doi:10.1016/j.jmb.2010.01.011 J. Mol. Biol. (2010) 397, 119–143

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size

Graham F. Hatfull[1]*, Deborah Jacobs-Sera[1], Jeffrey G. Lawrence[1], Welkin H. Pope[1], Daniel A. Russell[1], Ching-Chung Ko[1], Rebecca J. Weber[1], Manisha C. Patel[1], Katherine L. Germane[1], Robert H. Edgar[1], Natasha N. Hoyte[1], Charles A. Bowman[1], Anthony T. Tantoco[1], Elizabeth C. Paladin[1], Marlana S. Myers[1], Alexis L. Smith[1], Molly S. Grace[1], Thuy T. Pham[1], Matthew B. O'Brien[1], Amy M. Vogelsberger[1], Andrew J. Hryckowian[1], Jessica L. Wynalek[1], Helen Donis-Keller[2], Matt W. Bogel[1], Craig L. Peebles[1], Steven G. Cresawn[3] and Roger W. Hendrix[1]

[1]*Department of Biological Sciences, Pittsburgh Bacteriophage Institute, Pittsburgh, PA 15260, USA*

[2]*Franklin W. Olin College of Engineering, Needham, MA 02492, USA*

[3]*Department of Biology, James Madison University, Harrisonburg, VA 22807, USA*

Received 19 October 2009; received in revised form 8 December 2009; accepted 5 January 2010
Available online 11 January 2010

Mycobacteriophages are viruses that infect mycobacterial hosts. Expansion of a collection of sequenced phage genomes to a total of 60—all infecting a common bacterial host—provides further insight into their diversity and evolution. Of the 60 phage genomes, 55 can be grouped into nine clusters according to their nucleotide sequence similarities, 5 of which can be further divided into subclusters; 5 genomes do not cluster with other phages. The sequence diversity between genomes within a cluster varies greatly; for example, the 6 genomes in Cluster D share more than 97.5% average nucleotide similarity with one another. In contrast, similarity between the 2 genomes in Cluster I is barely detectable by diagonal plot analysis. In total, 6858 predicted open-reading frames have been grouped into 1523 phamilies (phams) of related sequences, 46% of which possess only a single member. Only 18.8% of the phams have sequence similarity to non-mycobacteriophage database entries, and fewer than 10% of all phams can be assigned functions based on database searching or synteny. Genome clustering facilitates the identification of genes that are in greatest genetic flux and are more likely to have been exchanged horizontally in relatively recent evolutionary time. Although mycobacteriophage genes exhibit a smaller average size than genes of their host (205 residues compared with 315), phage genes in higher flux average only 100 amino acids, suggesting that the primary units of genetic exchange correspond to single protein domains.
© 2010 Elsevier Ltd. All rights reserved.

Keywords: bacteriophage; genomics; tuberculosis; mycobacteriophage; evolution

Edited by J. Karn

Introduction

Bacteriophages are the most numerous biological entities in the biosphere, and their genetic diversity and abundant novel gene sequences suggest that they harbor the greatest unexplored reservoir of genetic information.[1,2] The phage population is not only large (estimated as a total of 10^{31} particles) but also dynamic, with as many as 10^{24} phage infections per second on a global scale.[3–6] Moreover, with a potentially early origin coinciding with the development of cellularity, phage evolution has likely been ongoing for at least 3 to 4 billion years.[7,8] It is

*Corresponding author. E-mail address: gfh@pitt.edu.

Abbreviations used: pham, phamily; ORF, open-reading frame; ANI, average nucleotide identities.

0022-2836/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.

Table 1. Genometrics of 60 sequenced mycobacteriophage genomes

Phage	Size (bp)	GC%	No. of ORFs	tRNA no.	tmRNA no.	Percentage coding	Ends	Accession no.	Cluster	Morphotype	Origin	Reference
Bethlehem	52,250	63.3	87	0	0	92.80	10-base 3′	AY500153	A1	Siphovirus	Bethlehem, PA	Hatfull *et al.*[16]
Bxb1	50,550	63.7	86	0	0	91.90	9-base 3′	AF271693	A1	Siphovirus	Bronx, NY	Mediavilla *et al.*[39]
DD5	51,621	63.4	87	0	0	94.04	10-base 3′	EU744252	A1	Siphovirus	Upper St. Clair, PA	This work
Jasper	50,968	63.7	94	0	0	94.86	10-base 3′	EU744251	A1	Siphovirus	Lexington, MA	This work
KBG	53,572	63.6	89	0	0	93.31	10-base 3′	EU744248	A1	Siphovirus	Kentucky	This work
Lockley	51,478	63.4	90	0	0	92.74	10-base 3′	EU744249	A1	Siphovirus	Pittsburgh, PA	This work
Solon	49,487	63.8	86	0	0	94.58	10-base 3′	EU826470	A1	Siphovirus	Solon, IA	This work
U2	51,277	63.7	81	0	0	90.11	10-base 3′	AY500152	A1	Siphovirus	Bethlehem, PA	Hatfull *et al.*[16]
Bxz2	50,913	64.2	86	3	0	91.73	10-base 3′	AY129332	A2	Siphovirus	Bronx, NY	Pedulla *et al.*[2]
Che12	52,047	62.9	98	3	0	94.11	10-base 3′	DQ398043	A2	Siphovirus	Chennai, India	Hatfull *et al.*[16]
D29	49,136	63.5	77	5	0	91.79	9-base 3′	AF022214	A2	Siphovirus	California	Hatfull *et al.*[16] Ford *et al.*[40]
L5	52,297	62.3	85	3	0	88.39	9-base 3′	Z18946	A2	Siphovirus	Japan	Hatfull *et al.*[42]
Pukovnik	52,892	63.3	88	1	0	93.11	10-base 3′	EU744250	A2	Siphovirus	Fort Bragg, NC	This work
Chah	68,450	66.5	104	0	0	94.99	Circ Perm	FJ174694	B1	Siphovirus	Ruffsdale, PA	This work
Orion	68,427	66.5	100	0	0	94.34	Circ Perm	DQ398046	B1	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16] Hatfull *et al.*[16]
PG1	68,999	66.5	100	0	0	94.55	Circ Perm	AF547430	B1	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Qyrzula	67,188	69.0	81	0	0	92.33	Circ Perm	DQ398048	B2	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16] Hatfull *et al.*[16]
Rosebush	67,480	69.0	90	0	0	95.55	Circ Perm	AY129334	B2	Siphovirus	Latrobe, PA	Pedulla *et al.*[2]
Phaedrus	68,090	67.6	98	0	0	94.65	Circ Perm	EU816589	B3	Siphovirus	Pittsburgh, PA	This work
Pipefish	69,059	67.3	102	0	0	95.66	Circ Perm	DQ398049	B3	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16] Hatfull *et al.*[16]
Cooper	70,654	69.1	99	0	0	96.44	Circ Perm	DQ398044	B4	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16] Hatfull *et al.*[16]
Nigel	69,904	68.3	94	1	0	96.48	Circ Perm	EU770221	B4	Siphovirus	Pittsburgh, PA	This work
Bxz1	156,102	64.8	225	35	1	92.00	Circ Perm	AY129337	C1	Myovirus	Bronx, NY	Pedulla *et al.*[2]
Cali	155,372	64.7	222	35	1	93.19	Circ Perm	EU826471	C1	Myovirus	Santa Clara, CA	This work
Catera	153,766	64.7	218	35	1	92.22	Circ Perm	DQ398053	C1	Myovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Rizal	153,894	64.7	220	35	1	93.52	Circ Perm	EU826467	C1	Myovirus	Pittsburgh, PA	This work
ScottMcG	154,017	64.8	221	35	1	93.01	Circ Perm	EU826469	C1	Myovirus	Pittsburgh, PA	This work

Spud	154,906	64.8	222	35	1	93.19	Circ Perm	EU826468	C1	Myovirus	Pittsburgh, PA	This work
Myrna	164,602	65.4	229	41	0	94.90	Circ Perm	EU826466	C2	Myovirus	Upper St. Clair, PA	This work
Adjutor	64,511	59.7	86	0	0	96.03	Circ Perm	EU676000	D	Siphovirus	Pittsburgh, PA	This work
Butterscotch	64,562	59.7	86	0	0	96.30	Circ Perm	FJ168660	D	Siphovirus	Pittsburgh, PA	This work
Gumball	64,807	59.6	88	0	0	96.11	Circ Perm	FJ168661	D	Siphovirus	Pittsburgh, PA	This work
P-lot	64,787	59.7	89	0	0	96.38	Circ Perm	DQ398051	D	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
PBI1	64,494	59.7	81	0	0	93.73	Circ Perm	DQ398047	D	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Troll4	64,618	59.6	88	0	0	90.92	Circ Perm	FJ168662	D	Siphovirus	Silver Springs, MD	This work
244	74,483	62.9	142	2	0	95.03	9-base 3′	DQ398041	E	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Cjw1	75,931	63.1	141	2	0	94.28	9-base 3′	AY129331	E	Siphovirus	Pittsburgh, PA	Pedulla *et al.*[2]
Kostya	75,811	62.9	143	2	0	92.74	9-base 3′	EU816591	E	Siphovirus	Washington, DC	This work
Porky	76,312	62.8	147	2	0	93.51	9-base 3′	EU816588	E	Siphovirus	Concord, MA	This work
Boomer	58,037	61.1	105	0	0	94.37	10-base 3′	EU816590	F1	Siphovirus	Pittsburgh, PA	This work
Che8	59,471	61.3	112	0	0	95.59	10-base 3′	AY129330	F1	Siphovirus	Chennai, India	Pedulla *et al.*[2]
Fruitloop	58,471	61.8	102	0	0	92.68	10-base 3′	FJ174690	F1	Siphovirus	Latrobe, PA	This work
Llij	56,852	61.5	100	0	0	95.46	10-base 3′	DQ398045	F1	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Pacc40	58,554	61.3	101	0	0	96.82	10-base 3′	FJ174692	F1	Siphovirus	Pittsburgh, PA	This work
PMC	56,692	61.4	104	0	0	94.60	10-base 3′	DQ398050	F1	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Ramsey	58,578	61.2	108	0	0	96.79	10-base 3′	FJ174693	F1	Siphovirus	White Bear, MN	This work
Tweety	58,692	61.7	109	0	0	95.78	10-base 3′	EF536069	F1	Siphovirus	Pittsburgh, PA	Pham *et al.*[25]
Che9d	56,276	60.9	111	0	0	95.41	10-base 3′	AY129336	F2	Siphovirus	Chennai, India	Pedulla *et al.*[2]
BPs	41,901	66.6	63	0	0	98.74	11-base 3′	EU568876	G	Siphovirus	Pittsburgh, PA	Sampson *et al.*[38]
Halo	42,289	66.7	64	0	0	99.56	11-base 3′	DQ398042	G	Siphovirus	Pittsburgh, PA	Hatfull *et al.*[16]
Konstantine	68,952	57.3	95	0	0	92.26	Circ Perm	FJ174691	H1	Siphovirus	Pittsburgh, PA	This work
Predator	70,110	56.3	92	0	0	91.84	Circ Perm	EU770222	H1	Siphovirus	Donegal, PA	This work
Barnyard	70,797	57.3	109	0	0	94.97	Circ Perm	AY129339	H2	Siphovirus	Latrobe, PA	Pedulla *et al.*[2]
Brujita	47,057	66.8	74	0	0	96.61	11-base 3′	FJ168659	I	Siphovirus	Virginia	This work
Che9c	57,050	65.4	84	0	0	94.75	10-base 3′	AY129333	I	Siphovirus	Chennai, India	Pedulla *et al.*[2]
Corndog	69,777	65.4	99	0	0	95.00	4-base 3′	AY129335	Single	Siphovirus	Pittsburgh, PA	Pedulla *et al.*[2]
Giles	53,746	67.5	78	0	0	94.45	14-base 3′	EU203571	Single	Siphovirus	Pittsburgh, PA	Morris *et al.*[24]
Omega	110,865	61.4	237	2	0	94.68	4-base 3′	AY129338	Single	Siphovirus	Upper St. Clair, PA	Pedulla *et al.*[2]
TM4	52,797	68.1	89	0	0	92.64	10-base 3′	AF068845	Single	Siphovirus	Colorado	Ford *et al.*[41]
Wildcat	78,296	56.9	148	24	1	92.19	11-base 3′	DQ398052	Single	Siphovirus	Latrobe, PA	Hatfull *et al.*[16]
Total	4,354,974		6834	301								
Average	72,582.9	63.4	113.9	5.02		94.18						

therefore perhaps not surprising that analysis of the approximately 600 sequenced bacteriophage genomes reveals that they have unusually high genetic diversity.[5,9] The majority of bacteriophages are double-stranded DNA tailed phages with genomes varying in size from ~15 to ~500 kbp.[10,11]

Bacteriophages exhibit specificity for their bacterial hosts, although host ranges vary enormously—from phages that infect only specific strains within a species to those that infect bacteria of different genera, although usually phylogenetically similar ones. Phages infecting distantly related bacterial hosts typically share little or no nucleotide sequence similarity, suggesting that host constraints present barriers to genetic exchange.[12] Nonetheless, despite a lack of any evident nucleotide sequence similarity, phages may encode protein products with significant amino acid sequence similarities, reflecting old but common origins.[4,12] Because genes or groups of genes often have distinctly different phylogenetic relationships, phage genomes are typically architecturally mosaic, with each genome corresponding to a unique combination of exchangeable modules.[2,12] However, the total number of such modules, the number of possible arrangements, and the factors constraining their exchange remain unclear.

