Virus Evolution; Bacterial Viruses

R W Hendrix, University of Pittsburgh, Pittsburgh, PA, USA

Glossary

Chromosome The physical DNA or RNA molecule present in the virion and containing the information of the genome. The chromosome can differ from the genome by, for example, having a terminal repetition of part of the genomic sequence or being circularly permuted relative to other chromosomes in the population.

Genome The totality of the genetic complement of a virus. The genome is a conceptual object usually expressed as a sequence of nucleotides.

Homologous Having common ancestry. Homology of two sequences is often inferred on the basis of a high percentage of identity between the sequences, but homology itself is either present or not and is not expressed as a percentage.

Novel joint A novel juxtaposition of sequences in a genome resulting from a nonhomologous recombination event.

Bacteriophage Evolution

Bacterial viruses ('bacteriophages' or 'phages' for short) have probably been evolving for 4 billion years or more, but it is only in recent years that we have come to a relatively detailed view of the genetic mechanisms that underlie phage evolution. Phages do not leave fossils in the conventional sense, but the sequences of phage genomes contain more detailed information about how the phages have evolved than any conventional fossil has. Because of the importance of sequence, our understanding of phage evolution has increased dramatically as more phage sequences have become available over the past few years. Even more important than the individual sequences, however, is the fact that we have multiple genome sequences that we can compare to each other. It is in such comparisons of different genomes that we see the unmistakable signatures of past evolutionary events, which would be completely invisible in an examination of a single genome alone.

Most of the work on phage evolution has addressed evolution of the double-stranded DNA (dsDNA) tailed phages, the members of the order Caudovirales, and most of the discussion here deals with that group. Toward the end of the article, we will consider evolution of other groups of phages as well as recent evidence suggesting deep evolutionary connections between some of the bacteriophage groups and viruses that infect members of the Eukarya and Archaea.

The Tailed Phage Population

Nothing in phage evolution makes sense except in the light of what we have learned about the nature of the global population of tailed phages. The most remarkable fact is that the size of the population is almost incomprehensibly large – a current estimate is that there are about 10^{31} individual phage particles on the planet. To give a feeling for the magnitude of this number, if these phages were laid end to end, they would reach into space a distance of ~200 million light years. This is apparently a very dynamic population; ecological studies of marine viruses suggest that the entire population turns over every few days. To replenish the population would then require about 10^{24} productive infections per second on a global scale. Each such infection is an opportunity for evolutionary change, either by mutation during replication or by recombination with the DNA in the cell, which will almost always include DNA of other phages in the form of resident prophages. At a lower frequency, cells may be co-infected by two different phages, affording a different opportunity for genetic exchange. As described below, the evolutionary events that are inferred from genome sequence comparisons include extremely large numbers of improbable events, and the only way these events could have given rise to the existing phage population is if they have had an extremely large number of opportunities to occur.

Evolutionary Mechanisms in the dsDNA Tailed Phages

The Nature of the Genomes

The genomes of the dsDNA tailed phages use DNA very efficiently, with typically about 95% of the sequence devoted to encoding proteins. Spaces between genes can often be identified as containing expression signals such as transcription promoters or terminators. Genes are typically arranged in groups that are transcribed together.

The specific types of genes in a genome, their numbers, and how they are arranged along the DNA are all highly variable among phages. Genome size is also variable, with the smallest known phages in this group having genomes of about 19 kbp and about 30 genes and the largest consisting genomes of 500 kbp and nearly 700 genes.

s0025

Types of Evolutionary Changes

p0025 Comparisons of genome sequences between phages reveal the sorts of changes in the genome sequence that in aggregate constitute phage evolution. The first of these is point mutation, in which one base pair is substituted for another. The number of point mutational differences seen between two homologous sequences varies from none to so many that our ability to detect any residual similarity of the sequence, often measured at the more sensitive level of the encoded amino acid sequence, has vanished. The amount of difference between two sequences due to point mutation is a function of how long it has been since they diverged from a common ancestral sequence, because such differences accumulate progressively over time. In practice, however, it is an unreliable measure of time for a number of reasons: we do not know the mutational rates for phages in a natural setting, we do not know that those rates have been constant over evolutionary time, and different genes accumulate changes at different rates because any mutations that are detrimental to the function of the encoded protein will be lost from the population by natural selection, and different proteins have different tolerances for mutational changes in their amino acid sequences.

