Published in final edited form as:

Res Microbiol. 2008 June ; 159(5): 340–348. doi:10.1016/j.resmic.2008.04.005.

Diversity among the tailed-bacteriophages that infect the *Enterobacteriaceae*

Sherwood R. Casjens

Department of Pathology, University of Utah School of Medicine, Salt Lake City, UT 84112, USA

Abstract

Complete genome sequences have been determined for seventy-three tailed-phages that infect members of the bacterial *Enterobacteriaceae* family. Biological criteria such as genome size, gene organization and gene orientation were used to place these phages into categories. There are thirteen such categories, some of which are themselves extremely diverse. The relationships between and within these categories are discussed with an emphasis on the head assembly genes. Although some of them are clearly homologues, suggesting a very ancient origin, there is little evidence for exchange of individual head genes between these phage categories. More recent horizontal exchange of phage tail fiber and early proteins between the categories occurs, but is probably not extremely rapid.

Keywords

Bacteriophage; Diversity; Evolution; Genomics; Horizontal exchange

1. Introduction

Bacteriophages are thought to be the most abundant “organisms” in Earth’s biosphere, and their diversity is becoming legendary among biologists. There are eight different major phyla of phages that have very different molecular lifestyles and are only extremely distantly related, if they are related at all [17]. This discussion will be limited to only one of these groups, the dsDNA tailed-phages or *Caudovirales*. The number of individual tailed-phage virions on Earth has been estimated to be over 10^{30} (one million trillion trillion) [59,67], and in addition, nearly as many phage genomes may be integrated into bacterial chromosomes as prophages [7]. Rohwer [55] calculated that over two billion phage novel phage gene types likely remain to be discovered, although nearly identical phage genes can be widely distributed [3]. The over three hundred genome sequences in the current database do show a huge amount of variation; there are many examples of genome pairs that have very little *apparent* homology as determined by searches for nucleotide or encoded protein sequence similarity. The above phage virion and gene numbers are incomprehensibly large, but can we understand this diversity by understanding a more limited number of phage *types*? Estimates of the number of extant bacterial species has ranged from ten million upwards [16,55], and, where studies have been done, multiple types of tailed-phages have been found that infect each bacterial species, so even the number of types could be extremely large.

Correspondence and reprints: Department of Pathology, Room 5200K EEJMRB, University of Utah School of Medicine, Salt Lake City, UT 84112, E-mail: Sherwood.casjens@path.utah.edu, Phone: 801-581-5980, FAX: 801-585-2417.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

In spite of their extreme diversity, the various tailed-phages do have features in common. For example, (i) they all by definition have tails, and (ii) all such phages examined (and presumably *all* of them) utilize a common mechanism of DNA packaging. In particular, they assemble a protein procapsid first and then the DNA molecule is pumped into the procapsid by an ATP cleavage-powered DNA translocase [11]. The proteins that make up the motor of this DNA pump are called terminase and portal protein. The large terminase subunit carries the ATPase site, and the portal protein forms the hole through which DNA enters the procapsid and is the site to which terminase binds during DNA packaging. Of all the proteins encoded by the tailed-phages, these are the only two that can be recognized by sequence similarity to be encoded by virtually *all* tailed-phages [7,10,43]. Thus, at least parts of head assembly and head structure appear to have been invented only once during evolution, and all extant individuals have a single common ancestor in terms of this aspect of their life cycle. Another head assembly protein that *may* be universally conserved is the coat protein that forms the icosahedral shell of the phage head. Tailed-phage coat proteins may all have a common fold, although in many cases similarity cannot be easily found when their amino acid sequences are compared. Information is rather meager at present but it supports this view; an atomic x-ray structure is only known for *Escherichia coli* phages HK97 coat protein and T4 vertex coat protein [19, 65], and they have the same fold. In addition, high resolution 3-dimensional reconstructions of virions of *Salmonella enterica* phage P22 and *Bacillus subtilis* phage ϕ 29 have shapes that appear to be identical to the HK97 and T4 coat proteins [34,45]. The detailed mechanisms of the other essential functions in the life cycles of tailed-phages, such as adsorption and injection (virion tails), DNA replication, control of phage gene expression, and lysis are sufficiently varied that it is clear that in these cases different tailed-phages can encode non-homologous but analogous proteins that carry out similar functions by different mechanisms.

2. Tailed-phage “types”?

We may begin to obtain a clearer picture of the nature of this diversity by examining in detail the tailed-phages that infect one or a few related bacterial species. Genome sequences of numerous phages with the same or related hosts are known in several bacterial phyla, but in this discussion the diversity of the tailed-phages that infect the Gram negative *Enterobacteriaceae* family is examined. Seventy-three *Enterobacteriaceae* tailed-phage genomes have been sequenced by a large number of workers (too numerous to cite here). These phages infect six of the forty-nine genera in this family – *Escherichia*, *Salmonella*, *Shigella*, *Erwinia*, *Klebsiella* and *Yersinia*.

