

Bacteriophages:

Phages exist on our planet not only in huge numbers but infinite variety as **Graham F. Hatfull** describes.

nature's most successful experiment

▲ Coloured electron micrographs of a variety of bacteriophages. From left to right: SEM of T phages attacking *Escherichia coli* (Eye of Science / SPL); TEM of lambda phages (CNRI / SPL); TEM of a cluster of P1 phages (Biozentrum / SPL); TEM of T phage particles attacking *E. coli* (Eye of Science / SPL).

You may well be under the impression that the largest number of undiscovered species – and the greatest pool of unknown genes – lie within the considerable biodiversity of the tropical rain forests. Not so. A compelling argument can be made that the biggest reservoir of unidentified genetic information is all around us, in the global population of bacteriophages.

How many phages are there?

There are two main components to this conclusion: the amazing abundance of phages in the environment and the emerging picture of their genetic diversity. Over the past few years it has been calculated that the total number of phage particles in the

biosphere is a stunning 10^{31} . This is a remarkable number, because it suggests that phages are a numerical majority of all biological entities – i.e. there are more phage particles than all other biological forms added together. If abundance can be equated with success, then phages represent the result of nature's most successful experiment! It is helpful to understand how this abundance is calculated. Samples from the environment can be stained with dyes that cause viruses and bacteria to fluoresce, and the number of particles counted using fluorescent microscopy. There are two main observations: that viral particles are typically present at 10^6 – 10^7 per ml seawater, and that there are 5- to 10-fold more viruses than bacteria. The viral abundance is broadly similar whether seawater or terrestrial

samples are analysed, and does not change significantly when comparing coastal, oceanic, surface or deep-water samples. A simple extrapolation to the total volume of seawater and inclusion of terrestrial counts leads to a total number of phage particles of 10^{31} . An independent estimate of the total number of bacteria in the biosphere arrived at 10^{30} , providing some comfort in the validity of these estimations.

A dynamic population

The abundance of phage particles is impressive, but would be rather less interesting if these were just a large number of a small number of types. However, there are clues suggesting that this is not the case. First, it has been estimated that there are approximately 10^{24} viral infections of bacteria per second on a global scale, suggesting that the entire phage population turns over every few days – far from being static, the population is highly dynamic, with each cycle of infections having the potential to generate altered or mutant particles. Second, it seems likely that the evolutionary origin of phages is not too distant from that of their bacterial hosts, so this dynamic relationship has been going on for at least 2 billion years!

What type of viral population has arisen, what do they look like, and how different are their genomes? Phages were among the first 'invisible' entities to be observed by electron microscopy, which showed them to be well-defined structures containing a protein head (capsid) surrounding the genetic material, attached to a tail. While these tailed phages containing double-stranded DNA (dsDNA) are perhaps the most common forms, a large variety of different morphotypes have been described, with some spectacular shapes described recently for viruses of archaeal hosts. The genetic material in phages can be either DNA or RNA, either in single- or double-stranded forms.

Genomics

Currently, just over 500 completely sequenced phage genomes are listed in the Genomes section at the National Center for Biotechnology Information (NCBI). The size of this collection has grown considerably over the past 5 years, but is still dwarfed by the number of sequenced bacterial genomes (currently ~750), even though these are 100 times larger. While these sequenced phage genomes represent only a

