRNA was isolated from the cytoplasmic RNA of M104E myeloma cells by oligo(dT) and size selection. This mRNA was analyzed by R-loop formation with cloned C_μ gene segments. Three C_μ gene segments were available (Calame et al., 1980), two derived from a genomic mouse sperm library (ChSpμ27 and ChSpμ7) and the third derived from a genomic library of mouse myeloma McPc603 tumor

(Ch603μ35). The three cloned DNA fragments differ in size, but are identical in and around the C_μ gene segment. Similar results were obtained for each clone.

Eighty to eighty-five percent of the molecules containing R-loops exhibited from two to four R-loops corresponding to the four C_μ domains, as reported previously (Calame et al., 1980). In addition, about 10-15% of these molecules contained another R-loop that was 0.38 ± 0.1 kb in size and was located 1.7 ± 0.2 kb 3' to the $C_\mu 4$ domain (Figures 1 and 2). The region of sequence corresponding to this R-loop will be denoted the M exon, for reasons which will become apparent below.

Although more than 25 R-loop molecules containing mRNA hybridized to the M exon were examined, we did not observe any in which the same RNA molecule had hybridized to all five exons. All 25 molecules had the structure shown in Figure 1. A collapsed segment of unhybridized RNA was observed on the 5' side of the M exon and another RNA molecule had hybridized to the exons encoding the four C_μ domains. This result is consistent with the possibility that the second type of mRNA encodes the four C_μ domains as well as the M exon, but that it was present in low amounts relative to the major species of μ mRNA. If so, the major μ species would saturate the four C_μ domains and the minor species would hybridize only to the M exon, producing the observed result. Thus, M104E cells may contain two classes of μ mRNA. The less abundant μ mRNA apparently has an additional exon at its 3' end.

M104E Cells Contain Two Classes of μ mRNA with Different 3' Ends

To obtain independent evidence for the two putative μ mRNA species, we fractionated M104E mRNA on a methylmercury hydroxide/agarose gel and transferred it to diazophenylthioether paper strips. These 'RNA blots' were probed by hybridization with ^{32}P-labelled cDNA plasmid p104Eμ12, which contains the μ_s coding sequence from the middle of the $C_\mu 2$ domain to the 3' untranslated ($3'UT_s$) segment (Figure 2

and Calame et al., 1980). This probe hybridized with two mRNA bands: a strong 2.4 kb band and a very weak 2.7 kb band (Figure 3).

As a specific probe for the 3' end of the μ_s sequence, we used HhaI fragment A from a subclone of a germline C_μ gene clone (Figure 2). This fragment contains most of the 3'UT_s segment (Early et al., 1980b). Fragment A hybridized to the 2.4 kb band but not to the 2.7 kb band (Figure 3). As a probe for the M exon observed in R-loop analysis, we used PstI fragment B from the same C_μ gene clone (Figure 2). This hybridized only to the 2.7 kb band (Figure 3).

Since p104Eμ12 hybridizes to both mRNAs while the 3'UT_s probe hybridizes only to the 2.4 kb species, we conclude that M104E contains two μ mRNAs, both of which contain at least some of the coding regions for the four C_μ domains. The two species differ in their 3' ends. The following μ mRNA structures can be proposed. The major, 2.4 kb species (μ_s) would encode the V region, the four C_μ domains, and the μ_s C-terminal segment, and would end with the 3'UT_s segment. The minor, 2.7 kb species would encode the same V region and four C_μ domains, and in addition would contain sequences from the M exon. It does not contain the 3'UT_s segment which lies between the $C_\mu 4$ domain and the M exon in the germline C_μ gene segment. This 2.7 kb species is the candidate for the μ_m mRNA.

The 2.7 kb mRNA Probably Encodes μ_m Chains

The B-cell lymphoma W279 expresses predominantly μ_m chains, although it makes some μ_s chains. The hybridoma MPC11xW279.2 produces only μ_s chains. The mRNA from W279 and from MPC11xW279.2 was analyzed by RNA blots (Figure 4), using the three μ gene probes used in Figure 3. The 2.4 kb band, which hybridized to the 3'UT_s but not to the M exon, was present in both cell lines. However, the 2.7 kb band, which hybridized to the M exon but not to 3'UT_s, appeared only in W279. The 2.7 kb band and the 2.4 kb band were also seen in the B-cell lymphoma

W231 (not shown), as has been reported by Perry and Kelley (1979), who suggested that they correspond to μ_m and μ_s mRNAs, respectively. Therefore, the 2.7 kb band is strong in W279 and W231, which produce predominantly μ_m chains; weak in M104E, which produces relatively few μ_m chains; and absent in the MPC11xW279.2 hybridoma, which produces no μ_m chains. This correlation constitutes convincing evidence that the 2.7 kb species is the mRNA for the μ_m chain.

cDNA Clones from the Two μ mRNAs

We have prepared μ cDNA clones from the mRNA of M104E cells (Calame et al., 1980). One clone, p104Eμ12 (μ12), was shown to encode the μ_s chain from the $C_\mu 2$ domain to the C-terminus. In view of the hybridization results presented above, this clone presumably represents the 2.4 kb μ_s mRNA, which is the major μ mRNA in the M104E tumor line. The DNA sequence of the μ12 clone from the $C_\mu 4$ domain to the end of the clone is presented in Figure 5. These cDNAs were generated by priming reverse transcriptase with oligo(dT) hybridized to the mRNA, followed by second-strand synthesis with DNA polymerase, and were completed by adding oligo(dC) tails to the double-stranded cDNA. Therefore, the $(dA)_{15}$ segment in the μ12 clone is likely to be derived from the poly(A) tail of the messenger RNA.

The μ12 clone has a stop codon immediately following codon 576, the C-terminus of the μ_s chain. There are 128 nucleotides comprising the $3'UT_s$ segment. This segment is rich in A and T; 33 out of the last 71 nucleotides before the poly(A) tail are A. This segment ends with $AATAAA(N)_{17}TC$-poly(A). The AATAAA sequence was first identified as a common sequence in 3' untranslated segments by Proudfoot and Brownlee (1976), and it has been found in all subsequent DNA sequences which precede messenger polyadenylation sites. Fitzgerald and Shenk (1979) have provided direct evidence that in the late RNAs of SV40, AATAAA

signals polyadenylation at a CA sequence some 13 nucleotides in the 3' direction. The 3' end of the μ_s mRNA is apparently typical with regard to this polyadenylation signal.

A second cDNA clone derived from M104E mRNAs, p104Eμ6 (μ6), was found by restriction mapping to share the $C_\mu 3$ and $C_\mu 4$ domains with the μ12 cDNA clone, but to have a different 3' portion (Figure 6). In the following paper, we demonstrate that the 3' portion of the μ6 cDNA clone represents the M exon of the genomic C_μ clone (Early et al., 1980b). Accordingly, μ6 was probably derived from the 2.7 kb μ_m mRNA. The possibility that the μ6 clone might be an aberration of cloning was ruled out by the independent isolation of another M104E cDNA clone, p104Eμ4, with an identical restriction map from the $C_\mu 3$ domain to the 3' terminus (data not shown). The p104Eμ4 and μ6 cDNA clones were the only two, out of 15 independent M104E μ cDNA clones analyzed, which contained this 3' segment.

The DNA sequence of the μ6 clone from the $C_\mu 4$ domain to the end of the clone is presented in Figure 6. Like the μ12 clone, the μ6 cDNA clone ends with an $AATAAA(N)_{17}TC$-poly(A) sequence, indicating that it too contains the 3' end of the corresponding mRNA sequence. The divergence of the μ6 and μ12 clones begins after codon 556 at the end of the $C_\mu 4$ domain. Where the μ12 clone has 20 C-terminal codons followed by 128 nucleotides of 3' UT_s sequence, μ6 has 41 C-terminal codons followed by 270 nucleotides of a different 3'UT sequence. This difference makes the μ6 mRNA 205 nucleotides longer than the μ12 mRNA, assuming that the poly(A) tracts and the sequences preceding the $C_\mu 4$ domain are identical in the two species. This is consistent with the observed sizes of 2.7 kb and 2.4 kb for the μ_m and μ_s mRNAs, respectively.

Discussion

The Predicted μ_m Polypeptide Sequence

The mRNAs coding for the μ_m and μ_s chains appear to be identical up to the 3' end of the coding region for the $C_\mu 4$ domain at codon 556. We can predict the amino acid sequence at the C-terminus of the μ_m chain from the μ6 DNA sequence (Figure 6). The μ_s chain has a hydrophilic C-terminal segment of 20 residues after the $C_\mu 4$ domain, which is required for μ_s secretion, whereas the μ_m chain has a C-terminal segment of 41 residues which we will denote the M (membrane) segment (Figure 7).

Structural studies on μ_m chains are in general agreement with the predicted properties of the M segment. (i) Mouse and human μ_m chains, synthesized in the presence of tunicamycin to block glycosylation, are 20-30 amino acids longer than similarly nonglycosylated μ_s chains (Vassalli et al., 1979; Sibley et al., 1980; A. Williamson, personal communication). (ii) The C-terminus of human μ_m chains does not contain a tyrosine residue which can be released by carboxypeptidase A (Williams et al., 1978). The μ_s chains do have C-terminal tyrosine (Figure 7). (iii) In the third paper of this series (Kehry et al., 1980), we present detailed studies using peptide map, amino acid sequence, and carboxypeptidase analyses, which strongly suggest that the M segment in Figure 7 is the mouse μ_m C-terminal sequence.

The M Segment Has the Properties of a Transmembrane Peptide

The predicted amino acid sequence of the M segment suggests that a membrane-bound IgM molecule is an integral membrane protein and that the M segment is the transmembrane portion of the μ_m polypeptide (Figure 8). The M segment may be divided into three regions. The first 12 amino acids include six glutamic acid residues and, accordingly, form a negatively charged region. The following 26 amino acids are all uncharged and include a very hydrophobic 11-residue core,

surrounded by two uncharged but hydrophilic regions containing serine and threonine residues. The last three C-terminal amino acids include two lysine residues, which form a short positively charged tail, presumably on the cytoplasmic side of the membrane. The M segment bears a striking resemblance to certain other transmembrane peptides. For example, the red blood cell protein glycophorin (Tomita and Marchesi, 1975) and the coat proteins of filamentous bacteriophages (Nakashima and Konigsberg, 1974; Snell and Offord, 1972) have acidic residues just outside the cell and basic residues just inside the cytoplasm, separated by a transmembrane portion of 18-23 uncharged amino acids. We have estimated the hydrophobicity of the proposed transmembrane portion of the M segment according to the method of Segrest and Feldman (1974), and it has the hydrophobicity characteristic of transmembrane peptides (Kehry et al., 1980).

An α helix requires about 24 uncharged residues to span the lipid bilayer and the extended β configuration about 11 residues. The M segment has a very hydrophobic core of 11 residues. However, the extended β configuration normally requires hydrogen bonding from a second polypeptide chain for stability, so we feel that an α helical configuration for the entire 26-residue uncharged sequence, as suggested in Figure 8, is the most likely arrangement for the M segment in the cell membrane.

Two Possible Functions for the M Segment

The M segment may serve as a membrane anchor during the synthesis of μ_m chains. As nascent μ_m chains are extruded into the endoplasmic reticulum, the hydrophobicity of the transmembrane portion of the M segment, as well as the presence of charged amino acids on both sides of the membrane, should serve to attach the μ_m chains firmly to the membrane.

The IgM molecule also may serve as a transmembrane signaling device to stimulate B-cell proliferation and differentiation upon binding of antigen. The cytoplasmic 'tail' of the M segment is short (three residues) but very basic (Figure 8). Perhaps these positively charged residues interact with elements of the cytoskeleton or other negatively charged cytoplasmic elements. Upon antigen binding, these interactions might be perturbed by cross-linking to other receptor IgM molecules. For example, when surface immunoglobulin molecules are cross-linked by antigen binding or other means, they aggregate into patches and become attached, either directly or indirectly, to submembrane actin (Flanagan and Koch, 1978). It is possible that the basic 'tail' of the μ_m chain may contribute to this attachment, and thus trigger the subsequent internalization of the cross-linked surface IgM.

The μ_s mRNA Has a Potential RNA Splicing Site at the End of the $C_\mu 4$ Domain

Previously, we have shown that the C_μ gene segment encodes the four C_μ domains in four separate exons bounded by RNA splicing sites (Calame et al., 1980). An RNA splicing site also appears to be present at codons 557-558 at the 3' end of the $C_\mu 4$ domain, where the μ_m (μ6) and μ_s (μ12) cDNA sequences diverge. The μ6 clone has the sequence TG/AGGGGG, while the μ12 clone has the sequence TG/GTAAAC. This sequence in the μ12 clone is identical with that of the chromosomal gene (Calame et al; 1980; Early et al., 1980b). It clearly resembles the consensus sequence for an 'upstream' RNA splicing site, AG/GTAAGT (Seif et al., 1979; Lerner et al., 1980; Rogers and Wall, 1980), where the stroke divides the exon from the intron. In the following paper (Early et al., 1980b), we show that this sequence in the C_μ gene segment is the 5' end of an intron which is spliced out to join the $C_\mu 4$ domain and the M coding segment.