p0030 The other major type of evolutionary change to a genome is DNA recombination. Particularly important is 'illegitimate' or 'nonhomologous' recombination, in which two different DNA sequences are joined together to form an association of sequences that did not exist before, known as a 'novel joint'. Nonhomologous recombination can join parts of two or more genomes into one new genome; it can also cause deletion or inversion of a sequence within a genome by mediating recombination between two sequences in the genome. Nonhomologous recombination can in principle – and, it appears, also in practice – produce virtually any novel joint that can be imagined. The results of laboratory experiments have shown that homologous recombination happens many orders of magnitude more frequently than nonhomologous recombination. Unlike nonhomologous recombination, homologous recombination does not leave a novel joint that can be detected in analysis of the sequence, but it has the potential to reassort any flanking novel joints, bringing them and their associated sequences together in new combinations and providing a mechanism for them to move rapidly through the population.

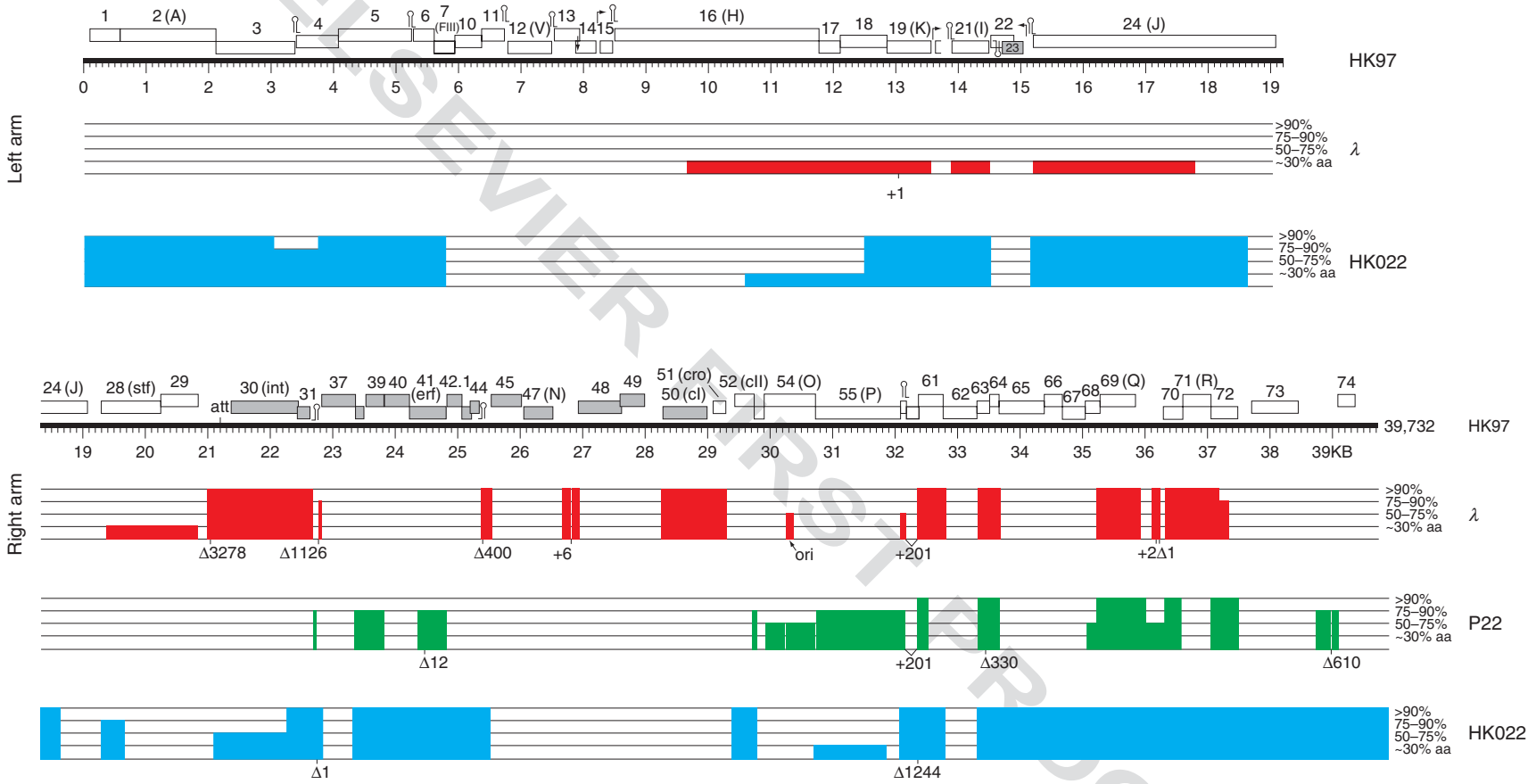
Genome Comparisons

s0030

p0035 There is typically nothing in a single phage genome sequence, viewed in isolation, that reveals the presence of changes in the phage's evolutionary past, of either mutational or recombinational origin. When multiple genomes are compared, a wealth of differences corresponding to such changes is revealed. **Figure 1** illustrates a comparison for a group of four rather similar genomes of phages infecting enteric bacterial hosts. The physical gene map shown is that of *Escherichia coli* phage HK97, and the histograms indicate the locations of sequences in the other phages that match HK97. It is apparent that the genome sequences match each other in a patchwork fashion, and the parts of the sequence that match the HK97 sequence are different for the three phages. In other words, the genomes are genetic mosaics with respect to each other. In a pairwise comparison of genomes, the transitions between where sequences match and where they do not are abrupt, even when they are examined at the level of the nucleotide sequence. This implies that there have been nonhomologous recombination events in the ancestry of the phages, creating novel joints in the resulting recombinant