Even in this limited context, however, the “simple” question of how many phage types exist contains significant difficulties. First, tailed-phages have limited morphological variation, yet they are currently classified by their tail type (long contractile, long non-contractile, or short; the *Myoviridae*, *Siphoviridae* and *Podoviridae* families, respectively). Tail morphology, although easy to observe, is not a particularly good indicator of overall relatedness [9,42]. This field of inquiry is left in the unfortunate position that in order to robustly understand diversity and evolutionary relationships, the individuals being compared must have completely sequenced genomes (or even more effort needs to be applied to detailed genetic and biochemical studies of their life cycles). Second, decisions must be made about how to define different tailed-phage “types.” This point is fraught with complexity and problems. If too general a feature is chosen, say “presence of a large terminase subunit gene,” then there is no discrimination and all tailed-phages belong to a single type. On the other hand if nearly complete identity is required, then almost every phage that has been examined in the laboratory to date constitutes a unique “type” by itself. Thus, some middle ground (the many details of which are admittedly debatable) must be found in which it makes biological sense to group individuals together in a manner in which groups are not so few as to hide the marvelous diversity of phages and are not so many as to be numerically overwhelming. Third, placing

different phages into such types, even after the types have been defined, is not always perfectly straightforward, since each of the types (as defined below) are known to have “mosaic” genomes. Genomes are mosaically related when in pair-wise comparison their genomes have genes of the same function aligned in the same order, but the sequence relatedness of genes of similar function varies from identity to having no recognizable relatedness [Refs. 6,7,9,27, 28,29 and numerous references therein]. Such mosaicism has the peculiar feature that two phages with very little actual sequence similarity can be included in the same “type” if they each have *different* similarities to another member of the group (e.g., phages P22 and N15 can both be considered to be members of the lambdoid type [Fig. 2 in Ref. 7]).

For purposes of argument here, we will define tailed-phage “types” as having similar genome size, gene function order, transcriptional pattern, as well as relatedness of their genes. In the case of the latter, we will focus on the head genes, since all tailed-phages carry a homologue of at least two of the head genes and they do not appear to be as rapidly horizontally transferred or as frequently separated by evolutionary processes as many of the other phage genes. So, for example, major differences in genome size and transcriptional pattern with few genes that have high sequence similarity will unambiguously constitute different types according to the above rules. This definition of “type” is purposefully somewhat imprecise in order to avoid various details that could confound such an analysis (e.g., see phages λ , N15 and ϕ KO2 below). This represents a more biologically informed approach than simple degree of sequence similarity.

One of the first phage types to be informally defined in this way was the “lambdoid” group. Member phages were initially included in this group by their similar lifestyles and especially their ability to form viable recombinants with λ during mixed infection [30], for example phage λ -434, λ -P22 and λ - ϕ 80 hybrids were isolated through co-infections in the laboratory [2,35, 60]. But what of phages that are observed to form hybrids not with λ itself but with phages that can hybridize with λ (P22-Fels-1 [15]), or phages that can be genetically engineered to form hybrids that are viable in the laboratory [5,52]? With the advent of rapid genome sequencing the de facto definition of “lambdoid” has become the group of phages that have very similar genome sizes, gene orders and transcription patterns, and have some syntenic, moderate to high sequence similarity to another member of the group (i.e., might *in theory* be able to form viable hybrids by *homologous* recombination) [25].

3. Completely sequenced tailed-phages that infect the *Enterobacteriaceae*

Fig. 1 diagrams the genomes of six *Enterobacteriaceae* tailed-phages that have chromosomes in the 38–62 kbp range. They all have head and tail genes in the same relative positions to each other – most head genes are transcriptionally upstream of the tail genes (a common arrangement, but not universal); however, these six phages have major differences in, for example, the location of the lysis genes, as well as overall gene organization and gene orientation/transcriptional direction. These six phages might therefore represent six different types by the rules suggested above; however, the similarity between their head genes caused us to make T1 and EcoM-GJ1 subgroups within a larger “type” (see below). Fig. 2 lists the seventy-three *Enterobacteriaceae* tailed-phages whose complete genome sequences are currently known (this discussion is limited to phages whose genomes are completely sequenced because of the difficulties discussed above). These phages can, by inspection, be placed into thirteen different “types” as defined above, that are typified by phages T1, T7, T4, T5, N4, Felix-O1, SETP3, 9NA, λ , P2, Mu, P1 and ϵ 15. These thirteen types each carry mostly genes that are not recognizably similar to each other; for example, their head shell coat proteins have very little, if any, recognizable amino acid sequence similarity. Each of the other sixty *Enterobacteriaceae* tailed-phages can be fit with some confidence into one of these groups (i.e., the members of each group are more like one another than they are like the other phages). In Fig. 2, these types are separated by thin horizontal lines, except for the two that are denoted

by shaded boxes; the latter can be subdivided into clear, rather different subtypes. The first of these are the T1 and ϕ EcoM-GJ1 subtypes of the “T1 type”; these are grouped together, in spite of some differences in gene organization because, unlike the other types, the head assembly genes of these two phages are moderately similar in sequence [33,54]. The members of the “lambdoid phage type” shown in the lower gray box (here the term “ λ -like” is reserved for the five phages on the λ line in the figure) all have very similar gene organization and transcription patterns, but the seven subtypes listed on different lines within the lambdoid shaded box have head assembly genes that are not or only barely recognizably similar in sequence comparisons.

The subtype discussion above points out that there are ambiguities and interpretational complexities in developing such a classification scheme. For example, although phages N15 and λ both have the typical lambdoid transcription pattern, and their late virion morphogenetic genes are very similar, they have very different early genes which program different DNA replication mechanisms and different prophage lifestyles (the λ prophage is integrated into the host chromosome while N15's is a linear plasmid) [51]. On the other hand, phages PY54 and ϕ KO2 have similar early genes and prophage lifestyles to N15, but their late operons are very different from that of λ and N15 [8,31]. Another somewhat ambiguous case is phage ϵ 15. Its genome is organized similarly, but not identically to the lambdoid phages (its lysis genes are in a different location), its morphogenetic genes are only extremely remotely related to any other phage type defined here [38]. On the other hand, there are a small number of ϵ 15 genes that are fairly close relatives to lambdoid phage genes; however, since these similar genes are not syntenically related to their lambdoid homologues, they are considered here to be examples of fairly recent horizontal exchange between phage types (see below), and ϵ 15 is not considered to be a lambdoid phage in this analysis. A final example of such difficulties is that phages T1, HK97 and λ , which all have long non-contractile tails, have some rather distantly related but clearly homologous tail genes; similarly the contractile tails of phages Mu, SfV and P2 are encoded by genes that also have some distantly related homologies. Tail genes have clearly undergone some horizontal exchanges relative to head genes over the eons, but our focus on head genes in the current discussion ignores this.