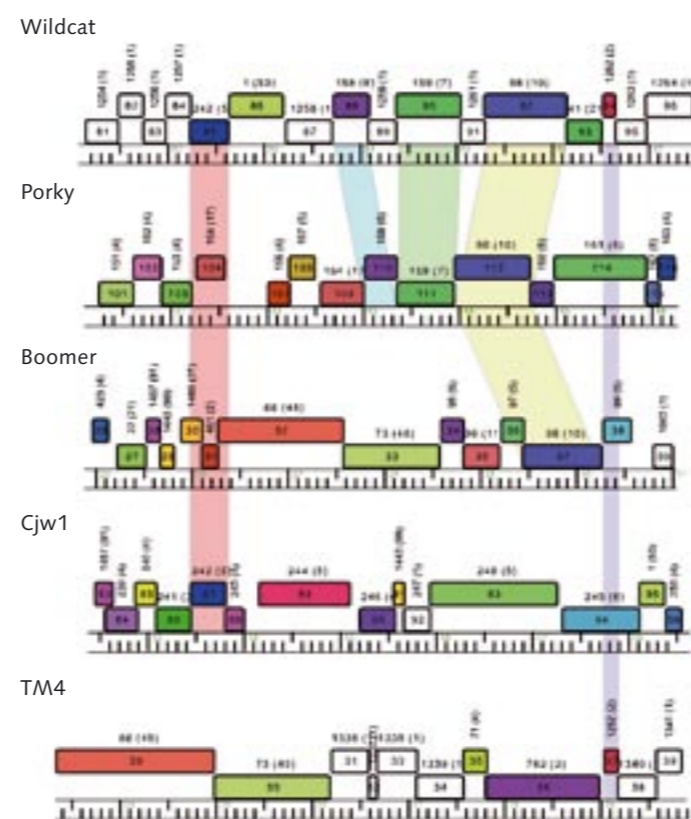
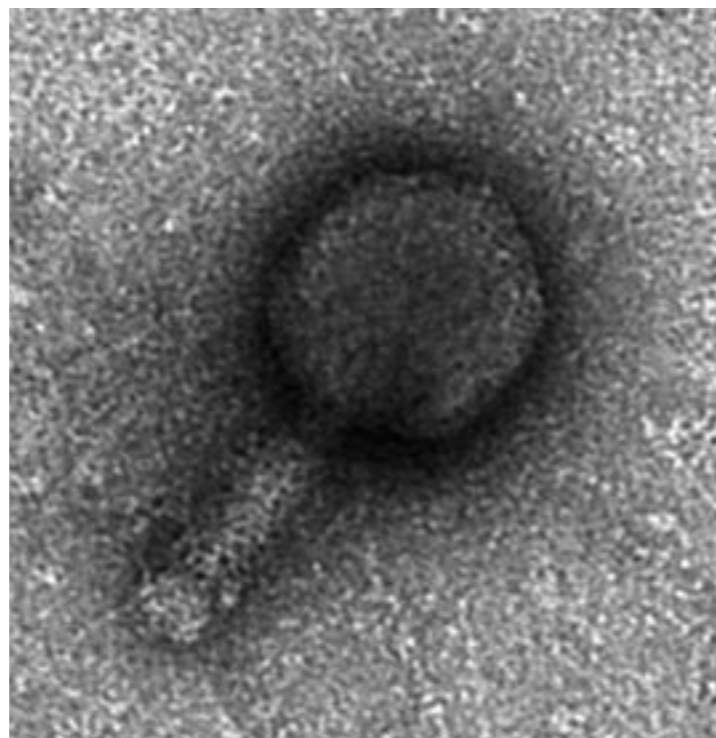
tiny slice of the total phage population, their comparative genomic analysis has provided some fascinating insights into the diversity of the population and the evolutionary mechanisms generating this diversity.

There are two general ways of thinking about the genetic relationships within the phage population. One is to consider the similarities between groups of phages that infect different bacterial hosts; the other is to compare genomes of phages that infect a common bacterial host. The 500 sequenced phage genomes correspond to about 70 different bacterial hosts (a small subset of the ~470 bacterial species that have been sequenced), and because phages probably exist for the vast majority of bacterial species, this is likely to correspond to a small and highly unrepresentative sample. Nonetheless, comparisons of these genomes show that they are highly diverse, and phages of one bacterial host will often have little or no recognizable nucleic acid sequence similarity with phages of other hosts. Although amino acid sequence comparison of predicted gene products can detect more distant relationships, relatively few genes emerge as having common evolutionary origins. However, it appears as though the more closely related the bacterial hosts, the greater the chance that their phages may share common genes. For example, phage PA6 of *Propionibacterium acnes* contains a number of genes that are shared by phages of the mycobacteria. Overall, it seems reasonable to suppose that phages form a continuum of relationships that will only be fully understood with a substantial increase in the number of sequenced genomes.

