Prophage genomics reveals patterns in phage genome organization and replication

Han Suh Kang[1], Katelyn McNair[2], Daniel A. Cuevas[3], Barbara A. Bailey[4], Anca M. Segall[1], Robert A. Edwards[1,2,3]

[1]Department of Biology, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182

[2]Department of Computer Science, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182

[3]Computational Sciences Research Center, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182

[4]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182

Abstract

Temperate phage genomes are highly variable mosaic collections of genes that infect a bacterial host, integrate into the host's genome or replicate as low copy number plasmids, and are regulated to switch from the lysogenic to lytic cycles to generate new virions and escape their host. Genomes from most Bacterial phyla contain at least one or more prophages. We updated our PhiSpy algorithm to improve detection of prophages and to provide a web-based framework for PhiSpy. We have used this algorithm to identify 36,488 prophage regions from 11,941 bacterial genomes, including almost 600 prophages with no known homology to any proteins. Transfer RNA genes were abundant in the prophages, many of which alleviate the limits of translation efficiency due to host codon bias and presumably enable phages to surpass the normal capacity of the hosts' translation machinery. We identified integrase genes in 15,765 prophages (43% of the prophages). The integrase was routinely located at either end of the integrated phage genome, and was used to orient and align prophage genomes to reveal their underlying organization. The conserved genome alignments of phages recapitulate early, middle, and late gene order in transcriptional control of phage genes, and demonstrate that gene order, presumably selected by transcription timing and/or coordination among functional modules has been stably conserved throughout phage evolution.

Introduction

Phages are major drivers in global biogeochemical cycles through the control they exert on bacterial populations. Despite their abundance in the environment they are under-represented in sequence databases and knowledge-bases. Phages' two life cycles, the lytic and lysogenic cycles, provide a unique window into phage genomics. During the lysogenic cycle, temperate phages integrate their genomes into their host's genome in a non-lethal fashion to generate a region of DNA known as a prophage – either as a plasmid (1) or by directly recombining with the host's genome (2, 3). The phage replicates in the host genome until external signals promote a transition to the lytic cycle, at which point the phage hijacks the host's cellular DNA replication machinery in order to replicate the phage genome, and adopts the host's protein synthesis machinery to create new phage particles. Eventually the phage lyse releases infective particles to find new hosts to attack.

The activation and movement of a prophage from one host to another enables phages to act as mediators of horizontal gene transfer, via either generalized or specialized transduction (4). This phage-mediated horizontal gene transfer may account for up to 10^{15} transduction events per second (5, 6). Phage-mediated horizontal gene transfer events are responsible for the spread of novel functions including the development of pathogenicity in bacterial strains such as *Escherichia coli* O157:H7 (5), *Vibrio cholerae* El Tor N16961 (7), *Streptococcus pyogenes* (8) and many *Staphylococcus aureus* strains (9). Phages may mediate the spread of antibiotic resistance in the mammalian intestine (10), and also transfer more benign genes such as those involved in photosynthesis (11).

We recently proposed a new model for phage host interactions that suggests at higher host densities ecological considerations favor the temperate lifestyle (12). The increased energy available at high host densities allows bacteria to support the replication of prophages, which in turn provide beneficial protection to the host via super-infection exclusion. The phages are using their host for replication, but are also providing protection for their host from other invading phages.

Identification of prophages in bacterial genomes is complicated by the problem that many phage genes do not share sequence homology with other proteins. In some cases a newly discovered phage does not carry any genes that share sequence similarity to any other gene in the available databases, highlighting how little we know about these viruses (13). The underrepresentation of well-annotated phage genes in available databases makes it difficult to assign functions to newly sequenced genomes. This is exacerbated as genome sequencing increases at a near-exponential rate, requiring reliance on automated annotation software that is dependent on primary sequence similarity to assign functions to new genes (14, 15). Most prophage identification algorithms rely on sequence similarity to known phage genes to identify prophage regions (16–20). This leads to a paradoxical problem of not being able to identify a prophage that does not contain any regions homologous to known phage proteins. Our prophage-finding program, PhiSpy, uses a novel algorithm that is capable of identifying a prophage even if there is no similarity to any of the proteins in the databases by analyzing a suite of genomic characteristics that separate phage-encoding regions from the bacterial core of the genome (21).

Here we present an analysis of over 11,000 bacterial genomes from which we identified 36,488 prophages. We used this data to identify a core phage genome organizational profile and demonstrate that a significant fraction of phage genomes adhere to this model. This genome structure allows us to predict the function of genes with no known homology solely based on their location in the genome. We identified the integrase genes in 15,162 prophages, and aligned those genes to infer phylogenetic relationships between prophages. Furthermore, we explored the connection between phage and host genomic characteristics to enable predictions of the bacteria that phages are infecting.

Materials and Methods

PhiSpy version 3.2 Improvements

Several improvements were made to the PhiSpy program, and we have released a new version of the software and a web site to provide an accessible interface to the program. The most important change we made was to allow PhiSpy to process genomes that were composed of multiple contigs, as the original version of PhiSpy would treat proteins from different contigs as adjoining proteins. The previous version of PhiSpy used a scanning window that looked 40 proteins ahead of the current protein being classified, but to remove the inherent direction bias, we changed the window to be 20 proteins on either side of the protein being classified. For proteins near the edges of contigs, which had fewer than 20 proteins on a side, we included all possible proteins. To update the list of prophage protein annotation keywords, all the phage genomes from PhAnToMe (http://www.phantome.org), and all the complete prokaryotic genomes from GenBank were downloaded. We compared the annotations of each set, and found 12 keywords unique to phage protein annotations: endolysin, terminase, baseplate, base plate, virion, antirepressor, excisionase, Cro-like repressor, cI-like repressor, rIIIA lysis, rI lysis, and rIIB lysis. These keywords were incorporated into PhiSpy for protein classification. In addition to these major changes, many other minor bug fixes were applied. A website was also built so that users can run PhiSpy online, without having to download and install the program and its dependencies: http://edwards.sdsu.edu/PhiSpy/. The source code for the updated version of PhiSpy is available from that website and from GitHub: https://github.com/linsalrob/PhiSpy.

Dataset generation

All bacterial genomes were extracted from the SEED database, which contains both sequence and annotation information that is used by PhiSpy in making prophage designations (14, 15). The bacterial genomes contained a representative of every genus currently known. The updated version of PhiSpy (version 3.2) was run on all bacterial genomes using the generic bacterial training set option, and the outputs were parsed to identify all predicted prophage regions (Fig. 1). All PhiSpy predictions are available from https://edwards.sdsu.edu/PhiSpy and https://figshare.com/articles/PhiSpy_prophage_predictions/3146656.

Raw sequence phage-host correlation studies

Prophage coordinates from PhiSpy were used to extract nucleotide sequence data from the original bacterial genome. The nucleotide sequences are available as FASTA files from https://edwards.sdsu.edu/PhiSpy and https://figshare.com/articles/PhiSpy_prophage_predictions/3146656. Putative tRNA-encoding genes were identified from those genomes using tRNAscan-SE (22). Modal codon usage was calculated from both the predicted prophage and the host genomes as described previously (23), and tRNA use and modal codon use were compared using the χ^2-test. The G+C content of the phage and host nucleotide sequences were calculated using custom-written Python scripts.

