Published in final edited form as:

J Mol Biol. 2009 April 24; 388(1): 48–70. doi:10.1016/j.jmb.2009.03.009.

The genome of *Bacillus subtilis* bacteriophage SPO1

Charles R. Stewart¹, Sherwood R. Casjens², Steven G. Cresawn^{3,6}, Jennifer M. Houtz³, Alexis L. Smith³, Michael E. Ford^{3,5}, Craig L. Peebles³, Graham F. Hatfull³, Roger W. Hendrix^{3,7}, Wai Mun Huang², and Marisa L. Pedulla^{3,4}

¹ Department of Biochemistry and Cell Biology, Rice University, Houston, TX 77251

² Pathology Department, University of Utah School of Medicine, Salt Lake City, UT 84112

³ Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260

Abstract

We report the genome sequence of *Bacillus subtilis* phage SPO1. The unique genome sequence is 132,562 bp, and DNA packaged in the virion (the chromosome) has a 13,185 bp terminal redundancy, giving a total of 145,747 bp. We predict 204 protein coding genes and five tRNA genes, and we correlate these findings with the extensive body of investigations of SPO1, including studies of the functions of the 61 previously defined genes and studies of the virion structure. 69% of the encoded proteins show no similarity to any previously known protein.

We identify 107 probable transcription promoters; most are members of the promoter classes identified in earlier studies, but we also see a new class that has the same sequence as the host sigma K promoters. We find three genes encoding potential new transcription factors, one of which is a distant homologue of the host sigma factor K. We also identify 75 probable transcription terminator structures. Promoters and terminators are generally located between genes and together with earlier data give what appears to be a rather complete picture of how phage transcription is regulated.

There are complete genome sequences available for five additional phages of Gram-positive hosts that are similar to SPO1 in genome size and in composition and organization of genes. Comparative analysis of SPO1 in the context of these other phages yields insights about both SPO1 and the other phages that would not be apparent from the analysis of any one phage alone. These include assigning identities and probable functions for several specific genes, and inferring evolutionary events in the phages' histories. The comparative analysis also allows us to put SPO1 into a phylogenetic context. We see a pattern similar to what has been noted in phage T4 and its relatives, in which there is minimal successful horizontal exchange of genes among a "core" set of genes that includes most of the virion structural genes and some genes of DNA metabolism, but there is extensive horizontal transfer of genes over the remainder of the genome. There is a correlation between genes in rapid evolutionary flux through these genomes and genes that are small.

Keywords

bacteriophage genomics; comparative genomics; phage SPO1; *B. subtilis* phage; SPO1 genome

⁷Corresponding author: Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, Phone: 412-624-4674, e-mail: rhx @pitt.edu.

⁴Present address: Montana Tech of the University of Montana, 1300 West Park St., Butte, MT 59701

⁵Present address: Cornell Medical School, New York, NY 10021

⁶Present address: Department of Biology, James Madison University, MSC 7801, Harrisonburg, VA, 22807

INTRODUCTION

Bacteriophage SPO1 is one of a group of large, lytic, tailed bacteriophages of *Bacillus subtilis*, which contain hydroxymethyluracil (hmUra) in place of thymine in their DNA^{1,2}. It was isolated in 1964 from soil near the Genetics Department of the University of Osaka, Japan³. This group includes *B. subtilis* phages SPO1, SP82, Φe, 2C, SP8, H1, Φ25, and SP5C, all of which are strikingly similar, but not identical, with respect to all parameters examined to date^{1; 4; 5}. In addition, five phage genomes from a wider range of Gram-positive hosts have been sequenced and have a significant level of similarity to SPO1 (*Staphylococcus* phages K, G1 and Twort, *Listeria* phage P100 and *Lactobacillus* phage LP65). SPO1 is by far the most intensively studied of all these phages, and it was the subject of pioneering studies in (1) the first demonstration of a cascade of sigma factors for regulation of a program of sequential transcriptional activation of gene expression⁶⁻⁹, (2) the first split gene discovered in Gram-positive bacteria, whose intron encodes a homing endonuclease that catalyzes the transfer of the intron to related genes lacking the intron^{10; 11}, (3) formation of concatemers from overlapping 13.2 kb terminally redundant sequences, which encode many of the proteins necessary for host-takeover¹²⁻¹⁴, and (4) the first example of a prokaryotic type II DNA binding protein encoded by a bacteriophage genome, which binds preferentially to DNA of that phage, causing DNA bending and facilitating diverse elements of the phage infection program, including the transition from middle to late gene expression^{2; 15}. In addition, SPO1's virion structure and morphogenesis have been studied extensively¹⁶⁻¹⁸.

As a tool for further understanding the biology of this wonderfully complex bacteriophage, we determined the complete nucleotide sequence of its genome. Among the new insights from the sequence are the following: (1) We identify three new putative transcription factors encoded by the SPO1 genome, each of a different type, and each offering promise of explanations for hitherto unexplained complexities in SPO1's program of gene activity. (2) Despite the similarity between the events of SPO1 infection and those of other large lytic phages, many SPO1 gene products show no detectable similarity to any other previously known protein, which emphasizes the extreme diversity among the extant tailed phages. (3) We clarify the phylogenetic position of SPO1 in relation to other large lytic tailed phages and to the more closely related large Gram positive phages with isometric heads and contractile tails. (4) We identify 58 predicted transcription units that cover the SPO1 genome, each terminated by a rho-independent transcription terminator, and initiated by one or more promoters of the three previously identified or the one newly-identified classes of promoters, accounting for transcription of all but three of the identified genes. (5) We identify genes specifying some, but not all, of the enzymes previously identified biochemically in studies of infection by SPO1 or hmUra-containing close relatives. (6) We show that early genes are unusually small, like those previously identified in the host-takeover module. Small genes are also over-represented among genes that are in relatively rapid flux through the genome. (7) By N-terminal sequencing, we identify the genes that encode four of the major structural proteins of the virion, including the major capsid protein, which is processed by removal of the N-terminal 23 amino acids.

RESULTS AND DISCUSSION

Characterization of the SPO1 Genome

Nucleotide sequence—The complete nucleotide sequence of the 132,562 bp SPO1 genome was determined; the DNA in the virion (the chromosome) has a 13,185 bp terminal redundancy, making the size of the linear virion DNA 145,747 bp. The sequences of several genome segments were determined previously^{9; 10; 13; 19-29}, some of which differ slightly from the present sequence. The G+C content of the SPO1 unique sequence is 40.0%, similar to the

43.5% of its *B. subtilis* host. The sequence has been deposited in GenBank, with Accession number FJ230960.

Predicted genes—Two hundred and four putative protein-coding genes and five tRNA genes were identified in the SPO1 genome, counting only once the thirty genes present in each copy of the terminal redundancy of the chromosome. These predicted genes are listed in SUPPLEMENTARY MATERIAL Table S1, and the genome map is shown in Figure 1. These 209 SPO1 genes are tightly packed in the genome; 45 genes overlap with the preceding gene by amounts ranging from 1 to 68 bp. Eighty-nine percent of the genome sequence is occupied by protein-coding genes, which is similar to, but a few percent lower than has been seen in other phages. Most intergenic spaces are occupied by putative transcriptional regulatory sequences (promoters and terminators), and almost all of the genome appears to be transcribed (see below). Promoters and terminators are more numerous in SPO1 than in most characterized phages, and this accounts for the slightly lower fraction of protein-coding sequence in SPO1. Gene names were assigned in order to retain, where possible, the numbering of previously identified genes^{2; 13; 30}, which were numbered 1 through 60, ordered from left to right on the genome, beginning at the left end of the non-redundant portion of the sequence. Other genes identified in the sequence were assigned decimal numbers, so as to maintain numerical order across the genome. This scheme preserves as much as possible the connections between the gene names and more than 40 years of experimental work on SPO1.

Essential Genes—Okubo *et al.*³⁰ isolated nonsense mutations in SPO1, sorted them into 36 complementation groups, and mapped them by recombination frequency, thereby defining genes 1 to 36. For many of these genes, several mutants were isolated, suggesting that a significant fraction of the essential genes in the SPO1 genome had been identified. Glassberg *et al.*³¹ searched extensively for temperature-sensitive and replication-deficient mutants, and all of these mutations were found to be in the genes already defined by Okubo *et al.*³⁰, supporting the notion that a large majority of essential genes have been identified. In addition, nonsense mutations have been introduced, by site-specific mutagenesis, into a number of SPO1 genes that had not been found in the searches for conditional lethal mutations. In every case except one, the nonsense mutation did not prevent plaque formation on non-suppressing strains, indicating that the encoded protein is not essential to SPO1 growth in the laboratory^{27; 32} (Stewart, C., Myles, B., Laughlin, L., Gupta, N., unpublished observations). The one exception we are aware of is the gene encoding TF1, in which nonsense mutations were so detrimental that they prevented growth even on suppressing strains³³, which accounts for the absence of TF1 mutants in the Okubo *et al.* collection. It thus appears that possibly as few as 20% of the SPO1 genes are essential. This may indicate either functional redundancy of genes (*e.g.*, gp44 and gp51, below), and/or that many SPO1 gene products are needed only under conditions different from those used in the laboratory. Table S2 lists these 37 essential genes. Eight of these genes had been sequenced previously, and thus could be located unequivocally in the genome sequence. With one exception, the other essential genes could not be located unequivocally in the genome sequence, so they were assigned decimal numbers as discussed in Materials and Methods. Table S2 shows, for each of the essential genes, which of the genes in the sequence could be its possible location.

Itinerant elements: introns and inteins—One of the earliest examples of a prokaryotic intron was discovered in gene 31 of SPO1¹⁰. While this group I intron in its DNA polymerase gene is the only apparent intron in SPO1, the set of six completely sequenced SPO1-related phages contains at least eighteen group I introns, of which fifteen are distinct sequences (these are listed in Table S3). Nine have been previously noted^{4; 10; 34; 35}. Including the SPO1 DNA polymerase intron, about half of the introns (eight of fifteen distinct cases) encode recognizable ORFs encoding putative homing endonucleases, mostly of the HNH endonuclease family^{36–}

³⁸. Six introns reside within DNA polymerase genes of four phages (SPO1, P100, with two identical intron pairs in G1 and K). The P100 DNA polymerase intron is clearly related to the first intron of G1 and K, although it lacks the internally encoded protein. Four related introns occur in the large terminase subunit genes, with two in Twort and one each in LP65 and P100. Three introns reside in ribonucleotide reductase genes of P100 and Twort³⁹, with two in the Twort gene. Three related small introns occur in the putative tail tube gene of Twort, Orf106 (also called Orf142)⁴⁰. Finally the lysin genes of K and G1 also contain an intron³⁴. Twort has the most with seven introns; P100, G1, and K have three each; while LP65 and SPO1 have only one each. Inteins are relatively rare in these phage genomes, with one example in SPO1 gene 34.33 and another appearing in Twort gene 6. Curiously, both genes encode a probable DNA helicase, but of different types.

tRNA genes—All of these completely sequenced SPO1-like phages contain multiple putative tRNA genes (Table S8). P100 has eighteen⁴¹, LP65 has seventeen⁴², SPO1, G1³⁴ and K³⁵ have five each, and Twort has two³⁴. The two tRNAs identified in Twort are quite similar to a pair of adjacent tRNAs in K and G1, which are however in a different position in the gene order in these phages. The tRNAs in SPO1 encoded by genes 2.32, 2.41 and 2.61 are closely related to bacterial tRNA genes, while the others make no significant database matches.

Gene product functions

Overview of analysis—We have employed a variety of strategies to determine the putative functions of the 209 SPO1 gene products, and the predicted assignments are listed in Table S1 and indicated in Figure 1. These include database comparison information, phenotypes of SPO1 mutants, and biochemical analysis of encoded proteins. In total, 64 genes are similar to genes in the extant sequence databases, and an overlapping set of 63 genes have been assigned experimentally determined or predicted functions; the functions of nineteen of these gene products (gp6.1, gp16.1, gp16.2, gp27, gp28, gp29, gp29.2, gp30, TF1, gp31, gp33, gp34, gp38, gp39, gp40, gp44, gp50, gp51, and gp56) have been studied experimentally, and the others have homology to proteins with known functions. (This list does not include those of the essential genes, discussed above, that were defined only by genetically mapped mutations that could not be located in a single gene in the genome sequence. The phenotypes caused by those mutations are listed in Table S2.) Thus 141 of the 204 SPO1 protein gene products (69%) show no similarity to any known protein. This high proportion of novel genes reflects the findings in other recently sequenced bacteriophage genomes^{43; 44} and supports the idea that bacteriophage gene discovery is still in its early stages⁴⁵. There is considerable clustering of genes with related functions in the SPO1 genome, but the clustering is far from absolute.

