Proc. Natl. Acad. Sci. USA
Vol. 75, No. 3, pp. 1485–1489, March 1978
Genetics

Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain

(λ light chain/hypervariable region/DNA sequencing/interspersed noncoding sequences)

SUSUMU TONEGAWA*, ALLAN M. MAXAM†, RICHARD TIZARD†, ORA BERNARD*, AND WALTER GILBERT†

* Basel Institute for Immunology, 487 Grenzacherstrasse, CH-4005 Basel, Switzerland; and † Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Contributed by Walter Gilbert, January 9, 1978

ABSTRACT **We have determined the sequence of the DNA of a germ-line gene for the variable region of a mouse immunoglobulin light chain, the $V\lambda_{II}$ gene. The sequence confirms that the variable region gene lies on the DNA separated from the constant region. Hypervariable region codons appear in the germ-line sequence. A sequence for the hydrophobic leader, 19 amino acids that are cleaved from the amino terminus of the protein, appears near, but not continuous with, the light chain structural sequence: most of the leader sequence is separated from the rest of the gene by 93 bases of untranslated DNA.**

Immunoglobulin molecules are synthesized by special differentiated cells which produce and export a protein specialized to bind an antigen. That binding site is a pocket between two subunits, a light chain and a heavy chain. The part of the light chain in contact with the antigen, the amino-terminal region, termed the variable region, differs extensively from one immunoglobulin to another; the carboxy-terminal half of the light chain, the constant region, is invariant and, in mice, falls into three classes, one κ and two λ. The variable region itself, though different in different immunoglobulins, shows certain consistencies in structure: three hypervariable regions embedded in a framework (1). The hypervariable regions span five to ten amino acids each around positions 30, 55, and 90 in the immunoglobulin chain and contain most of the variation. In the three-dimensional structure of an immunoglobulin, they are the segments in contact with antigen. Differences in sequence in framework regions provide a classification of light chains into different subgroups (2). Some of these differences are a change in only a single amino acid; others are more extensive. Estimates of the largest number of different variable region subgroups that can appear with single constant regions range from 10 to 100. (A similar structure holds for the heavy chains; the first 110 amino acids constitute a variable domain and the next 330, a series of constant domains.)

Hozumi and Tonegawa (3) showed, by restriction enzyme analysis and hybridization mapping techniques, that in the germ line, in embryonic tissue, the sequences coding for the variable region and those coding for the constant region of an immunoglobulin light chain lie separate from one another. These two genetic regions move during differentiation of lymphocyte precursors and come closer together to form a single transcription unit that will produce the final immunoglobulin molecule. Experiments with cloned immunoglobulin genes have since demonstrated this movement more directly and revealed a surprising structure for the active gene. Brack, Lenhard-Schuller, and Tonegawa (ref. 4; unpublished data) analyzed a gene isolated from a myeloma, by R-loop mapping and by heteroduplex comparisons with embryonic genes. They showed that the variable and constant region genes move from distant positions in the embryonic DNA to less distant positions in the producing cell but do not become contiguous. The variable and constant region coding sequences in the myeloma are separated still by 1250 bases. Presumably the DNA rearranges to bring these two regions close enough to lie on a single messenger precursor. One hypothesizes that this precursor then loses the additional sequences as the messenger matures in the nucleus.

In this paper we report the sequence of a germ-line gene for a mouse λ light chain variable region. We determined the sequence of this DNA for two reasons: to establish that the DNA coded for a variable region truly in isolation from any constant portion, and to determine whether a germ-line gene contained hypervariable as well as framework sequences.

