BACTERIOPHAGE EVOLUTION AND THE ROLE OF PHAGES IN HOST EVOLUTION

Roger W. Hendrix

4

BACTERIOPHAGE POPULATION

The tailed double-stranded DNA bacteriophages have been evolving for perhaps 3 billion years or more, but it is only in very recent years that we have come to the beginning of a real understanding of the genetic mechanisms behind that evolution as well as an appreciation for the major role that phages have in the evolution of their bacterial hosts. Because the host range of a typical phage is narrow, estimates of the abundance of phages in the environment based on the number of plaques formed on a few bacterial strains that can be grown in the laboratory have been low by many orders of magnitude. It was only when environmental samples were examined directly by electron microscopy that it became clear not only that tailed phages are remarkably abundant in the environment but that they probably constitute a numerical majority of organisms on the planet. In the first such measurements (2), the concentration of particles with the characteristic morphology of tailed phages in Norwegian fjord water was about 10^7 per ml. Subsequent measurements from several other environmental sources have found similarly large, and in some cases, substantially larger numbers (31). Estimates of the total global population of phages can be made from these measurements and from the sizes of the different environmental compartments, and estimates are on the order of 10^{31} total viral particles. This is a truly astronomical number, in that 10^{31} tailed phages, laid end to end, would extend into space to a distance of 200 million light years. In all of the environmental samples examined, there were roughly 5 to 10 phage particles for every bacterial cell, which is the basis for the claim above that phages are a majority of the organisms on Earth.

The significance of 10^{31} phage virions in biological terms begins to become clear when one considers what is being learned about the dynamics of this population. Estimates of the longevity of phages in the environment (30) suggest that the entire population turns over every few days. The resynthesis of this huge population on such a short time scale requires on the order of 10^{24} infections per s, 24 h a day and 7 days a week. Furthermore, even though phages grow clonally for the most part, almost

Roger W. Hendrix, Pittsburgh Bacteriophage Institute and Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260.

Phages: Their Role in Bacterial Pathogenesis and Biotechnology
Edited by Matthew K. Waldor, David I. Friedman, and Sankar L. Adhya

every one of these 10^{24} infections per s offers an opportunity for genetic exchange with other phage sequences. This is because the great majority of bacterial genome sequences examined carry from one to many prophages, with which an infecting phage genome can in principle recombine (4, 5). This roiling global-scale orgy of bacteriophage sex has been going on for a very long time—possibly as long as there have been bacterial cells and almost certainly for a time measured in billions of years (10).

The amazing size of the global phage population, together with its similarly amazing dynamism, has a profound qualitative influence on how one can conceptualize the evolution of the population and its interactions with its hosts. As discussed below, comparative analyses of phage genome sequences imply that a major component of phage evolution is large numbers of intrinsically very improbable events. This view of the manner in which phage evolution occurs becomes plausible only with the realization that there has been ample opportunity for these events to have occurred, despite their improbable nature.

A window into the genetic structure and mechanisms of evolution of this population is afforded by the recent sequencing of a significantly large number of tailed phage genomes. At this writing, there are approximately 200 complete tailed phage genome sequences in the public databases; the rate of increase is increasing, and larger genomes (>100 kbp) are increasingly well represented. In addition, there are comparable numbers of prophage sequences that are found in the genomic sequences of bacteria and, sometimes, archaea. These numbers, of course, seem almost hopelessly small relative to the total number of phages in the global population. It remains to be seen how closely these few hundred genomes approach being representative of the whole population of 10^{31}. However, the prominence of horizontal exchange in phage evolution, as described below, gives hope that the phage gene pool is well stirred across the population and that individual phage genome sequences are closer to representative than would be expected in a more strictly clonal population.