This splice point can be used to define precisely the 3' end of the $C_\mu 4$ domain, just as other splice points have defined the junctions between the other immunoglobulin

domains (Bernard et al., 1978; Seidman et al., 1978; Sakano et al., 1979; Tucker et al., 1979; Calame et al., 1980). The $C_\mu 4$ splice point occurs in the same phase with respect to the coding frame as the other inter-domain splice points.

All Heavy Chain Genes May Have the Same RNA Splicing Sequence for Joining C_H and M Exons

All immunoglobulin heavy chains that have been sequenced at the DNA or protein level, other than the μ_m chain, are derived from secreted immunoglobulins. However, they all have a glycine-lysine sequence precisely at the end of the last domain, which could be encoded by the same RNA splicing site that joins the $C_\mu 4$ exon to the M exon (Table 1). In genomic $\gamma 1$ (Honjo et al., 1979) and $\gamma 2b$ (Tucker et al., 1979) clones, as in the μ_s clone (Calame et al., 1980; Early et al., 1980b), this dipeptide has been shown to be encoded by the same sequence, GGTAAA, which forms the upstream splice site at the 3' end of the $C_\mu 4$ exon. (The final lysine is absent from mature γ chains, presumably due to posttranslational proteolysis.) The δ chains are predominantly membrane-bound, and membrane forms of γ and α chains have been observed in some cell lines (Vassalli et al., 1979; M. Kuehl and C. Word, personal communication; R. Lynch, personal communication). Moreover, these δ and γ molecules have properties suggesting that they contain hydrophobic segments (Vassalli et al., 1979), which could be homologous to the M segment described for μ_m chains. Thus, it appears likely that most, if not all, immunoglobulin heavy chains will have this potential RNA splice site, which presumably permits each C_H coding region to be associated with an M sequence and, accordingly, to generate a membrane-bound receptor immunoglobulin.

Evolution of Immunoglobulin Gene Segments

An active immunoglobulin gene consists of a series of exons, encoding the signal peptide, the V domain, and from one to four C domains. The RNA splice sites

between the exons are all in the same phase with respect to the coding frame. Thus, the V and C coding segments all appear to represent duplications of a single ancestral exon (Sakano et al., 1979). With the discovery of an additional RNA splice from the $C_\mu 4$ exon to the M exon, also in the same coding phase, it appears that the exon encoding the ancestral domain may already have been flanked by RNA splice sites before any duplications occurred. Two findings support the idea that the $C_\mu 4$-M splice represents an ancient feature of immunoglobulin genes. (i) Both μ_m and μ_s chains appear to be present in sharks, the most primitive species studied to date (Marchalonis, 1977). (ii) The other heavy chain classes may perform the same RNA splice from the 3' terminal C_H exon to the M exon, as described above.

Therefore, we propose the following scheme for the origin of the immunoglobulin genes. The primordial immunoglobulin polypeptide consisted of three segments: a signal peptide (P), which caused the nascent polypeptide to enter the endoplasmic reticulum; a single immunoglobulin-like domain (Ig); and a transmembrane peptide (M), which anchored this primitive receptor protein to the cell surface. Correspondingly, the primordial immunoglobulin gene consisted of three exons, P, Ig, and M. The RNA splice points between the P-Ig and Ig-M exons were fortuitously in the same coding phase. Thus, duplication of the central Ig exon with some flanking sequences could occur within the transcription unit and the duplicated copies would be joined in tandem by RNA splicing. We believe that the V and C region domains of immunoglobulins evolved from the initial duplication of the Ig exon. The light and heavy chain genes then evolved separately from duplications of the entire transcription unit. During subsequent evolution, the light chains evidently lost the M segment, since they are never integral membrane proteins. The ancestral heavy chain gene underwent further duplications of the

C exon within the same transcription unit, to generate a C_H gene segment coding for four C_H domains. This enlarged transcription unit retained the separate M exon at the 3' end.

This process of gene expansion by internal duplication of exons is probably common in evolution. Preliminary restriction maps have been reported for three other genes whose polypeptide products show evidence of multiple internal duplications: conalbumin or ovotransferrin (MacGillivray et al., 1977; Cochet et al., 1979), serum albumin (McLachlan and Walker, 1977; Sargent et al., 1979), and collagen (Tolstoshev et al., 1980). All three genes were found to consist of numerous (>13) exons.

If other integral membrane proteins evolved from the primordial immunoglobulin gene, they also may contain M segments. The T-cell differentiation antigen Thy-1 (Campbell et al., 1979) and the heavy chains of the major histocompatability antigens (Orr et al., 1979a and b) show some homology with immunoglobulins. The T-cell antigen receptor is an additional candidate for homology relationships with immunoglobulins (Binz and Wigzell, 1975; Rajewsky and Eichmann, 1977; Eichmann, 1978). Although we would not necessarily expect M segments on these proteins to retain detectable sequence homology with the M segment of the μ_m chain, they could be identified as C-terminal transmembrane polypeptide segments encoded by separate exons.

Experimental Procedures

Growth of Cells

M104E was grown as a solid tumor in BALB/c mice. W231, W279, and the hybridoma MPC11xW279.2 were grown in tissue culture in Dulbecco's modified Eagle's medium plus penicillin, streptomycin, 20% fetal calf serum, 4 mM glutamine, and 50 μM β-mercaptoethanol.

R-loop Mapping and 'RNA Blots'

Messenger RNA was isolated from M104E tumors as described in Wall et al. (1977) and from the tissue culture lines as described in Gilmore-Hebert and Wall (1979). Heavy chain mRNA for R-loop formation was poly(A)-selected on an oligo(dT)-cellulose column and size-selected on a sucrose gradient. R-looping was performed as in Thomas et al. (1976) as modified by Kaback et al. (1979).

mRNA for 'RNA blots' (Alwine et al., 1977) was selected on an oligo(dT)-cellulose column. It was electrophoresed on a 1% agarose gel containing 5 mM methylmercury hydroxide in buffer E (Alwine et al., 1977). The gel was washed at room temperature for 1 hr in 50 mM NaOH/5 mM β-mercaptoethanol, then for 3 x 20 min in 140 mM sodium citrate-phosphate buffer, pH 4.0. The second of the citrate-phosphate washes contained 1 μg/ml ethidium bromide and the gel was then photographed under UV illumination. The third citrate-phosphate wash contained 7 mM sodium iodoacetate. The gel was then washed for 2 x 30 min in 20 mM sodium citrate-phosphate buffer, pH 4.0, and transferred to diazophenylthioether paper in the same buffer as in Alwine et al. (1977).

Aminophenylthioether paper was prepared by an unpublished procedure of B. Seed. It was stored sealed at -20°C and activated to make diazophenylthioether paper immediately before transfer. Following transfer, the paper was pre-hybridized and hybridized as in Wahl et al. (1979). DNA plasmids and fragments for hybridization were labelled by nick translation (Rigby et al., 1977) with α^{32}P-dCTP (Amersham) to a specific activity of 4-10 x 10^7 cpm/μg. Hybridized paper was washed in 2 x SSC/0.1% SDS for 3 x 5 min at 25°C then in 0.1 x SSC/0.1% SDS for 3 x 30 min at 50°C. It was then autoradiographed at -70°C on Kodak XRP-5 film with a Dupont Lightning-Plus intensification screen.

Growth of Recombinant DNA

Recombinant phage were grown in E. coli DP50supF, and plasmid pSpμ A1, a subclone from ChSpμ 7, was grown in E. coli χ1776. These chromosomal gene clones were under P2 EK2 containment. Recombinant plasmids containing cDNA were grown in E. coli HB101 under P2 EK1 containment. All procedures were in accordance with the NIH Guidelines on Recombinant DNA.

Restriction Mapping

The restriction map of plasmid p104Eμ6 (μ6) was established by means of parallel digests of pBR322, μ6, and isolated PstI fragments of μ6. These digests were performed with each enzyme singly and, where necessary, in combination with XbaI, MspI (HpaII), and other enzymes. Several close pairs of sites were resolved only by DNA sequencing. HpaII sites were located using the isoschizomer MspI. MspI and MboII were purchased from New England Biolabs and used as suggested by the supplier. Other enzymes were made by M. Komaromy and used in standard buffers. Digestion products were analyzed on 6% polyacrylamide gels. The digests of pBR322 served as size markers (Sutcliffe, 1978).

DNA Sequencing

DNA fragments were eluted from 6% polyacrylamide gels as in Maxam and Gilbert (1977), except that incubation was for two days at 42°C. γ^{32}P-ATP synthesis, phosphatase treatment, 5' end-labelling, and cleavage or strand separation were performed as in Maxam and Gilbert (1980). ^{32}P (carrier-free) was purchased from ICN Pharmaceuticals, bacterial alkaline phosphatase (BAPF) from Worthington, and T4 polynucleotide kinase from Boehringer-Mannheim or PL Biochemicals. Sequencing was performed according to Maxam and Gilbert (1977), using the G (alternative), A > C, T+C, and C reactions. Samples were dissolved in urea-dye

mixture (without NaOH) and electrophoresed on 8% and 20% polyacrylamide-urea gels (Maxam and Gilbert, 1980) of thickness 0.32 mm (Sanger and Coulson, 1978). Gels were autoradiographed at -70°C on Dupont Cronex 4 film, with a Dupont Lightning-Plus intensification screen when necessary.

Acknowledgements

We are grateful to W. Raschke for providing the cell lines and unpublished data on their characteristics. We also thank B. Seed for communicating a method for making diazophenylthioether paper, and M. Komaromy for providing restriction enzymes. This work was supported by NIH grants AI 13410 and CA 12800 to R. W., and NIH grant AI 16590 to L. H. P. E. is supported by NIH training grant GM 07616. K. C. is supported by NIH fellowship GM 05442.

References

Alwine, J. C., Kemp, D. J., and Stark, G. P. (1977). Proc. Nat. Acad. Sci. USA 74, 5350-5354.

Anderson, J., Buxbaum, J., Citronbaum, R., Douglas, S., Forui, L., Melchers, F., Peruis, B., and Stott, D. (1974). J. Exp. Med. 140, 742-763.

Bergman, Y., and Haimovich, J. (1978). Eur. J. Immunol. 8, 876-880.

Bernard, O., Hozumi, N., and Tonegawa, S. (1978). Cell 15, 1133-1144.

Binz, H., and Wigzell, H. (1975). J. Exp. Med. 142, 197-211.

Blobel, G., and Dobberstein, B. (1975). J. Cell Biol. 67, 835-851.

Burrows, P., LeJeune, M., and Kearney, J. F. (1979). Nature 280, 838-840.

Calame, K., Rogers, J., Early, P., Davis, M., Livant, D., Wall, R., and Hood, L. (1980). Nature, in press.

Campbell, D. G., Williams, A. F., Bayley, P. M., and Reid, K. B. M. (1979). Nature 282, 341-342.

Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F., and Chambon, P. (1979). Nature 282, 567-574.

Cory, S., and Adams, J. M. (1980). Cell 19, 37-51.

Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L., and Hood, L. (1980). Nature 283, 733-739.

Della Corte, E., and Parkhouse, R. M. E. (1973). Biochem. J. 136, 597-606.

Early, P. W., Davis, M. M., Kaback, D. B., Davidson, N., and Hood, L. (1979). Proc. Nat. Acad. Sci. USA 76, 857-861.

Early, P., Huang, H., Davis, M., Calame, K., and Hood, L. (1980a). Cell 19, 981-992.

Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980b). Cell, submitted.

Eichmann, K. (1978). Adv. in Immunol. 26, 195-254.

Eshhar, Z., Blatt, C., Bergman, Y., and Haimovich, J. (1979). J. Immunol. 122, 2430-2434.

Fitzgerald, M., and Shenk, T. (1979). Ann. N.Y. Acad. Sci. (Genetic Variation of Viruses), in press.

Flanagan, J., and Koch, G. L. E. (1978). Nature 273, 278-281.

Gilmore-Hebert, M., and Wall, R. (1979). J. Mol. Biol. 135, 879-891.

Honjo, T., Obata, M., Yamawaki-Kataoka, Y., Kataoka, T., Kawakami, T., Takahashi, N., and Mano, Y. (1979). Cell 18, 558-568.

Kaback, D. B., Angerer, L. M., and Davidson, N. (1979). Nucleic Acids Res. 6, 2499-2517.

Kehry, M., Sibley, C., Fuhrman, J., Schilling, J., and Hood, L. E. (1979). Proc. Nat. Acad. Sci. USA 76, 2932-2936.

