Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 5355–5358, September 1983
Genetics

DNA inversions in the chromosome of *Escherichia coli* and in bacteriophage Mu: Relationship to other site-specific recombination systems

(bacteriophage Mu G inversion/DNA sequence/phase variation)

RONALD H. A. PLASTERK, AD BRINKMAN, AND PIETER VAN DE PUTTE

Laboratory of Molecular Genetics, State University of Leiden, Department of Biochemistry, Wassenaarseweg 64, 2333 AL Leiden, The Netherlands

Communicated by Allan Campbell, May 12, 1983

ABSTRACT **The gene product of bacteriophage Mu *gin* catalyzes a 3,000-base-pair inversion in the DNA of the phage, thus changing its host range. In some strains of *Escherichia coli* there is a function that can complement Mu *gin* mutations. This function (*pin*) was cloned and shown to catalyze an inversion of 1,800 base pairs in the adjacent *E. coli* DNA (P region). *pin*⁻ derivatives carry the P region frozen in the (+) or (−) orientation. The function of the switch is not yet clear. The sequences of *gin* and *pin* were determined; they exhibit 70% homology. The sequences around the recombination sites of Gin and Pin are also largely homologous; a consensus sequence is derived for the recombination sites of Gin and Pin, and of Hin in *Salmonella typhimurium*. The amino acid sequences of Gin, Pin, Hin, and TnpR are compared, and the evolutionary relationship between these prokaryotic site-specific recombination systems is discussed.**

Inversions of DNA segments in prokaryotes have been found to serve different functions such as change of the host range of bacteriophage Mu (1–3) and change of the flagellar antigen of *Salmonella typhimurium* (4). The NH_2-terminal part of the Mu tail fiber gene is located in the noninverting DNA, and two different COOH-terminal parts of the gene are spliced to the constant part by inversion of the G region (2). In *S. typhimurium*, a promoter located in the invertible DNA can turn on genes in the adjacent DNA (4). Although function and genetic organization of invertible DNA are different in these cases, all genes catalyzing inversions in prokaryotes have been shown to complement each other [*gin* of Mu, *hin* of *S. typhimurium*, and *cin* of phage P1 which is closely related to Mu (5–7)].

We recently found a function in the chromosome of *Escherichia coli* (*pin*) that complements Mu *gin* mutations and catalyzes the inversion of a 1,800-base-pair (bp) region of DNA (unpublished data). This function is found in *E. coli* strains HB101 and CSH520. To investigate this gene further, we cloned it and showed that Pin is responsible for the inversion of a region which we named P region. To compare the different DNA-invertase genes we determined the sequences of *pin* and *gin*. The comparison defines regions in the genes which are conserved in the invertases Gin, Pin, and Hin.

It was previously noted that the proteins Hin and TnpR are 33% homologous (8). These genes do not complement each other, however. TnpR catalyzes deletions (resolution of cointegrates of transposon Tn3), whereas Gin preferentially catalyzes inversions (9). The conserved regions in the DNA-invertase genes (*hin*, *pin*, and *gin*) are compared to the *tnpR* sequence to find regions homologous in both types of recombinases.

gin of Mu has been shown to be expressed at a very low level, probably due to a low efficiency of both transcription- and translation-initiation (9, 10). The nucleotide sequence of the promoter and translation-initiation site are discussed in the light of this low level of expression.

The invertible G region of Mu is flanked by short inverted repeats, the size of which has been estimated to be 50 bp by using electron microscopy (11). The inverted repeats of the invertible region of *S. typhimurium* are 14 bp long (4). We compared the sites where Gin, Pin, and Hin act. The putative promoters of *gin*, *pin*, and *hin* overlap these recombination sites.

MATERIALS AND METHODS

Bacterial Strains and Plasmids. The strains used during this study were *E. coli* C (R. Sinsheimer) and the *E. coli* K-12 strains HB101 (*recA*, *pin*+), KMBL1164 (*pin*, Δ *lac-prox*111, our laboratory), and JM101 (12). Phages used were Mu *gin*10 (our laboratory) and the M13 vectors mp8 and mp9 (12). The plasmids used were the cloning vectors pBR322 and pACYC177.