Genome comparisons show that module boundaries commonly correlate with gene boundaries, and sometimes with domain boundaries of the encoded proteins.[12] While recombination could be targeted to gene borders *via* short, shared boundary sequences,[13,14] the majority of exchange events probably occur by illegitimate recombination events using little or no sequence similarity.[1,12] In this second model, any correspondence of exchange boundaries with gene extremities would result from selection for gene function, with most genetic exchange events generating genomic trash.[1] A role for lambda Red-like recombination systems has been proposed for mediating exchange events between diverse sequences to contribute to mosaic architectures.[15] Because illegitimate recombination is more likely to occur between phage genomes and the much larger bacterial genomes, phages acquire and transmit host genes and play major roles in the evolution of their bacterial hosts.[2,16,17]

An additional view of phage diversity can be obtained by comparative genomic analysis of phages that infect a common bacterial host and therefore have the potential to be in direct genetic interaction with one another. Collections of double-stranded DNA tailed phages infecting *Mycobacteria*,[16] *Pseudomonas*,[18] *Staphylococcus*,[19] dairy bacteria,[20] and enteric bacteria[21] have been described. We previously reported the genomic comparison of 14 mycobacteriophages that can be propagated on *Mycobacterium smegmatis*,[2] as well as an expanded analysis of 30 genomes of mycobacteriophages.[16] Among the initial 14 phages, there was little identifiable similarity at the nucleotide sequence level, except between phages L5 and D29, and, to a lesser extent, Bxz2.[2] Although the collection of 30 genomes showed a high level of genetic diversity, additional groups of genomes with some identifiable nucleotide sequence similarity could be recognized.[16] The putative gene products of these 30 phages were grouped into phamilies (phams) of related sequences, and the genomes were examined by gene content comparison. Six clusters of related genomes were revealed (Clusters A–F), encompassing 21 of the 30 genomes, plus 9 that were singletons.[16] However, this clustering does not display the complete phylogenetic history of these phages because each genome also contains examples of genes that have been exchanged horizontally between differently clustered phages. Lawrence *et al.*[22] have noted the need for a reticulate taxonomic approach that accommodates the pervasive mosaicism, and a graph-based approach has been described to classify phage genomes in a reticulate manner.[23]

Here, we report an expansion of the collection of sequenced mycobacteriophage genomes to a total of 60. Each of the newly sequenced phages was isolated by direct plating of environmental samples on lawns of *M. smegmatis* mc^{2}155, purified, sequenced, annotated, and compared. We present four approaches to assort these 60 genomes into clusters and subclusters according to their relatedness and use these cluster relationships to identify genes that are likely to be in more rapid genetic flux than others—being either more frequently lost from phage genomes or gained from genomes outside of their cluster. These rapid flux genes are unusually small—only about 50% of the length of the average mycobacteriophage gene, suggesting that bacteriophage genes are on average only two-thirds the size of bacterial host genes because of the dominant role that horizontal genetic exchange plays in their evolution plus the propensity for these readily exchanged genes to be small.

Results and Discussion

Mycobacteriophage isolation and sequencing

Using *M. smegmatis* mc^{2}155 as a host, we isolated new mycobacteriophages by direct plating of environmental samples (soil, compost, mulch, etc.) on bacterial lawns, followed by plaque purification and amplification; samples were from various sources across the United States, although most were from the greater Pittsburgh, PA, region (Table 1). The genomes of 28 of these were sequenced using a shotgun sequencing strategy and automated Sanger sequencing, which, together with the previously described genomes,[16,24,25] raised the total number of completely sequenced mycobacteriophages to 60, the largest collection of phages known to infect a common host and more than 10% of the 554 (as of December 2009) phage genomes deposited in the National Center for Biotechnology Information phage genome database†. Average genome length (72.6 kbp) and GC% content (63.4%) are not

† http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=10239&type=6&name=Phages

significantly different from previous analyses,[2,16] nor are the extremes of variance in length and GC% (genome lengths range from 41,901 to 164,602 bp, and GC% values range from 56.3% to 69.1%) (Table 1). Thirty-five of the 60 genomes have defined genome ends with 3′ single-stranded extensions from 4 to 14 bp, and 25 are circularly permuted and presumed to be terminally redundant (Table 1). Nineteen genomes encode tRNAs ranging from 1 to 41 genes, and 7 of these genomes also encode a tmRNA. The average number of protein-coding genes per genome is 114 (Table 1), and the average gene length is 616 bp, about two-thirds that of mycobacterial genes as reported previously.[16] Genome maps for all 60 genomes were created using the Phamerator program and annotated according to the comparative analyses and functional characterization described in further detail below. The complete genome maps are shown in Fig. S1.

Mycobacteriophage virion morphologies

The virion morphologies of all 60 phages for which genome sequences have been obtained were examined by electron microscopy (Fig. 1; Fig. S2). All 60 are tailed phages belonging to either the Siphoviridae morphotypes (53 examples) or the Myoviridae morphotypes (7 examples); none is a Podoviridae morphotype. Fifty of the siphoviruses contain isometric heads with diameters varying from 55 to 60 nm, while three (Corndog, Che9c, and Brujita) have prolate heads (Fig. S2 and Table S1); all 7 myoviruses have similarly sized isometric heads (85.9 nm in diameter) (Fig. S2 and Table S1). Tail lengths are highly variable, spanning greater than a 2-fold range (from 135 to 350 nm) (Table S1). In general, capsid volumes predicted from transmission electron microscopy correlate with genome size, indicating similar DNA packaging densities.

Assembly of open-reading frames into phams

We previously described the assembly of mycobacteriophage open-reading frames (ORFs) into phams of related sequences.[16] We performed pham assembly of all 6858 putative ORFs encoded by the 60 mycobacteriophage genomes using an automated program, Phamerator (S.G.C., M.W.B., R.W.H., & G. F.H., unpublished results). All ORFs were compared with all other mycobacteriophage ORFs using both ClustalW and BlastP, and any two ORFs with a score of 25% amino acid identity or an *E*-value of 0.0001 or better were grouped into the same pham. This generated a total of 1523 phams, similar to the 1536 reported for 30 mycobacteriophages,[16] and although the numbers and sizes of the phams have not changed significantly with the doubling of the numbers of genomes, the proportion of orphams (the 699 phams containing only a single gene) is somewhat lower (reduced from 50.4% to 46.1%) (see below for further explanation). A complete list of pham assignments and other pham characteristics is included in Table S2.

Automated pham assembly results in the generation of several very large phams primarily due to the modular construction of some ORFs, as noted

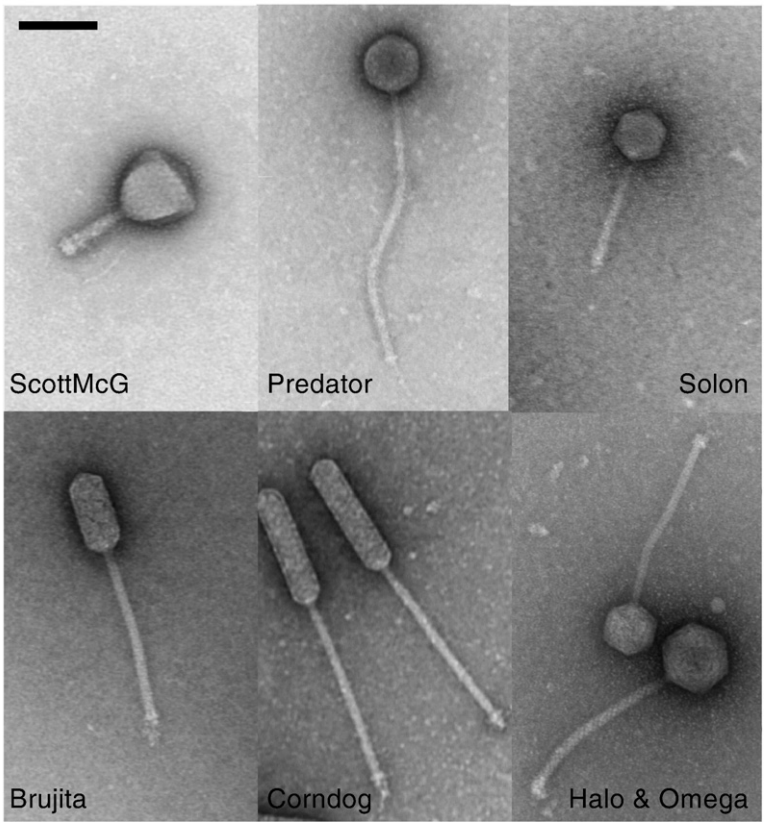


Fig. 1. Mycobacteriophage morphotypes. Representatives of each of the different mycobacteriophage morphotypes are shown. Of the 60 sequenced phages, 7 exhibit myoviral morphotypes with isometric heads (e.g., ScottMcG), and the other 53 all have siphoviral morphotypes. Three of the siphoviruses contain prolate heads, ranging from a length/width ratio of ~2.5:1 (e.g., Brujita) to that of 4:1 (Corndog). Tail lengths also vary by greater than 2-fold, from the Cluster A phages (e.g., Solon, tail shaft average of 113 nm) to the Cluster H phages (e.g., Predator, tail shaft average of 293 nm). Bar corresponds to 100 nm. Morphotypes of all 60 phages are shown in Fig. S2, and virion dimensions are listed in Table S1.

previously,[16] because inclusion of an ORF into a pham requires only that it share significant similarity to one other ORF. Thus, gene A may match gene B and gene B may match gene C but genes A and C may share no direct relationship. This situation arises for phams with genes containing inteins or homing domains, or combinations of domains with distinct evolutionary histories.

Three phams (Pham1406, Pham1410, and Pham1396) have more than 250 members. These were manually deconvoluted into subphams (Pham1406-1, etc.), with each gene assigned to no more than a single subpham (Table S3). The largest pham, Pham1406 (454 members), was deconvoluted into a total of 40 subphams, of which the largest is Pham1406-11, containing 39 members (Table S3). Pham1406 is of special interest since it contains mostly virion structural proteins, including putative tail fibers, whose genetic modularity has been noted previously.[26] However, it also contains capsid and major tail subunit proteins because these contain common C-terminal extensions in some genomes (e.g., Bxb1).[27,28] Pham1410 is the second largest pham (292 members—some of which appear to contain HNH motifs that are expected to have greater mobility throughout the phage population) and was deconvoluted into a total of 62 subphams, of which 20 are orphams, apparently included because of similarity scores close to the threshold level (Table S3). Pham1396 (269 members) was deconvoluted to a total of 86 subphams, some of which (e.g., Pham1396-53) are known to function as

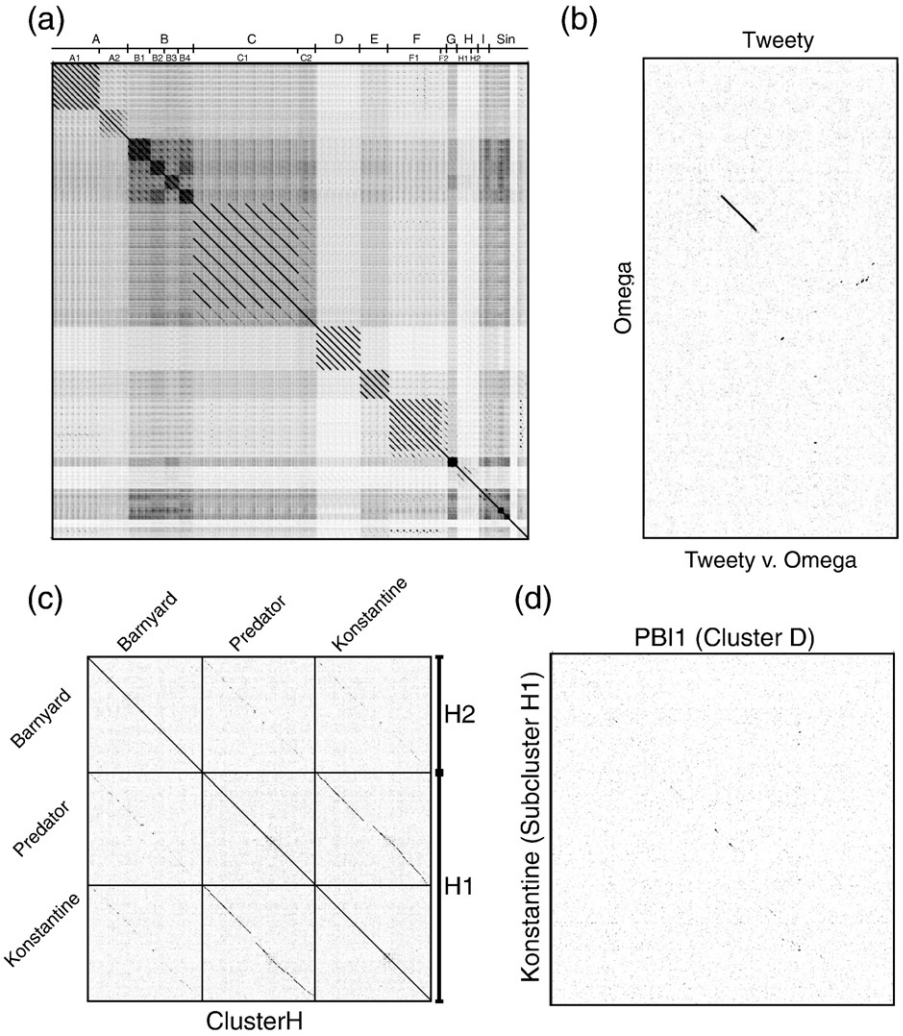


Fig. 2. Nucleotide sequence comparisons of mycobacteriophage genomes. (a) Dot plot of all 60 sequenced mycobacteriophage genomes displayed using Gepard.[68] Individual genome sequences were concatenated into a single sequence arranged such that related genomes were adjacent to each other. The assignment of clusters and subclusters is shown at the top. (b) Dot plot of Omega and Tweety showing segments of ~6.5 kbp that are very similar. Omega and Tweety have not been grouped in the same cluster because the similarity does not span >50% of the genomes. (c) Dot-plot analysis of Cluster H genomes. Predator and Konstantine are more closely related to each other than to Barnyard or other phages and constitute sub-Cluster H1. Barnyard (sub-Cluster H2) is included within the H cluster because its similarity to other Cluster H spans more than 50% of the genome even though its relationship to Konstantine and Predator is weak. (d) Dot plot of Konstantine (sub-Cluster H1) and PBI1 (Cluster D) showing a weak relationship that does not warrant inclusion in the same cluster.

recombination directionality factors for tyrosine integrases (Table S3). We identified nine additional but smaller phams that warranted similar deconvolution (Pham12, Pham13, Pham66, Pham86, Pham1219, Pham1429, Pham1944, Pham2292, and Pham2330), and these are also shown in Table S3. Considering these deconvolutions, the total number of phams and subphams is 1723; the total number of orphams is 773 (44.8% of the total).