Kehry, M., Ewald, S., Douglas, R., Sibley, C., Raschke, W., Fambrough, D., and Hood, L. (1980). Cell, submitted.

Koshland, M. E. (1975). Adv. in Immunol. 20, 41-69.

Laskov, R., Kim, K. J., and Asofsky, R. (1979). Proc. Nat. Acad. Sci. USA 76, 915-919.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980). Nature 283, 220-224.

Low, T. L. K., Liu, Y.-S. V., and Putnam, F. W. (1976). Science 191, 390-392.

MacGillivray, R. T. A., Mendez, E., and Brew, K. (1977). Proteins of Iron Metabolism, E. B. Brown et al., eds. (New York: Grune and Stratton), pp. 133-141.

Marchalonis, J. J. (1977). Immunity in Evolution (Cambridge, Mass. Harvard University Press).

Maxam, A. M., and Gilbert, W. (1977). Proc. Nat. Acad. Sci. USA 74, 560-564.

Maxam, A. M., and Gilbert, W. (1980). Methods in Enzymology 65, L. Grossman and K. Moldave, eds. (New York: Academic Press) in press.

McLachlan, A. D., and Walker, J. E. (1977). J. Mol. Biol. 112, 543-558.

Melcher, U., and Uhr, J. W. (1973). J. Exp. Med. 138, 1282-1287.

Mestecky, J., and Schrohenloher, R. E. (1974). Nature 249, 650-652.

Nakashima, Y., and Konigsberg, W. (1974). J. Mol. Biol. 88, 598-600.

Orr, T. H., Lancet, D., Robb, R. J., Lopez de Castro, J. A., and Strominger, J. L. (1979a). Nature 282, 266-270.

Orr, T. H., Lopez de Castro, J. A., Parham, P., Ploegh, H. L., and Strominger, J. L. (1979b). Proc. Nat. Acad. Sci. USA 76, 4395-4399.

Parkhouse, R. M. E., Lifter, J. L., and Choi, Y. S. (1979). B Lymphocytes in the Immune Response. M. Cooper et al., eds. (Elsevier-North Holland, Inc.) pp. 33-38.

Perry, R. P., and Kelley, D. E. (1979). Cell 18, 1333-1339.

Proudfoot, N. J., and Brownlee, G. G. (1976). Nature 263, 211-214.

Rajewsky, K., and Eichmann, K. (1977). Contemporary Topics in Immunobiol. 7, 69-112.

Raschke, W. C., Mather, E. L., and Koshland, M. E. (1979). Proc. Nat. Acad. Sci. USA 76, 3469-3473.

Rigby, P. W. J., Dieckmann, M., Rhodes, C., and Berg, P. (1977). J. Mol. Biol. 113, 237-251.

Rogers, J., and Wall, R. (1980). Proc. Nat. Acad. Sci. USA, in press.

Sakano, H., Rogers, J. H., Huppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R., and Tonegawa, S. (1979). Nature 277, 627-633.

Sanger, F., and Coulson, A. R. (1978). FEBS Letters 87, 107-110.

Sargent, T. D., Wu, J.-R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A., and Bonner, J. (1979). Proc. Nat. Acad. Sci. USA 76, 3256-3260.

Segrest, J. P., and Feldmann, R. J. (1974). J. Mol. Biol. 87, 853-858.

Seidman, J. G., Max, E. E., and Leder, P. (1979). Nature 280, 370-375.

Seif, I., Khoury, G., and Dhar, R. (1979). Nucleic Acids Res. 6, 3387-3398.

Sibley, C., Ewald, S., Kehry, M., Douglas, R., and Hood, L., submitted.

Snell, D. T., and Offord, R. E. (1972). Biochem. J. 127, 167-178.

Sutcliffe, J. G. (1978). Nucleic Acids Res. 5, 2721-2728.

Thomas, M., White, R. L., and Davis, R. W. (1976). Proc. Nat. Acad. Sci. USA 73, 2294-2298.

Tolstoshev, P., Boyd, C. D., Schafer, M. P., Trapnell, B. C., Coon, H. C., Kretschmer, P. J., Nienhuis, A. W., and Crystal, R. G. (1980). Miami Winter Symposia, Vol. 17, W. J. Whelan and J. Schultz, eds. (New York: Academic Press) in press.

Tomita, M., and Marchesi, V. T. (1975). Proc. Nat. Acad. Sci. USA 72, 2964-2968.

Tucker, P. W., Marcu, K., Newell, N., Richards, J., and Blattner, F. R. (1979). Science 206, 1303-1306.

Vassalli, P., Tedghi, R., Lisowska-Bernstein, B., Tartakoff, A., and Jaton, J.-C. (1979). Proc. Nat. Acad. Sci. USA 76, 5515-5519.

Vitetta, E. S., Baur, S., and Uhr, J. W. (1971). J. Exp. Med. 134, 242-264.

Wahl, G. M., Stern, M., and Stark, G. R. (1979). Proc. Nat. Acad. Sci. USA 76, 3683-3687.

Wall, R., Lippman, S., Toth, K., and Federoff, N. (1977). Anal. Biochem. 82, 115-129.

Warner, N. L., Leary, J. F., and McLaughlin, S. (1979). B Lymphocytes in the Immune Response, M. Cooper et al., eds. (Elsevier-North Holland, Inc.) pp. 371-378.

Williams, P. B., Kubo, R. T., and Grey, H. M. (1978). J. Immunol. 121. 2435-2439.

Yuan, D., Uhr, J. W., Knapp, M. R., Slavin, S., Strober, S., and Vitetta, E. S. (1979). B Lymphocytes in the Immune Response. M. Cooper et al., eds. (Elsevier-North Holland, Inc.) pp. 23-31.

Table 1. Sequences at the End of the Final Domain in Immunoglobulin Heavy Chains

	Protein Sequences
μ	-- Thr Gly Lys Pro Thr --
γ1	-- Pro Gly Lys*
γ2b	-- Pro Gly Lys*
α	-- Ala Gly Lys Pro Thr --
ε	-- Pro Gly Lys
	DNA Sequences
μ	-- ACT GGT AAA CCC ACA --
γ1	-- CCT GGT AAA TGA
γ2b	-- CCG GGT AAA TGA
α	-- GCN GGN AAR CCN ACN --
ε	-- CCN GGN AAR TRR
Consensus splice site	A GGT AAG TA

Sequences are shown for one representative of each heavy chain class sequenced. These are the contiguously encoded sequences in μ, γ1 and γ2b, and presumptively so in α and ε. The mouse μ sequences are from cloned cDNA plasmid μ12 (this paper). Mouse γ1 and γ2b sequences are from cloned chromosomal DNA (Honjo et al., 1979; Tucker et al., 1979). The C-terminal lysine (*) in γ chains is not found in the mature proteins. Sequences for human α and ε chains were derived from protein (Low et al., 1976). The consensus sequence for upstream RNA splice sites (Rogers and Wall, 1980) is aligned with the splice site observed in μ6. The underlined GT is the obligatory dinucleotide at the beginning of an intron. R = A or G, N = any nucleotide.

Figure Legends

Figure 1. Electron Micrograph of R-Loops Formed with M104E mRNA on μ Genomic DNA

The 9.8 kb EcoRI fragment of Ch603μ35 Calame et al., 1980) was reacted under R-loop conditions with size-selected, polyadenylated mRNA from M104E. One RNA molecule is hybridized to the four C_μ domains (C_μ1-4). The V region and the poly(A) tail are presumably represented by the collapsed knobs of RNA at each end. Another RNA molecule is hybridized to the M exon, 1.7 kb 3' to the C_μ4 domain. This R-loop too displays a knob at each end, the larger of which is interpreted as containing RNA encoding the V and C_μ1-4 domains (see text).

Figure 2. The Germline C_μ Gene Segment

The diagram shows the germline C_μ gene segment as cloned in ChSpμ7 and sub-cloned in pSpμA1 (Calame et al., 1980). The left-hand Eco RI site is synthetic. Solid boxes indicate the four C_μ domain exons and the μ_s 3' sequences, while the dashed box indicates the M exon seen in R-loops. Below the map are indicated the probes used for RNA blots: cDNA clone p104Eμ12 (μ12), Hha I fragment A, and Pst I fragment B.

Figure 3. RNA Blots of mRNA from MOPC 104E

Cytoplasmic polyadenylated RNA from M104E cells was transferred to an RNA blot and hybridized to ^{32}P-labeled μ12 (two exposures), Hha I fragment A (3'UT_s probe), and Pst I fragment B (M exon probe). The genomic locations of the probes are shown in Figure 2. The sizes of the two μ mRNAs were deduced from calibration with ribosomal RNA markers visualized by ethidium bromide staining.

Figure 4. RNA Blots of mRNA from the B-Cell Lymphoma WEHI 279 and the Hybridoma MPC11xW279.2

^{32}P-labeled probes were the same as in Figure 3.

Figure 5. Restriction Map and DNA Sequence from cDNA Clone μ12 (μ_s)

Arrows below the map indicate sequenced regions. The encoded amino acid sequence corresponds to that of the M104E μ_s chain (Kehry et al., 1979). The $(dA)_{15}$ segment represents the poly(A) of the mRNA, while the $(dC)_{13}$ linker and the reconstituted plasmid Pst I site are synthetic. The restriction map and C-terminal coding sequence are from Calame et al. (1980). They are reproduced here for comparison with the μ6 clone (Figure 6).

Figure 6. Restriction Map and DNA Sequence from cDNA Clone μ6 (μ_m)

There are no sites for Eco RI, Hha I or Taq I in the inserted DNA. Arrows below the map indicate sequenced regions. The encoded amino acid sequence is the predicted sequence for the μ_m chain. The $(dA)_{75\pm3}$ represents the poly(A) of the mRNA, while the $(dC)_9$ and the reconstituted plasmid Pst I site are synthetic.

Figure 7. Carboxy-Terminal Amino Acid Sequences of μ_s and μ_m Chains

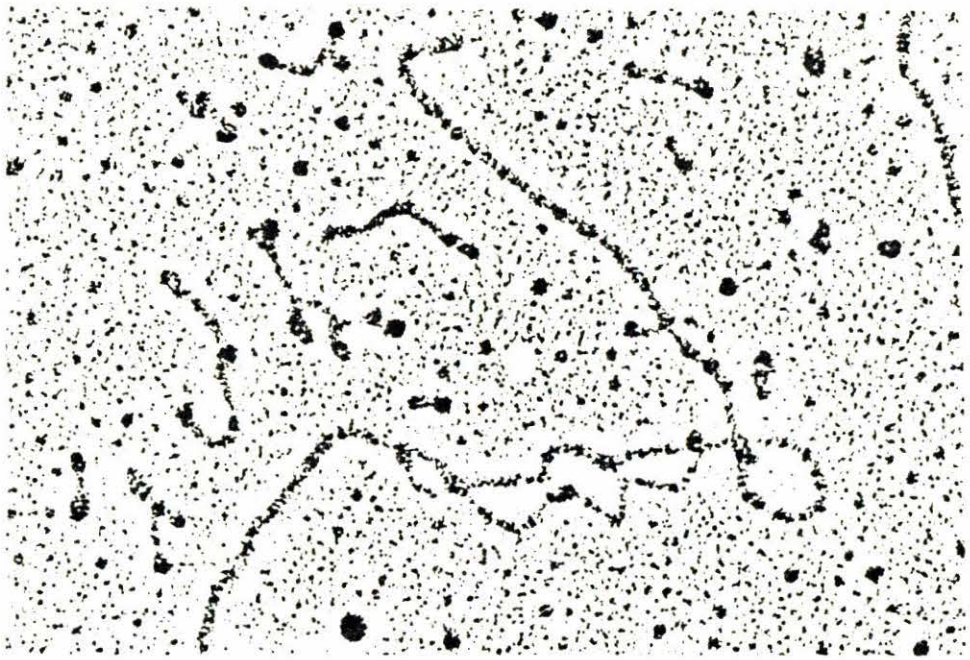
The μ_s sequence (upper) was established by Kehry et al. (1979) and confirmed by the DNA sequence of the μ12 clone. The 20-residue tail following the $C_\mu 4$ domain has no homology with the immunoglobulin domains, and includes the cysteine (boxed) which links μ_s chains to the J chain in secreted IgM (Mestecky and Schrohenloher, 1974) and the asparagine site of carbohydrate attachment (boxed CHO; Kehry et al., 1980). The μ_m sequence (lower) is predicted from the DNA sequence of

