

A processed human immunoglobulin ϵ gene has moved to chromosome 9

(pseudogene/human IgE/chromosome mapping)

JAMES BATTEY*†, EDWARD E. MAX‡, WESLEY O. MCBRIDE§, DAVID SWAN§, AND PHILIP LEDER*†

*Laboratory of Molecular Genetics, National Institute of Child Health and Human Development; †Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases; ‡Laboratory of Biochemistry, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205; and §Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115

Contributed by Philip Leder, June 4, 1982

ABSTRACT Processed genes—genes that resemble processed RNA transcripts rather than interrupted genomic sequences—have been identified as dispersed members of several gene families. Here we describe a processed gene that is one of the three human IgE-like sequences present in the human genome. The processed IgE gene has precisely lost its three intervening sequences, thereby fusing its four coding domains. The homology of the gene to its functional counterpart ends in an adenine-rich tail followed by an 11-base-pair sequence that is directly repeated 150 base pairs 5' to its first coding domain. In addition, the processed gene is located on human chromosome 9 rather than on chromosome 14, the site of the active immunoglobulin locus. The structure and evident mobility of this sequence support the concept that sequences can move about in the genome via RNA intermediates and that processed genes are a prominent feature of genomic structure.

Many genes of eukaryotes form families of homologous sequences that include both functional genes and related sequences that appear to be inactive and are therefore called pseudogenes (1–6). One group of eukaryotic pseudogenes, called “processed genes” (7), has specific structural features suggesting that they were generated from the functional gene sequence via an RNA intermediate. Processed genes have precisely lost the intervening sequences found in the corresponding functional genes and retain homology to the 3' end of the functional gene through the polyadenylation signal (A-A-T-A-A) to the site of poly(A) addition. In at least two cases—human immunoglobulin λ (7) and human β -tubulin (8)—a cluster of adenines resembling a poly(A) tail follows this break in homology. In addition, where the determination has been made, these processed genes are not found on the chromosome that encodes their normal counterparts (7, 9, 10). These features have suggested that processed genes may arise from a processed RNA species originally transcribed from a functional gene and conveyed to a new chromosomal location (2, 3, 7–9). Moreover, based on early observations, it seemed that processed genes represent a significant element of genomic structure (2, 7, 9).

We have recently reported the cloning of two human ϵ genes and the determination of their sequences. One is a functional IgE gene with four constant region domains. The other is a pseudogene that has lost the first two coding domains and the 5'-flanking sequences found adjacent to the functional gene (11). Here we describe a third human ϵ -like gene, pseudo- ϵ -2, that contains sequences homologous to the four constant region domains of human IgE but precisely lacks the intervening sequences that interrupt the functional gene. Homology to the

functional gene extends into the 3'-untranslated region, ending 15 nucleotides beyond the polyadenylation signal at which a 32-base-long adenine-rich sequence is found. An 11-base-long sequence that occurs exactly at the end of the adenine-rich tail is repeated [9/11-base-pair (bp) homology] 150 bp 5' to the 5' homology border, suggesting a target site repeat generated by integration into a staggered break in the DNA duplex. Nucleotide sequence analyses carried out separately by Honjo *et al.* (20) on an apparently identical human sequence has suggested a similar conclusion. In addition, although the functional gene for the IgE constant region is located on chromosome 14 in the human immunoglobulin heavy chain cluster (12), we find this processed gene to be located on human chromosome 9, establishing that this sequence is dispersed from its functional IgE gene homologue.

MATERIALS AND METHODS

Cloning and Subcloning. The 9.0-kilobase (kb) *Bam*HI fragment containing pseudo- ϵ -2 was cloned from purified human placental genomic DNA fragments by using the functional IgE sequence as a probe (11). The phage clone was subsequently subcloned into pBR322 for heteroduplex analysis and sequence analysis.

Heteroduplex Analysis. Heteroduplexes were prepared by alkali denaturation of *Cla*I-cleaved pBR322 subclones containing the functional IgE gene and pseudo- ϵ -2 and subsequent renaturation in 50% formamide/12.5 mM EDTA/56 mM NaCl/100 mM Tris base, pH 8.5, at 37°C for 1 hr (11).