Genome clustering

The 60 mycobacteriophage genomes are clearly not uniformly diverse, and we have assorted them into clusters of related genomes. This sorting is not simple, however, because three classes of relationships were observed. The first two cases reflect the extremes of the relationships, where genomes are either very closely related and clearly belong to the same cluster or those for which no relationship is seen and can be considered in different clusters. The third class includes those with more complex relationships, and these fall into three main subclasses: where nucleotide similarity is detected across large genome segments but the relationship is very weak; where short segments of very high levels of sequence similarity are found; and where there is little or no evident nucleotide sequence in common but genomes share a large number of genes encoding proteins with detectably related amino acid sequences.

A primary utility of clustering the genomes is to facilitate identification of genes and modules that have been exchanged between genomes by lateral gene transfer in recent evolutionary time and that contribute to the mosaic architectures of phage genomes. Because of the prominent role of horizontal genetic exchange,[2] this clustering does not represent a phylogenetic or taxonomic grouping but rather provides a framework for reflecting their overall genome relationships and for identifying genes that have been recently exchanged and their genomic context. Clustering does not substitute for a reticulate taxonomy, which more accurately describes the global relationships.[22,23]

We have used four approaches to assort the 60 genomes into clusters according to their relatedness: dot-plot comparison of all genomes with one another, pairwise average nucleotide identities (ANI), pairwise genome map comparisons, and gene content analysis.

Genome clustering: dot-plot analyses

The primary criterion we have chosen for placing two genomes in the same cluster is that they show evident sequence similarity in a dot plot that spans more than 50% of the smaller of the two genomes (Fig. 2a). This generates nine clusters (Clusters A–I) that incorporate 55 of the 60 genomes; five phages (TM4, Wildcat, Giles, Omega, and Corndog) are not closely related to any of the other phage genomes by this comparison and are included in a single category of singleton genomes (Table 2). The dot-plot analysis shown in Fig. 2a reveals that even the relationships within a cluster are often non-uniform, and we have further subdivided five clusters (A, B, C, F, and H) into a total of 12 subclusters; the total number of groupings (clusters and subclusters, including each of the five singleton genomes) is 21 (Table 2), a reflection of the overall high degree of diversity of these phages. The additional analyses below support these cluster/subcluster assignments.

Genome clustering: ANI

The second clustering approach we have used is comparison of ANI (Table 3), and this agrees well with the dot-plot analyses. First, it should be noted

Table 2. Assignment of mycobacteriophage genomes into clusters and subclusters

A	B	C	D	E	F	G	H	I	Singletons
A1 Bxb1 Bethlehem U2 DD5 Jasper KBG Lockley Solon	**B1** Chah Orion PG1 **B2** Rosebush Qyrzula **B3** Pipefish Phaedrus **B4** Cooper Nigel	**C1** Bxz1 Catera Cali Rizal ScottMcG Spud **C2** Myrna	PBI1 Plot Adjutor Butterscotch Gumball Troll4	Cjw1 244 Kostya Porky	**F1** Che8 PMC Llij Boomer Fruitloop Pacc40 Ramsey Tweety **F2** Che9d	Halo BPs	**H1** Predator Konstantine **H2** Barnyard	Che9c Brujita	Corndog Omega TM4 Wildcat Giles
A2 D29 L5 Bxz2 Che12 Pukovnik									

Phages are color-coded according to their order of isolation as follows:
First fourteen sequenced mycobacteriophage genomes
Next sixteen (total 30) sequenced mycobacteriophage genomes
Next thirty (total 60) sequenced mycobacteriophage genomes

that genomes that are not clustered together in the dot-plot analyses exhibit ANI values in the range of 53%–59% (a complete set of all 3600 ANI values is shown in Table S4), with these relatively high values reflecting in part the high GC% content (Table 1). In contrast, intra-subcluster values can

Table 3. ANIs shared by mycobacteriophages

Cluster A

	Bethlehem	Bxb1	DD5	Jasper	KBG	Lockley	U2	Solon	Che12	D29	L5	Pukovnik	Bxz2
Bethlehem	100	89.4	92.6	90.7	93.8	92.4	94.5	94.4	61.9	61.6	61.2	61.5	63.1
Bxb1		100	88.5	86.4	88.9	89.0	89.7	90.7	63.7	63.0	62.8	62.8	64.6
DD5			100	91.8	91.8	91.1	92.5	93.0	61.5	61.4	61.4	61.6	63.1
Jasper				100	88.6	92.2	90.2	90.7	61.9	61.5	61.0	61.8	62.6
KBG					100	91.3	93.4	93.8	62.5	61.5	61.5	61.1	63.7
Lockley						100	92.2	92.5	61.9	61.5	61.5	61.5	62.9
U2							100	93.1	61.8	61.6	62.1	61.6	63.5
Solon								100	62.4	62.0	62.1	61.7	63.3
Che12									100	79.4	81.3	75.1	67.3
D29										100	94.4	76.0	67.8
L5											100	75.6	67.5
Pukovnik												100	67.3
Bxz2													100

Subcluster A1; Yellow. Subcluster A2; Turquoise

Cluster B

	Chah	Orion	PG1	Qyrzula	Rosebush	Phaedrus	Pipefish	Cooper	Nigel
Chah	100	99.7	99.6	62.3	64.3	62.7	62.6	67.2	66.5
Orion		100	**99.8**	62.2	64.3	62.8	62.6	66.8	66.4
PG1			100	62.1	64.3	62.6	62.5	66.7	66.3
Qyrzula				100	**97.9**	65.3	65.5	65.1	64.6
Rosebush					100	64.7	64.7	65.2	64.5
Phaedrus						100	**96.6**	63.1	62.7
Pipefish							100	63.3	62.8
Cooper								100	**87.2**
Nigel									100

Average Nucleotide Identity determined by DNA Master Genome Comparison function.
Subcluster B1: Red. Subcluster B2: Purple. Subcluster B3: Green. Subcluster B4: Yellow

Cluster C

	Bxz1	Cali	Catera	Rizal	ScottMcG	Spud	Myrna
Bxz1	100	98.2	98.2	98.9	98.2	98.8	66.6
Cali		100	98.3	98.8	98.1	98.4	66.6
Catera			100	99.2	99.2	98.9	66.5
Rizal				100	98.7	99.2	66.7
ScottMcG					100	99.2	66.7
Spud						100	66.8
Myrna							100

Subcluster C1: Green. Subcluster C2: Purple

Cluster D

	Adjutor	PBI1	PLot	Butterscotch	Troll4	Gumball
Adjutor	100	99.9	98.2	99.5	98.5	98.1
PBI1		100	98.3	99.5	98.6	98.1
Plot			100	98.4	98.0	97.6
Butterscotch				100	98.7	98.0
Troll4					100	97.6
Gumball						100

Cluster E

	244	Cjw1	Kostya	Porky
244	100	98.7	95.0	98.1
Cjw1		100	95.0	97.9
Kostya			100	95.1
Porky				100

Table 3 (*continued*)

Cluster F

	Boomer	PMC	Llij	Che8	Tweety	Fruitloop	Ramsey	Pacc40	Che9d
Boomer	100	94.2	90.4	85.2	87.1	86.7	91.0	88.6	71.2
PMC		100	97.6	88.3	90.2	89.6	90.4	92.8	71.6
Llij			100	88.0	89.5	89.3	88.9	91.4	68.5
Che8				100	85.4	83.1	86.5	88.0	78.2
Tweety					100	94.1	88.8	88.8	68.0
Fruitloop						100	87.6	87.8	65.9
Ramsey							100	90.0	71.8
Pacc40								100	73.7
Che9d									100

Subcluster F1: Turquoise. Subcluster F2: Yellow

Cluster G

	BPs	Halo
BPs	100	99.0
Halo		100

Cluster H

	Predator	Konstantine	Barnyard
Predator	100	73.4	57.9
Konstantine		100	58.4
Barnyard			100

Subcluster H1: Yellow. Subcluster H2: Red

Cluster I

	Che9c	Brujita
Che9c	100	76.8
Brujita		100

Singletons

	Corndog	Giles	TM4	Wildcat	Omega
Corndog	100	57.1	59.0	55.5	57.0
Giles		100	58.4	53.9	54.7
TM4			100	55.6	56.9
Wildcat				100	53.7
Omega					100

Average Nucleotide Identities determined by DNA Master Genome Comparison function.

be as high as 99.8% (Table 3), although the ANI values vary greatly for different clusters and subclusters.

In Cluster A, eight of the genomes (Bethlehem, Bxb1, DD5, Jasper, KBG, Lockley, U2, and Solon) have pairwise ANI values between 88.5% and 94.5%, and four of the genomes (Che12, D29, L5, and Pukovnik) share ANI values between 75.1% and 94.4%. However, none of the pairwise ANI of genomes across the two groups exceeds 63.7% ANI, and these values thus support the division of Cluster A into at least two subclusters (A1 and A2). The positioning of Bxz2 in sub-Cluster A2 represents a conundrum; it shares higher ANI values for the A2 cluster (67.3%–67.8%) than for the A1 phages (62.6%–64.6%), and all of these are substantially lower than the pairwise ANI values among Che12, D29, L5, and Pukovnik (75.1%–94.4%). Bxz2 could conceivably be placed into a third subcluster, but we have currently placed it within the A2 subcluster, to which it is most closely related. This situation is a good illustration of the somewhat arbitrary nature of this clustering process; it usefully reflects the fact that some genomes are close relatives of others, but clustering is an imperfect process with the boundaries between groups of phages being ill-defined because of the prominent role of horizontal genetic exchange in phage genome evolution.

The use of ANI values also introduces additional complexities. For example, although the singleton Omega genome is not closely related to other mycobacteriophages, it shares a 6.1- to 8.3-kbp segment that is very closely related (95% identical with Tweety) with the Cluster F genomes (Fig. 2b) and clearly represents a relatively recent exchange event. As a consequence, the overall ANI values between Omega and the Cluster F genomes are fairly high (65.6%–74.3%) even though they do not fulfill the criterion of sharing evident sequence similarity spanning 50% of the genomes. Next, although Predator and Konstantine (sub-Cluster H1) are closely related and share 73.4% ANI, Barnyard (sub-Cluster H2) shares only 57.9% and 58.4% ANI

with Predator and Konstantine (sub-Cluster H1), respectively (Fig. 2c). These values are within the range observed for unrelated genomes (Table S4), but the inclusion of Barnyard in Cluster H is justified by dot-plot analysis (Fig. 2c), showing that, although the relationship to Konstantine and Predator is weak, it spans >50% of the genomes. In contrast, while Konstantine has a similar ANI value to Cluster D phages (57.0% with PBI1), the dot-plot relationship is extremely weak (Fig. 2d).

Genome clustering: gene content analysis

A third approach to genome clustering is a gene content analysis based on scoring whether the genomes contain a member of each of the protein phams and representing them using the program Splitstree as described previously[16] (Fig. 3). The resulting pattern is in good agreement with the analyses from dot-plot and ANI comparisons and supports the overall cluster and subcluster groupings (Table 2). The subdivision of Clusters A, B, C, and F is clearly delineated, and the more distant relationships between Clusters D, H1, and H2 described above are further substantiated. We note further that while Bxz1, Spud, Catera, Rizal, Cali, ScottMcG, and Myrna warrant being in the same cluster (C), Myrna (Cluster C2) is a distant relative of the C1 phages.

Genome clustering: pairwise genome analyses

The fourth approach to representing the genome relationships of clusters is pairwise alignment and correlation of regions of genome similarity with gene location (Fig. 4). This is especially useful for displaying segments of similarity between more distantly related genomes, as well as revealing departures among more closely related genomes. For example, the subclustering of the Cluster B genomes is clearly illustrated, and the locations of genome differences within each subcluster are delineated (Fig. 4). In addition, one example of a relatively recent exchange of genes between one subcluster (B1; PG1 genes *33*–*35*) and another (B2; Rosebush genes *33*–*35*) is evident. Overall, these comparisons show the closeness of relationships within Clusters C1, D, E, and G, as well the relatively weaker ones within Clusters F, H, and I. The obvious disadvantage of this approach is that the presentation is limited to pairwise display comparisons, and thus only a subset of the interesting and complex relationships between phages such as in Cluster F can be shown in a single representation (Fig. 4).

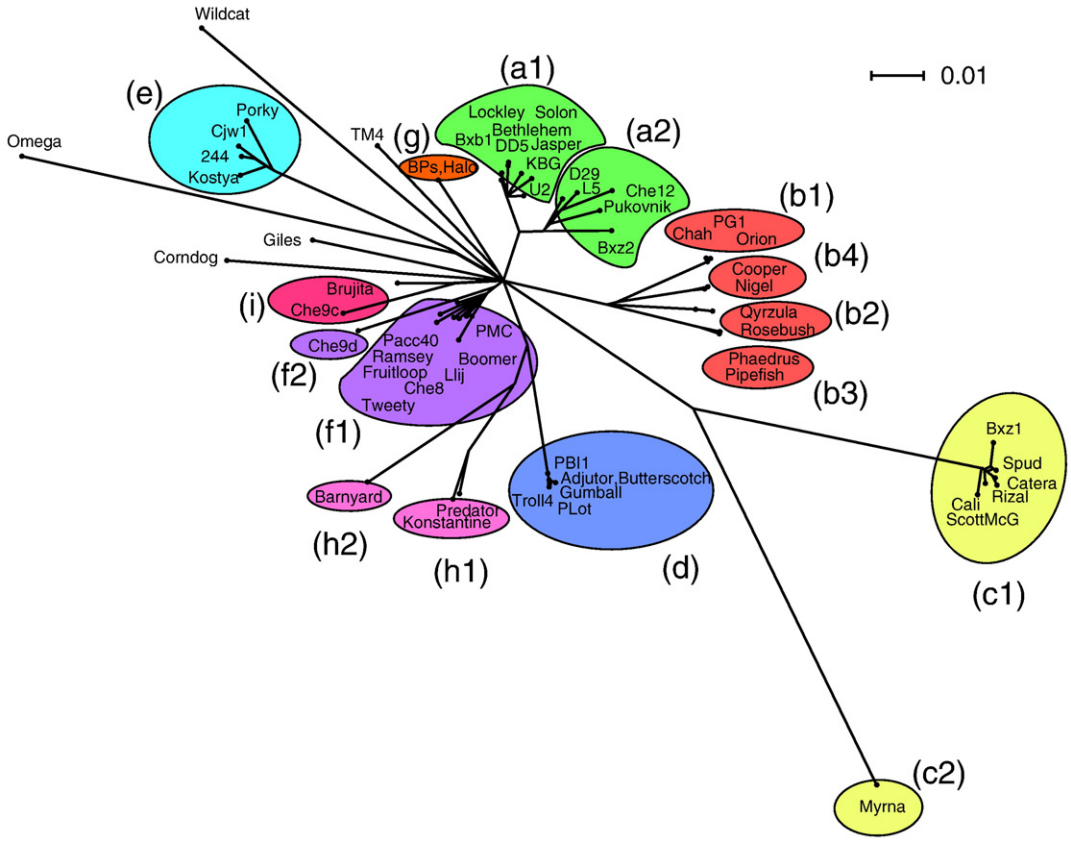


Fig. 3. SplitsTree representation of mycobacteriophage relationships. All 6858 mycobacteriophage predicted protein products were assorted into 1523 phams according to shared sequence similarities. Each genome was then assigned a value reflecting the presence or absence of a pham member, and the genomes were compared and displayed using SplitsTree.[71] The clusters and subclusters derived from dot-plot analyses are annotated. The scale bar indicates 0.01 substitutions/site.