A leading model for the behavior of the hypervariable regions, suggested by Kabat and his coworkers (1, 5), is that hypervariable region DNA is extraneous to the germ-line gene and inserts, like a small version of phage lambda, into receptor sites in the gene. This model would explain why different immunoglobulins in different subgroups occasionally have identical hypervariable regions. If this model were true, we would expect a germ-line gene to be a framework surrounding attachment sites rather than hypervariable sequences. Alternative models for the production of different immunoglobulins within one subgroup envisage the germ-line variable gene as a complete, inherited sequence that could be expressed with a unique antigenic specificity, a germ-line idiotype. This specificity will then change by mutation in somatic cells. Either these changes are localized in hypervariable regions by some process that makes the DNA in these areas unusually labile or, alternatively, mutations might arise anywhere in the variable gene but appear only in the hypervariable areas because the selection pressure matching immunoglobulin to antigen will select primarily for mutations in the antigen combining site. Determination of the sequence of a germ-line immunoglobulin light chain gene answers some of these questions.

MATERIALS AND METHODS

The recombinant phage λgt-WES-Ig 13, containing a 4800-base-pair insert of mouse DNA carrying a Vλ gene (6, 7), was grown under P3-EK2 conditions in Basel. A total of 5 mg of purified phage DNA was used for the sequence determination in Cambridge. Milligram amounts of the phage DNA were cleaved with a mixture of *Eco*RI and *Hae* III restriction endonucleases and the three cleavage products of the mouse insert, 750, 1500, and 2500 base pairs long, were separated on a 6 × 200 × 400 mm slab gel polymerized from 5% (wt/vol) acrylamide/0.17% (wt/vol) bisacrylamide/50 mM Tris-borate (pH

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

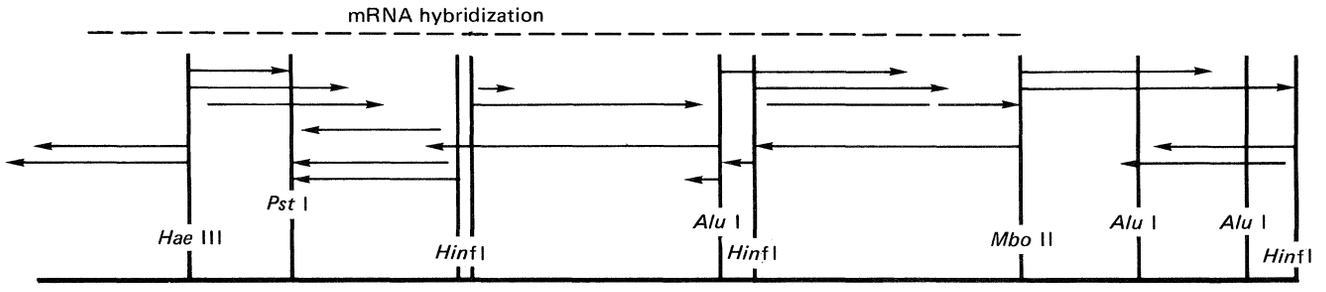

FIG. 1. Restriction endonuclease (7), mRNA hybridization (7), and DNA sequence map of the mouse embryonic $V\lambda_{II}$ region. The direction and the extent of the sequence determinations are shown by the arrows.

8.3)/1 mM EDTA, and extracted as described (8), except that the eluate was passed through a Millipore filter before ethanol precipitation.

The two larger *Hae* III fragments were then cleaved with *Hinf*, *Mbo* II, *Alu* I, or *Pst* I (9), end-labeled, and cleaved again, and their sequences were determined precisely as described (8) except for the following modifications. Tris·HCl buffer (50 mM, pH 9.5) was used instead of glycine/NaOH in the kinase reaction. One-tenth the carrier DNA (1 μg) and tRNA (5 μg) in all reactions for sequence determination produced sharper bands on autoradiograms. Addition of only sodium acetate, EDTA, and tRNA (no magnesium acetate) after the hydrazinolysis reactions and careful rinsing of the final ethanol precipitate to remove any residual sodium acetate eliminated occasional aberrations in the pyrimidine cleavage patterns. The "alternative guanine cleavage" (8) was used for guanine-only base-specificity.