DNA Sequence Analysis. DNA sequence analysis was carried out by using both the chemical cleavage method of Maxam and Gilbert (13) and the M13 dideoxynucleotide technique. For dideoxy analysis, the M13 kit and protocol supplied by Bethesda Research Laboratories was used. The resulting fragments were resolved on 8% or 20% acrylamide/8 M urea gels.

Rodent-Human Hybrid Cell Chromosomal Mapping. Analysis of *Bam*HI-digested genomic DNA from rodent-human hybrid cell lines was carried out by Southern blotting using the nick-translated functional IgE gene [the 2.6-kb *Bam*HI fragment (11)] as a probe. The methods and cell lines are essentially as described for the chromosomal mapping of the human κ and λ immunoglobulin genes (14).

RESULTS AND DISCUSSION

Identification and Cloning of the Processed Gene. The human genome contains three sequences that show homology in genomic blot experiments to a functional human IgE constant region probe (11). These homologous sequences are found on

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.

Abbreviations: bp, base pair(s); kb, kilobase(s).

unique *Bam*HI restriction fragments, 2.6, 6.0, and 9.0 kb long. The functional IgE gene is located on the 2.6-kb *Bam*HI fragment and a pseudogene, pseudo- ϵ -1, is found on the 6.0-kb *Bam*HI fragment. Both the functional gene and pseudo- ϵ -1 are linked to IgA constant region genes. The complete nucleotide sequence of the functional IgE gene has been determined, establishing its four-domain structure (11). Sequence determination of pseudo- ϵ -1 showed a deletion of the 5'-flanking sequence along with the first two coding domains present in the functional ϵ gene (11). The third sequence, located on the 9.0-kb *Bam*HI fragment, could not be linked to an IgA constant region gene and, therefore, was suspected of having been conveyed from the active IgE locus to a new chromosomal site. This 9.0-kb fragment was cloned into ϕ CH28 from genomic *Bam*HI fragments of human placental DNA that were size purified by preparative agarose gel electrophoresis. Clones containing ϵ -related sequences were identified by using the 2.6-kb *Bam*HI fragment (functional human ϵ gene) as a probe.

Heteroduplex Comparison of the Functional and Processed IgE Genes. The sequences of the functional and the newly cloned genes could be compared most readily by direct visualization of heteroduplex structures formed between them. Accordingly, the 9.0-kb *Bam*HI fragment containing pseudo- ϵ -2 was subcloned into pBR322 and hybridized to a pBR322 subclone of the 2.6-kb *Bam*HI fragment containing the functional gene. A representative heteroduplex is shown in Fig. 1. Comparison of these two sequences shows nonhomology at both the 5' and 3' ends of the subclones and four homologous regions separated by three small deletion loops. Measurement of 20 such heteroduplex structures established that the 5' border of

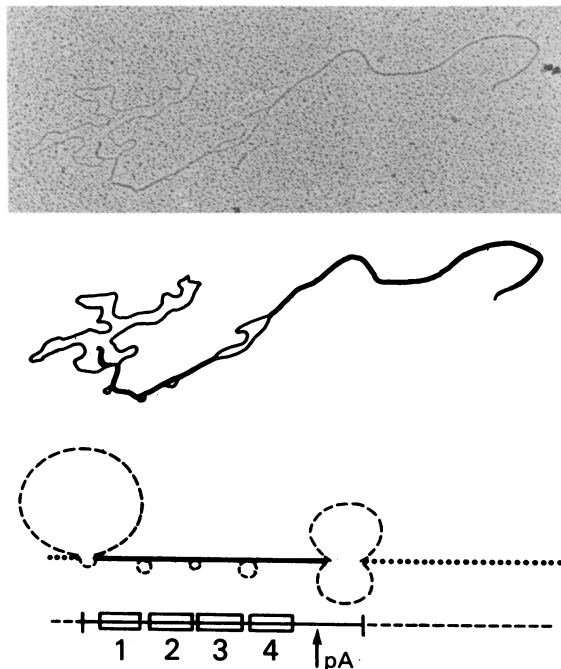


FIG. 1. Heteroduplex showing regions of homology between pseudo- ϵ -2 and the functional ϵ gene. The electron micrograph and tracing below show a representative heteroduplex between pBR322-derived plasmids containing the 2.6-kb *Bam*HI fragment (functional gene) and the 9-kb *Bam*HI fragment (processed pseudogene). As interpreted in the diagram below, the heteroduplex shows nonhomology at the 5' ends of the two cloned regions (left), the homologous region interrupted by three small deletions, and the nonhomologous 3' ends (right), beginning at the poly(A) addition site (indicated as pA). Dashed lines represent single-stranded DNA, dark solid lines represent homologous double-stranded heteroduplex, and dotted lines are double-stranded pBR322.

homology is located near the 5' end of the first coding domain of the functional gene. The 3' break in homology is near the polyadenylation signal in the functional gene. In addition, the three small deletion loops interrupting the homologous central region are located at the positions of the three intervening sequences in the functional gene. These data are consistent with the identification of pseudo- ϵ -2 as a processed pseudogene.

