

# Multiple related immunoglobulin variable-region genes identified by cloning and sequence analysis

( $\kappa$  light chain/subgroups/DNA sequencing)

J. G. SEIDMAN, AYA LEDER, MARSHALL H. EDGELL, FRED POLSKY, SHIRLEY M. TILGHMAN, DAVID C. TIEMEIER, AND PHILIP LEDER

Laboratory of Molecular Genetics, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20014

Communicated by DeWitt Stetten, Jr., May 18, 1978

**ABSTRACT** We have identified at least six *EcoRI* fragments of mouse DNA that encode variable-region gene sequences closely related to the mouse  $\kappa$  light chain, MOPC-149. Two of these fragments have been cloned, and the entire nucleotide sequence of the variable-region genes encoded on each has been determined. Both genes encode closely related variable-region sequences extending from codon position 1 through position 97. Neither fragment encodes a constant-region sequence. Although both genes are closely related, they differ from one another and from the sequence expressed in the MOPC-149 cell from which they were cloned. These few differences cluster within the complementarity-determining regions although several occur in framework sequences as well. We therefore conclude that an antibody-producing cell contains genetic information corresponding to its expressed sequence and several other closely related but silent sequences. These initial results raise the possibility that similar sets of genes might exist corresponding to each of the many subgroups already identified among mouse  $\kappa$  light chains. If true, this would further suggest that the mouse genome might be rich enough in variable-region genes so as to encode a major portion of the variable-region repertoire.

The unique pattern of antibody structure led Dreyer and Bennett (1) to propose that the constant and variable regions of the immunoglobulin light chain are encoded separately in chromosomal DNA. Their argument has been supported by the finding that there are few copies of light chain constant-region genes (2-14) and, more recently, by evidence derived directly from cloned immunoglobulin gene fragments (15, 16). It has further been suggested that variable- and constant-region gene sequences are rearranged during the somatic differentiation of immunocytes so as to bring constant- and variable-region genes together (17, 18). The question of whether this rearrangement of genes is essential for their expression remains unanswered. Also unanswered is the important question of how the diversity represented in immunoglobulin variable regions is produced—through evolutionary or somatic mechanisms or a combination of both. Several useful models that define contending germ-line and somatic mutation hypotheses have been advanced to explain this interesting genetic phenomenon (19-31).

In order to begin to distinguish between these models (the germ-line model requires many variable region genes; the somatic model, few), we have cloned (32) two closely related  $\kappa$  variable-region genes derived from the mouse plasmacytoma, MOPC-149. We have determined the entire nucleotide sequence of the coding segments and portions of the flanking sequences of both these genes and compared them to one another and to a cloned cDNA sequence corresponding to a variable-region gene expressed in this cell line. Among other

inferences we may draw from the structure of these genes, the identification of three closely related genetic sequences in a committed, immunoglobulin-producing cell allows us to begin to estimate a minimum number for variable regions encoded in the mouse genome. Such estimates (see below) indicate that the mouse genome is likely to be a repository of many  $\kappa$ -type variable-region genes, so many as to possibly encode a major portion of the variable-region repertoire.

## MATERIALS AND METHODS

**Preparation and Analysis of *EcoRI* Fragments of MOPC-149 DNA.** The *EcoRI* restriction fragments from 50 mg of MOPC-149 DNA were fractionated by RPC-5 column chromatography (33) and subsequently by agarose gel electrophoresis (34). The DNA fragments, embedded in agarose gels, were denatured in alkali, neutralized, and transferred to nitrocellulose sheets as described by Southern (35). The nitrocellulose sheets were presoaked in Denhardt's solution (36) (0.02% Ficoll/0.02% pyrrolidine/0.02% bovine serum albumin) for 3 hr and hybridized with an appropriate <sup>32</sup>P-labeled probe. The hybridization solution contained 10 ng of DNA (labeled by the nick translation reaction to 40-80 cpm/pg) per ml, 0.1% sodium dodecyl sulfate (NaDodSO<sub>4</sub>), 0.09 M sodium citrate/0.9 M NaCl, triple-strength Denhardt's solution, and 50  $\mu$ g of *Escherichia coli* DNA and 50  $\mu$ g of salmon sperm DNA per ml. After hybridization, the sheets were washed four times in 0.1% NaDodSO<sub>4</sub>/1.5 mM sodium citrate/15 mM NaCl at 53° for 30 min per wash and twice more in the same solution without NaDodSO<sub>4</sub> and then were exposed to Kodak XR1 film backed by Cronex intensifier screens. Preparative fractionation of appropriate RPC-5 fractions was carried out by using an electronic preparative electrophoresis apparatus (34).