For determination of long sequences, 200 nucleotides or so from the labeled end, products of sequencing reactions were loaded once or twice on the 20% gel described (8) and once on a 1.5 × 200 × 400 mm 10% gel polymerized from 9.5% (wt/vol) acrylamide/0.5% (wt/vol) bisacrylamide/7 M urea/100 mM Tris-borate (pH 8.3)/2 mM EDTA and electrophoresed at a high enough voltage (800–1100 V) and current to keep the surface of the gel at or above 50° during the run.

RESULTS AND DISCUSSION

Tonegawa *et al.* (6) isolated a derivative of phage lambda containing a 4800-base-pair insert of mouse DNA that hybridizes to the variable region half of a λ_I mRNA. A partial restriction map was constructed for the right third of this fragment, from a *Hae* III cut to the *Eco*RI end: the region in which hybridization to a messenger shows the putative variable region gene to lie (7). We chemically determined the sequence (8) of the relevant restriction fragments. In so doing, we could immediately establish the position of the variable region gene by comparison of the DNA sequence to the known protein sequences. Fig. 1 shows the restriction map of this area and indicates the experimental strategy that developed and checked the sequence. Fig. 3 shows the sequence of some 750 base pairs around the variable region gene.

Identification of Variable Region. Although the original clone was identified by hybridization to a message for a λ_I immunoglobulin, the sequence on the DNA corresponds more closely to a λ_{II} variable region. λ_I is a minor group in the mouse; about 5% of the light chains have this variable region attached to a λ_I constant region. The sequences of 18 λ_I chains from mouse myelomas have been determined (2, 10–12). Twelve of these are identical and the six others differ by one, two, or three amino acids in the hypervariable regions. λ_{II} accounts for only 1–2% of mouse immunoglobulin (H. N. Eisen, personal communication); the sequence of only one example, MOPC 315, has been determined (13). The variable regions of λ_I and λ_{II} chains are very similar; their frameworks differ in seven places. [The two λ constant regions differ at about 29 places (13).] Fig. 2 schematically compares the protein sequence determined by the DNA and those of the λ_I and λ_{II} variable regions, while Fig. 3 shows the entire sequence. Specifically, at framework residues 16 (Glu or Gly), 19 (Thr or Ile), 62 (Ala or Val), 71 (Asn or Asp), 85 (Glu or Asp), and 87 (Ile or Met), the DNA sequence corresponds to a λ_{II} chain (underlined). Thus we believe that we have determined the sequence of the λ_{II} germ-line gene.

There is one exception to this framework agreement: at position 38, λ_I has Val and λ_{II} has Ile, while the DNA sequence at this point predicts Val. We regard the DNA sequence as unambiguous. Since the protein sequence suggests that the expressed gene in MOPC 315 has Ile at this point, we have an excellent candidate for a mutational change in a *framework* region. This interpretation would suggest that single amino acid changes in the framework regions of immunoglobulins are mutations and do not represent different germ-line genes, reducing sharply the number of different subgroups of κ chains that one might estimate to exist in the germ line.

The λ subgroups in the mouse may represent a degenerate example: there appear to be only single variable-region genes for λ_I and λ_{II} (restriction enzyme analysis by M. Hirama and S. Tonegawa, unpublished). Presumably the other possible λ frameworks have been lost in the mouse.