An interesting feature of this type of grouping is that tail morphology, the feature that is currently used at the highest level to classify tailed-phages into three families, *myoviridae*, *siphoviridae* and *podoviridae*, is not a major factor. In fact, several of the types and even subtypes defined here, the lambdoid type and T1-like and HK97-like subtypes, have members that have different tail morphologies. T1 has a long non-contractile tail while PY100 has a contractile tail [57], and λ and HK97 long non-contractile tails, while SfV and P22 have a long contractile and a short tail, respectively. P22, for example, has head and tail genes that are not recognizably similar to those of λ , but the two phages can form viable hybrids and have gene organization, transcription pattern and early genes that are extremely similar; indeed, two cognate regulatory proteins, the λ gene *Q* protein and P22 gene 23 protein (96% identical), are known to have the same DNA binding/transcription antitermination specificity [2,53].

We can observe from this analysis that every *Enterobacteriaceae* tailed-phage is not equally different from every other one, and phages do, in fact, fall quite naturally into biological groupings that are more like each other than they are like the other enteric phages. We can't know for sure at this point, but the fact that a majority of the enteric phages in Fig. 2 fall unambiguously into types with multiple members suggests that a limited number of *Enterobacteriaceae* tailed-phage types exist; just any random combination of phage genes which might come together to form a set of functions required “to be a tailed-phage” does not make an evolutionarily viable phage. The identification of such types is likely not near saturation, since a number of the types are represented by a single individual, and several of these have been characterized only recently (e.g., ES18, 9NA, N4, Felix-O1, ϕ EcoM-GJ1,

ε15). Analysis of more *Enterobacteriaceae* phage genomes may confirm the current apparent existence of discrete types, but it is also likely to discover new types and could muddy the waters with “hybrids” that fall between the current types defined here? We can also see from Fig. 2 that there is a clean division between the lytic temperate phages, i. e., none of these phages appear to have undergone a recent conversion between the temperate and lytic lifestyles. Although this may seem self-obvious or trivial, it is worth noting because it seems theoretically quite simple for a temperate phage to lose its ability to form a lysogen, and rather closely related lytic and temperate phages are known that infect the mycobacteria [20].

4. Variation within tailed-phage types

But what of variation within types? Fig. 3 shows the head gene regions of the seven lambdoid subtypes mentioned above (Fig. 2). The differences include the number of gene products involved and differences in mechanisms of assembly (e.g., use or not of scaffolding proteins and proteolytic cleavage during assembly), in addition to the very different amino acid sequences of their head proteins. The difference among the head genes in Fig. 3 exemplifies the much-discussed mosaic variation within types, but there is also simple sequence divergence. Let us consider the phages whose coat proteins are most like those of the very well-studied phage λ. The database of related *Enterobacteriaceae* temperate phage nucleotide sequences is actually much richer than is indicated in Fig. 2, which only shows those fully functional phages whose genomes have been completely sequenced. Partial genome sequences have been determined for additional phages, and prophages related to these fully functional phages are found in most genome sequences of their host bacterial species [4,7]. A significant fraction of these putative prophages appear, by inspection of their sequences, to be potentially fully functional phages, although nearly all have not been studied in this regard. On the other hand, many of these prophages have suffered deletions and are no longer fully functional. However, in the cases that have been studied, many of the *apparently* intact genes in such defective prophages have been found to actually be functional [summarized in 7]. It is therefore not unreasonable to assume that prophage genes that retain their full length relative to known functional homologues are likely to be functional. In addition, it has been argued that removal of prophage genes by deletion is faster than their obliteration by point mutation [7,41]. Thus, it is reasonable to include such genes (even if partially deleted) in diversity and evolutionary studies, since even if a few unselected missense mutations had occurred in the gene after its integration with the prophage, these should not be sufficient to have an obfuscating effect on the analysis (except in the interpretation of relationships among closely related sequences).

As an example of an analysis of all the extant sequences of one lambdoid phage subtype, the phage λ coat protein (gene *E* protein) was used as a probe in a BLASTp [1] search of the current sequence database. The top 57 matches were the five phages listed on the phage λ line of Fig. 2 and 52 putative prophages in *E. coli*, *Shigella flexneri*, *Shigella boydii*, *Shigella sonnei*, and *Salmonella enterica* genomes. Fig. 4 shows a neighbor-joining tree of these λ-like coat proteins, and it indicates that they fall into three rather robustly delineated branches (and two additional sequences outside of these groups). Two of these coat protein branches, I and II, are typified by phages λ and 21, respectively, and no previously studied phage is present in branch III. The amino acid sequences of category I and II coat proteins are on average about 60–65% identical, and category III proteins are on average about 40–45% identical to the other two types. The variation within coat protein categories I, II and III ranges up to about 13%, 17% and 11% different, respectively. Nearly all of the prophages that harbor these genes have not been fully analyzed, but in many cases (all of those analyzed) there are nearby syntenic λ-like head genes. Thus, it appears that the λ-like coat protein genes must have been diversifying within the *Enterobacteriaceae* for a rather long time to accumulate this many differences, and at least in this superficial analysis have not been exchanging with coat genes of the other lambdoid subtypes such as the P22-like or 933W-like phages. A parallel analysis of the coat protein of

the twenty-one phages and prophages of the P22-like lambdoid subtype indicates a very similar although even more divergent tree that contains three robustly delineated P22-like coat protein categories that range from only about 15% to 30% identical. Again, these prophages have other syntenic P22-like head assembly genes nearby (S. Casjens, to be published elsewhere). The coat proteins of phage λ and P22 have no recognizable amino acid sequence similarity, although they may well be ancient homologues that have diverged beyond the point of our being able to recognize such similarity (above).