Recombinant DNA Techniques. Plasmid DNA was isolated as described (13). Restriction enzyme reactions, BAL31, and T4 DNA ligase reactions were performed as advised by the suppliers (Boehringer Mannheim, BioLabs, P-L Biochemicals, Amersham). DNA was analyzed on 1% agarose gels (2).

DNA Sequence Analysis. DNA was cloned into M13 vector mp8 or mp9, and clones were checked for the insert by isolation of replicative form DNA 4 hr after infection (13). Polymerization and electrophoresis were done as described (14, 15).

DNA Heteroduplexing. Plasmid DNA was heteroduplexed as described (1) and visualized by using the Philips EM 300.

RESULTS

Cloning of *pin* and the Invertible P Region. A plasmid containing the *pin* gene was isolated from the colony bank of Clarke and Carbon (16). This colony bank contains chromosomal DNA of *E. coli* cloned into the *Eco*RI site of ColE1 (see legend to Fig. 1). Heteroduplexing of the circular plasmid DNA revealed a 1,800-bp invertible segment (unpublished data). By heteroduplexing of plasmid DNA digested with *Hin*dIII or *Sma* I, we mapped the invertible segment (P region) relative to these sites (Fig. 1). The *Sma* I site is in the vector DNA (17), and the *Hin*dIII site is in the chromosomal insert. The *Hin*dIII–*Sma* I fragment containing the P region was cloned into pACYC177 digested with *Hin*dIII and *Sma* I (pGP300). The *Hin*dIII–*Xho* I fragment of this plasmid was cloned subsequently into pBR322 digested with *Hin*dIII and *Sal* I. A restriction map was made of the resulting plasmid pGP301 (Fig. 2). All plasmid isolates contained 50% in the (+) and 50% in the (−) orientation. Apparently, *pin*

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.

Abbreviation: bp, base pair(s).

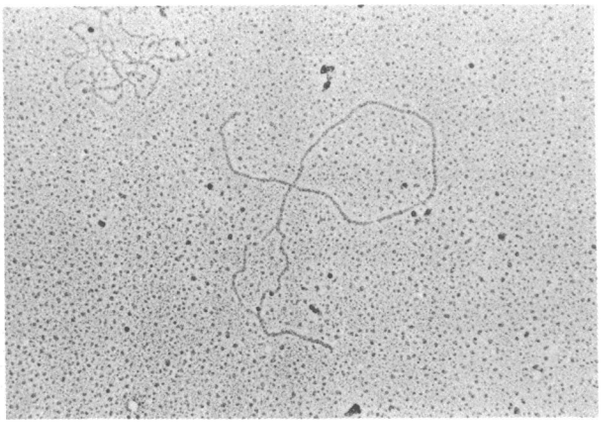

FIG. 1. Heteroduplex of *Hin*dIII-digested DNA of pM6, the plasmid from clone 6-13 of the colony bank of Clarke and Carbon (16). It has a chromosomal insert of *E. coli* strain CSH520 in the *Eco*RI site of ColE1. By using pBR322 as an internal size marker, the size of the invertible region and the distances from it to the *Hin*dIII site and the *Sma* I site were determined. The *Sma* I site is in the vector DNA (17); the *Hin*dIII site was mapped in the chromosomal insert.

is located on the subcloned fragment.

A unique *Bgl* II site was mapped just outside of the invertible region. In analogy with the Mu G region, it was expected that the *pin* gene would map in this region. This was investigated by filling the *Bgl* II site by using the Klenow fragment of DNA polymerase I. The resulting plasmid is *pin*$^-$. In this plasmid, the P region was frozen in the (+) or the (−) orientation, showing that indeed *pin* is responsible for the P inversion.