**Cloning and Identification of Variable-Region Gene-Containing Hybrid Phage.** Purified *EcoRI* fragments were ligated into the EK2 vector,  $\lambda$ gtWES- $\lambda$ B, transfected into the EK2 host *E. coli* LE392 as described (32, 33). Immunoglobulin-containing fragments were identified by *in situ* hybridization of nitrocellulose filter blots of transfection plates containing 500-2000 plaques (37). Hybrid phage were selected and grown preparatively on *E. coli* DP50supF. For sequence determination, the entire K2 *EcoRI* fragment [3.0 kilobases (kb)] was subcloned in pMB9 and a 4.1-kb *HindIII* fragment of K3 containing the entire variable-region sequence was subcloned in pBR322 (38) in *E. coli*  $\chi$ 1776. Procedures used conformed to the *NIH Guidelines for Recombinant DNA Research*.

**Restriction Fragments from Different Parts of the Light Chain Structural Gene.** The plasmid pCR1-K40 containing 700 base pairs of the MOPC-149 light chain mRNA sequence has been described (39). Hereafter, this plasmid will be called K149 for simplicity. The light chain sequences in K149 corre-

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.

Abbreviations: NaDodSO<sub>4</sub>, sodium dodecyl sulfate; kb, kilobases.

spond to sequences encoding amino acid 44 through the 3' end of the MOPC-149 light chain mRNA. The *Hha* I fragment (39) containing all of these sequences was used as the "V + C" probe. The variable-region probe was obtained by cloning (using synthetic oligonucleotide linkers containing the *Bam*HI site that were purchased from Collaborative Research, Cambridge, MA) a *Hae* III fragment to *Hinc*II fragment of K149 DNA in pBR322 (38). This fragment contained immunoglobulin sequences that encode amino acids 44–90. Two constant-region fragments, a *Hinc*II (amino acid 125)-to-*Hae* III (amino acid 196) fragment and an 800-base-pair *Hae* III fragment (containing sequences encoding amino acid 196 through the 3' untranslated region) were also cloned separately in pBR322.

**Electron Microscopic Analysis of R Loops.** MOPC-149 mRNA (15 g/ml) (13) was annealed with the cloned *Eco*RI fragments (50  $\mu$ g/ml) in 70% formamide/0.5 M NaCl/0.1 M *N*-[tris(hydroxymethyl)methyl]glycine (Tricine), pH 8.0, at 52° for 12 hr. The samples were spread, stained, shadowed, visualized and measured as described (40).

**Restriction Site Analysis and Sequence Determination.** Restriction sites were determined (32, 39) and the nucleotide sequence was determined as described (41).

## RESULTS AND DISCUSSION

**Identification of Variable-Region Gene-Containing Fragments.** One of the major problems encountered in quantitating mouse  $\kappa$ -type variable-region genes arises from the very diversity that these genes appear to encode. A hybridization probe generated from an mRNA corresponding to one  $\kappa$ -light chain must encode a sequence so different from other known  $\kappa$  light chains that the extent of cross hybridization between such diverse genes, the "range" of the probe, is difficult to assess. Thus, even if there were hundreds or thousands of variable-region genes in the mouse genome, a probe directed against any

one of them would be expected to detect only those sequences that are closely homologous. If extensive somatic mutational mechanisms generated diversity from a single sequence, there might be only one such gene per genome. On the other hand, if extensive diversity is already encoded into the mouse genome by evolutionary mechanisms, there would be many.