Genome heatmaps

High-quality prophage genome predictions were defined as those where at least one protein was annotated as an integrase, a capsid, a portal, a terminase, and a tail protein. The annotated genes in all the prophages were tabulated to identify a suite of genes that occur frequently in phage genome annotations (integrase, portal, tail, holin, transposase, terminase, endonuclease, capsid, antitermination, Holliday junction resolvase, baseplate assembly protein, primosomal protein, protease-like, lysin, toxin-encoding, and lysozyme).

To generate the genome alignment, each gene was assigned a unique letter indicating its function. A modified Clustal Omega alignment algorithm that scores +1 for matches and 0 for mismatches was used to align the annotated genomes.

Integrase phylogenomics

A total of 24,936 integrase (Int) gene sequences were extracted from the high-quality manually curated dataset and clustered using CD-HIT (24) at a 90% 24,936 similarity threshold (an approximate virus “species” demarcation (16, 18) to produce 8,536 protein clusters. A representative of each cluster was used in the phylogenetic assessment of the 24,936 genes. In addition, short motifs were identified in the Int genes by using *k*-mer profiles for *k* from 4 to 7 amino acids long.

This approach allowed us to identify shared amino acid motifs including the eight-residue motif at the active site of the tyrosine recombinases. One hundred randomly selected integrases were aligned using Clustal Omega (25, 26) and visualized using FigTree (27). To compare the evolution of phage and bacterial proteins, all 24,936 phage Int proteins were aligned in a pair-wise fashion using the Smith-Waterman algorithm and the percent identity between each pair of proteins was calculated. Bacterial taxonomy was taken from the NCBI taxonomy database (28).

Computational materials

Computational work was conducted on the Edwards Bioinformatics Lab compute cluster. Standard Python libraries were used for file parsing and Biopython version 1.65 was used to parse GenBank files. Scripts written for file parsing and data analysis are available in the github repository: https://github.com/hkang408/ProphageGenomics and from https://github.com/EdwardsLab.

Results and Discussion

Prophage genometrics

At the time of analysis, the SEED contained 11,941 bacterial genomes spanning 34 different phyla of Archaea and Bacteria (Fig. 2). A total of 36,465 prophages were identified in 9,883 bacterial hosts (82.76% of the bacteria have at least one prophage). Slightly fewer than one half of the prophages that were identified (15,765 prophages) contained an annotated Int protein, suggesting that they are phages with the potential to integrate. The mechanism of integration of the remaining phages could not be easily identified (see below). A total of 1,511,644 prophage ORFs were identified – of these, 45.7% or 798,235 ORFs had no known function. In addition, 567 genomic regions had no similarity to any of the proteins in the SEED non-redundant database (15) although these putative prophages had many other hallmarks of prophages such as shorter genes, many consecutive genes on the same DNA strand, and %GC compositional skews. A total of 7,341 portal genes, 6,228 holin genes, 14,771 capsid genes, and 59,010 tail-related genes were identified in the prophage dataset (Table 1).

As we have shown previously, some families of genes are easier to identify than others, and gaps remain in our understanding of families of core phage proteins that we cannot readily identify from sequence homology alone (29–31).

Across the most highly represented phyla, there were an average of between 2.7 (Actinobacteria) and 3.8 (Proteobacteria) prophages found per genome. Within phyla there were also differences (although not statistically significant) in the prophage genome content. For example, *Sodalis glossinidius*, a proteobacterium, had the most prophages at 29, while several phyla (Caldiserica, Deinococcus–Thermus, Dictyoglomi, Elusimicrobia, and Fibrobacteres) had no prophages, Spirochetes had the shortest prophage genomes on average (25,595 bp), and Bacteroidetes had the longest (46,059 bp) (Fig. 3). Across all of the prophages, the mean and median prophage genome length was 36,804 bp and 32,352 bp, respectively. Shorter prophage genomes may indicate that the phage is no longer viable and is in the process of being degraded or removed from the host's genome (32). Alternatively, different genome lengths may represent different packaging strategies used by prophages associated with different phyla (33).

Several prophage genome characteristics correlate with the host's characteristics (34). For example, the prophage's and host's GC content correlated strongly ($R^2 = 0.95$; Fig. 4). As we have shown previously, the GC content of phage genomes can be accurately used to predict the phyla of the host associated with a phage (to >95% accuracy), however GC content can not be accurately used to predict the phages' host at lower taxonomic levels (35). Even though %GC content correlates at the phyla level, there are not enough degrees of freedom in %GC calculations to separate thousands of phages and their hosts. Percent GC is basically a *k*-mer DNA profile with $k = 1$. Higher *k*-mer DNA profiles increase the correlation between phage and host, but when *k* exceeds 7 nucleotides the data become sparse and over fit the predictions (35).

Phage integrase

Phage integrases are responsible for catalyzing the site-specific recombination event that inserts the phages into the hosts' genome (36–41), as well as during replication to help in phage and plasmid genome maintenance. A total of 18,352 prophages (43%) contained an integrase-like protein from one of the two main families – the tyrosine and serine recombinases – named after the catalytic residue involved in recombination. The tyrosine recombinases include the well-studied integrase employed by phage λ to insert its own genome into the bacterial chromosome (2), while the serine recombinases are exemplified by the integrases of *Streptomyces* phage phiC31 (42) and *Mycobacteria* phage Bxb1 (43). The 24,936 integrase sequences were clustered at a 90% sequence identity threshold using CD-HIT, which resulted in 8,536 clusters. Representatives from each cluster were used to generate a smaller dataset with which to work.

Five-mer profiles of the amino acid sequences were used to identify an eight-residue motif in one of the catalytic region of tyrosine recombinases with a consensus sequence of **HDLRH**TF**A** (Fig. 5A). The very strong conservation of five of the amino acids (bolded) suggests that they participate in the catalytic activity of these recombinases. Several of these amino acids were identified earlier as invariant among tyrosine integrases or nearly so (44–46), and both of the histidines, the leucine, and the arginine have been experimentally demonstrated to participate in DNA recognition, binding, and/or catalysis (41). Alignment of one hundred randomly selected tyrosine integrases from the condensed integrase set shows that the proteins that contained the motif were grouped together (Fig. 5B), offering a potential route to identify tyrosine recombinases.

tRNA presence in prophages compared to tRNA integration occurrences

Many phages encode tRNA genes, which enhance phage fecundity (47, 48). Several hypotheses have been proposed to explain the presence of tRNA genes in phage genomes, including providing tRNAs to compensate for missing or weakly expressed genes in the host (49). A previous study of 37 phage genomes from 15 bacterial genomes demonstrated that phage tRNA genes compensate for codon bias between host and phage, especially when a codon is common in the phage and rare in the host (50). In addition, tRNA genes carried by the phage may replace those interrupted during phage insertion into the genome at tRNA genes (51, 52). It is also possible that encoding multiple tRNA genes provides the phage genome alternative attachment sites to integrate into the host's genome in case the preferred attachment site is mutated and/or occupied by another phage.