MECHANISMS OF PHAGE EVOLUTION INFERRED FROM SEQUENCE DATA

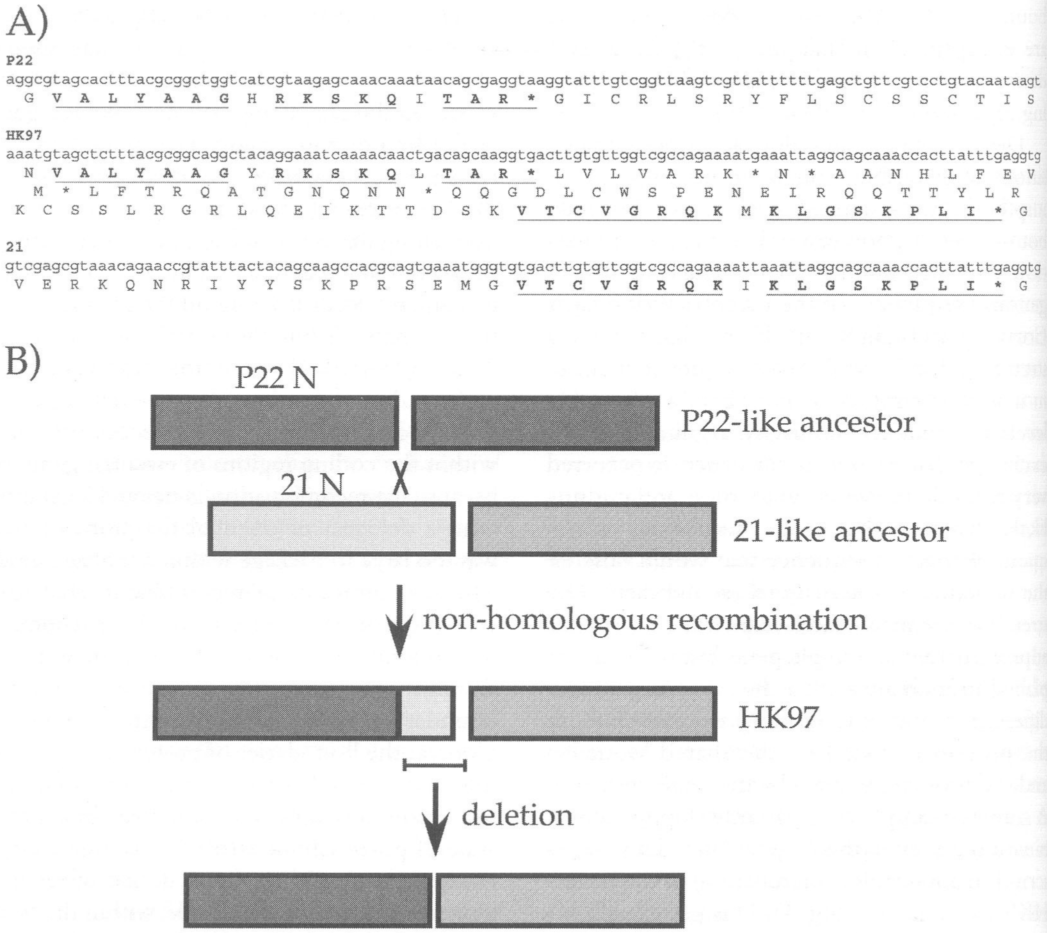
Mosaicism and Nonhomologous Recombination

If two phage genome sequences are compared and there is any detectable sequence similarity, then they are seen to be related in a mosaic fashion. In other words, there are stretches of sequence that match well, with abrupt transitions to regions with no detectable similarity or sometimes a different level of similarity. These transition points are considered the products of nonhomologous recombination events in the ancestry of one of the phages being compared; in this sense, they are fossils of past events in the history of the genome. It is noteworthy that, in general, these transitions are completely inapparent when a single genome is examined in isolation and only become detectable when two or more genomes are compared. It must also be true that the recombination points revealed by any pairwise comparison of genomes is only a subset of the nonhomologous recombination events in the history of either phage, as any recombination event that lies in the ancestry of both phages will not show up as a transition in the sequence comparison.

The sites of nonhomologous recombination detected in this way are not distributed randomly over the phage genome; rather, they are overwhelmingly located at gene boundaries (14, 24). In aggregate, these exchange points define modules of sequence that have moved horizontally between phage genomes in the evolutionary past. For the most part, the modules of exchange, so defined, are individual genes, but sometimes the modules are clusters of genes, most notably the structural genes for the heads or the tails. While the recombination sites usually do not fall within protein coding regions of the sequence, there are informative exceptions to this rule. Among these, the integrases of phages lambda and HK022 and the Erf recombination proteins of phages P22 and HK97, among others, show clear evidence of recombination within their coding regions, and in each of these cases, the exchange point corresponds to a well-defined protein domain boundary (14, 32). Genes encoding tail fibers are exceptional in that they often show evidence of multiple exchanges within their coding regions.