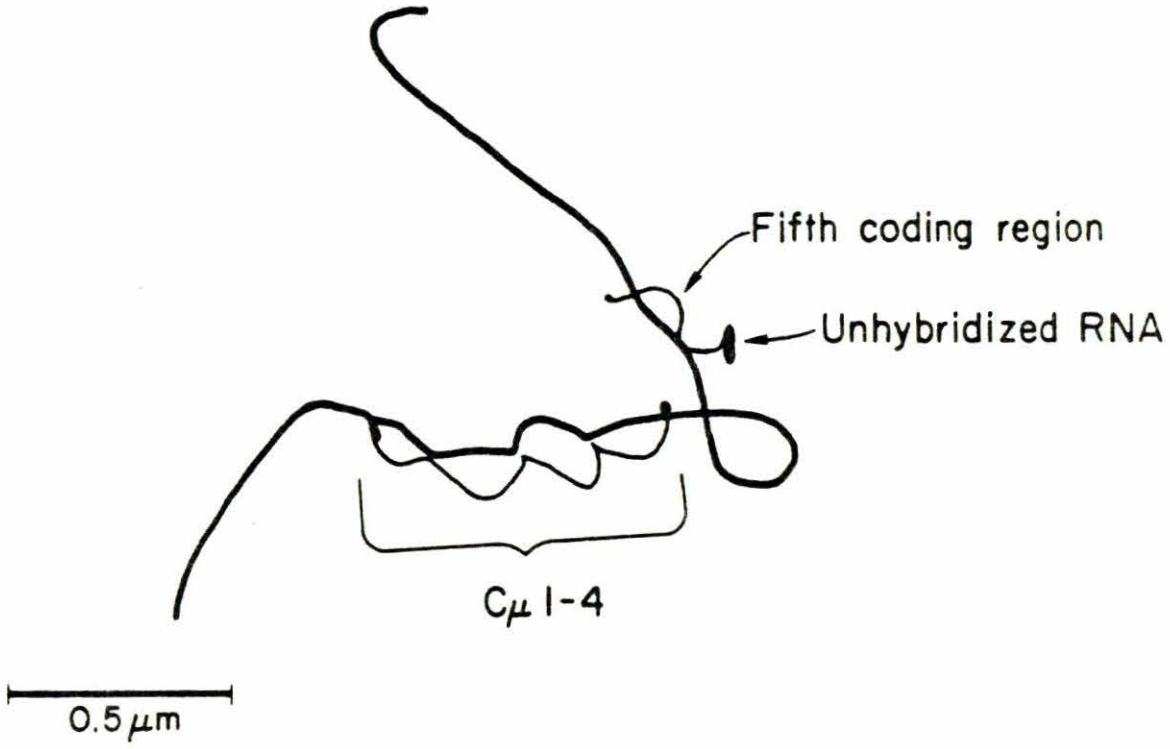
Figure 7 (continued)

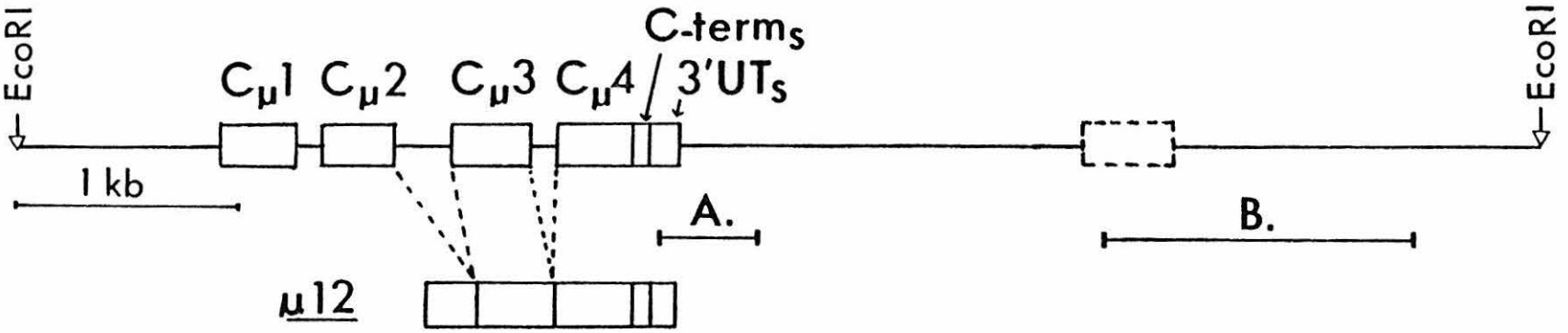
the μ6 clone. The 41-residue M segment has no homology with the immunoglobulin domains or the μ_s C-terminal segment, and includes the long hydrophobic sequence (underlined). Charged residues are boxed.

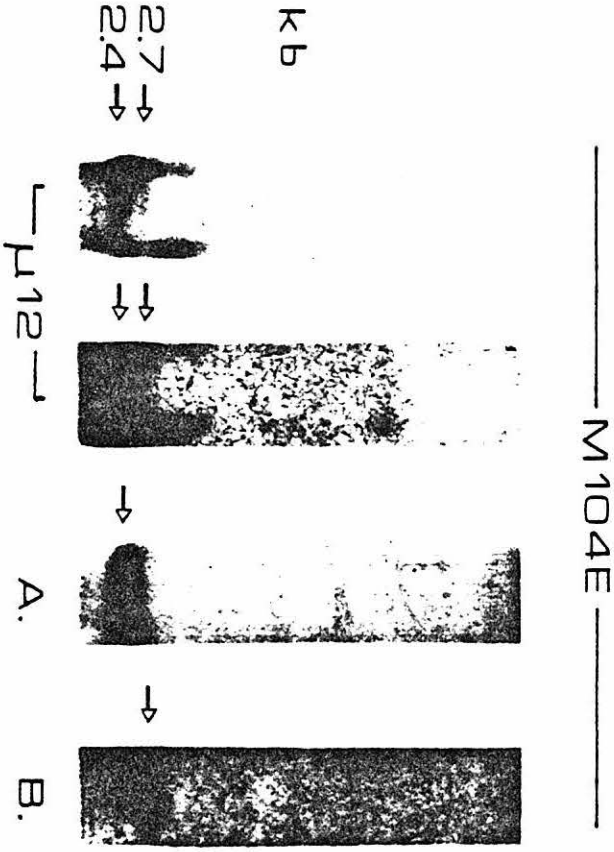
Figure 8. Possible Configuration of the M Segment in the Cell Membrane

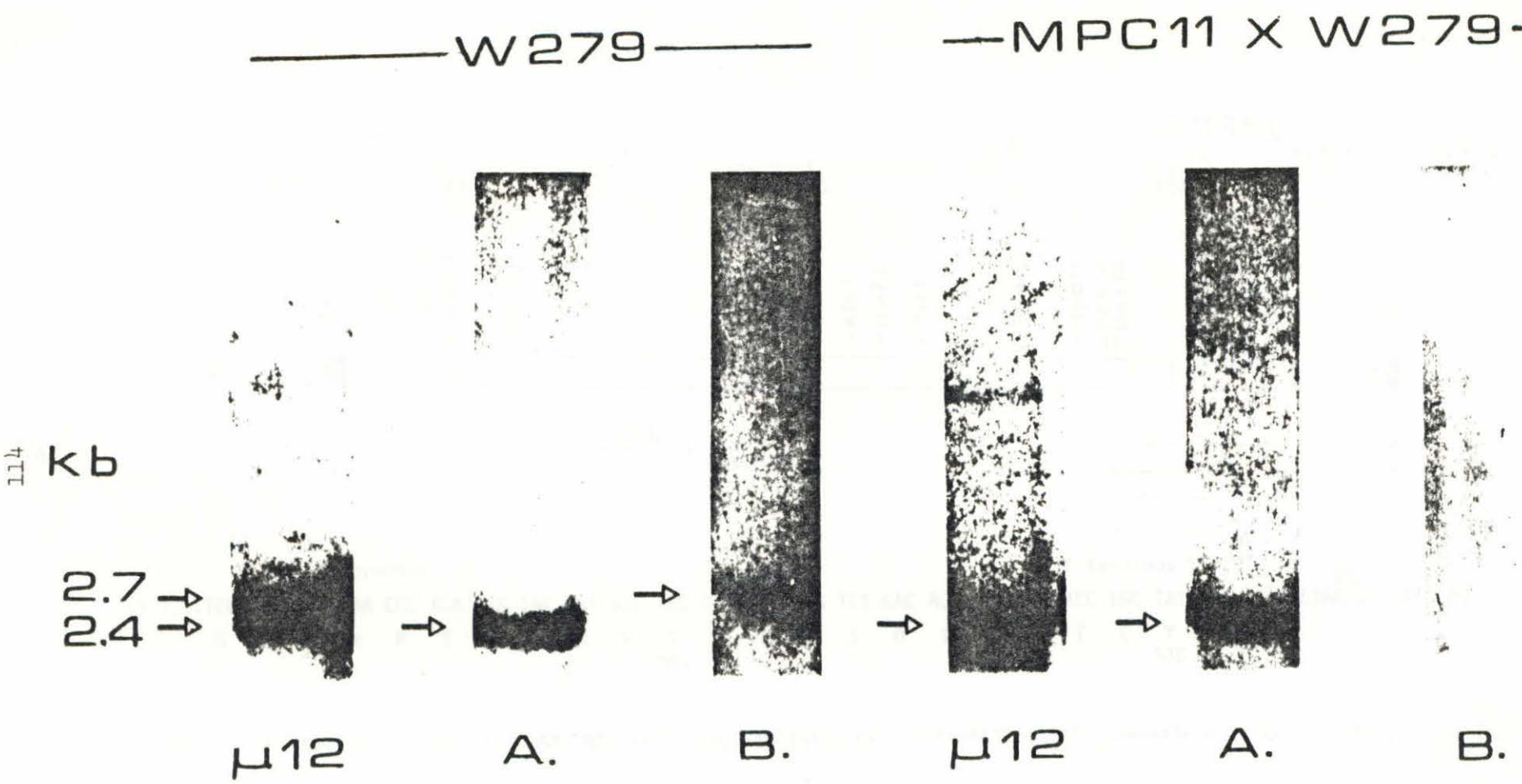
The μ_m chain can be anchored in the cell membrane by the M segment. The uncharged 26-residue sequence in the M segment is envisaged as an α helix spanning the membrane, as indicated by the coiled segment in the diagram. Positive and negative charged residues which flank the transmembrane sequence are indicated.

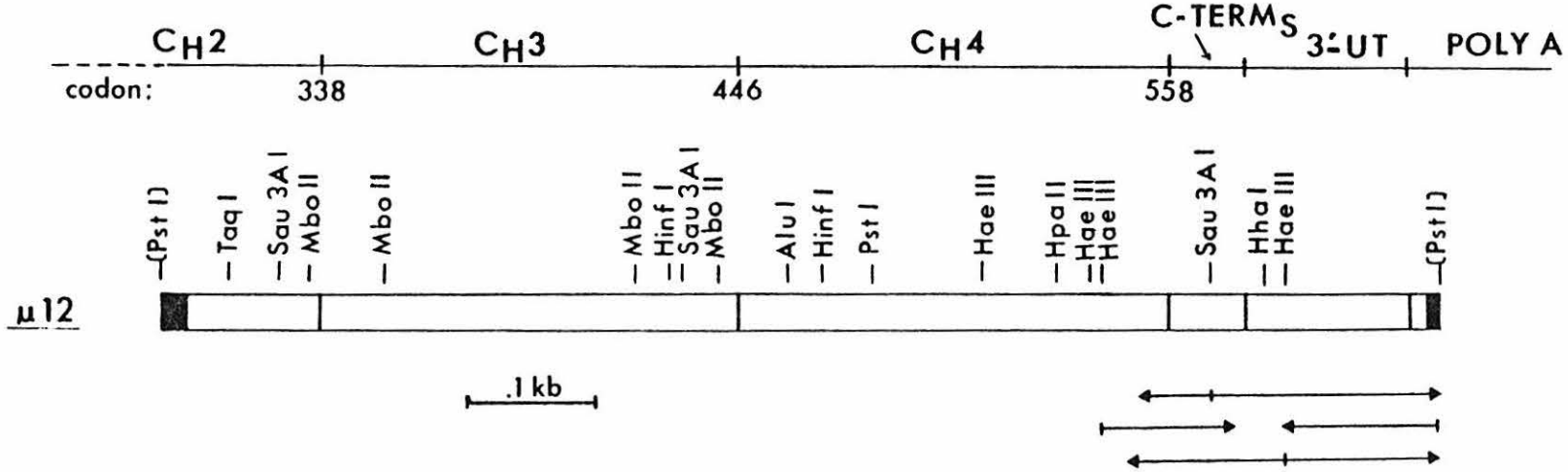



Fifth coding region
Unhybridized RNA
Cμ 1-4
0.5 μm


EcoRI
Cμ1 Cμ2 Cμ3 Cμ4
C-terms
3'UTs
EcoRI
1 kb
A.
B.
μ12


M104E
kb
2.7
2.4
μ12
A.
B.


W279
MPC11 X W279
kb
2.7
2.4
μ12
A.
B.
μ12
A.
B.