Nucleotide Sequence of the *E. coli pin* Gene and the Mu *gin* Gene. To investigate the inversion systems further, we determined the nucleotide sequence of *pin* and *gin*. By using exonuclease BAL31, deletions were made from the *Bgl* II site in *pin* and *Bam*HI linkers were inserted. A similar procedure was followed for *gin*: deletions were made from the *Eco*RI site in PGP204 (18) and *Eco*RI linkers were inserted. After cloning of the various deletion derivatives into M13 vectors, sequences were determined. The nucleotide sequence of *pin* and adjacent regions is shown in Fig. 3. The sequence of *gin* is in Fig. 4.

Several features of the *gin* sequence are as follows.

1. The initiation triplet is GTG instead of ATG. This may contribute to the low translation efficiency of the gene (9). The initiation triplet is preceded by a Shine–Dalgarno sequence.

2. The termination codons are found at the end of *gin*, each of which immediately follows a methylation site (G-A-T-C). As shown elsewhere (18), methylation of these sites, which are in the promoter region of the Mu *mom* gene, is essential for *mom* transcription. The overlap of the *gin* terminus and the *mom* methylation-dependent promoter may indicate some coordinate expression, but no evidence for this has been found. A Shine–Dalgarno sequence is found 90 bp downstream, followed by an ATG triplet and open reading frame.

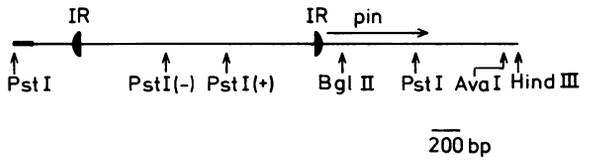

FIG. 2. Restriction map of the chromosomal DNA carrying the P region cloned in pGP301. The ends of the invertible region are indicated by ◖ and ◗. The location of *pin* was determined by nucleotide sequence analysis (see below). The thick line represents ColE1 DNA. It is separated from the insert by a poly(T) tail of approximately 100 bp (as follows from sequence analysis).

```
GAAAAGCGGACCATTGCATTTCAGCCAGCC
          12        22        32
TGTTGGCGGATGCTGAAGGCCACGGAACCG
          42        52        62
GGACACCAATAGGTAATGCAGAGCCTTCTC
          72        82        92
CCAAACCAACGTTTATGAAAATGAAGAAAT
         102       112  -10  122
AACAAGCAAATGGCATCATTCCTGCTTTTA
         132       142       152
    SD                Met Leu Ile Gly Tyr
CCAGGGGGATTTAACATGCTTATTGGCTAT
         162       172       182
 Val Arg Val Ser Thr Asn Asp Gln Asn Thr
GTACGCGTATCAACAAATGACCAGAACACA
         192       202       212
 Asp Leu Gln Arg Asn Ala Leu Asn Cys Ala
GATCTACAACGTAATGCGCTGAACTGTGCA
  BglII  222       232       242
 Gly Cys Glu Leu Ile Phe Glu Asp Lys Ile
GGATGCGAGCTGATTTTTGAAGACAAGATA
         252       262       272
 Ser Gly Thr Lys Ser Glu Arg Pro Gly Leu
AGCGGCACAAAGTCCGAAAGGCCGGGACTG
         282       292       302
 Lys Lys Leu Leu Arg Thr Leu Ser Ala Gly
AAAAAACTGCTCAGGACATTATCGGCAGGT
         312       322       332
 Asp Thr Leu Val Val Trp Lys Leu Asp Arg
GACACTCTGGTTGTCTGGAAGCTGGATCGG
         342       352       362
 Leu Gly Arg Ser Met Arg His Leu Val Val
CTGGGGCGTAGTATGCGGCATCTTGTCGTG
         372       382       392
 Leu Val Glu Glu Leu Arg Glu Arg Gly Ile
CTGGTGGAGGAGTTGCGCGAACGAGGCATC
         402       412       422
 Asn Phe Arg Ser Leu Thr Asp Ser Ile Asp
AACTTTCGTAGTCTGACGGATTCAATTGAT
         432       442       452
 Thr Ser Thr Pro Met Gly Arg Phe Phe Phe
ACCAGCACACCAATGGGACGCTTTTTCTTT
         462       472       482
 His Val Met Gly Ala Leu Ala Glu Met Glu
CATGTGATGGGTGCCCTGGCTGAAATGGAG
         492       502       512
 Arg Glu Leu Ile Val Glu Arg Thr Lys Ala
CGTGAACTGATTGTTGAACGAACAAAAGCT
         522       532       542
 Gly Leu Glu Thr Ala Arg Ala Gln Gly Arg
GGACTGGAAACTGCTCGTGCACAGGGACGA
         552       562       572
 Ile Gly Gly Arg Arg Pro Lys Leu Thr Pro
ATTGGTGGACGTCGTCCCAAACTTACACCA
         582       592       602
 Glu Gln Trp Ala Gln Ala Gly Arg Leu Ile
GAACAATGGGCACAAGCTGGACGATTAATT
         612       622       632
 Ala Ala Gly Thr Pro Arg Gln Lys Val Ala
GCAGCAGGAACTCCTCGCCAGAAGGTGGCG
         642       652       662
 Ile Ile Tyr Asp Val Gly Val Ser Thr Leu
ATTATCTATGATGTTGGTGTGTCAACTTTG
         672       682  HindII   692
 Tyr Lys Arg Phe Pro Ala Gly Asp Lys ***
TATAAGAGGTTTCCTGCAGGGGATAAATAA
         702  PstI 712       722
AGTTAAAGACACTTTGTGTACAAAAGAAAG
         732       742       752
TAAAACAACAGCCAACTTGTTCGAATTTTA
         762       772       782
TCAATAAAAGTCAGTATTGTCGTGAAAATT
         792       802       812
```