DNA restriction fragments bearing elements of these genes can be identified by using *in situ* hybridization (35) in conjunction with well-defined, pure hybridization probes. We have used recombinant DNA techniques to prepare variable- and constant-region hybridization probes derived from a cloned, reverse transcript of MOPC-149 mRNA (39). The variable-region probe corresponds to MOPC-149 mRNA from amino acid codons 44 through 90. Cloned probes corresponding to the constant region from codons 125 through 196 and 196 through the 3'-untranslated region were also prepared. The variable-region probe was used to detect *Eco*RI fragments of plasmacytoma MOPC-149 DNA that carried related variable-region gene sequences, and analogous studies were carried out with the constant-region probe.

Due to the complexity of the mouse genome and expected complexity of the variable-region sequences, an *Eco*RI digest of genomic DNA was separated in two dimensions—RPC-5 column chromatography and agarose gel electrophoresis—in order to carry out high-resolution analysis while also purifying appropriate fragments for subsequent cloning steps. After the second dimension of fractionation (the analytical gel), the DNA fragments were transferred to nitrocellulose filters (35) and hybridized to a <sup>32</sup>P-labeled variable-region probe. At least six discrete *Eco*RI fragments hybridized to the variable region probe, four quite strongly (Fig. 1). The fragments ranged in size from approximately 3 to 13 kb and the largest and smallest (labeled K2 and K3 in Fig. 1) were selected for further purification by preparative gel electrophoresis (34).

