Cell, Vol. 108, 13–16, January 11, 2002, Copyright ©2002 by Cell Press

Phage Genomics: Small Is Beautiful

Minireview

Harald Brüssow[1,3] and Roger W. Hendrix[2]
[1] Nestlé Research Center
Nestec Ltd.
Vers-chez-les-Blanc
CH-1000 Lausanne 26
Switzerland
[2] Pittsburgh Bacteriophage Institute and
Department of Biological Sciences
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

All the world's a phage.
—W. Shakespeare

There are similarities between the diseases of animals and man and the diseases of beer and wine.
—L. Pasteur

The Age of Genomics dawned only gradually for bacteriophages. It was 1977 when the genome of phage ϕX174 was published and 1983 when the "large" genome of phage λ hit the streets. More recently, the pace has quickened, so that we now have over 100 complete phage genomes and can expect thousands in a very few years. These sequences have been marvelously informative for the biology of the individual phages, but with the advent of high volume sequencing technology, the real excitement for phage biology is that it is now possible to analyze the sequences together and thereby address—for the first time at whole genome resolution—a set of fundamental biological questions related to populations: What is the structure of the global phage population? What are its dynamics? How do phages evolve? This is Comparative Genomics with a capital "C".

Dramatis Personae

Bacteriophages are not a homogenous group. They are currently classified on the basis of their genome (ss versus ds, RNA versus DNA) and their morphology into ten phage families (The Encyclopaedia of Virology, Academic Press, 1999, provides many useful entries for a first orientation on phages). Their diversity is also reflected by the diversity of genome sizes, which ranges from barely 4 kb to up to 600 kb (a mycobacteriophage). 96% of all bacterial viruses are tailed phages (Caudovirales, the focus of our writing), which come as Myo-, Sipho- and Podoviridae on the basis of tail morphology. Myoviridae have a long contractile tail (Figure 1); Sipho- and Podoviridae have long and short noncontractile tails, respectively. While this classification is quite popular, its evolutionary meaning is far from being clear. Podovirus P22 and Siphovirus λ share such a related structural gene map that their taxonomic distinction becomes questionable. The minimal genome of tailed phages encodes DNA packaging, head, tail and tail fibers, DNA replication, transcription regulation, and lysis genes. These genetic functions can be squeezed into a 20 kb DNA genome as demonstrated by c2 Siphovirus and ϕ29 Podovirus. As the genome size increases, the virion morphology gets more complicated, and the phage interferes more with cellular activities. However, we are far from understanding this in detail. Despite decades of research, only about 130 from the estimated 230 genes of Myovirus T4 have assigned functions. Tailed phages have been described in many phylogenetic divisions of Eubacteria. Curiously, viruses from one branch of Archaea (Euryarchaeota) resemble tailed phages in morphology and genome organization (Pfister et al., 1998). There is also evidence for shared protein folds between phages and eukaryotic viruses (Benson et al., 1999), pointing to very distant ancestral relationships (tailed phages; Herpes- and PRD1-like phages; Adeno- and ϕ6-like phages; and Reo-viruses).

As the list of phage genomes has grown, we have also come to appreciate the almost inconceivably large size of the global phage population, and specifically the population of dsDNA tailed phages. There are an estimated $>10^{30}$ tailed phages in the biosphere (Figure 2), and since phage particles typically outnumber prokaryotic cells by about 10-fold in environmental samples, the tailed phages probably constitute an absolute majority of "organisms" on our home planet in sheer numbers. This is evidently a very ancient population as well—an informed guess would be that the ancestors of phages originated well before the three contemporary domains of life separated—though it must be admitted that accurate numbers are not available. In a somewhat different arena, the discovery that phages are so numerous in the environment is gradually leading to a realization that they have a major role in important environmental processes such as carbon and energy cycling in the oceans.

The Impact of Comparative Genomics

Analysis of phage genomic data is starting to give a clearer view, on the one hand, of the great genetic diversity of this population and, on the other, of the remarkable underlying similarities. Much of the current discussion aims to reconcile these two superficially different views of the phage world, and to use that as a basis for inferring ancient and contemporary mechanisms of phage evolution.