SPO1 head assembly genes—SPO1 virions are large and complex. They have an isometric T=16 icosahedral head 108 nm in diameter^{17; 46}, a contractile tail about 140 nm long, a hexagonal baseplate, and bushy tail fibers or spikes¹⁸. The complex virions may contain as many as 53 different types of protein molecules¹⁶. Yet mutations in only 21 of the genetically identified complementation groups affect assembly, and only eleven genes in the SPO1 genome encode proteins with detectable similarity to putative morphogenetic proteins of other phages (Table S1). Furthermore, we find that the virion assembly gene cluster described below contains only 32 genes; thus, even though some assembly genes may be outside this cluster, the earlier estimates of the number of virion structural genes seem likely to be somewhat high (we note the possibility that some of the virion proteins could be present in multiple post-translationally processed forms).

To identify some SPO1 virion protein genes unambiguously, we separated the proteins of purified virions by SDS polyacrylamide gel electrophoresis and attempted to determine the N-terminal amino acid sequence of five of these proteins. Figure 2 shows such a gel with the

amino acid sequences that were obtained; three sequences clearly identify genes *6.1*, *16.1*, *29.2* as encoding major virion components. A fourth sequence from a gel band of ~95 kDa gave a weak signal of NH₂-XXXNK. This matches four N-terminal sequences in the SPO1 proteome, but only the product of gene *16.2* has a large enough predicted mass (110 kDa) to account for the mobility of the gel band, with the other three possibilities ranging in size from 73 to 171 amino acids. We therefore tentatively identify gp16.2 as a virion component. A fifth protein of about 18 kDa did not give a signal, despite its high abundance, and it may have a blocked N-terminus.

The product of gene *6.1* is by far the most abundant protein in virions and so is certainly the major head shell (coat) protein; its T=16 geometric arrangement⁴⁶ predicts that the virion contains 955 coat protein molecules if five of the predicted subunits are replaced at the tail vertex, as is the case with other tailed phages. It is 26% identical to a protein (named Cps) similarly identified as the coat protein of *Listeria monocytogenes* phage A511⁴⁷, and these two have similar levels of identity to the annotated (putative) coat proteins of *Listeria* phage P100⁴¹, *Staphylococcus* phages Twort³⁴, K³⁵ and G1 (Kwan *et al.*, 2005), and *Lactobacillus* phage LP65⁴². *Leptospira* phage LE1 coat protein is recognizably similar but is somewhat less closely related⁴⁸. The SPO1, K and A511 putative coat proteins have N-termini that do not correspond to the beginning of the gene that encodes them, and the mature protein's N-terminal amino acid is encoded by codon 24 of the coat protein gene in all three cases. Parker and Eiserling¹⁶ previously concluded that the major coat protein of SPO1 is proteolytically cleaved after its synthesis, that about 2 kDa are removed from the primary translation product, and that like other cleaved phage coat proteins, the cleavage is dependent upon head assembly¹⁶. Our results agree in that 2326 Da are removed from the N-terminus. The cleavage is between residues K23 and A24 in SPO1 and residues K23 and S24 in K and A511. Comparison of their coat proteins suggests that, like SPO1, K and A511, the other five SPO1-like coat proteins likely also undergo a similar coat protein cleavage: all five have a potential cleavage site K-S sequence at amino acids 24 (P100, Twort, G1), 42 (LP65) or 51 (LE1); however their coat proteins have not been studied in this regard. Proteolytic removal of the N-terminus of phage coat proteins is common, and many phage coat proteins enjoy assembly-dependent cleavage during virion assembly (reviewed in ref. ⁴⁹). Such coat protein cleavages are typically performed by phage-encoded proteases, and interestingly, the SPO1 gene *3.2* product (and its homologues in P100, Twort, K, G1 and LP65) has sequence similarity to the N-terminal region of tailed phage procapsid proteases of the HK97 family; this corresponds to the catalytic domain of these proteases.⁵⁰ The SPO1-family protease gene is located two genes transcriptionally upstream from the coat gene, a typical location for such genes.

The similarity between the SPO1 coat protein and that of phage LE1 is of interest because the 73,623 bp LE1 genome^{48; 49} is about one-half the size of the SPO1 genome. The heads of the tailed phages are typically packed with DNA to similar densities⁴⁹, and the reported 85 nm diameter of the LE1 head⁵¹ would give it about half the internal volume of the SPO1 head. The observation that lattice constants for tailed phage capsid lattices are very similar in different phages, strongly suggests that SPO1 and LE1 have different triangulation (T) numbers⁵². Current evidence suggests that all tailed phage coat proteins *may* be homologous, including those of phages T4, HK97, P22, ϕ 15 and ϕ 29 (T=13 [prolate], 7, 7, 7, and 3 [prolate] respectively), since they have similar folds^{53; 54; 55; 56}, but it is not always possible to infer homology between these coat proteins on the basis of sequence alone. SPO1 and LE1 appear to provide an example of capsid protein sequences that have diverged recently enough that they retain detectable sequence similarity, and yet have evolved to assemble capsids with different T numbers. Other pairs of phage capsids with such a relationship are only recently coming to light; perhaps phages T4 and Syn9 provide the clearest example. The capsid protein sequences of these phages are 38% identical but the triangulation numbers are prolate T=13

and isometric T=16, respectively⁵⁷. The existence of such pairs of phages informs how we can think about the evolution of capsid size, and therefore genome size, in phages.

The head genes, and to a slightly lesser extent the tail genes, are strongly conserved among the six phages in the SPO1-like phages (Figure 3), and this means that features of the genes or of the gene organization may stand out because of their conservation even though they might escape notice in the examination of a single genome sequence. A surprising feature of the head gene region is that the genes for the head maturation protease and the major head subunit have sequence homology (are classified in the same sequence family, see below), a relationship that has not been seen in other phages. Examination of these sequences shows that there is a short region of rather weak sequence similarity between the two proteins, as shown in Figure 4a. This sequence similarity is slight enough that it would likely have been discounted if it had been detected in a single genome, but the relationship is conserved across all six of these phages, lending support to its significance. The part of the SPO1 protease sequence that matches most other phage maturation proteases does not overlap with the part of the sequence that is homologous to the head protein (Figure 4a). The part of the sequence that matches other proteases corresponds to the N-terminal portion of those enzymes, the locus of the catalytic activity⁵⁰. This is reminiscent of the case of the proteases of phages λ and Mu, which contain not the head protein sequence but the scaffolding protein sequence C-terminal to the protease sequence⁵⁸. It has been suggested in these cases that the shared sequence between the protease and scaffolding protein may facilitate assembly of the protease into the procapsid through homotypic interactions between protease and scaffolding protein⁴⁹. In an analogous way, we can imagine that the SPO1 protease might assure its assembly into the procapsid through homotypic interactions between the protease and the major head protein. A somewhat different similarity to the SPO1 case is seen in a sequence relationship found in phage T4 and some of its relatives (R. Hendrix and C. Georgopoulos, unpublished). Here a short sequence (in gene 61.1) matching the corresponding part of the head protein is found not as part of the protease protein but encoded as a separate short protein (Figure 4b). What the functional relationships are, if any, among these different examples of sequence relationships remains to be determined, but we note that comparative sequence analysis as described here is a productive method for bringing such relationships to light for further examination.

Portal proteins form the hole through which DNA enters and leaves the capsid, and the genes encoding portals in other phages are typically located immediately upstream of the procapsid protease gene⁴⁹. SPO1 gene 3.1, which is in this position, has convincing portal protein homology. Other than a large terminase subunit gene (see below), there are no other matches in the SPO1 genome to known head assembly genes; however, an approximately 28 kDa protein was shown to undergo assembly-dependent proteolytic degradation, and it is not present in virions¹⁶. We find that gene 3.3 lies in the position that is typical of scaffolding protein genes (between the portal and coat protein genes), and like other known scaffolding proteins gp3.3 (MW=26.5 kDa) is hydrophilic and predicted to have a high α -helix content; it is common for scaffolding proteins to be degraded by a phage-encoded protease⁴⁹. Thus, gene 3.3 seems likely to encode the head assembly scaffolding protein.

SPO1 tail assembly genes—Proteins from several substructures of the contractile tail of SPO1 have been analyzed¹⁶. It was concluded that there might be as many as 34 proteins in the tail. Fourteen of the genetically defined SPO1 genes (“7” through “20”) have an effect on tail assembly; these original genes all lie in the interval from gene 3.2 through 19.4 (Table S2). Twelve SPO1 gene products have similarity to tail proteins of other phages (Table S1). Most convincing is that gene 9.1 encodes the tail sheath protein. Although phage tail proteins are notoriously variable in sequence, sheath protein is one of the most conserved (presumably because it carries out a critical structurally-conserved rearrangement during DNA injection), and SPO1 gp9.1 has similarity to proteins in many known members of the *Myoviridae*,

including weak but convincing similarity to the well-known sheath proteins from phages T4, P2 and PBSX (14–16% identity over most of the protein in these cases). In addition, Parker and Eiserling^{16; 18} showed that the SPO1 sheath protein is about 60 kDa, and the sequence predicts gp9.1 to be 61.3 kDa. Gene *11.2* has similarity to known tapemeasure proteins (proteins that determine tail length) including 17% identity over 418 amino acids of the experimentally studied tapemeasure of phage TP901–1⁵⁹. We note that there is a reasonable correlation between the length of this protein and tail length in the known SPO1-like phages, ranging from tail lengths of 140 nm and 890 amino acids in SPO1 to 210 nm and 1351 amino acids in phage K⁶⁰. The ratios of these numbers imply ~1.5Å of tail length per amino acid of tapemeasure protein, in agreement with what has been seen for other phages^{61; 62; 63}, and so is consistent with the inference that the tapemeasure protein acts as an α -helix.

In phages with contractile tails, the tail tube subunit protein is typically encoded by a small gene following the tail sheath gene. SPO1 gene *10.1* fits this description, despite its lack of a sequence match to any experimentally verified tail tube subunit, and there are homologues at the corresponding position in all six SPO1-like phages (Figure 5). The observations that SPO1 gene *10.1* is in the expected position and has the expected size for a tail tube gene, that there are homologues in the corresponding positions of the other five phages, and that the other two SPO1 genes in this region that might be candidates for the tail tube gene can be assigned different, position-appropriate functions (see below), lead us to propose that SPO1 gp10.1 and its homologues in the other phages are the tail tube subunits. We suspect that this protein is the abundant ~18 kDa virion protein from which we were unable to obtain an N-terminal sequence (above).

In many phages, the tail tube subunit gene is followed by a pair of overlapping open reading frames whose expression includes a programmed translational frameshift^{64; 65}. This frameshift transfers a fraction of the ribosomes translating the upstream reading frame into the second open reading frame (ORF). The proteins produced have an essential role in facilitating assembly of the tail tube around the tail length tapemeasure protein. The frameshifting genes are typically transcribed immediately upstream of the tail length tapemeasure gene. In SPO1 there are just two ORFs (genes *10.2* and *11.1*) between the putative tail tube gene (*10.1*) and the tapemeasure protein gene (*11.2*) (Figure 5), and these overlap in such a way that a +1 frameshift would shift between them, so they are attractive candidates for the SPO1 frameshifting genes. There is in fact a potential +1 “slippery site” 7 codons upstream from the termination codon of the upstream gene where a tRNA^{Lys} reading AAA could slip forward to AAG (Figure 6). Because a +1 frameshift has a smaller signature in the sequence than the more common -1 frameshift⁶⁶, this identification is weak in the absence of corroborating information. However, the other five phages in the group also have two overlapping ORFs immediately upstream from their tapemeasure protein genes, in the same frame relationship as the SPO1 genes, and all have a similar potential +1 slippery site near the termination codon of the upstream gene (Figure 6). Another feature shared by all six phages is that the downstream ORF of the putative frameshifting pair lacks a convincing Shine-Dalgarno translational start sequence. This is striking in these phages which otherwise have Shine-Dalgarno sequences with a very strong match to consensus; this buttresses the contention that ribosomes enter these ORFs by a mechanism other than standard initiation. We conclude that all of these phages very likely use the frameshift mechanism with a +1 frameshift between the two genes identified in this analysis. The comparative analysis is especially helpful for phages K, G1, Twort, and LP65, since these phages have insertions of one to several extra genes between their putative tail tube genes and frameshift genes.