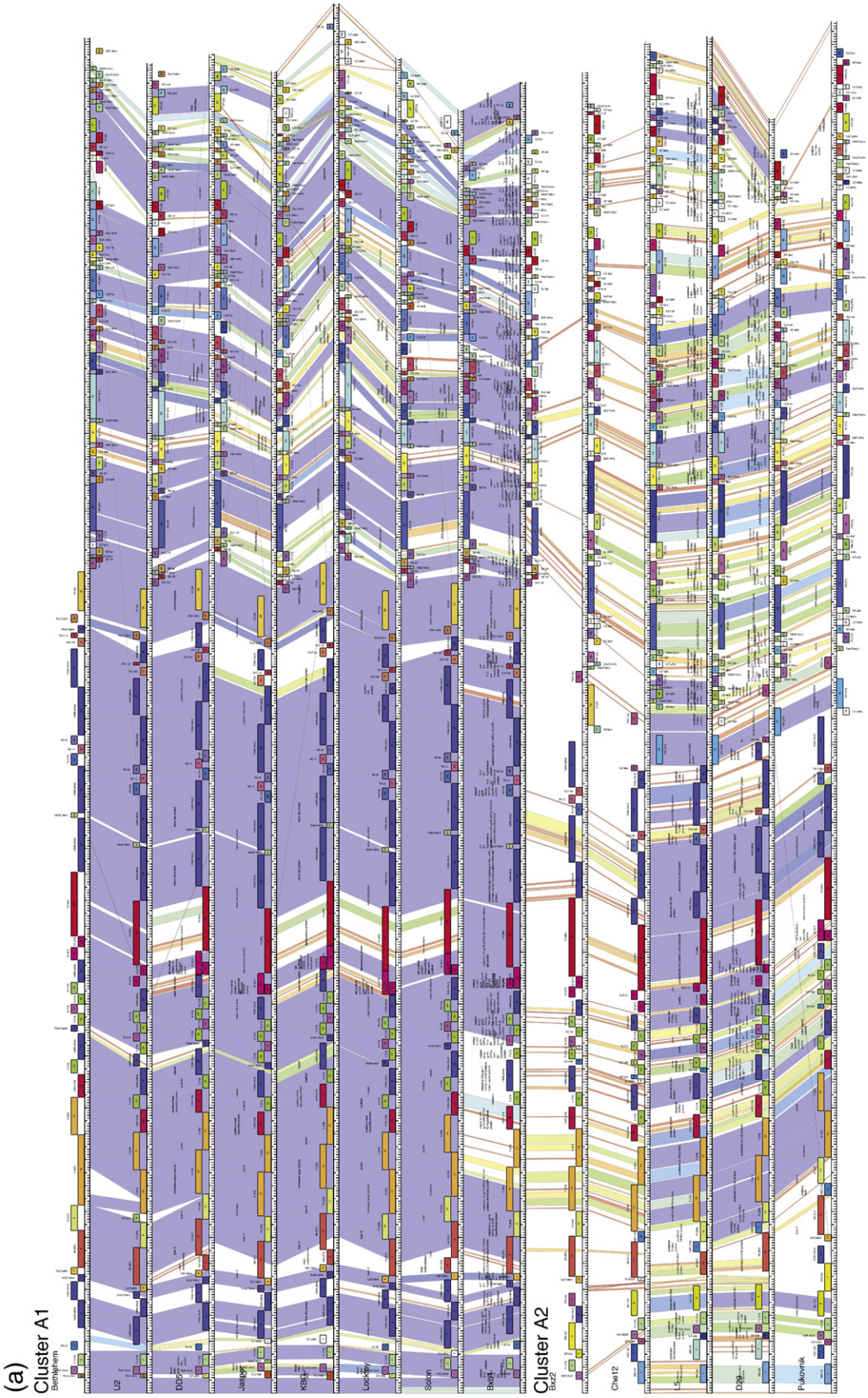


Fig. 4 (*legend on page 133*)

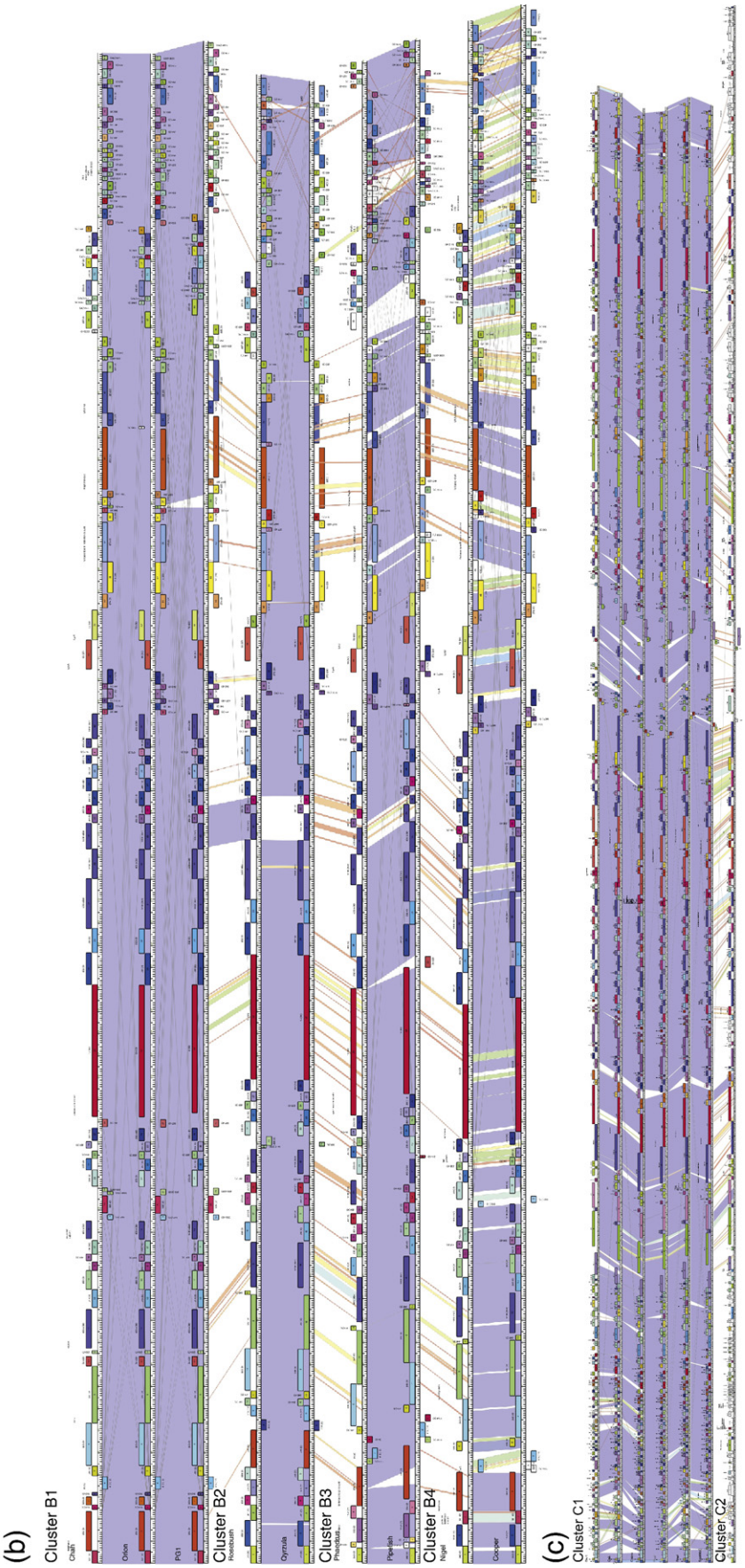


Fig. 4 (*legend on page 133*)

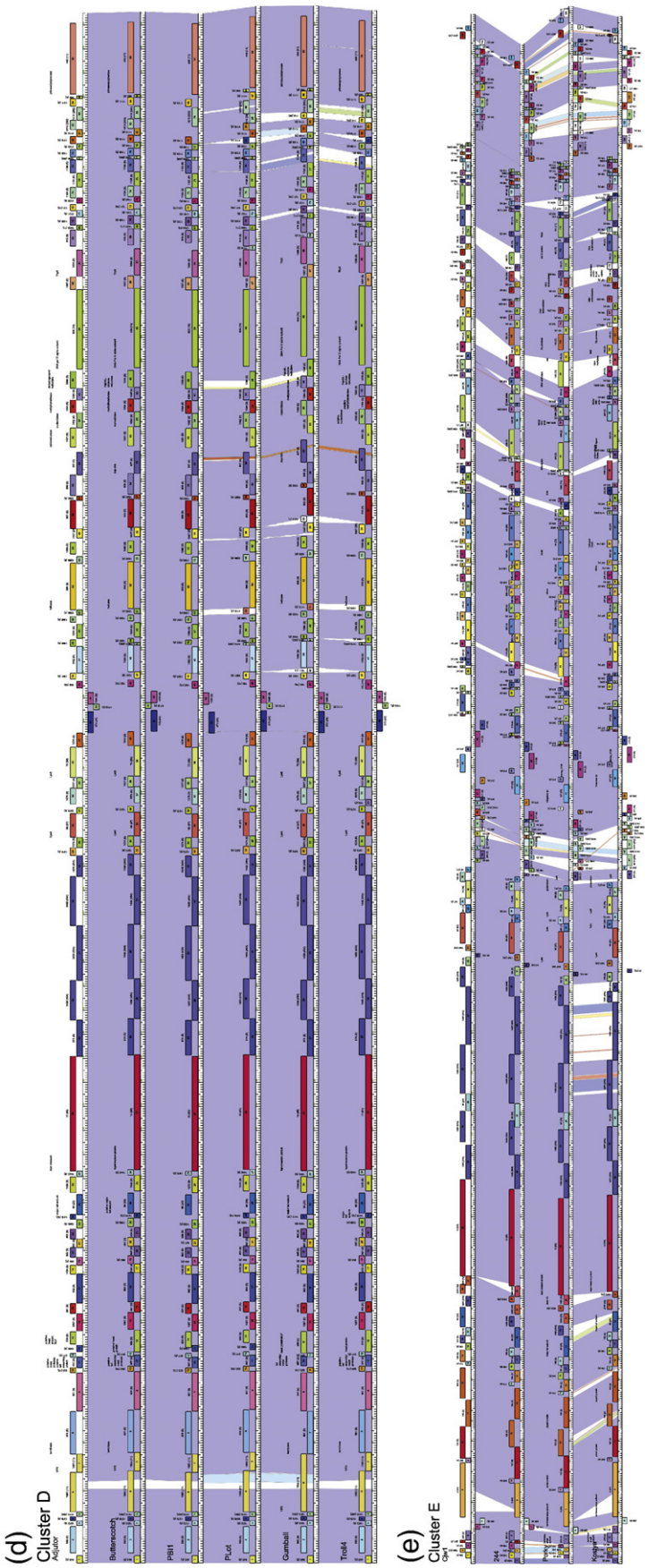


Fig. 4 (*legend on page 133*)