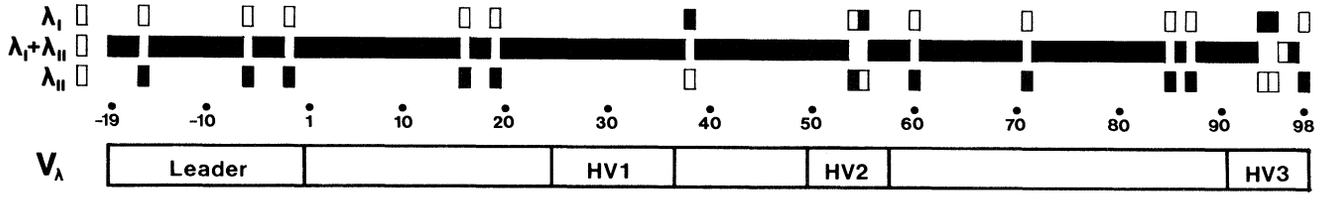

FIG. 2. Schematic comparison of the variable regions of mouse λ_I and λ_{II} immunoglobulin light chains with the DNA of the embryonic Vλ gene. The almost continuous bar in the middle represents an amino acid sequence common to most of the light chains, while boxes set apart from it mark positions at which they differ. Wherever the bar or a box is shaded, the sequence of that protein corresponds with the sequence of the DNA. Of 15 positions that distinguish the two variable regions, the DNA matches λ_{II} (MOPC 315) at 11 and λ_I (MOPC 104E) at 4. The actual sequences are given in Fig. 3.

