Genes Whose Mission Is to Jump

Phase Variation: Evolution of a Controlling Element

Melvin Simon, Janine Zieg, Michael Silverman
Gail Mandel, Russell Doolittle

Summary. Phase variation in bacteria is regulated by homologous recombination at a specific DNA site. This recombinational event causes the inversion of a 970-base-pair DNA sequence that includes the promoter necessary for transcription of a flagellar gene. The invertible segment is flanked by two sites that are necessary for the inversion and contains a gene (*hin*) whose product mediates the inversion event. The *hin* gene shows extensive homology with the *TnpR* gene carried on the Tn3 transposon. It is also homologous with the *gin* gene carried on bacteriophage mu. These relationships suggest that the phase variation system may have evolved by the association of a transposon with a resident gene and the subsequent specialization of these elements to regulate flagellar antigen expression.

A great deal of genetic evidence suggests that gene rearrangements may be involved in regulating the nature of gene expression (*1*-*3*). Until recently, exploring this hypothesis has been difficult since testing it requires that the DNA corresponding to different forms of the gene be isolated and compared. The advent of new techniques in gene cloning has made it possible to do such experiments, and has led to the characterization of mobile elements in prokaryotes and eukaryotes. In general, these mobile elements or transposons have the ability to become associated with specific genes and to translocate them, as well as to induce various genetic rearrangements, including insertions, deletions, and inversions (*4*, *5*). McClintock's work in maize (*6*) suggested that mobile elements have components that control gene expression and that they could serve as a source for genetic regulatory units. Thus, elements of a transposon might be recruited by a gene system so that some of the properties of the transposon would be incorporated into the regulatory circuit for specific gene expression. McClintock suggested that these kinds of controlling elements are important in regulating cell differentiation and development.

Phase variation in *Salmonella* (*7*) follows a pattern indicating that it might be mediated by a classical controlling element. This system involves the alternation of expression of bacterial flagellar antigens. There are two genes, called *H1* and *H2*, that encode the protein flagellin, which is the main component of the flagellar filament. These genes are expressed alternatively, and a site adjacent to the *H2* gene is involved in regulating which of the genes is to be expressed. When this site is in the *H2* "on" state, the *H2* gene is expressed and an adjacent gene *rh1* is coordinately expressed. The product of the *rh1* gene acts as a repressor of the *H1* gene (*8*). Therefore, only *H2* type flagella are formed. When the controlling site is in the *H2* "off" configuration, neither the *H2* gene nor the *rh1* gene is expressed. The absence of repressor allows the transcription of the *H1* gene and the cells form *H1* type flagella. The controlling event is the inversion of a 970-bp (base pair) DNA sequence adjacent to the *H2* gene (*9*) (Fig. 1). The invertible region contains a promoter element necessary for the initiation of transcription of the *H2* gene (*10*). When this region is in the *H2* "on" orientation, the promoter is connected to the *H2* gene and transcription can proceed. When this region is inverted, the promoter is disconnected from the *H2* gene and transcription can occur in the opposite direction. Thus, the inversion behaves like a flip-flop switch activating and inactivating *H2* gene transcription. Depending on the frequency of switching, a fraction of the population expresses one flagellar antigen while the rest of the cells express the other flagellar antigen. The phase switch provides a model for examining the mechanism of recombinational regulation of gene expression.

Functional Components of the Phase Switch

The switch can be analyzed by determining the nature of the components that make it up and the molecular events that are involved in the transition from one state to the other. At least three components can be defined (*11*) with genetic and molecular techniques. Their relative map positions (Fig. 1) include:

M. Simon is a professor and, J. Zeig, M. Silverman, and G. Mandel are research investigators in the Department of Biology at the University of California, San Diego, La Jolla 92037. R. Doolittle is a professor in the Department of Chemistry at the University of California, San Diego, La Jolla 92037. M. Simon and M. Silverman are members of the Agouron Institute, La Jolla 92037.

0036-8075/80/0919-1370$01.25/0

1) Two sites, IRR and IRL (inverted repeat right and inverted repeat left, respectively), which map at the extreme ends of the invertible region. The deletion of either of these sites completely eliminates the ability of the region to invert.

2) A function that is essential for inversion occupies a genetic map position corresponding to two-thirds of the invertible region. Mutations or deletions that map within this function reduce the frequency of inversion by four orders of magnitude. This is a *trans*-acting function, that is, the defect can be corrected by the presence of an intact copy of the invertible region, suggesting that it is a gene that codes for a polypeptide necessary for the inversion process. We refer to the gene as *hin* since its product appears to be involved in mediating *H2* specific inversion.