In this search with the λ coat protein, the most closely related sequences outside the *Enterobacteriaceae* part of the tree are from phages and prophages that infect(ed) *Vibrionaceae* and *Burkholderiaceae*. These bacterial families are in the orders *Gammaproteobacteria* and *Betaproteobacteria*, respectively, and so are not extremely distantly related to the *Gammaproteobacteria*/ *Enterobacteriaceae* whose phages are under discussion here. We also note that, in Fig. 4, all the sequences of *Salmonella* phage and prophage λ -like coat protein sequences (phage Gifsy-1 and three putative prophages) cluster together on a single robust branch. This suggests, perhaps surprisingly, that these phages (at least the λ -like coat protein gene) may have moved to *Salmonella* recently and only once, and they have *not* been moving *rapidly* back and forth, even between these closely related bacterial host species. The *E. coli* and *Shigella* phages and prophages are intermingled on the tree. But this situation is similar to the *coli* and *Shigella* bacteria themselves, which probably in fact constitute a single species [50]; so this does not indicate recent host species switching among the λ -type phages. Parallel searches of the sequence database with the phage P2 coat protein (gene N protein) finds about one hundred matching proteins with >35% identity. Many of these are in prophages in, for example, the genera *Xanthomonas*, *Burkholderia*, *Escherichia*, *Salmonella*, *Vibrio*, and *Haemophilus*. Here again, there is a significant clustering of host genus according to coat protein subtype [47] (S. Casjens, unpublished). These kinds of analyses suggest that host switching by phages, although certainly theoretically possible and an attractive mode of evolutionary spread and exchange of phage (and other) genes, does not happen with extremely high frequency, even among rather closely related *Enterobacteriaceae* hosts.

5. Genetic exchange between types

Several rather convincing examples of fairly recent horizontal exchange have been discovered between the *Enterobacteriaceae* tailed-phage types defined here. The first of these to be discovered was the similarity between the tail fiber assembly (Tfa) proteins of phage λ and phage T4. They are only about 40% similar in amino acid sequence, including conservative substitutions [22], but the λ and T4 proteins can functionally substitute for one another [23, 44]. Clearly there has been genetic contact between the lambdoid and T4-like phage types since the divergence of their coat proteins. Other examples are the similar tailspike proteins of phages of different types, namely phages P22 [58], SP6 [14,56], Det7 [62], SETP3 (deduced from Accession No. EF177456) and 9NA ([66]; S. Casjens and R. Hendrix, unpublished). These tailspikes bind and cleave the polysaccharide portion of the external lipopolysaccharide during the adsorption process [reviewed in 63], and it appears that this successful adsorption mechanism has spread among phage types that infect *Salmonella* and *E. coli*. Finally, the high similarity of ϵ 15 genes 44 and 45 to the non-syntenic phage P22 *eaE* and *eaD* genes (92% protein identity) and genes 46 and 47 to the non-syntenic phage ES18 genes 39 and 38 (87% protein identity) indicate recent genetic contact between the types represented by these phages. Exchange of genes between types is not common – certainly not frequent enough to severely disrupt our definition of phage types – but it clearly can and does occur.

6. Relationship of *Enterobacteriaceae* tailed-phages to phages that infect other bacterial families

Are all bacterial families host to a similar variety of tailed-phages? Those that have been examined serve as hosts for multiple virion morphotypes, but in no other families have phages been as extensively sequenced as in the *Enterobacteriaceae*. The following families are currently the next best studied: *Staphylococcaceae* [e.g., Ref. 39], *Streptococcaceae* (numerous workers), *Mycobacteriaceae* [24,48] and *Pseudomonadaceae* [e.g., Ref. 40] have 43, 36, 31 and 21 complete tailed-phage genome sequences in the current database, respectively. In each case, these clearly fall into multiple types, but they have not been analyzed by the criteria used here. For example, matrix similarity plots of the genomes of the Gram-positive *Mycobacteriaceae* and *Staphylococcaceae* phage genome sequences indicate that they fall into at least thirteen and seven different “sequence types” respectively, and their genomes range from 42 to 156 kbp and 15 to 138 kbp, respectively [24,39]. Both groups include phages with different virion morphologies, and temperate and lytic members have been characterized. However, this kind of comparison may not recognize relationships between more distantly related phages like HK97, Φ 27 and ϕ KO2 which are grouped together here in Fig. 2 on the basis of gene organization and head *protein* relationships rather than *nucleotide* sequence relationships; the latter only recognizes rather close relationships, and even the subtypes discussed above here are sufficiently divergent that their head gene similarities are not recognizable in matrix plots. Among the eighteen *Lactococcus* (a genus in the *Streptococcaceae* family) phages studied, Deveau et al. [13] recognized eight types with long non-contractile tails and two with short tails. Thus, the *Enterobacteriaceae* tailed-phage diversity situation seems, at our current level of understanding, typical of bacterial families.