FIG. 3. The nucleotide sequence of *pin* and the predicted amino acid sequence. The inverted repeat adjacent to the invertible region is indicated by a line. The putative promoter and the Shine–Dalgarno sequence (*SD*) are indicated.

3. The clone in which the BAL31-generated deletion extends to site 10 (indicated in Fig. 4) is *gin*$^+$, whereas the clone in which the deletion extends to site 66 is *gin*$^-$. *gin* expression was tested by a complementation test as described (9). The promoter of *gin* thus maps at least partially between sites 10 and 66. Indeed, a sequence can be found that shows good homology with the −10 consensus sequence for *E. coli* promoters (indicated in Fig. 4). This promoter was previously designated as the putative *gin* promoter on the basis of the DNA sequence.*

Comparison of the sequences of *pin* and *gin* permits the following conclusions to be drawn.

1. The sequence of *pin* is colinear with *gin* but is 9 triplets shorter at the COOH terminus (Fig. 5).

2. The initiation triplet of *pin* is ATG. It is not known if the

* Kahmann, R., EMBO Meeting on Bacteriophage Mu, May 11–15, 1981, Texel, the Netherlands.

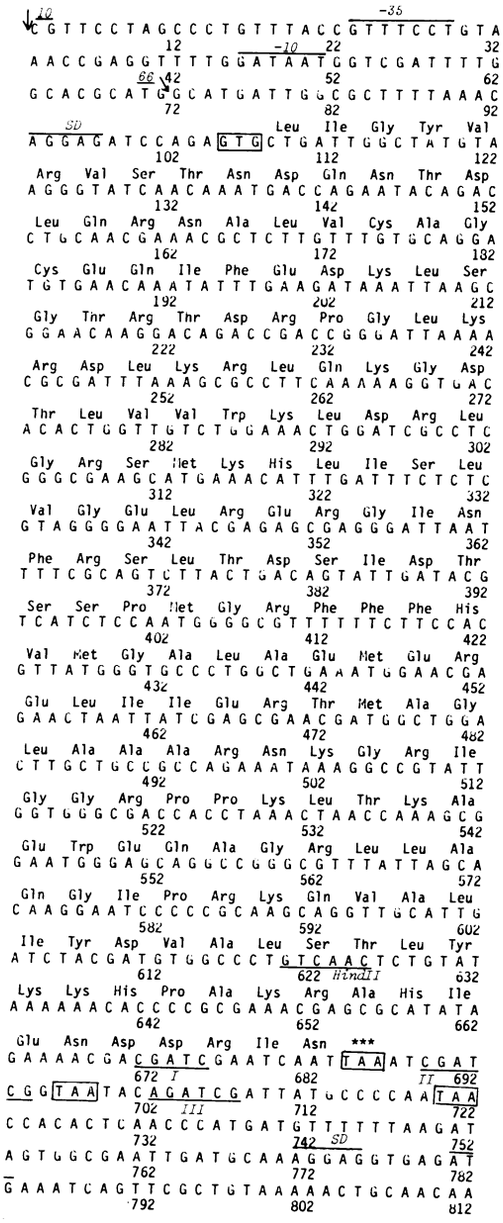

FIG. 4. The nucleotide sequence of *gin* and the predicted amino acid sequence. Arrows *10* and *66* indicate the end points of the BAL31-generated deletions. The promoter of *gin* is genetically mapped between these two sites, and the putative −35 and −10 sequences are indicated. The three G-A-T-C sites at the end of *gin* are underlined. The probable start of a gene downstream from the *mom* promoter is indicated.

translation efficiency is higher than in the case of *gin*, which starts with a GTG triplet.

3. At the end of *pin* no sequences resembling the three G-A-T-C sites or the *mom* promoter are found. Also, the leader sequence of *pin* is different from that of *gin*.

4. A Pribnow box is found 60 bp upstream of the *pin* ATG triplet. However the −35 region shows no significant homology with the consensus sequence in *E. coli*.