A set of 17,539 tRNA genes were identified from the phages in our total dataset and condensed to a subset of 3,833 tRNA genes (as described for the integrase genes). As expected, the tRNA gene profiles in the phages reflect the number of synonymous codons for each tRNA (Fig. 6). However, we found that tRNA-Met is one of the most commonly carried tRNA genes among the prophages we studied. Phages often encode a peptide deformylase enzyme that removes the formyl group from newly synthesized proteins (53, 54). We propose that the methionine tRNA genes encoded by the prophages are included to increase the overall rate of translation when the phage enters the lytic cycle. Together, the peptide deformylase and tRNA-Met suggest that phage protein synthesis saturates the host's protein synthesis machinery and that, in many cases, optimal production of phage progeny requires protein translation capacity which exceeds that provided by the host.

We explored the hypothesis that the tRNA genes compensate for codon bias between the host and the phage. For a tRNA gene found in a prophage, the anticodon was converted to the genomic DNA codon it’s associated with. This codon’s modal usage was calculated for the prophage’s genome, as well as for the host’s genome without the prophage sequence. The codon’s modal usage in the phage and host was compared using the χ^2-test. We found that 41.3% of the codons associated with the tRNA found in a prophage had statistically higher modal usage in the phage than in the host (Fig. 6).

We applied the same approach to compare the phage’s modal usage of the same codon with an expected usage assuming an even distribution given the number of synonymous codons available. This comparison revealed that 31% of the codon associated with the tRNA found in a prophage had a higher usage in the phage than expected by chance. Together, these results suggest that phage-encoded tRNA are supplementing limitations in the hosts translational machinery to ensure that phage particles are produced with high efficiency upon induction.

Genome Architecture

Phages have a mosaic genome architecture (55, 56), in which the genome is organized into groups or clusters of genes. Mosaicism has been proposed to be driven by illegitimate recombination, but it is not clear from previous genomic analyses whether the mosaic boundaries are between clusters of related genes or between individual genes (32, 53, 57).

We identified the fourteen most commonly occurring and well-annotated genes within the prophage dataset (Table 1). The majority (66.9%; 10,542/15,765) of these predicted prophages had the integrase gene located near one of the ends of the prophage genome. We generated a genome data set where all the genomes were organized with the integrase at the same end of the genome, in some cases by reverse-complementing the genome to place the integrase gene consistently at the left end of every genome, and in other cases by recircularizing the genome *in silico* and introducing a break that placed the integrase at the left end.

The resulting alignment allowed us to compare genome architecture across all of the prophages, and revealed that gene order is conserved across phage genomes (Fig. 7a), suggesting that there is an advantageous order of clusters of related function within the genomes of prophages. To highlight the clusters, we used a modified Smith-Waterman approach, essentially ascribing a single letter to each of the functions and then aligning the genomes based on the order of the functions with a +1 score for a match and no penalty for a mismatch (Fig. 7b). This shows that the gene order across 1,537 phage genomes is highly conserved, reminiscent of the early—middle—late gene organization of many well characterized phages and eukaryotic viruses (58–60). We also visualized the presence of each gene at its location in the genome using a gene density plot (Fig. 8), which highlights the separation of the genes based on location in the genome. The combined results from these alignments strongly indicate that the mosaicism in phage genomes occurs at the gene-cluster level, and that phage genomes are not comprised of random hybrids of many genomes. Instead, recombination (illegitimate or otherwise) appears to be constrained to result in a highly conserved gene order where the early genes, including those encoding entry and DNA replication functions, precede those encoding packaging functions, which in turn precede tail formation and host lysis functions.

The phage genome is therefore not a collection of genes that have been acquired at random by recombination, but rather analogous to a computer motherboard, where modules may be replaced provided they are slotted into the appropriate gene expression framework. Thus, a viable phage is more likely to be produced if genes are placed in the appropriate modules.

Conclusions

Here we presented an analysis of over 11,000 bacterial genomes from which we identified 36,488 prophages. Many phages appear to be limited by initiation of translation by the host's machinery, and may increase translation rates by carrying their own tRNA genes, effectively increasing the availability of both tRNAs loaded with methionine and of peptide deformylase. We have also demonstrated that phages maintain a highly conserved gene order that suggests phage genome mosaicism is limited to clusters of conserved genes rather than individual genes.

Table 1. Frequency of identification of different genes among the prophages. ORFs were identified from 36,465 prophages.

Function	**Number of observations**
Integrase	24,936
Single-stranded binding protein	2,556
Endonuclease	4,043
DNA Polymerase	5,011
Protease	6,575
Terminase	15,657
Portal	7,341
Capsid	14,771
Tail-related	59,010
Baseplate	3,728
Holin	6,228
Lysin	5,983
Toxic	132
Transposase	5,139
Hypothetical	737,992
Other	645,956
Total	**1,511,644**

Figure legends

Fig. 1. Workflow of the methodology used to identify prophages in complete microbial genomes.

Fig. 2. Bacterial genomes and the number of prophage regions per genome (blue bars). The phylogenetic tree is based on the NCBI taxonomy and was visualized using iTOL (http://itol.embl.de/). Selected genera have been annotated on the tree.

Fig. 3. (A) Box and whisker plot denoting number of prophages identified per host, grouped by six phyla with most prophages. **(B)** Length (bp) of prophages identified grouped by six phyla with most prophages. In both A and B plots, the center line denotes the median of the distribution while box edges denote one standard deviation above or below the median.

Fig. 4. The %GC content of prophages shows strong correlation to %GC content of their hosts (R^2 = 0.94735). There is no correlation between prophage %GC content and prophage length (upper panel).

Fig. 5 (A) Residue frequency of the eight-mer motif contained in the catalytic domain of phage HP1-like integrases. The bars represent the frequency with which each amino acid was found at each of the eight positions in the motif. The consensus sequence is HDLRHTFA. **(B)** A phylogenetic tree built from one hundred randomly selected integrases shows clusterings of integrase containing the eight-mer motif (highlighted in red).

Fig. 6. The frequency of different tRNAs identified within prophage regions. For each amino acid, the total number of phage-encoded tRNA is shown in parentheses above each graph. The percentage of those tRNAs for each of the possible anticodons is shown in the figure.

Fig. 7. (A) Conserved prophage genome organization. Proteins in the phage genomes are colored by their annotation (integrase, portal, tail, holin, transposase, terminase, endonuclease, capsid, anti-termination, Holliday junction resolvase, baseplate assembly protein, primosomal, protease-like, lysin, toxin-encoding genes, and lysozyme), and the phage genomes were normalized to the largest prophage.

(B) Phage gene heatmap aligned using a modified Clustal alignment algorithm, as described in Materials and Methods.

Fig. 8. Gene density plots of the occurrence of phage marker genes along the genomes of 1,537 phage genomes. The genomes were rearranged as described in the text to place the integrase gene at the 5' end of the phage genome, and then genes were displayed as ordered by the modal frequency of each gene.