Are the tailed-phage types discussed here limited to the *Enterobacteriaceae*? Or should they include phages that encompass a wider range of hosts? This question is even more fraught with problems than attempts to define biological types within the *Enterobacteriaceae* phages. In theory, viruses would seem to have the evolutionary capacity to gain the ability to infect new hosts, some of which might not be close relatives of their original host. Thus, phage genes might be expected to be able to flow at some rate between phages whose hosts are not closely related [26], and indeed the phage lifestyles and encoded proteins represented among the *Enterobacteriaceae* phages are not limited to phages that infect the *Enterobacteriaceae*. For a few proteins, like large terminase subunit and portal protein, homologues appear to be universally present in tailed-phage virions [7,9], although clear cases of horizontal transfer of the genes for these proteins have been difficult to document. It is not the purpose of this report to discuss this aspect of phage diversity in detail, but it is worth noting that if more highly diversified (than terminase and portal) proteins like the phage head shell coat protein are used as BLASTp probes, matches are found outside the *Enterobacteriaceae*. For example, as is shown in Fig. 4, the λ coat protein matches a few prophages in the *Vibrionaceae* and a *Burkholdariaceae* phage. Similarly, *Pseudomonas aeruginosa* (*Pseudomonadaceae* family) phage D3 has a very similar gene organization (except its lysis genes are in a different location) and has low to moderate sequence similarity in many of its genes to those of the lambdoid phages [37]. Host diversity is also seen in the search with the phage P2 coat protein (above), where numerous matches are found to prophages in the *Xanthomonadaceae*, *Pseudomonadaceae*, *Vibrionaceae* and *Burkholdariaceae*, and phages ϕ CTX [46], ϕ Mha-PH101 [32], K139 [36] and ϕ RSA1 [21], which infect members of *Pseudomonadaceae*, *Pasteurellaceae*, *Vibrionaceae* and *Burkholdariaceae*, respectively, have similarities in both gene organization and gene sequence to *E. coli* phage P2 type phages. All of these families, like *Enterobacteriaceae*, are members of the order *Gammaproteobacteria* except the last one, which is in the order *Betaproteobacteria*. Finally, large lytic phages with head genes that are recognizably related to those of *E. coli* phage T4 have been discovered that infect

cyanobacteria, a major bacterial phylum that is only very distantly related to the *Proteobacteria* [12,18,64]. Thus, phages with features of the types defined here do infect hosts from different bacterial phyla, but tailed-phages are very ancient, and it is not yet entirely clear whether the current state of affairs is due to long co-evolution with their hosts, horizontal transfer due to phage host switching or (most likely) both. Further research, and no doubt much more tailed-phage comparative genomics, will be required to gain an understanding of the differences among phages of the same “type” that infect very closely related bacteria relative to the differences among these phage and relatives of similar lifestyle that infect more distantly related bacteria.

7. Conclusions

Seventy-three tailed-phages that infect members of the bacterial *Enterobacteriaceae* family currently have completely sequenced genomes. Criteria such as genome size, gene organization and gene orientation can be used to place these phages into biologically sensible categories. There are thirteen such “types”, some of which are themselves extremely diverse. Analysis of variation within and among these types suggests that genetic exchange does occur between the “types”, but is not frequent enough to seriously disrupt this sort of classification. The *Enterobacteriaceae* tailed-phages are especially informative in such comparative studies because of the immense amount of molecular genetic research that has gone into gaining an understanding of the functions of many of their genes. Similar analyses of phages that infect other bacterial phyla are severely limited by the large fraction of genes whose functions are unknown.

Acknowledgements

I thank Roger Hendrix, Graham Hatfull, Guy Plunkett III, Andrew Kropinski and Robert Villafane for permission to include mention of unreported phage sequences in Fig. 2. The author’s research is supported by NIH research grant IRO1AI074825.

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402. [PubMed: 9254694]
2. Botstein D, Herskowitz I. Properties of hybrids between Salmonella phage P22 and coliphage lambda. *Nature* 1974;251:584–589. [PubMed: 4608090]
3. Breitbart M, Miyake JH, Rohwer F. Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett* 2004;236:249–256. [PubMed: 15251204]
4. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H. Prophage genomics. *Microbiol. Mol. Biol. Rev* 2003;67:238–276. [PubMed: 12794192]
5. Casjens S, Eppler K, Parr R, Poteete AR. Nucleotide sequence of the bacteriophage P22 gene *l9* to *3* region: identification of a new gene required for lysis. *Virology* 1989;171:588–598. [PubMed: 2763468]
6. Casjens S, Hatfull G, Hendrix R. Evolution of dsDNA tailed-bacteriophage genomes. *Sem. Virol* 1992;3:383–397.
7. Casjens S. Prophages in bacterial genomics: What have we learned so far? *Molec. Microbiol* 2003;249:277–300. [PubMed: 12886937]
8. Casjens S, Gilcrease EB, Huang WM, Bunney KL, Pedulla ML, Ford ME, et al. The pKO2 linear plasmid prophage of *Klebsiella oxytoca*. *J. Bacteriol* 2004;186:1818–1832. [PubMed: 14996813]
9. Casjens SR. Comparative genomics and evolution of the tailed-bacteriophages. *Curr. Opin. Microbiol* 2005;8:451–458. [PubMed: 16019256]