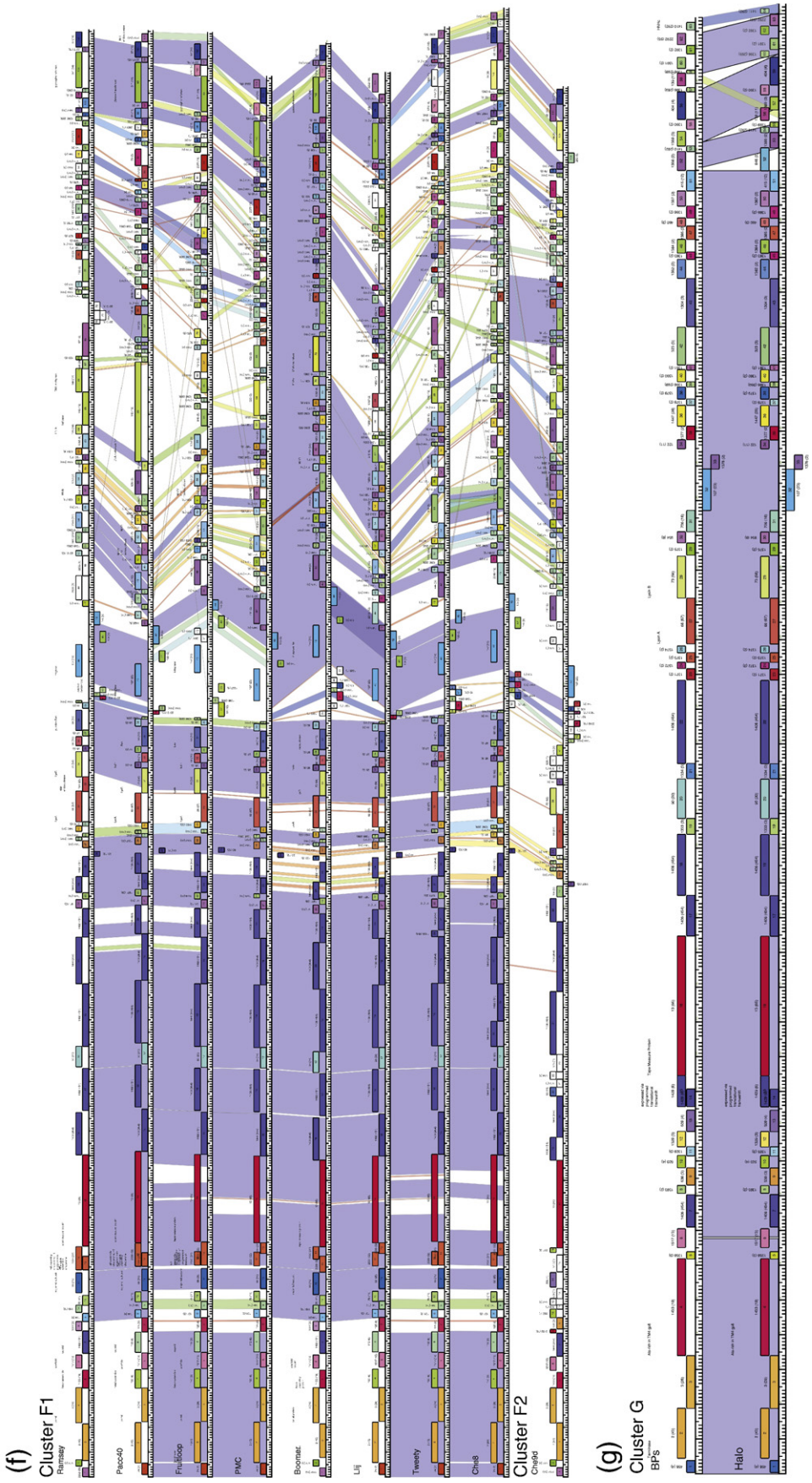


Fig. 4 (*legend on next page*)

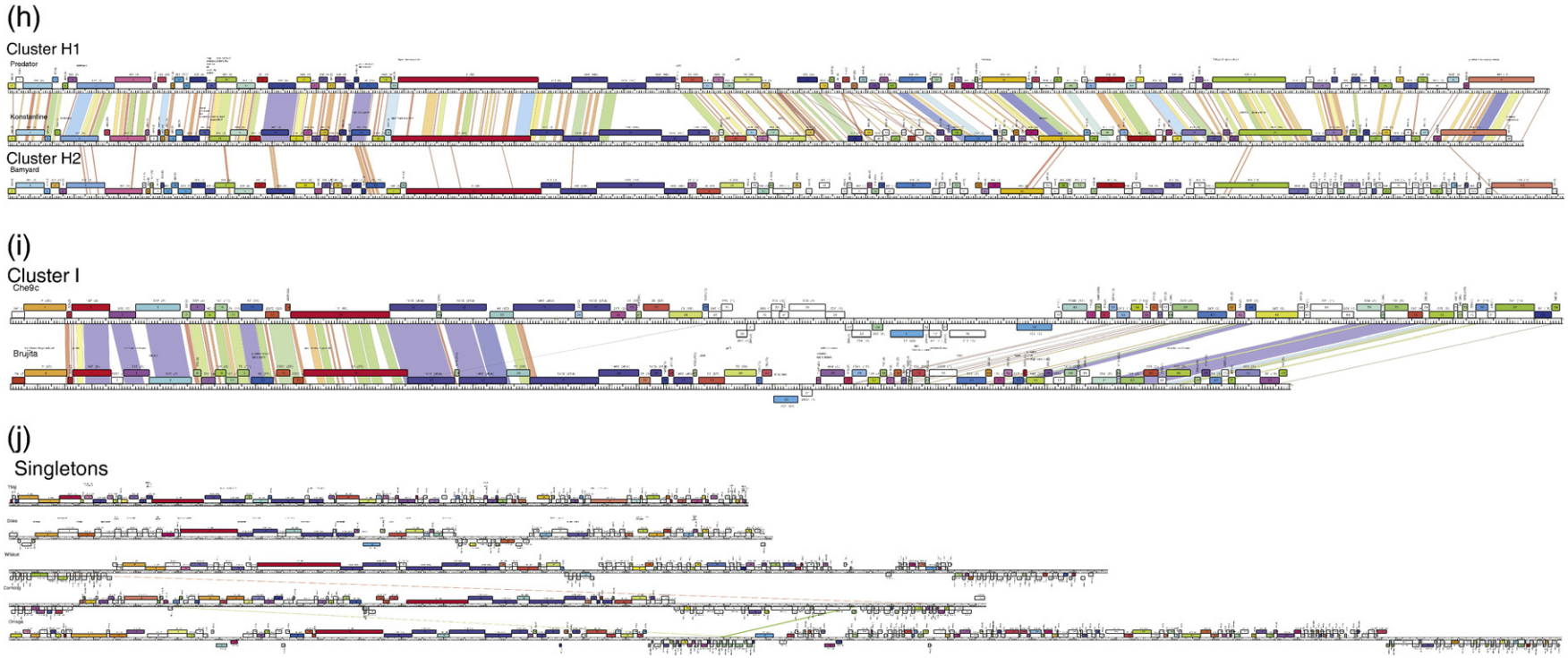


Fig. 4. Pairwise alignment of clustered mycobacteriophage genomes. Each of the mycobacteriophage genome clusters is displayed showing segments of nucleotide sequence similarity between adjacently displayed genomes. The strength of the relationship is represented by shading according to the color spectrum, with purple being the highest. The order of the genomes displayed within each subcluster is as follows: A1: Bethlehem, U2, DD5, Jasper, KBG, Lockley, Solon, Bxb1; A2: Bxz2, Che12, L5, D29, Pukovnik; B1: Chah, Orion, PG1; B2: Rosebush, Qyrzula; B3, Phaedrus, Pipefish; B4: Nigel, Cooper; C1: Bxz1, Cali, Catera, Rizal, Spud, ScottMcG; C2: Myrna; D: Adjutor, Butterscotch, PBI1, Plot, Gumball, Troll4; E: Cjw1, 244, Porky, Kostya; F1: Ramsey, Pacc40, Fruitloop, PMC, Boomer, Llij, Tweety, Che8; F2: Che9d; G: BPs, Halo; H1: Predator, Konstantine; H2: Barnyard; I: Che9c, Brujita; Singletons: TM4, Giles, Wildcat, Corndog, Omega. Detailed maps of individual genomes are shown in Fig. S1.

Revision of prior genome cluster designations

The specific cluster designations closely reflect those reported previously based on the comparative analysis of 30 of these genomes.[16] One notable departure occurs in Cluster F, which previously included only PMC and Llij but did not include Che8 or Che9d. The methods described here all justify inclusion of both Che8 and Che9d in Cluster F, although Che9d is clearly a more distant relative than the other Cluster F phages, justifying its placement into a separate subcluster (F2) (Fig. 4; Table 3). Cluster F is one of the more diverse groups in this set, and the combination of methods described here offers greater reliability in the cluster assignments.

Seven other phages (Barnyard, Corndog, Che9c, Halo, Omega, TM4, and Wildcat) were previously classified as singletons, and Corndog, Omega, TM4, and Wildcat remain in this category even though the number of sequenced genomes has doubled. Clustering relatives of Barnyard (Konstantine, Predator)

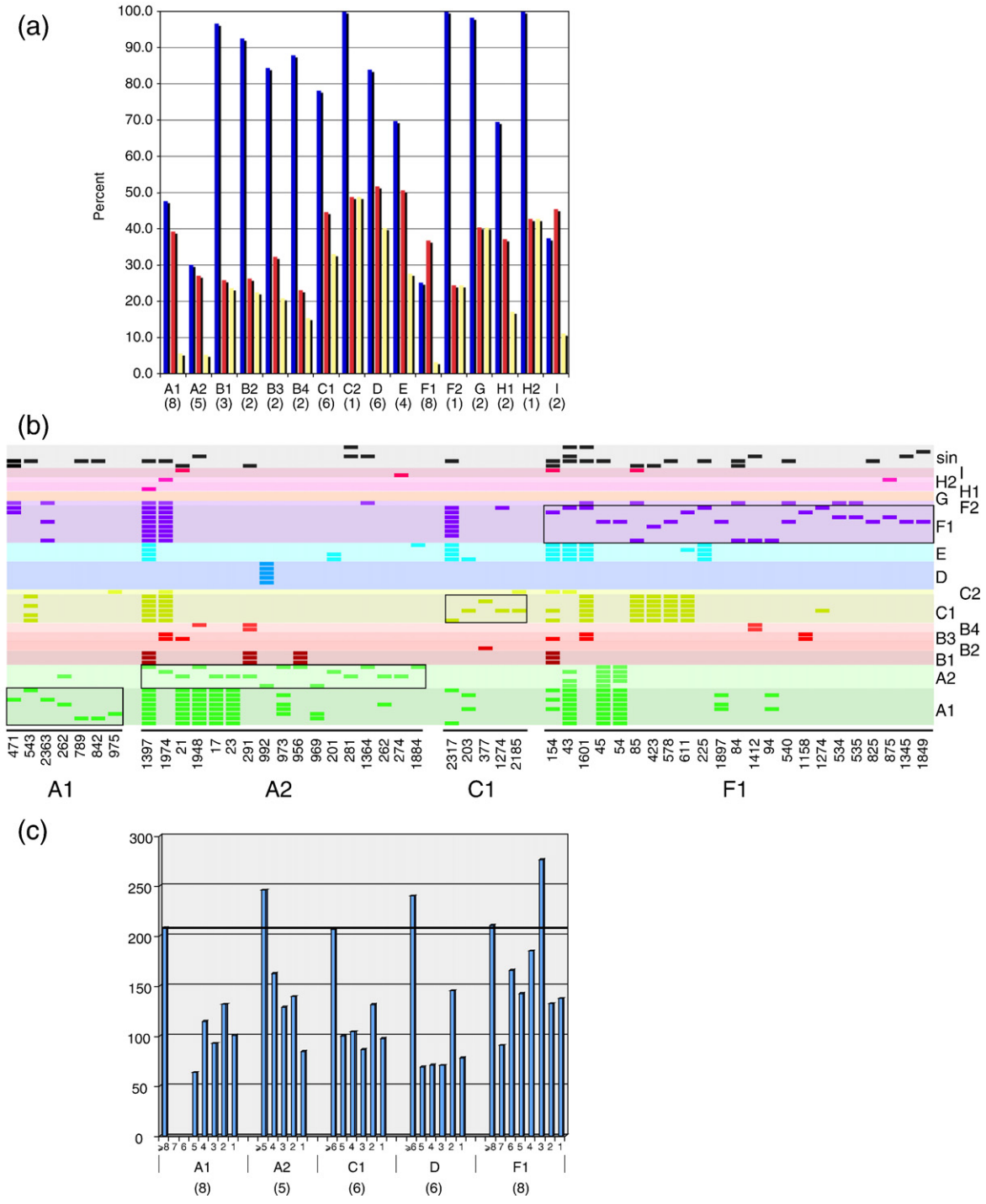


Fig. 5 (*legend on next page*)

and Che9c (Brujita) have now been identified, and we presume that increased sampling will eventually identify relatives of all the currently singleton genomes. A notation of the representation of cluster members at different stages in their discovery is summarized in Table 2.

Cluster diversity

The pairwise ANI values show that some clusters are much more diverse than others. To quantify the extent of diversity within clusters and subclusters, we have determined the proportion of cluster-universal phams—those phams represented in a cluster/subcluster that are present in all constituent genomes in that cluster/subcluster (blue bars in Fig. 5a). This value ranges from Cluster G, in which 98% of the phams are present in both of the constituent genomes, to the F1 subcluster, in which only 25% of the phams are shared (Fig. 5a). Clusters/sub-Clusters A1, I, A2, and F1 each shares fewer than 50% of all phams (48%, 38%, 30%, and 25%, respectively), while B1, B2, B4, D, B3, C1, E, and H1 all have values greater than 50% (97%, 92%, 88%, 84%, 84%, 78%, 70%, and 69%, respectively) (Fig. 5a). The rank ordering of the clusters/subclusters (containing two or more genomes, from least to most diverse) is thus G<B1<B2<B4<B3<D<C1<E<H2<A1<I<A2<F1.

The pairwise comparison maps (Fig. 4) provide information about the distribution of intracluster diversity within the genomes. In the highly diverse Cluster F1, for example, it is evident that the diversity is not uniform throughout the genomes, with the leftmost regions containing the virion structure and assembly genes being better conserved than the rightmost genomic segments (Fig. 4). This reflects previous studies describing strong conservation of the order of virion structure and assembly genes in Siphoviridae and the paucity of horizontal exchange within the groups of structural genes. This has been ascribed to their co-evolution resulting from close interaction of the protein products.[29] In contrast, the units contributing to the mosaic architecture of the non-structural genes are commonly just single genes.[2] Other examples for which the virion structure and assembly genes are more highly conserved than the non-structural genes are provided in Clusters/sub-Clusters A1, A2, B4, and I. It is noteworthy that the extent of cluster diversity does not simply reflect the number of genomes present. For example, the six genomes in Cluster D share 84% of the total phams, while the five genomes of Cluster A2 share only 30% (Fig. 5a).

Intercluster relationships

An alternative perspective on cluster relationships is provided by considering not just which phams are common to all cluster members but also the proportion of cluster-unique phams—those present only within one cluster or subcluster and not present in other mycobacteriophages. For example, in Clusters D and E—both of which have modest diversity levels (83% and 70% of phams present in all genome constituents, respectively), 50% of the total phams represented are cluster unique and not found in other mycobacteriophage genomes (red bars in Fig. 5a). In general, all of the cluster/subcluster groups contain a minimum of 20% of total phams that are specific to that cluster/subcluster, and no obvious relationship appears between cluster diversity and the proportion of cluster-specific phams.