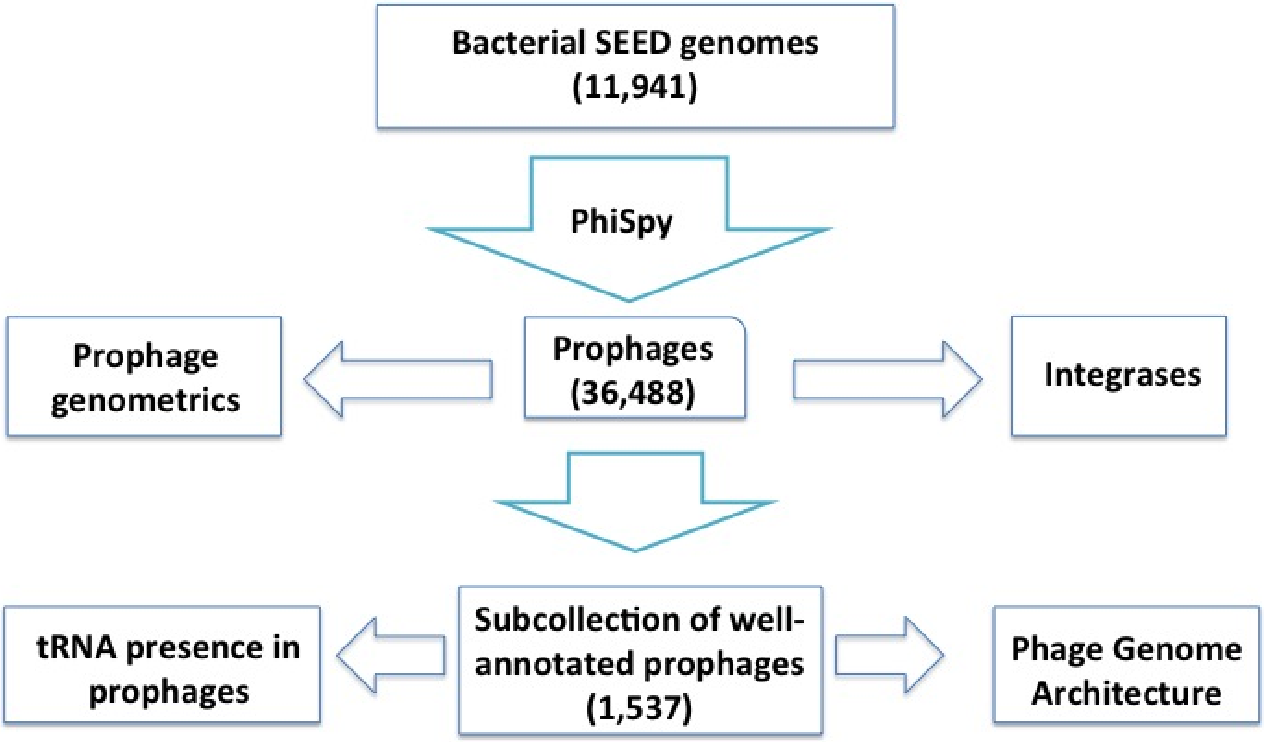

Bacterial SEED genomes (11,941)
PhiSpy
Prophage genometrics
Prophages (36,488)
Integrases
tRNA presence in prophages
Subcollection of well-annotated prophages (1,537)
Phage Genome Architecture

Fig. 1.

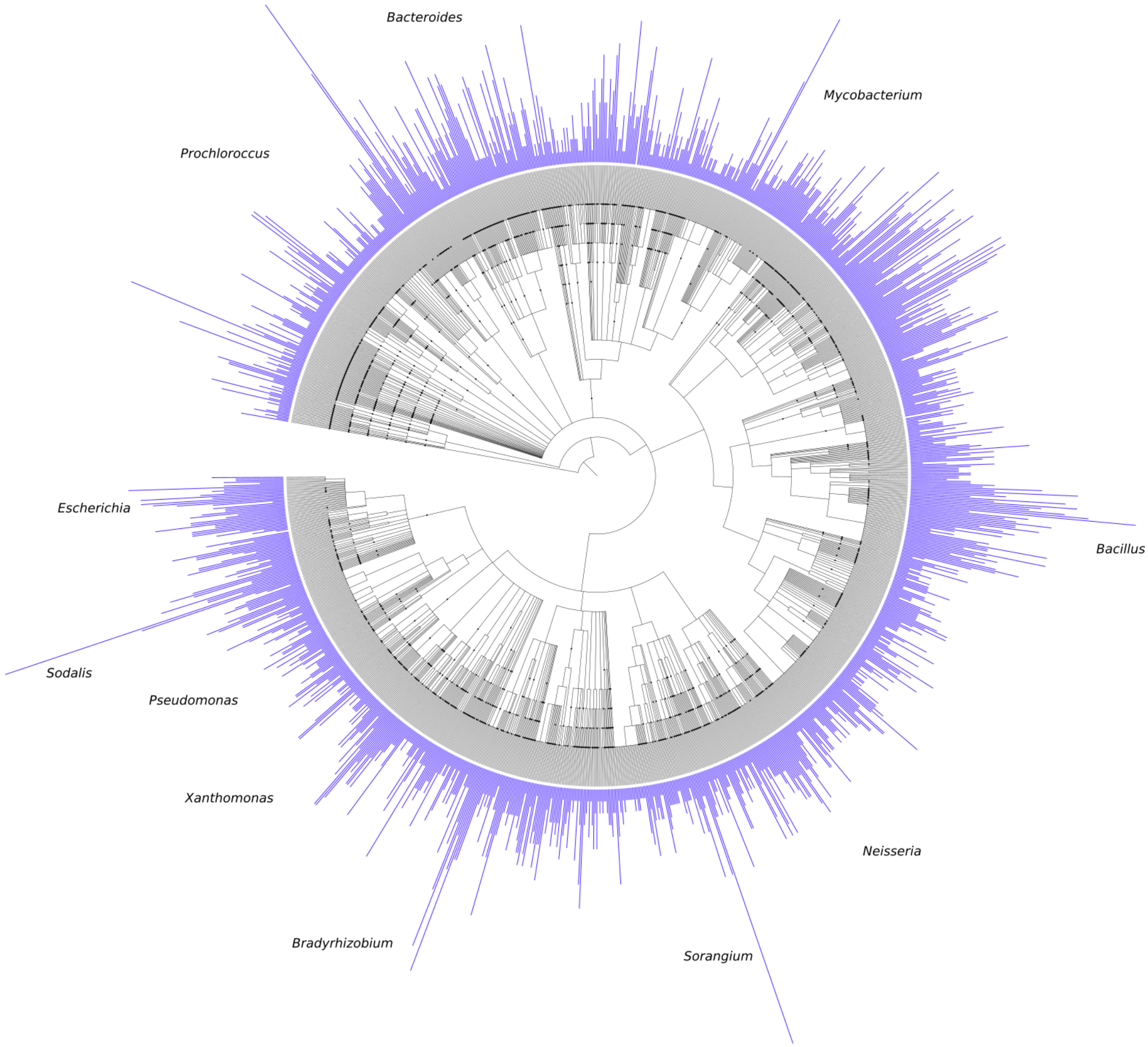

Bacteroides
Mycobacterium
Prochloroccus
Escherichia
Bacillus
Sodalis
Pseudomonas
Xanthomonas
Neisseria
Bradyrhizobium
Sorangium

Fig. 2.