10. Casjens SR, Gilcrease EB, Winn-Stapley DA, Schicklmaier P, Schmieger H, Pedulla ML, et al. The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J. Bacteriol* 2005;187:1091–1104. [PubMed: 15659686]
11. Catalano, C., editor. *Viral genome packaging machines: genetics, structure, and mechanism*. Georgetown, TX: Landes Bioscience; 2005.
12. Comeau AM, Bertrand C, Letarov A, Tetart F, Krisch HM. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* 2007;362:384–396. [PubMed: 17289101]
13. Deveau H, Labrie SJ, Chopin MC, Moineau S. Biodiversity and classification of lactococcal phages. *Appl. Environ. Microbiol* 2006;72:4338–4346. [PubMed: 16751549]
14. Dobbins AT, George M Jr, Basham DA, Ford ME, Houtz JM, Pedulla ML, et al. Complete genomic sequence of the virulent *Salmonella* bacteriophage SP6. *J. Bacteriol* 2004;186:1933–1944. [PubMed: 15028677]
15. Droffner ML, Yamamoto N. Analysis of proteins induced by the *Salmonella typhimurium* Phage P221, a hybrid between serologically and morphologically unrelated phages P22 and Fels 1. *J. Gen. Virol* 1982;59:377–385. [PubMed: 7042911]
16. Dykhuizen DE. Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* 1998;73:25–33. [PubMed: 9602276]
17. Fauquet, C.; Mayo, M.; Maniloff, J.; Desselbergre, U.; Ball, A., editors. *Virus Taxonomy*. vol. VIII. Amsterdam: Elsevier Academic Press; 2005.
18. Filee J, Tetart F, Suttle CA, Krisch HM. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U S A* 2005;102:12471–12476. [PubMed: 16116082]
19. Fokine A, Leiman PG, Shneider MM, Ahvazi B, Boeshans KM, Steven AC, et al. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl. Acad. Sci. U S A* 2005;102:7163–7168. [PubMed: 15878991]
20. Ford ME, Sarkis GJ, Belanger AE, Hendrix RW, Hatfull GF. Genome structure of mycobacteriophage D29: implications for phage evolution. *J. Mol. Biol* 1998;279:143–164. [PubMed: 9636706]
21. Fujiwara A, Kawasaki T, Usami S, Fujie M, Yamada T. Genomic characterization of *Ralstonia solanacearum* phage ϕ RSa1 and its related prophage (ϕ RSX) in strain GMI1000. *J. Bacteriol* 2008;190:143–156. [PubMed: 17965167]
22. George DG, Yeh LS, Barker WC. Unexpected relationships between bacteriophage lambda hypothetical proteins and bacteriophage T4 tail-fiber proteins. *Biochem. Biophys. Res. Commun* 1983;115:1061–1068. [PubMed: 6226290]
23. Hashemolhosseini S, Stierhof YD, Hindennach I, Henning U. Characterization of the helper proteins for the assembly of tail fibers of coliphages T4 and lambda. *J. Bacteriol* 1996;178:6258–6265. [PubMed: 8892827]
24. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, et al. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet* 2006;2:e92. [PubMed: 16789831]
25. Hendrix, R.; Casjens, S. Bacteriophage λ and its genetic neighborhood. In: Calendar, R., editor. *The Bacteriophages*. 2nd Edition. New York City, NY: Oxford Press; 2006. p. 409–447.
26. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. USA* 1999;96:2192–2197. [PubMed: 10051617]
27. Hendrix RW. Bacteriophages: evolution of the majority. *Theor. Popul. Biol* 2002;61:471–480. [PubMed: 12167366]
28. Hendrix RW. Bacteriophage genomics. *Curr. Opin. Microbiol* 2003;6:506–511. [PubMed: 14572544]
29. Hendrix RW, Hatfull GF, Smith MC. Bacteriophages with tails: chasing their origins and evolution. *Res Microbiol* 2003;154:253–257. [PubMed: 12798229]
30. Hershey, A.; Dove, W. introduction to lambda. In: Hershey, A., editor. *The bacteriophage lambda*. Cold Spring harbor, NY: The Cold Spring Harbor Laboratory; 1971.