Local Neighborhoods

Comparisons among the "lambdoid" phages of enteric hosts show that the genomes, which share overall gene organization, are mosaic with respect to each other. Points of recombination have been identified at gene boundaries. A recent report (Clark et al., 2001) shows the presence of some short regions of sequence homology between gene modules in lambdoid coliphages. These linker sequences could promote genetic reassortment (modular exchanges) through homologous or site-specific recombination. In an alternative model, nonhomologous recombination occurs indiscriminately and pervasively across the genome, followed by stringent selection for

[3] Correspondence: harald.bruessow@rdls.nestle.com

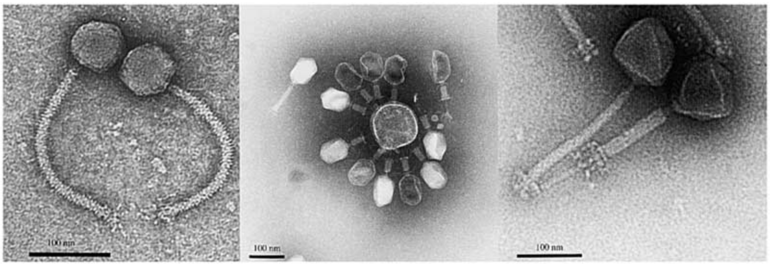

Figure 1. Electron Microscopy of Negative Stained Bacteriophages

Current sequencing efforts target phages with large genome size in Gram-positive bacteria (left: *Mycobacterium* phage CJW1; right: *Lactobacillus* phage LP65) to explore their relationship with Myoviridae from Gram-negative bacteria (center: a field isolate of a T4-like phage from a diarrhea patient attacking an *E. coli* cell remnant). Scale bars = 100 nm. (EM: M.L. Dillmann, S. Chennoufi, and M. Pedulla).

functional phages (Juhala et al., 2000). The second part of that process should eliminate most products of nonhomologous recombination within coding regions, leaving the gene-boundary recombinants and thereby giving the overall process an undeserved appearance of order and purpose.

Homologous recombination can occur every time a phage infects a cell carrying a prophage with appropriate homologies. These events serve not to create new mosaic boundaries but to rapidly reassort existing gene modules within the population.

In addition to the lambdoid group of phages from Gram-negative hosts, there are two groups with comparable numbers of complete genome sequences: the "dairy phages" (sensu lato, phages of lactic acid bacteria, and evolutionary relatives; Brüssow, 2001) and the mycobacteriophages (Ford et al., 1998), infecting distinct clusters of Gram-positive bacteria. These groups largely resemble the lambdoid group in their mosaic relationships, arguing for comparable mechanisms of gene exchange. There is however a quantitative difference among the groups, with the lambdoids displaying the most flamboyant mosaicism and the dairy phages being the most homogeneous.

Acquisition of New Phage Genes

The discussion above addresses reassortment of existing phage genes, but how do novel genes enter the phage genome in the first place? In a group of dairy phages that each carry novel genes in the context of a mostly shared genome organization, the novel genes are all located near the prophage attachment site, suggesting that they may have been acquired by imprecise prophage excision. In other cases, individual genes are inserted (by mechanism unknown) into interior parts of the genome (Juhala et al., 2000). Many genes that can be identified as recent acquisitions are transcriptionally autonomous and expressed from the repressed prophage; they may provide a selective benefit to the host and consequently a disincentive to prophage deletion. Others may enhance lytic growth of the phage. One speculation is that the entire genome could have arisen one gene at a time over phage evolution, with constant influx of novel genes coupled with selective retention of those that increase the genome's aggregate fitness.

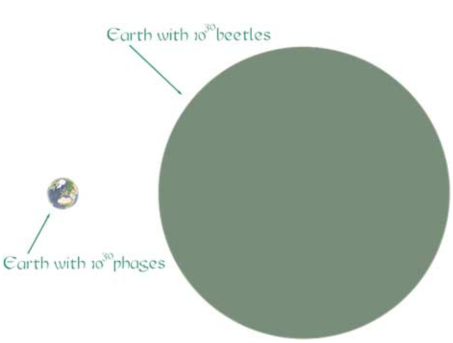

Figure 2. "The Creator Must Have an Inordinate Fondness for Beetles" (J.B.S. Haldane)

Tailed phages are vastly more abundant than we had imagined. The figure illustrates the consequences if 10^{30} phages were to be transmogrified into Haldane's beetles.

The Global Scene

How diverse is the phage population, and how are phages related to each other? Considering just the tailed phages, the well-characterized examples of these phages cluster into groups. A given bacterial species may be host to several of these groups; e.g., λ, T4, T7, Mu, P2, and N4-like phages and others all grow on *E. coli*. It is not yet clear how far across the phylogenetic field of hosts a given phage group can range. Sequence-related P2-like phages can infect at least a moderately large collection of bacterial genera (Nakayama, et al., 1999). T4-like phages infecting as distinct hosts as proteobacteria and cyanobacteria still shared some protein sequence identity (Hambly et al., 2001). It is currently being explored whether phages with a similar genome organization as T4 exist in Gram-positive bacteria (Figure 1). In fact, phages with a nearly identical gene order for DNA packaging and head and tail morphogenesis genes to that of phage λ were identified in Gram-positive bacteria and even Archaea (Brüssow and Desiere, 2001), while sequence similarity between them had nearly been erased.