phages. These novel joints are detected when they are compared with a sequence that did not suffer that particular recombination event. In the comparisons shown in **Figure 1**, there is evidence for about 75 ancestral nonhomologous recombination events, occurring in either the ancestry of HK97 or in that of one of the three phages being compared. It is worth noting that this probably underestimates the number of such events, because any novel joints that lie in the common ancestry of all these phages would not have been detected.

p0040 A striking feature of the novel joints in the sequence revealed by such genome sequence comparisons is that they fall predominantly at gene boundaries. While some are precisely at the gene boundaries, some also fall in the middle of spaces between genes or even a few codons into the upstream or downstream gene's coding region. More rarely, we can see novel joints in the interior of a gene's coding region. In the case of some such genes for which the encoded proteins are well characterized, the novel joints in the DNA sequence fall at a position corresponding to a domain boundary of the protein. Another feature of the genome comparisons that extends across all the groups of phages examined to date is that there are clusters of genes that are never, or rarely, separated by nonhomologous recombination. These are typically genes whose protein products are known to function together, such as the genes encoding the structural proteins of the head.

p0045 A frequent consequence of nonhomologous recombination among viruses is the transfer of genetic material from the genomes of one viral lineage into the genomes of a different viral lineage, a process known as 'horizontal



f0005 Au3

Figure 1 The horizontal black line, scaled in kilobase pairs, represents the genome sequence of *Escherichia coli* phage HK97. The genes are indicated above the line by rectangles, with open rectangles indicating genes transcribed rightward and shaded rectangles genes transcribed leftward. The colored histograms indicate the areas of sequence similarity between HK97 and the three phages indicated to the right of each histogram: *E. coli* phages lambda and HK022 and *Salmonella enterica* phage P22. Adapted from Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, and Hendrix RW (2000) Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. *Jounal of Molecular Biology* 299: 27–52.

