Exploring the Mycobacteriophage Metaproteome: Phage Genomics as an Educational Platform

Graham F. Hatfull[1,2*], Marisa L. Pedulla[1,2¤a], Deborah Jacobs-Sera[1,2], Pauline M. Cichon[1,2], Amy Foley[3], Michael E. Ford[1,2¤b], Rebecca M. Gonda[1,2], Jennifer M. Houtz[1,2], Andrew J. Hryckowian[1,2], Vanessa A. Kelchner[1,2¤c], Swathi Namburi[1,2], Kostandin V. Pajcini[1,2¤d], Mark G. Popovich[4], Donald T. Schleicher[1,2], Brian Z. Simanek[1,2¤e], Alexis L. Smith[1,2], Gina M. Zdanowicz[1,2¤f], Vanaja Kumar[5], Craig L. Peebles[1,2], William R. Jacobs Jr.[6,7], Jeffrey G. Lawrence[1,2], Roger W. Hendrix[1,2]

1 Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, 2 Pittsburgh Bacteriophage Institute, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America, 3 Hampton High School, Allison Park, Pennsylvania, United States of America, 4 Lincoln High School, Elwood City, Pennsylvania, United States of America, 5 Tuberculosis Research Center, ICMR, Chetpet, Chennai, India, 6 Department of Microbiology and Immunology, Albert Einstein College of Medicine, New York, New York, United States of America, 7 Howard Hughes Medical Institute, Albert Einstein College of Medicine, New York, New York, United States of America

Bacteriophages are the most abundant forms of life in the biosphere and carry genomes characterized by high genetic diversity and mosaic architectures. The complete sequences of 30 mycobacteriophage genomes show them collectively to encode 101 tRNAs, three tmRNAs, and 3,357 proteins belonging to 1,536 "phamilies" of related sequences, and a statistical analysis predicts that these represent approximately 50% of the total number of phamilies in the mycobacteriophage population. These phamilies contain 2.19 proteins on average; more than half (774) of them contain just a single protein sequence. Only six phamilies have representatives in more than half of the 30 genomes, and only three—encoding tape-measure proteins, lysins, and minor tail proteins—are present in all 30 phages, although these phamilies are themselves highly modular, such that no single amino acid sequence element is present in all 30 mycobacteriophage genomes. Of the 1,536 phamilies, only 230 (15%) have amino acid sequence similarity to previously reported proteins, reflecting the enormous genetic diversity of the entire phage population. The abundance and diversity of phages, the simplicity of phage isolation, and the relatively small size of phage genomes support bacteriophage isolation and comparative genomic analysis as a highly suitable platform for discovery-based education.

Introduction

Approximately 10^{31} tailed bacteriophages are estimated to be on planet Earth, representing the majority of all biological entities in the biosphere [1,2]. Phages exert considerable influence on the microbial community [3] and cohabit with their bacterial hosts in highly dynamic relationships [3–5]. While phages show only somewhat limited morphological variation, the genomes of the approximately 300 completely sequenced double-strand DNA (dsDNA) phage genomes reveal both a high level of genetic diversity and a high proportion of genes that are dissimilar to any that have been previously sequenced [6–10]. Furthermore, phage genomic architecture is characterized by a high degree of mosaicism that likely arises from extensive horizontal genetic exchange occurring over perhaps as many as 3 billion years [9, 11–13].

dsDNA tailed phages are a substantial proportion of the total phage population, and the number of completely sequenced genomes has grown dramatically over the past 5 years to more than 150. These infect a wide variety of bacterial hosts, but the phages of enteric bacteria, bacteria relevant to the dairy industry, *Mycobacterium* spp., *Pseudomomas* spp., *Staphylococcus* spp., and marine bacteria are prevalent among the completely sequenced dsDNA tailed phage genomes [6,14–17]. These genomes vary in size from 25 to greater than 200 kilobase-pairs (kbp) and are harbored by viruses with a variety of virion morphologies [18]. The high degree of phage genetic diversity is indicated not only by the

sequenced genomes of individual phages but also by viral metagenomic approaches showing that the majority of genes sequenced from uncultured viral libraries have no significant similarity to known genes [19–23].

Synopsis

Bacteriophages are viruses that infect bacterial hosts and are estimated to be the most numerous biological entities in the biosphere. Insights into the genetic diversity of the bacteriophage population and the evolutionary mechanisms that give rise to it can be obtained using comparative genomic analyses. The genomic analysis of 30 complete mycobacteriophages—viruses that infect mycobacterial hosts—reveals them to be genetically diverse and to contain many previously unidentified genes. The high diversity and relatively small genome sizes of these phages provide an ideal platform for introducing high school and undergraduate students to the research laboratory, isolating and naming novel viruses, and determining their genomic sequences. The thrill of discovering new viruses and previously unidentified genes, coupled with ownership of individual phage projects, provides strong motivations for students to engage in and pursue scientific research.

Mycobacteriophages—viruses of the mycobacteria—have facilitated the development of mycobacterial genetic systems [24–27] and provided insights into viral diversity and the evolutionary mechanisms that generate them [6,28–31]. The comparative genomic analysis of 14 mycobacteriophages showed that they have relatively large genomes (average length, 70 kbp), contain large numbers of previously unidentified genes, and are highly diverse at both the nucleotide and amino acid sequence levels [6]. Moreover, phage genome architectures are pervasively mosaic; each genome apparently contains a unique combination of individual modules, of which the majority correspond to one or a small cluster of genes. Generation of these mosaic genomes reflects a high level of horizontal genetic exchange within the phage population, with illegitimate recombination events underlying the generation of new module boundaries. This model is supported by the identification of a small number of exchange events that have occurred relatively recently and for which there is no evidence of targeting to specific locations, such as gene extremities [6,7].

Since the majority of newly sequenced phage genes, including those from mycobacteriophages, are not closely related to known gene sequences, their functions remain largely unknown [15,19]. In many, but not all, dsDNA tailed phages, the genes encoding the virion structure and assembly functions are clustered into long operons with a well-defined gene order, and these genes can thus be more easily identified [32]. In phages adorned with a long flexible tail, the length of the tail is determined by the *tmp* gene, whose length corresponds directly to the tail length [6,33]. Since it is common for these phages to have tail lengths in excess of 100 nm, the *tmp* gene is frequently the longest open reading frame in the genome [6]. Outside of the structural operons, gene functions are well understood only in those phages that have been thoroughly dissected genetically and biochemically, and many phages contain large numbers of relatively small open reading frames of unknown function [6,34].

While the characteristic mosaic architecture of phage genomes can be explained by abundant horizontal genetic exchange events in their evolutionary history, little is known about which genomes participate in these events or the rates at which exchange occurs. Phages infecting the same host are expected to be more likely to exchange genes, but phages can readily switch or expand their host-range by a variety of

mechanisms [35,36]. Moreover, some phages have very broad host-ranges [37]. In addition, capsid packaging imposes nonselective constraints on genome length that are quite distinct from the functional advantages and disadvantages of phage genes. Little is known about the rates of nonselective acquisition, or the subsequent fates of such genes in the phage population, although it is likely that gene acquisition and loss contribute significantly to the evolution of bacterial genomes [38–40].

Bacteriophages exchange genes not only among themselves but also with their host chromosomes, and it is common for phages to carry genes that profoundly influence the physiology of their hosts [41–44]; these include genes that confer bacterial pathogenicity, such as the toxins responsible for pathogenesis of cholera, diphtheria, and shigellosis [45,46]. If bacteriophage mosaicism is largely generated by a process of illegitimate recombination, then the acquisition of genes from the host is not surprising [6]. However, for host-like genes identified by phage genomics, it is unclear what advantage such genes may confer upon lysogenic hosts or whether they remain in the phage population for long or short periods.

We have expanded our collection of complete mycobacteriophage genomes to a total of 30 and utilized the enlarged set of gene sequences to explore their genetic diversity. The organization of protein-encoding genes into "phamilies" of related sequences provides insights into which genes are most prevalent in these phages, whether there are signature sequences that are characteristic of mycobacteriophages, and how phamily size corresponds to phage and bacterial homologs outside of the mycobacteriophage group. The genetic diversity, abundance of novel genes, and technical feasibility of phage isolation and genomic characterization strongly support the utilization of phage discovery and comparative genomics as an effective educational platform for undergraduate and high school students.