3) A site contained within the invertible element that maps between the *hin* gene and IRR. Mutations or deletions at this site eliminate expression of the *H2* gene, but have no effect on the frequency of inversion. This site corresponds to the promoter for the *H2* gene that is contained within the inversion region.

The structural gene for the *H2* flagellin protein maps at a position outside of the invertible region (Fig. 1). Deletions or insertions in this region eliminate *H2* gene expression, but have no effect on the frequency of inversion.

These functions can be further resolved by determining the DNA base sequence of the invertible region. The DNA sequences that roughly correspond to each of the genetic functions are summarized in Fig. 2. We were able to determine the base sequence of all four products carrying the crossover points that result from the inversion (*12*). These sequences indicate that the inversion event occurs within the 14-bp inverted repeat sequences found at either end of the invertible region. These sequences must correspond at least in part to IRL and IRR and they represent the switching points. A homologous recombination event within these sites would result in the inversion of the 970-bp sequence between them. Figure 2 also shows the sequence for the *hin* gene. The data indicate that the only sequence which can be extensively translated corresponds to a polypeptide of 189 amino acids.

The sequence of the structural gene for *hin* starts at a position 75 bp within the invertible region. It is possible that the promoter that initiates transcription for *hin* could overlap with the inverted repeat sequence (IRL). If the *hin* gene

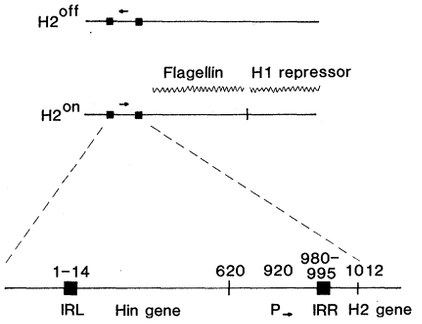


Fig. 1. Analysis of the components of the phase variation system. The top two lines schematically illustrate the mechanism of phase transition. The boxes indicate the invertible region and the horizontal arrow indicates the orientation of the DNA sequence. The lower line represents a map of the genetically defined functions that participate in switching. The numbers refer to the nucleotide sequence (Fig. 2).

product mediates inversion by binding to the IRR and IRL sites, it could also influence the rate of its own synthesis, since, in either orientation, one of these sites would be necessary for transcription of the *hin* gene. Figure 2 also shows the DNA sequence corresponding to the beginning of the *H2* gene. The structural gene sequence begins 17 bp outside of the invertible region. Thus, while the entire gene is contained outside the invertible region, the promoter that initiates transcription for the *H2* gene must be contained within the invertible region. This is consistent with the genetic map position specified by deletions that eliminate gene expression, but leave the *H2* structural gene intact.