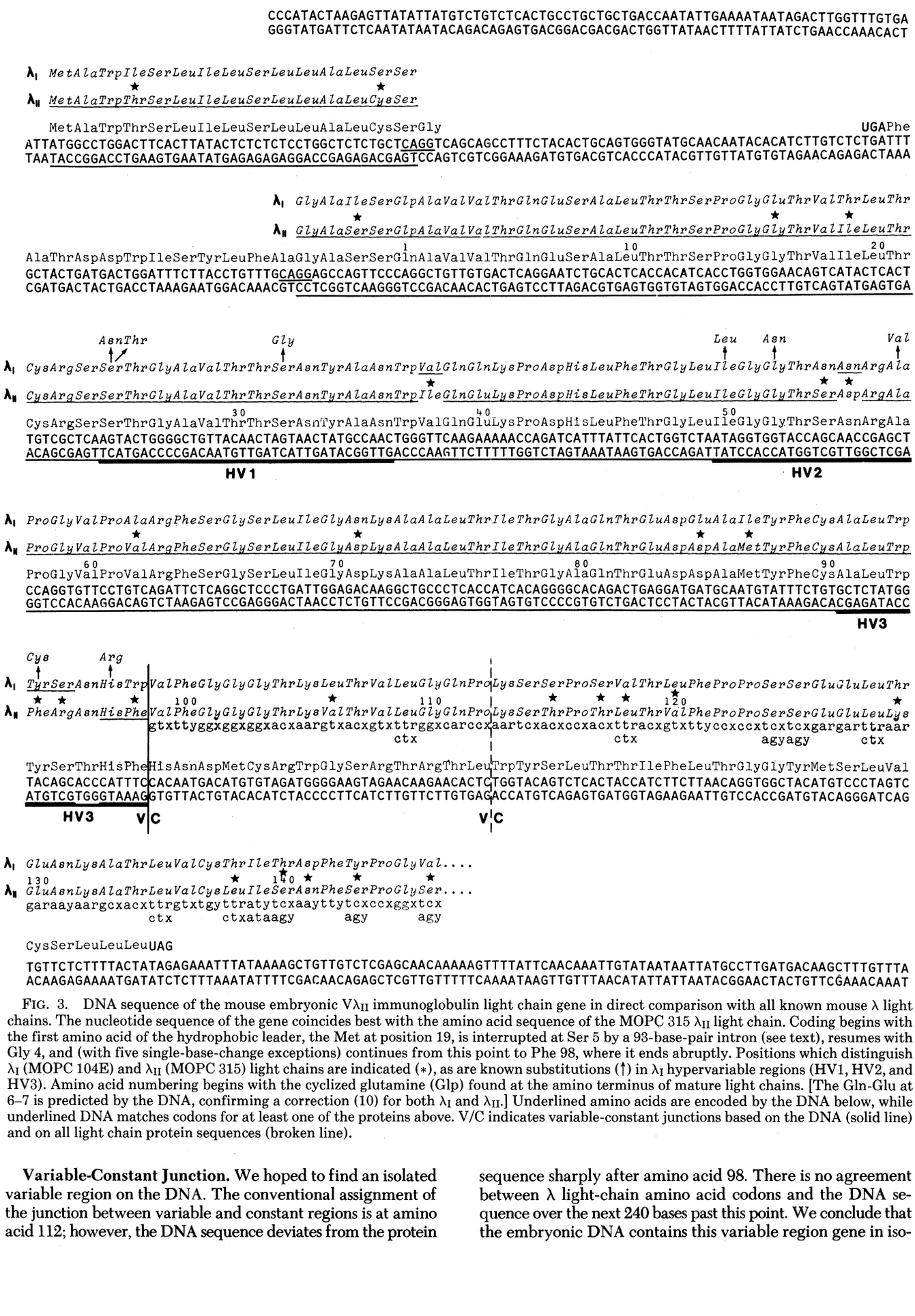

FIG. 3. DNA sequence of the mouse embryonic $V\lambda_{II}$ immunoglobulin light chain gene in direct comparison with all known mouse λ light chains. The nucleotide sequence of the gene coincides best with the amino acid sequence of the MOPC 315 λ_{II} light chain. Coding begins with the first amino acid of the hydrophobic leader, the Met at position 19, is interrupted at Ser 5 by a 93-base-pair intron (see text), resumes with Gly 4, and (with five single-base-change exceptions) continues from this point to Phe 98, where it ends abruptly. Positions which distinguish λ_I (MOPC 104E) and λ_{II} (MOPC 315) light chains are indicated (*), as are known substitutions (↑) in λ_I hypervariable regions (HV1, HV2, and HV3). Amino acid numbering begins with the cyclized glutamine (Glp) found at the amino terminus of mature light chains. [The Gln-Glu at 6–7 is predicted by the DNA, confirming a correction (10) for both λ_I and λ_{II}.] Underlined amino acids are encoded by the DNA below, while underlined DNA matches codons for at least one of the proteins above. V/C indicates variable-constant junctions based on the DNA (solid line) and on all light chain protein sequences (broken line).

Variable-Constant Junction. We hoped to find an isolated variable region on the DNA. The conventional assignment of the junction between variable and constant regions is at amino acid 112; however, the DNA sequence deviates from the protein sequence sharply after amino acid 98. There is no agreement between λ light-chain amino acid codons and the DNA sequence over the next 240 bases past this point. We conclude that the embryonic DNA contains this variable region gene in iso-

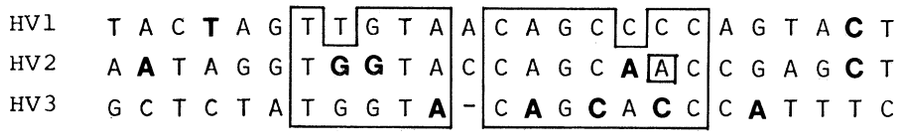

FIG. 4. Double and triple DNA sequence homologies in the $V\lambda_{II}$ gene corresponding to immunoglobulin light chain hypervariable regions. The bottom strand of HV1 and top strands of HV2 and HV3 (Fig. 3) have been aligned 5′ to 3′, with the bases that could change to produce known λ_I and inferred λ_{II} amino acid substitutions shown in boldface.

lation from the constant region. No obvious element identifies the boundary as unusual; we cannot interpret a feature of the DNA sequence that shows how the structures are brought together. The identification of amino acid 112 as the variable-constant junction is based on analogy with human sequences. We do not know, for the mouse λ_{II} chains, that the junction should be there: we interpret our sequence as showing that the variable-constant junction of λ_{II} gene follows amino acid 98.