Despite their nonrandom locations, the positions of recombination junctions often appear imprecise, for example, with respect to gene boundaries. A good example is the recombination event implied by a comparison of the genome sequences of the mycobacteriophages Corndog and Che8 (24). These phages share a stretch of 368 bp with 100% sequence identity in a background of, at best, barely detectable levels of sequence similarity, arguing that the exchange that produced this structure occurred very recently in evolutionary time, and it is unlikely that there has been time for any subsequent changes in sequence that would obscure the structure of the original recombinant. The stretch of identical shared sequence corresponds approximately to a single gene, but the ends are placed in such a way that the two phages show different extents of C-terminal sequence in both the protein encoded by the shared sequence and the protein encoded by the gene upstream. A similar example of apparently sloppy recombination is seen in the *N* gene (encoding a transcription antitermination function) of the phages HK97 and 933W (Fig. 1). This gene has a sequence similar to that of the homologous gene of phage P22, including a termination codon at the expected position. However, the termination codon is followed immediately by a sequence in a different reading frame that is nearly identical to the last 17 sense codons plus the termination codon of the *N* gene of phage 21. The simplest explanation for the origin of this unusual structure, I would argue, is that there was a nonhomologous recombination event that was slightly out of register with respect to gene position in the ancestry of these phages, leading to a short quasi-duplication in the sequence. It seems unlikely that the headless tag end of a phage 21-like *N* gene would provide any useful function or selective benefit to the phages carrying it, but at the same time, it must not cause a significant selective detriment or it would not have survived in the population long enough for us to detect it.

The observations described above can be accommodated by a two-step Darwinian model of the evolution of phage genomes in which diversity in the phage population is initially generated by nonhomologous recombination between phage genomes, occurring at positions that are essentially random both in their positions along the genome and in the alignment of the recombining genomes. This is then followed by natural selection acting on the recombinants to eliminate all but those with no functional deficit. One would expect the vast majority of such random recombinants to be defective, either because of disruptive recombinations within the coding regions of essential genes or because the recombination generated a genome with a deletion of essential functions or that was too large to package within the phage head. The tiny minority of recombinants that survive selection will be those that have recombination junctions positioned so that they do no damage, i.e., those that are located at gene boundaries or, if they are within coding regions, at the boundaries of protein domains. In this case, as with other examples of evolution, the action of natural selection gives an appearance of purposefulness to what is essentially a random process. The relative deficit of recombination junctions, for example, within the head genes, is explained by the fact that the head proteins have coevolved to interact intimately with each other and so cannot successfully be mixed and matched with other lineages of head proteins. There is evidence that recombination per se is not excluded from the head gene region (14, 20); rather, almost all nonhomologous recombinations in the head region produce functional deficits and are presumably rapidly lost from the population. The case of tail fiber genes, in which multiple recombination junctions can be seen within the coding region, is rationalized by the fact that tail fiber proteins generally have an approximately one-dimensional fold, which means that the sort of intimate interactions between distant parts of the protein that would be disrupted by recombination are rare. Thus, it is imagined that the highly mosaic nature of many tail fiber genes is a closer approximation to the primary product



```
A)
P22
aggcgtagcactttacgcggctggtcatcgtaagagcaaacaaataacagcgaggtaaggtatttgtcggttaagtcgttattttttgagctgttcgtcctgtacaataagt
 G V A L Y A A G H R K S K Q I T A R * G I C R L S R Y F L S C S S C T I S

HK97
aaatgtagctctttacgcggcaggctacaggaaatcaaaacaactgacagcaaggtgacttgtgttggtcgccagaaaatgaaattaggcagcaaaccacttatttgaggtg
 N V A L Y A A G Y R K S K Q L T A R * L V L V A R K * N * A A N H L F E V
  M * L F T R Q A T G N Q N N * Q Q G D L C W S P E N E I R Q Q T T Y L R
K C S S L R G R L Q E I K T T D S K V T C V G R Q K M K L G S K P L I * G

21
gtcgagcgtaaacagaaccgtatttactacagcaagccacgcagtgaaatgggtgtgacttgtgttggtcgccagaaaattaaattaggcagcaaaccacttatttgaggtg
V E R K Q N R I Y Y S K P R S E M G V T C V G R Q K I K L G S K P L I * G
```

FIGURE 1 Hypothesis on the origin of the sequence around the end of the HK97 *N* gene. (A) The DNA sequence crossing the ends of the HK97 *N* gene is shown, with three frames of translation represented below. The N protein amino acid sequence is shown in the top translation frame and ends in the middle of the line. The corresponding part of the P22 *N* gene region is shown at the top, with amino acid identities indicated in bold and underlined. The end of the 21 *N* gene is shown at the bottom, with the termination codon near the right end of the line and identities indicated as described above. (B) The out-of-register nonhomologous recombination event that is hypothesized to have given rise to the HK97 sequence in this region is shown in the first reaction. The second reaction shows a hypothetical future deletion that would give this region of HK97 the appearance of having arisen through a single in-register nonhomologous recombination event.

of nonhomologous recombinations than is the case for the more stringently selected sequences in most of the rest of the genome.