Results/Discussion

Isolation and Characterization of Mycobacteriophages

We previously reported the genome sequences of mycobacteriophages L5 [28], D29 [30], TM4 [31], and Bxb1 [29], along with a comparative genomic analysis of these and ten additional mycobacteriophages: Barnyard, Bxz1, Bxz2, Che8, Che9c, Che9d, Cjw1, Corndog, Omega, and Rosebush [6]. We have isolated an additional 16 mycobacteriophages using the methods described previously [6] from a variety of sources; Che12 was isolated in Chennai, India; Bethlehem and U2 are from Bethlehem, Pennsylvania, United States; 244 is from Connecticut, United States; and Catera, Cooper, Halo, Llij, Orion, PBI1, PG1, Pipefish, P-Lot, PMC, Qyrzula, and Wildcat are from Pittsburgh, Pennsylvania, United States, and the surrounding areas. Each of these 16 new phages was plaque-purified, grown in quantity, banded through an equilibrium CsCl density gradient, and used to isolate virion DNA. DNA was hydrodynamically sheared, cloned into plasmid vectors, sequenced, and assembled as described previously [6]. Many of these phages were isolated, named, and characterized by undergraduate and high school students, taking advantage of phage discovery and genomics as a platform for integrating scientific research and education as described in detail below.

Table 1. Features of Completely Sequenced Mycobacteriophage Genomes

Phage	GC%	Size (bp)	Number of ORFs	Number of tRNAs[a]	Number of tmRNAs[b]
L5	62.3	52,297	90	3	0
D29	63.5	49,136	77	5	0
TM4	68.1	52,797	92	0	0
Bxb1	63.6	50,550	86	0	0
Bxz1	64.8	156,102	225	30	1
Bxz2	64.2	50,913	86	3	0
Che8	61.3	59,471	112	0	0
Che9c	65.4	57,050	84	0	0
Corndog	65.4	69,777	122	0	0
Omega	61.4	110,865	237	2	0
CJW1	63.1	75,931	141	2	0
Che9d	60.9	56,276	111	0	0
Rosebush	69.0	67,480	90	0	0
Barnyard	57.3	70,797	109	0	0
PG1	66.5	68,999	100	0	0
Bethlehem	63.2	52,250	87	0	0
Wildcat	56.9	78,296	148	22	1
Cooper	69.1	70,654	99	0	0
Halo	66.7	42,289	65	0	0
U2	63.7	51,277	81	0	0
LLIJ	61.5	56,851	100	0	0
PBI1	59.7	64,494	81	0	0
Che12	62.9	52,047	98	3	0
P-Lot	59.7	64,787	89	0	0
Qyrzula	69.0	67,188	81	0	0
Orion	66.5	68,427	100	0	0
PMC	61.4	56,692	104	0	0
Catera	64.7	153,766	218	29	1
244	62.9	74,483	142	2	0
Pipefish	67.3	69,059	102	0	0
Average	63.7	69,033	112		
Total		2,071,001	3,357	101	3

[a]Number of tRNAs predicted by tRNA-Scan-SE with a COVE score = 20.
[b]Number of tmRNAs predicted by Aragorn v1.1.
DOI: 10.1371/journal.pgen.0020092.t001

Mycobacteriophage Genometrics

Our view of the general features of mycobacteriophage genomes changes only modestly with a doubling of the number of genome sequences available (Table 1). The total sequence information is increased from 979,434 bp to 2,071,001 bp but the average genome length has not changed significantly. The largest of the newer genomes is Catera (153.7 kbp), slightly smaller than Bxz1 (156.1 kbp), the largest previously sequenced mycobacteriophage. One of the newly sequenced genomes, Halo, is only 42.3 kbp long, substantially smaller than any other mycobacteriophage genome (Table 1). Neither the base composition (average %GC content, 63.7%) nor the average ORF length (600 bp) has changed significantly (Table 1). These genomes also encode 101 tRNAs and three tmRNAs. The distribution of these small translation-system RNAs is quite uneven, with three phages (Bxz1, Wildcat, and Catera) contributing 80% of the total tRNAs and all three tmRNAs. Since the primary focus of this paper is to report on the diversity of mycobacteriophage gene phamilies and the use of phage genomics as a discovery-based educational platform, further details of the individual phages and their genomes will be reported elsewhere.

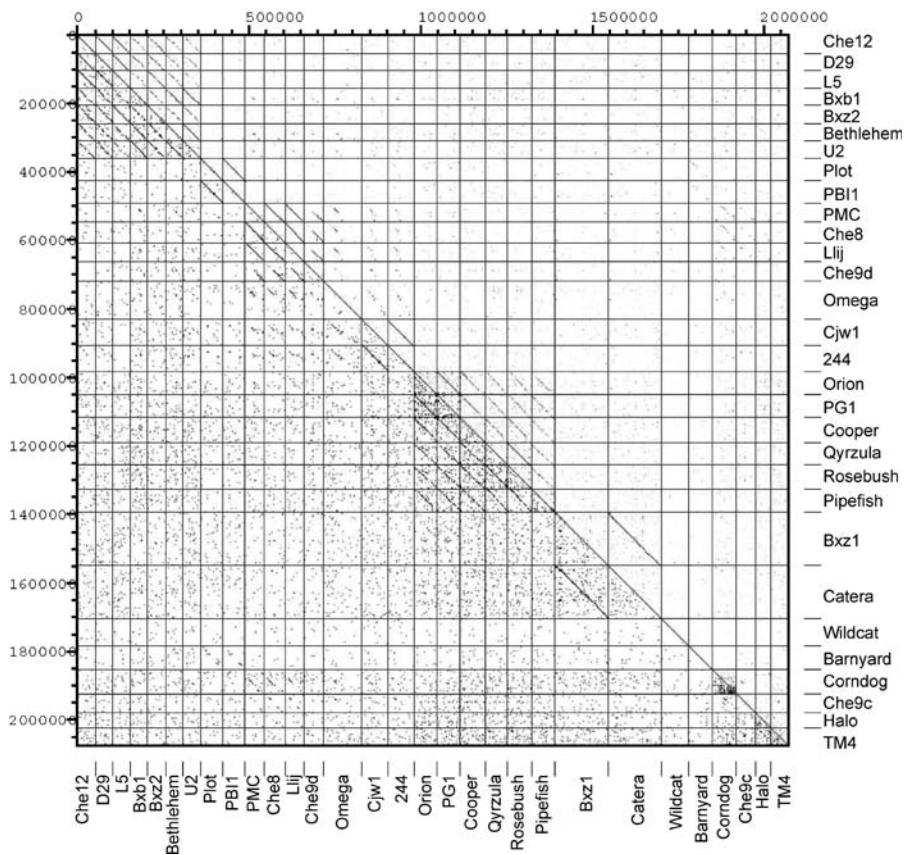## Nucleotide Sequence Diversity of Mycobacteriophages

Comparison of the 30 mycobacteriophages with each other at the nucleotide level reveals considerable overall diversity, with small groups having recognizable sequence similarity (Figure 1). The most numerous phage genome cluster contains seven that are more closely related to each other than to other phages; these include L5, the first sequenced mycobacteriophage genome [28], D29 [30], Bxb1 [29], Bxz2 [6], Che12, Bethlehem, and U2. The next most numerous group contains six members, including the previously described Rosebush [6], Orion, PG1, Cooper, Qyrzula, and Pipefish. Phages PMC, Che8, and Llij form another cluster, with parts of Che9d and Omega having similarity to these; two-member groups are formed by phages P-Lot and PBI1, Cjw1 and 244, and Bxz1 and Catera. Phages showing little or no nucleotide sequence similarity to any of the others are TM4, Halo, Che9c, Barnyard, Corndog, and Wildcat.

The profile of nucleotide sequence similarities differs in several notable ways from the previously reported comparison of only 14 genomes [6]. While the overall diversity remains very high, 11 of the newly sequenced genomes have recognizable nucleotide sequence similarity to at least one of the previously reported genomes. A particular surprise is the finding that five of the newly sequenced genomes have detectable nucleotide sequence similarity to the previously sequenced genome of phage Rosebush (of which no closely related genomes had been described), although, with the exception of Qyrzula, the degree of similarity is low. In contrast, the Bxz1/Catera and Cjw1/244 groups are more closely related than any other pairs of mycobacteriophage genomes. Both pairs exhibit greater than 90% nucleotide sequence identity with a number of small insertions or deletions accounting for much of the difference. In contrast to other groups of phages that have been analyzed, the diversity at the nucleotide sequence level of the mycobacteriophages appears to be greater than that of either dairy phages [14] or staphylococcal phages [17], although it is not yet clear whether this reflects underlying biological diversity differences or just the relatively small numbers that have been analyzed and the isolation approaches utilized.

There is no clear correlation between membership in these clusters of similar phages and the phages' geographic origins. Although all the members of the Rosebush group came from the environs of Pittsburgh, the seven members of the L5 cluster came from Japan, New York City, California, Chenai, India, and Bethlehem, PA. The smaller groups are also mostly geographically diverse. Our view at the current level of data collection is that the clusters of similar phages are widely distributed geographically.