Hypervariable Regions. Sequences corresponding to hypervariable regions appear in this germ-line gene. The first hypervariable region in the DNA sequence fits both the λ_I and λ_{II} proteins (Fig. 3). The second hypervariable region corresponds to λ_{II} at position 54 (Ser rather than Asn) to λ_I at 55 (Asn rather than Asp) and to both elsewhere. However, at the third hypervariable region, the MOPC 315 protein deviates from our DNA sequence at three amino acids, two of which correspond to λ_I chains. Our finding that the hypervariable regions are in the germ-line DNA rules out models in which DNA is inserted into a pre-existing germ-line sequence. Differences among mouse λ light chains in the hypervariable regions are thus most likely due to the gene for the final protein having accrued one or several mutations before or during the antigen-driven expansion of the pool of precursor cells specialized to make that particular immunoglobulin.

There are no features in these hypervariable regions that are so dramatic as to define a mechanism by which these regions might be more labile than the rest of the DNA. We do not find extensive palindromes surrounding these hypervariable regions (14, 15). (Of the four palindromes suggested in ref. 14, we find only one in the second hypervariable region.) There is, however, a similarity in the DNA structure of the three hypervariable regions. Fig. 4 matches the bottom strand of the first hypervariable region with the top strand of the second and third. These similarities may be simply evolutionary relics, but it is not impossible that they could serve as a recognition site for an enzyme that would cleave or modify the DNA in order to make the sequence labile.

At three positions in the hypervariable regions and one in the framework, this $V\lambda_{II}$ gene deviates from λ_{II} (MOPC 315) expectations and agrees with the most common λ_I residues. This may reflect a past duplication that gave rise to both the λ_I and λ_{II} germ-line genes.

$V_{\lambda_{II}}$ Gene Contains a Genetic Discontinuity—93 Nontranslated Bases in Leader Region. The light chain is synthesized as a longer precursor (16, 17), containing at its amino terminus a hydrophobic peptide that, presumably, is involved in moving this protein through the cell membrane (18, 19). As the protein matures, this precursor is cleaved at a glutamine (20), which cyclizes into a pyrrolidone carboxylic acid (10). Sequences of the 19-amino-acid long signal peptides for mouse λ light chains have been worked out by Schecter and his coworkers (20, 21) by translating purified immunoglobulin messenger *in vitro*. The DNA, however, codes for only four of the expected amino acids preceding the initial glutamine (Fig. 3) and then deviates entirely from the precursor protein. Ninety-three bases earlier in the DNA, we find a sequence for the first 15 amino acids of the light chain precursor, including an ATG for the initial methionine. When this match was first made, only the λ_I precursor sequence was available, and that deviates from the DNA in three places. However, after the DNA sequence was established, Burstein and Schecter (ref. 20; personal communication) worked out the hydrophobic leader sequence for the λ_{II} light chain; that protein sequence agrees with the DNA at all points. Therefore the DNA sequence was a predictor of the protein sequence. The extra region of 93 bases cannot be translated into protein; it contains a stop signal in phase. This 93-base region is not an artifact of the cloning. It contains a unique *Pst* I cut that can be exhibited in uncloned embryonic DNA (M. Hirama and S. Tonegawa, unpublished).

We believe the functioning gene in the myeloma will consist of the precursor region followed first by the 93-base-long interspersed DNA, followed by the variable region gene, then by a 1250-base-long piece of noncoding DNA (4), and last, by the constant region gene. We call such an additional piece of DNA that arises within a gene an *intron* (for intragenic region or intracistron) and thus look upon the structure of this gene as leader(45)-intron(93)-variable (306)-intron(1250)-constant(348). The current level of analysis cannot exclude still more small introns within the constant region nor specify the exact location of the second intron.

Maturation. Does the sequence for the 93-base intron show us how this area can be skipped or excised from the message? There is only a short repetition at its ends, a CAGG sequence repeated in the correct phase. A four-base repeat, however, is not specific enough to be used by the RNA polymerase as a signal to stop and start up again on an RNA-priming model. The signals to eliminate the unwanted region may be contained in the base sequence and structure of the messenger. A hairpin structure could hold the point of splicing in its stem, but that would necessitate ligation from one chain across to the opposite side of the helix. Fig. 5 shows the one well-placed hairpin we find in the sequence; it is neither convincing nor impossible. An alternative structure would be to use a sequence elsewhere in the messenger as a template to bring into juxtaposition the bonds that are to be split and rejoined.