Homologous Recombination, Point Mutation, Insertion, and Deletion

There are other important components of phage evolution in addition to nonhomologous recombination. Homologous recombination almost certainly occurs many orders of magnitude more frequently than nonhomologous recombination, but because it reconstitutes the original sequence, it does not leave a direct fossil of its occurrence in the form of a mosaic joint in the sequence. However, homologous recombination does have the potential to reassort the mosaic joints that flank the region of homology in the two recombining genomes and thus has the potential to mediate the rapid dissemination through the phage population of the mosaic joints created by nonhomologous recombination. Homologous recombination (and nonhomologous recombination as well) also has the potential to produce large-scale rearrangements of genomes to produce novel combinations of gene types and novel genome organizations. The *Escherichia coli* phage N15, for example, has head and tail genes with high sequence similarities to those of phage lambda, but the early gene region of the genome clearly has a very different ancestry from that of lambda.

Point mutation is also a major aspect of phage evolution: genes such as those encoding capsid components, which have apparently conserved functions, have diverged to the point that distant members of the family can have 15% amino acid sequence identity or less remaining. The degree to which point mutations generate novel functions and the degree to which they represent functionally neutral changes are not clear. However, one possible example of a novel function generated through point mutations is seen with the large group of major head proteins that are detectably related to the major head protein from the phage HK97. The HK97 protein forms autocatalytic covalent cross-links as the final step in head maturation. The corresponding proteins for some of the other phages in this group are also shown to cross-link, and these proteins, but not those which are known not to cross-link, share a characteristic trio of amino acids in the sequence alignment, namely, a lysine and an asparagine that form the cross-link and a glutamate that has an essential catalytic role (6, 7, 29; my unpublished observations).

Insertion and deletion, which are particular manifestations of nonhomologous recombination, round out the evolutionary mechanisms that are apparent from genome sequence comparisons. Deletions clearly offer a mechanism by which genes may be removed from a genome, but beyond that they may be an important part of the mechanism by which phage genomes have acquired their efficiently organized, seemingly premeditated, genome arrangements. One would expect, for example, that deletions removing nonessential DNA between essential genes would not be counterselected, and thus that essential genes would come, over evolutionary time, to be tightly packed on the genome, as is generally seen to be the case. In such a scenario, deletions may continue to occur between the genes as long as there is any DNA left to delete, but any deletions that stray into the essential gene coding sequences will be rapidly counterselected. An interesting variant of this argument is seen for a few cases in which there is a short quasi-duplication in the sequence of the sort that would be produced by a nonhomologous recombination between two genomes that is slightly out of register with respect to the positions of the genes. A deletion of the duplicated DNA back to the boundaries of the functional genes would leave a sequence junction that would appear to be the product of a perfectly in-register nonhomologous recombination event (Fig. 1B). It seems likely that many of the fossils of nonhomologous recombination that are detected at gene boundaries are the products of multiple events, in which an initial "sloppy" recombination event is subsequently cleaned up by one or more deletion events.

The insertion of novel DNA into a genome sequence can, in principle, occur by one or a series of nonhomologous recombinations. This must certainly occur, in much the same way that one can imagine random nonhomologous recombination, in conjunction with selection for undiminished function, causing the substitution of one gene for another. However, there is a recently recognized class of insertions into phage genomes for which it is possible to entertain the possibility of a more directed mechanism of insertion. These insertions are known as "morons" (units of more DNA) (11, 14). A moron is typically a sequence of a few hundred base pairs containing one or, less frequently, two plausible genes. The coding region in most cases is flanked by a putative promoter upstream and a stem-loop factor-independent transcription terminator downstream. Morons are identified as genes located between two genes whose homologs are adjacent in a comparison phage. In many cases the moron DNA has a distinctly different G+C content than the DNA flanking it, arguing for its recent introduction into the genome. The mechanism by which morons enter the genome, whether by random recombination events or a more directed mechanism, is unclear. They are typically found very neatly inserted between the flanking genes—that is, with essentially no extra DNA beyond the transcription control sequences—so an argument can be made that their recent entry into the genome has left too little time for them to have undergone the multistep mechanism of sloppy recombination followed by deletions to remove the extra DNA that was described above. This argument is weakened, however, by the fact that there is a lack of reliable estimates of either the rate at which the G+C content of newly introduced DNA comes to resemble that of the surrounding sequence or the rate at which deletions remove extra DNA. In any case, it is clear that morons represent one way that a phage genome can acquire a novel sequence.