SPO1 gene *12.1* protein and its LP65 homologue have a domain (from about amino acid 400 to 600) that has similarity to domains that are thought to be involved with cell wall binding; searches with the parallel domains of Twort or K yield proteins that contain lysozyme/

peptidase/amidase domains; such proteins have been reported to be part of the tail's injection machinery in several phages⁶⁷. Genes *14.1* through *16.1* encode proteins that have weak but convincing similarity to phage baseplate and tail fiber proteins, including gpW and gpJ of phage P2 and gp8 of T4. The product of gene *16.2* has similarity to proteins of unknown function in the tail regions of some other phages, and we gave evidence above that it is likely a virion component.

It is not clear which SPO1 genes encode the brushy tail fibers or spikes, but there are several possible candidates. Gene *17.2* and *18.1* proteins have short patches of similarity to other phage tail fiber proteins, in particular to phage ϕ 29 tail appendage (gp12) protein and the tail fibers of Mu-like phages, and the products of genes *18.3* and *19.1* have longer regions of similarity to ϕ 29 tail appendage protein. Some phage tailspikes have the ability to cleave bacterial surface polysaccharides, and these proteins have an unusual β -helix protein fold⁶⁸. We find that both gp18.3 and gp19.1 have very high β -helix predictions over most of their lengths (estimated at the <http://betawrap.lcs.mit.edu/> website⁶⁹). Gp19.1 has strong similarity to phage SPP1 gp33 and to a phage SP β protein. In addition, the products of genes *2.43*, *2.8*, *12.1*, and *34.16* show sequence similarity to proteins known to bind to cell walls and/or to cleave macromolecules that are components of cell walls (Table S1; gene *12.1* appears to have a particularly variable domain structure among the SPO1-like phages, see Figure S1), but it remains unknown at this time whether these are virion components involved in cell attachment and/or DNA entry, important for cell lysis, or have other functions.

SPO1 terminase and mature DNA ends—The 145,747 bp, non-permuted DNA molecule of the SPO1 virion has a very long 13,185 bp direct terminal redundancy (TR). Since the concatemers that result from SPO1 DNA replication have only one copy of the TR between flanking copies of the non-TR part of the genome sequence, the SPO1 terminal redundancy must be generated in concert with DNA packaging, but the mechanism by which it is generated remains unclear^{12; 13; 70; 71}.

Phage terminases are the enzymes that recognize and cleave the DNA concatemer to virion chromosome length and are a critical part of the motor that inserts the chromosome into the preformed procapsid during packaging⁷². DNAs within the particles of tailed phages have a variety of different end structures (reviewed in ref. ⁷³), and the large subunits of the terminases that create similar DNA end structures typically form related sequence groups that correlate with the types of DNA ends they create⁷⁴. The putative large terminase subunits of SPO1 (encoded by gene *2.11*), K, LP65, G1, Twort, A511 and P100, when compared to other phage terminases in a neighbor-joining tree, form a single well-separated branch which has very high bootstrap support (the large terminase subunits have 30–40% identical amino acid sequence within this group; Figure S2). The seven phages listed above have other similarities (below) including genome sizes, and this terminase similarity suggests that all these phages' virion DNAs may have similar long, terminally redundant end sequences. This has in fact recently been demonstrated experimentally for phages A511, P100 and K⁷⁵. We note that phage LE1 has a terminase gene that does not fall within the SPO1 group (see below), even though some of its other head and tail assembly genes are similar. The putative large terminase subunits of *E. coli* phage T5 (encoded by its gene *144*)⁷⁶ and *Bacillus thuringiensis* phage 0305 ϕ 8–36 (encoded by its gene *117*)⁷⁷, which have 10,139 bp and 6,479 bp long terminal redundancies, respectively, are not close relatives of the terminases of the SPO1 group. Thus, there appear to be multiple types of terminases that generate long terminal repeats, just as there are multiple types that create some other end structures⁷⁴.

SPO1 virion assembly gene arrangement—Most of the genes with identifiable roles in structure and assembly of the virion map in a contiguous block of 32 genes bounded on the left by gene *3.1* (portal) and on the right by gene *19.1* (possible tail fiber). As with many other

phages^{49; 78}, these genes are not only clustered separately from genes with non-assembly functions, but they are grouped within that cluster into head genes (on the left) and tail and tail fiber genes (on the right). Within those groupings we are able to recognize several features of gene order that are widely conserved among tailed phages, most notably the portal–protease–scaffolding–coat genes in the head gene region and the sheath–tube–chaperones–tape measure genes in the tail gene region. This section of the genome is known to be transcribed late in the infection cycle, as expected by analogy with other tailed phages^{14;79}. However, this region does not encompass all of the virion protein genes. A clear example of an assembly gene outside the main cluster is gene *2.11*, encoding the large terminase subunit, which appears to have been separated from the main cluster by a ~15 kb insertion (discussed below). Mutants in genetically defined genes *1* through *20* are deficient in virion assembly^{2; 80}, and this implies an assembly gene, corresponding to original gene *1*, to the left of the large terminase subunit gene, and another corresponding to original gene *20*, beyond the right end of the main assembly/structure cluster. Further, we found that gene *29.2* encodes a 19.8 kDa protein that is a fairly abundant component of the virion (above). The role of gp29.2 is not known, but in T4, the only large phage where such proteins have been characterized in detail, decoration (external surface accessory) protein and internal protein genes are not clustered with the rest of the morphogenetic genes. SPO1 does appear to have a decoration-like protein on its surface that is present in fairly large numbers⁴⁶, and it could be gp29.2. We note that it is common for the presence of decoration protein genes to be variable in closely related phages⁸¹, so it is not surprising that the other SPO1-like phages do not have a recognizable gene *29.2* homologue. Related phage LP65 also encodes a major virion protein outside the head and tail gene clusters, which does not have a homologue in SPO1⁴². Finally, a mutation in original gene *35* causes a defect in virion assembly²; it may correspond to gene *35.1*, *36.1*, *36.2* or *36.3*. These genes are located in a putative late transcription unit, suggesting that one or more of them may also specify a virion structural protein.

Lysis—Genes *19.2* and *19.3* specify proteins with sequence similarity to holins and lysins, respectively, of other phages that infect Gram-positive bacteria. The putative lysin, gp19.3, also makes sequence matches to a number of cell wall hydrolyzing enzymes of Gram-positive bacteria. The patterns of sequence matches with all of these enzymes argue that the N-terminal portion of gp19.3, roughly the first 160 amino acids, has the cell wall hydrolytic activity and the remainder, amino acids ~160 to 336, forms a peptidoglycan binding domain. A mutant in original gene *19* is deficient in both tail formation and production of lytic enzyme^{2; 30}. Possible locations of gene *19* mutations are in *19.1–19.5*, but *19.5* is unlikely since it specifies a putative helicase. If gp19.3 were also a tail component, that might account for the deficiency of the gene *19* mutant in both lytic enzyme and tail formation. Possible alternatives involve polar effects in the *19.1–19.4* region.

DNA Metabolism—The SPO1 genome is replicated from at least two origins, one in the 105–108 kbp region, the other to the left of about 45 kbp⁸². The specific origin sequences have not been identified, and there is no sequence in the SPO1 genome that is very similar to the *B. subtilis* replication origin or termination sites^{83; 84}.

Conditional lethal mutations in nine original genes, *21*, *22*, *23*, *27*, *28*, *29*, *30*, *31*, and *32*, completely prevent SPO1 DNA synthesis and plaque-formation under restrictive conditions³⁰. The gene *28* mutation has its effect by preventing expression of middle genes, which include genes specifying the replication enzymes. The gene *27* mutation also prevents expression of certain late genes, and it is not established whether either of its effects is indirect.

The products of SPO1 original gene *21* (located in any of genes *19.5–21.95*) and original gene *32* (located in *32.1*, *32.2* or *32.4*) are necessary for initiation of replication (and gene *32* also for resolution of concatemers)⁸⁵. Original gene *21* most likely corresponds to gene *21.1*, since

gp21.1 has moderate similarity to phage T4 helicase gp41, which is an essential component of the T4 replication machinery. There is no functional information about any of the three genes in which the original gene 32 mutation might be located.

The products of genes 22, 30, and 31 are required for the elongation stage of SPO1 DNA synthesis⁸⁵. A gene 22 mutation also prevented expression of a subset of late genes⁸⁶, possibly as an indirect effect of the replication block. Possible locations of the original gene 22 mutation are in genes 21.9, 21.95 or 22.1. Gene 21.95 is most likely, since 22.1 has been tentatively assigned to original gene 23 (below), and gp21.9 shows sequence similarity to a protein needed for recombination and not for replication; gp21.95 shows sequence similarity to primases known to participate in replication. Nothing further is known about the activity of gp30, which has no recognizable sequence similarity to known proteins. Finally, gene 31 encodes a protein shown to have DNA polymerase activity *in vitro*⁸⁷, with similarity to the DNA polymerase and 3'-5' exonuclease domains of *E. coli* DNA polymerase I, plus a domain similar to a bacterial uracil-DNA glycosylase. We suppose that the latter domain may be involved in discriminating between the hmUra-containing SPO1 DNA and thymine-containing host DNA. Also, gp32.85 is homologous to the 5'-3' exonuclease domain of *E. coli* DNA polymerase I.

Enzymes involved in synthesis of hydroxymethyluracil nucleotides include dCMP deaminase, dUMP hydroxymethylase and HMdUMP kinase. Gp21.8 is about 30% identical to Gram-positive bacterial dCMP deaminases, and no other SPO1 gene is related to dCMP deaminases; but we note that gp21.8 is about 25% smaller than those bacterial enzymes. The SPO1 dCMP deaminase is known to be expressed at middle times⁸⁸, whereas only an early and a late promoter (see below) are in the position to transcribe gene 21.8. Gene 29 encodes dUMP hydroxymethylase (deoxyuridylate hydroxymethyl-transferase), as shown experimentally by Okubo *et al.*³⁰ and Wilhelm and Ruger²⁸. Gene 23 likely encodes HMdUMP kinase, since the original gene 23 mutant failed to complement an SP82 mutant that is deficient in that enzyme⁸⁹ (Stewart, C. R. and Franck, M., unpublished results). Possible locations for this mutation are gene 22.1 or 23.1. Gp23.1 shows no recognizable similarity to any known protein; however, gp22.1 has moderate (up to 30%) amino acid sequence identity to bacterial dUTPases (deoxyUTP pyrophosphatases) and about 30% identity to T5 dUTPase^{76; 90}. It seems possible that gp22.1 is actually an HMdUMP kinase, and its similarity to dUTPases arises because the reactions that the latter catalyze are similar to the reverse of the HMdUMP kinase reaction. Thus we believe that gene 22.1 most likely encodes the HMdUMP kinase.

Proteins involved in diminution of the thymidine nucleotide pool, which have been observed in infection by SPO1 or other closely-related hmUra-containing phages, include dTTPase/dUTPase, dTMPase, and an inhibitor of thymidylate synthetase^{2; 91}.

Since intensive searches have failed to find additional genes with conditional lethal, replication-deficient mutations, it is likely that at least most other SPO1 gene products are not essential for DNA replication. Several functions that might be expected to be required for DNA replication have not been identified in our analysis of the SPO1 genome. None of the proteins specified by SPO1 shows significant similarity to any single-strand DNA binding protein, sliding clamp, or sliding clamp loader, all of which are specified by the genome of bacteriophage T4, and participate directly in T4 DNA replication (Ref. 92 and references therein), nor does any SPO1 gene specify a DNA topoisomerase. These functions are likely either performed by novel phage proteins or by host proteins.

The products of genes 21.3 and 21.9 are similar to phage T4 gp47 and gp46, respectively, which are subunits of its recombination nuclease, and to various other nucleases involved in DNA repair and/or recombination. Mutations inactivating the corresponding T4 genes are recombination-deficient⁹², but no similar studies have been done with SPO1.