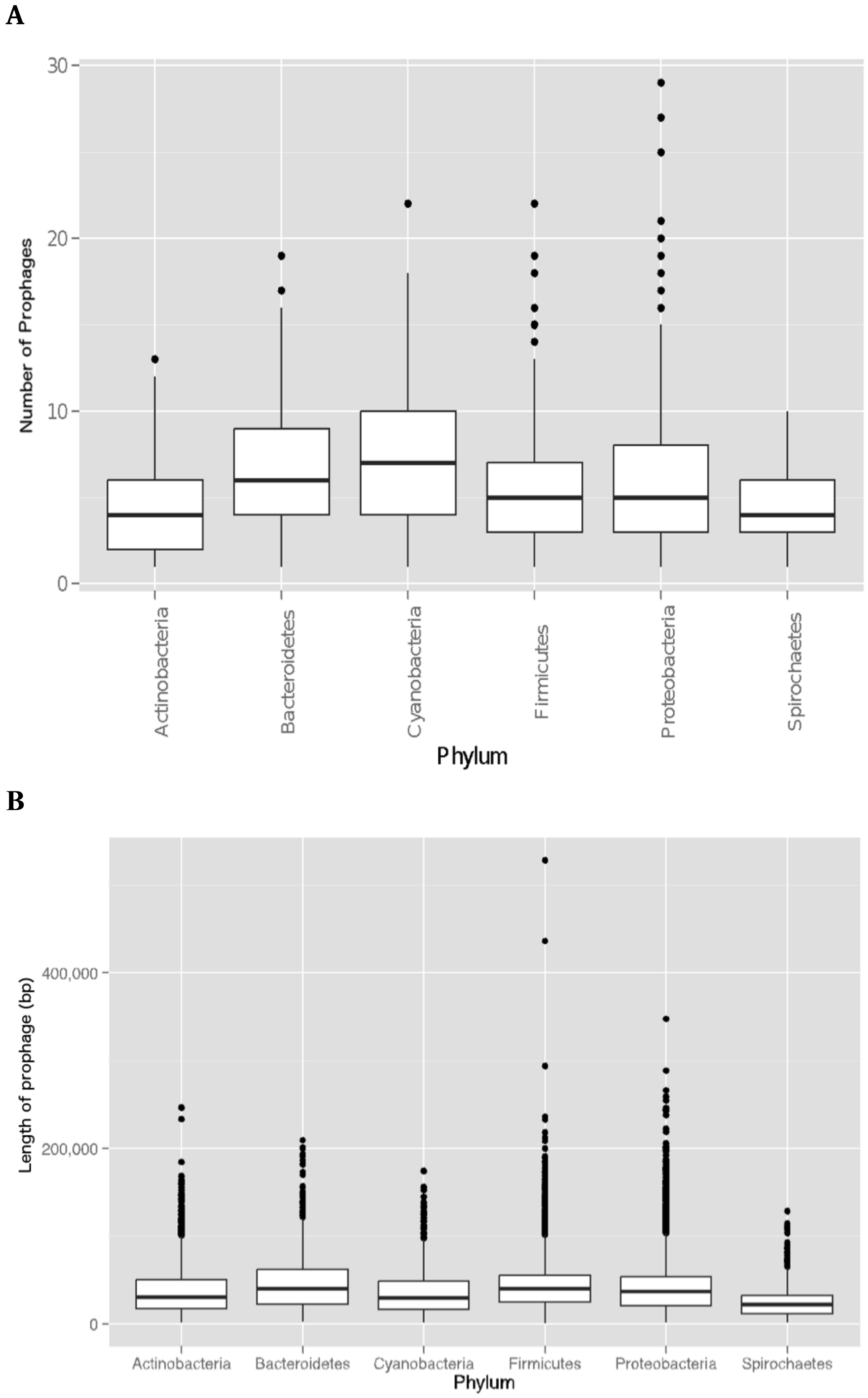

A
Number of Prophages
0
10
20
30
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Proteobacteria
Spirochaetes
Phylum
B
Length of prophage (bp)
0
200,000
400,000
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Proteobacteria
Spirochaetes
Phylum

Fig. 3.

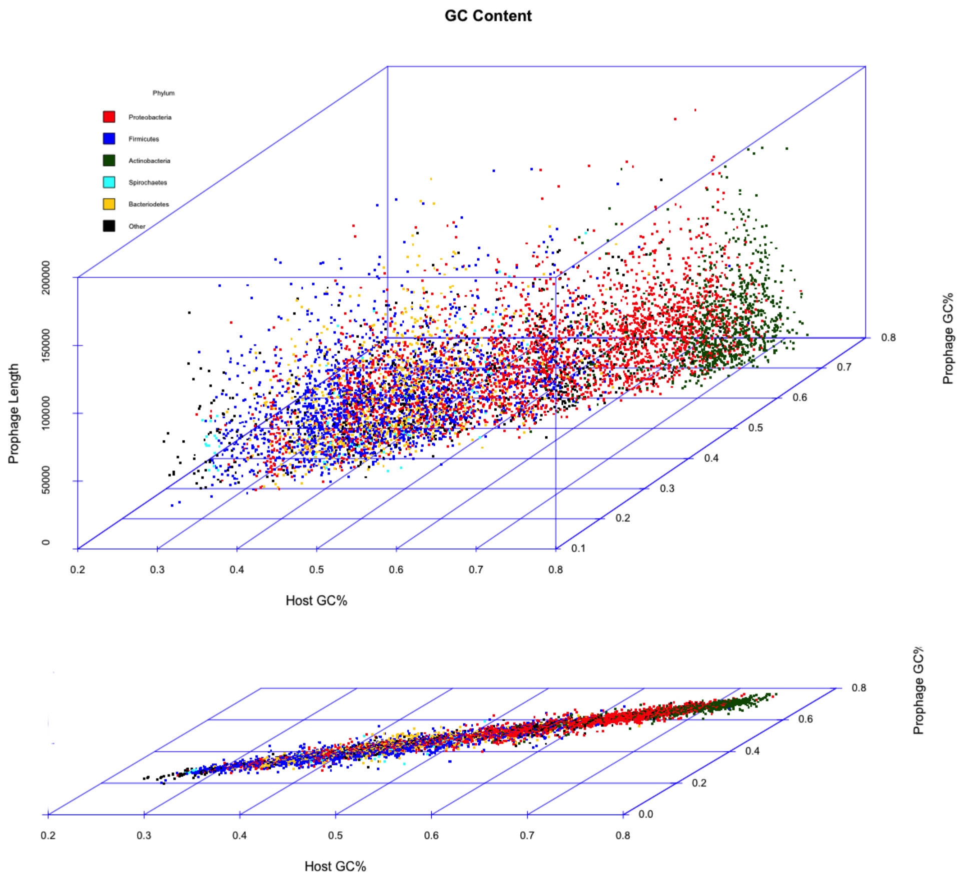

GC Content
Phylum
🟥 Proteobacteria
🟦 Firmicutes
🟩 Actinobacteria
🟦 Spirochaetes
🟨 Bacteriodetes
⬛ Other
Prophage Length
0
50000
100000
150000
200000
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Host GC%
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Prophage GC%
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Host GC%
0.0
0.2
0.4
0.6
0.8
Prophage GC%

Fig. 4.