In each cluster/subcluster, there are subsets of phams that are cluster identifiers—those phams that are present in all genomes within a cluster and are not found elsewhere (yellow bars in Fig. 5a). In some subclusters, such as A1, A2, and F1, these are quite rare (<6%), in part reflecting the relatively high diversity of those subclusters. In contrast, 40% of the

Fig. 5. Cluster diversity and intercluster relationships. (a) Distribution of cluster-universal, cluster-unique, and cluster-identifier phams. Cluster-universal phams (blue bars) are defined as those that are present within all genome members within a cluster or subcluster (as shown below the *x*-axis with the number of genomes), and their proportion of the total number of phams in that cluster or subcluster is shown as a percentage. Cluster-unique phams (red bars) are defined as those that are present within that cluster or subcluster and are not present in other mycobacteriophages, and their proportion of the total number of phams in that cluster or subcluster is shown as a percentage. Cluster-identifier phams (yellow bars) are defined as those that are found in all genomes within a cluster or subcluster but absent from all other mycobacteriophages. (b) Some phams are present in only one genome within a cluster/subcluster, and these are candidates for being acquired relatively recently by horizontal genetic exchange. A subset of these have one or more relatives in other cluster/subcluster genomes as illustrated for the four subclusters (A1, A2, C1, and F1) that contain at least five genome members (see Table 2). Along the *x*-axis, each of the phams (grouped by the subcluster containing just the single member) is shown, with bars above indicating which other genomes contain homologues and to which cluster they belong. The 60 genomes are listed vertically and arranged into clusters as shown on the right. The locations of the relatives of these putative newly acquired genes are distributed among the mycobacteriophage genomes, suggesting that they have been acquired from multiple sources and not from any single prominent genome cluster. It is noteworthy that no relatives are seen in Cluster G and that Cluster D only has relatives for the Pham992 member present in one A2 cluster member (D29). Gene members of each Pham and their specific genome and cluster locations are listed in Table S2. (c) Average protein size of phams distributed in different numbers of genomes within clusters/subclusters A1, A2, C1, D, and F1. For each pham, the average protein length (in amino acid residues) is plotted as a function of how many genomes the pham is present in. The total number of genomes within each cluster/subcluster is shown in parentheses. The average length of all mycobacteriophage predicted proteins is shown by the horizontal bar. Note that phams present in only a subset of the cluster genomes are substantially smaller, with the exception of one category in Cluster F1. However, there is only a single gene member in this category.

Table 4. Pham assignments of virion structure and assembly genes in mycobacteriophage clusters

Cluster	Small terminase	Large terminase	Portal	Protease	Scaffold	Capsid	Major tail subunit	'G'	'T'	Tmp
Cluster A		Pham2	Pham3	Pham1433	Pham5	Pham1406-1	Pham1406-2	Pham12-1	Pham12-2	Pham13
Cluster B	Pham931	Pham394	Pham347	?		Pham1406-18	Pham1406-19	?	?	Pham13
Cluster C1										Pham13
Cluster C2									Pham511	Pham13
Cluster D		Pham891	Pham901	Pham857		Pham862?	Pham86-1	Pham1439	Pham1440	Pham13
Cluster E	Pham209	Pham2	Pham147	?		Pham164	Pham86-2	Pham2330-1	Pham2330-2	Pham13
Cluster F1										
B&R[a]	Pham512	Pham2	Pham3	Pham1433	Pham1517	Pham1406-1	Pham86-2	Pham2330-1	Pham2330-3	Pham13
Others[b]	Pham74	Pham2	Pham3	Pham100	Pham1517	Pham118	Pham86-2	Pham2330-1	Pham2330-3	Pham13
Cluster F2	Pham512	Pham2	Pham3	Pham1433	Pham1517	Pham1406-1	Pham509	Pham510	Pham511	Pham13
Cluster G	Pham456	Pham2	Pham3	Pham1433	Pham1517	Pham1406-1	Pham509	Pham1429-1	Pham1429-2	Pham13
Cluster H		Pham891	Pham901	Pham857		Pham862?	Pham86-1	Pham1439	Pham1440	Pham13
Cluster I										
Cluster I - Che9c	Pham294	Pham2	Pham147	Pham315		Pham322	Pham86-2	Pham2330-1	Pham2330-5	Pham13
Cluster I - Brujita	Pham74	Pham2	Pham147	Pham315		Pham322	Pham86-2	Pham2330-1	Pham2330-4	Pham13
Corndog	Pham456	Pham2	Pham457	Pham459		Pham462	Pham86-2	Pham2330-1	Pham2330-3	Pham13
Giles		Pham2	Pham1311	Pham150		Pham1326	?		Pham12-2	Pham13
TM4	Pham456	Pham2	Pham3	Pham1433	Pham1517	Pham1406-1	Pham509	Pham1429-1	Pham1429-2	Pham13
Omega		Pham2	Pham457	Pham315		Pham748	Pham86-2	Pham2330-1	Pham2330-3	Pham13
Wildcat		Pham2	Pham3	Pham1216	Pham1517	Pham1219-1	Pham1219-2	Pham1224?	Pham1225?	Pham13

Tmp indicates tape-measure protein; 'G' and 'T' refer to putative analogues of phage lambda gpG and gpT, respectively, that are expressed *via* a programmed translational frameshift. ?, unknown.
[a] Mycobacteriophages Boomer and Ramsey.
[b] Cluster F1 phages other than Boomer and Ramsey.

total phams in Clusters D and G are cluster identifiers.

Identification and characterization of newly acquired genes

A principal feature of genome clustering described above is that it facilitates the identification and characterization of those genes that are most likely to have been exchanged horizontally in their relatively recent evolutionary history. While each cluster contains a number of phams that are present in all cluster members as discussed above (Fig. 5a), the remaining phams are present in only a subset of the genomes. These correspond to genes that are in greatest evolutionary flux. The lack of full representation could result from loss of a gene from one or more genomes or alternatively from recent acquisition by horizontal genetic exchange. While both explanations could account for phams that are present in only one genome, these are strong candidates for recent acquisition events.

A subset of the phams that are present in only a single member genome of a cluster/subcluster also have one or more pham members in other clusters/subclusters. We have examined these more closely to explore whether there are patterns of exchange that might reveal the origins of these genes (Fig. 5b). Interestingly, the genomes and clusters containing related pham members are broadly represented, supporting the idea that all of these genomes have been in genetic communication, albeit in more distant evolutionary history. For example, the 23 phams represented by a single gene within Cluster F1 have relatives within most of the other clusters/subclusters and singleton genomes (Fig. 5b). However, we note that although about one-half of the 17 phams represented by a single gene in sub-Cluster A2 are also found in sub-Cluster A1, only 1 of the 7 phams of this type present in one of the A1 genomes is also present in the A2 cluster (Fig. 5b). Furthermore, Clusters G and D are notably underrepresented in that there are no relatives of any of the genes in this classification in the Cluster G genomes and Cluster D contains relatives of just one of the phams, Pham992 (Fig. 5b). Genomes in these clusters might thus enjoy a higher degree of isolation than other mycobacteriophages, perhaps as a result of host specificity or geographical or environmental influences.

It was reported previously that genes within the SPO1 family of phage genomes that are not related to other members are in general smaller than those that are, with the implication that genes that are moving between genomes on a rapid time scale are small.[30] The clustering of mycobacteriophage genomes enables us to extend this type of analysis to multiple genome sets. Specifically, we have grouped phams within clusters/subclusters according to their extent of representation within their specific cluster/subcluster (i.e., whether they are present in all or just a subset of genomes within that cluster/subcluster) and determined their average lengths (Fig. 5c). This analysis shows that phams that are present in all members of a cluster/subcluster are at or close to the average number of codons for all 1523 genes (205 codons) but that phams represented in a

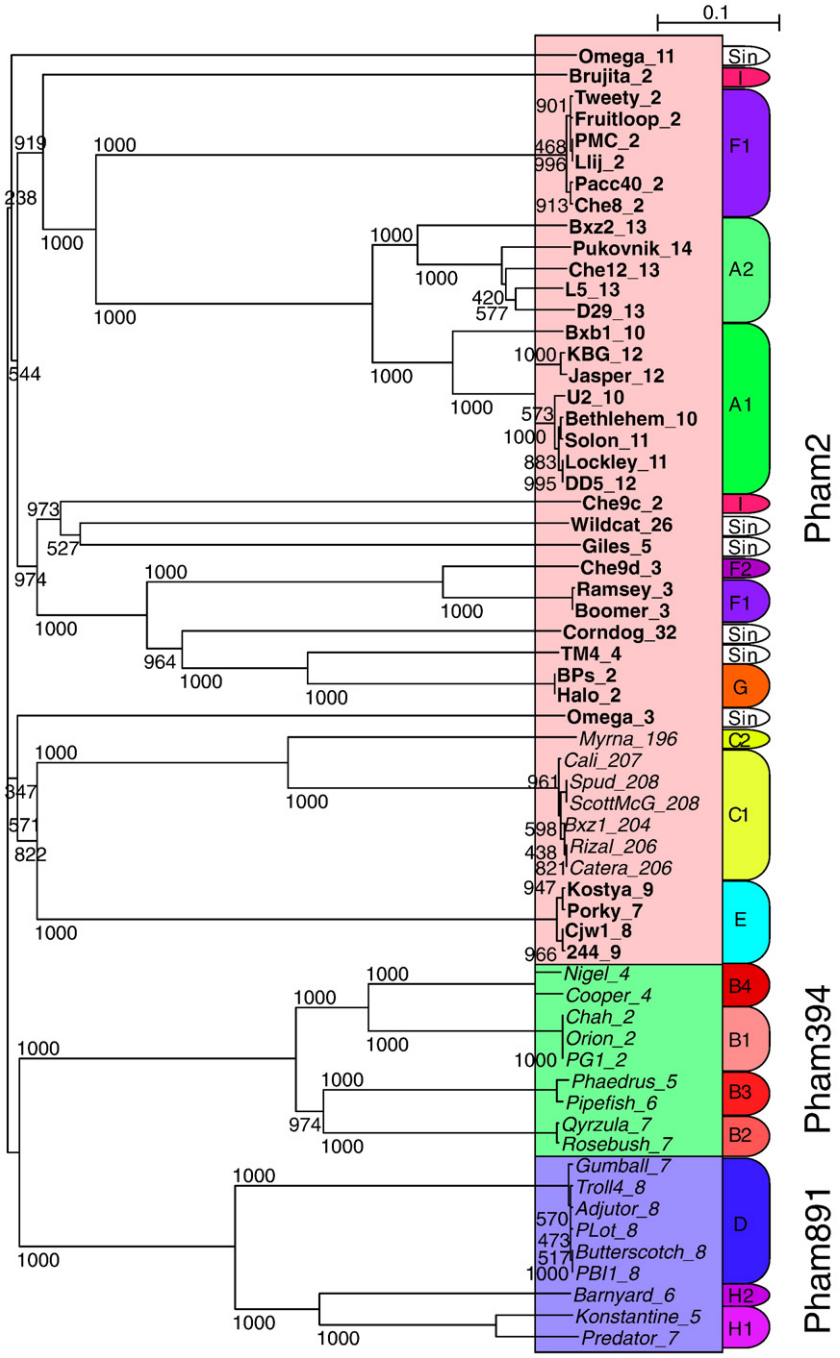


Fig. 6. Phylogenetic relationships of mycobacteriophage terminases. The protein sequences of all members of Pham2, Pham394, and Pham891 were aligned using ClustalX,[72] and the tree is represented by Njplot.[73] The members of Pham2, Pham394, and Pham891 are shown in red, green, and blue boxes, respectively. Cluster designations of individual genomes are shown on the right; singleton phages are notated as Sin. Phage genes corresponding to genomes with defined cohesive termini are shown in bold type, and those with terminally redundant ends are shown in italic type. Note that the cluster C phages are only included in Pham2 because of the presence of an intein that is related to inteins in other Pham2 members. Bootstrap values were derived from 1000 iterations. Scale bar represents the estimated number of changes per site.

subset of the cluster genomes are substantially smaller (Fig. 5c). This is certainly true of phams represented in only a single genome within a cluster and thus more likely to have been acquired by recent horizontal genetic exchange, but this is also observed for all phams not present in all cluster genomes. This is consistent with the hypothesis that all genes active in genetic flux—both loss from a genome and acquisition—are smaller than the average gene. Furthermore, the size differences are substantial, with the sub-represented phams being generally at least 25% smaller than the average of those represented in all cluster/subcluster members (Fig. 5c).

Genomic architectural features of mycobacteriophage clusters

The grouping of genomes into clusters enables simpler representations of overall genome architectural features. First, the Cluster C virions all have myoviral morphotypes, relatively large capsids, and longer genomes (Table 1). The C1 subcluster genomes are extremely similar to one another, with greater than 98% pairwise ANI values (Table 3), and yet the single C2 subcluster phage, Myrna, is clearly a distant relative. A plausible explanation is that Myrna only relatively recently acquired the ability to infect *M. smegmatis*, and it remains to be seen if other C cluster genomes that are distinct from the C1 subcluster are isolated in the future. The virion structural genes are not well defined in any of the C cluster phages, but they do not appear to enjoy the tight linkage and synteny seen with siphoviral virion structure genes. Furthermore, it is not obvious from the pairwise map representations that structural genes are better conserved between the C1 and C2 genomes than the non-structural genes.