Genes in Pieces. We hypothesize that most higher cell genes will consist of informational DNA interspersed with silent sequences. The eukaryotic cistron is a transcription unit containing alternate regions to be excised from the messenger, the introns, and regions left to be expressed, exons. [There are many recent examples of this structure (4, 22–26); others are reviewed in ref. 27.] Thus the gene is a mosaic: sequences corresponding to expressed functions held in a matrix, the latter possibly 10 times larger than the coding components. This model immediately resolves two puzzles. Heterogeneous nuclear RNA would be the long transcription products from which the much smaller ultimate messengers are spliced. The extra DNA in higher cells would arise because the genes, the transcription units, are much larger than those for a single polypeptide product.

Assume that the splicing mechanism is general, independent of the gene structure, and recognizes simply a unique secondary structure in the RNA. Of what benefit then are the infilling sequences? They speed evolution. Single base changes not only

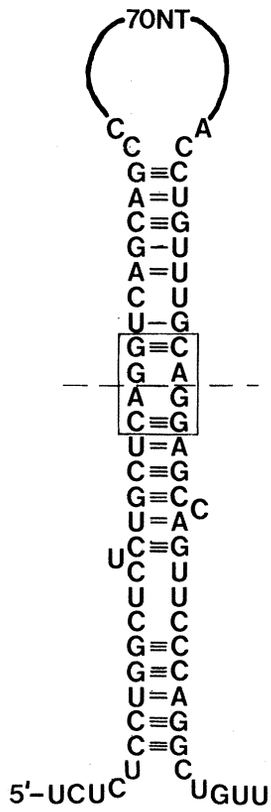

FIG. 5. Hypothetical hairpin structure in λ_{II} light chain precursor mRNA with the nontranslated sequences to be excised (intron) above the dashed line and coding sequences to be joined (exon) below. The base-paired sequences are from the ends of the 93-base noncoding segment in Fig. 3. Proper splicing of this precursor RNA would require a symmetric pair of cuts in the antiparallel sequences boxed in the stem, followed by ligation of the strands below to preserve one copy of the CAGG sequence. The strands to be joined, however, are not favorably oriented for ligation.

can alter single amino acids, but now, if they occur at the boundaries of introns, can change the splicing to add or delete a string of amino acids. This generates a more rapid search through the space of protein molecules. Second, the splicing process need not be 100% efficient. For example, changes in silent positions, third base positions in codons, can alter both the pattern and the efficiency of splicing so that the product of a single transcription unit can be two polypeptide chains, one being the original gene product and the second, also synthesized at high frequency, the new product. Evolution can seek new solutions without destroying old. This resolves a classic problem: one thought that the organism had to create a second copy of an essential gene in order to mutate it to a new function. The intronic model eliminates any special duplication. The extra material, scattered widespread across the genome, can be called into action at any time. After a new gene function appears, there can be selective pressure for duplication. One consequence of the intron model is that the dogma of one gene, one polypeptide chain disappears. A gene, a contiguous region on DNA, now corresponds to one transcription unit, but that transcription unit can correspond to many polypeptide chains, of related or differing functions.

Since the gene is now spread out over a larger region of DNA, recombination between exons will be enhanced. Furthermore, if the exons correspond to identifiable functions put together by splicing to form a special combination in a finished protein, then recombination between these regions can sort their functions independently. In the λ light chain a hydrophobic precursor sequence is separated by an intron from its follower region; recombination could combine this leader sequence with some other protein. One might anticipate that middle repetitive sequences within introns will provide hot spots for recombination to reassort the exonic sequences.

Still another picture emerges if we hypothesize that specific new splicing patterns can be turned on by special gene products, providing developmental control. A differentiation pathway can be defined by the appearance of a new splicing enzyme.