Comparison of *gin* and *pin* to Other Site-Specific Recombinases. The sequences of Hin and TnpR have been shown to contain 33% homology (8). However, these functions do not complement each other (20). The sequence in Tn3 previously indicated as being homologous to the Hin inverted repeat (4) turned out not to be the TnpR recombination site (21). *hin* and *gin* do complement each other (5, 22), as *pin* and *gin* do.

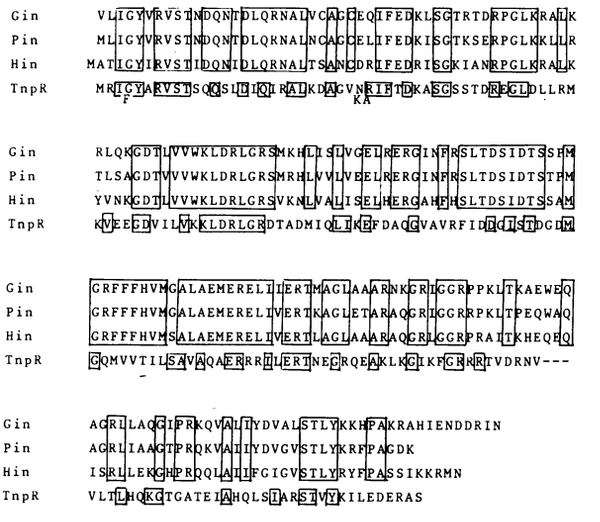

FIG. 5. Alignment of the amino acid sequences of Pin, Gin, Hin, and TnpR. The amino acid sequences of Pin and Gin are derived from the nucleotide sequences that we determined; the Hin sequence is from Zieg and Simon (4), and TnpR sequence is from Heffron *et al.* (19). The top three lines show the DNA-invertase proteins; the amino acids present in all three proteins are boxed. In the bottom line, the homology between TnpR and Hin [as first demonstrated by Simon *et al.* (8)] is indicated by boxes in the TnpR sequence. The amino acids are designated by the standard one-letter symbols.

We compared the amino acid sequences of these three DNA-invertases to see which regions are conserved (Fig. 5). As expected, there is a considerable degree of homology between Pin and Gin (70%). Homology is 62% for Hin and Gin and 60% for Pin and Hin. It seems that Pin and Gin are more closely related to each other than either of them is to Hin. Many differences in the sequence are found in the third nucleotide of codons, leaving the amino acid sequence unaltered. This implies that there has been a selective pressure to keep the proteins functional. This is an important conclusion because no function has yet been found for P inversion. In Fig. 6 the regions of homology in the three DNA-invertases are boxed. The amino acid sequence of TnpR is compared to this homology. It is apparent that especially the longer stretches of homology between TnpR and Hin are conserved among all DNA-invertases. Probably these are regions that are essential for the function of both types of recombination enzymes.