More generally, if we compare the genomes of two tailed phages from distant points on the diversity landscape—either two phages with distant hosts or phages from two different groups infecting the same host—we typically find that there is little or no sequence similarity apparent between them, despite a sometimes closely related gene map. For example, structural phage genes within a given phage lineage are found in a strikingly conserved order; as another illustration, sequence-unrelated phage proteins showed similar secondary structure predictions and proteolytic processing sites. The synteny argument became a popular tool in phage genomics. This suggests either very ancient divergence or rather spectacular evolutionary convergence. The divergence option (that is, that all the tailed phages share common ancestry) is beginning to gain convincing support, based on more sensitive search algorithms that show previously invisible protein sequence relationships as well as the possibility of linking two very dis-

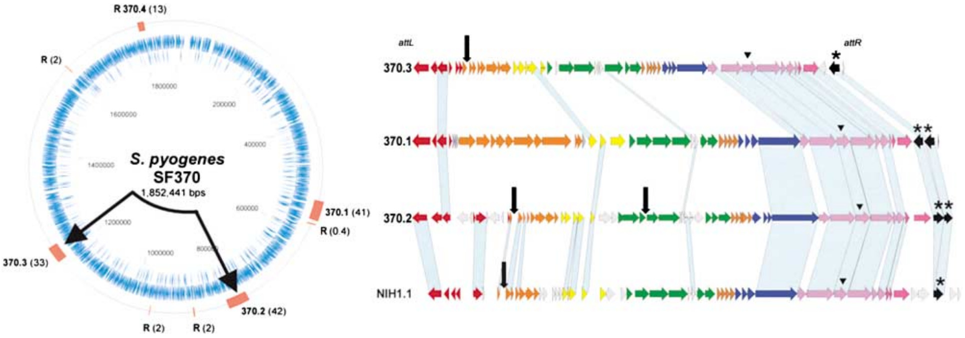

Figure 3. Prophage Genomics

Left: Location and relative size (in kb) of prophages and prophage remnants R (in red) on the *Streptococcus pyogenes* strain SF370 genome map. Comparison with the unfinished *S. pyogenes* strain Manfredo sequence (Sanger Center) identified a possible site of genome rearrangement due to putative crossover events between two prophages (indicated by linked arrows). Right: The genome maps of three SF370 prophages and a prophage from *S. pyogenes* strain NIH1 are aligned with the attachment sites at the left and right ends. Prophage NIH1.1 has been identified as a genetic marker for recently emerged clinical isolates of *S. pyogenes* in Japan. The phage modules are color-coded (red: lysogeny, orange: DNA replication, yellow: transcriptional regulation(?), green: DNA packaging and head, brown: head-to-tail, blue: tail, mauve: tail fiber, dark pink: lysis, black: superantigen/ mitogenic factor genes). Likely prophage-inactivating mutations are indicated by large vertical arrows (replisome organizer, portal protein); an asterisk marks phage genes that potentially contribute to the virulence of the lysogenic host; and the phage hyaluronidase is labeled by a triangle. Regions of DNA sequence similarity between the prophages are connected by blue shading (drawing based on entries from the database by C. Canchaya).

tantly related sequences by chains of similar sequences from other phages. Significantly, when analyzing the structural gene clusters from dairy phages, gradients of relatedness were detected that correlated with the evolutionary relationships between the bacterial hosts. Graded relatedness is a hallmark of Darwinian evolution. On the other hand, isolated sequence-related genes are often found between quite distinct phages. This argues that horizontal exchange of sequences also occurs, though with reduced frequency, across the entire span of known tailed phages (Hendrix et al., 1999). Just as with individual genes, it appears that functionally related groups of genes can swap horizontally through the population. For example, there are rather close relationships between some early genes of P2-like phages and those of certain dairy phages currently placed into two distinct phage families (Myo- versus Siphoviridae). For scientists bent on bringing order out of chaos, phages' pervasive horizontal swapping has the interesting implication that it is not possible even in principle to represent their history with a simple branching phylogeny.

Prophage Genomics

The peculiar lifestyle of temperate phages, with alternating phases of lytic growth and prophage existence, submits phage DNA to paradoxical selection pressures; it also makes phages fascinating objects for biologists interested in the evolution of parasitic DNA. The selfish and mutualistic aspects of prophages for bacteria emerges from prophage genomics (Desiere et al., 2001).