Mycobacteriophage Genes and Gene Phamilies

The 30 mycobacteriophage genomes encode a total of 3,357 open reading frames (Table 1). As expected, the newly sequenced genomes possess a mosaic architecture similar to those described previously, with modules—frequently containing just a single gene—shared by otherwise distantly related phages. In order to better understand this genetic diversity, we have assembled all 3,357 open reading frames into gene phamilies, i.e., groups of related sequences, using the criteria that an encoded protein must share amino acid sequence similarity at an E-value of 0.001 or better or 25% amino acid identity across its length with at least one other

Figure 1. Nucleotide Sequence Comparison of 30 Mycobacteriophage Genomes as Illustrated in a Dotter Plot Using a Sliding Window of 25 bp [63]
The lower triangle represents the relationships at an elevated level of gray-scale relative to the upper triangle, revealing weaker sequence relationships.
DOI: 10.1371/journal.pgen.0020092.g001

member of the phamily (Table S1). This generates a total of 1,536 phamilies, with a mean phamily size of 2.19 genes (Figure 2). However, more than half of the phamilies (774, 50.3%) contain just a single constituent gene, and 88% of the genes are contained within phamilies containing three or fewer members (Figure 2).

Surprisingly, there are only three phamilies that contain members in all 30 of the mycobacteriophage genomes. Since these phamilies have the potential to contain genes or domains that are present in all mycobacteriophage genomes—and may thus correspond to mycobacteriophage signatures—we have examined these more closely. The first of these, Pham7, contains Lysin A genes, one of two putative lysins encoded by the mycobacteriophages [47]. The second lysin, Lysin B (Pham9), is also present in 27 of the 30 genomes (Figure 2); however, both of these phams are highly diverse, and they appear to be composed of subgenic modules with reasonably defined boundaries (Figure 3A). While each of the Pham7 members has sequence similarity to at least one other member of the phamily, no single sequence element within Pham7 is present in all 30 genomes (Figure 3A).
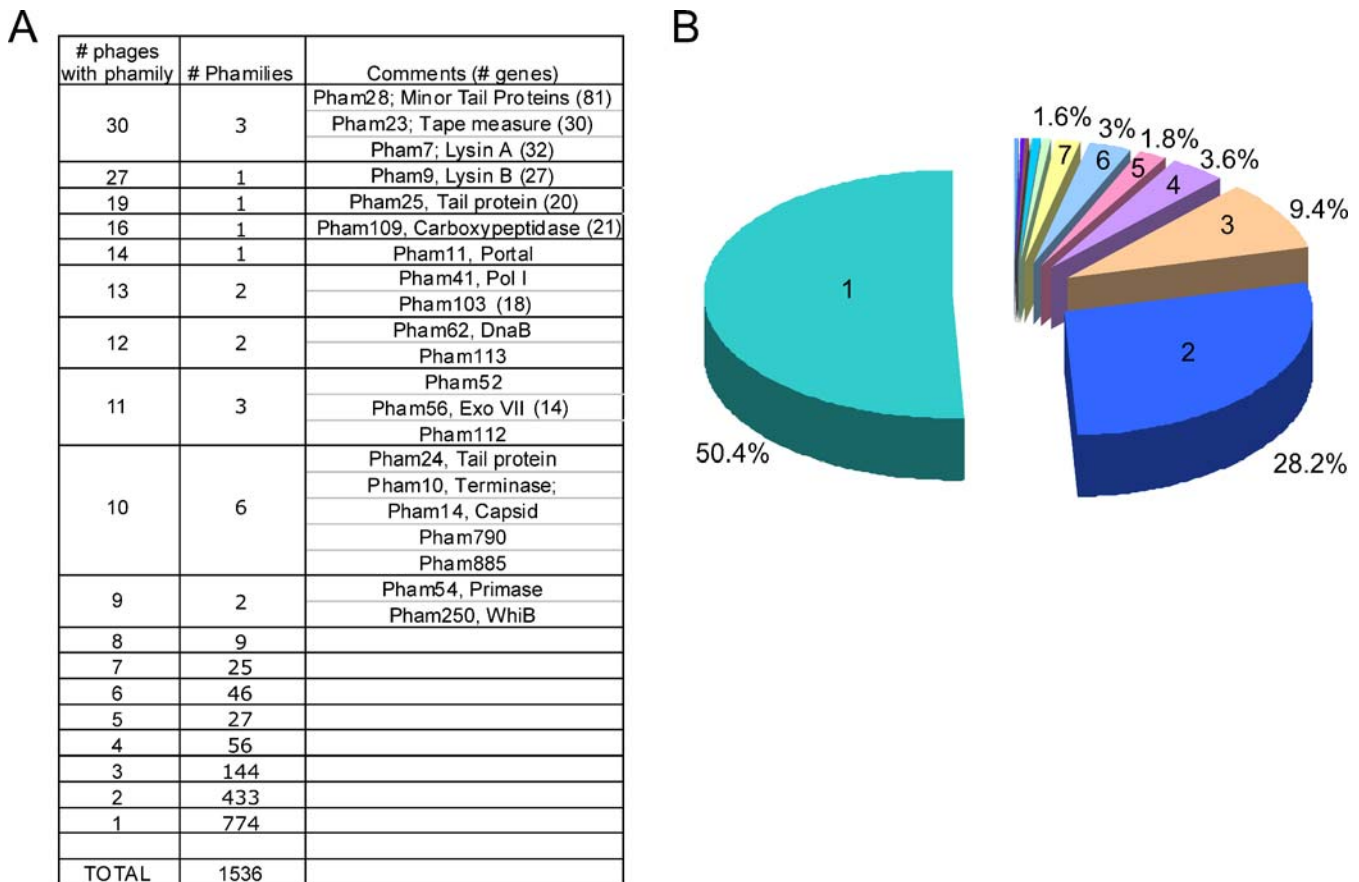
Both of the remaining 30-genome phamilies correspond to tail proteins. One of these encodes the tape-measure protein (Tmp) (Pham23) that plays a role in tail assembly and determines the lengths of noncontractile tails [6,33]. This phamily is also highly diverse, but the relationships among the members are complicated, and no well-defined boundaries between shared modules could be recognized; this

complexity is illustrated by the long branch lengths within a phylogenetic analysis (Figures 3B). The second tail protein phamily (Pham28) is equally complicated, with ill-defined module boundaries (Figure 3C), and a total of 81 genes fall into this phamily (Figure 2A). Nevertheless, as seen in Pham7, no single sequence element in either of these tail protein phamilies is present in all 30 genomes.

It is noteworthy that the six most abundant phamilies (Phams 28, 23, 7, 9, 25, and 109; Figure 2A) appear to be highly represented not just because of the conservation of essential functions but also because of their highly divergent and modular natures. These phamilies correspond to phage functions—predominantly tail proteins and lysis proteins—that are expected to be intimately involved in interacting with their bacterial hosts, and high diversity among phage tail proteins has been described previously [48,49]. Although the Pham109 members are related to carboxypeptidases (Figure 2A), Pham109 genes are typically located among tail genes [6,29]. Thus, we postulate that they are likely to be structural components of tails.

Phamily-Based Clustering of Mycobacteriophage Genomes

The comparison of mycobacteriophage genomes at the nucleotide level (Figure 1) reveals not only considerable genetic diversity but also small groups of phages that appear to be more closely related to each other than they are to other mycobacteriophages. To explore this further, we examined

A

# phages with phamily	# Phamilies	Comments (# genes)
30	3	Pham28; Minor Tail Proteins (81)
		Pham23; Tape measure (30)
		Pham7; Lysin A (32)
27	1	Pham9, Lysin B (27)
19	1	Pham25, Tail protein (20)
16	1	Pham109, Carboxypeptidase (21)
14	1	Pham11, Portal
13	2	Pham41, Pol I
		Pham103 (18)
12	2	Pham62, DnaB
		Pham113
11	3	Pham52
		Pham56, Exo VII (14)
		Pham112
10	6	Pham24, Tail protein
		Pham10, Terminase;
		Pham14, Capsid
		Pham790
		Pham885
9	2	Pham54, Primase
		Pham250, WhiB
8	9	
7	25	
6	46	
5	27	
4	56	
3	144	
2	433	
1	774	
TOTAL	1536	