Other genes whose putative products might plausibly be involved in replication or other aspects of DNA metabolism, but which have not been shown to be essential for DNA replication, include genes 2.9 and 4 (ribonucleoside diphosphate reductase alpha and beta subunits, respectively), 19.5 (helicase), 32.85 (exonuclease), 32.95 (DNA ligase), and 34.33 (UvrD/Rep family helicase). The two ribonucleoside diphosphate reductase genes have sequence affinities that suggest an evolutionary history different from those of the homologous genes of other phages in the SPO1-like group. Curiously, a nonsense mutation unequivocally located in gene 4, based on marker rescue activity of restriction fragments⁹³, was deficient in head formation². Since that is an unlikely effect of ribonucleoside diphosphate reductase deficiency, especially since the host genome should also provide this activity, this surprising phenotype may be due to polar effects on downstream gene(s) 5.1 and/or 5.2.

Host takeover—SPO1 infection directs a major remodeling of the host cell, subverting its biosynthetic machinery to the purposes of the phage. Many of the 24 genes of the “host-takeover module,” which occupies most of the terminal redundancy, have been identified as playing major roles in redirecting metabolism to phage production^{13; 27; 32; 94}. Specific gene products have been shown to: inhibit host cell division; shut off host DNA and RNA synthesis; regulate the timing of those shutoffs; and regulate the expression of the genes causing host-takeover. The genes of the host-takeover module (almost all of which are early genes) are unusual in that they are smaller than average and have promoters and ribosome-binding sites characteristic of very highly expressed genes. In SUPPLEMENTARY MATERIAL, we discuss these unusual features in comparison with other groups of genes.

Transcription of the SPO1 genome

Promoter—Sequences characteristic of SPO1 early, middle, and late promoters were identified as described in Materials and Methods. Since some genes had either no recognizable promoter, or only a promoter that was highly divergent from the consensus, and since SPO1 gp2.21 showed sequence similarity to *B. subtilis* sigma K, we searched for sigma K promoter-like sequences upstream of each such gene and found several. The positions of all of these 107 putative promoters are shown in Figure 1 and Tables S4 and S6, and their sequences are listed in Table S4. Among these are 16 early, 21 middle, and 5 late promoters whose activity had been demonstrated experimentally *in vivo* and/or *in vitro* 9; 23; 26; 29; 32; 95; 96.

Transcription terminators—Figure 1 and Tables S5 and S6 indicate the positions of 75 putative rho-independent transcription terminators that have been identified in the SPO1 genome, including 15 that had been identified previously^{13; 97}, and their sequences are shown in Table S5. Most of these terminators are positioned to cause termination shortly after the end of a gene, and, in most cases, there is at least one promoter overlapping, or just downstream of, the terminator, placed so it would begin transcription before the next gene. As noted in SUPPLEMENTARY MATERIAL, several of these putative terminators have unusual features, such as exceptionally long loops in their predicted stem-loop structure.

Transcription units—Figure 1 and Table S6 show the assignment of the above promoters and terminators to 58 transcription units, using the criteria discussed in SUPPLEMENTARY MATERIAL. Since many have multiple promoters, these “units” may include several lengths of mRNA. Only four genes, 1.4, 2.43, 5.9, 28.2, are not included in the transcription units identified in this way. Transcription of these genes must depend on either leakage through the preceding terminator or initiation at unrecognized promoters.

Newly identified transcription factors—The SPO1 genome sequence suggests that three new transcriptional regulatory factors, beyond those previously studied, may be encoded by

the SPO1 genome. This is not surprising, since the previously known transcription factors were insufficient to account for all of the complexities of SPO1's transcription program.

(i) Possible new sigma factor: SPO1 has been unusual among characterized bacteriophages in that it is known to encode two sigma factors: the product of gene 28, which directs the transcription of middle genes, and the products of genes 33 and 34, which together direct the transcription of late genes⁸. The nucleotide sequence reveals a third apparent sigma factor, gp2.21, which shows homology to sigma factor K of *B. subtilis* and other bacteria. The sequence also shows sequences identical to those of sigma K promoters, located at positions such that the initiation of transcription at those positions, possibly by the action of gp2.21, could account for several experimentally observed transcripts that cannot presently be explained by the known sigma factors and their cognate promoters. These considerations are discussed in detail in SUPPLEMENTARY MATERIAL.

(ii) Possible new chromatin-like DNA-binding protein: SPO1 has also been unusual among characterized bacteriophages in that it is known to encode a chromatin-like DNA-binding protein. This protein, called TF1, binds preferentially to double-stranded hmUra-DNA, such as SPO1's, causes DNA bending¹⁵, and is required for shutoff of expression of certain middle genes, and for activation of expression of certain late genes³³. The genome sequence identifies a second chromatin-like DNA-binding protein, which, to the best of our knowledge, is unprecedented among bacteriophages. SPO1 gp34.25 shows very high identity to α/β -type small acid-soluble spore proteins, including the *B. subtilis* *spsB* gene product. These proteins bind to spore DNA, modifying its conformation, and their effects are modulated by HBSu, the major host chromatin protein, which is homologous to TF1 (Ross and Setlow, 2000). Thus, it is plausible that gp34.25 might bind to SPO1 DNA, modifying its structure and, thereby, its gene expression. In SUPPLEMENTARY MATERIAL, we discuss how gp34.25 might account for a previously unexplained gene regulatory event.

(iii) Putative gp44/gp51-type transcription factor: The products of SPO1 genes 44, 50 and 51 are required for the normal transition from early to middle gene expression, and for other gene regulatory events³². A double mutation inactivating genes 50 and 51 has the same effect on gene expression as a single mutation inactivating gene 44, and gp44 and gp51 are 46% identical over a 117 amino acid segment that constitutes most of gp51. The genome sequence reveals a third related gene. Gp25.1 has 34–35% identity to gp44 and gp51 over most of the length of each of those proteins, suggesting that it might also play a role in regulating transcription. In SUPPLEMENTARY MATERIAL, we discuss how gp25.1 might account for the currently unexplained delay in the onset of delayed-early gene expression.

An SPO1-like superfamily of phages: evolution and relationships to other phages

The large lytic bacteriophages—The relationships among the large, lytic tailed phages are complex, and many aspects of their life cycles are not well understood. Among such phages, the genomes of a number of phages similar to the well known *E. coli* phage T4 (166 kbp) have been sequenced, and these phages form the best characterized group of related, large phages, termed the “T4-like superfamily”^{98–105}. To date, sequenced phages in this group infect hosts in the Proteobacteria and Cyanobacteria. About 45 “core genes” are found in all members of the superfamily, each of which has about 300 genes¹⁰³. (The exact number of core genes, so defined, can vary somewhat depending on the phylogenetic breadth of the group of phages being considered). The core genes encode mostly proteins that are required for virion assembly and DNA replication. Genome comparisons and phylogenetic analyses have shown that the core genes experience little or no horizontal exchange among phages in this group; it is likely that exchange of genes encoding parts of complex protein machines is limited by the intimate interactions among the parts^{98; 106}. In contrast, the remainders of their genomes are occupied

by genes that appear to exchange in and out of the genomes frequently, as shown by the fact that in most cases they occur in only one member or a subset of the most closely related members of the group. These exchangeable genes most likely confer a selective benefit on the specific phages in which they occur, as for example the photocenter genes of some of the cyanophages^{57; 107; 108}, but in most cases the functions of these accessory genes are not known, and the possibility that some of them confer no selective advantage cannot be ruled out.

There are no doubt many other such superfamilies of large tailed phages which remain to be elucidated with further analysis of phage diversity. For example, such a group may be represented by the two *Pseudomonas aeruginosa* phages, ϕ KZ¹⁰⁹ and EL¹¹⁰, whose genomes are over 210 kbp in length and whose related coat protein sequences (for example) are not recognizably similar to those of T4 or SPO1. The coat proteins of *Yersinia enterocolitica* phage ϕ R1–37¹¹¹ (genome 270 kbp), *B. thuringiensis* phage 0305 ϕ 8–36⁷⁷ (218 kbp) and *Thermus thermophilus* phage ϕ YS40¹¹² (152 kbp) are each not obviously similar to any other phage coat protein and so could be prototypical of other such groups. We argue below that SPO1 and its relatives constitute a newly identified superfamily analogous to the T4 superfamily.

Similar gene organization in the SPO1 superfamily—The current database contains complete genome sequences for five phages that have substantial similarity to SPO1 – *Listeria* phage P100⁴¹, *Staphylococcus* phages Twort³⁴, K³⁵ and G1³⁴, and *Lactobacillus* phage LP65⁴². As this report was being written, the genome sequences of two SPO1-like phages were reported, the 142 kbp genome sequence of *Enterococcus* phage ϕ EF24C¹¹³ (Accession No. NC_009904) and the 134 kbp genome of *Listeria* phage A511⁷⁵ (Accession No. DQ003638). These genomes were not included in our comparative analysis with the exception of information previously available for A511⁴⁷. Other phages that likely fit into this group, as judged by genome sequence similarities, are *Staphylococcus* phages ϕ 812, SK311, U16 and ϕ 131¹¹⁴. Comparisons among these genomes place SPO1 in a phylogenetic context and give examples of evolutionary events in the past history of these phages. Such comparisons also reveal which genes are most strongly conserved among members of the group of phages and therefore are most likely to have functions critical to their lifestyle.

To facilitate these comparisons we used the computer program Phamerator, which will be described in detail elsewhere (S. Cresawn, in preparation); in brief, it carries out all pairwise comparisons among the amino acid sequences of the annotated protein-coding genes of (in this case) the six completely sequenced SPO1-like phages and classifies them into related groups called “Phamilies” that share significant sequence similarity as determined by the BLASTp and CLUSTAL W algorithms. This analysis calculates that these six phage genomes encode 611 different protein types (Phamilies) among the 1,114 genes; the members of each Phamily are listed in Table S7. Of these Phamilies, 444 have only one member (*i.e.*, the gene is unique among the six phages), 167 have more than one member, and 12 have seven or more genes. In the last category, this analysis identifies a few Phamilies of proteins that contain members which appear on other grounds not to be orthologous — for example, they differ substantially in size or domain structure, or they have numerous members at different locations. The major examples of such complex Phamilies are discussed in more detail in SUPPLEMENTARY MATERIAL (see also Figure S1). Manual inspection of possible orthologies within these complex Phamilies was performed in all cases, and we believe this does not confound any conclusions drawn below.

The physical maps of the six putative SPO1-like phages are aligned on the major capsid genes (gene 6.1 for SPO1) in Figure S3. It is clear from this alignment that their gene orders have very significant similarities. For SPO1, 26% of its protein-coding genes (53 out of 204) have homologues in one or more of the other phages in the group. Pairwise comparisons show that the different phages have between 27 and 141 gene Phamilies in common (Table 1). All six

phages share 26 genes (manually parsed from 21 Phamilies, see above and SUPPLEMENTARY MATERIAL) in common, and nearly all of these homologous genes have similar organization within the SPO1-like phage group. The genes that are shared by all or most of the phages are concentrated over somewhat more than half the length of the genome and consist primarily of morphogenetic (head, tail, tail fiber) genes and DNA metabolism genes. Additional genes are present in each genome that match a smaller fraction of the six genomes, and each genome contains genes that match none of the other five phages (indicated in Figure S3 by white gene boxes). Table 1 also shows that the six SPO1-like phages are not equidistantly related to one another; G1, Twort and K have more genes in common with each other than they do with the others. This closer relationship is perhaps not surprising, since these three phages infect the same *Staphylococcus* host.

The SPO1-like phages clearly do not fall into the only previously defined superfamily of large lytic phages, the T4 superfamily. Their gene order and gene relatedness characteristics are well outside the boundaries of the T4 superfamily of phages. For example, there is little to no statistically significant sequence similarity between the SPO1-like phage coat proteins and the T4-like coat proteins; the most divergent coat proteins within the T4 superfamily are 20–25% identical, as are the most divergent members of the SPO1-like group (SPO1 and LP65 coat proteins are 24% identical). Among the SPO1 genes, only gene *21.1* has its closest relatives in the T4 superfamily (and they are weak matches at <27% identity). Thus, these two groups appear, at our present state of understanding, to form largely self-contained and non-overlapping tailed phage types, which nonetheless each occupy similarly wide spans of diversity. In addition, we have found no close homologies or overall organizational similarities between SPO1-like and any other large phage genomes. Thus, SPO1-like phages clearly form a definable superfamily; they have similar genome sizes and virion morphologies, and they share many more homologous genes (in largely the same organization) with other members of the group than they do with any other phage whose genome sequence is known.