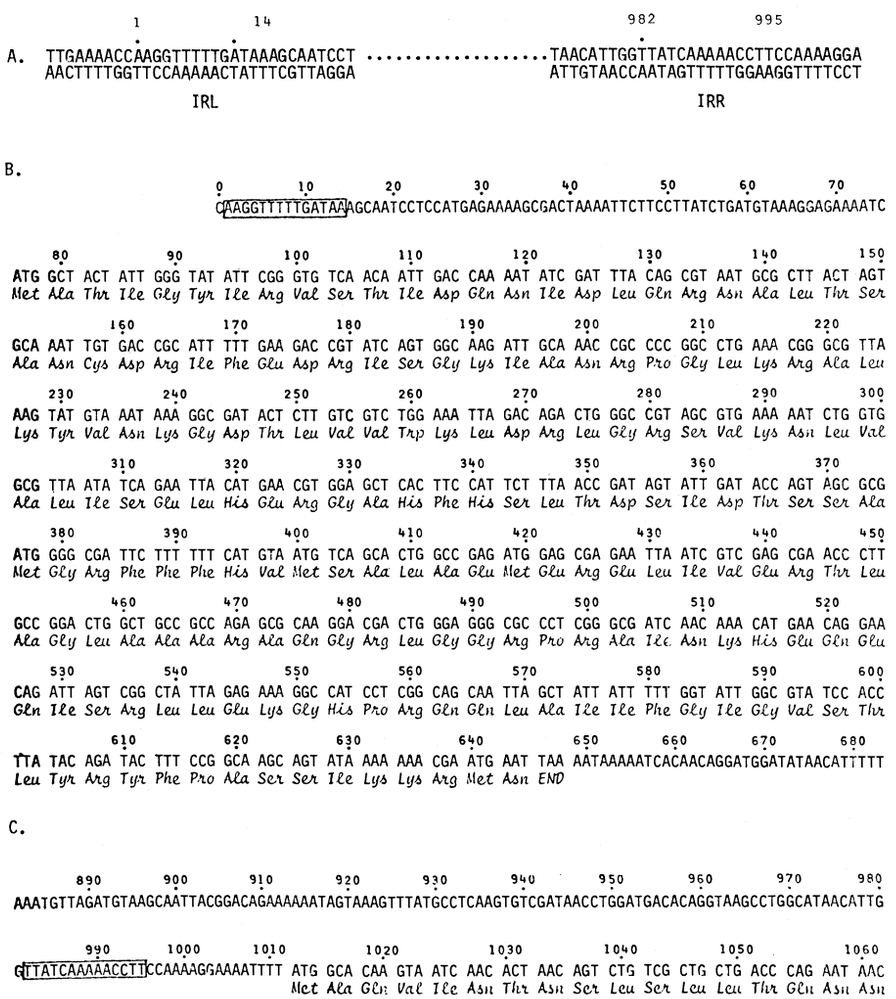


Fig. 2. The nucleotide sequence of the components of the switch. (A) The nucleotide sequence corresponding to the crossover points where inversion occurs. (B) The sequence corresponding to the *hin* gene. (C) The nucleotide sequence corresponding to the *H2* gene promoter, IRR, and the beginning of the *H2* structural gene. Abbreviations; A, adenine; T, thymine; G, guanine; C, cytosine; Ala, alanine; Asn, asparagine; Asp, aspartic acid; Arg, arginine; Cys, cysteine; Glu, glutamic acid; Gln, glutamine; Gly, glycine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; Val, valine.

Polypeptide Product of the *hin* Gene

We have used DNA carrying the *hin* gene sequence and mutations in this gene to define the nature of the polypeptide product of the *hin* gene. Figure 3 shows that an in vitro, coupled transcription-translation system is stimulated to synthesize a 19,000-dalton polypeptide when DNA fragments corresponding to the invertible region are added. The same polypeptide is seen when the system is programmed with plasmids that carry the invertible region; however, plasmids carrying DNA with deletions in the *hin* gene do not show the 19,000-dalton polypeptide. In order to examine the in vivo product of the *hin* gene, minicell preparations were used. Minicells containing plasmids that carry the invertible region synthesize a 19,000-dalton polypeptide, whereas minicells containing plasmids that had deletions, insertions, or mutations in the *hin* gene did not synthesize the 19,000-dalton polypeptide. Furthermore, a naturally occurring variant (*13*), called vH2$^-$, that lacks the ability to show phase variation was cloned. When DNA carrying the vH2$^-$ region is used in similar programming experiments, the 19,000-dalton polypeptide is not seen. Using Southern hybridization techniques we showed that vH2$^-$ corresponds to a deletion of a 100-bp DNA sequence that maps within the *hin* gene. We conclude that the presence of the 19,000-dalton polypeptide correlates with the presence of a function required to mediate the *H2* specific inversion event. This function is encoded by the *hin* gene.

Taken together the data give us a picture of the minimal number of components involved in the phase controlling element. Inversion apparently requires at least the presence of the *hin* gene product, the intact IRL and IRR sequences, and the inverted repeat configuration of these sequences. The simplest explanation of the data is that a homologous recombination event within the inverted repeat sequences allows reversible inversion to occur. It is not immediately clear what factors are involved in controlling the frequency of inversion. The amount of the *hin* gene product available may be critical and the *hin* gene product may act to repress its own synthesis. There may be other sequences adjacent to IRL and IRR that influence the rate of transcription of the *hin* gene. Furthermore, there may be other gene products that also affect the rate of *hin* transcription and the frequency of inversion. The *hin* gene product could directly mediate the inversion event. On the other hand, it may behave as a cofactor for a more general enzyme specified by a gene that is involved in site-specific homologous recombination.