B

Figure 2. Size and Distribution of Mycobacteriophage Phamilies

All 3,357 mycobacteriophage genes were assorted into 1,536 phamilies based on amino acid sequence similarity with a BLAST E value of 0.001 or better to at least one other member of the phamily.
(A) The distribution of phamilies is shown ranked according to the number of mycobacteriophage genomes containing at least one phamily member. Examples of specific phams and the total number of mycobacteriophage genes within that pham are shown.
(B) Pie-chart representation of the phamily-size distribution. Phamilies with eight or more members represent about 2% of the total.
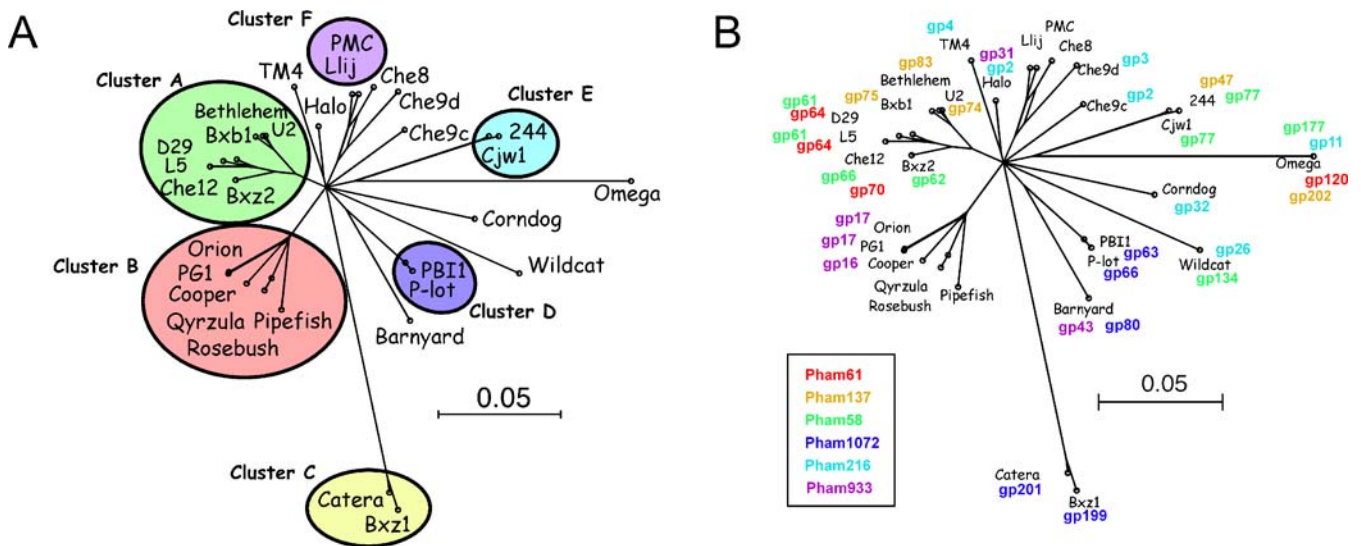DOI: 10.1371/journal.pgen.0020092.g002

the relationships among these phages by asking whether or not each genome contains a member of each gene phamily by using the program Splitstree [50], which accommodates alternative phylogenetic relationships, to express these data (Figure 4A). This analysis reveals six clearly defined groups of genomes that we have termed clusters A through F. Cluster A contains the previously characterized phages L5, D29, Bxb1, and Bxz2, plus the newly sequenced genomes of Bethlehem, Che12, and U2. The second most numerous cluster (cluster B) includes six genomes, of which only one (Rosebush) was previously characterized [6]; the remaining four clusters are closely related pairs of genomes (Figure 4A).

The relationships presented in Figure 4A help to organize the discussion of the general features of these genomes, and we note that this approach bears some resemblance to the phage proteomic tree described previously [51], although the Splitstree presentation enables at least partial inclusion of phylogenetic ambiguities that arise from the comparison of genomes that are pervasively mosaic in their architecture. But while these representations reflect the global relationships of the phages, it is important to note that they represent aggregate representations of the evolutionary histories of these viruses, and this ignores the numerous constituent phamilies that have phylogenies that are distinct from that

shown in Figure 4A. The six phamilies (Phams 61, 137, 58, 1,072, 216, and 933) shown in Figure 4B illustrate this problem, underscoring the important role of horizontal genetic exchange in phage evolution.

An alternative representation of phage genome relationships utilizes phamily circles to identify the participants and indicate the strength of relationships within each phamily (Figure 5). For example, Pham58 is present in eight genomes, with the strongest relationships between the phamily members in four of the seven components of cluster A and both members of cluster E. In contrast, Pham61 members are also in three of the same four members of cluster A as Pham58, but these are all closely related to a Pham61 member in Omega. This representation also illustrates the phylogeny of the intein present in the Pham216 member in Omega, which differs substantially from the remainder of the phamily members. While the phamily circles shown in Figure 5 obviously represent only a subset of all of the possible genetic relationships in the bacteriophages, a hypothetical extension to all of the 762 phamilies that contain two or more members would constitute a complete and accurate representation of the phylogenies of the protein-encoding genes of these genomes. The future development of automated circle drawing software should greatly facilitate this.

Figure 3. Complex Relationships within Highly Abundant Mycobacteriophage Phamilies

(A) Complex relationships among members of the Pham7 (Lysin A) phamily. The output of a BLAST comparison of Wildcat gp49 against other mycobacteriophage proteins shows that only 16 other mycobacteriophage proteins are matched and that these correspond to different parts of Wildcat gp49. Colored bars represent the strength of the matches, with red being the strongest, followed by purple, blue, and black.

(B) Phylogenetic relationships between members of mycobacteriophage Pham23 (tape-measure protein; Tmp). Amino acid sequences for each of the 30 constituent members of Pham23 were aligned using ClustalW and the unrooted phylogenetic relationships represented using NJTree. Bootstrap values from 1,000 reiterations are shown.

(C) Chimerism in Pham28 (minor tail) proteins. Llij gp18 is related to both gp18 and gp19 of phage Che8 at high levels of amino acid sequence identity, and these proteins are related in turn to other members of Pham28 as shown.

DOI: 10.1371/journal.pgen.0020092.g003

How Many Mycobacteriophage Phams Exist?

To estimate the total number of phams shared among mycobacteriophages, we calculated the number of phams found among randomly chosen subsets of the 30 phages described above (Figure 6). A hyperbolic curve was fit to the data, where PhamMax is the maximum number of phams in the population being sampled, and K__Phage is the number of phages that must be sampled to uncover one-half of the phams. We found that a curve with PhamMax = 3,064 and K__Phage = 29 fit the data well (Figure 6). That is, this sample of 30 phages contains representatives of only half of the predicted total number of 3,064 phams to be found among mycobacteriophages, and sequencing a further 100 mycobacteriophage genomes would probably uncover members of an additional approximately 1,000 phams, assuming that phages that are closely related to known phages are not excluded from the sample. We note that on average the addition of a thirtieth phage to a randomly selected group of 29 phages leads to identifying approximately 25 new phams.

Genetic Novelty of Phage Phamilies

The organization of mycobacteriophage genes into phamilies simplifies the process of determining how these are related to genomes of other phages and their bacterial hosts. For example, of the 1,536 phamilies, only 230 (15%) have recognizable sequence similarity (Blast E value of 0.001 or better) to existing (nonmycobacteriophage) database entries (Figure 7). Remarkably, over 85% of these mycobacterio-

Figure 4. Representation of Mycobacteriophage Clusters Using Splitstree

(A) The relationships between 30 mycobacteriophages are represented by Splitstree representation of a dataset in which each of the 1,536 gene phamilies is annotated as being either present or absent in each of the 30 genomes. Clusters A through F of genomes that are more closely-related to each other than to other mycobacteriophages are shown by colored circles.

(B) The distribution of the members of six phamilies on the Splitstree representation in (A) illustrates that individual phamilies have notably different evolutionary histories than the aggregate representation.

DOI: 10.1371/journal.pgen.0020092.g004

phage phamilies, a total of 1,306, represent previously unidentified genetic sequences, consistent with the idea that phages may represent the largest reservoir of unexplored genes in the biosphere [6, 19]. Interestingly, only 126 (54.8%) of the 230 phamilies with matches correspond to genes found in other phages, whereas 104 (45.2%) are related to nonphage (predominantly bacterial) genomes; 57 (24.8%) of the 230 phamilies are found in both phage and bacterial genomes (Figure 7). Some the phams listed as exclusively matching bacterial genomes may be identifying unrecognized or unannotated prophages.