transfer' or 'lateral transfer' of genes. This is the process that gives rise to the mosaicism in the genomes discussed above. It also means that different parts of a genome may, and most often do, have different evolutionary histories. This last fact leads to an interesting and still unresolved difficulty for viral taxonomists. That is, attempts to represent the relationships among viruses by a hierarchical taxonomy, as is conventionally done, necessarily fail to capture the multiple different sets of hierarchical relationships displayed by the individual exchanging genetic modules, deriving from their different evolutionary histories. The viruses, of course, are unaware of human attempts to classify them and so are unaffected by the resulting controversies.

s0035

Evolutionary Mechanisms

p0050 When the mosaicism of phage genomes was first seen in DNA heteroduplex mapping experiments in the late 1960s, it was proposed that there might be special sites in the DNA, possibly at gene boundaries, that served as points of high-frequency nonhomologous recombination, either through short stretches of homology or through a site-specific recombination mechanism. This view was formulated as the 'modular evolution' model of phage evolution, in which it was proposed that exchange between genomes took place repeatedly at these special sites, leading to the observed mosaic relationship between genomes.

p0055 With the current availability of many more genomes and data at the level of nucleotide sequence, it has become clear that, while the mosaic results of phage evolution are much as described nearly 40 years ago, the mechanisms by which that state is achieved are fundamentally different than what had been proposed. That is, the observations are best explained by a model of rampant nonhomologous recombination among phage genomic sequences, with the sites of recombination being distributed, to a first approximation, randomly across the sequence with no regard for gene boundaries or other features of the sequence. Most recombinants produced this way will be nonfunctional because, for example, they have glued together parts of two encoded functional proteins into a nonfunctional chimera. The tiny fraction of such recombinants that are fully functional and competitive in the natural environment will be those that have recombination joints where they do no harm, like at gene boundaries; the rest will be rapidly eliminated from the population by natural selection. We would also expect that two recombining genomes would not typically be lined up in register, and this would most often lead to nonfunctional recombinants. There are a few examples of genomes with short quasi-duplications that can be explained by a slightly out of register recombination event, supporting this supposition.

p0060 Nonhomologous recombination also has the potential to bring novel genes into a genome and to mediate large-scale reorganizations of genomes, as seen for example in *E. coli* phage N15 which has head and tail structural genes homologous to those of phage lambda and, in the other half of the genome, genes from a very different, apparently plasmid-like, source. Another example is the *Shigella* phage SfV, which has head genes in the phage HK97 sequence family and tail genes like those of phage Mu.

p0065 The entire process described above can be viewed as a classical Darwinian scenario in which tremendous diversity is generated in the population by a combination of point mutations and homologous and nonhomologous recombination, and the stringent sieve of natural selection acts to eliminate all but the fittest recombinants. The survivors have largely, though not entirely, lost the appearance of having been generated by what is essentially a random mutational process.

Evolution of Other Phage Types

s0040

p0070 Although most of the genome comparisons that have led to our current understanding of phage evolution have been carried out with the tailed phages, there has been some work done with other groups. The most extensive has been done with members of the families *Microviridae* and the *Inoviridae*, phages with small, circular single-stranded DNA (ssDNA) chromosomes, typified by the well-studied *E. coli* phages ϕX174 and M13. For both of these groups there is evidence for horizontal exchange of sequences, as in the tailed phages, but it is quantitatively much less prominent than what is seen in the tailed phages. There are several differences between the tailed phages and these two groups of smaller phages, some or all of which may contribute to the differences in evolutionary outcomes seen between them. These include 10- to 100-fold differences in the sizes of the genomes and the fact that the members of the *Microviridae* and *Inoviridae* do not encode recombination functions, among others.

p0075 Another study looked at evidence for evolutionary change among members of the family *Cystoviridae*, a group of phages with a double-stranded RNA (dsRNA) genome, divided into three physical segments. Comparisons among environmentally derived genome sequences showed no evidence for recombinational exchange within the segments but clear indications of reassortment of the segments. Thus it appears that evolution in this group of viruses achieves the same general outcome as evolution of the tailed phages, namely horizontal exchange of genetic modules, but by a very different mechanism. Analogous horizontal exchange by reassortment of the segments of a segmented genome has been extensively characterized

in the influenza viruses, a group of animal viruses with a single-stranded RNA (ssRNA) genome of eight segments.