Nucleotide Sequence Analysis of the Processed Gene: Comparison with the Functional IgE Gene. The partial nucleotide sequence of pseudo- ϵ -2 is shown in Fig. 2. The boxed portions of the sequence represent regions homologous to the four domains (CH1-CH4) and the 3'-untranslated region (3'-UT) of the functional ϵ gene. The two sequences show extensive homology over these regions. However, multiple small insertions and deletions in the pseudo- ϵ -2 sequence create termination codons and missense regions rendering it a pseudogene, a sequence incapable of encoding a normal ϵ constant region polypeptide.

A sequence comparison between pseudo- ϵ -2 and the functional IgE gene at the coding domain borders verifies that the intervening sequences are precisely deleted from pseudo- ϵ -2 (Fig. 3). On the 5' side of the sequence, homology ends abruptly at the 3' border of the intervening sequence that flanks the first coding region (CH1) of the functional IgE gene. On the 3' side, the homology stops 15 nucleotides 3' to the polyadenylation signal (A-A-T-A-A-A). The pseudogene sequence 3' to this break in homology is a 32-bp sequence that is 75% adenine, showing no significant homology to the corresponding sequence in the functional gene. Each of these points of divergence between the functional gene and pseudo- ϵ -2 can be generated by established eukaryotic RNA processing reactions: removal of intervening sequences and polyadenylation.

The 5' Sequences of Processed Genes Show No Homology to Sequences Immediately 5' to the Functional IgE Gene. Comparison between the pseudo- ϵ -2 nucleotide sequence 5' to the CH1 homology region and the functional IgE sequence 5' to CH1 shows no significant homology. A consensus RNA splice acceptor site (C-A-C-A-G- \downarrow) (15) is found immediately 5' to CH1 in the functional gene, as expected. Pseudo- ϵ -2 contains no such sequence, suggesting that a new "domain" has been created 5' to CH1 in pseudo- ϵ -2, and resulting in removal of the intervening sequence and its splice acceptor site. The origin of the sequences in this new pseudo- ϵ -2 domain remains a matter for speculation. It is possible that the sequences comprising the new domain are part of a mobile genetic element responsible for creating the new pseudogene. Alternatively, homology to the sequence comprising the new domain may exist further 5' to the functional gene than our available sequence. If this is correct, such a domain must have been carried in over a great distance, because heteroduplexes formed between phage clones containing pseudo- ϵ -2 and clones containing the functional gene fail to show any homology between the new domain and sequences as far as 12 kb 5' to the functional gene (data not shown).

An 11-Nucleotide Sequence 3' to Poly(A) Is Repeated 5' to CH1 in Pseudo- ϵ -2. Several interesting sequences are found in the region 5' to the CH1 homology region of pseudo- ϵ -2. The nucleotide sequence immediately 3' to the poly(A) sequence (C-C-T-A-G-A-G-G-A-A) is repeated directly with 80% fidelity approximately 150 bp 5' to the CH1 homology region, as shown in Fig. 2. The occurrence of a repeat sequence 5' and 3' to the processed gene sequence is reminiscent of the direct repeats known to flank both eukaryotic and prokaryotic transposable elements. By analogy, the location of the direct repeats would then define the 5' and 3' borders of the reintegrated sequence. Their generation may have resulted from DNA repair of a stag-