$C_\mu 4$ | C-terminus

(5')...TCC ACT | GGT AAA CCC ACA CTG TAC AAT GTC TCC CTG ATC ATG TCT GAC ACA GGC GGC ACC TGC TAT | TGA CCATGCTAGCGCTCAACCAG-

Sau3AI

C-terminus | $3'UT_S$ HhaI

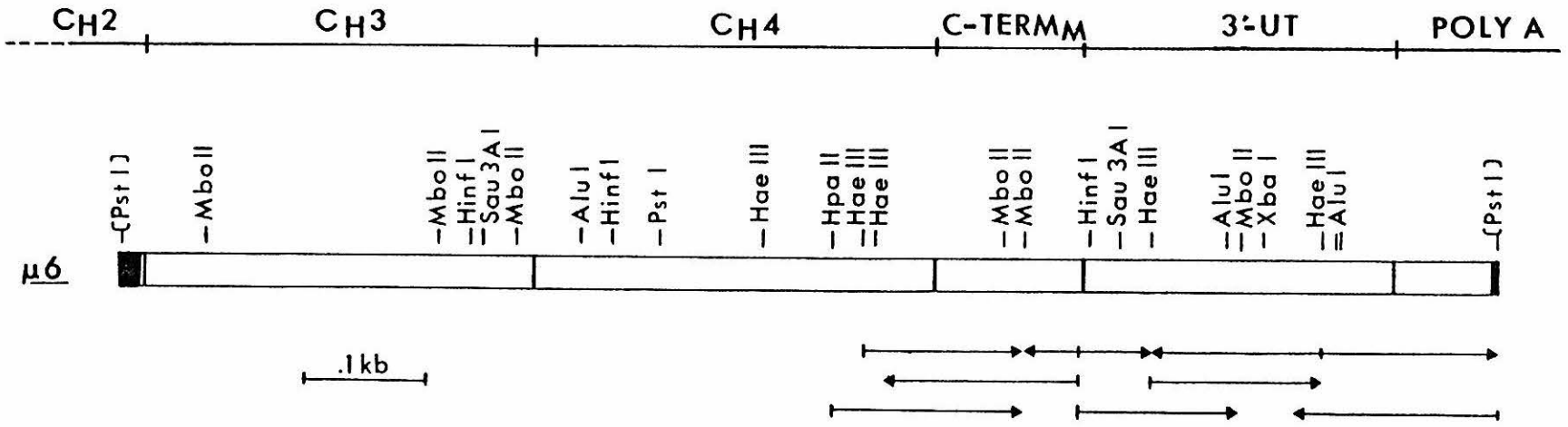
S T G K P T L Y N V S L I M S D T G G T C Y

555 565 576

HaeIII

-GCAGGCCCTGGGTGTCTAGTTGCTCTGTGTATGCAAACTAACCATGTCAGAGTGAGATGTTGCATTTTATAAAAATTAGAAATAAAAAAAATCCATTCAAACGTC-$(A)_{15}$-$(C)_{13}$-TGCAG-(3')

PstI



HpaII HaeIII HaeIII $C_\mu 4$ | M

(5')-CC GGA GAG ACC TAT ACC TGT GTT GTA GGC CAC GAG GCC CTG CCA CAC CTG GTG ACC GAG AGG ACC GTG GAC AAG TCC ACT | GAG-

G E T Y T C V V G H E A L P H L V T E R T V D K S T | E

531 540 550 557

MboII MboII

-GGG GAG GTG AAT GCT GAG GAG GAA GGC TTT GAG AAC CTG TGG ACC ACT GCC TCC ACC TTC ATC GTC CTC TTC CTC CTG AGC CTC TTC-

G E V N A E E E G F E N L W T T A S T F I V L F L L S L F

558 570 580 586

M HinfI $3'UT_m$ Sau3AI HaeIII

-TAC AGC ACC ACC GTC ACC CTG TTC AAG GTG AAA | TGA CTCTCAGCATGGAAGGACAGCAGAGACCAAGAGATCCTCCCACAGGGACACTACCTCTGGGCCTGGG-

Y S T T V T L F K V K |

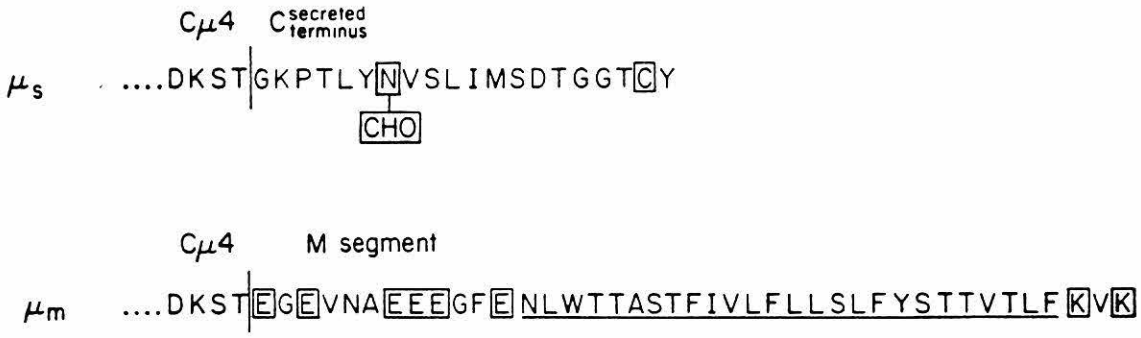
587 597

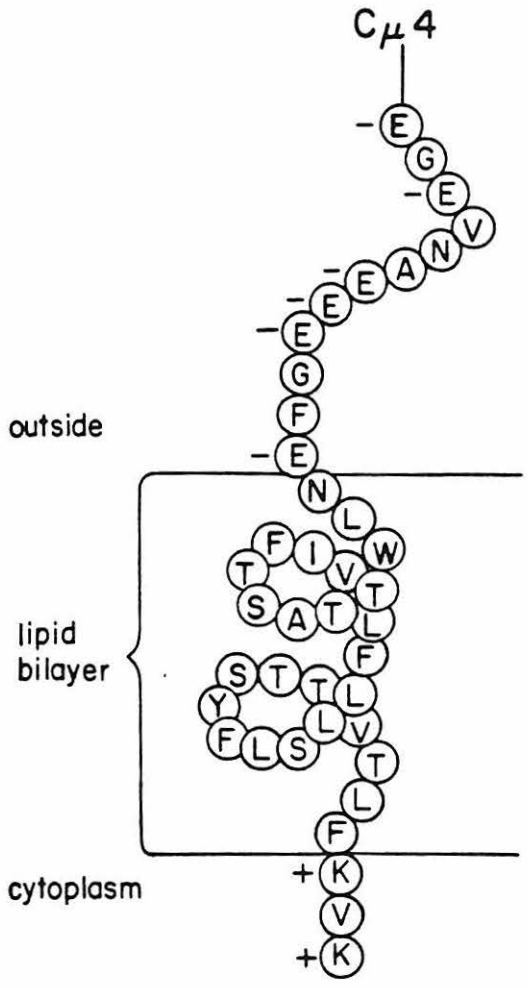
AluI MboII XbaI

-ATACCTGACTGTATGACTAGTAAACTTATTCTTACGTCTTTCCTGTGTTGCCCTCCAGCTTTTATCTCTGAGATGGTCTTCTTTCTAGACTGACCAAAGACTTTTTGTCAACTTG-

HaeIII AluI AluI PstI

-TACAATCTGAAGCAATGTCTGGCCCACAGACAGCTGAGCTGTAAACAAATGTCACATGGAAATAAATACTTTATCTTGTGAACTC-$(A)_{75\pm3}$-$(C)_9$-TGCAG-(3')


Cμ4
C secreted terminus
μs
....DKSTGKPTLYNVSLIMSDTGGTCY
CHO
Cμ4
M segment
μm
....DKSTEGEVNAEEEGFENLWTTASTFIVLFLLSLFYSTTVTLFKVK


Cμ4
outside
lipid
bilayer
cytoplasm

Two mRNAs can be Produced From a Single Immunoglobulin μ Gene by Alternative RNA Processing Pathways (II)

P. Early, J. Rogers,* M. Davis, K. Calame, M. Bond, R. Wall* and L. Hood

Division of Biology

California Institute of Technology

Pasadena, CA 91125

Cell (June 1980), in press.

*Molecular Biology Institute and Department of Microbiology and Immunology

UCLA School of Medicine

Los Angeles, CA 90024

Running Title: Immunoglobulin μ Gene Expression

Summary

As shown in the accompanying paper (Rogers et al., 1980), μ chains of the membrane-bound (μ_m) and secreted (μ_s) forms of IgM are encoded by two species of mRNA. Cloned cDNAs produced from the two μ mRNAs of M104E mouse myeloma tumors differ only at their 3' ends, which encode either the μ_m or μ_s C-terminus. In this paper, we show that both μ_m and μ_s mRNAs are produced from transcripts of a single μ gene. The last 187 nucleotides of μ_s mRNA are derived from DNA contiguous with the 3' end of the sequence encoding the $C_\mu 4$ domain. The μ_m cDNA clone does not include these 187 nucleotides, but instead contains 392 nucleotides derived from two exons located 1850 bp 3' to the $C_\mu 4$ sequence. Comparison of genomic and cDNA sequences shows that in μ_m mRNA, an RNA splice of 1850 nucleotides joins a site in the coding sequence at the end of $C_\mu 4$ with a site at the beginning of the first membrane-specific exon. A second RNA splice of 118 nucleotides joins sequences transcribed from the first and second membrane-specific exons.

The differences observed between μ_m and μ_s cDNAs suggest that developmental control of the site at which poly(A) is added to transcripts of the μ gene determines the relative levels of μ_m or μ_s chain synthesis. We discuss possible models for the control of μ gene transcripts and the significance of this form of developmentally regulated RNA processing for the evolution of eukaryotic "split genes".

Introduction

IgM antibody molecules exist in two separate forms, as monomeric membrane-bound receptors on lymphocytes and as secreted pentameric effectors of humoral immunity in the bloodstream. The μ heavy chains of IgM molecules contain four constant region domains, plus a short C-terminal segment (Kehry et al., 1979). In secreted IgM molecules, the C-terminal segment of μ chains forms disulfide bonds to a J chain linking each pentamer together. Monomeric IgM molecules on the surfaces of lymphocytes are attached to the cell membrane near their C-terminal ends. Comparative analyses of membrane-bound (μ_m) and secreted (μ_s) μ chain polypeptides suggest that they differ only in their respective C-terminal segments (Kehry et al., 1980; reviewed in Rogers et al., 1980).

The preceding paper presents sequences from two cDNA clones (μ6 and μ12) which separately encode the μ_m and μ_s chains from M104E mouse myeloma tumors (Rogers et al., 1980). Both cDNA clones contain identical sequences encoding the $C_\mu 3$ and $C_\mu 4$ domains. However, the μ_m and μ_s cDNA clones encode distinct C-terminal segments of 20 amino acids in the case of μ_s chains and 41 amino acids in μ_m chains (Rogers et al., 1980). The C-terminal segment encoded by the μ_s cDNA clone is identical to that determined from the complete amino acid sequence of M104E μ chains (Kehry et al., 1979). While the four C_μ domains are encoded by separate exons (Gough et al., 1980; Calame et al., 1980), previous evidence has suggested that the coding sequence for the μ_s C-terminal segment is contiguous with the 3' end of the $C_\mu 4$ exon (Calame et al., 1980). R-loop mapping experiments suggest that the μ_m C-terminal segment is encoded by a separate exon in the C_μ gene (Rogers et al., 1980).

In this paper, we examine the genomic origins of μ_m and μ_s mRNAs. We find that both mRNAs are transcribed from the same μ gene and that two additional

RNA splices generate μ_m mRNA. DNA rearrangement does not play a direct role in controlling μ_m or μ_s synthesis. A form of developmentally regulated RNA processing is responsible for determining the relative levels of μ_m and μ_s mRNAs produced in different cell types.

Results

Both μ_m and μ_s mRNAs Must be Derived from a Single Germline C_μ Gene

In determining the genomic origins of μ_m and μ_s mRNAs, we asked whether the two mRNAs are derived from one or two C_μ genes. As stated elsewhere (Calame et al., 1980), the haploid Balb/c mouse germline genome contains a single C_μ gene. Ten independently isolated germline C_μ genomic clones all contained the same C_μ gene and flanking sequences (M. Davis, K. Calame, P. Early, unpublished results). Hybridization of labeled μ_s (μ12) plasmid DNA to Southern (1975) blots of germline or embryo DNAs digested with one of four different restriction enzymes showed only the pattern of bands expected from the cloned germline C_μ gene (Davis et al., 1980; M. Davis, unpublished results). These observations, plus the fact that restriction maps and the DNA sequences of large portions of the μ_s and μ_m cDNA clones are identical (Rogers et al., 1980), lead to the conclusion that both the μ_m and μ_s constant regions must be encoded by the exons of the single germline C_μ gene.