One striking interpretation, with this new picture, would be that the simultaneous expression of IgM and IgD with the same idiotype by a single cell (28) will turn out to be due to a V_H gene translocated near the C_H genes and transcribed together with C_μ and C_δ into a single V_H-μ-δ precursor, alternate splicings then providing the two products. The switch from IgM to IgG may be a new translocation of a V_H gene, or it may be a new processing of a V_H-μ-δ-γ precursor to produce a V_H-γ product.

We are grateful for the expert assistance of Rita Lenhard-Schuller. We thank I. Schecter for the communication of results before publication and P. Slonimsky, B. Blomberg, D. Wiley, and S. Harrison for discussions. W.G. is an American Cancer Society Research Professor. Part of this work was supported by National Institutes of Health Grant GM 09541-16.

1. Wu, T. T. & Kabat, E. A. (1970) *J. Exp. Med.* **132,** 211–250.
2. Cohn, M., Blomberg, B., Geckeler, W., Raschke, W., Riblet, R. & Weigert, M. (1974) in *The Immune System; Genes, Receptors, Signals*, eds. Sercarz, E. E., Williams, A. R. & Fox, C. F. (Academic, New York), pp. 89–117.
3. Hozumi, N. & Tonegawa, S. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 3628–3632.
4. Brack, C. & Tonegawa, S. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5652–5656.
5. Wu, T. T., Kabat, E. A. & Bilofsky, H. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 5107–5110.
6. Tonegawa, S., Brack, C., Hozumi, N. & Schuller, R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3518–3522.
7. Tonegawa, S., Brack, C., Hozumi, N. & Pirrotta, V. (1977) in *Cold Spring Harbor Symp. Quant. Biol.*, in press.
8. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
9. Roberts, R. J. (1977) in *Critical Reviews in Biochemistry,* ed. Fasman, G. D. (CRC Press, Cleveland, OH), pp. 123–164.
10. Weigert, M. G., Cesali, I. M., Yonkovich, S. J. & Cohn, M. (1970) *Nature* **228,** 1045–1047.
11. Apella, E. (1971) *Proc. Natl. Acad. Sci. USA* **68,** 590–594.
12. Cesari, I. M. & Weigert, M. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 2112–2116.
13. Dugan, E. S., Bradshaw, R. A., Simms, E. S. & Eisen, H. N. (1973) *Biochemistry* **12,** 5400–5416.
14. Leder, P., Honjo, T., Seidman, J. & Swan, D. (1976) *Cold Spring Harbor Symp. Quant. Biol.* **41,** 855–862.
15. Wuilmart, C., Urbain, J. & Givol, D. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 2526–2530.
16. Milstein, C., Brownlee, G., Harrison, T. M. & Mathews, M. B. (1972) *Nature New Biol.* **239,** 117–120.
17. Swan, D., Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 1967–1972.
18. Blobel, G. & Dobberstein, B. (1975) *J. Cell Biol.* **67,** 835–851.
19. Blobel, G. & Dobberstein, B. (1975) *J. Cell Biol.* **67,** 852–862.
20. Burstein, Y. & Schecter, I. (1977) *Biochem. J.* **165,** 347–354.
21. Burstein, Y. & Schecter, I. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 716–760.
22. Berget, S. M., Moore, C. & Sharp, P. A. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 3171–3175.
23. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. (1977) *Cell* **12,** 1–8.
24. Klessig, D. F. (1977) *Cell* **12,** 9–21.
25. Breathmark, R., Mandel, J. L. & Chambon, P. (1977) *Nature* **270,** 314–319.
26. Doel, M. T., houghton, M., Cook, E. A. & Carey, N. H. (1977) *Nucleic Acids Res.* **4,** 3701–3713.
27. Williamson, B. (1977) *Nature* **270,** 295–297..w
28. Rowe, D. S., Hug, K., Forni, L. & Permis, B. (1973) *J. Exp. Med.* **138,** 965–972.