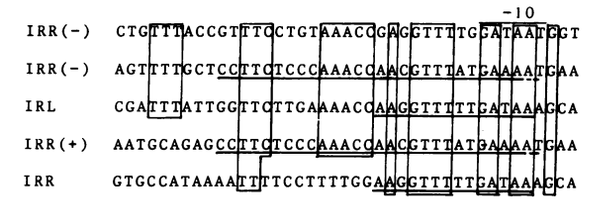

FIG. 6. The recombination sites of *gin*, *pin*, and *hin*. The *hin* inverted repeat is from Zieg and Simon (4). In order to align it with the inverted repeats of *gin* and *pin*, the *hin* inverted repeat must be inverted with regard to the invertible segment. The extent of the inverted repeats is indicated by horizontal lines. The exact end points of the inverted repeat of *gin* have not been determined; only the end point of the left side of the inverted repeat of *pin* is known. Homologous bases are in boxes. The putative −10 promoter sequences are indicated by a line. It is clear that homology between the recombination sites is found outside of the inverted repeats of *pin* and *hin* but only in one of the two sites. The putative −35 sequences of *hin* are thus in the DNA outside of the inverted repeat—i.e., in the noninverting *S. typhimurium* DNA. IRL, IRR, left and right inverted repeats; (+), (−), orientations of inverting DNA.

The Recombination Sites. Fig. 6 shows the recombination sites of Gin and Pin (see Figs. 3 and 4) aligned with those of Hin (4). The inverted repeats of Hin show homology with those of Gin and Pin for 8 of 14 bp. The inverted repeat of Pin extends longer than that of Hin and shows more homology with the Gin inverted repeat. The consensus inverted repeat for all three DNA-invertases is -A-GTTT--GA-AA.

Homology is also found between the inverted repeats of Gin and Pin and the DNA just outside *one* of the two Hin inverted repeats. These sequences of extra homology may turn out to be essential for the inversion reaction. The situation is not symmetrical: these sequences are found only near the inverted repeat on one side of the invertible region.

Alignment with the internal resolution site of TnpR is not possible. The 6-bp palindromic sequence cut by TnpR (T-T-A-T-A-A) (23) is not found. This lack of homology between the recombination sites is sufficient to explain why the DNA-invertases and the resolvase do not complement each other (20).

The promoter of *gin*, as we mapped it, overlaps with the inverted repeat. Because the Pribnow box is conserved in both the *pin* and *hin* inverted repeats, it may also be expected that in these cases the promoter overlaps the recombination sites. No other reasonable fit with the −10 consensus can be found preceding *pin* in the noninverting DNA. If the *hin* promoter indicated by us is correct, it has different −35 regions depending on the orientation of the invertible DNA. This apparently influences the level of expression of *hin* because the inversion rates from (+) to (−) and vice versa are slightly different (8). In the case of *pin*, the −35 regions are within the inverted repeat, and the inversion rates in both directions are equal.

DISCUSSION

Genetic switches by DNA inversions such as described here have different functions and appear in different genetic elements. However, the inversion systems are closely related in both the recombination sites and the amino acid sequence of the recombinases. Whereas Pin, Gin, and Hin are more than 60% homologous at the amino acid level, the sequences surrounding the genes are different. The Hin system is located on the *S. typhimurium* chromosome, the Gin system is in the phage Mu genome (which itself is a transposon), and the Pin system is in the *E. coli* chromosome. We have indications that *pin* is on a defective prophage or cryptic plasmid (unpublished data). Possibly the Pin system, whatever its function, represents the "missing link" between the inverting DNA in the Mu "transposable phage" and the *S. typhimurium* chromosomal inversion.

Another known site-specific recombination system, cointegrate resolution of Tn3 by TnpR, differs from the inversion systems in several respects: (*i*) TnpR preferentially catalyzes deletions (20), whereas Gin preferentially catalyzes inversions (9); (*ii*) TnpR and the DNA-invertases do not complement each other (20); and (*iii*) the sequences of the recombination sites are not homologous (21, 24).