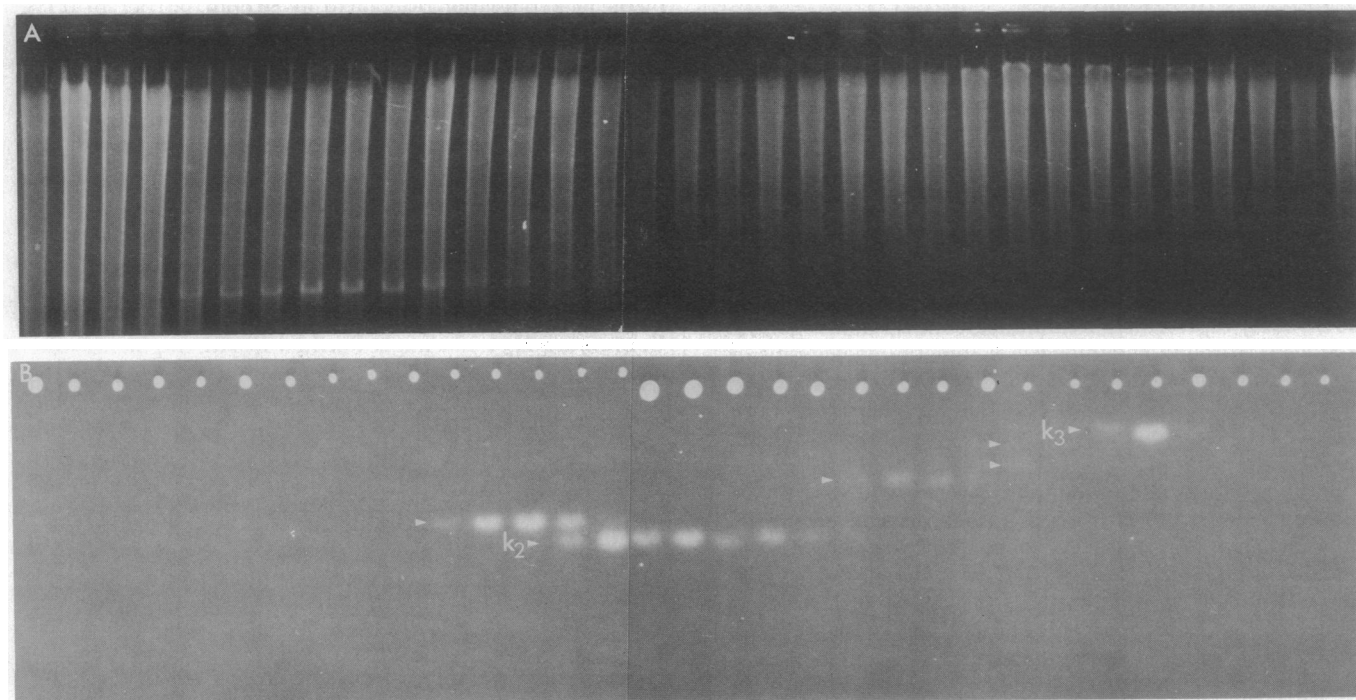


FIG. 1. Identification of *Eco*RI fragments from MOPC-149 DNA that carry variable-region gene sequences cross-hybridizing to a MOPC-149 variable-region probe. (A) The *Eco*RI-digested mouse DNA was fractionated first by RPC-5 column chromatography and then by agarose gel electrophoresis and stained with ethidium bromide. (B) The fragments were transferred to a nitrocellulose filter and hybridized to a variable region probe and identified by autoradiography. The variable-region probe contained sequences that encode amino acids 44–90 of the MOPC-149 variable region (39). The K2 (3 kb) and K3 (13 kb) fragments are indicated. Six fragments were easily visualized on the original radioautograph; two (just under K3) are only faintly seen in the reproduction.

**Cloning the Variable-Region Gene-Containing Fragments.** The observation that at least six *EcoRI* DNA fragments hybridized to the MOPC-149 variable-region sequence raised the possibility that each of these fragments might contain elements of one or more variable-region genes and, moreover, that these fragments might be so closely related to one another so as to constitute a subgroup of individually encoded, but closely related, variable-region genes. To test this possibility and to determine the molecular basis for the observed hybridization, fragments K2 (3 kb) and K3 (13 kb) (Fig. 1), which had been purified approximately 1000-fold, were cloned in the  $\lambda$ gtWES- $\lambda$ B EK2 vector system and detected directly on transfection plates by using nitrocellulose filter blots and *in situ* hybridization with a variable-region probe. Three independent (nonsibling) clones were detected among about 5000 plaques in the case of K2 and two independent clones were detected among 2000 plaques in the case of K3.

**Both Cloned Fragments Encode Only a Single Variable-Region Gene Sequence Related to MOPC-149.** That each cloned fragment contained variable-region sequences, but no constant-region sequence, was determined by *in situ* hybridization of each to cloned, isolated fragments of the MOPC-149 cDNA clone corresponding to separated variable- and constant-region sequences (data not shown). [Because no *EcoRI* site occurs between the variable- and constant-region portions of the MOPC-149 mRNA sequence (39), these sequences must not be contiguous with the constant-region gene in MOPC-149 genomic DNA.] That only a single homologous variable-region gene sequence was present in each cloned fragment was demonstrated by annealing MOPC-149 light chain mRNA to each fragment and visualizing the resultant R loops (40) in the electron microscope. Each fragment contained a single, uninterrupted R loop (approximately 300 base pairs in length) from which hung an RNA tail (Fig. 2). That the tail corresponded

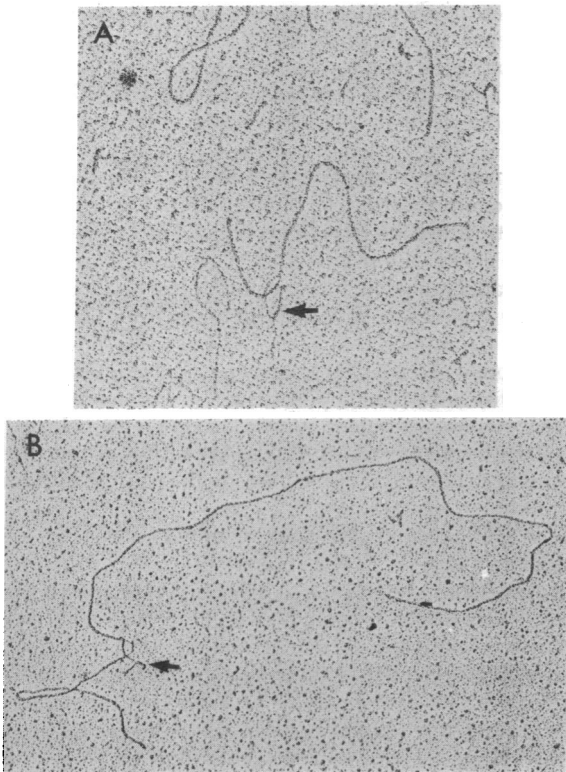


FIG. 2. R loops formed between the cloned K2 and K3 fragments and MOPC-149 light chain mRNA. (A) The 3-kb K2 fragment; (B) the 13-kb K3 fragment.

to the 3' (or constant-region) portion of the mRNA was demonstrated by annealing simian virus 40 molecules covalently linked to poly(T) tails and visualizing the resulting structure (electron micrographs not shown) and by the sequence analysis described below. Thus, both cloned fragments were judged to contain one MOPC-149-like variable-region sequence but no constant-region sequence.