Substantial groups of sequenced genomes have now been accumulated that share common hosts, with more than 20 genomes for each of *Burkholderia*, enterobacteria, lactococci, mycobacteria, pseudomonads, staphylococci and streptococci. An emerging theme is that these groups are also quite divergent, although there are clusters of phages that are more closely related to each other than to others of that host. This probably reflects a similar, but perhaps more focussed view to that of the phage population as a whole, with the clusters probably representing unequal sampling

▲ Virion morphology of mycobacteriophage Bx21. Electron microscopy shows that Bx21 contains an icosahedral capsid containing a dsDNA genome, and a contractile tail. *Graham Hatfull*

► Mosaicism in mycobacteriophage genomes. Short segments of five mycobacteriophage genomes are shown with the genes represented as coloured boxes. Each gene has been assigned to a phamily (Pham) of related sequences and the Pham number presented above the box, along with the number of genes in that Pham in parentheses; genes of the same Pham are coloured accordingly. Genes without any relatives elsewhere in the collection of 50 completely sequenced mycobacteriophage genomes are shown as white boxes. Phages Porky, Boomer, Cjw1 and TM4 all contain at least one gene that is related to one in Wildcat in this region, and coloured stripes indicate these relationships. Careful inspection will reveal additional genes shared by some of these phages. *Graham Hatfull*



of the population rather than specific and stable population structures. Viral metagenomic studies – in which DNA fragments from the total collection of viruses in a sample are sequenced – support this view of a highly diverse phage population.

Mosaicism

One of the clearest and most amazing aspects of phage genomes is their mosaic architecture. Comparative analysis reveals segments or modules that are shared by two or more phages, but which are flanked by different modules. These modules can be groups of genes – especially those whose functions need to work together, such as the virion structural genes – or single genes. In the mycobacteriophages, this single-gene mosaicism within the non-structural genes is particularly prevalent with large groups of contiguous genes having distinct and different evolutionary histories. The diversity of phage genomes suggests that the number of different modules is very large and that only a very small subset have as yet been identified. Each individual phage genome can thus be viewed as a unique assembly of modules, and the combinatorial possibilities are enormous.

Evolutionary mechanisms

The evolutionary mechanisms that give rise to these mosaic genome structures is not clear, although both homologous and illegitimate recombination appear to be involved. One particular mystery

is how the junctions between modules are generated, and although these could arise from homologous recombination between short conserved boundary sequences, these are not found in most phage genomes. An alternative explanation is that illegitimate recombination occurs at randomly chosen positions and with randomly chosen partners, which could be other phage genomes, plasmids or the bacterial chromosome. Most of these events will generate genomic trash, but with selection for appropriately-sized genomes that can be packaged into capsids, and for maintenance of a functional set of genes, viable progeny could arise. The overall process is not expected to be efficient or frequent, but with such a dynamic population evolving over such a long period of time, this would not seem to be a problem. Other events such as transposition and site-specific recombination will also generate further rearrangements, and homologous recombination mediates exchanges at common gene sequences.

Gene functions

What do all these phage genes do? While the genes required for virion structure and assembly can often be recognized – especially since they are often arranged in common gene orders – along with some recombination and DNA replication activities, the functions of most other phage genes remains largely unknown. When

they are compared against databases such as GenBank, matches revealing putative functions are relatively rare, and typically fewer than 30% of genes in newly sequenced phage genomes can be assigned putative functions. Genomics and bioinformatics are thus unlikely to provide a comprehensive understanding of phage gene functions, and experimental approaches are clearly needed.

Phage genomics are clearly in their infancy. The current state of the field provides a tantalizingly tasty appetizer, but with such a vast landscape of unexplored phages yet to be studied, one cannot but feel that the main course is yet to come. In the next few years, the advent of ultra-high throughput DNA sequencing technologies will generate a multitude of new phage genome sequences, and functional and structural genomic approaches will help us to elucidate their biology. While we may never know for sure what the whole phage population looks like, perhaps we can at least learn enough to appreciate how much of it we really don't understand.

Graham F. Hatfull

Department of Biological Sciences and Pittsburgh Bacteriophage Institute, University of Pittsburgh, 4249 5th Avenue, Pittsburgh, PA 15260, USA (t +1 412 624 4350; f +1 412 624 4870; e gfh@pitt.edu)