It was known for a while that a number of famous bacterial toxins like the food poisoning botulinus toxin are actually phage encoded. A prominent late arrival to that list was cholera toxin. *Vibrio cholerae* is actually a fascinating case of how multiple phages contribute to bacterial pathogenicity. It is now clear that phages are a major source of horizontally transferred DNA in bacteria and that prophage DNA accounts in many bacteria for important interstrain genetic variability. In a number of bacterial genomes, 3% to 10% of the DNA is contributed by prophages. Sequencing the genome of an emerging food pathogen, *E. coli* O157, for example, identified 18 prophages or prophage remnants. Prophage DNA accounts for half of the 1.3 Mb of DNA found in O157 but absent from reference strain *E. coli* K12 (Ohnishi et al., 2001). These prophages belong to different phage lineages and encode many likely virulence factors. One could argue that the acquisition of prophage DNA has played a decisive role in the emergence of this food pathogen. This is not an isolated case. *Salmonella typhimurium*, another important food pathogen, possesses a variable assortment of prophages which apparently represent a transferable repertoire of pathogenic determinants (Figueroa-Bossi et al., 2001). The importance of prophage genes for the in vivo virulence of *Salmonella* was demonstrated by inactivation studies of selected prophage genes.

Streptococcus pyogenes is a Gram-positive human pathogen with a highly variable clinical symptomatology, ranging from harmless to life-threatening. *S. pyogenes* strain SF370 contains 8 prophage elements (Ferretti et al., 2001): five prophages have suffered massive deletions, two carry point mutations in key phage genes, and only one prophage is inducible (Figure 3). Similar trends for prophage inactivation exist in other bacteria. A high deletion rate may be a bacterial defense against a high rate of DNA influx by dangerous foreign DNA elements. This hypothetical "cleansing" activity could account for the compact, pseudogene-free nature of most bacterial genomes (Lawrence et al., 2001). Homologous recombination between prophages sharing some DNA homology and residing in the same host (Figure 3) can lead to host genome rearrangements and new associations with lysogenic conversion genes in phages.

The prophage virulence genes differ frequently in GC content from those of the phage and bacterial genome. This leads to provocative questions. Are phages major drivers of the evolution of bacterial pathogens by contributing foreign genes that increase the competitiveness of the lysogen in its ecological niche? Can the variable clinical potential of some protean bacterial pathogens be interpreted by possession of a specific prophage set? Many bacterial pathogens show a very dynamic epidemiology; is the replacement of older by newer strains linked to the acquisition of new prophages, as suggested for newly emerged *S. pyogenes* strains in Japan (Inagaki et al., 2000) (Figure 3)? Upon changed growth conditions, are prophage genes prominently upregulated genes, as proposed in a genome-wide microarray for *S. pyogenes*?

It is increasingly becoming clear that gut commensals and pathogenic bacteria have much in common. Despite their large numbers, gut bacteria do not lead a cozy life. They are under selective pressure of a T cell independent mucosal IgA response to which they respond by constantly changing their surface polysacharides. Interestingly, sequenced *Lactobacillus* commensals contain multiple prophage genomes that showed lysogenic conversion genes related to those of prophages from *S. pyogenes*. Also, a classical dairy strain like *Lactococcus lactis* IL1403 contains six prophage genomes (Chopin et al., 2001). In view of this heavy prophage load, one might suspect that prophages contribute to the evolutionary success of lactic acid bacteria living in strikingly distinct environments.

Practical Applications

The idea to use phages for therapeutical purpose dates back to the discovery of phages. T4-like phages, for example, can be isolated from stool samples of diarrhea patients. Some isolates infect and lyse a large range of pathogenic *E. coli* strains in vitro (Figure 2, center) and survive an unprotected gastrointestinal passage in mice. Uncontrolled clinical trials suggested that staphylococcal phages could have a role in the treatment of severe skin infections. Phage therapy with infectious phages is limited by the host range of the phage isolate. Comparative phage genomics leads approaches to achieve host range extension by the manipulation of tail fiber genes (Tétart et al., 1998). Recent animal experiments show great promise for the prevention and treatment of *S. pyogenes* and *S. pneumoniae* infections with an isolated phage enzyme from tailed phages, the phage lysin that attacks the bacterial cell wall (Loeffler et al., 2001). Due to their modular structure and high species specificity, phage lysins are also potential tools for rapid detection of bacterial pathogens in the clinic and the environment. Small-genome phages do not encode lysins, but protein inhibitors that block steps in the synthesis of cell wall precursor molecules, raising the possibility of DNA-encoded oligopeptide antibiotics (Bernhardt et al., 2001).