**Nucleotide Sequences of the Two Cloned Variable-Region Genes and MOPC-149 Expressed Sequence: Silent Genes.** The orientation and extent of the variable-region gene encoded in each fragment was determined by direct sequence analysis by using the procedure of Maxam and Gilbert (41). The restriction sites and fragments used to determine and check both sequences are shown in Fig. 3. The entire sequence of the K2 variable-region gene, including flanking sequences, 497 bases in all, was determined and is shown in Fig. 4. The entire sequence of the K3 variable-region gene including flanking sequences was also determined and is compared to the K2 sequence in the same figure.

A variable region gene is easily distinguished within each determined sequence by identifying in-phase codons corresponding to the initial pentapeptide of the MOPC-149 light chain, Asp-Ile-Gln-Met-Thr (labeled from position 1 in Fig. 4; ref. 42) and following this reading frame through 97 codon positions in both gene sequences. The amino acid sequences generated by each gene are in substantial (but not complete) agreement (points of sequence divergence are discussed below) with the initial 23 amino acid positions that have been determined for six  $\kappa$  light chain polypeptides, MOPC-31B, MOPC-178, MOPC-31C, TEPC-173, RPC-23, and MOPC-149 (43).

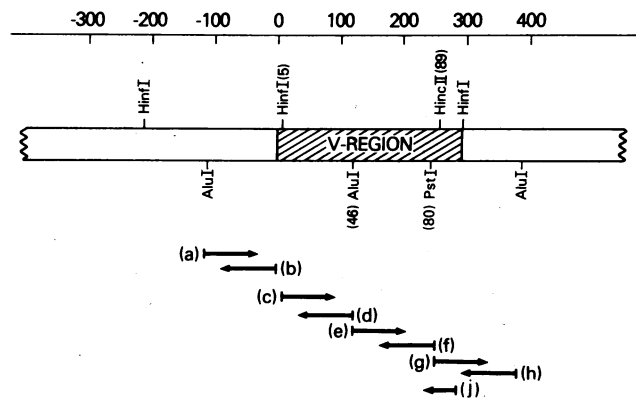


FIG. 3. Sequencing strategy used to determine the variable region gene sequences in K2 and K3. A 1800-base-pair *Hae* III fragment was isolated from pMB9-K2 and a 1200-base-pair *Hae* III fragment was cleaved from pBR322-K3 which contained the entire variable-region genes found on the K2 and K3 fragments. These *Hae* III fragments were cleaved with *Hinc*II, *Alu* I, or *Hinf*I, end-labeled with  $^{32}$ P by the procedure of Maxam and Gilbert (41), and then further cleaved with a second enzyme. The two  $^{32}$ P-labeled ends were separated on a polyacrylamide gel, eluted, and subjected to the nucleotide sequencing reaction described by Maxam and Gilbert (41). Specifically, a and d were obtained by cleavage of an *Alu* I fragment by the action of *Hinf*I, b was derived by cleavage of a *Hinf*I fragment with *Alu* I, e and h were ends of an *Alu* I fragment cleaved with *Pst* I, and f and g were obtained by cleaving two *Hae* III/*Hinc*II fragments with *Alu* I. The ends (g and h) were not used to determine the K3 sequence; instead, the K3 sequence was determined by cleavage of the *Hinc*II fragment with *Pst* I (j). The numbers above the figure give the approximate length of the fragments (expressed in base pairs) and the numbers next to some of the restriction sites refer to the amino acid location of the cleavage site. The length of nucleotide sequence determined from each end is indicated by the arrows. The nucleotide sequence at each restriction site, used as an end, was taken from published work, except that for the *Hinc*II site the recognition sequence was shown to be GTCAAC by end-group analysis.



insertion and corrected by a single base deletion. Of the 336 coding bases compared between both genes, 314 are identical.