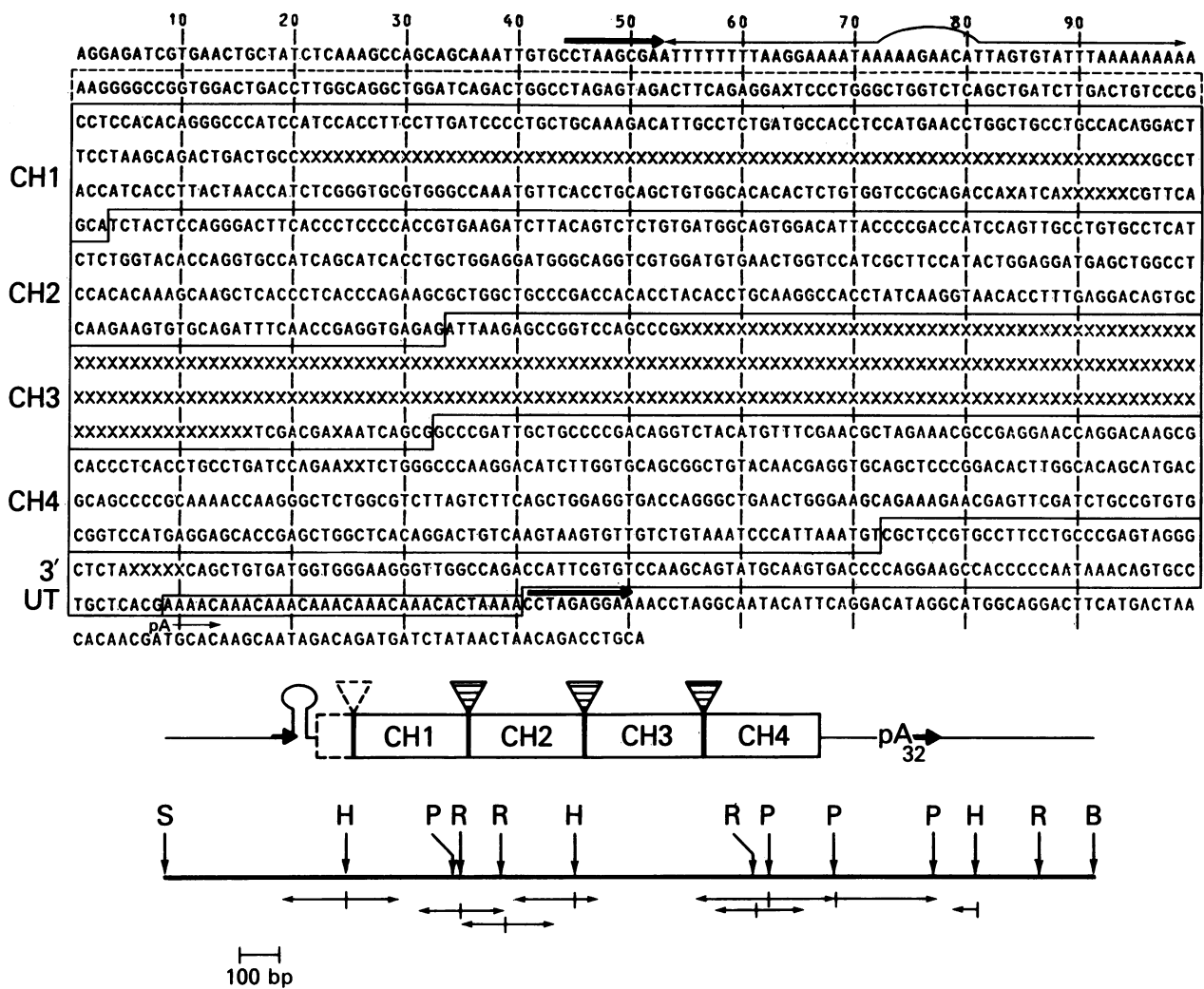


FIG. 2. Partial nucleotide sequence of pseudo- ϵ -2. (Upper) Sequence. Homologous regions to the four domains (CH1-CH4) of the functional gene are indicated by solid boxes. A dashed box indicates the new domain created during formation of the processed pseudogene. A solid box also surrounds the sequence homologous to the 3'-untranslated (3-UT) sequence of the functional gene. The poly(A)-like sequence flanking the 3'-untranslated homology is indicated by a small box and pA→. The direct repeats flanking the 5' and 3' regions of the gene are indicated by dark arrows above the sequence. The palindrome in sequences 5' to the CH1 homology is indicated by a light arrow. (Middle) Schematic of pseudo- ϵ -2. Triangles indicate positions of removed intervening sequences. The four homologous domains are indicated by solid boxes, and the new domain is indicated by a dashed box. The position of the poly(A)-like sequence is designated pA₃₂. The direct repeat is shown by heavy arrows and the position of the palindrome is shown just 3' to the 5' direct repeat. (Lower) Restriction map and sequence analysis strategy for pseudo- ϵ -2 drawn to the same scale as the schematic. Horizontal arrows indicate the direction and extent of analysis from the designated sites. B, *Bam*HI; H, *Hin*FI; P, *Pvu* II; R, *Rsa* I; S, *Sst* I.