31. Hertwig S, Klein I, Schmidt V, Beck S, Hammerl JA, Appel B. Sequence analysis of the genome of the temperate *Yersinia enterocolitica* phage PY54. *J. Mol. Biol* 2003;331:605–622. [PubMed: 12899832]
32. Highlander SK, Weissenberger S, Alvarez LE, Weinstock GM, Berget PB. Complete nucleotide sequence of a P2 family lysogenic bacteriophage, var ϕ MhaA1-PHL101, from *Mannheimia haemolytica* serotype A1. *Virology* 2006;350:79–89. [PubMed: 16631219]
33. Jamalludeen N, Kropinski AM, Johnson RP, Lingohr E, Harel J, Gyles CL. Complete genomic sequence of bacteriophage ϕ EcoM-GJ1, a novel phage that has myovirus morphology and a podovirus-like RNA polymerase. *Appl. Environ. Microbiol* 2008;74:516–525. [PubMed: 18039824]
34. Jiang W, Li Z, Zhang Z, Baker ML, Prevelige PE Jr, Chiu W. Coat protein fold and maturation transition of bacteriophage P22 seen at subnanometer resolutions. *Nat. Struct. Biol* 2003;10:131–135. [PubMed: 12536205]
35. Kaiser AD, Jacob F. Recombination between related bacteriophages and the genetic control of immunity and prophage localization. *Virology* 1957;4:509–517. [PubMed: 13507311]
36. Kapfhammer D, Blass J, Evers S, Reidl J. *Vibrio cholerae* phage K139: complete genome sequence and comparative genomics of related phages. *J. Bacteriol* 2002;184:6592–6601. [PubMed: 12426348]
37. Kropinski AM. Sequence of the genome of the temperate, serotype-converting, *Pseudomonas aeruginosa* bacteriophage D3. *J. Bacteriol* 2000;182:6066–6074. [PubMed: 11029426]
38. Kropinski AM, Kovalyova IV, Billington SJ, Patrick AN, Butts BD, Guichard JA, et al. The genome of ϵ 15, a serotype-converting, Group E1 *Salmonella enterica*-specific bacteriophage. *Virology* 2007;369:234–244. [PubMed: 17825342]
39. Kwan T, Liu J, DuBow M, Gros P, Pelletier J. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc. Natl Acad. Sci. U S A* 2005;102:5174–5179.
40. Kwan T, Liu J, Dubow M, Gros P, Pelletier J. Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. *J. Bacteriol* 2006;188:1184–1187. [PubMed: 16428425]
41. Lawrence JG, Hendrix RW, Casjens S. Where are the bacterial pseudogenes? *Trends Microbiol* 2001;9:535–540. [PubMed: 11825713]
42. Lawrence JG, Hatfull G, Hendrix R. The imbroglios of viral taxonomy: genetic exchange and the failings of phenetic approaches. *J. Bacteriol* 2002;184:4891–4905. [PubMed: 12169615]
43. Mitchell MS, Matsuzaki S, Imai S, Rao VB. Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. *Nucleic Acids Res* 2002;30:4009–4021. [PubMed: 12235385]
44. Montag D, Schwarz H, Henning U. A component of the side tail fiber of *Escherichia coli* bacteriophage lambda can functionally replace the receptor-recognizing part of a long tail fiber protein of the unrelated bacteriophage T4. *J. Bacteriol* 1989;171:4378–4384. [PubMed: 2526805]
45. Morais MC, Choi KH, Koti JS, Chipman PR, Anderson DL, Rossmann MG. Conservation of the capsid structure in tailed dsDNA bacteriophages: the pseudoatomic structure of ϕ 29. *Mol. Cell* 2005;18:149–159. [PubMed: 15837419]
46. Nakayama K, Kanaya S, Ohnishi M, Terawaki Y, Hayashi T. The complete nucleotide sequence of ϕ CTX, a cytotoxin-converting phage of *Pseudomonas aeruginosa*: implications for phage evolution and horizontal gene transfer via bacteriophages. *Molec. Microbiol* 1999;31:399–419. [PubMed: 10027959]
47. Nilsson AS, Haggard-Ljungquist E. Evolution of P2-like phages and their impact on bacterial evolution. *Res. Microbiol* 2007;158:311–317. [PubMed: 17490863]
48. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. *Cell* 2003;113:171–182. [PubMed: 12705866]
49. Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 2001;409:529–533. [PubMed: 11206551]
50. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl. Acad. Sci. USA* 2000;97:10567–10572. [PubMed: 10954745]
51. Ravin V, Ravin N, Casjens S, Ford ME, Hatfull GF, Hendrix RW. Genomic sequence and analysis of the atypical temperate bacteriophage N15. *J. Mol. Biol* 2000;299:53–73. [PubMed: 10860722]

52. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol* 1991;222:67–88. [PubMed: 1942069]
53. Roberts, JW.; Roberts, CW.; Hilliker, S.; Botstein, D. Transcription termination and regulation in bacteriophages P22 and lambda. In: Losick, R.; Chamberlin, M., editors. *RNA polymerase*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory; 1976. p. 707-718.
54. Roberts MD, Martin NL, Kropinski AM. The genome and proteome of coliphage T1. *Virology* 2004;318:245–266. [PubMed: 14972552]
55. Rohwer F. Global phage diversity. *Cell* 2003;113:141. [PubMed: 12705861]
56. Scholl D, Kieleczawa J, Kemp P, Rush J, Richardson CC, Merrill C, et al. Genomic analysis of bacteriophages SP6 and K1-5, an estranged subgroup of the T7 supergroup. *J. Mol. Biol* 2004;335:1151–1171. [PubMed: 14729334]
57. Schwudke D, Ergin A, Michael K, Volkmar S, Appel B, Knabner D, et al. Broad-host-range *Yersinia* phage PY100: genome sequence, proteome analysis of virions, and DNA packaging strategy. *J. Bacteriol* 2008;190:332–342. [PubMed: 17965162]
58. Steinbacher S, Baxa U, Miller S, Weintraub A, Seckler R, Huber R. Crystal structure of phage P22 tailspike protein complexed with *Salmonella* sp. O-antigen receptors. *Proc. Natl. Acad. Sci. USA* 1996;93:10584–10588. [PubMed: 8855221]
59. Suttle CA. Viruses in the sea. *Nature* 2005;437:356–361. [PubMed: 16163346]
60. Szpirer J, Thomas R, Radding C. Hybrids of bacteriophages lambda and ϕ 80. A study of nonvegetative functions. *Virology* 1969;37:585–593. [PubMed: 4976535]
61. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
62. Walter M, Fiedler C, Grassl R, Biebl M, Rachel R, Herms-Parrado XL, et al. Structure of the receptor-binding protein of bacteriophage Det7: a podoviral tail spike in a myovirus. *J. Virol* 2008;82:2265–2273. [PubMed: 18077713]
63. Weigele PR, Scanlon E, King J. Homotrimeric, β -stranded viral adhesins and tail proteins. *J. Bacteriol* 2003;185:4022–4030. [PubMed: 12837775]
64. Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF, et al. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ. Microbiol* 2007;9:1675–1695. [PubMed: 17564603]
65. Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* 2000;289:2129–2133. [PubMed: 11000116]
66. Wollin R, Eriksson U, Lindberg AA. *Salmonella* bacteriophage glycanases: endorhamnosidase activity of bacteriophages P27, 9NA, and KB1. *J. Virol* 1981;38:1025–1033. [PubMed: 7017163]
67. Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev* 2000;64:69–114. [PubMed: 10704475]