Intra-SPO1 superfamily diversity—The six SPO1-like phages discussed above occupy a substantial span of diversity in spite of the fact that we group them into a single superfamily. This diversity is manifest at three levels – sequence *per se*, gene content and gene arrangement. First, individual genes have different nucleotide (and encoded protein amino acid) sequences due to simple sequence divergence. We do not discuss this divergence in detail, but note that (for example) for the two most divergent members, SPO1 and LP65, the coat proteins are only 24% identical. This extent of divergence is typical between SPO1 and LP65 orthologues, suggesting that the SPO1 phage superfamily has existed and been undergoing divergence for a long time. Second, as discussed above, there are very significant differences in gene content among the superfamily members; indeed, a majority of genes are different between the more distantly related members.

On average, the SPO1 genes that have matches in one or more of the other five superfamily members are larger than those with no matches. Figure 7 shows the size distributions of the SPO1 genes with and without such matches as a function of gene size. The genes without matches are strongly skewed toward smaller gene sizes. Of the 141 SPO1 genes with 200 or fewer codons, 23 (16%) match at least one gene in the other five phages, while 37 of the 61 genes larger than 200 codons (61%) make such a match. The genes without matches in the other phages are, we suggest, genes that exchange in and out of the phage genomes on an evolutionarily rapid time scale, since a gene uniquely present in SPO1, for example, was either absent from the last common ancestor of the group of phages and acquired more recently by SPO1 or present in the common ancestor and lost more recently in the other phages. Thus we conclude that the genes that exchange on a rapid evolutionary time scale are smaller on average than those with more permanent residence. (A caveat for this sort of analysis is that the method for detecting matches must not itself have a length bias or it can introduce a result such as that

shown in Figure 7 artifactually. The BLAST algorithm does have such a length bias, but PHAMERATOR also uses CLUSTAL W in such a way—scoring percent amino acid identity in pair-wise alignments—that we believe it should identify most pairs without substantial length bias). As noted in SUPPLEMENTARY MATERIAL, there is also a rough correlation between small genes and genes that are transcribed early.

Third, a few genome rearrangements can be recognized when the genomes are compared. The great majority of genes that match between any two genomes in the SPO1 superfamily have similar locations, but there are examples of genes that have moved in one genome relative to another. An example of such a difference is seen in a cluster of genes corresponding to SPO1 genes *12.1*, *13.1*, *13.2* and *14.1*. They are present in the same positions in SPO1, K, G1, Twort and P100, but in LP65 they are in a position about 36 kbp leftward from their homologues in the other phages and have suffered an inversion as well as a permutation of order.

There are also several examples among these phages of apparent, large, relatively recent (non-intron) insertion/deletion in one phage relative to another. These features are especially well-defined in a comparison between phages K and G1, since they are more than 99% identical in nucleotide sequence over 93% of the phage K genome. Against this background, 233 bp from the middle of phage K gene *108* are replaced in phage G1 by 7,381 bp; these 7,381 bp contain 22 predicted small genes, only four of which match genes in the other phages. These 22 genes fit the pattern described above for genes that exchange in and out of the genomes rapidly, but it is not clear whether they represent a recent addition to the G1 genome or a recent subtraction from the K genome.

The most clearly defined apparent insertion in the SPO1 genome is between the terminase and portal genes (genes *2.11* and *3.1*); we suggest that this feature illustrates how SPO1 has acquired novel functions in its recent evolutionary past. In phage LP65 the homologous terminase and portal genes are adjacent, as in many phages, but in SPO1 these genes are separated by 14,929 bp of sequence which contains fifteen predicted protein-coding genes and the five SPO1 tRNA's. (Similarly, phages Twort, K, G1 and P100 have related 3–5 kbp sequences in this interval that are different from the SPO1 sequence.) The SPO1 protein-coding genes in this interval include four that have inferred functions based on sequence matches to the database but that do not make sequence matches to any of the other phages in this comparison group. These genes include the putative σ^K -like transcription factor gp2.21, the ParB-like (gp2.12) and PhoH-like (gp2.62) proteins, and a protein (gp2.52) with sequence similarities to the “Hep-Hag family” of bacterial outer membrane carbohydrate-binding proteins. This stretch of putatively inserted DNA in the SPO1 genome might therefore plausibly be inferred to be responsible for some phenotypic properties of SPO1 that differentiate it from other members of the superfamily. A striking feature of the eight protein-coding genes in this region that make significant matches to sequences in the databases is that seven of them make their strongest (and numerous) matches to bacterial sequences; the same is true of three of the five tRNA sequences. Matches to phage sequences, when present, are less strong than the matches to bacterial genes, with the one exception of gp2.43, which also has very strong bacterial matches. These observations argue that much of this fifteen kbp segment of SPO1 DNA was most likely acquired in SPO1's ancestry primarily from bacterial rather than phage sources. However, the fact that there are eleven phage-specific promoters annotated in this region argues that this sequence has adapted to become a functional part of the phage genome in the time since its entry into the phage universe.

Relationships among large phage superfamilies—The overall pattern of relationships among the genomes of SPO1 and its relatives discussed here is reminiscent of what has been described for the T4 superfamily of phages (above). Our analysis of the SPO1 superfamily, as well as preliminary phylogenetic analysis, shows that the presence of a core set of virion

assembly genes and a few DNA metabolism genes appears to be constant, and they have remained in approximately constant, but not always contiguous, locations (Figure 8), while the presence of other (often smaller) genes varies widely among the group members. It is of interest to note that, as with the T4 group, there is no evidence of exchange of core genes among these SPO1-like phages, since the trees of the core genes we have examined are strikingly similar (Figure 9). Thus, the two characterized superfamilies of large, virulent tailed phages for which there are enough genomes to compare profitably, typified by T4 and SPO1, are only extremely distantly related and yet have similar relationships between gene sequences and gene organization within each group.

There is growing evidence that some virion structural proteins have common ancestries extending across all tailed phages⁴³. Current sampling of phage genome sequences is extremely sparse, particularly for large phage genomes, and it covers a very limited range of host species, but the evidence cited above for the T4 and SPO1 groups strongly suggests that they are now essentially genetically isolated from each other. However, there are already indications that these groups may not be as monolithic nor as isolated as they may seem from our current vantage point. In the large cyanophages, which have been included in the T4-type superfamily as the “Exo T-evens” sub-type, less than the full set of the T4 core genes can be recognized. Whether the unseen core gene functions are missing from these phages, are present but were acquired horizontally from a very distant gene pool, or are simply diverged past recognition, remains to be determined. For the SPO1 group there is the curious case of *Leptospira* phage LE1 mentioned above, in which the major head protein and seven other putative virion structural proteins are clear homologues of proteins encoded by the SPO1 group (Figure 3), in the context of a phage with a genome half the size of SPO1’s that infects a spirochete and has no other clear homologies to the SPO1 group. As we learn more phage genome sequences, including those of phages infecting more diverse hosts, it will be interesting to see how many more such superfamilies emerge and especially, as sampling of the population becomes denser, whether the boundaries separating them become sharper or more blurred.

Equally interesting, we suggest, will be the non-core genes. These genes are clearly exchanging horizontally at a relatively rapid rate over evolutionary time, and their diversity is such that there is not yet any reliable estimate of how many different kinds of these genes are potentially available to any one phage. We imagine that the flux of these genes through the large phage genomes on an evolutionary time scale allows the phages to adapt to new ecological situations. As we have proposed for phages with smaller genomes^{115; 116}, such as the lambdoid phages, we suggest that the differential rate of flux of the core and non-core genes through the large phage genomes is not the result of different rates of recombination in the two groups of genes. Rather we believe that there is rampant and indiscriminant recombination across the entire genomes of these phages (on an evolutionary timescale), but that the resulting recombinant phages survive only when the associations of the core genes mandated by natural selection remain intact.

MATERIALS AND METHODS

Sequencing

Phage SPO1 from C. Stewart’s strain collection was plaque-purified and propagated in *B. subtilis* CB10⁸⁵. Approximately 10 µg of purified phage SPO1 DNA were sheared hydrodynamically using a Hydroshear instrument (Gene Machines) and repaired using *E. coli* Klenow DNA Polymerase and T4 phage DNA Polymerase according to the manufacturers’ recommendations. These 1–3 Kbp fragments were then ligated into plasmid pBluescript, cleaved with *EcoRV* and transformed into *E. coli* strain XL1BlueF’. Individual plasmid clone inserts were sequenced using outside primers with ABI377 and ABI3100 instruments, and the resulting sequence was assembled with CONSED¹¹⁷. When approximately 8-fold redundancy

in bulk sequencing was attained, oligonucleotide primers were used with genomic template to resolve sequence ambiguities and join contigs to generate a single accurate genome sequence. The boundaries of the terminal redundancy were determined by sequencing toward the ends of the virion DNA, using appropriate primers. Such sequencing reactions are expected to give a mixture of two sequences, one from one copy of the terminal redundancy that terminates at the physical end of the DNA, and another from the other terminal redundancy that continues into the abutting non-redundant sequence. We took advantage of the fact that Taq polymerase makes a large untemplated A peak at the position following the last base of the sequence and searched for positions near the predicted ends where the signal strength dropped by about half following a large untemplated A peak.

Gene identification

The SPO1 genome sequence was scanned for open reading frames, using computer programs Glimmer^{118; 119}, GeneMark¹²⁰, and DNA Strider¹²¹, and for tRNA genes using tRNA Scan-SE with a low threshold score setting (<http://lowelab.ucsc.edu/tRNAscan-SE/>), as well as ARAGORN (<http://bioinfo.thep.lu.se>). The predicted gene locations are listed in Table S1. Because of the stringent requirement for substantial ribosome-binding sites (RBS) able to pair with the 16S rRNA in *B. subtilis*¹²², the appropriate positioning of a sufficiently strong RBS was considered to be a positive identification of a translation start (RBS sequences are listed in Table S6). A few “genes” with questionable RBSs are discussed in the SUPPLEMENTARY MATERIAL.

Nearly half of the previously identified genes (genes “1–3”, “5–26”, and “32”) had been located only by their position on the genetic map as determined by recombination frequency². In some cases, further positional information was obtained from the size of truncated proteins produced by nonsense mutants^{123; 124}) or from the relative effectiveness of complementation or marker rescue by specific cloned fragments^{93; 94}. Many of these genetically identified genes could not be correlated unequivocally with a particular ORF in the genome sequence. For instance, gene *11* must be one of the two ORFs between genome coordinates 50,414 and 51,365. In such a case, we have called those two ORFs *11.1* and *11.2*, with no gene being called simply gene *11*. Similarly, we only know that gene *10* lies between genes 9 and *11*; thus, it could be ORF *9.1*, *10.1*, *10.2*, or *11.1*. Thus, ORF *11.1* could be either of original genes *10* or *11*. There was a good, but not definitive, argument for assigning gene 2 to a single ORF, so that ORF was called gene 2.

Identification of transcription terminators

Possible rho-independent transcription terminators were identified by the STEMLOOP program (Wisconsin Package Version 9.0, Genetics Computer Group (GCG), Madison, WI) as stem-loop structures with stems at least five contiguous base pairs long. These were screened manually for the presence of a downstream run of T's. Some additional putative terminators were found by visual examination of intergenic regions for strings of T's preceded by a stem-loop. 75 putative terminators were thus found in the SPO1 genome, and are listed in Tables S5 and S6. Those not previously named¹³ were numbered T1 to T61 from left to right across the genome. Evaluation of the effectiveness of specific terminators and unusual features of certain terminators are reported in the SUPPLEMENTARY MATERIAL. Since the *rho* gene of *B. subtilis* is dispensable (unlike that of *E. coli*), and since only the *rho* gene itself has been shown to have a Rho-dependent terminator¹²⁵, it is likely that few, if any, SPO1 terminators are Rho-dependent, and we did not search for such terminators.