A similar relationship emerges from a comparison of K2 and K3 genes to a third variable-region gene sequence that must be present in MOPC-149 genomic DNA, the expressed MOPC-149 mRNA sequence. Of the 44 codons compared between MOPC-149 and K2, two coding differences occur within the seven-codon-long second complementarity-determining region and one within the 36 framework codons compared. The same region compared between MOPC-149 and K3 reveals six coding differences, four of which occur within the seven-codon-long second complementarity-determining region. Of the 132 bases compared, 125 are identical between K2 and MOPC-149; 120 are identical between K3 and MOPC-149. Differences between the genes exist in framework sequences but appear to cluster in the complementarity-determining region.

**A New Definition of Variable-Region Subgroup.** From the foregoing we may conclude that a probe corresponding to a given variable-gene sequence detects other closely related variable-gene sequences. The three gene sequences characterized here occur within distinct and, in the case of K2 and K3, nonallelic DNA fragments present both in plasmacytoma and embryonic DNA (unpublished data). It is likely that, of the four other fragments identified, at least the two strongly hybridizing ones represent additional but also closely related variable-region gene sequences. This variable-region probe thus appears to be detecting a family of closely related variable-region genes and provides a working definition of a subgroup of variable-region genes here consisting of at least six distinct but related (by hybridization homology under standard conditions) genes.

**Germ-Line or Somatic Mutation or Both?** If each  $\kappa$  light chain representing a subgroup were to identify a discrete set of five genomic genes, the 25–30 subgroups already identified from amino acid sequence studies would already account for 125–150 distinct variable-region genes. The rate at which distinct  $\kappa$  light chains are being discovered in BALB/c plasmacytomas indicates that the variable region isotopes could easily approach 100 (M. Potter, personal communication). This suggests that genomic DNA, both germ-line and committed cell, will turn out to be quite rich in variable-region genes.

Thus, a closely related set of multiple germ-line genes—although encoding the bulk of immunoglobulin diversity—might be unstable in both germ-line and somatic cells. Clearly, the analysis of additional cloned genes derived from undifferentiated or plasmacytoma cells should allow us to decide if this is so. In the meantime, it seems most likely that evolution has played a major role in generating diverse variable-region genes, but the possibility that somatic mutation may further intensify this diversity cannot yet be ruled out.

We are grateful to Barbara Norman and Marion Nau for their many contributions to this work. We are also grateful to Margery Sullivan for her able advice and assistance in our electron microscopic studies and to Terri Broderick for her expert assistance in the preparation of this manuscript. We are also grateful to Dr. Michael Potter who provided the original plasmacytomas used in these studies.

1. Dreyer, W. J. & Bennett, J. C. (1965) *Proc. Natl. Acad. Sci. USA* **54**, 864–868.
2. Faust, C. H., Diggelman, H. & Mach, B. (1974) *Proc. Natl. Acad. Sci. USA* **2491–2495**.
3. Honjo, T., Packman, S., Swan, D., Nau, M. & Leder, P. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 3659–3663.
4. Leder, P., Honjo, T., Packman, S., Swan, D., Nau, M. & Norman, B. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 5109–5114.
5. Stavnezer, J., Huang, R. C. C., Stavnezer, E. & Bishop, J. M. (1974) *J. Mol. Biol.* **88**, 43–63.
6. Storb, U. (1974) *Biochem. Biophys. Res. Commun.* **57**, 31.
7. Tonegawa, S., Steinberg, C., Dube, S. & Bernadini, A. (1974) *Proc.*