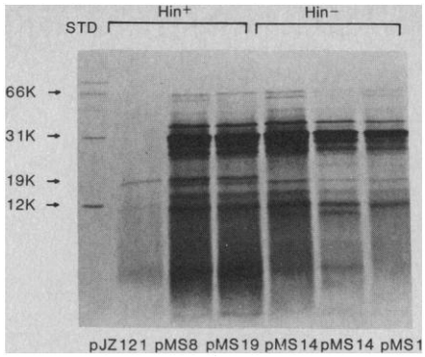


Fig. 3. Polypeptide synthesis directed by recombinant plasmids. The synthesis of [^{35}S]methionine-labeled polypeptides was programmed for a cell-free transcription-translation system, with several plasmids, and with DNA fragments that carried sequences corresponding to the *hin* gene. The column labeled pJZ121 represents a 1.6-kbp DNA fragment that includes the sequences that correspond to the *hin* gene. It was prepared by digestion of plasmid pJZ121 (*9*). Plasmids pMS8 and pMS19 contain the *H2* gene and an intact *Hin* region. Plasmids pMS14 and pMS1 contain the *H2* gene inserts derived by cloning the Hin$^-$ region from *H2* deletion mutants λ*fla*378 and λ*fla*381 (*11*) into plasmid pBR322.

Comparison with Other Site-Specific Recombination Systems

The results of this analysis of the molecular basis for phase variation presents some interesting questions, for example, how this system evolved, and what its relationship is to other similar systems. In a formal sense, phase variation may be classed with other site-specific homologous recombination events. Some of these events are diagrammed in Fig. 4. One system where site-specific inversion is known to regulate gene expression is the G loop in bacteriophage mu. This appears to be analogous to the phase variation system. The beta region of the mu genome contains a sequence of 3000-bp called the G loop, which is known to undergo inversion (*14*). There is evidence suggesting the presence of inverted repeat sequences at either end of the 3000-bp region. The inversion of the G loop is mediated by a polypeptide encoded by a gene called *gin* (G inversion) that maps adjacent to the invertible region. The inversion controls the expression of two sets of genes that are required for making the polypeptides involved in specifying the host range of mu (*15*). When the G loop is in one orientation (G+) the S and U polypeptides are synthesized and the resulting bacteriophage grow efficiently on *Escherichia coli* K12 as host. When the G loop is in the opposite orientation (G−), two other genes, *S′* and *U′*, are expressed and the phage now grow efficiently on *Citrobacter* or *Shigella* host strains. Other phage have analogous regions. A 3000-bp invertible loop with a great deal of homology to mu has been shown to exist in bacteriophage P1 and bacteriophage P7.

It is possible to test whether components of the phase variation system that regulated *H2* gene expression and elements of the G-loop system of bacteriophage mu are related to each other. Kutsukake and Iino showed that bacteriophages mu and P1 were able to cause phase variation to occur in vH2$^-$ mutants (*16*), suggesting that the bacteriophage supplied a function impaired by the vH2$^-$ mutation. We have also shown that lysogens of P1 and mu can complement Hin$^-$ mutants and restore phase variation. Furthermore, Kamp and Kahmann (*17*) found that the presence of DNA carrying the *hin* gene could restore function of Gin$^-$ mutants of mu. Thus, the *hin* function of the *Salmonella* phase variation system and the *gin* function of bacteriophage mu must be highly homologous.