If phages are the friend of medical bacteriologists, they are the foe of dairy microbiologists. When they get access to the dairy factory, they multiply on the lactic acid bacteria that ferment milk into yogurt or cheese, leading to production delays or even product loss. Phage genomes can be exploited to construct genetic traps for phages. For example, phage-resistant bacterial starters have been designed that contain the cloned phage origin of replication on a plasmid. When this cell is infected, the phage drives the replication of the plasmid, and no longer drives its own DNA replication (Brüssow, 2001).

Finally, phages have been a valuable source of enzymes for a variety of purposes, as any recombinant DNA jockey can attest. As more of the genetic diversity of phages is discovered, the phage gene pool increasingly comes to resemble the much admired gene pool of the tropical rain forest as a potential resource for serving human ends.

Outlook

There is now interest not only in sequencing new phage genomes but in doing so from as diverse an array of phages as possible. Phage biologists are also beginning to explore methods to sample phage sequences from environmental sources without introducing the severe bias of asking them to grow on culturable bacteria. Combining such an approach with techniques like microarray analysis promises to give a better picture of the diversity of the phage population and also to provide tools to ask how that diversity changes over space and time. The ease and low cost of phage sequencing, combined with the extensive knowledge on model phages, could give phage genomics a lead role in population genetics, the evolution of simple DNA genomes, and the modeling of a realistic DNA sequence space. In fact, small is not only beautiful in aesthetic terms, but also immensely handy in practical terms, allowing even small laboratories a scientific living in the era of genomics research.

Selected Reading

Benson, S.D., Bamford, J.K., Bamford, D.H., and Burnett, R.M. (1999). Cell *98*, 825–833.

Bernhardt, T.G., Wang, I.-N., Struck, D.K., and Young, R. (2001). Science *292*, 2326–2329.

Brüssow, H. (2001). Annu. Rev. Microbiol. *55*, 283–303.

Brüssow, H., and Desiere, F. (2001). Mol. Microbiol. *39*, 213–222.

Chopin, A., Bolotin, A., Sorokin, A., Ehrlich, S.D., and Chopin, M.-C. (2001). Nucleic Acids Res. *29*, 644–651.

Clark, A.J., Inwood, W., Cloutier, T., and Dhillon, T.S. (2001). J. Mol. Biol. *311*, 657–679.

Desiere, F., McShan, W.M., van Sinderen, D., Ferretti, J.J., and Brüssow, H. (2001). Virology *288*, 325–341.

Ferretti, J.J., McShan, W.M., Ajdic, D., Savic, D.J., Savic, G., Lyon, K., Primeaux, C., Sezate, S., Suvorov, A.N., Kenton, S., et al. (2001). Proc. Natl. Acad. Sci. USA *98*, 4658–4663.

Figueroa-Bossi, N., Uzzau, S., Maloriol, D., and Bossi, L. (2001). Mol. Microbiol. *39*, 260–271.

Ford, M.E., Sarkis, G.J., Belanger, A.E., Hendrix, R.W., and Hatful, G.F. (1998). J. Mol. Biol. *279*, 143–164.

Hambly, E., Tétart, F., Desplats, C., Wilson, W.H., Krisch, H.M., and Mann, N.H. (2001). Proc. Natl. Acad. Sci. USA *98*, 11411–11416.

Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999). Proc. Natl. Acad. Sci. USA *96*, 2192–2197.

Inagaki, Y., Myouga, F., Kawabata, H., Yamai, S., and Watanabe, H. (2000). J. Infect. Dis. *181*, 975–983.

Juhala, R.J., Ford, M.E., Duda, R.L., Zoulton, A., Hatfull, G.F., and Hendrix, R.W. (2000). J. Mol. Biol. *299*, 27–51.

Lawrence, J.G., Hendrix, R.W., and Casjens, S. (2001). Trends Microbiol. *9*, 535–540.

Loeffler, J.M., Nelson, D., and Fischetti, V.A. (2001). Science *294*, 2170–2172.

Nakayama, K., Kanaya, S., Ohnishi, M., Terawaki, Y., and Hazashi, T. (1999). Mol. Microbiol. *31*, 399–419.

Ohnishi, M., Kurokawa, K., and Hayashi, T. (2001). Trends Microbiol. *9*, 481–485.

Pfister, P., Wasserfallen, A., Stettler, R., and Leisinger, T. (1998). Mol. Microbiol. *30*, 233–244.

Tétart, F., Desplats, C., and Krisch, H.M. (1998). J. Mol. Biol. *282*, 543–556.