Population Structure and Diversity

With the recent availability of numerous genome sequences, it is just now beginning to be possible to glimpse the types and diversity of tailed phages in the biosphere and to understand something of the genetic structure of the population. The most obvious messages from the genome comparisons are as follows. First, there has been a very large amount of horizontal exchange of genetic material between phages, probably by the mechanisms outlined above. Second, the lineages of the genes seen in phages are extremely old, as judged by the high degree of divergence in their sequences. Exactly how old they are is not clear because it has not been possible to calibrate the mutational clock for phages. However, current guesses, tentative though they are, place the origin of tailed phages as far back as the time of divergence of bacteria from the archaeal and eukaryotic lineages or earlier (10).

Given the enormous population of tailed phages on Earth, the facts that a phage infects a bacterial cell on the order of 10^{24} times per s and that phages have been at this activity for a period on the order of 10^{17} s, and the fact that most bacteria have one or more prophages in their genome with which an infecting phage may recombine, one might expect that the genes of the global phage population would by now be thoroughly mixed into a uniformly varying genetic equivalent of a smooth chocolate pudding. There is, in fact, evidence of horizontal exchange of phage genes across the entire gamut of sequenced phages (12), arguing that the entire population of tailed phages is a single genetic population. However, the sequence comparisons also make it clear that the chocolate pudding is somewhat lumpy, that is, that the rate of gene exchange can be very different between different pairs of points in sequence space and that the types of phage genomes that can be found do not vary smoothly across sequence space. This statement must be qualified by the fact that our current sampling of the global phage population is extremely sparse, and it is not uncommon for a newly sequenced phage to have some properties that are associated with one well-studied phage and other properties characteristic of a different well-studied phage that until that point was not thought to have any features in common with the first (18, 26).

It is believed that there are at least two factors that can restrict the horizontal flow of genes across the expanses of phage sequence space. The first of these is that phages tend to have rather narrow host ranges. For this reason, one would expect that phages with high levels of similarity in sequence and genome organization would most often infect the same or similar host cells, in which they would have maximum chances of encountering each other, and this expectation is generally supported by the data. Over the longer evolutionary term, genes appear able to breach these host range barriers, but long journeys across phylogenetic space evidently occur in a large number of small steps, as judged by the fact that homologous genes found in phages whose hosts are phylogenetically distant are never very close in sequence (12). A second factor that resists gene flow is the differences among phage lifestyles. For example, *E. coli* K-12 is a host to both phage lambda and phage T4, two phages with very different lifestyles, and there is very little detectable sequence similarity between the two phages. It is not the case that lambda and T4 are unable to exchange DNA: they share the C-terminal end of a tail fiber protein and the adjacent tail fiber assembly chaperone, and the exchange took place recently enough in their ancestries that the chaperone proteins from the two phages are functionally interchangeable (9). One can imagine, rather, that the requirements for success in the two ecological niches that these phages occupy are sufficiently different that almost all DNA exchanges that have occurred (with the evident exception of that of the tail fiber sequences) have failed to provide enough benefit to the recipient phage to persist in that genome. One factor that apparently does not provide a barrier to gene exchange is geographic separation. Phage lambda, isolated as a prophage from a hospital patient in California in 1922, and phage HK97, isolated in the 1970s from a pig sty in Hong Kong, have immunity repressor sequences with 99% identical amino acid sequences and 97% identical nucleotide sequences in the context of genomes with large regions of no detectable identity. More strikingly, mycobacteriophages Corndog and Che8 were isolated in Pittsburgh, Pa., and Chennai, India, and share a stretch of 368 bp with 100% identity. Both of these examples and others that could be cited argue that the time required for phages to traverse large geographical distances is comparable to or shorter than the time required for the phages to accumulate mutations.