It is clear from this analysis as well as that of other phage genomes that there is little overlap between the metaproteome of phages and that of prokaryotes. Moreover, typically, greater than 80% of genes within bacterial genomes can be assigned to clusters of orthologous groups, and in some cases (e.g., *Buchnera*) virtually every gene can [52, 53]; this is in contrast to the high proportion of mycobacteriophage genes (approximately 50%) that are unrelated to other genes (i.e., ORFans). Grouping of the phage genes into phamilies is therefore justified similarly to the separate ordering of eukaryotic genes into eukaryotic clusters of orthologous genes [53]. We also note that only a small proportion of mycobacteriophage phamilies (approximately 20%) would qualify as adding to or expanding the current cluster of orthologous groups database.

Phamily Flux

Examination of the phams that match other phage or nonphage genes, or both, reveals qualitative differences among these groups, particularly in regard to the numbers of mycobacteriophage genomes within each pham. For example, the average pham size (defined as the number of mycobacteriophage genomes containing a phamily member) of all phams with matches to nonmycobacteriophage genes is

3.88, compared to 1.83 for those with no matches (Figure 7). Thus, if a mycobacteriophage gene has sequence similarity to any nonmycobacteriophage gene, it will have more relatives within the group of mycobacteriophages than one that does not.

The average pham size is different for the groups of phams that match phage and nonphage genes (Figure 7). However, the average pham size for the 126 phams matching all other phage genes is 4.67, substantially larger than the average size for all phams (2.19) or those with matches (3.88) (Figure 7); the subset of these that specifically matches both phage and nonphage sequences is even larger, with an average pham size of 5.12 (Figure 7). On the other hand, the average pham size for the 104 phams that exclusively match nonphage genes is only 2.91.

What is the basis for different distributions of pham sizes? While the dataset remains relatively small and the 30 genomes are not all equally different from each other (Figures 1 and 4), we postulate that the differences in average pham sizes reflect the extent to which these phams provide functions of general utility to the mycobacteriophages (i.e., high pham size) or provide specialist functions to smaller numbers of mycobacteriophages that are required to infect specific hosts to or survive within particular environmental or biological niches. Examples of the pham group that match other phages include both structure and assembly genes (e.g., capsids, portals, terminases) as well as nonstructural genes (e.g., *ruvC,* integrases, excises), but these phams presumably provide important functions to both mycobacteriophages and other phages. In contrast, the phams matching nonphage gene products with smaller pham sizes include gene products such as WhiB, Glutaredoxins, FtsK, Lsr2, DinD, DinG, Ro, and PurA; these presumably provide more specialist functions. It seems likely that these specialist functions have been acquired directly from host genomes, and they are relatively rare
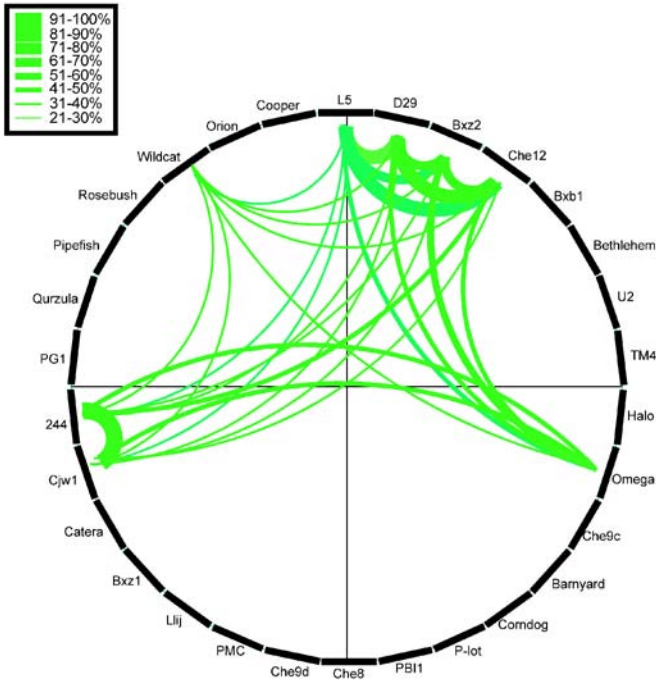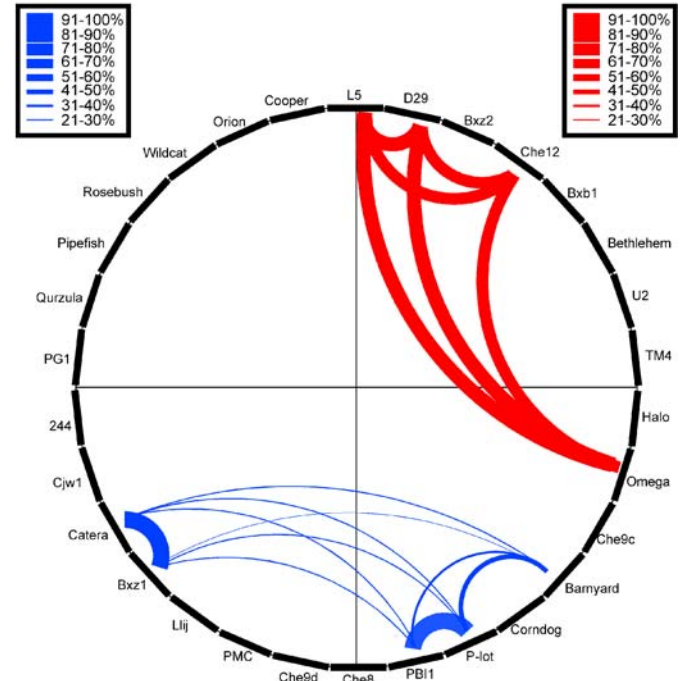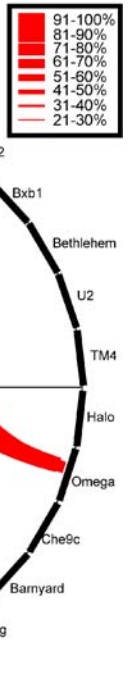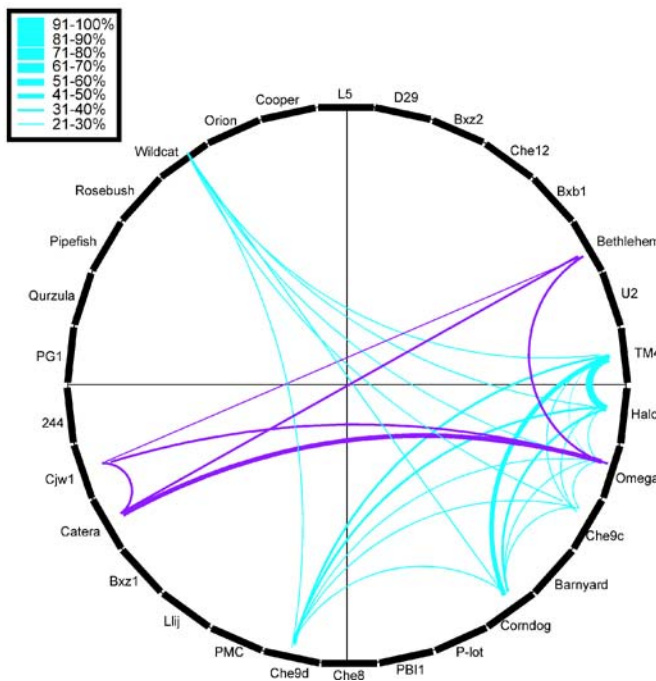
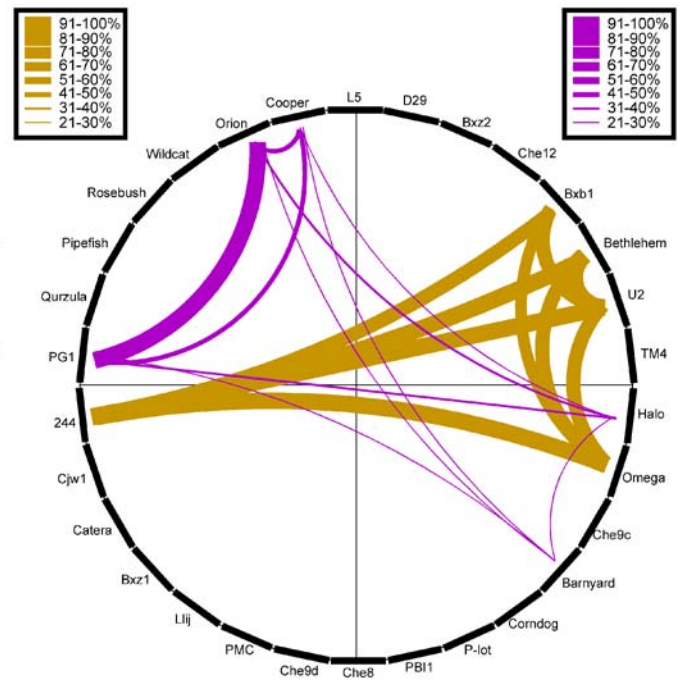**Figure 5.** Phamily Circle Representations of Phamily Relationships

All 30 genomes are shown around the circumference, and the phamily members are linked by a line with the width representing the degree of similarity. Phamily circles of Pham58 (upper left), Pham61 and Pham1072 (upper right), and Pham137 and Pham993 (lower right) are shown using different colors for different Phams. Pham216 (bottom left) is shown in turquoise, with the intein present within the Omega phamily member (which has a different set of relationships) shown in purple.