The 3' Terminal Sequences of μ_m and μ_s cDNAs Hybridize to Separate Exons in the C_μ Gene

In order to determine the germline locations of the sequences present in the μ_m and μ_s cDNA clones, we hybridized the labeled cDNA clones to a Southern blot of restriction fragments from Ch$Sp\mu$7, a germline C_μ genomic clone (Davis et al., 1980). A restriction map of this C_μ clone (Calame et al., 1980) is shown in Figure 1.

The 4.3 kb XbaI fragment of ChSpμ7 indicated in Figure 1 was isolated, digested with various restriction enzymes, and the fragments separated on 8% polyacrylamide gels. Southern blots of these gels were hybridized with either μ_m (μ6) or μ_s (μ12) plasmid DNAs which had been labeled with ^{32}P by nick translation (Maniatis et al., 1975).

As seen in Figure 2, both the μ_s and μ_m cDNA clones hybridized with restriction fragments derived from the $C_\mu 3$ and $C_\mu 4$ exons. The μ_s cDNA clone also hybridized with restriction fragments containing the $C_\mu 2$ exon and secreted C-terminal segment sequences (Calame et al., 1980; Rogers et al., 1980). The μ_s cDNA clone (μ12) does not contain V_H or $C_\mu 1$ sequences (Calame et al., 1980) and the μ_m cDNA clone (μ6) does not contain V_H, $C_\mu 1$, or $C_\mu 2$ sequences (Rogers et al., 1980).

The μ_m cDNA clone did not hybridize with restriction fragments containing the secreted C-terminal sequences 3' to $C_\mu 4$ (Figure 2). However, the μ_m cDNA clone did hybridize with restriction fragments derived from a region at the 3' end of the 4.3 kb XbaI fragment, about 2 kb 3' to the $C_\mu 4$ exon, as can be seen by comparing the fragment sizes in Figure 2 with the restriction map in Figure 1. The μ_m cDNA clone also hybridized with the 1.4 kb XbaI fragment (Figure 1) located immediately 3' to the 4.3 kb XbaI fragment in the germline C_μ clone, ChSpμ7 (data not shown). The μ_s cDNA clone did not hybridize with this 1.4 kb XbaI fragment. These results, in conjunction with R-loop mapping (Rogers et al., 1980), suggest that the different C-terminal and 3' untranslated sequences of the μ_m and μ_s cDNA clones are encoded by separate exons about 2 kb apart within the germline C_μ gene.

The Specific μ_m mRNA Sequences Are Derived from Two Exons Located 1850 Base Pairs 3' to the $C_\mu 4$ Exon

The region of the ChSpμ7 germline C_μ clone surrounding the 3' end of the

4.3 kb XbaI fragment (Figure 1) was sequenced by the technique of Maxam and Gilbert (1977) to precisely define the relationship of this portion of the genome to the μ_m cDNA clone. As shown in Figure 3, this region includes two exons which encode the complete M segment and contain the 3' untranslated sequences of the μ_m cDNA clone. These will be referred to as the M exons. The first M exon encodes amino acids 557-595 of the M104E μ_m chain, while the second M exon encodes amino acids 596-597, and also contains the termination codon (UGA) and 3' untranslated region of the μ_m cDNA clone. An intron of 118 nucleotides separates the two M exons. The RNA splice site at the 5' end of the first M exon and the splice sites between the two M exons obey the GT. . .AG rule (Breathnach et al., 1978; Catterall et al., 1978).

The μ_s mRNA Is Derived from Sequences Contiguous with the $C_\mu 4$ Exon

Figure 4 shows the nucleotide sequence of a region in the ChSpμ7 C_μ clone contiguous with the 3' end of the $C_\mu 4$ exon. As suggested elsewhere by restriction mapping (Calame et al., 1980), this region includes the C-terminal sequence of the μ_s cDNA clone (μ12) directly adjoining the $C_\mu 4$ exon. The entire C-terminal and 3' untranslated regions of the μ_s cDNA clone μ12 (Calame et al., 1980; Rogers et al., 1980) are present as a continuous nucleotide sequence which terminates 187 nucleotides 3' to the $C_\mu 4$ exon.

The sequence of the ChSpμ7 genomic clone which includes the sequence at the 3' end of μ12 shows no extensive homology with the two M exons or the sequence following them. The genomic sequence following the 3' end of the M exons also is notably more T-rich than the sequence following the 3' end of μ12 (Figure 3). However, both the μ_m (μ6) and μ_s (μ12) cDNA clones (Rogers et al., 1980) contain the apparent poly(A) addition signal AATAAA located 19 nucleotides 5' to the terminal poly(A) sequence (Proudfoot and Brownlee, 1976). Poly(A) addition

to both μ_m and μ_s mRNAs occurs within the sequence TCACT located at the 3' termini of both cDNA sequences in the genomic clone. The presence of poly(A) in both cDNA clones and the similarity of the 3' terminal sequences to other poly(A) addition sites (Proudfoot and Brownlee, 1976) indicate that the μ6 and μ12 cDNA clones end at the normal 3' termini of μ_m and μ_s mRNAs, respectively.

The μ_m and μ_s mRNAs Are Derived from Transcripts of One μ Gene

Having defined the relationship of the μ_m and μ_s cDNA clones to the exons of the germline C_μ gene, it is necessary to consider whether rearrangement of the distinct μ_m and μ_s DNA sequences occurs in the genomes of cells producing secreted or membrane-bound IgM molecules. In an IgM-producing cell, DNA rearrangements will have created a V_H gene associated with a C_μ gene (Early et al., 1980; Davis et al., 1980). This type of DNA rearrangement presumably occurs only on one chromosome, since a single V_H region is expressed by each cell (reviewed by Goding, 1978). Additional DNA rearrangements within the expressed C_μ gene might lead to μ_m or μ_s synthesis. If this were the case, differentiated B cells which produce both μ_m and μ_s mRNAs, such as those in the M104E myeloma tumor (Rogers et al., 1980), could be expected to contain two distinct copies of the expressed C_μ gene.

The existence of two distinguishable copies of the C_μ gene in M104E DNA was tested in the Southern blot shown in Figure 5. For comparison, both embryo and M104E DNAs were digested with EcoRI, electrophoresed on an agarose gel, blotted to a nitrocellulose filter and hybridized with labeled μ12 plasmid DNA. Only one C_μ band is visible in either embryo or M104E DNA. This observation eliminates the possibility that the μ_m and μ_s mRNAs of the M104E myeloma tumor are produced from two copies of the expressed C_μ gene with distinguishable EcoRI cleavage patterns. Note that this result excludes any deletions within one of a putative two copies of the expressed C_μ gene, perhaps to remove the M exons,

since any such deletions would alter the size of the EcoRI fragment containing the resultant C_μ gene (Figure 1).

In Figure 5, the single C_μ EcoRI fragment in M104E DNA is about 9.9 kb in length, compared to 12.2 kb for the germline fragment. The lack of a band of germline length suggests that the unexpressed C_μ gene has been lost in M104E, a phenomenon frequently seen with myeloma tumors (Davis et al., 1979; Gough et al., 1980; Cory and Adams, 1980). The reduction in length of the C_μ EcoRI fragment in the expressed M104E C_μ gene is not the direct result of V_H joining, since the J_H gene segment (J_{H107}) used in M104E is located 5' to the germline C_μ EcoRI fragment (Early et al., 1980). Similar alterations in a C_μ EcoRI fragment, evidently not the direct result of V_H joining, have been seen in HPC 76, an IgM-producing myeloma tumor (Gough et al., 1980). In that case, analysis of a C_μ genomic clone from the myeloma tumor shows that the alterations have not affected C_μ gene organization 3' to $C_\mu 4$, including the M exons (compare the restriction map of our germline C_μ clone in Figure 1 with that in Figure 4 of Gough et al., 1980).

Figure 5 also shows that the 9.9 kb C_μ EcoRI fragment in M104E DNA is nearly identical in size to a cloned C_μ EcoRI fragment in Ch603μ35. The cloned Ch603μ35 C_μ gene (Calame et al., 1980) was derived from the M603 IgA-producing myeloma tumor, which contains low levels of the germline 12.2 kb C_μ EcoRI fragment (Figure 5). The Ch603μ35 clone has deleted DNA 5' to the C_μ gene during cloning, a phenomenon which we have consistently observed when recombinant Charon 4A phage containing the C_μ gene are grown in DP50supF (Davis et al., 1980; M. Davis, unpublished results). The deletions 5' to the C_μ gene which occurred in cloning Ch603μ35 are similar in size to the apparent deletions in the M104E and HPC76 myeloma tumors. In the myeloma tumors, as in the bacterial hosts of Ch603μ35, deletions 5' to the C_μ gene may not be specifically involved in the regulation

of μ gene expression, but only reflect a propensity for certain sequences (perhaps repeats) to undergo deletion during their replication over many cell generations. However, we cannot rule out the possibility that DNA rearrangements 5' to the C_μ gene, while not altering the organization of the M exons or the μ_s C-terminal coding sequence, could indirectly affect the relative levels of synthesis of the two forms of μ mRNA.

Normal mouse spleen lympocytes synthesize both μ_m and μ_s polypeptide chains (Vassalli et al., 1979). Spleen lymphocytes selected for IgM synthesis by the fluorescence-activated cell sorter contain expressed μ genes with the same arrangement of μ_m and μ_s sequences seen in the C_μ germline clone depicted in Figure 1 (P. Early and C. Nottenburg, unpublished observations). These observations and those discussed for the IgM-producing myeloma tumors indicate that DNA rearrangements do not directly determine μ_m or μ_s chain synthesis. Accordingly, both μ_m and μ_s mRNAs must be produced from transcripts of a single μ gene.

Discussion

RNA Splicing from the 3' $C_\mu 4$ Boundary Occurs Only in μ_m mRNA

The sequences of the C_μ genomic clone and the μ_s cDNA clone both contain CTGGTAAAC at the boundary of $C_\mu 4$ and the μ_s C-terminal segment(Figure 4). By contrast, the μ_m cDNA sequence at the same position, CTGAGGGGG (Rogers et al., 1980), is evidently the result of an RNA splice between the $C_\mu 4$ sequence and the sequence TCCCTTCATAG/AGGGGG at the 5' boundary of the first M exon (Figure 3). A slash is used to denote the splice junction, in accordance with the GT. . .AG rule (Breathnach et al., 1978; Catterall et al., 1978).

The μ_s C-terminal coding sequence which follows the $C_\mu 4$ splice site (CTG/GTAAAC) and the intron sequence at the 5' boundary of the first M exon (TCCCTTCATAG/AG)

both exhibit complementarity to the U1 sequence, as has been noted for other splicing junctions (Reddy et al., 1974; Lerner et al., 1980; Rogers and Wall, 1980). This suggests that U1 or a similar small nuclear RNA may be involved in the splicing of μ_m mRNA. The 1850 nucleotide splice to generate μ_m mRNA is the first example in eukaryotic cells of an RNA splice to remove sequences which would otherwise encode a polypeptide, i.e., the C-terminal segment of the μ_s chain.

Figure 6 shows a diagram of the splicing patterns deduced for μ_m and μ_s mRNAs. Since the μ12 and μ6 cDNA clones lack the variable region and some constant region sequences, it cannot be conclusively stated that the μ_m and μ_s mRNAs from the M104E myeloma tumor are identical 5' to $C_\mu 3$. However, available information on the sizes of the two forms of μ mRNA and the nature of the polypeptides they produce suggests that the μ_m and μ_s mRNAs are identical from their 5' ends to codon 556 (Figure 4) (Rogers et al., 1980; Kehry et al., 1980).

The μ_m mRNA is produced by two more RNA splices than is μ_s mRNA (Figure 6). The first of these splices occurs at the boundary of $C_\mu 4$ with the μ_s C-terminal sequence, so that the coding region for the μ_s C-terminal segment is eliminated from μ_m mRNA. Instead, a longer hydrophobic membrane-bound C-terminal segment, the M segment, is derived from the two M exons located 1850 bp 3' to the $C_\mu 4$ exon. The precursor to μ_s mRNA does not undergo RNA splicing at the 3' end of the $C_\mu 4$ sequence (Figure 6). The μ_s mRNA terminates 187 nucleotides 3' to the $C_\mu 4$ exon, eliminating the possibility of an RNA splice to the M exon sequence.