ROLE OF PHAGES IN HOST EVOLUTION

Phages have been understood to influence the evolution of their hosts since early experiments showed that the presence of the phage T2 could select for T2-resistant *E. coli* (21). Numerous other examples can be cited to show how the cell defends against infecting phages, and interestingly, many of these are provided by genes of other phages residing in the cell as prophages. They include such things as prophage immunity, exclusion systems like the lambda *rex* genes, which excluded T4*rII* mutants in Benzer's classical experiments (1), restriction enzymes, and a variety of others. However, the benefits that prophages supply to their hosts are hardly restricted to defense against infection by other phages. The best known examples of other functions of this sort are probably toxin genes of human and agricultural pathogens that are carried into the cell as part of the genome of a prophage. Thus, the toxin genes for botulism, diphtheria, ovine foot rot, and cholera, among many others, are prophage genes. In most cases, these toxin genes are expressed from the otherwise repressed prophage (but for a somewhat more complex example, see references 27 and 28). It is generally assumed that the toxin gene confers a selective benefit on the host cell by making it a more effective pathogen.

The ability to be an effective human pathogen is presumably of great importance to the small number of bacteria that have adopted that lifestyle, and these same bacteria are understandably of great interest to the scientists who study them (who are themselves typically of the human persuasion); these considerations can thus explain the large amount of attention that these bacteria and their phage-derived toxin genes have received. Similarly, phages that provide their hosts with defenses against infection by other phages produce phenomena that are readily apparent in a laboratory setting and are attractive targets for experimental investigation. However, if the phages that infect these well-studied strains of bacteria so frequently carry genes that are beneficial to the host bacteria on their prophage DNAs, it seems almost certain that similar interactions must take place in the vast expanses of bacterial phylogenetic space that are less well characterized experimentally. One can imagine that regardless of the ecological niche a bacterium may occupy, there are likely temperate phages that carry genes that can be brought into the bacterial genome as part of a prophage and that can provide a selective advantage to the lysogenic cell in its particular ecological niche.

Assuming that this scenario is correct, one can wonder, first, what benefit the phage derives from this arrangement that makes it worthwhile to carry genes that directly benefit the host, and second, how a given phage "knows" which genes will benefit its particular host. Perhaps the most obvious answer to the first question is that anything that contributes to the successful propagation of the host genome contributes ipso facto to the propagation of the phage DNA that is carried as a prophage. A somewhat less straightforward view, though not mutually exclusive with the first, sees the beneficial gene(s) as a sort of "rent" that the prophage pays to the host for the benefit of not being evicted. It has been argued that the property of bacterial recombination systems of being able to work with very short stretches of homology may have been selected because it allows them to remove, by deletion, dangerous genetic parasites such as prophages and transposons (19). If a prophage that is a potential target of the host deletion machinery carries a gene that is beneficial to the host, then deleting the prophage, while it may improve the cell's chances of avoiding death by prophage induction, comes at the cost of the loss of the beneficial gene. Cells that have not deleted the prophage, in this view, will have a selective advantage relative to those that have deleted the prophage and thus will tend to persist.

Many of the beneficial genes carried by prophages appear to be morons, i.e., the genes described above that have entered the genome recently and are typically flanked by a transcription promoter and a terminator. It is believed that morons are added to the phage genome repeatedly over evolutionary time and that the ones that survive in the genome will tend to be those that provide a selective benefit to the phage, either directly during lytic growth or, possibly more frequently, indirectly by benefiting the host bacterium as discussed above. In this way, the phage genome will preferentially end up carrying those genes that benefit the particular strain of bacteria that the phage infects. One can carry this line of argument farther by suggesting a sort of "moron cycle" by which phages acquire, as morons, novel genes that will be beneficial to the host and contribute them in a multistep process to the bacterial genome. This formulation assumes that a prophage in a bacterial genome is subject to deletion, either in part or in its entirety. Deletions that remove the beneficial gene(s) that the prophage carries will put the cell at a selective disadvantage, and disadvantaged cells will tend to be lost from the population. On the other hand, deletions that remove only nonbeneficial phage genes will be either approximately selectively neutral for the cell, or if they decrease the ability of the prophage to kill the cell by induction, selectively positive. Over the long term, the expectation is that all of the prophage genes will be deleted, with the exception of those that provide a benefit to the host. At this point, the remaining beneficial gene will no longer be in the context of other phage genes and thus will no longer be identifiable as having a different history than any other bacterial gene. It may be possible to test the reality of this scenario by examining the gene content of defective prophages that have lost some but not all of their genes; a survey of a small number of lambda-like defective prophages (19) seems consistent with this view in that they preferentially lack the genes that seem most likely to be harmful to the host. However, much more data will be needed to make a clear case.