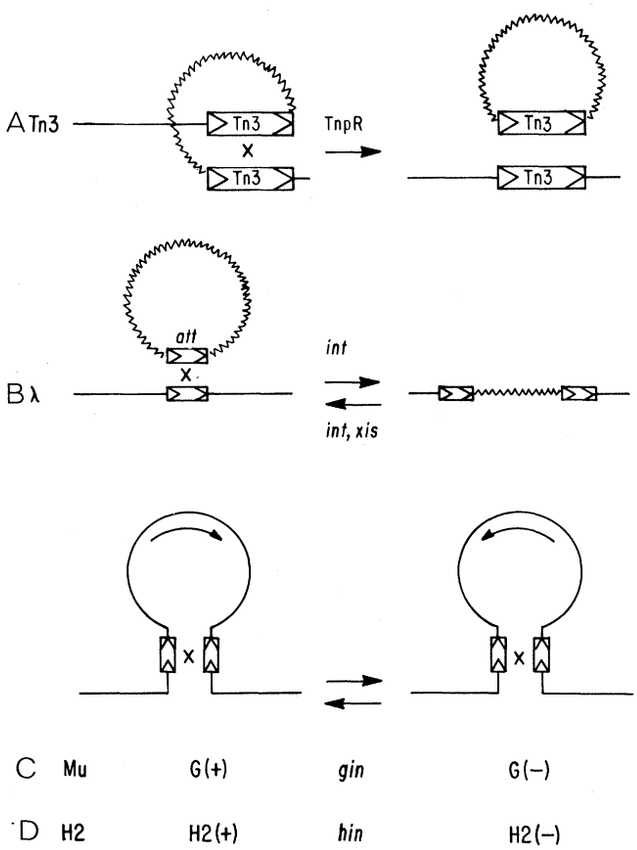
We can look for further analogies in other systems that involve site-specific recombination events. These need not be inversions since the configuration of the site at which the recombination event occurs affects the nature of the resulting products. Thus, if sites are in the direct repeat configuration, a homologous recombination event would lead to a deletion rather than an inversion. One example of a system where site-specific recombination between direct repeats occurs is the resolution of cointegrate structures of Tn3. Transposition appears to involve the replication of the transposable element (*18*), so that a copy is inserted at the new site. This reaction is mediated by a transposase, the product of the *TnpA* gene carried on the Tn3 transposon (*19*). If the transposon is involved in mobilizing adjacent genes, the product of the transposition event contains two directly repeated copies of the transposon which flank the non-transposon DNA (Fig. 4). Such cointegrate structures are usually not found with transposon Tn3 because the transposon contains a specific site (IRS) and a gene (*TnpR*) that codes for a product that mediates site-specific recombination between directly repeated IRS sites (*20*) resulting in the resolution of the cointegrate structure (Fig. 4). The entire sequence of the transposon Tn3 has been

determined (*21*). The *TnpR* gene encodes a protein that contains 185 amino acids and has a number of functions. In addition to being involved in the resolution of cointegrate structures, it acts as a repressor of the gene *TnpA* that codes for transposase and it acts as autogenous repressor to regulate its own synthesis.

We compared the DNA sequence and the putative amino acid sequence of the TnpR protein to those of the Hin protein. Optimum alignment of the sequence with a computer program revealed considerable homology between the two proteins; only two small gaps exist in the alignment (Fig. 5). It is apparent that there is considerable identity between the Hin polypeptide and the TnpR polypeptide. In fact, 64 of 185 amino acids are identical. Table 1 shows a comparison of a variety of other proteins and their degree of amino acid identity established in the same manner. Thus the TnpR protein and the Hin polypeptide appear to share a common ancestor. We have tested Tn3 to see if it will replace *hin* function, but we could not detect a TnpR-mediated H-loop inversion. There is another transposon, gamma-delta, found in *E. coli*, that is highly homologous to Tn3. The DNA sequence of the gene in gamma-delta that corresponds to the *TnpR* gene of Tn3 has been determined (*22*). When the amino acid sequence of this gene is compared to that of TnpR, 90 percent of the amino acids are identical for three-quarters of the protein starting from the NH_2-terminal end. The last quarter of the protein shows only 50 percent identity. A similar pattern is found in the comparison of the TnpR protein and the Hin protein where, from the NH_2-terminal, the first three-quarters of the polypeptide has 44 percent identity while the last quarter has only 26 percent identity. We conclude that, while *TnpR* may not have sufficient homology with *hin* to replace it functionally, nevertheless the two polypeptides are related.

Table 1. Proteins related by common ancestry and their percent identity.

Related proteins		Percent identity
Carboxypeptidase A (bovine)	Carboxypeptidase B (bovine)	47
Hemoglobin α chain (human)	Hemoglobin β chain (human)	45
Chymotrypsinogen A (bovine)	Trypsinogen (bovine)	42
Lysozyme (chicken)	Lactalbumin (bovine)	35
t Antigen (SV40)	t Antigen (polyoma)	29
Myoglobin (human)	Hemoglobin α chain (human)	28
Azurin (*Pseudomonas aeruginosa*)	Plastocyanin (spinach)	24
α Lytic protease (*Myxobacter*)	Chymotrypsinogen A (bovine)	18
Hin (*Salmonella typhimurium*)	TnpR (Tn3)	36