DOI: 10.1371/journal.pgen.0020092.g005
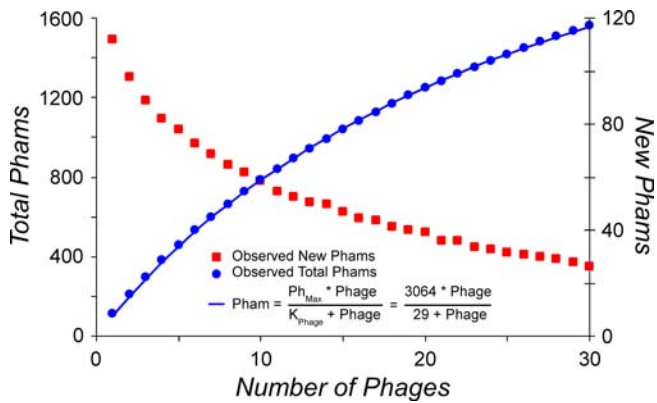
Figure 6. Estimating the Number of Mycobacteriophage Phams

A subset of the 30 phages was randomly selected without replacement, and the total number of Phams was determined; this was repeated 10,000 times with the mean shown as a blue circle. For each subset, an additional phage was then randomly chosen, and the average number of new Phams found in that phage was determined; these data are shown as red squares. The total number of Phams was fit to a hyperbolic function, with the best-fit equation determined by least-squares regression.

DOI: 10.1371/journal.pgen.0020092.g006

because of their restricted utility to a subset of phages, rather than their lack of opportunity to spread more broadly among this group of phages that infect at least one common host (i.e., *Mycobacterium smegmatis*).

Constraints on the Modular Construction of Phage Genomes

If there are approximately 3,000 different sequence phamilies among phages that infect *M. smegmatis*, how many different ways can these be combined to make a functional phage genome? If we assume that an average genome contains 100 genes, then there are approximately 10^{350} possible ordered phamily combinations. Since there are estimated to be 10^{31} phage particles in the biosphere and mycobacteriophages represent only a fraction of these, it is clear that only a miniscule fraction of the conceivable gene combinations and organizations have been used. More important, we expect that there are strong functional constraints imposed on the generation of competent phage genomes. These constraints can be grouped into four main categories. First, each genome must encode all necessary functions not provided by the host, mandating inclusion of virion structural and assembly genes, lysis genes, and possibly some DNA replication functions. Thus, a specific subset of phamilies with required functions must be present in each genome. Second, specific gene combinations are required when a gene can provide functional benefits only in combination with another gene. One example might be the joint action of holin and lysin components of the lysis machinery. Third, some phamilies may encode alternative functions, where the inclusion of a member of one phamily (e.g., a virion capsid) precludes the inclusion of any members from another phamily (e.g., a nonhomologous capsid protein). Fourth, gene organization is likely to be an important constraint, for example, allowing cotranscription of genes that need to be coregulated.

Clearly, the 30 mycobacteriophage genomes are not simply the result of random assortment of phamilies. At both the
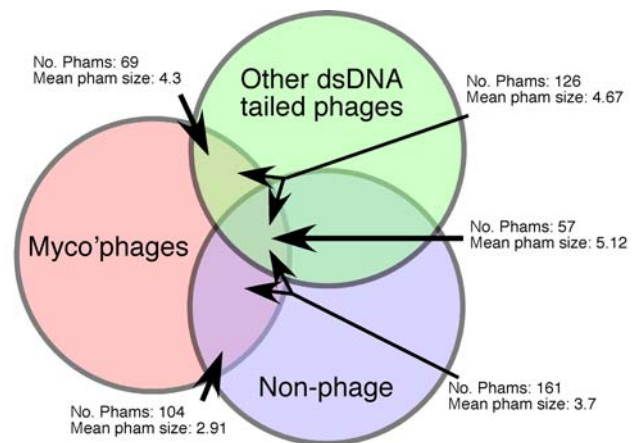
Total No. Myco'phage Phams with matches: 230 (15% total Phams)
Average size of all Phams with matches: 3.88
Average size of all Phams without matches: 1.83

Number of other (i.e. non myco'phage) dsDNA tailed phages: 143
Mean Pham size in Myco'phage genomes: 2.16

Figure 7. Relationships between Mycobacteriophage Phams and Previously Sequenced Proteins

The number and size of mycobacteriophage phamilies with sequence similarity to nonmycobacteriophage genes are shown. The numbers of Phams shared by mycobacteriophages, other phages, and nonphage genomes are shown, along with the average pham size, defined as the number of mycobacteriophage genomes containing at least one member of that phamily. The red circle represents mycobacteriophage genomes, the green circle represents all dsDNA phage genomes other than the mycobacteriophages, and the blue circle represents all nonphage genomes. The number of phams shared between these groups and the mean mycobacteriophages pham size of those phams are shown, with arrows indicating whether they are shared by mycobacteriophages (red circle), nonmycobacteriophage phage genomes (green circle), or nonphage genomes (blue circle).

DOI: 10.1371/journal.pgen.0020092.g007

nucleotide (Figure 1) and protein (Figure 4) levels, we can identify clusters of phages that exhibit greater levels of similarity to each other than to the larger population; for example, 21 of the 30 phages fall into six separate clusters at the nucleotide sequence and the phamily inclusion levels used here (Figure 4). While it is likely that this clustering reflects in part the particular host and the isolation procedures used, it also suggests that there are pools of related phages that enjoy success within the current environment. This is reminiscent of what has been seen for phages of other hosts, including those of *Escherichia coli,* where separate clusters are represented by phages lambda, T4, T7, and P2. The types of phages defined by the clusters of mycobacteriophages are not obviously correlated with the types seen in the coliphages, suggesting the possibility that each host type has a distinct set of phage types associated with it. We suspect instead that each of the clusters defined by the current collection of genome sequences is part of a continuum of types extending over a range of hosts, with no sharply definable boundaries. However, more data will be required to clarify this question.

It seems likely that an additional constraint on the assortment of phamilies is that their distribution within genomes is not independent of each other. For example, examination of the data set reveals that Pham56 is found in diverse bacteriophages but only in those that carry the

Table 2. Isolation, Sequencing, and Annotation of Mycobacteriophage Genomes

Phage	Isolated by	Status	Sequencer[a]	Annotator[a]
Catera	Andrew Hrykowian	High school student	AH	AH
Halo	Brian Simanek	High school student	BS	BS
Wildcat	Joe Gross	High school student	PBIGC	MLP
	Jake Falbo	High school student		
Pipefish	Amy Foley	High school teacher	AF/MP	AF/MP
	Mark Popovich	High school teacher		
244	Rebecca Gonda	Undergraduate student	RG	RG
Cooper	Bill Brucker	Undergraduate student	PBIGC	MLP
Llij	Gina Zdanowicz	Undergraduate student	GZ	GZ
Orion	Don Schleicher	Undergraduate student	DS	DS
PMC	Polly Cichon	Undergraduate student	PC	PC
Qyrzula	Swathi Namburi	Undergraduate student	SN	SN
	Kostandin Pajcini	Undergraduate student	KP	KP
Bethlehem	Vanessa Kelchner	Research assistant	VK	VK
U2	Vanessa Kelchner	Research assistant	VK	VK
Che12	Prabakaran Loganathan	Graduate student	PBIGC	MLP
PBI1	Lori Bibb	Graduate student	PBIGC	JH
	John Lewis	Graduate student		
	Curt Wadsworth	Graduate student		
PG1	Lori Bibb	Graduate student	PBIGC	MLP
	John Lewis	Graduate student		
	Curt Wadsworth	Graduate student		
P-Lot	Rajeswari Dandapani	Volunteer	PBIGC	MLP and CB group

[a]Initials indicate author primarily responsible for sequencing and annotation. PBIGC, Pittsburgh Bacteriophage Institute Genome Center; CB group, a group of ten University of Pittsburgh undergraduate students taking a course in computational biology. Phage genome sequence determination and annotation were considered a contribution warranting authorship of this communication. Other authors contributed to experimental design, data analysis, and manuscript preparation.
DOI: 10.1371/journal.pgen.0020092.t002

Pham25 minor tail protein, suggesting both a role for Pham56 in virion assembly and a lack of function in genomes lacking Pham25. The nine members of Pham297 are not only found in Pham25-bearing genomes, but they are adjacent to the Pham25-encoding genes, suggesting an even more intimate association between these proteins. The coassociation of genes among bacteriophage genomes can only be assessed using a large collection of diverse genomes, and this demonstrates how datasets such as this one can be used to uncover plausible gene functions and interactions.