In addition to the probability that phages donate individual genes to a bacterial genome by the process described above, there is rather more straightforward evidence that phages have been the source of whole sets of virion structural genes, which appear to have been appropriated by the host. The best known of these are sets of phage tail genes that produce bacteriocins that look like phage tails. *Pseudomonas aeruginosa* has two sets of such tail genes (23). One encodes a structure called the R-pyocin, which resembles the contractile tail of phage P2 in morphology as well as in organization and the sequence of the encoding genes. The second specifies the F-pyocin, which closely resembles the flexible tails of phage lambda by the same criteria. These phage tail structures are apparently weapons of competition that act against closely related bacteria by adsorbing their ends to a cell and collapsing its membrane potential, presumably by making a hole through the cell envelope as if they were acting as part of an intact virion. The two sets of tail genes are encoded adjacent to each other in the bacterial genome, positioned between genes of the *trp* operon, and their expression is coordinately regulated. It seems likely that these genes started out their association with the cell as parts of prophages but that they now function strictly as part of the bacterial genome.

A somewhat different example of phage structural genes that were apparently coopted for the benefit of the host cell is the gene transfer agents (GTA) of *Rhodobacter capsulatus*. GTAs were first identified as agents of generalized transduction between conspecific *Rhodobacter* cells and were later shown to be small-headed tailed-phage-like particles that package 4.5-kbp pieces of genomic double-stranded DNA derived from locations distributed uniformly around the *Rhodobacter* genome (17). The genes encoding the particles are located in a contiguous block of 15 kbp in the cell's genome. They include an apparently complete set of phage head and tail genes arranged in canonical order, with some showing high levels of similarity to known phage genes, but there are no candidate early phage genes in the vicinity (16). The expression of the genes is controlled in response to environmental conditions and is mediated by a two-component signal transduction system (15). It is clear that the GTAs cannot function as a normal phage because their small heads can only hold about 30% of the DNA that encodes the particles. However, as with the pyocins discussed above, it seems most likely that the GTAs were derived from a complete prophage and have now been taken over by the cell for its own benefit. There are GTA-like particles described for different bacterial genera (13, 25) and for an archaeon (3, 8), arguing against the possibility that the *Rhodobacter* GTAs simply happened to be caught midway through the process of genetic decay. Rather, it appears likely that the GTAs are being maintained by selection for a beneficial function that they supply to the host. What that function might be is not known; perhaps the most plausible guess is that they afford a means of genetic exchange among members of the host cell population, thereby helping to avoid the mutational collapse implied by Müller's ratchet (22).

PROSPECTS AND CAVEATS

Our understanding of how phages evolve began in the late 1960s with the heteroduplex mapping of the chromosomes of phage lambda and some close relatives, showing that these molecules are mosaic with respect to each other. That understanding has acquired a considerable amount of breadth and specificity with the availability of multiple whole genome sequences and with a better understanding of the size and activity of the global population of phages. However, as the reader will appreciate from the foregoing discussion, our understanding of evolutionary processes is based almost entirely on inferences about past events based on sequences of contemporary genomes. Furthermore, in most cases the number of individual examples that are available for a particular class of "fossil" (for example, the case illustrated in Fig. 1) can be quite small. For this reason, the robustness of our understanding of evolutionary processes in phages and of the interaction of phages with their hosts' evolution can be expected to improve substantially as more genome sequences become available. Better computer-based methods of analysis can also be expected to improve our ability to analyze the data and to handle the rapidly increasing volume of data available for analysis. At the same time, it will be important to remember that many of the most telling features of sequence comparisons were not anticipated and therefore that their discovery would have been difficult to automate.

In addition to broadening our understanding of the sorts of evolutionary mechanisms discussed here, there are two other areas of investigation that will clearly benefit from continuing work on comparative phage genomics. First, we will improve our understanding of the magnitude of the diversity of phage gene sequences. It is already clear that the diversity of gene types in the phage population is huge and probably larger than that in any other compartment of the biosphere, but the limits of that diversity have not yet been glimpsed. One of the most important benefits of future work in this area will be a better understanding of the extent of the phage gene pool and the ways that phage genes are moving around among the phage population and between phages and their hosts. A second area of considerable promise will be to extend what has been learned about the evolution of "ordinary" phages to phages with much larger genomes. Genome sequences for such phages are just now becoming available, and the largest genome to date is 10 times as big as that of phage lambda, with a corresponding increase in gene number. Understanding the life strategies that support the maintenance of such large numbers of genes promises to be both fascinating and informative.