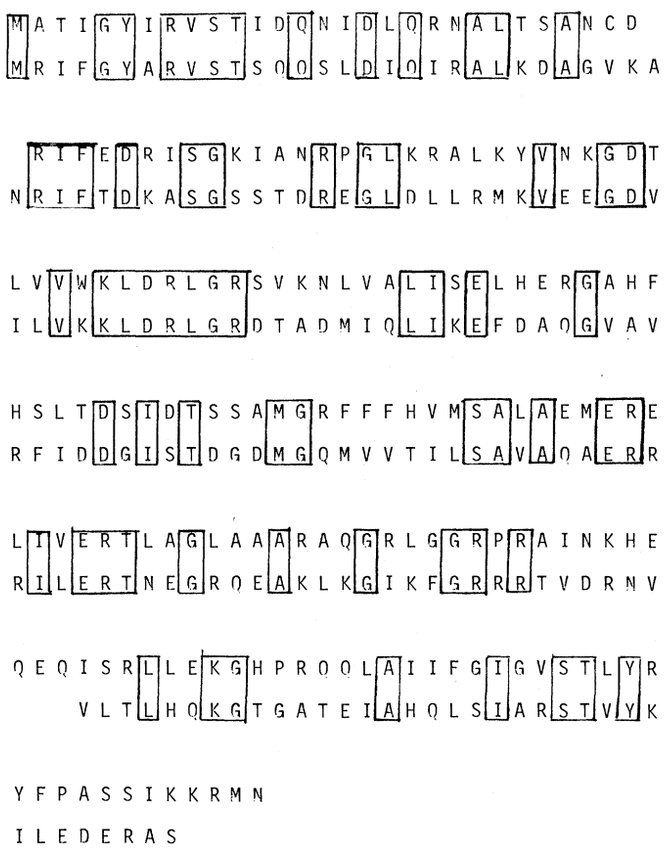


Fig. 4 (left). Systems in which site-specific recombination occurs. (A) Resolution of Tn3 cointegrate structure. The Tn3 cointegrate is shown on the left. This structure can be formed as a result of transposition. Thus, for example, if Tn3 initially associated with a plasmid (wavy line) transposes onto the genome (straight line) the resulting structure has two directly repeated copies of the Tn3 flanking the plasmid DNA. The resolution products of this structure are shown on the right. (B) The integration and excision of bacteriophage lambda (wavy line) from a specific attachment site (*att*) on the bacterial genome. (C) The inversion of the G loop in bacteriophage mu. (D) The H loop in the phase variation system. Fig. 5 (right). Amino acid similarities between the TnpR polypeptide and the Hin polypeptide sequence. The amino acid sequence of the Hin protein is shown on top and the TnpR sequence below it. Amino acid identities are marked by the boxes enclosing the amino acids in the two proteins. One-letter abbreviations for the amino acid residues indicated are: A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; Q, glutamine; E, glutamic acid; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.

Integration and excision of bacteriophage lambda also involves a site-specific recombinational event (Fig. 4). The core sequences at the lambda attachment site and in the corresponding sequences in the phage itself, bear marked similarities to the inverted repeat sequence in the phase variation system of *Salmonella* (*12*). We therefore compared the amino acid sequence of the Hin protein to that of the polypeptide corresponding to the *int* and *xis* genes. Together with host determined factors, these gene products mediate the integration and excision of bacteriophage lambda. We could find no significant similarity between the *hin* polypeptide sequence and the *int* and *xis* polypeptide sequence (*23*). Some of the factors determined by host genes that are necessary for integration, for example, the *him* A factor could be homologous to *hin* gene product. On the other hand, these two systems may not be directly related. There are probably a number of different families of sites and corresponding polypeptides that catalyze site-specific recombination. Parts of these systems could interact or they may be totally independent of each other.

From the point of view of the evolution of controlling elements, it is clear that there is a relationship between Tn3, bacteriophage mu, and the phase variation system. It has been suggested that there are similarities between the transposase genes specified by the *TnpA* gene in Tn3 and the *A* and *B* genes required for the transposition of bacteriophage mu (*21*). One might imagine that bacteriophage mu could have arisen from the insertion of a temperate phage genome into a site between the *TnpA* and *TnpR* genes of an ancestral Tn3. In fact, several scenarios are possible.

Phase variation appears to have evolved as the result of interchanges between mobile elements and resident genes. Site-specific recombination functions associated with transposons and perhaps viruses have been recruited to allow a new kind of regulatory mechanism that increases the diversity of flagellar antigen types and presumably provides survival value to the bacterial colony. These functions may still retain the ability to interact with other controlling elements, and there may be other regulatory units that are coordinated with the phase switch (*6*). These kinds of mechanisms could be useful in aiding bacteria to adapt to new environments and to interact with a variety of surfaces.