Fig. 1. Genome maps of six tailed-phages that infect enterobacteria

Genomes of the six phages named on the left are shown with gene functions above (one function, like “Heads”, can indicate a cluster of genes that are involved in that function); functions in gray text are poorly documented. The maps are oriented with DNA packaging genes at the left end if the genomes are circularly permuted, and asterisks (*) indicate the virion chromosomes that have specific termini. Transcriptional direction (not RNA polymerase start or stop locations) is indicated below by gray arrows (known or putative early functions) and black arrows (known or putative late functions); NAEnz, nucleic acid or nucleotide metabolism genes; RNPase, RNA polymerase genes; att site, site of integration of prophage into host chromosome; pac site, packaging machinery recognition site where it is known. The accession Nos. for these genome sequences and their annotation are as follows: T1, AY216660; ϕ EcoM-GJ1, EF460875; T7, NC_001604; P2, AF063097; 933W, AF125520; Mu, AF083977.

	<i>Escherichia</i> <i>Shigella</i>	<i>Erwinia</i> <i>Klebsiella</i>	<i>Yersinia</i>	<i>Salmonella</i>
Lytic	T1/Rtp/TLS*/JK06*		PY100	
	φEcoM-GJ1			
	T7/T3/K1-5/K1E/K1F	Era103*(E)	φA1122/φYe203-12/Berlin*	SP6
	T4/RB32*/RB43/RB49/ RB69/JS98/φ1*			Det7 [†]
	T5			H8
	N4*			
				Felix-O1*
				KS7*/SETP3*
				9NA [†]
Temperate	λ/N15/21^/φ80 [@]			Gifsy-1
	HK97/HK022/φP27/SfV	φKO2(K)	PY54	ST64B
	HK620/Sf6/CUS-3			P22/ST64T/ST104/ε34 [#]
				ES18
	cdt-Iφ			Fels-1/Gifsy-2
	933W/VT2-Sa/stx1φ/stx2φ-I/ stx2φ-II/stx2-86*/Min27*			
	BP-4795			
	P2/186/φW*		L-413C	Fels-2/SopEφ/PsP3*
	Mu			
	P1			
	φV10*			ε15

Fig. 2. The tailed-phages of the *Enterobacteriaceae* with completely sequenced genomes

The seventy-three *Enterobacteriaceae* phages with completely sequenced genomes as of March 2008 are listed; all are known to be fully functional infectious phages except φKO2 which forms virions upon induction, but either they are defective or a susceptible host has not been found [8]. Thin lines separate phage types and subtypes as defined in the text; thick lines separate the lytic and temperate phages; the gray box in the temperate section highlights the “lambdoid” phages with subtypes on different lines, and in the lytic section highlights the phages with T1-like head genes (see text). Phages are listed below the bacterial genus they infect with slashes (/) separating their names (*Shigella* and *Escherichia* phages are indicated in the same column since these “genera” are not cleanly separated in evolutionary trees [50]); *, sequence deposited in GenBank but not yet published; †, S. Casjens, R. Hendrix and G. Hatfull, unpublished; ^, R. Hendrix and G. Hatfull, personal communication; @, G. Plunkett III, personal communication; # A. Kropinski, R. Villafane and S. Casjens, unpublished; (E), infects *Erwinia amylovora*; (K), isolated from *Klebsiella oxytoca*.

Fig. 3. Head gene maps of the seven lambdoid phage "head types"

Rectangles denote genes. Genes with similar known function or putative function as predicted from protein amino acid similarity are shown in the same color (which does not indicate sequence similarity); white genes have no known or predicted function. The putative protease/coat protein of the Gifsy-2 and BP-4795 genomes is predicted by sequence to contain these two functions, but this has not yet been proven in either case. At the right, the prototypical member of each group is indicated; other members with similar head genes are listed in Fig. 2. Black diamonds below a gene indicates that it is not present in all functional phage members of the subtype.

A neighbor-joining tree constructed by CLUSTAL X [61] is shown with bootstrap values (out of 1000) on each branch; horizontal branch lengths are proportional to extent of differences. At the right, λ -like coat protein categories I, II and III are indicated. The five fully functional phage coat proteins are indicated by “host species.phage name.gene name” and each prophage coat protein is denoted by its “host species.host strain name.gene number portion of the GenBank locus_tag”, except in the case of the *E. coli* strain ELD933 prophages [49] where the locus_tag number is replaced by the prophage name 933X or 9933O. Host bacterial species are abbreviated as follows: Ec, *E. coli*; Se, *Salmonella enterica*; Sf, *Shigella flexneri*; Sb, *Shigella boydii*; Ss, *Shigella sonnei*; Vh, *Vibrio harveyi*; Va, *Vibrio alginolyticus*; Bc,

Burkholderia cepacia. The asterisks (*) note that eight additional sequences that are identical to the Gifsy-1 coat protein have been reported in various *S. enterica* sequences, as have three additional sequences identical to the *E. coli* prophage 933O coat protein and these identical sequences are not shown in the figure; “trunc” indicates that the gene is truncated in the GenBank sequence.