Deep Evolutionary Connections

p0080 Although there are some genes that are found in all tailed phages – for example, genes encoding the major capsid protein – these are very diverse in sequence. This diversity is sufficiently great that it is not possible, on the basis of their sequences, to group them into a single sequence family, at either the DNA or the protein sequence levels. Thus the subsets of capsid protein sequences that share demonstrable sequence similarity may represent independent nonhomologous lineages of capsid proteins, or alternatively they may all be part of the same lineage but have diverged in sequence to the point that in many cases no surviving sequence similarity can be demonstrated. However, recent structural work on the polypeptide folds of capsid proteins appears to be extending our analytical reach farther back in evolutionary time and giving a resolution to this question. High-resolution X-ray structures of the capsid proteins of phages HK97 and T4 demonstrate that they have a common fold which might indicate common ancestry despite the absence of any surviving sequence similarity. Cryoelectron microscopic structures of phages representing two more capsid protein sequence groups, P22 and ϕ29, make a less rigorous but still convincing case that these capsid proteins also share in a common polypeptide fold and so are most likely part of the same lineage. Remarkably, a similar experiment on (the animal virus) herpes simplex virus argues that the shell-forming domain of its major capsid protein also shares the tailed phage capsid protein fold and so may be part of the same lineage.

p0085 A similar and more completely documented case for a virus lineage extending across the cellular domains of the hosts they infect can be made for a group that includes bacteriophage PRD1 (a member of the family *Tectiviridae*), archaeal virus STIV, and eukaryotic viruses adenovirus, PBCV1 (*Phycodnaviridae*), and mimivirus. This is again based on shared capsid protein folds, in this case an unusual double 'jellyroll' fold. Finally, there is compelling structural similarity between the cystoviruses, the group of phages with segmented dsRNA chromosomes, and a characteristic double protein shell discussed above, and the reoviruses of plant and animal hosts.

p0090 The simplest interpretation of these observations, though not the only possible interpretation, is that there were already viruses resembling these different types of contemporary viruses prior to the divergence of cellular life into the three current domains. The different types of viruses in this view would have co-evolved with their hosts as the hosts divided into domains and descended to the present time, with in the end nothing to indicate their common origins except the shared structure of their capsid proteins. Whatever the truth of this matter, an important caveat to such an interpretation is that the evidence for common viral ancestries across host domains is at present based primarily on conserved properties of capsid proteins, and so any conclusions about common lineages, strictly speaking, only apply to capsid protein lineages. However, despite such reservations, the data make an intriguing though still tentative case that viruses resembling at least three groups of contemporary viruses had already evolved into forms something like those of contemporary viruses before the three domains of cellular life had begun to separate, $\sim$3.5 billion years ago.

See also: Origin of viruses (00709); Evolution of viruses (00706); Quasispecies (00481).

Further Reading

Au2

Brüssow H and Hendrix RW (2002) Phage genomics: Small is beautiful. *Cell* 108: 13–16.

Casjens SR (2005) Comparative genomics and evolution of the tailed-bacteriophages. *Current Opinion in Microbiology* 8: 451–458.

Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* 399: 541–548.

Hatfull GF, Pedulla ML, Jacobs-Sera D, *et al.* (2006) Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *P LoS Genetics* 2: e92. http://genetics.plosjournals.org/perlserv/?request=get-document&doi=10.1371%2Fjournal.pgen.0020092 (accessed Jun. 2007).

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, and Hendrix RW (2000) Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. *Jounal of Molecular Biology* 299: 27–52.

Nolan JM, Petrov V, Bertrand C, Krisch HM, and Karam JD (2006) Genetic diversity among five T4-like bacteriophages. *Virology Journal* 3: 30–44.

Suttle CA (2005) Viruses in the sea. *Nature* 437: 356–361.