Recent studies of the structure of transposons and some viruses in eukaryotic systems indicate that they bear structural and functional similarities to bacterial transposons. Functions related to these mobile elements may also be found to be associated with specific recombinational events that affect gene expression in eukaryotes. In this regard, it may be more than coincidental that there is marked similarity between the sequences found at the switching points in the phase variation system

$$\overset{956}{\dot{C}}ACAGGT \ldots 23 \text{ nucleotide gap} \ldots \overset{985}{\dot{T}}CAAAAACC$$

as shown in Fig. 2 and the sequence at the point where site-specific recombination occurs to join the V and J regions in the formation of immunoglobulins (*24*)

$$CACAGTG \ldots 23 \text{ nucleotide gap} \ldots ACAAAAACC$$

These kinds of similarities may underlie some basic requirement for the physical association of DNA strands with each other or with proteins involved in site-specific recombination events.

The techniques and materials now available should allow us to determine the biochemical nature of the components involved in site-specific recombination events. Comparisons with other systems should yield further information about both the regulation of these events and the role that recombinational processes play in regulating gene expression, cell differentiation, and development.

References and Notes

1. B. McClintock, *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13 (1952).
2. M. Green, in *DNA Insertion Elements, Plasmids, and Episomes*, A. I. Bukhari, J. A. Shapiro, S. L. Adhya, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1977), pp. 437–445.
3. P. Starlinger and H. Saldler, *Curr. Top. Microbiol. Immunol.* **75**, 111 (1976).
4. N. Kleckner, *Cell* **11**, 11 (1977).
5. M. Calos and J. Miller, *ibid.* **20**, 579 (1980).
6. B. McClintock, *Cold Spring Harbor Symp. Quant. Biol.* **21**, 197 (1957).
7. J. Lederberg and T. Iino, *Genetics* **41**, 743 (1956).
8. H. Fujita, S. Yamaguchi, T. Iino, *J. Gen. Microbiol.* **76**, 127 (1973).
9. J. Zieg, M. Silverman, M. Hilmen, M. Simon, *Science* **196**, 170 (1977); J. Zieg, M. Hilmen, M. Simon, *Cell* **15**, 237 (1978).
10. M. Silverman, J. Zieg, M. Hilmen, M. Simon, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 391 (1979).
11. M. Silverman and M. Simon, *Cell* **19**, 845 (1980).
12. J. Zieg and M. Simon, *Proc. Natl. Acad. Sci. U.S.A.* **77**, 4196 (1980).
13. T. Iino, *Genetics* **46**, 1465 (1961).
14. E. Daniell, J. Abelson, J. Kim, N. Davidson, *Virology* **69**, 2823 (1972).
15. D. Kamp, R. Kahmann, D. Zipser, T. Broker, L. Chow, *Nature (London)* **271**, 577 (1978); M. Howe, *ibid.*, p. 608; P. Van de Putte, S. Cramer, M. Giphart-Gassler, *ibid.* **286**, 218 (1980).
16. K. Kutsukake and T. Iino, *ibid.* **284**, 479 (1980).
17. D. Kamp and R. Kahmann, in preparation.
18. N. Grindley and D. Sherratt, *Cold Spring Harbor Symp. Quant. Biol.* **43**, 1257 (1979); J. Shapiro, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 1933 (1979).
19. R. Gill, F. Heffron, S. Falkow, *Nature (London)* **282**, 797 (1979); J. Chou, P. Lemaux, M. Casadaban, S. Cohen, *ibid.*, p. 801.
20. R. Gill, F. Heffron, G. Dougan, S. Falkow, *J. Bacteriol.* **136**, 742 (1978); A. Arthur and D. Sherratt, *Mol. Gen. Genet.* **175**, 267 (1979).
21. F. Heffron, B. McCarthy, H. Ohtsubo, E. Ohtsubo, *Cell* **18**, 1153 (1979).
22. R. Reed and K. Shibuya, in preparation.
23. H. Miller and D. Friedman, *Cell* **20**, 711 (1980).
24. H. Sakano, R. Maki, Y. Kurosawa, W. Roeder, S. Tonegawa, *Nature (London)* **286**, 676 (1980); P. Early and L. Hood, *Cold Spring Harbor Symp. Quant. Biol.*, in press.

25. Supported by a grant from the genetics section of the National Science Foundation.

4 August 1980