There are also important similarities.

1. TnpR shows homology to Hin, Gin, and Pin, found predominantly within some longer stretches of homology among the DNA-invertases. Two tyrosine residues are found in all four proteins (Fig. 5). One of these (or both) may play a role in the protein–DNA interactions (21, 25).
2. The promoters of *tnpR* and *gin*, and probably of *pin* and *hin*, overlap the corresponding recombination sites.
3. Tn*3* resolution, G inversion, and possibly P inversion are site-specific recombinations *within* a transposable or excisable element. These similarities lead to the conclusion that this site-specific recombination system has been incorporated as a kind of "module" into different complex genetic structures. Homology between these modules is retained by selective pressure to keep them functional.

We thank several people for their contribution to this work: Steve Cramer for the electron microscopy and Thyra Ilmer, Martha Vollering, and Harry Vrieling for their assistance at various stages. The sequence of *gin* has independently been determined by Dr. R. Kahmann, and several features of it were described at the EMBO Mu workshop in Texel, May 1981. We thank Dr. J. Brouwer, Dr. M. Giphart-Gassler, and R. S. Haxo for critical reading of the manuscript and Mrs. N. van Hoek for typing it.

1. Van de Putte, P., Cramer, S. & Giphart-Gassler, M. (1980) *Nature (London)* **286**, 218–222.
2. Giphart-Gassler, M., Plasterk, R. H. A. & Van de Putte, P. (1982) *Nature (London)* **297**, 339–342.
3. Kamp, D., Kahmann, R., Zipser, D., Broker, T. R. & Chow, L. T. (1978) *Nature (London)* **271**, 577–580.
4. Zieg, J. & Simon, J. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4196–4201.
5. Kutsukake, K. & Iino, T. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7338–7541.
6. Chow, L. T. & Bukhari, A. I. (1976) *Virology* **74**, 242–248.
7. Iida, S., Meyer, J., Kennedy, K. E. & Arber, W. (1983) *EMBO J.* **1**, 1445–1453.
8. Simon, M., Zieg, J., Silverman, M., Mandel, G. & Doolittle, R. (1980) *Science* **209**, 1370–1374.
9. Plasterk, R. H. A., Ilmer, T. & Van de Putte, P. (1983) *Virology* **127**, 24–36.
10. Kwoh, D. Y. & Zipser, D. (1981) *Virology* **119**, 291–296.
11. Hsu, M. & Davidson, N. (1972) *Proc. Natl. Acad. Sci. USA* **69**, 2823–2827.
12. Messing, J. & Vierra, J. (1982) *Gene* **19**, 269–276.
13. Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513–1523.
14. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
15. Sanger, F. & Coulson, A. R. (1978) *FEBS Lett.* **87**, 107–110.
16. Clarke, L. & Carbon, J. (1976) *Cell* **9**, 91–99.
17. Schumann, W. & Bade, E. G. (1979) *Mol. Gen. Genet.* **169**, 97–105.
18. Plasterk, R. H. A., Vrieling, H. & Van de Putte, P. (1983) *Nature (London)* **301**, 344–347.
19. Heffron, F., McCarthy, J., Ohtsubo, M. & Ohtsubo, E. (1979) *Cell* **18**, 1153–1163.
20. Reed, R. R. (1981) *Cell* **25**, 713–719.
21. Reed, R. R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3428–3432.
22. Kahmann, R. & Kamp, D. (1981) *Mol. Gen. Genet.* **184**, 564–566.
23. Reed, R. R. & Grindley, N. D. F. (1981) *Cell* **25**, 721–728.
24. Grindley, N. D. F., Lauth, M. R., Wells, R. G., Wityk, R. J., Salvo, J. J. & Reed, R. R. (1982) *Cell* **30**, 19–27.
25. Tse, Y. S., Kirkegaard, K. & Wang, J. C. (1980) *J. Biol. Chem.* **225**, 5560–5565.