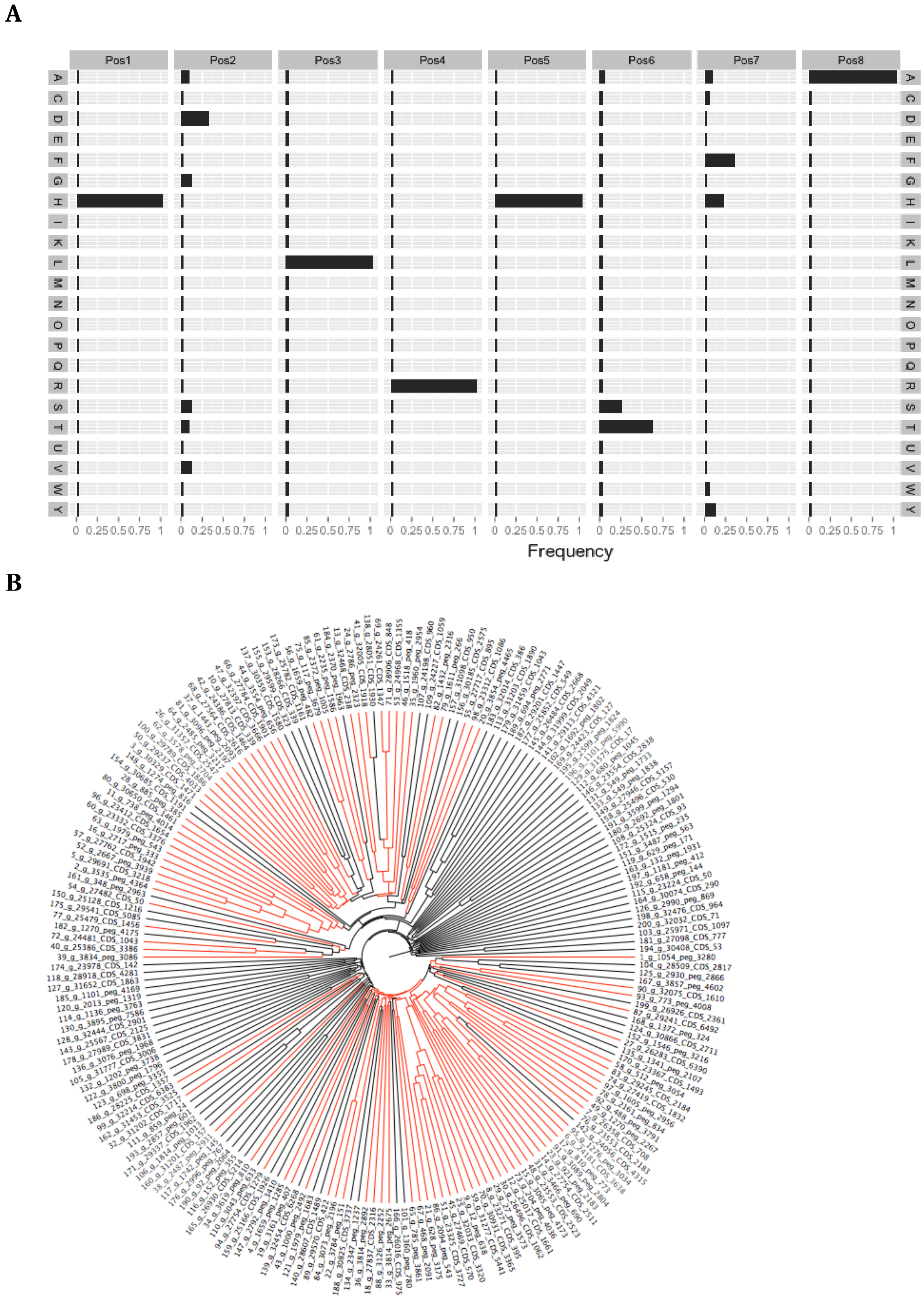

A
Pos1
Pos2
Pos3
Pos4
Pos5
Pos6
Pos7
Pos8
A
C
D
E
F
G
H
I
K
L
M
N
O
P
Q
R
S
T
U
V
W
Y
0 0.25 0.5 0.75 1
Frequency
B

Fig. 5.

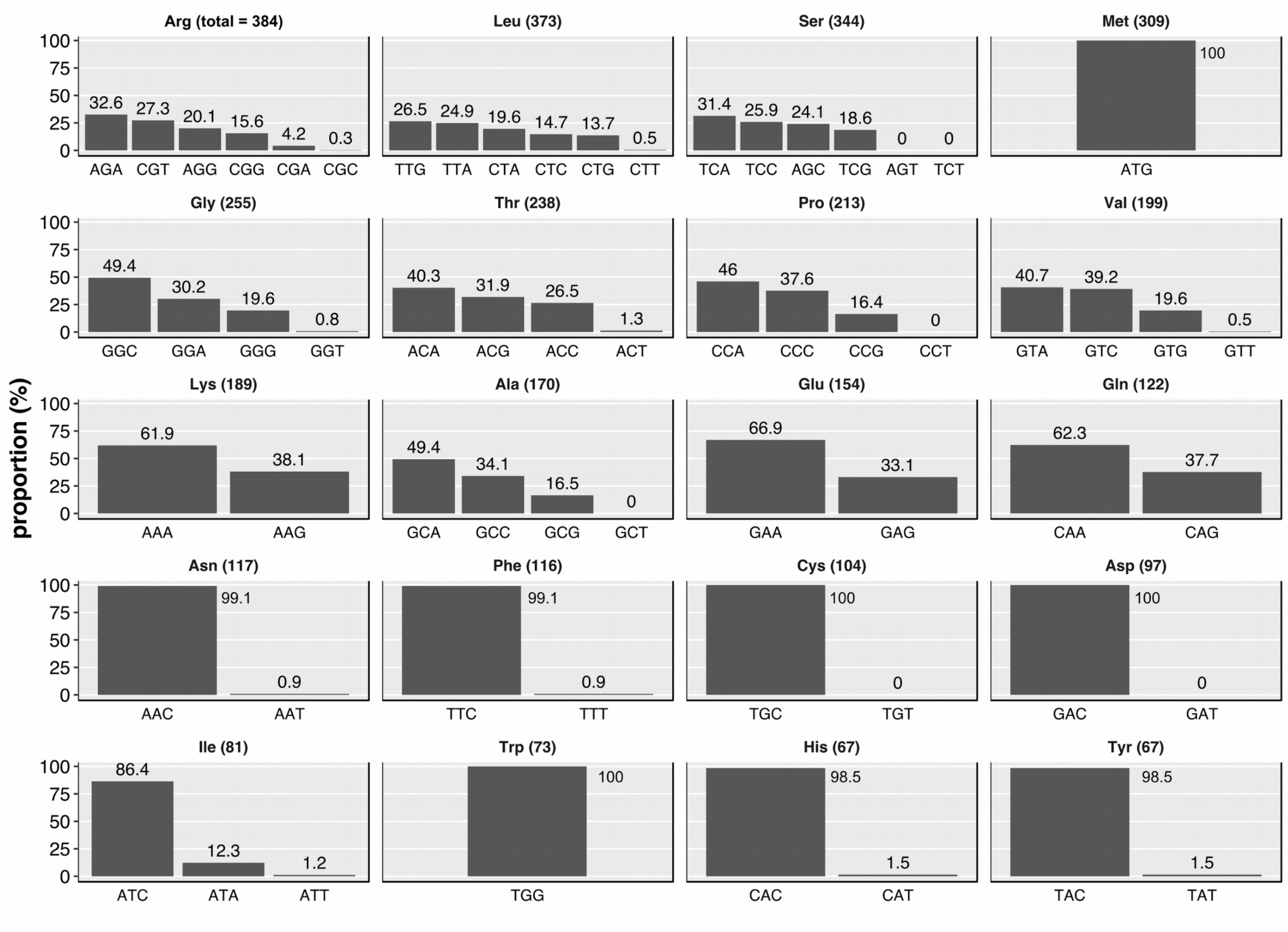
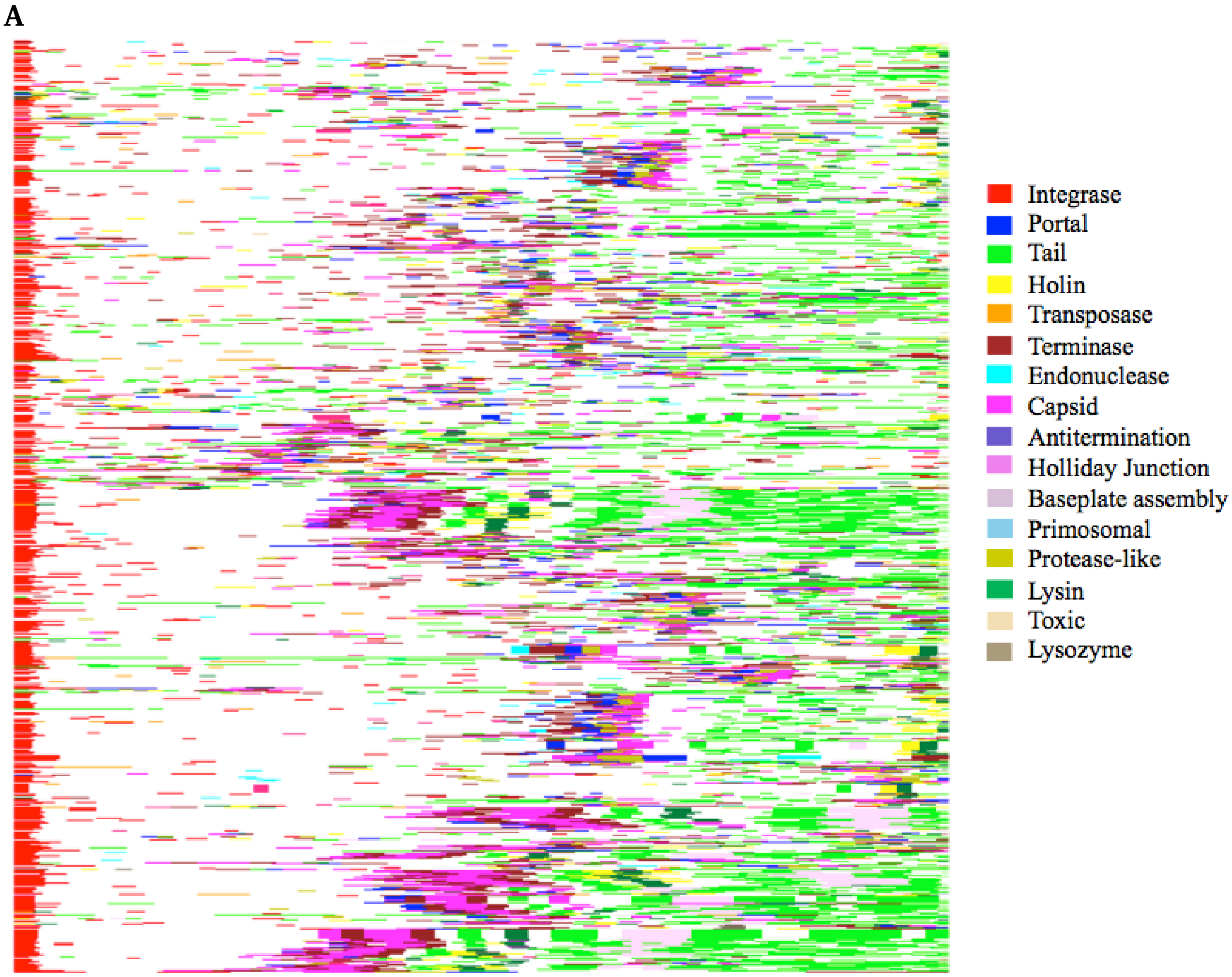