Developmental Regulation of RNA Processing Directs the Synthesis of μ_m or μ_s mRNAs

Although production of μ_m as opposed to μ_s mRNA involves two additional RNA splices, it is unlikely that any developmental regulation of the splicing process

itself takes place in the μ transcripts. The addition of poly(A) to the 3' ends of nuclear RNAs is a rapid process which generally precedes RNA splicing, as shown in adenovirus 2 (Nevins and Darnell, 1978) and SV40 (Lai et al., 1978), as well as immunoglobulins (Schibler et al., 1978; Gilmore-Hebert and Wall, 1979). This observation suggests that the control of μ_s mRNA synthesis depends on the addition of poly(A) to RNA transcripts 187 nucleotides 3' to the $C_\mu 4$ exon. When polyadenylation at this point occurs before splicing to the M exons can take place, μ_s mRNA is produced. Otherwise, elongation of the transcript continues and RNA splicing produces μ_m mRNA.

Nuclear RNA 3' ends for poly(A) addition may be created either by cleavage of nascent transcripts or by termination of transcription. The 3' ends of some adenovirus 2 late RNAs in vitro appear to be produced by termination (J. Manley, personal communication). On the other hand, RNA cleavage generates the 3' ends of a majority of late transcripts in adenovirus 2 (Nevins and Darnell, 1978; Fraser et al., 1979) and SV40 (Ford and Hsu, 1978; Lai et al., 1978). In these cases, transcription continues for more than 1000 nucleotides 3' to some cleavage sites. Accordingly, either RNA cleavage or termination of transcription could be the mechanism producing the 3' ends of μ mRNAs.

The control of polyadenylation in μ transcripts could be exercised by either positive or negative regulators of RNA cleavage or termination. These would presumably interact differentially with nucleotides in the C_μ gene near either the first (μ_s) or second (μ_m) sites of polyadenylation. Fusion of the B lymphoma W279, which produces both μ_m and μ_s mRNAs, with the myeloma cell line MPC11 results in a hybridoma line which only synthesizes μ_s mRNA (Raschke et al., 1979; Rogers et al., 1980). This observation suggests that a positive regulator contributed by the myeloma cell is responsible for inducing polyadenylation at the first site to

produce μ_s mRNA. In plasma cells secreting IgM, the change from predominantly μ_m to predominantly μ_s mRNA synthesis might occur by the production or release of a positive regulator molecule which either enhances cleavage of transcripts 187 nucleotides 3' to the $C_\mu 4$ sequence or interacts with RNA polymerase II to cause termination of transcription at this point. Various ratios of μ_m and μ_s mRNAs could be synthesized, depending on the concentration of the regulator.

Perhaps the simultaneous expression of μ and δ chains with identical V_H regions in a single B lymphocyte at an early stage of development (reviewed by Goding, 1978) can be explained by an extension of some μ transcripts to include the C_δ sequences, which might be spliced directly to V_H, eliminating the intervening C_μ sequences. This idea is particularly attractive because the C_δ gene segment is located about 6 kb 3' to the M exons in germline DNA (K. Moore and T. Hunkapiller, personal communication), whereas the spacing between the C_γ genes is more than 20 kb (M. Davis and S. Kim, personal communication).

RNA Splicing and the Evolution of Developmentally Regulated Polyadenylation Sites

The presence of alternative membrane and secreted C-terminal sequences adds flexibility to the C_μ gene. The products of a single expressed μ gene can exist either as membrane-bound or serum antibodies. The evolution of this dual role for a single gene apparently depended on the existence of RNA splicing between exons encoding functionally distinct portions of the μ polypeptide. An early ancestor of the C_μ gene probably included an M exon (Rogers et al., 1980). Messenger RNA produced from this gene encoded an exclusively membrane-bound cell surface receptor protein. This hypothesis is supported by phylogenetic studies on certain invertebrates which indicate that immune-like reactions are mediated by cell-surface recognition molecules and not by secreted proteins (Marchalonis and Cone, 1973).

The evolution of a second role for the ancestral μ polypeptide, that of a secreted protein, probably began with a mutation in the intron preceding the M exon. This mutation caused some RNA transcripts of the ancestral μ gene to be polyadenylated prior to the M exon. Messenger RNA produced from these transcripts would no longer have undergone splicing to the M exon, but instead would have contained a new 3' terminus derived from part of the former intron. When translated into protein, the new type of mRNA would have produced the N-terminal portion of the ancestral μ polypeptide, but with the M segment replaced by a series of amino acids derived from the former intron sequence. This novel form of the ancestral μ polypeptide could be secreted, since it retained an N-terminal "signal peptide" but lacked the membrane-binding M segment. Subsequent evolution presumably optimized the sequence of the μ_s C-terminal segment for interactions with an aqueous environment and J chain (Koshland, 1975). In time, developmental control regulating polyadenylation at either the first or second site in the C_μ gene was also acquired.

Developmental regulation of polyadenylation sites may also have evolved in other eukaryotic "split genes". The potential flexibility this provides for gene expression may be an evolutionary advantage for the eukaryotic form of gene organization. Such flexibility would be in addition to the role that separated exons may have played as independent genetic elements, able to recombine to form new "split genes" (Gilbert, 1978; Davis et al., 1979). Perhaps as the result of such processes, the M exons or similar sequences may be found to occur in a variety of genes encoding membrane-bound proteins.

Materials

The germline C_μ clone ChSpμ7 was isolated from a Charon 4A phage library derived from BALB/c mouse sperm DNA partially digested with HaeIII and AluI and ligated

to synthetic EcoRI linkers (Maniatis et al., 1978; Davis et al., 1980). The cells used to prepare DNA were assayed by light microscopy to be greater than 99% sperm (Joho et al., 1980). In addition to the 6.7 kb EcoRI fragment shown in Figure 1, ChSpμ7 contains an 8 kb Eco RI fragment 3' to the C_μ gene (Calame et al., 1980). Ch603μ35 was isolated from a Charon 4A phage library containing M603 myeloma DNA partially digested with EcoRI (Early et al., 1979; Calame et al., 1980). Recombinant plasmids μ12 and μ6 contain M104E μ_s and μ_m cDNA sequences cloned in the PstI site of pBR322 by dG:dC tailing (Calame et al., 1980; Rogers et al., 1980).

Restriction endonucleases and polynucleotide kinase were obtained from New England Biolabs and Boehringer/Mannheim. BA85 Nitrocellulose filter sheets were from Schleicher and Schuell, and γ-^{32}P-ATP (~9000 Ci/mmole) was from ICN. Manipulations of organisms containing recombinant DNA were performed under P2/EK2 or P2/EKI containment conditions in compliance with the NIH guidelines.

Acknowledgements

This work was supported by NIH grant AI 16590 to L.H., and NIH grants AI 13410 and CA 12800 to R.W. P.E. and M.D. are supported by NIH Training Grant GM07616. K.C. is supported by NIH Fellowship GM 05442. We thank Marilyn Kehry and Richard Douglas for helpful discussions.

References

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978). Proc. Nat. Acad. Sci. USA 75, 4853-4857.

Calame, K., Rogers, J., Early, P., Davis, M., Livant, D., Wall, R. and Hood, L. (1980). Nature, in press.

Catterall, J. F., O'Malley, B. W., Robertson, M. A., Staden, R., Tanaka, Y. and Brownlee, G. G. (1978). Nature 275, 510-513.

Cory, S. and Adams, J. M. (1980). Cell 19, 37-51.

Davis, M., Early, P., Calame, K., Livant, D. and Hood, L. (1979). In Eukaryotic Gene Regulation, R. Axel, T. Maniatis and C. F. Fox, eds. ICN-UCLA Symposium (Academic Press, New York), pp. 393-406.

Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L. and Hood, L. (1980). Nature 283, 733-739.

Early, P. W., Davis, M. M., Kaback, D. B., Davidson, N. and Hood, L. (1979). Proc. Nat. Acad. Sci. USA 76, 857-861.

Early, P., Huang, H., Davis, M., Calame, K. and Hood, L. (1980). Cell 19, 981-992.

Ford, J. P. and Hsu, M.-T. (1978). J. Virology 28, 795-801.

Fraser, N. W., Nevins, J. R., Ziff, E., Darnell, J. E. (1979). J. Mol. Biol. 129, 643-656.

Gilbert, W. (1978). Nature 271, 501.

Gilmore-Hebert, M. and Wall, R. (1979). J. Mol. Biol. 135, 879-891.

Goding, J. W. (1978). In Contemporary Topics in Immunobiology, vol. 8, N. L. Warner and M. D. Cooper, eds. pp. 203-243. (Plenum, New York)

Gough, N. M., Kemp, D. J., Tyler, B. M., Adams, J. M. and Cory, S. (1980). Proc. Nat. Acad. Sci. USA 77, 554-558.

Joho, R., Weissman, I. L., Early, P., Cole, J. and Hood, L. (1980). Proc. Nat. Acad. Sci. USA 77, 1106-1110.

Kehry, M., Sibley, C., Fuhrman, J., Schilling, J. and Hood, L. (1979). Proc. Nat. Acad. Sci. USA 76, 2932-2936.

Kehry, M., Ewald, S., Douglas, R., Sibley, C., Raschke, W., Fambrough, D. and Hood, L. (1980). Cell, submitted.

Koshland, M. E. (1975). Adv. in Immunol. 20, 41-69.

Lai, C.-J., Dhar, R. and Khoury, G. (1978). Cell 14, 971-982.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. and Steitz, J. A. (1980). Nature 283, 220-224.

Maniatis, T., Jeffrey, A. and Kleid, D. G. (1975). Proc. Nat. Acad. Sci. USA 72, 1184-1188.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. and Efstratiadis, A. (1978). Cell 15, 687-701.

Marchalonis, J. J. and Cone, R. E. (1973). Aust. J. Exper. Biol. Med. Sci. 51, 461-488.

Maxam, A. M. and Gilbert, W. (1977). Proc. Nat. Acad. Sci. USA 74, 560-564.

Nevins, J. R. and Darnell, J. E. (1978). Cell 15, 1477-1493.

Proudfoot, N. J. and Brownlee, G. G. (1976). Nature 263, 211-214.

Raschke, W. C., Mather, E. L. and Koshland, M. E. (1979). Proc. Nat. Acad. Sci. USA 76, 3469-3473.

Reddy, R., Ro-Choi, T. S., Henning, D. and Busch, H. (1974). J. Biol. Chem. 249, 6486-6494.

Rogers, J., Early, P., Calame, K., Carter, C., Bond, M., Hood, L. and Wall, R. (1980). Cell, submitted.

Rogers, J. and Wall, R. (1980). Proc. Nat. Acad. Sci. USA, in press.

Schibler, U., Marcu, K. B., Perry, R. P. (1978). Cell 15, 1495-1509.

Southern, E. M. (1975). J. Mol. Biol. 98, 503-517.

Vassalli, P., Tedghi, R., Lisowska-Bernstein, B., Tartakoff, A. and Jaton, J.-C. (1979). Proc. Nat. Acad. Sci. USA 76, 5515-5519.

Figure Legends

Figure 1. Restriction Map of the C_μ Gene in the Germline ChSpμ7 Clone

The EcoRI fragment shown is 6.7 kb long, with the direction of transcription from left to right. The triangle at the left end signifies a synthetic EcoRI site created by the EcoRI linker library technique (Maniatis et al., 1978). Raised boxes are exons and the horizontal lines are introns or nontranscribed sequences. Shading denotes 3' untranslated sequences present in either μ_m or μ_s mRNAs. The expanded diagram in the lower portion of the figure shows the M exon restriction fragments detected by hybridization to the μ_m cDNA clone μ6 (Figure 2).

Figure 2. Localization of μ_m Sequences in the Germline C_μ Gene

The 4.3 kb XbaI fragment of ChSpμ7 (Figure 1) was digested with the indicated restriction enzymes and electrophoresed on 8% polyacrylamide gels with pBR322 markers. Southern blots from separate gels were hybridized with either μ_m (μ6) or μ_s (μ12) cDNA plasmids labeled with ^{32}P by nick translation. Restriction fragments derived from $C_\mu 3$ and $C_\mu 4$ exons hybridized to both plasmids (Calame et al., 1980; Rogers et al., 1980; J. Rogers, unpublished results) and are not labeled. Note that the gel on the right (μ12) has been electrophoresed longer than the gel on the left (μ6). As indicated, restriction fragments which hybridized only to μ12 originate either from the $C_\mu 2$ exon or the specific μ_s 3' sequence (Calame et al., 1980; Rogers et al., 1980). The fragments whose lengths are shown hybridized only with μ6, and all of these can be derived from the restriction map of the M exons shown in Figure 1.

Figure 3. Sequence of the M Exons

The μ_m amino acids encoded by these exons are numbered by homology with human μ chains (Kehry et al., 1979). M104E μ_m mRNA actually encodes 593 amino acids, excluding the N-terminal signal peptide. The UGA termination codon is boxed. GT. . .AG sequences at RNA splice sites are underlined. The 3' terminus of μ_m cDNA (μ6) is indicated by the poly(A) addition point. A hexanucleotide sequence associated with polyadenylation is underlined (Proudfoot and Brownlee, 1976). The lower portion of the figure shows the strategy used in determining this sequence.