Phage Discovery as an Educational Tool

The newly discovered phages described here were isolated and characterized by junior members of the laboratory, including high school and undergraduate students (Table 2). The educational benefits of performing scientific research have been reported previously [54,55], although identifying suitable research projects and laboratory environments represents a significant challenge to the research community. The Phagehunters Program developed at the University of Pittsburgh—in which students discover and genomically characterize their own bacteriophages—provides a particularly strong combination of attributes that maximize the educational benefits within a research environment; we note that a successful program with some common features also was developed by R. Young and colleagues [56]. While there may be many others, we have identified seven key attributes of this project that facilitate successful outcomes, and these are described in Table 3. Identifying the key features of projects that are well suited for student research should facilitate the identification and development of other effective programs.

Two particularly important features of this educational platform are the strong emphasis on scientific discovery and project ownership. As evident from the comparative genomic analysis described above, there are three main discovery elements. First, there is an excellent prospect of each student isolating a phage from the environment that is different to all previously described viruses. Second, within each genome there is an opportunity to discover new genes that are distinct to all previously identified genes. Third, many of these phage genomes surprisingly contain homologs of known genes that have not been previously found in phage genomes. The high diversity of the mycobacteriophage population (Figure 1), the preponderance of novel genes (Figure 7), and the mosaic architecture of these genomes provide a high promise of discovery for each participating student.

The opportunity for students to discover novel genes and viruses is important since it is stimulating and highly motivating, providing a strong impetus for students to become engaged in scientific research and to maintain their involvement even through the more challenging aspects of their projects. The modest genome size of the phages is such that a single student can reasonably manage an individual phage, and project ownership adds motivation and commitment. The prospect of naming their new phage generates considerable interest among these novice scientists while also offering opportunities to learn about the justification of a nonsystematic nomenclature for viruses that contain genomes with highly individualized mosaic architectures. This educational platform is clearly not restricted to mycobacteriophages and can be readily extended to the isolation and characterization of environmental phages for other nonpathogenic hosts; comparative genomic analyses of dairy,

Table 3. Seven Helpful Attributes for High School and Undergraduate Research Projects

Attribute	Description (Phage-Hunting Examples)
1	Technical simplicity, especially at initial stages
	The initial phage isolation procedures are technically approachable, and while they become more sophisticated with genomic characterization, they are within the grasp of pre-college and undergraduate students working within a pre-established genomics infrastructure.
2	Conceptual simplicity and minimal background requirements
	To begin phage hunting, no deep conceptual understanding beyond the high school curriculum is required, although conceptual appreciation is acquired throughout the project.
3	Compatibility with flexible scheduling
	The varied demands on students' time requires scheduling accommodations, and phage isolation and characterization offers flexible timing.
4	Multiple achievement milestones
	Significant research and educational accomplishments accrue at distinct steps, such as phage isolation, DNA purification, microscopy, sequence determination, and annotation. Achieving each milestone is a success.
5	Parallel project structure enabling greater numbers of students
	While sacrificing some student-initiated research planning, parallel projects facilitate the involvement of multiple students and are well suited to peer and near-peer mentoring.
6	Real research that is publishable and interesting to others
	The value of the scientific findings is validated by peer-reviewed publication.
7	Project ownership provides strong motivation
	Identifying a new biological entity is an exciting prospect, and the discovery by students of their own new virus—virtually guaranteed given the genetic diversity of bacteriophages and their great abundance—is highly motivating. Furthermore, naming new viruses using nonsystematic nomenclature—justified by the nonhierarchical taxonomic relationships—is fun!

staphylococcal, and pseudomonal phages show that these also are genetically diverse and typically harbor a high proportion of novel genes [14,16,17].

The development of multiple projects with parallel structures as described here brings both advantages and disadvantages. The main disadvantage is that since the research path is well established, students generally have only limited opportunities for experimental planning and design. However, there are also considerable advantages to parallel project structures. First, it provides opportunities for a greater number of students to participate than if each were independently structured. Second, it facilitates the training of students through peer and near-peer mentoring systems, and we find that mentorship of high school students by undergraduates is a particularly effective combination. Moreover, undergraduate student mentors can be readily trained using the Entering Mentoring program developed by Pfund and colleagues [57]. The parallel project structure, coupled with this mentoring system and the technical simplicity of the initial project stages, represents essential ingredients for enabling students in the early stages of their educational development to engage in scientific research.

Finally, this phage discovery educational platform requires only modest prior comprehension of biological facts and concepts. This simplifies access of young students to scientific research and provides opportunities to students who do not necessarily excel in more traditional classroom settings. The platform offers numerous opportunities for students to learn concepts in microbiology, ecology, genetics, computational biology, and evolution within an inquiry-driven environment and is fully inclusive of a diverse variety of learning styles. Additionally, the significant bioinformatics component of the program appeals to students with computer science and engineering backgrounds, and in doing so, it creates a diverse research group that offers advantages both to the participants and the research agenda itself. Detailed protocols are available at http://www.pitt.edu/~gfh.

Concluding Remarks

The comparative genomic analysis of 30 mycobacteriophage genomes provides important new insights into the diversity and architecture of phage genomes and offers insights about gene exchange between phage genomes and between phages and their hosts. It is likely that these general features will be shared by most other phages, and the recent comparative analysis of 27 *Staphylococcus* and 18 *Pseudomonas* phages also shows relatively high genetic diversity [16,17]. Phage isolation and genomics is a powerful educational platform that provides research opportunities to students from diverse educational backgrounds. The high diversity of the phage population offers the excitement that each student can isolate a unique virus and uses genomic approaches to understand the relationship of the newly discovered phage to the broader biological world. The ability of students to contribute successfully to achieving the key scientific goals of understanding viral diversity, and the underlying evolutionary mechanisms that give rise to it, suggests that phage isolation and characterization can be used broadly for educational purposes.

Materials and Methods

Phage isolation. Phages were isolated from the following locations: Cooper, Halo, Llij, Orion, PBI1, PG1, P-Lot, Pipefish, PMC, and Qyrzula were from Pittsburgh; Bethlehem and U2, Bethlehem, Pennsylvania, Unites States; Catera and Wildcat, Latrobe, Pennsylvania, United States; 244, Connecticut; and Che12, Chennai, India. Samples from various sources were extracted with phage buffer, plated directly onto solid overlays containing 0.35% agar and *Mycobacterium smegmatis* mc²155, and incubated at 37 °C for 24 h as described previously, with the exception of Che12, which was isolated using *Mycobacterium tuberculosis* as a host [6]. Individual plaques were picked and purified through several rounds and purified by CsCl equilibrium density gradient centrifugation.

Genome sequencing and analysis. Approximately 10 µg of purified phage DNA was sheared hydrodynamically and repaired, and 1- to 3-kbp fragments inserted into plasmid pBluescript (Stratagene, La Jolla, California, United States). Individual clones were sequenced using ABI3730 or ABI3100 instruments (Applied Biosystems, Foster City, California, United States), and the sequences were assembled into a single or small number of contigs [58]. At approximately 8-fold redundancy, sequence data from oligonucleotide primers used with phage template DNA generated a single contig and resolved sequence ambiguities. Genome termini were sometimes identifiable as an overabundance of clone ends, and a comparison of the sequences generated using primers annealed to ligated and unligated phage DNA allowed determination of the molecular ends, where possible.

Sequence assembly was performed using the Phred/Phrap/Consed suite of programs and annotated using a variety of programs including DNA Master (available from http://cobamide2.bio.pitt.edu), Genemark [59], and Glimmer [60]. tRNA and tmRNA genes were identified using tRNA-Scan-SE and ARAGORN [61,62]. BLAST analyses were performed either locally or remotely at the National Center for Biotechnology Information and were used to assemble a database of related sequences in Microsoft Excel.

Supporting Information

Table S1. The Pham Database

Each of the putative proteins encoded by 30 mycobacteriophage genomes are tabulated according to sequence similarity and assembled into phamilies. Putative pham functions are also listed.

Found at DOI: 10.1371/journal.pgen.0020092.st001 (250 KB XLS).