gered chromosomal break at the site into which pseudo- ϵ -2 was integrated. Direct repeats flanking processed genes have been observed in several other pseudogenes. The human immunoglobulin λ pseudogene $\lambda\psi 1$ contains a 9-bp sequence (G-A-T-G-T-G-A-A-T) immediately 3' to poly(A) that is directly repeated 150 bp 5' to the processed gene (7). Similarly, the human β -tubulin processed gene contains an 11-bp sequence (G-C-T-G-A-G-T-G-T-C) immediately 3' to poly(A) that is directly repeated 170 bp 5' to the processed coding sequence (8). Direct-repeat sequences have also been found flanking three human SnRNA pseudogenes (6). One of these SnRNA pseudogenes, U1.101, has a 16-bp repeat that immediately follows a poly(A) sequence on the 3' side. Reiterated *Alu* sequences are flanked by direct repeats and in at least one instance this repeat immediately follows a poly(A) sequence (16).

An ≈ 50 -bp sequence marked by runs of adenines and thymines is located immediately 3' to the direct repeat situated 150 bp 5' to the CH1 homology in pseudo- ϵ -2 (Fig. 2). This 50-bp region contains an inverted complement consisting of mostly

adenines and thymines. The significance of these sequences is unclear, as no such palindromic structure is found in the sequences located in an analogous position in either the $\lambda\psi 1$ or the β -tubulin processed genes.

Pseudo- ϵ -2 Is Both Processed and Dispersed. Two previously described processed pseudogenes, the mouse α -globin pseudogene $\alpha\psi 1$ (9) and the human $\lambda\psi 1$ (7), were shown to be located on different chromosomes than their corresponding active genes. The human heavy chain immunoglobulin constant region locus is known to be located on chromosome 14 at band q32 (12). To test whether pseudo- ϵ -2 is located on chromosome 14 or dispersed to a different chromosome, rodent-human hybrid cell lines that retain different human chromosomes (14) were used. The presence or absence of the characteristic 9.0-kb *Bam*HI fragment containing pseudo- ϵ -2 was correlated with the presence or absence of each human chromosome in the hybrid cell line. The results for four sets of rodent-human hybrid lines (A-D) are shown in Table 1. In all four sets of hybrid lines, the concordance of pseudo- ϵ -2 is highest with chromo-

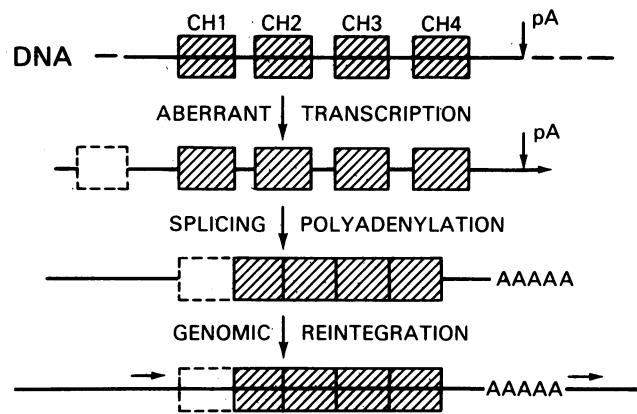


FIG. 4. Model for generation of pseudo- ϵ -2 from the functional ϵ gene. The top line shows the functional ϵ gene with its four domains (CH1-CH4), polyadenylation site (pA), and flanking DNA. The second line shows an aberrant transcript of this gene. The dashed box is the new domain that is formed during pseudogene generation. The third line shows the processed RNA generated from the aberrant transcript during RNA processing reactions. The adenines at the 3' end represent the poly(A) sequence. The bottom line shows the re-integrated gene, pseudo- ϵ -2. The small arrows flanking the gene 5' and 3' represent the target site duplication created by repair of a staggered chromosomal break at the site of reintegration.

was proposed as a possible explanation for the mouse α -globin processed pseudogene (2). Thus, a processed RNA transcript could provide a template for gene conversion of a heteroduplex structure formed between the transcript and the functional cellular gene. The looped-out single-stranded DNA from the intervening sequences would then be nicked and removed, and the DNA would be sealed by repair mechanisms. Although this mechanism may explain some processed genes [e.g., the rat insulin gene that lacks one of the two ancestral intervening sequences (19)], it cannot easily account for the poly(A)-like region of pseudo- ϵ -2 and thus appears an unlikely mechanism for the origin of this processed gene.