Identification of promoters

Three types of promoter had been identified previously in the SPO1 genome, early, middle and late (reviewed in ref. ²). Identification of promoters in the newly sequenced portions of the

genome was complicated by the fact that many promoters diverge substantially from their consensus sequences, but still have demonstrable promoter activity. SPO1 early promoters (consensus TTGAC(A/T) – 17–19 bp spacer – (T/A/C)ATAAT) are utilized by host RNA polymerase with host sigma factor A. Many of the *B. subtilis* ϕ A promoters whose activity has been demonstrated have –35 and –10 sequences that diverge significantly from this consensus^{126–128}. The same is true for the previously known SPO1 middle promoters (consensus AGGAGA, 17–19 bp spacer, TTTNTTT^{9; 23}, and late promoters (consensus CGTTAGA, 17–19 bp spacer, GATATT²⁹). For example, the newly sequenced promoter P_{E14}, which has three divergent elements, was shown to have promoter activity⁹⁵, and a demonstrably active middle promoter, P_{M8}, also has three divergent elements^{23; 95}. (Each base within the –35 or –10 regions that was different from the consensus base for that position was counted as one divergent element. If the spacing between –35 and –10 was <16 or >19, that was counted as one divergent element, and spacings of <14 or >21 were counted as two.) The five previously described late promoters²⁹ are all substantially divergent; one has four divergent elements. (Such divergence among late promoters is not unusual; T4 late promoters do not have a recognizable –35 consensus⁹²). Because of this extensive divergence, it was not possible to comprehensively identify all promoters of each type simply by computational means (see also SUPPLEMENTARY MATERIAL). Therefore, we focused our analysis on intergenic regions, scanning each such region visually for each type of promoter.

New putative early, middle and late promoters are named P_E—, P_M—, and P_L—, respectively. We have included early and middle promoters having up to three divergent elements, and late promoters with as many as five. Since there is no evidentiary basis for limiting the acceptable divergence, and since highly divergent promoters can be difficult to recognize, it is quite possible that we have not found all functional promoters or that some of the proposed promoters actually have no promoter activity. As discussed in SUPPLEMENTARY MATERIAL, we also identified eight putative sigma K promoters, which we hypothesize are utilized by the newly identified putative sigma factor, gp2.21. A total of 107 putative promoters were thus identified and are listed in Table S4.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

RWH, GFH and SRC were supported by NIH grants GM51975, AI28927 and AI074825, respectively. MLP was supported by Grant Number P20-RR-16455-05, -06, -07, and -08 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH). CRS was supported by grants 003604-0020-1999 and 003604-0035-2001 from the Texas Advanced Technology Program.

Abbreviations

hmUra	hydroxymethyluracil
RBS	ribosome binding site
ORF	open reading frame

References

1. Hoet PP, Coene MM, Cocito CG. Replication cycle of *Bacillus subtilis* hydroxymethyluracil-containing phages. *Annu Rev Microbiol* 1992;46:95–116. [PubMed: 1444272]
2. Stewart, CR. SPO1 and related bacteriophages. In: Sonenshein, AL.; Joch, JA.; Losick, R., editors. *Bacillus subtilis* and other Gram positive bacteria. American Society of Microbiology; Washington, D.C: 1993.
3. Okubo S, Strauss B, Stodolsky M. The possible role of recombination in the infection of competent *Bacillus subtilis* by bacteriophage deoxyribonucleic acid. *Virology* 1964;24:552–62. [PubMed: 14240402]
4. Goodrich-Blair H, Shub DA. The DNA polymerase genes of several HMU-bacteriophages have similar group I introns with highly divergent open reading frames. *Nucleic Acids Res* 1994;22:3715–21. [PubMed: 7937082]
5. Lawrie JM, Downard JS, Whiteley HR. *Bacillus subtilis* bacteriophages SP82, SPO1, and Φ e: a comparison of DNAs and of peptides synthesized during infection. *J Virol* 1978;27:725–37. [PubMed: 100607]
6. Fujita DJ, Ohlsson-Wilhelm BM, Geiduschek EP. Transcription during bacteriophage SPO1 development: mutations affecting the program of viral transcription. *J Mol Biol* 1971;57:301–17. [PubMed: 4996235]
7. Gage LP, Geiduschek EP. RNA synthesis during bacteriophage SPO1 development: six classes of SPO1 RNA. *J Mol Biol* 1971;57:279–97. [PubMed: 4996233]
8. Losick R, Pero J. Cascades of Sigma factors. *Cell* 1981;25:582–4. [PubMed: 6793235]
9. Lee G, Pero J. Conserved nucleotide sequences in temporally controlled bacteriophage promoters. *J Mol Biol* 1981;152:247–65. [PubMed: 6276567]
10. Goodrich-Blair H, Scarlato V, Gott JM, Xu MQ, Shub DA. A self-splicing group I intron in the DNA polymerase gene of *Bacillus subtilis* bacteriophage SPO1. *Cell* 1990;63:417–24. [PubMed: 2119891]
11. Landthaler M, Lau NC, Shub DA. Group I intron homing in *Bacillus* phages SPO1 and SP82: a gene conversion event initiated by a nicking homing endonuclease. *J Bacteriol* 2004;186:4307–14. [PubMed: 15205433]
12. Cregg JM, Stewart CR. Terminal redundancy of “high frequency of recombination” markers of *Bacillus subtilis* phage SPO1. *Virology* 1978;86:530–41. [PubMed: 96587]
13. Stewart CR, Gaslightwala I, Hinata K, Krolikowski KA, Needleman DS, Peng AS, Peterman MA, Tobias A, Wei P. Genes and regulatory sites of the “host-takeover module” in the terminal redundancy of *Bacillus subtilis* bacteriophage SPO1. *Virology* 1998;246:329–40. [PubMed: 9657951]
14. Pero J, Hannett NM, Talkington C. Restriction cleavage map of SPO1 DNA: general location of early, middle, and late genes. *J Virol* 1979;31:156–71. [PubMed: 116010]
15. Geiduschek EP, Schneider GJ, Sayre MH. TF1, a bacteriophage-specific DNA-binding and DNA-bending protein. *J Struct Biol* 1990;104:84–90. [PubMed: 2128458]
16. Parker ML, Eiserling FA. Bacteriophage SPO1 structure and morphogenesis. III. SPO1 proteins and synthesis. *J Virol* 1983;46:260–9. [PubMed: 6827651]
17. Parker ML, Ralston EJ, Eiserling FA. Bacteriophage SPO1 structure and morphogenesis. II. Head structure and DNA size. *J Virol* 1983;46:250–9. [PubMed: 6402606]
18. Parker ML, Eiserling FA. Bacteriophage SPO1 structure and morphogenesis. I. Tail structure and length regulation. *J Virol* 1983;46:239–49. [PubMed: 6402605]
19. Costanzo M, Hannett N, Brzustowicz L, Pero J. Bacteriophage SPO1 gene 27: location and nucleotide sequence. *J Virol* 1983;48:555–60. [PubMed: 6413701]
20. Costanzo M, Pero J. Structure of a *Bacillus subtilis* bacteriophage SPO1 gene encoding RNA polymerase sigma factor. *Proc Natl Acad Sci U S A* 1983;80:1236–40. [PubMed: 6402778]
21. Costanzo M, Brzustowicz L, Hannett N, Pero J. Bacteriophage SPO1 genes 33 and 34. Location and primary structure of genes encoding regulatory subunits of *Bacillus subtilis* RNA polymerase. *J Mol Biol* 1984;180:533–47. [PubMed: 6441846]
22. Greene JR, Brennan SM, Andrew DJ, Thompson CC, Richards SH, Heinrikson RL, Geiduschek EP. Sequence of the bacteriophage SPO1 gene coding for transcription factor 1, a viral homologue of the

- bacterial type II DNA-binding proteins. *Proc Natl Acad Sci U S A* 1984;81:7031–5. [PubMed: 6438630]
23. Scarlato V, Greene JR, Geiduschek EP. Bacteriophage SPO1 middle transcripts. *Virology* 1991;180:716–28. [PubMed: 1840706]
 24. Scarlato V, Gargano S. The DNA polymerase-encoding gene of *Bacillus subtilis* bacteriophage SPO1. *Gene* 1992;118:109–13. [PubMed: 1324872]
 25. Scarlato V, Sayre MH. Sequence of the bacteriophage SPO1 gene 30. *Gene* 1992;114:115–9. [PubMed: 1587473]
 26. Talkington C, Pero J. Distinctive nucleotide sequences of promoters recognized by RNA polymerase containing a phage-coded “sigma-like” protein. *Proc Natl Acad Sci U S A* 1979;76:5465–9. [PubMed: 118447]
 27. Wei P, Stewart CR. A cytotoxic early gene of *Bacillus subtilis* bacteriophage SPO1. *J Bacteriol* 1993;175:7887–900. [PubMed: 8253678]
 28. Wilhelm K, Ruger W. Deoxyuridylate-hydroxymethylase of bacteriophage SPO1. *Virology* 1992;189:640–6. [PubMed: 1641983]
 29. Costanzo, MD. Ph.D. thesis. Harvard University; 1983. Bacteriophage SPO1 regulatory genes.
 30. Okubo S, Yanagida T, Fujita DJ, Olsson-Wilhelm BM. The genetics of bacteriophage SPO1. *Biken J* 1972;15:81–97. [PubMed: 4628367]
 31. Glassberg J, Slomiany RA, Stewart CR. Selective screening procedure for the isolation of heat- and cold-sensitive, DNA replication-deficient mutants of bacteriophage SPO1 and preliminary characterization of the mutants isolated. *J Virol* 1977;21:54–60. [PubMed: 401897]
 32. Sampath A, Stewart CR. Roles of genes 44, 50, and 51 in regulating gene expression and host takeover during infection of *Bacillus subtilis* by bacteriophage SPO1. *J Bacteriol* 2004;186:1785–92. [PubMed: 14996809]
 33. Sayre MH, Geiduschek EP. TF1, the bacteriophage SPO1-encoded type II DNA-binding protein, is essential for viral multiplication. *J Virol* 1988;62:3455–62. [PubMed: 2841496]
 34. 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;102:5174–9. [PubMed: 15788529]
 35. O’Flaherty S, Coffey A, Edwards R, Meaney W, Fitzgerald GF, Ross RP. Genome of staphylococcal phage K: a new lineage of Myoviridae infecting gram-positive bacteria with a low G+C content. *J Bacteriol* 2004;186:2862–71. [PubMed: 15090528]
 36. Gorbalenya AE. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci* 1994;3:1117–20. [PubMed: 7920259]
 37. Huang H, Yuan HS. The conserved asparagine in the HNH motif serves an important structural role in metal finger endonucleases. *J Mol Biol* 2007;368:812–21. [PubMed: 17368670]
 38. Shub DA, Goodrich-Blair H, Eddy SR. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem Sci* 1994;19:402–4. [PubMed: 7817395]
 39. Landthaler M, Begley U, Lau NC, Shub DA. Two self-splicing group I introns in the ribonucleotide reductase large subunit gene of *Staphylococcus aureus* phage Twort. *Nucleic Acids Res* 2002;30:1935–43. [PubMed: 11972330]
 40. Landthaler M, Shub DA. Unexpected abundance of self-splicing introns in the genome of bacteriophage Twort: introns in multiple genes, a single gene with three introns, and exon skipping by group I ribozymes. *Proc Natl Acad Sci U S A* 1999;96:7005–10. [PubMed: 10359829]
 41. Carlton RM, Noordman WH, Biswas B, de Meester ED, Loessner MJ. Bacteriophage P100 for control of *Listeria monocytogenes* in foods: genome sequence, bioinformatic analyses, oral toxicity study, and application. *Regul Toxicol Pharmacol* 2005;43:301–12. [PubMed: 16188359]
 42. Chibani-Chennoufi S, Dillmann ML, Marvin-Guy L, Rami-Shojaei S, Brüssow H. *Lactobacillus plantarum* bacteriophage LP65: a new member of the SPO1-like genus of the family Myoviridae. *J Bacteriol* 2004;186:7069–83. [PubMed: 15489418]
 43. Casjens SR. Comparative genomics and evolution of the tailed-bacteriophages. *Curr Opin Microbiol* 2005;8:451–8. [PubMed: 16019256]

44. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM, Hryckowian AJ, Kelchner VA, Namburi S, Pajcini KV, Popovich MG, Schleicher DT, Simanek BZ, Smith AL, Zdanowicz GM, Kumar V, Peebles CL, Jacobs WR Jr, Lawrence JG, Hendrix RW. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet* 2006;2:e92. [PubMed: 16789831]
45. Rohwer F. Global phage diversity. *Cell* 2003;113:141. [PubMed: 12705861]
46. Duda RL, Hendrix RW, Huang WM, Conway JF. Shared architecture of bacteriophage SPO1 and herpesvirus capsids. *Curr Biol* 2006;16:R11–3. [PubMed: 16401408]
47. Loessner MJ, Scherer S. Organization and transcriptional analysis of the *Listeria* phage A511 late gene region comprising the major capsid and tail sheath protein genes *cps* and *tsh*. *J Bacteriol* 1995;177:6601–9. [PubMed: 7592439]
48. Bourhy P, Frangeul L, Couve E, Glaser P, Saint Girons I, Picardeau M. Complete nucleotide sequence of the LE1 prophage from the spirochete *Leptospira biflexa* and characterization of its replication and partition functions. *J Bacteriol* 2005;187:3931–40. [PubMed: 15937155]
49. Casjens, S.; Hendrix, R. Control mechanisms in dsDNA bacteriophage assembly. In: Calendar, R., editor. *The Bacteriophages*. Vol. 1. Plenum Press; New York City: 1988. p. 15-91.
50. Liu J, Mushegian A. Displacements of prohead protease genes in the late operons of double-stranded-DNA bacteriophages. *J Bacteriol* 2004;186:4369–75. [PubMed: 15205439]
51. Saint Girons I, Margarita D, Amouriaux P, Baranton G. First isolation of bacteriophages for a spirochaete: potential genetic tools for *Leptospira*. *Res Microbiol* 1990;141:1131–8. [PubMed: 2092364]
52. Caspar D, Klug A. Physical principles in the construction of regular viruses. *Cold Spr Harbor Symp Quant Biol* 1962;27:1–27.
53. Fokine A, Leiman PG, Shneider MM, Ahvazi B, Boeshans KM, Steven AC, Black LW, Mesyanzhinov VV, Rossmann MG. 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;102:7163–8. [PubMed: 15878991]
54. Wikoff WR, Liljas L, Duda RL, Tsuruta H, Hendrix RW, Johnson JE. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* 2000;289:2129–33. [PubMed: 11000116]
55. Jiang W, Baker ML, Jakana J, Weigele PR, King J, Chiu W. Backbone structure of the infectious ϵ 15 virus capsid revealed by electron cryomicroscopy. *Nature* 2008;451:1130–4. [PubMed: 18305544]
56. 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;18:149–59. [PubMed: 15837419]
57. Weigele PR, Pope WH, Pedulla ML, Houtz JM, Smith AL, Conway JF, King J, Hatfull GF, Lawrence JG, Hendrix RW. Genomic and structural analysis of Syn9, a cyanophage infecting marine *Prochlorococcus* and *Synechococcus*. *Environ Microbiol* 2007;9:1675–95. [PubMed: 17564603]
58. Morgan G, Hatfull G, Casjens S, Hendrix R. Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in *Haemophilus*, *Neisseria* and *Deinococcus*. *J Mol Biol* 2002;317:337–359. [PubMed: 11922669]
59. Pedersen M, Ostergaard S, Bresciani J, Vogensen FK. Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901–1 involved in tail length determination and baseplate assembly. *Virology* 2000;276:315–28. [PubMed: 11040123]
60. Rees PJ, Fry BA. The morphology of staphylococcal bacteriophage K and DNA metabolism in infected *Staphylococcus aureus*. *J Gen Virol* 1981;53:293–307. [PubMed: 6455500]
61. Katsura I, Hendrix RW. Length determination in bacteriophage lambda tails. *Cell* 1984;39:691–8. [PubMed: 6096021]
62. Katsura I. Determination of bacteriophage lambda tail length by a protein ruler. *Nature* 1987;327:73–5. [PubMed: 2952887]
63. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF. Origins of highly mosaic mycobacteriophage genomes. *Cell* 2003;113:171–82. [PubMed: 12705866]

64. Levin ME, Hendrix RW, Casjens SR. A programmed translational frameshift is required for the synthesis of a bacteriophage λ tail assembly protein. *J Mol Biol* 1993;234:124–39. [PubMed: 8230192]
65. Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol Cell* 2004;16:11–21. [PubMed: 15469818]
66. Baranov PV, Fayet O, Hendrix RW, Atkins JF. Recoding in bacteriophages and bacterial IS elements. *Trends Genet* 2006;22:174–81. [PubMed: 16460832]
67. Moak M, Molineux IJ. Peptidoglycan hydrolytic activities associated with bacteriophage virions. *Mol Microbiol* 2004;51:1169–83. [PubMed: 14763988]
68. Weigele PR, Scanlon E, King J. Homotrimeric, β -stranded viral adhesins and tail proteins. *J Bacteriol* 2003;185:4022–30. [PubMed: 12837775]
69. Bradley P, Cowen L, Menke M, King J, Berger B. BETAWRAP: successful prediction of parallel β -helices from primary sequence reveals an association with many microbial pathogens. *Proc Natl Acad Sci U S A* 2001;98:14819–24. [PubMed: 11752429]
70. Cregg JM, Stewart CR. EcoRI cleavage of DNA from *Bacillus subtilis* phage SPO1. *Virology* 1978;85:601–5. [PubMed: 96586]
71. Levner MH, Cozzarelli NR. Replication of viral DNA in SPO1-infected *Bacillus subtilis*. I. Replicative intermediates. *Virology* 1972;48:402–16. [PubMed: 4623500]
72. Catalano, CE. Viral genome packaging machines. In: Catalano, CE., editor. *Viral genome packaging machines: genetics, structure, and mechanism*. Landes Bioscience; Georgetown, TX: 205. p. 1–5.
73. Casjens, S.; Gilcrease, D. Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. In: Clokie, M.; Kropinski, A., editors. *Bacteriophages: methods and protocols*. Humana Press; Totowa, NJ: 2008. in press
74. Casjens SR, Gilcrease EB, Winn-Stapley DA, Schicklmaier P, Schmiegier H, Pedulla ML, Ford ME, Houtz JM, Hatfull GF, Hendrix RW. The generalized transducing *Salmonella* bacteriophage ES18: complete genome sequence and DNA packaging strategy. *J Bacteriol* 2005;187:1091–104. [PubMed: 15659686]
75. Klumpp J, Dorscht J, Lurz R, Biemann R, Wieland M, Zimmer M, Calendar R, Loessner MJ. The terminally redundant, nonpermuted genome of *Listeria* bacteriophage A511: a model for the SPO1-like myoviruses of gram-positive bacteria. *J Bacteriol* 2008;190:5753–65. [PubMed: 18567664]
76. Wang J, Jiang Y, Vincent M, Sun Y, Yu H, Bao Q, Kong H, Hu S. Complete genome sequence of bacteriophage T5. *Virology* 2005;332:45–65. [PubMed: 15661140]
77. Thomas JA, Hardies SC, Rolando M, Hayes SJ, Lieman K, Carroll CA, Weintraub ST, Serwer P. Complete genomic sequence and mass spectrometric analysis of highly diverse, atypical *Bacillus thuringiensis* phage 0305 ϕ 8–36. *Virology* 2007;368:405–21. [PubMed: 17673272]
78. Casjens S, Hatfull G, Hendrix R. Evolution of dsDNA tailed-bacteriophage genomes. *Sem Virol* 1992;3:383–397.
79. Talkington C, Pero J. Restriction fragment analysis of the temporal program of bacteriophage SPO1 transcription and its control by phage-modified RNA polymerases. *Virology* 1977;83:365–79. [PubMed: 412317]
80. Fujita, DJ. Ph.D thesis. University of Chicago; 1971. Studies on conditional lethal mutants of bacteriophage SPO1.
81. Gilcrease EB, Winn-Stapley DA, Hewitt FC, Joss L, Casjens SR. Nucleotide sequence of the head assembly gene cluster of bacteriophage L and decoration protein characterization. *J Bacteriol* 2005;187:2050–7. [PubMed: 15743953]
82. Glassberg J, Franck M, Stewart CR. Multiple origins of replication for *Bacillus subtilis* phage SPO1. *Virology* 1977;78:433–41. [PubMed: 405794]
83. Duggin IG, Matthews JM, Dixon NE, Wake RG, Mackay JP. A complex mechanism determines polarity of DNA replication fork arrest by the replication terminator complex of *Bacillus subtilis*. *J Biol Chem* 2005;280:13105–13. [PubMed: 15657033]
84. Mackiewicz P, Zakrzewska-Czerwinska J, Zawilak A, Dudek MR, Cebrat S. Where does bacterial replication start? Rules for predicting the *oriC* region. *Nucleic Acids Res* 2004;32:3781–91. [PubMed: 15258248]

85. Glassberg J, Franck M, Stewart CR. Initiation and termination mutants of *Bacillus subtilis* bacteriophage SPO1. *J Virol* 1977;21:147–52. [PubMed: 401896]
86. Gage LP, Geiduschek EP. RNA synthesis during bacteriophage SPO1 development. II. Some modulations and prerequisites of the transcription program. *Virology* 1971;44:200–10. [PubMed: 4999125]
87. De Antoni GL, Besso NE, Zanassi GE, Sarachu AN, Grau O. Bacteriophage SPO1 DNA polymerase and the activity of viral gene 31. *Virology* 1985;143:16–22. [PubMed: 3933176]
88. Shub DA. Bacteriophage SPO1 DNA- and RNA-directed protein synthesis *in vitro*: comparison with *in vivo* control. *Mol Gen Genet* 1975;137:171–80. [PubMed: 810656]
89. Kahan E. Early and late gene function in bacteriophage SP82. *Virology* 1971;46:634–7. [PubMed: 5002938]
90. Warner HR, Thompson RB, Mozer TJ, Duncan BK. The properties of a bacteriophage T5 mutant unable to induce deoxyuridine 5'-triphosphate nucleotidohydrolase. Synthesis of uracil-containing T5 deoxyribonucleic acid. *J Biol Chem* 1979;254:7534–9. [PubMed: 381286]
91. Dunham LF, Price AR. Deoxythymidine triphosphate-deoxyuridine triphosphate nucleotidohydrolase induced by *Bacillus subtilis* bacteriophage Φ e. *Biochemistry* 1974;13:2667–72. [PubMed: 4367957]
92. Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Ruger W. Bacteriophage T4 genome. *Microbiol Mol Biol Rev* 2003;67:86–156. [PubMed: 12626685]
93. Curran JF, Stewart CR. Transcription of *Bacillus subtilis* plasmid pBD64 and expression of bacteriophage SPO1 genes cloned therein. *Virology* 1985;142:98–111. [PubMed: 2414903]
94. Curran JF, Stewart CR. Cloning and mapping of the SPO1 genome. *Virology* 1985;142:78–97. [PubMed: 2997983]
95. Chelm BK, Romeo JM, Brennan SM, Geiduschek EP. A transcriptional map of the bacteriophage SPO1 genome. III. A region of early and middle promoters (the gene 28 region). *Virology* 1981;112:572–88. [PubMed: 6266141]
96. Romeo JM, Brennan SM, Chelm BK, Geiduschek EP. A transcriptional map of the bacteriophage SPO1 genome. I. The major early promoters. *Virology* 1981;111:588–603. [PubMed: 18635059]
97. Brennan SM, Chelm BK, Romeo JM, Geiduschek EP. A transcriptional map of the bacteriophage SPO1 genome. II. the major early transcription units. *Virology* 1981;111:604–628. [PubMed: 18635060]
98. Filee J, Baptiste E, Susko E, Krisch HM. A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Mol Biol Evol* 2006;23:1688–96. [PubMed: 16782763]
99. Nolan JM, Petrov V, Bertrand C, Krisch HM, Karam JD. Genetic diversity among five T4-like bacteriophages. *Virol J* 2006;3:30. [PubMed: 16716236]
100. Petrov VM, Nolan JM, Bertrand C, Levy D, Desplats C, Krisch HM, Karam JD. Plasticity of the gene functions for DNA replication in the T4-like phages. *J Mol Biol* 2006;361:46–68. [PubMed: 16828113]
101. Hambly E, Tetart F, Desplats C, Wilson WH, Krisch HM, Mann NH. A conserved genetic module that encodes the major virion components in both the coliphage T4 and the marine cyanophage S-PM2. *Proc Natl Acad Sci U S A* 2001;98:11411–6. [PubMed: 11553768]
102. Comeau AM, Chan AM, Suttle CA. Genetic richness of vibriophages isolated in a coastal environment. *Environ Microbiol* 2006;8:1164–76. [PubMed: 16817925]
103. 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;362:384–96. [PubMed: 17289101]
104. 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;102:12471–6. [PubMed: 16116082]
105. Hambly E, Suttle CA. The virosphere, diversity, and genetic exchange within phage communities. *Curr Opin Microbiol* 2005;8:444–50. [PubMed: 15979387]
106. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A* 1999;96:3801–6. [PubMed: 10097118]

107. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol* 2006;4:e234. [PubMed: 16802857]
108. Mann NH, Clokie MR, Millard A, Cook A, Wilson WH, Wheatley PJ, Letarov A, Krisch HM. The genome of S-PM2, a “photosynthetic” T4-type bacteriophage that infects marine *Synechococcus* strains. *J Bacteriol* 2005;187:3188–200. [PubMed: 15838046]
109. Mesyanzhinov VV, Robben J, Grymonprez B, Kostyuchenko VA, Bourkaltseva MV, Sykilinda NN, Krylov VN, Volckaert G. The genome of bacteriophage ϕ KZ of *Pseudomonas aeruginosa*. *J Mol Biol* 2002;317:1–19. [PubMed: 11916376]
110. Hertveldt K, Lavigne R, Pleteneva E, Sernova N, Kurochkina L, Korchevskii R, Robben J, Mesyanzhinov V, Krylov VN, Volckaert G. Genome comparison of *Pseudomonas aeruginosa* large phages. *J Mol Biol* 2005;354:536–45. [PubMed: 16256135]
111. Kiljunen S, Hakala K, Pinta E, Huttunen S, Pluta P, Gador A, Lonnberg H, Skurnik M. Yersiniophage ϕ R1–37 is a tailed bacteriophage having a 270 kb DNA genome with thymidine replaced by deoxyuridine. *Microbiology* 2005;151:4093–102. [PubMed: 16339954]
112. Naryshkina T, Liu J, Florens L, Swanson SK, Pavlov AR, Pavlova NV, Inman R, Minakhin L, Kozyavkin SA, Washburn M, Mushegian A, Severinov K. *Thermus thermophilus* bacteriophage ϕ YS40 genome and proteomic characterization of virions. *J Mol Biol* 2006;364:667–77. [PubMed: 17027029]
113. Uchiyama J, Rashel M, Maeda Y, Takemura I, Sugihara S, Akechi K, Muraoka A, Wakiguchi H, Matsuzaki S. Isolation and characterization of a novel *Enterococcus faecalis* bacteriophage ϕ EF24C as a therapeutic candidate. *FEMS Microbiol Lett* 2008;278:200–6. [PubMed: 18096017]
114. Pantucek R, Rosypalova A, Doskar J, Kailerovala J, Ruzickova V, Borecka P, Snopkova S, Horvath R, Gotz F, Rosypal S. The polyvalent staphylococcal phage ϕ 812: its host-range mutants and related phages. *Virology* 1998;246:241–52. [PubMed: 9657943]
115. Hendrix RW. Bacteriophages: evolution of the majority. *Theor Popul Biol* 2002;61:471–80. [PubMed: 12167366]
116. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 2000;299:27–51. [PubMed: 10860721]
117. Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res* 1998;8:195–202. [PubMed: 9521923]
118. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* 1999;27:4636–41. [PubMed: 10556321]
119. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* 1998;26:544–8. [PubMed: 9421513]
120. Borodovsky M, McInish J. GeneMark: parallel gene recognition for both strands. *Comput Chem* 1993;17:123–133.
121. Douglas SE. DNA Strider. A Macintosh program for handling protein and nucleic acid sequences. *Methods Mol Biol* 1994;25:181–94. [PubMed: 8004164]
122. Mountain, A. Gene expression systems for *Bacillus subtilis*. In: Harwood, CR., editor. *Bacillus*. Plenum; New York City, NY: 1989. p. 73-114.
123. Shub DA. Nature of the suppressor of *Bacillus subtilis* HA101B. *J Bacteriol* 1975;122:788–90. [PubMed: 805128]
124. Reeve JN, Mertens G, Amann E. Early development of bacteriophages SP01 and SP82G in minicells of *Bacillus subtilis*. *J Mol Biol* 1978;120:183–207. [PubMed: 417185]
125. Helmann, JD.; Moran, CP. RNA polymerase and sigma factors. In: Sonenshein, AL.; Hoch, JA.; Losick, R., editors. *Bacillus subtilis and its closest relatives: From genes to cells*. American Society for Microbiology; Washington, D.C: 2002. p. 289-312.
126. Helmann JD. Compilation and analysis of *Bacillus subtilis* sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res* 1995;23:2351–60. [PubMed: 7630711]
127. Jarmer H, Larsen TS, Krogh A, Saxild HH, Brunak S, Knudsen S. Sigma A recognition sites in the *Bacillus subtilis* genome. *Microbiology* 2001;147:2417–24. [PubMed: 11535782]

128. Makita Y, Nakao M, Ogasawara N, Nakai K. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res* 2004;32:D75–7. [PubMed: 14681362]
129. Earnshaw W, Casjens S, Harrison S. Assembly of the head of bacteriophage P22, X-ray diffraction from heads, proheads and related structures. *J Mol Biol* 1976;104:387–410. [PubMed: 781287]
130. 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;22:4673–80. [PubMed: 7984417]
131. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* 2005;3:e144. [PubMed: 15828858]

Figure 1. Gene and transcription unit map of the SPO1 genome

The sequence of the SPO1 genome is represented with the two copies of its terminal direct repeats highlighted with yellow boxes. Predicted genes are indicated by rectangular boxes; rightward-transcribed (green) are denoted above the kbp scale and leftward-transcribed (red) genes are below. Black and blue arrows denote rightward and leftward transcription units (defined in the text), vertical green and red lines mark predicted promoters and terminators, respectively. E, M, L, and K represent early, middle, late, and sigma K promoters, respectively. Selected gene names are shown on the gene rectangles, and transcript numbers are shown above each transcript. Established or predicted functions of the encoded proteins are indicated below the gene rectangles. Those predicted only from sequence comparisons are underlined.

Figure 2. SPO1 virion proteins

The protein components of SPO1 virions (purified by CsCl density step gradient centrifugation¹²⁹) are displayed by SDS-12.5% polyacrylamide gel electrophoresis and stained with Coomassie brilliant blue¹²⁹. A molecular weight scale is shown on the left, and the five bands subjected to N-terminal amino acid sequence analysis are marked on the right with the SPO1 gene that is thought to encode them and their N-terminal amino acid sequence in parentheses.

Figure 3. Similar gene organization in the morphogenetic gene clusters of the SPO1-like phages
 The head and tail gene regions are shown for each of the SPO1-like phages and phage LE1. Genes are represented by rectangles; each phage's genes are indicated by a different color except that white boxes indicate genes that have no matches; red genes are inside introns, and terminase, coat and tail sheath genes are yellow for orientation. Blue lines connect homologous genes. Beneath each gene cluster is a scale in kbp. Note that all the homologous genes have similar locations except for LP65 genes *129–132* and *134* which are located far from this cluster, and SPO1 gene *2.11* which is separated by a large insertion (see text). Above the SPO1 map the putative functions are shown (“tail” refers to genes of unknown function with homologues in tail gene clusters of phages outside the SPO1-like phage group).

Figure 4. Sequence relationship between SPO1 head and protease proteins

A. The shaded rectangle shows the location and extent of sequence similarity between the major capsid protein and the head maturation protease of SPO1. Sequence matching other phage proteases, corresponding to the catalytic domain, and the site of maturation cleavage, are indicated. **B.** Similar relationship between the sequences of the phage T4 major capsid protein and a small, separately encoded T4 protein gp61.1.

Figure 5. Tail gene comparisons

Alignment of a portion of the tail gene regions of the SPO1-like group of phages analyzed here. Orthologous genes, represented as rectangles filled with the same color, have been classified into the same Phamily by the computer program Phamerator, and they are connected in the figure by red lines. SPO1 gene *10.1* and its homologues in the other phages are proposed to encode the tail tube protein. SPO1 genes *10.2* and *11.1* are proposed as the genes encoding the tail assembly chaperones related by a programmed translational frameshift. This figure has been extracted from the figure showing the alignment of the complete genomes of these six phages, shown as Figure S3 in SUPPLEMENTARY MATERIAL. Gene names are shown inside the gene rectangle; Phamily number is shown above each gene; the three lavender rectangles in Twort gene *106* represent introns. For more details, see SUPPLEMENTARY MATERIAL.

<u>PHAGE</u>							<u>POSITION</u>
SPO1	GGA	GAA	GCU	AAAG	GAA	AUA	–7
K/G1	UUA	AAC	UCC	UUUC	GUU	ACU	–3
P100/A511	CUA	GAC	ACG	UUUC	GCU	ACU	–3
LP65	UAC	GAU	ACG	UUUC	UCA	AAA	–7
Twort	UUA	AAU	ACC	UUUC	GUU	ACU	–3

Incoming frame / \ Outgoing frame
 Frameshift
 site

Figure 6. Tail gene frameshifting

The mRNA sequences shown are from the overlap region of the two ORFs in the phage tail genes that are proposed to sponsor a +1 programmed translational frameshift (see text). Bases to the left are grouped to indicate the codons in the incoming frame and bases to the right indicate the outgoing frame. A tRNA in the P site of the ribosome (tRNA-Lys for SPO1 and tRNA-Phe for the other phages) is proposed to shift one base forward at the frameshift site. Numbers in the “Position” column indicate the position of the frameshift site relative to the termination codon of the upstream gene, in codons.

Figure 7. Size distributions of SPO1 genes

Bars indicate the number of SPO1 genes in each size category; black, genes that have no detected homologues in the other phages of the SPO1 superfamily; gray, genes with at least one homologue in this group of phages.

Figure 8. Comparative organization of SPO1 and closely related genomes

The genomes of the six phages named at the left are shown above each of their published kbp scales. The orange boxes on the SPO1 genome mark its long terminal repeats (the possible terminal redundancies of the other five phages are not indicated). The genes with the following functions are found in the regions indicated by colored boxes above the scales as follows: virion assembly (yellow); DNA metabolism (red); lysis (green); and tRNAs (blue with number of tRNA genes indicated). Red lines connect genes that lie in parallel locations, present in most members of the group, and that map outside the virion assembly gene cluster (see Figure 3 for such relationships within the latter cluster); blue lines indicate some examples of genes present in most of the genomes that are in different locations. Green shading between genomes denotes recognizable overall nucleotide sequence similarity. Note that phages K and G1 are nearly identical except for three insertions (noted by black triangles) in G1 relative to K. Black arrows below the kbp scales denote the general direction of transcription in the various regions of each genome.

Figure 9. Neighbor-joining trees of four SPO1 core gene products

Clustal W¹³⁰ was used to construct neighbor-joining trees of the indicated six proteins. The scales of the trees are approximately equal, with 5% difference indicated by a bar near each tree. Non-SPO1-like phage LE1 is included in the four virion structural protein trees (coat, portal, sheath and baseplate proteins); phage A511 (a recently sequenced close relative of phage P100⁴⁷) genes are included where they were available; orthologues from T4 and T4-like subgroups¹⁰⁴ pseudo-T4 (phage RB49), Schizo-T4 (phage Aeh1), and Exo-T4 (phages PSSM4¹³¹ and Syn9⁵⁷) are included for comparison in the two lower trees; if the T4-like orthologues are included in the top four trees they also fall outside the SPO1-like branches (not shown). Bootstrap values from 1000 repetitions are shown for each branch. Branching orders

are very similar for the six SPO1-like phage proteins in all six trees with only minor differences in branch lengths. The SPO1 group and T4 group orthologues (including the virion proteins in the upper four trees, not shown) formed separate branches with very high bootstrap support.

Table 1
Orthologous Gene Phamilies in the SPO1-like phages¹

	LP65	G1	K	Twort	P100	size(kbp) ²	No. of genes ³
LP65						132	169
G1	36					139	207
K	35	141				127 ⁴	176 ⁴
Twort	37	101	97			131	182
P100	32	64	63	59		131	178
SPO1	27	30	27	30	29	132	204

¹ This table shows the number of Phamerator gene Phamilies in common in pairwise comparisons of the SPO1-like phages. In a few cases complex Phamilies can be parsed into more than one orthologous sub-Phamily, so the values in the table are a slight underestimate of the number of orthologous genes in these phages (see SUPPLEMENTARY MATERIAL).

² Genome size does not include second copy of terminal redundancy.

³ Protein-encoding genes. Some re-analysis of the previously published genomes was carried out, which in several cases increased the number of predicted genes from the previously published value.

⁴ Genome size and gene number for phage K are slightly higher than shown here due to the fact that a portion of the genome sequence was not determined³⁵. Comparison with the phage G1 sequence (Figure S3) suggests that the missing segment of the K sequence may be about 2.9 kbp with about 8 genes.