proportion (%)
0
25
50
75
100
Arg (total = 384)
32.6
27.3
20.1
15.6
4.2
0.3
AGA
CGT
AGG
CGG
CGA
CGC
Leu (373)
26.5
24.9
19.6
14.7
13.7
0.5
TTG
TTA
CTA
CTC
CTG
CTT
Ser (344)
31.4
25.9
24.1
18.6
0
0
TCA
TCC
AGC
TCG
AGT
TCT
Met (309)
100
ATG
Gly (255)
49.4
30.2
19.6
0.8
GGC
GGA
GGG
GGT
Thr (238)
40.3
31.9
26.5
1.3
ACA
ACG
ACC
ACT
Pro (213)
46
37.6
16.4
0
CCA
CCC
CCG
CCT
Val (199)
40.7
39.2
19.6
0.5
GTA
GTC
GTG
GTT
Lys (189)
61.9
38.1
AAA
AAG
Ala (170)
49.4
34.1
16.5
0
GCA
GCC
GCG
GCT
Glu (154)
66.9
33.1
GAA
GAG
Gln (122)
62.3
37.7
CAA
CAG
Asn (117)
99.1
0.9
AAC
AAT
Phe (116)
99.1
0.9
TTC
TTT
Cys (104)
100
0
TGC
TGT
Asp (97)
100
0
GAC
GAT
Ile (81)
86.4
12.3
1.2
ATC
ATA
ATT
Trp (73)
100
TGG
His (67)
98.5
1.5
CAC
CAT
Tyr (67)
98.5
1.5
TAC
TAT

Fig. 6.


A
Integrase
Portal
Tail
Holin
Transposase
Terminase
Endonuclease
Capsid
Antitermination
Holliday Junction
Baseplate assembly
Primosomal
Protease-like
Lysin
Toxic
Lysozyme