Bands comigrating with the ϵ (11) and λ processed pseudogenes (G. Hollis, personal communication) are found in blots of genomic DNA from all human individuals (>20) examined to date and from chimpanzee. The mouse α -globin processed gene occurs in a variety of mouse species (R. Taub and A. Leder, personal communication). Although processed genes appear to be a common element in many eukaryotic gene families, both human processed genes analyzed in this laboratory were formed before the divergence of chimpanzees and humans. From these

examples, it appears that processed gene formation and movement is not a frequent event in the lifetime of a single individual. Instead, processed pseudogenes accumulate in the genome over tens of millions of years.

We are grateful to Marion Nau for her invaluable assistance in the gene cloning portions of these experiments, to Barbara Norman for her technical expertise, to Terri Broderick for her expert help in the preparation of the manuscript, and to Greg Hollis for helpful discussion. We also wish to acknowledge the support of the American Cancer Society (ACSPF 2057 to J.B.) and the American Business Cancer Research Fund.

- Jacq, D., Miller, J. P. & Brownlee, G. C. (1977) *Cell* **12**, 109-120.
- Nishioka, Y., Leder, A. & Leder, P. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2806-2809.
- Vanin, B. F., Goldberg, G. I., Tucker, P. W. & Smithies, O. (1980) *Nature (London)* **286**, 222-226.
- Bentley, D. L. & Rabbitts, T. H. (1980) *Nature (London)* **288**, 730-733.
- Firtel, R. A., Timon, R., Kimmel, A. R. & McKeown, M. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6206-6210.
- VanArsdell, S. W., Denison, R. A., Bernstein, L. E., Weiner, A. M., Manser, T. & Gasteland, R. F. (1981) *Cell* **26**, 11-17.
- Hollis, G. F., Hieter, P. A., McBride, C. W., Swan, D. & Leder, P. (1982) *Nature (London)* **296**, 321-325.
- Wilde, C. D., Crowther, C. E., Cripe, T. P., Lee, M. G. & Cowen, N. J. (1982) *Nature (London)* **297**, 83-84.
- Leder, A., Swan, D., Ruddle, F., D'Bustachio, P. & Leder, P. (1981) *Nature (London)* **293**, 196-200.
- Popp, R. A., Lalley, P. A., Whitney, J. E., III, & Anderson, W. F. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 6362-6366.
- Max, B. B., Battey, J., Kirsch, I. R., Ney, R. I. & Leder, P. (1982) *Cell* **29**, 691-699.
- Kirsch, I. R., Morton, C. C., Nakahara, K. & Leder, P. (1982) *Science* **216**, 301-303.
- Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
- McBride, O. W., Hieter, P. A., Hollis, G. P., Swan, D., Otey, M. C. & Leder, P. (1982) *J. Exp. Med.* **155**, 1480-1490.
- Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349-383.
- Jagadeeswaran, P., Forget, B. G. & Weissman, S. M. (1981) *Cell* **26**, 141-142.
- Goff, S. P., Gilboa, B., Witte, O. N. & Baltimore, D. (1980) *Cell* **22**, 777-785.
- Lueders, K., Leder, A., Leder, P. & Kuff, B. (1982) *Nature (London)* **295**, 426-428.
- Lomedico, P., Rosenthal, N., Efstradiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545-558.
- Honjo, T., Ishida, N., Kataoka, T., Nakai, S., Nikaido, T., Nishida, Y., Noma, Y., Obata, M., Sakoyama, Y., Shimizu, A., Takahashi, N., Takeda, S., Ueda, S., Yamawaki-Kataoka, Y. & Yaoita, Y. (1982) in *Proceedings of the Nobel Symposium* (Plenum, New York), in press.