Accession Numbers

GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession numbers for phages are L5 (Z18946), D29 (AF022214), Bxb1 (AF271693), TM4 (AF068845), Barnyard (AY129339), Bxz1 (AY129337), Bxz2 (AY129332), Che8 (AY129330), Che9c (AY129333), Che9d (AY129336), Corndog (AY129335), Cjw1 (AY129331), Omega (AY129338), Rosebush (AY129334), Catera (DQ398053), Halo (DQ398042), Wildcat (DQ398052), Pipefish (DQ398049), 244 (DQ398041), Cooper (DQ398044), Llij (DQ398045), Orion (DQ398046), PMC (DQ398050), Qyrzula (DQ398048), Bethlehem (AY500153), U2 (AY500152), Che12 (DQ398043), PBI1 (DQ398047), PG1 (AF547430), and P-Lot (DQ398051).

Acknowledgments

References

1. Suttle CA (2005) Viruses in the sea. Nature 437: 356–361.
2. Wommack KE, Colwell RR (2000) Virioplankton: Viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64: 69–114.
3. Hambly E, Suttle CA (2005) The viriosphere, diversity, and genetic exchange within phage communities. Curr Opin Microbiol 8: 444–450.
4. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541–548.
5. Wilhelm SW, Suttle CA (1999) Viruses and nutrient cycles in the sea. Bioscience 49: 781–788.
6. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, et al. (2003) Origins of highly mosaic mycobacteriophage genomes. Cell 113: 171–182.
7. Hendrix RW (2002) Bacteriophages: Evolution of the majority. Theor Popul Biol 61: 471–480.
8. Brussow H, Hendrix RW (2002) Phage genomics: Small is beautiful. Cell 108: 13–16.
9. Hendrix RW (2003) Bacteriophage genomics. Curr Opin Microbiol 6: 506–511.
10. Hendrix RW, Hatfull GF, Smith MC (2003) Bacteriophages with tails: Chasing their origins and evolution. Res Microbiol 154: 253–257.
11. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. Proc Natl Acad Sci U S A 96: 2192–2197.
12. Hendrix RW (1999) The long evolutionary reach of viruses. Curr Biol 9: 914–917.
13. Bamford DH, Burnett RM, Stuart DI (2002) Evolution of viral structure. Theor Popul Biol 61: 461–470.
14. Brussow H (2001) Phages of dairy bacteria. Annu Rev Microbiol 55: 283–303.
15. Casjens SR (2005) Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microbiol 8: 451–458.
16. Kwan T, Liu J, Dubow M, Gros P, Pelletier J (2006) Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. J Bacteriol 188: 1184–1187.
17. Kwan T, Liu J, DuBow M, Gros P, Pelletier J (2005) The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. Proc Natl Acad Sci U S A 102: 5174–5179.
18. Maniloff J, Ackermann HW (1998) Taxonomy of bacterial viruses: Establishment of tailed virus genera and the order Caudovirales. Arch Virol 143: 2051–2063.
19. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504–510.
20. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250–14255.
21. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, et al. (2004) Diversity and population structure of a nearshore marine sediment viral community. Proc R Soc Ser B 271: 565–574.
22. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 185: 6220–6223.
23. Cann AJ, Fandrich SE, Heaphy S (2005) Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes 30: 151–156.
24. Hatfull GF (2004) Mycobacteriophages and tuberculosis. In: Eisenach K, Cole ST, Jacobs WR Jr, McMurray D, editors. Tuberculosis. Washington (D.C.): ASM Press.
25. Hatfull GF (2004) Mycobacteriophages. In: Calendar R, editor. The bacteriophages. New York: Oxford University Press. pp. 602–620.
26. Hatfull GF (1994) Mycobacteriophage L5: A toolbox for tuberculosis. ASM News 60: 255–260.
27. McNerney R (1999) TB: The return of the phage. A review of fifty years of mycobacteriophage research. Int J Tuberc Lung Dis 3: 179–184.
28. Hatfull GF, Sarkis GJ (1993) DNA sequence, structure and gene expression of mycobacteriophage L5: A phage system for mycobacterial genetics. Mol Microbiol 7: 395–405.
29. Mediavilla J, Jain S, Kriakov J, Ford ME, Duda RL, et al. (2000) Genome organization and characterization of mycobacteriophage Bxb1. Mol Microbiol 38: 955–970.
30. Ford ME, Sarkis GJ, Belanger AE, Hendrix RW, Hatfull GF (1998) Genome structure of mycobacteriophage D29: Implications for phage evolution. J Mol Biol 279: 143–164.
31. Ford ME, Stenstrom C, Hendrix RW, Hatfull GF (1998) Mycobacteriophage TM4: Genome structure and gene expression. Tuber Lung Dis 79: 63–73.
32. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, et al. (2000) Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 299: 27–51.
33. Katsura I, Hendrix RW (1984) Length determination in bacteriophage lambda tails. Cell 39: 691–698.
34. Brussow H, Desiere F (2001) Comparative phage genomics and the evolution of Siphoviridae: Insights from dairy phages. Mol Microbiol 39: 213–222.
35. Liu M, Gingery M, Doulatov SR, Liu Y, Hodes A, et al. (2004) Genomic and genetic analysis of Bordetella bacteriophages encoding reverse transcriptase-mediated tropism-switching cassettes. J Bacteriol 186: 1503–1517.
36. Liu M, Deora R, Doulatov SR, Gingery M, Eiserling FA, et al. (2002) Reverse transcriptase-mediated tropism switching in Bordetella bacteriophages. Science 295: 2091–2094.
37. Jensen EC, Schrader HS, Rieland B, Thompson TL, Lee KW, et al. (1998) Prevalence of broad-host-range lytic bacteriophages of *Sphaerotilus natans, Escherichia coli,* and *Pseudomonas aeruginosa*. Appl Environ Microbiol 64: 575–580.
38. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8: 504–508.
39. Lawrence JG, Hendrix RW, Casjens S (2001) Where are the pseudogenes in bacterial genomes? Trends Microbiol 9: 535–540.
40. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.
41. Brussow H, Canchaya C, Hardt WD (2004) Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev 68: 560–602.
42. Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brussow H (2003) Phage as agents of lateral gene transfer. Curr Opin Microbiol 6: 417–424.
43. Boyd EF, Davis BM, Hochhut B (2001) Bacteriophage-bacteriophage

interactions in the evolution of pathogenic bacteria. Trends Microbiol 9: 137–144.

44. Casjens S (2003) Prophages and bacterial genomics: What have we learned so far? Mol Microbiol 49: 277–300.

45. Wagner PL, Waldor MK (2002) Bacteriophage control of bacterial virulence. Infect Immun 70: 3985–3993.

46. Mekalanos JJ, Rubin EJ, Waldor MK (1997) Cholera: Molecular basis for emergence and pathogenesis. FEMS Immunol Med Microbiol 18: 241–248.

47. Garcia M, Pimentel M, Moniz-Pereira J (2002) Expression of mycobacteriophage Ms6 lysis genes is driven by two sigma(70)-like promoters and is dependent on a transcription termination signal present in the leader RNA. J Bacteriol 184: 3034–3043.

48. Desplats C, Krisch HM (2003) The diversity and evolution of the T4-type bacteriophages. Res Microbiol 154: 259–267.

49. Tetart F, Desplats C, Krisch HM (1998) Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: Recombination between conserved motifs swaps adhesin specificity. J Mol Biol 282: 543–556.

50. Huson DH (1998) SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 14: 68–73.

51. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: A genome-based taxonomy for phage. J Bacteriol 184: 4529–4535.

52. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.

53. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4: 41.

54. Lapatto D (2004) Survey of Undergraduate Research Experiences (SURE): First findings. Cell Biol Educ 3: 270–277.

55. Seymour E, Hunter A-B, Laursen SL, DeAntoni T (2004) Establishing the benefits of undergraduate research for undergraduates in the sciences: First findings from a three-year study. Sci Educ 88: 493–594.

56. Summer EJ, Gonzalez CF, Carlisle T, Mebane LM, Cass AM, et al. (2004) *Burkholderia cenocepacia* phage BcepMu and a family of Mu-like phages encoding potential pathogenesis factors. J Mol Biol 340: 49–65.

57. Pfund C, Maidl Pribbenow C, Branchaw J, Miller Lauffer S, Handelsman J (2006) Professional skills. The merits of training mentors. Science 311: 473–474.

58. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. Genome Res 8: 195–202.

59. Borodovsky M, McIninch J (1993) Recognition of genes in DNA sequence with ambiguities. Biosystems 30: 161–171.

60. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641.

61. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.

62. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32: 11–16.

63. Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167: GC1–GC10.