ACKNOWLEDGMENTS

I thank Sherwood Casjens, Graham Hatfull, Jeffrey Lawrence, Maggie Smith, and other members of the phage evolution community for ongoing stimulating discussions and ideas.

Work in my laboratory on this topic is supported by grants from the National Institutes of Health (GM51975) and the National Science Foundation (EF0333112).

REFERENCES

1. **Benzer, S.** 1955. Fine structure of a genetic region in bacteriophage. *Proc. Natl. Acad. Sci. USA* **41:**344.
2. **Bergh, O., K. Y. Borsheim, G. Bratbak, and M. Heldal.** 1989. High abundance of viruses found in aquatic environments. *Nature* **340:**467–468.
3. **Bertani, G.** 1999. Transduction-like gene transfer in the methanogen *Methanococcus voltae. J. Bacteriol.* **181:**2992–3002.
4. **Canchaya, C., C. Proux, G. Fournous, A. Bruttin, and H. Brussow.** 2003. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67:**238–276.
5. **Casjens, S.** 2003. Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49:**277–300.
6. **Duda, R. L.** 1998. Protein chainmail: catenated protein in viral capsids. *Cell* **94:**55–60.
7. **Duda, R. L., J. Hempel, H. Michel, J. Shabanowitz, D. Hunt, and R. W. Hendrix.** 1995. Structural transitions during bacteriophage HK97 head assembly. *J. Mol. Biol.* **247:**618–635.
8. **Eiserling, F., A. Pushkin, M. Gingery, and G. Bertani.** 1999. Bacteriophage-like particles associated with the gene transfer agent of Methanococcus voltae PS. *J. Gen. Virol.* **80:**3305–3308.
9. **Hashemolhosseini, S., Y. D. Stierhof, I. Hindennach, and U. Henning.** 1996. Characterization of the helper proteins for the assembly of tail fibers of coliphages T4 and lambda. *J. Bacteriol.* **178:** 6258–6265.
10. **Hendrix, R. W.** 1999. Evolution: the long evolutionary reach of viruses. *Curr. Biol.* **9:**R914–R917.
11. **Hendrix, R. W., J. G. Lawrence, G. F. Hatfull, and S. Casjens.** 2000. The origins and ongoing evolution of viruses. *Trends Microbiol.* **8:**504–508.
12. **Hendrix, R. W., M. C. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull.** 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. USA* **96:**2192–2197.
13. **Humphrey, S. B., T. B. Stanton, N. S. Jensen, and R. L. Zuerner.** 1997. Purification and characterization of VSH-1, a generalized transducing bacteriophage of *Serpulina hyodysenteriae. J. Bacteriol.* **179:**323–329.
14. **Juhala, R. J., M. E. Ford, R. L. Duda, A. Youlton, G. F. Hatfull, and R. W. Hendrix.** 2000. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299:**27–51.
15. **Lang, A. S., and J. T. Beatty.** 2002. A bacterial signal transduction system controls genetic exchange and motility. *J. Bacteriol.* **184:**913–918.
16. **Lang, A. S., and J. T. Beatty.** 2001. The gene transfer agent of Rhodobacter capsulatus and "constitutive transduction" in prokaryotes. *Arch. Microbiol.* **175:**241–249.
17. **Lang, A. S., and J. T. Beatty.** 2000. Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of Rhodobacter capsulatus. *Proc. Natl. Acad. Sci. USA* **97:**859–864.
18. **Lawrence, J. G., G. F. Hatfull, and R. W. Hendrix.** 2002. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184:**4891–4905.
19. **Lawrence, J. G., R. W. Hendrix, and S. Casjens.** 2001. Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* **9:**535–540.
20. **Liu, J., and A. Mushegian.** 2004. Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. *J. Bacteriol.* **186:**4369–4375.
21. **Luria, S. E.** 1945. Mutations of bacterial viruses affecting their host range. *Genetics* **30:**84.
22. **Müller, H. J.** 1950. Our load of mutations. *Am. J. Hum. Gen.* **2:**111–176.
23. **Nakayama, K., K. Takashima, H. Ishihara, T. Shinomiya, M. Kageyama, S. Kanaya, M. Ohnishi, T. Murata, H. Mori, and T. Hayashi.** 2000. The R-type pyocin of Pseudomonas aeruginosa is related to P2 phage, and the F-type is related to lambda phage. *Mol. Microbiol.* **38:**213–231.