- Natl. Acad. Sci. USA* **71**, 4027–4031.
8. Leder, P., Honjo, T., Swan, D., Packman, S., Nau, M. & Norman, B. (1976) in *Molecular Approaches to Immunology*, eds. Smith, E. E. & Ribbons, D. W., (Academic, New York), pp. 173–188.
9. Rabbitts, T. H., Jarvis, J. M. & Milstein, C. (1975) *Cell* **6**, 5–12.
10. Rabbitts, T. H. & Milstein, C. (1975) *Eur. J. Biochem.* **52**, 125–133.
11. Farace, M.-G., Aellen, M.-F., Briand, P.-A., Faust, C. H., Vassalli, P. & Mach, B. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 727–731.
12. Honjo, T., Packman, S., Swan, D. & Leder, P. (1976) *Biochemistry* **15**, 2780–2785.
13. Honjo, T., Swan, D., Nau, M., Norman, B., Packman, S., Polsky, F. & Leder, P. (1976) *Biochemistry* **15**, 2775–2779.
14. Tonegawa, S. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 203–207.
15. Tonegawa, S., Brack, C., Hozumi, N. & Schuller, R. (1977) *Proc. Natl. Acad. Sci. USA* **8**, 3518–3522.
16. Brack, C. & Tonegawa, S. (1977) *Proc. Natl. Acad. Sci. USA* **12**, 5652–5655.
17. Hozumi, N. & Tonegawa, S. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 3628–3632.
18. Tonegawa, S., Hozumi, N., Matthysens, G. & Schuller, R. (1976) *Cold Spring Harbor Symp. Quant. Biol.* **41**, 877–889.
19. Burnet, F. M. (1959) *The Clonal Selection Theory of Acquired Immunity* (Cambridge Univ. Press, New York).
20. Lederberg, J. (1959) *Science* **129**, 1649–1653.
21. Szilard, L. (1960) *Proc. Natl. Acad. Sci. USA* **46**, 293–302.
22. Smithies, O. (1963) *Nature* **199**, 1231–1236.
23. Brenner, S. & Milstein, C. (1966) *Nature* **211**, 242–243.
24. Edelman, G. M. & Gally, J. A. (1967) *Proc. Natl. Acad. Sci. USA* **57**, 353–358.
25. Lennox, E. S. & Cohn, M. (1967) *Annu. Rev. Biochem.* **36**, 365–402.
26. Smithies, O. (1967) *Science* **157**, 267–273.
27. Gally, J. A. & Edelman, G. M. (1970) *Nature* **227**, 341–348.
28. Jerne, N. K. (1970) in *Immune Surveillance*, eds. Smith, R. T. & Landy, M. (Academic, New York), pp. 345–363.
29. Cohen, M. (1971) *Ann. N.Y. Acad. Sci.* **190**, 529–584.
30. Jerne, N. K. (1971) *Eur. J. Immunol.* **1**, 1–9.
31. Brown, D. D. (1972) in *Molecular Genetics and Development Biology*, ed. Sussman, M. (Prentice-Hall, Englewood Cliffs, NJ), pp. 101–125.
32. Tilghman, S. M., Tiemeier, D. C., Polsky, F., Edgell, M. H., Seidman, J. G., Leder, A., Enquist, L. W., Norman, B. & Leder, P. (1977) *Proc. Natl. Acad. Sci. USA* **10**, 4406–4410.
33. Tiemeier, D. C., Tilghman, S. M. & Leder, P. (1977) *Gene* **2**, 173–191.
34. Polsky, F. I., Edgell, M., Seidman, J. G. & Leder, P. (1978) *Anal. Biochem.*, **87**, 397–410.
35. Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503–517.
36. Denhardt, T. (1966) *Biochem. Biophys. Res. Commun.* **23**, 641–646.
37. Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180–182.
38. Bolivar, E., Rodriguelez, R. L., Green, P. J., Behach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H. & Falkow, S. (1977) *Gene* **2**, 95–113.
39. Seidman, J. G., Edgell, M. H. & Leder, P. (1978) *Nature* **271**, 582–585.
40. White, R. L. & Hogness, D. S. (1977) *Cell* **10**, 177–192.
41. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
42. Hood, L., McKean, O., Farnsworth, V. & Potter, M. (1973) *Biochemistry* **12**, 741–749.
43. Kabat, E. A., Wu, T. T. & Bilotsky, H. (1976) *Variable Regions of Immunoglobulin Chains* (Medical Computer Systems; Bolt, Barenak and Newman, Cambridge, MA).
44. Tonegawa, S., Maxam, A. M., Tizard, R., Bernard, O. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1485–1489.
45. Milstein, C., Brownlee, G. G., Harrison, T. M. & Matthews, M. B. (1972) *Nature (London) New Biol.* **239**, 117.
46. Swan, D., Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 1967–1971.
47. Schechter, I. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 2256–2260.
48. Rose, S. M., Kuehl, W. M. & Smith, G. P. (1977) *Cell* **12**, 453–462.