All of the other 53 mycobacteriophages have siphoviral morphotypes, and the virion structure and assembly genes are arranged in the highly conserved arrangements found in all such phages. One of the most obvious components is the tape-measure gene, typically readily identifiable as the longest gene in the genome, reflecting the relatively long phage tails (Table S1). While there is strong synteny of the structural genes, the sequence diversity is high, and mosaic relationships are evident. For example, when the pham assignments are presented along with their putative functional roles, the use of different functional cassettes is evident (Table 4). There are, for example, as many as 10 phams/subphams encoding putative capsid subunits, even though viral capsid proteins, including HK97, T4, P22, ϕ29, and Herpes virus, share a common fold.[31–36] There are also 7 phams encoding major tail subunit genes, and there is little or no correspondence between the particular capsid and major tail subunit phams in each cluster/subcluster (Table 4). The diversity among putative large terminase subunits is not so great, with only 3 phams used (Pham2, Pham394, and Pham891) (Table 4), and there is a good correlation between the Pham distribution and the types of genome ends as noted previously.[37] For example, all of the genomes encoding members of Pham394 and Pham891 have terminally redundant ends, while most members of Pham2 have defined cohesive ends. The exceptions to this are the genes encoded by the terminally redundant Cluster C genomes, which only assemble into Pham2 because some members (Catera, Rizal, ScottMcG, and Spud) contain an intein that is also present in Omega gp11, Kostya gp9, and Cjw1 gp8. The Cluster C Pham2 extein sequences are not, however, related to terminases. A phylogenetic reconstruction of the mycobacteriophage terminases is shown in Fig. 6. Finally, we note that there are 6 phams encoding portal proteins (Table 4) and that, similar to the major tail subunit phams, these do not correlate closely within clusters/subclusters with the capsid subunits. It has been noted previously that the genes encoding the DNA packaging system (terminase and portal) are among the best conserved of the tailed-phage-encoded proteins,[11] and it is therefore notable that such extensive variation is seen within these mycobacteriophage genomes.

Mycobacteriophage gene functions

We noted previously that only 15% of the phams identified in the comparative analysis of 30 mycobacteriophage genomes matched existing database entries.[16] Because of expansion of the extant sequence databases and the increase in the number of mycobacteriophage genomes, we have repeated the database searches. Using the set of 1523 phams, we found that 287 (18.8%) matched at least one non-mycobacteriophage entry at an *E*-value greater than 0.001 (Table S2). Forty percent of these match proteins of unknown functions that are annotated as conserved hypotheticals (many of which may be prophage-encoded genes in sequenced bacterial genomes) such that only 11.3% of all 1523 phams currently can be assigned putative functions based on sequence similarity to proteins of known functions. We have identified another set of 20 phams that do not match database entries above the *E*-value threshold but do match a conserved domain. Twenty-one additional phams were assigned putative functions according to their positions within structural gene operons.

In view of the high genetic diversity, abundance of genes of unknown function, and mosaic architectures, we reevaluated the mycobacteriophage genomes for evidence of mobile genetic elements. Recently, we described a new class of ultra small elements, mycobacteriophage mobile elements, that are present in many of the mycobacteriophages[38] but completely absent from the host genomes. Likewise, there is a notable absence of any of the transposons identified in mycobacterial genomes in the phage genomes. We note that although mobile elements are not typically associated with phage genomes, there are numerous examples.[19,29,43–49] Analysis of the mycobacteriophage phams revealed two (Pham789 and Pham1062) that have sequence similarity to putative transposases and likely correspond to additional mobile elements. There are two members of Pham789 (Bethlehem gp71 and Omega gp21) with weak sequence similarity to IS110-like elements, although the ORFs are small (~250 codons) relative to other IS110 family transposases (~400 amino acids). Pham1062 has only a single member (Llij gp83) and contains Transposase-2 and Transposase-35 superfamily conserved motifs with strong similarity to members of the large IS200 family; the closest relative is a putative transposase in *Nocardia farcinica* with which Llij gp83 shares 73% amino acid identity. With only three genes of a total of 6858 mycobacteriophage ORFs with identifiable

sequence similarity to the multitude of known prokaryotic transposons, this would appear to be a highly underrepresented class. A more abundant group of elements includes proteins containing HNH homing endonuclease domains, and at least six phams/subphams have these motifs (members of Pham453, Pham154, Subpham1410-24, Pham1421, Pham126, and Subpham2292-1), including over 50 genes in total. We note that HNH-containing proteins are common residents of phage genomes, including the T-even myoviruses.[50]

There are three phams (Pham2, Pham394, and Pham1944) in which one or more member contains an intein. Two of these (Pham2 and Pham394) encode large terminase subunits. Within Pham2, 6 of the 43 members contain an intein although they are distributed across different clusters [one in A1 (Bethlehem gp2), two in E (Cjw1 gp8, Kostya gp9), two in C1 (Catera gp206, ScottMcG gp208), and one singleton (Omega gp11)]; as noted above, the extein components of the Cluster C Pham2 entries are not terminases (Fig. 6). Only one of the nine members of Pham394 (Pipefish gp6) contains an intein, although it is quite distinct in sequence from any of those in Pham2. Pham1944 includes the recombination directionality factor of Bxb1 (gp47),[51] and three relatives contain inteins (Bethlehem gp51, KBG gp53, U2 gp50); these are grouped into Subpham1944-1 (Table S3). A related intein is also present in Cali gp3 (Cluster C), which likely encodes a nucleotidyltransferase; similar genes lacking this intein are present in the other Cluster C genomes and Wildcat gp58; these constitute subpham1944-2 (Table S3). The mycobacteriophage intein profile in general reflects that of the broader phage population (the intein database currently lists 36 phage-encoded inteins‡) in which phage-encoded inteins are predominantly found in DNA polymerase, ribonucleotide reductase, primase/helicase, thymidylate synthase, and terminase genes.[52–54] The Bethlehem gp51 intein has recently been shown to be the prototype member of a new class III group of inteins.[55] Finally, we note that we have yet to identify introns in any of the mycobacteriophage genomes even though there are many examples of introns in bacteriophages of other hosts.[30,56–58]

Conclusions

The increase in the number of available mycobacteriophage genomes to 60 gives better understanding of the genetic diversity of the phages that infect *M. smegmatis*, but it also begins to reveal information about the genetic structure of the population of these phages. The most obvious feature of our sample of the population is its grouping into clusters. The fact that the different methods we used to define the clusters give similar (although not identical) groupings argues that the clusters have a degree of biological reality, but in that context, there are a large number of genes that do not follow the clustering, owing to their horizontal mobility between the clusters or into one or more of the clusters from outside sources. Thus, the clusters, although biologically meaningful, are separated by boundaries that are not very sharply defined and, we suspect, are intrinsically incapable of being sharply defined. This situation is reminiscent of what is seen in phages of enteric hosts, where distinct types (analogous to our clusters) can be recognized (e.g., T4, λ, Mu, T7, P22, etc.), but as more genome sequences are determined, the individual types become more diverse and more examples of horizontal exchange of genes are seen. It is perhaps surprising that we do not see, among the 60 mycobacteriophages, any examples of large-scale hybrids of the established clusters, analogous to enteric phage N15 (head and tail genes such as phage λ, early genes such as a non-integrating plasmid)[59] or SfV (head genes and early genes such as lambdoid phage HK97, tail genes such as Mu),[22] but we think it is likely that such "hybrids" will be seen as more sequences are determined.

Previously, in grouping 30 sequenced mycobacteriophages, we placed them, with nine being singletons, into six clusters.[16] A doubling of the number of sequenced genomes has increased the number of major clusters to nine, as a result of newly discovered relatives of three genomes that were previously singletons. Furthermore, because no new singleton was discovered, this would suggest that most of the major clusters have been identified. Alternatively, the observation that the fraction of protein phams that have only one member has decreased only incrementally with a doubling in population size argues that we have only begun to scratch the surface of sequence diversity in these phages.

It is not yet clear to what degree this population of 60 phages—phages that grow on one strain of *M. smegmatis* and that were mostly isolated from one geographical location—is representative of mycobacteriophages as a whole. Most of those we examined (37/60) were isolated in the vicinity of Pittsburgh, PA, but the remaining 23 were isolated from India, Japan, and nine states in the United States. The latter group fit into the clusters discussed here, and we are not able to detect any features of their sequences or genome organization that would distinguish them from the Pittsburgh phages. Thus, we favor the view that the phage types defined by the clusters have a global distribution, as has been suggested earlier for phage sequences found in four widely separated marine environments.[60] A separate question is how widely the clusters we define here for phages that grow on a particular strain of *M. smegmatis* are shared with phages that infect other hosts. Of the 60 phages examined here, only those in Cluster A1 and TM4 also efficiently infect *M. tuberculosis*,[38] arguing that there are some similarities in the kinds of phages that infect these two mycobacterial species. Somewhat further afield, we have compared the sequences of the 60 mycobacteriophages with some of the sequenced phages of

‡ http://www.neb.com/neb/inteins.html

Streptomyces (R.W.H., G.F.H., & M. Smith, unpublished observations). In pairwise comparisons, we typically found a small number of genes that match weakly between the two phages, but their orders on the genome maps are often not preserved. *Mycobacterium* and *Streptomyces* are both members of the actinomycetes and so are not very distant from each other phylogenetically, but preliminary comparisons suggest that the phages that infect *Streptomyces* are unlikely to fit into the same clusters as the mycobacterial phages considered here.

The observation that genes in greater genetic flux than the majority of genes are relatively small is consistent with a model in which the majority of horizontal exchange events between phage genomes involve illegitimate recombination events. Because such events require little sequence specificity, most will occur within coding sequences (especially since most of the genome space is protein coding), and there will be a strong tendency toward acquisition of the smallest independent functional domains. Structural studies suggest that protein domains are commonly as small as 60 residues,[61] in reasonable agreement with our finding that genes with the greatest likelihood of recent acquisition average less than 100 amino acids (Fig. 5c). The finding that the exchange of genes primarily involves small segments corresponding perhaps to a single domain helps explain a long-standing yet puzzling feature of phage genomes—that the average phage gene size is only about two-thirds that of the bacterial host.

Comparative genomics of the T-even phages has identified a number of highly divergent hyperplastic regions (HPRs) that contain large numbers of genes of unknown function but may confer adaptations to the host.[62] These phages have a core of commonly conserved genes with which there are no obvious counterparts in the mycobacteriophages. However, the multitude of small genes—especially those populating the right arms of the siphoviral phages (in all clusters except for Cluster C)—is reminiscent of the HPR genes, and it seems likely that they share the property of relatively recent acquisition and the functions of host adaptations.

The proposal that phage genomic mosaicism may be mediated by lambda Red-like recombinases catalyzing homeologous recombination events raises the question as to what mycobacteriophages encode related enzymes.[15] As reported previously, Che9c gp61 is a RecT-like protein that catalyzes recombination between relatively short DNA segments,[63,64] and a total of five mycobacteriophages encode related enzymes (Pham324: Che9c gp61, Brujita gp43; Pham1304: BPs gp43, Halo gp43, Giles gp53); an Erf-like protein also is encoded by the singleton Wildcat (gp64). Eleven mycobacteriophages encode RecA-related proteins (Pham161), all within Clusters C and E. Nearly two-thirds of the mycobacteriophages therefore do not have genes encoding identifiable recombinases, and the question arises as to whether there are new classes of these enzymes that remain to be discovered.

Finally, while this genome-wide view of these mycobacteriophage genomes provides a broad look at their comparative relationships and structures, the large number of different genes, the high genetic diversity, and the abundance of genes of unknown functions mean that there is a wealth of information in the detailed genome structures that has yet to be analyzed. With the development of tools for functional genomic dissection,[65] the prospects are good for positioning this genomic information in the context of the biology of these bacteriophages.

Materials and Methods

Phage isolation, genome sequencing, and analysis

Phages were isolated from various environmental sources as listed in Table 1. Samples were extracted with phage buffer, plated directly on solid overlays containing 0.35% agar and *M. smegmatis* mc^{2}155, and incubated at 37 °C for 24 h as described previously.[16] Individual plaques were picked, passaged through several rounds, amplified, and purified using CsCl equilibrium density gradient centrifugation. DNA preparation, genome sequencing, and bioinformatic analysis were performed as previously described.[16]

During analysis, two previously reported genome sequences were revised. Mycobacteriophage Wildcat was corrected by removal of 145 nucleotides that were errantly included at one end of the genome; the corrected genome length is 78,296. There was no change in gene annotation. The reported sequence of mycobacteriophage Giles contained a 766-bp region errantly included at one end of the genome; the revised sequence is 53,746 bp and lacked the previously annotated gene *79*. GenBank files for both genomes have been corrected.

Genome annotation used a variety of programs, including DNA Master (available online§), Genemark,[66] Glimmer,[67] and Gepard.[68] tRNA and tmRNA genes were identified using tRNAscan-SE (with a relaxed Cove cutoff score of 2) and ARAGORN.[69,70] Table 1 excludes tRNA matches to the *att*P site of F1 cluster phages and a putative *att*P site in Che9c. The default Aragorn settings were used for tRNA and tmRNA searches. The program Phamerator (S.G.C., M.W.B., R.W.H., & G.F.H., unpublished data) was used to assemble ORFs into phams using both a ClustalW cutoff value of 27.5% amino acid identity and a BlastP score of 0.0001. An output showing the assignments of ORFs to phams is shown in Table S2, along with the summaries of BlastP searches of all phams against the GenBank database and putative functional assignments. Twelve phams that contained large numbers of genes and that upon inspection were complex and did not all correspond to a single sequence type were identified. This situation typically arose from one or more genes being hybrids, matching two or more genes that are not related to each other. These complex phams were manually deconvoluted using BlastP searches and by grouping genes into subphams, placing the hybrid genes into a single, randomly designated subpham.

The Pham number designations differ from those reported previously,[16,24,25] reflecting a transition from a manual organization into phams into a fully automated

§ http://cobamide2.bio.pitt.edu/

system using Phamerator. The Phamerator program is written to maintain the present Pham designations when additional genomes are added to the database. However, some renumbering is unavoidable due to circumstances in which genes previously placed in different Phams may be joined into new Phams (S.G.W., R.W.H., & G.F.H., unpublished observations).

Electron microscopy was performed by placing a suspension of virion purified through a CsCl gradient onto a sample grid with a carbon-coated nitrocellulose film, staining with 2% uranyl acetate, and examining the grid in an FEI Morgagni 268 transmission electron microscope equipped with an AMT digital camera system.

Accession numbers

The accession numbers for phages are as follows: 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; P-Lot, DQ398051; Adjutor, EU676000; Boomer, EU816590; BPs, EU203571; Brujita, FJ168659; Butterscotch, FJ168660; Cali, EU826471; Chah, FJ174694; DD5, EU744252; Fruitloop, FJ174690; Giles, EU203571; Gumball, FJ168661; Jasper, EU744251; KBG, EU744248; Konstantine, FJ174691; Kostya, EU816591; Lockley, EU744249; Myrna, EU826466; Nigel, EU770221; Pacc40, FJ174692; Phaedrus, EU816589; Porky, EU816588; Predator, EU770222; Pukovnik EU744250; Ramsey, FJ174693; Rizal, EU826467; ScottMcG, EU826469; Solon, EU826470; Spud, EU826468; Troll4, FJ168662; and Tweety, EF536069.