Figure 4. Genomic Sequence Encoding the μ_s C-terminal Segment

Amino acids are numbered by homology with human μ chains (Kehry et al., 1979). M104E μ_s chains actually contain 572 amino acid residues. The UGA termination codon is boxed. The GT boundary of the RNA splice which produces μ_m mRNA is underlined. The 3' terminus of μ_s cDNA (μ12) is indicated by the poly(A) addition point. A hexanucleotide sequence associated with polyadenylation is underlined (Proudfoot and Brownlee, 1976). The lower portion of the figure shows the sequencing strategy used for the genomic clone. Other nucleotides in this sequence were supplied from the sequence of μ12 (Calame et al., 1980; Rogers et al., 1980). As we have shown, the restriction maps of the μ12 cDNA clone and the genomic C_μ clone are identical in this region (Calame et al., 1980).

Figure 5. Southern Blot of C_μ Gene in Genomic DNAs

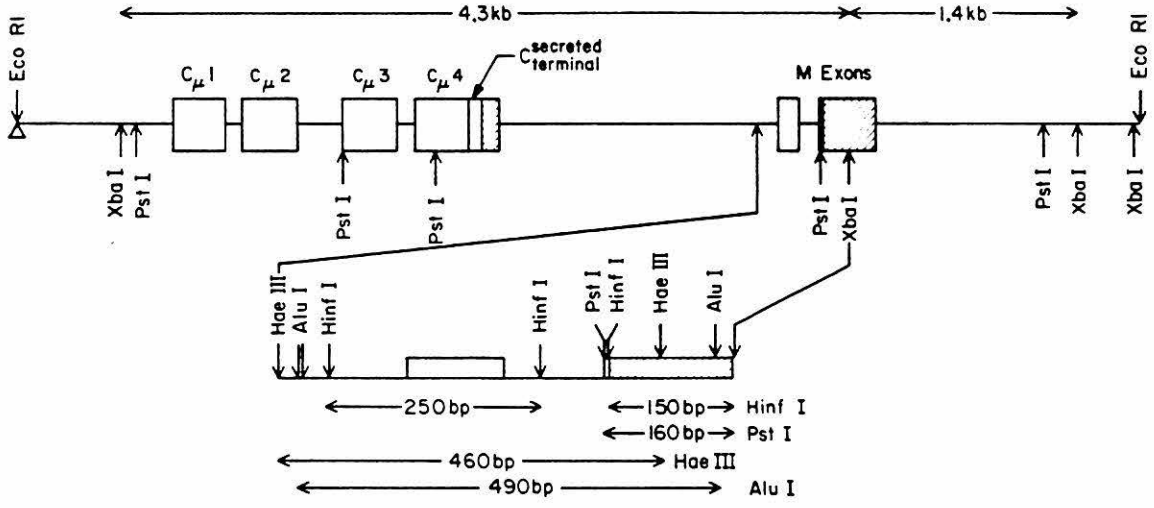
About 10 μg each of EcoRI digested DNA from BALB/c mouse embryos, M104E IgM-producing myeloma tumors or M603 IgA-producing myeloma tumors was electrophoresed on a 0.7% agarose gel. A Southern blot from this gel was hybridized

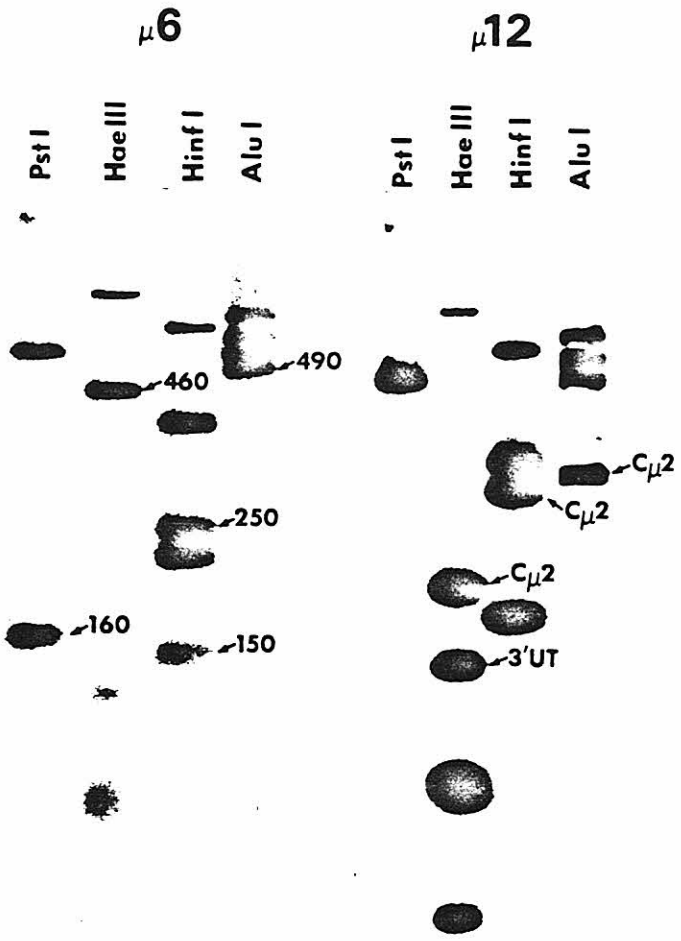
Figure 5 (continued)

to μ12 plasmid DNA labeled with ^{32}P by nick translation. Sizes of the C_μ EcoRI fragments, determined from parallel EcoRI fragments of λ DNA, can be compared to the 9.8 kb C_μ EcoRI fragment in Ch603μ35 (Calame et al., 1980).

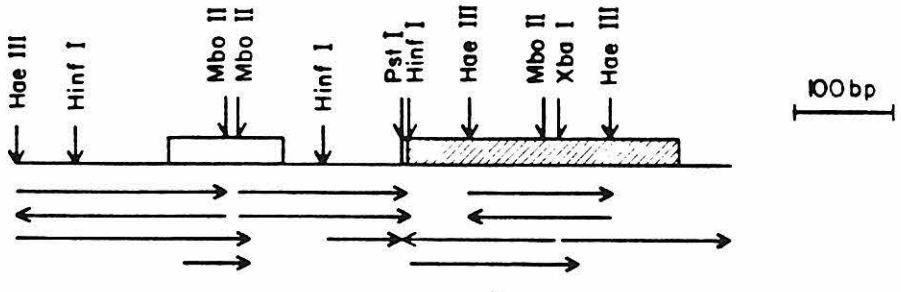
Figure 6. Splicing Patterns Deduced for μ_m and μ_s mRNAs

The μ_m and μ_s mRNAs are assumed to be identical 5' to $C_\mu 4$, although the μ6 and μ12 cDNA clones do not contain this complete sequence. Raised boxes indicate exons. 3' untranslated sequences are shaded. P refers to the signal peptide exon and V to the rearranged V_H exon. Bent lines indicate RNA splicing between exons.

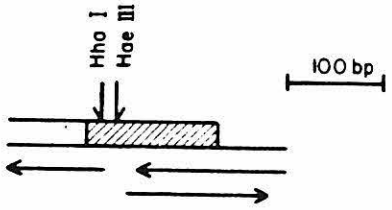

4.3kb
1.4kb
Eco RI
Eco RI
$C_\mu 1$
$C_\mu 2$
$C_\mu 3$
$C_\mu 4$
$C_{secreted\ terminal}$
M Exons
Xba I
Pst I
Pst I
Pst I
Pst I
Xba I
Pst I
Xba I
Xba I
Hae III
Alu I
Hinf I
Hinf I
Pst I
Hinf I
Hae III
Alu I
250bp
150bp → Hinf I
160bp → Pst I
460bp → Hae III
490bp → Alu I

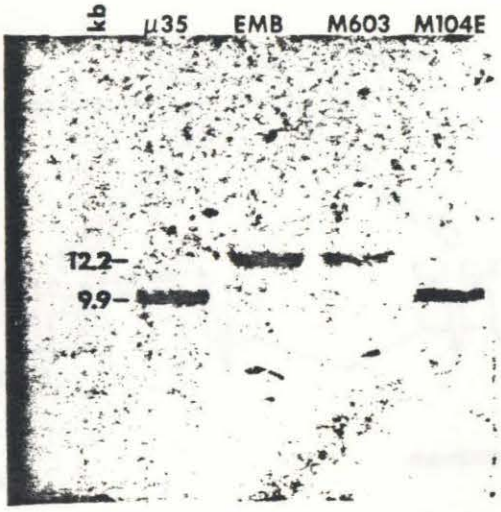

μ6
μ12
Pst I
Hae III
Hinf I
Alu I
490
460
250
160
150
Cμ2
Cμ2
Cμ2
3'UT

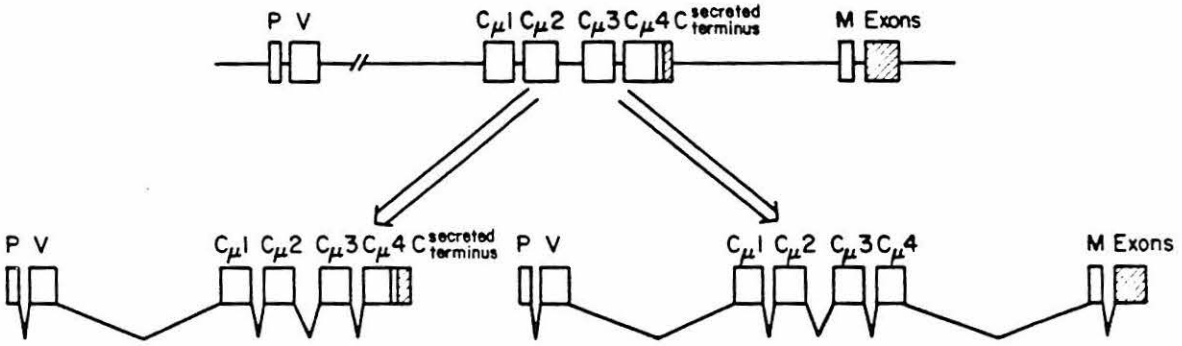
```
5'  GGCCTGTTCTGTGCCTCCGTCTAGCTTGAGCTATTAGGGGACCAGTCAATACTCGCTAAGATTCTCCAGAACCAT

    CAGGGCACCCCAACCCTTATGCAAATGCTCAGTCACCCCAAGACTTGGCTTGACCCTCCCTCTCTGTGTCCCTTC
       557
       GluGlyGluValAsnAlaGluGluGluGlyPheGluAsnLeuTrpThrThrAlaSerThrPheIleValLeu
    ATAGAGGGGGAGGTGAATGCTGAGGAGGAAGGCTTTGAGAACCTGTGGACCACTGCCTCCACCTTCATCGTCCTC
                                                          595
    PheLeuLeuSerLeuPheTyrSerThrThrValThrLeuPheLys
    TTCCTCCTGAGCCTCTTCTACAGCACCACCGTCACCCTGTTCAAGGTAGTATGGTTGTGGGGCTGAGGACACAGG

    GCTGGGACAGGGAGTCACCAGTCCTCACTGCCTCTACCTCTACTCCCTACAAGTGGACAGCAATTCACACTGTCT
                  596597
                  ValLys
    CTGTCACCTGCAGGTGAAATGACTCTCAGCATGGAAGGACAGCAGAGACCAAGAGATCCTCCCACAGGGACACTA

    CCTCTGGGCCTGGGATACCTGACTGTATGACTAGTAAACTTATTCTTACGTCTTTCCTGTGTTGCCCTCCAGCTT

    TTATCTCTGAGATGGTCTTCTTTCTAGACTGACCAAAGACTTTTTGTCAACTTGTACAATCTGAAGCAATGTCTG

                                                                    poly(A)
    GCCCACAGACAGCTGAGCTGTAAACAAATGTCACATGGAAATAAATACTTTATCTTGTGAACTCACTTTATTGTG

    AAGGAATTTGTTTTGTTTTTCAAACCTTTCCTGCGGTGTTGACAG  3'
```



```
      550
      ArgThrValAspLysSerThrGlyLysProThrLeuTyrAsnValSerLeuIleMetSerAspThrGlyGlyThr
5'    AGGACCGTGGACAAGTCCACTGGTAAACCCACACTGTACAATGTCTCCCTGATCATGTCTGACACAGGCGGCACC

         576
      CysTyr
      TGCTATTGACCATGCTAGCGCTCAACCAGGCAGGCCCTGGGTGTCTAGTTGCTCTGTGTATGCAAACTAACCATG

                                                                    poly(A)
      TCAGAGTGAGATGTTGCATTTTATAAAAATTAGAAATAAAAAAAATCCATTCAAACGTCACTGGTTTTGATTATA

      CAATGCTCATGCCTGCTGAGACAGTTGTGTTTTGCTTGCTCTGCACACACCCTGC  3'
```




kb
μ35
EMB
M603
M104E
12.2
9.9



secreted mu mRNA

membrane mu mRNA