Acknowledgements

This work was supported in part by a grant to the University of Pittsburgh by the Howard Hughes Medical Institute in support of G.F.H. under the institution's Professors Program. Support was also provided by grants from the National Institutes of Health to R.W.H. (GM51975) and G.F.H. (AI28927). We thank Christina Ferreira for superb technical assistance. We also acknowledge the following students and teachers who contributed to genome annotation and analysis: (1) Anand Naranbhai and Melisha Sukkhu (Brujita), Natasha Pillay and Reevanan Naidoo (Gumball), and Fortunate Ndlandla, Karnishree Govender, and Mantha Makume (Butterscotch), all of whom were participants in the 2008 KwaZulu-Natal Research Institute for TB and HIV (K-RITH) Phage Discovery Workshop at the Nelson R. Mandela School of Medicine led by G.F.H., D.J.S., William R. Jacobs, Jr., Michelle Larsen, and A. Wilhelm Sturm; (2) Tom Bogen, Gary Osowick, and Greg King (Fruitloop), Rachael Becker, Beth Smyder, and Sandy Breitenbach (Pacc40), Sue Glennon, Susan Offner, Carol Seemuller, and Sandy Wardell (Ramsey), Kathy VanHoeck, Jen Gordinier, Chris Bogiages, and Blair Buck (Solon), Jerry Fuelling, Joan-Beth Gow, Sue Lentz, and Bill Welch (Troll4), all of whom participated in a teacher phagehunting workshop (2008) at the University of Pittsburgh; (3) Roger Chambers and Dalton Paluzzi (Lockley) and Chris Lyons and Dan Altman (Adjutor), who participated in the Science Education Alliance Pilot Course (2008) at the University of Pittsburgh; and (4) Sam Miake-Lye and Dr. Susan Offner (KBG) at Lexington High School, Lexington, MA.

We also thank the high school teachers and their classroom phagehunters from the following for isolation of phages: (1) Upper St. Clair High School, Pittsburgh, PA, for DD5 and Myrna; Greater Latrobe Junior High School, Latrobe, PA, for Fruitloop; Lexington High School, Lexington, MA, for Jasper; Champlin Park High School, Champlin Park, MN, for Ramsey; York Community High School, Elmhurst, IL, for Solon; and St. Andrew's Episcopal School, Rockville, MD, for Troll4.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2010.01.011

References

1. Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor. Popul. Biol.* **61**, 471–480.
2. Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A. *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–182.
3. Brussow, H. & Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell*, **108**, 13–16.
4. Hendrix, R. W. (2003). Bacteriophage genomics. *Curr. Opin. Microbiol.* **6**, 506–511.
5. Hatfull, G. F. (2008). Bacteriophage genomics. *Curr. Opin. Microbiol.* **11**, 447–453.
6. Wilhelm, S. W., Jeffrey, W. H., Suttle, C. A. & Mitchell, D. L. (2002). Estimation of biologically damaging UV levels in marine surface waters with DNA and viral dosimeters. *Photochem. Photobiol.* **76**, 268–273.
7. Hendrix, R. W., Hatfull, G. F. & Smith, M. C. (2003). Bacteriophages with tails: chasing their origins and evolution. *Res. Microbiol.* **154**, 253–257.
8. Hendrix, R. W., Lawrence, J. G., Hatfull, G. F. & Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508.
9. Casjens, S. R. (2008). Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. *Res. Microbiol.* **159**, 340–348.
10. Hendrix, R. W. (2009). Jumbo bacteriophages. *Curr. Top. Microbiol. Immunol.* **328**, 229–240.
11. Casjens, S. R. (2005). Comparative genomics and evolution of the tailed-bacteriophages. *Curr. Opin. Microbiol.* **8**, 451–458.
12. Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E. & Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the

world's a phage. *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.
13. Susskind, M. M. & Botstein, D. (1978). Molecular genetics of bacteriophage P22. *Microbiol. Rev.* **42**, 385–413.
14. Clark, A. J., Inwood, W., Cloutier, T. & Dhillon, T. S. (2001). Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. *J. Mol. Biol.* **311**, 657–679.
15. Martinsohn, J. T., Radman, M. & Petit, M. A. (2008). The lambda red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. *PLoS Genet.* **4**, e1000065.
16. Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E. *et al.* (2006). Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet.* **2**, e92.
17. Lawrence, J. G. & Hendrickson, H. (2003). Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**, 739–749.
18. Kwan, T., Liu, J., Dubow, M., Gros, P. & Pelletier, J. (2006). Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. *J. Bacteriol.* **188**, 1184–1187.
19. 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. USA*, **102**, 5174–5179.
20. Brussow, H. (2001). Phages of dairy bacteria. *Annu. Rev. Microbiol.* **55**, 283–303.
21. Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* **49**, 277–300.
22. Lawrence, J. G., Hatfull, G. F. & Hendrix, R. W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184**, 4891–4905.
23. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**, 762–777.
24. Morris, P., Marinelli, L. J., Jacobs-Sera, D., Hendrix, R. W. & Hatfull, G. F. (2008). Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by illegitimate recombination. *J. Bacteriol.* **190**, 2172–2182.
25. Pham, T. T., Jacobs-Sera, D., Pedulla, M. L., Hendrix, R. W. & Hatfull, G. F. (2007). Comparative genomic analysis of mycobacteriophage Tweety: evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. *Microbiology*, **153**, 2711–2723.
26. Desplats, C. & Krisch, H. M. (2003). The diversity and evolution of the T4-type bacteriophages. *Res. Microbiol.* **154**, 259–267.
27. Hatfull, G. F. (2006). Mycobacteriophages. In (Calendar, R., ed.), pp. 602–620, Oxford University Press, New York, NY.
28. Fraser, J. S., Yu, Z., Maxwell, K. L. & Davidson, A. R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J. Mol. Biol.* **359**, 496–507.
29. Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F. & Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* **299**, 27–51.
30. Stewart, C. R., Casjens, S. R., Cresawn, S. G., Houtz, J. M., Smith, A. L., Ford, M. E. *et al.* (2009). The genome of *Bacillus subtilis* bacteriophage SPO1. *J. Mol. Biol.* **388**, 48–70.
31. Morais, M. C., Choi, K. H., Koti, J. S., Chipman, P. R., Anderson, D. L. & Rossmann, M. G. (2005). Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of phi29. *Mol. Cell*, **18**, 149–159.
32. Wikoff, W. R., Liljas, L., Duda, R. L., Tsuruta, H., Hendrix, R. W. & Johnson, J. E. (2000). Topologically linked protein rings in the bacteriophage HK97 capsid. *Science*, **289**, 2129–2133.
33. Fokine, A., Leiman, P. G., Shneider, M. M., Ahvazi, B., Boeshans, K. M., Steven, A. C. *et al.* (2005). Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl Acad. Sci. USA*, **102**, 7163–7168.
34. Duda, R. L., Hendrix, R. W., Huang, W. M. & Conway, J. F. (2006). Shared architecture of bacteriophage SPO1 and herpesvirus capsids. *Curr. Biol.* **16**, R11–R13.
35. Hendrix, R. W. (1999). The long evolutionary reach of viruses. *Curr. Biol.* **9**, 914–917.
36. Krupovic, M. & Bamford, D. H. (2009). Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nat. Rev. Microbiol.* **7**, 250; author reply, 250.
37. Casjens, S. R. & Gilcrease, E. B. (2009). Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods Mol. Biol.* **502**, 91–111.
38. Sampson, T., Broussard, G. W., Marinelli, L. J., Jacobs-Sera, D., Ray, M., Ko, C. C. *et al.* (2009). Mycobacteriophages BPs, Angel and Halo: comparative genomics reveals a novel class of ultra-small mobile genetic elements. *Microbiology*, **155**, 2962–2977.
39. Mediavilla, J., Jain, S., Kriakov, J., Ford, M. E., Duda, R. L., Jacobs, W. R., Jr. *et al.* (2000). Genome organization and characterization of mycobacteriophage Bxb1. *Mol. Microbiol.* **38**, 955–970.
40. Ford, M. E., Stenstrom, C., Hendrix, R. W. & Hatfull, G. F. (1998). Mycobacteriophage TM4: genome structure and gene expression. *Tuber. Lung Dis.* **79**, 63–73.
41. Ford, M. E., Sarkis, G. J., Belanger, A. E., Hendrix, R. W. & Hatfull, G. F. (1998). Genome structure of mycobacteriophage D29: implications for phage evolution. *J. Mol. Biol.* **279**, 143–164.
42. Hatfull, G. F. & Sarkis, G. J. (1993). DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Mol. Microbiol.* **7**, 395–405.
43. Iida, S., Meyer, J. & Arber, W. (1978). The insertion element IS1 is a natural constituent of coliphage P1 DNA. *Plasmid*, **1**, 357–365.
44. Gertman, E., White, B. N., Berry, D. & Kropinski, A. M. (1986). IS222, a new insertion element associated with the genome of *Pseudomonas aeruginosa*. *J. Bacteriol.* **166**, 1134–1136.
45. Walter, T. M. & Aronson, A. I. (1991). Transduction of certain genes by an autonomously replicating *Bacillus thuringiensis* phage. *Appl. Environ. Microbiol.* **57**, 1000–1005.
46. Woods, D. E., Jeddeloh, J. A., Fritz, D. L. & DeShazer, D. (2002). *Burkholderia thailandensis* E125 harbors a temperate bacteriophage specific for *Burkholderia mallei*. *J. Bacteriol.* **184**, 4003–4017.
47. Lo, T. C., Shih, T. C., Lin, C. F., Chen, H. W. & Lin, T. H. (2005). Complete genomic sequence of the temperate bacteriophage PhiAT3 isolated from *Lactobacillus casei* ATCC 393. *Virology*, **339**, 42–55.
48. Casjens, S., Winn-Stapley, D. A., Gilcrease, E. B.,

Morona, R., Kuhlewein, C., Chua, J. E. *et al.* (2004). The chromosome of *Shigella flexneri* bacteriophage Sf6: complete nucleotide sequence, genetic mosaicism, and DNA packaging. *J. Mol. Biol.* **339**, 379–394.
49. Chibani-Chennoufi, S., Dillmann, M. L., Marvin-Guy, L., Rami-Shojaei, S. & Brussow, H. (2004). *Lactobacillus plantarum* bacteriophage LP65: a new member of the SPO1-like genus of the family Myoviridae. *J. Bacteriol.* **186**, 7069–7083.
50. Nolan, J. M., Petrov, V., Bertrand, C., Krisch, H. M. & Karam, J. D. (2006). Genetic diversity among five T4-like bacteriophages. *Virol. J.* **3**, 30.
51. Ghosh, P., Wasil, L. R. & Hatfull, G. F. (2006). Control of phage Bxb1 excision by a novel recombination directionality factor. *PLoS Biol.* **4**, e186.
52. Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L. & Hilario, E. (2002). Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* **56**, 263–287.
53. Lazarevic, V., Soldo, B., Dusterhoft, A., Hilbert, H., Mauel, C. & Karamata, D. (1998). Introns and intein coding sequence in the ribonucleotide reductase genes of *Bacillus subtilis* temperate bacteriophage SPbeta. *Proc. Natl Acad. Sci. USA*, **95**, 1692–1697.
54. Perler, F. B. (2002). InBase: the intein database. *Nucleic Acids Res.* **30**, 383–384.
55. Tori, K., Dassa, B., Johnson, M. A., Southworth, M. W., Brace, L. E., Ishino, Y. *et al.* (2010). Splicing of the mycobacteriophage Bethlehem DnaB intein: identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285**, 2515–2526.
56. Sandegren, L. & Sjoberg, B. M. (2004). Distribution, sequence homology, and homing of group I introns among T-even-like bacteriophages: evidence for recent transfer of old introns. *J. Biol. Chem.* **279**, 22218–22227.
57. Haugen, P., Simon, D. M. & Bhattacharya, D. (2005). The natural history of group I introns. *Trends Genet.* **21**, 111–119.
58. Tourasse, N. J. & Kolsto, A. B. (2008). Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. *Nucleic Acids Res.* **36**, 4529–4548.
59. Ravin, V., Ravin, N., Casjens, S., Ford, M. E., Hatfull, G. F. & Hendrix, R. W. (2000). Genomic sequence and analysis of the atypical temperate bacteriophage N15. *J. Mol. Biol.* **299**, 53–73.
60. Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C. *et al.* (2006). The marine viromes of four oceanic regions. *PLoS Biol.* **4**, e368.
61. Veretnik, S., Bourne, P. E., Alexandrov, N. N. & Shindyalov, I. N. (2004). Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.* **339**, 647–678.
62. Comeau, A. M., Bertrand, C., Letarov, A., Tetart, F. & Krisch, H. M. (2007). Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology*, **362**, 384–396.
63. van Kessel, J. C. & Hatfull, G. F. (2008). Efficient point mutagenesis in mycobacteria using single-stranded DNA recombineering: characterization of antimycobacterial drug targets. *Mol. Microbiol.* **67**, 1094–1107.
64. van Kessel, J. C. & Hatfull, G. F. (2007). Recombineering in *Mycobacterium tuberculosis*. *Nat. Methods*, **4**, 147–152.
65. Marinelli, L. J., Piuri, M., Swigonova, Z., Balachandran, A., Oldfield, L. M., van Kessel, J. C. & Hatfull, G. F. (2008). BRED: a simple and powerful tool for constructing mutant and recombinant bacteriophage genomes. *PLoS ONE*, **3**, e3957.
66. Borodovsky, M. & McIninch, J. (1993). Recognition of genes in DNA sequence with ambiguities. *Biosystems*, **30**, 161–171.
67. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641.
68. Krumsiek, J., Arnold, R. & Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, **23**, 1026–1028.
69. Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964.
70. Laslett, D. & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16.
71. Huson, D. H. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68–73.
72. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H. *et al.* (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
73. Perriere, G. & Gouy, M. (1996). WWW-query: an online retrieval system for biological sequence banks. *Biochimie*, **78**, 364–369.