B

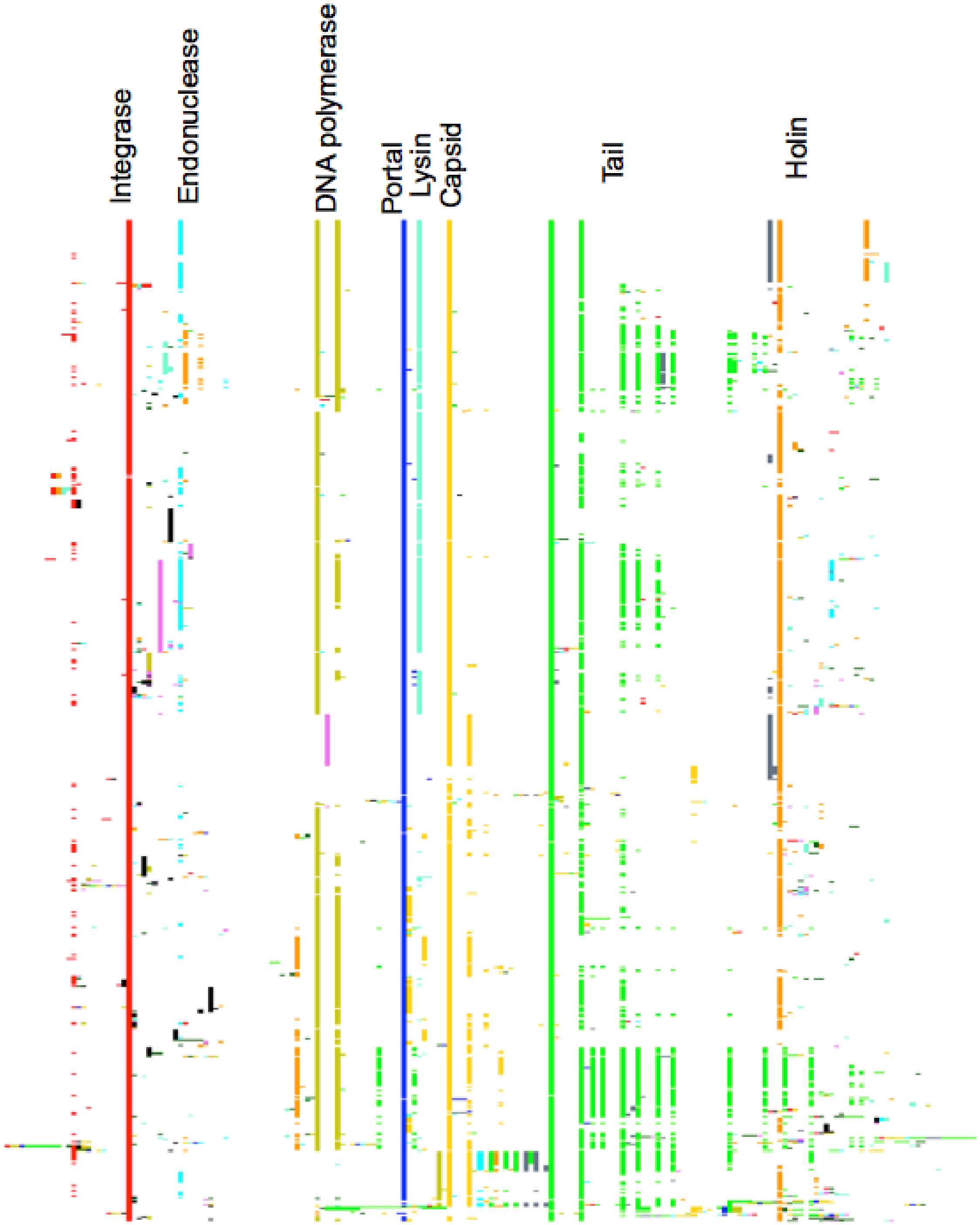

Integrase
Endonuclease
DNA polymerase
Portal
Lysin
Capsid
Tail
Holin

Fig. 7.

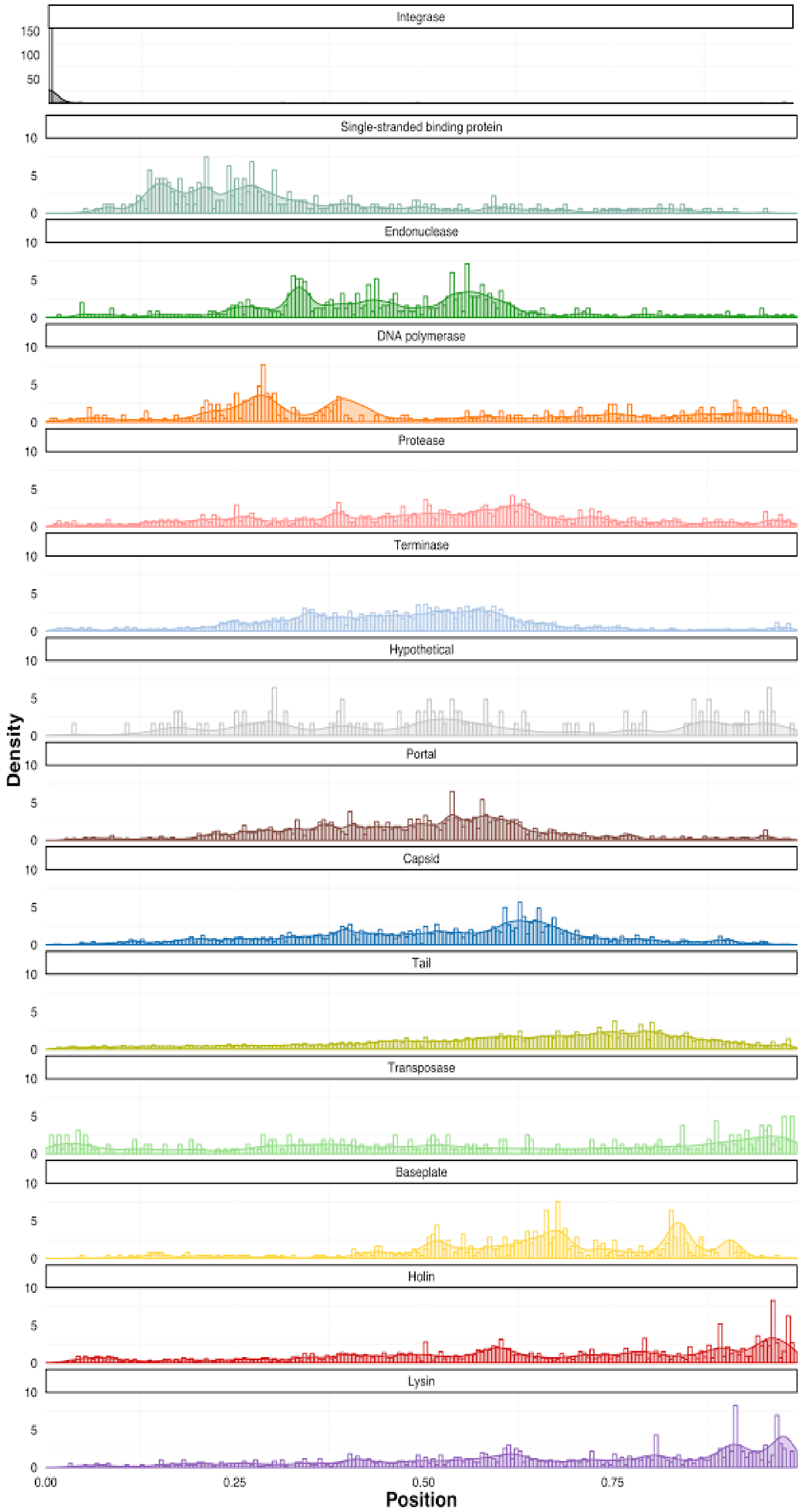

Integrase
150
100
50
Single-stranded binding protein
Endonuclease
DNA polymerase
Protease
Terminase
Hypothetical
Portal
Capsid
Tail
Transposase
Baseplate
Holin
Lysin
10
5
0
Density
0.00
0.25
0.50
0.75
Position

Fig. 8.

References

1. Prentki,P., Chandler,M. and Caro,L. (1977) Replication of the prophage P1 during the cell cycle of Escherichia coli. *Mol. Gen. Genet. MGG*, **152**, 71–76.

2. Campbell,A.M. (1992) Chromosomal insertion sites for phages and plasmids. *J. Bacteriol.*, **174**, 7495–7499.

3. Campbell,A. (2007) Phage integration and chromosome structure. A personal history. *Annu. Rev. Genet.*, **41**, 1–11.

4. Maloy,S.R., Stewart,V.J. and Taylor,R.K. (1996) Genetic Analysis of Pathogenic Bacteria: A Laboratory Manual Cold Spring Harbor Laboratory Press.

5. Brüssow,H., Canchaya,C. and Hardt,W.-D. (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev. MMBR*, **68**, 560–602, table of contents.

6. Bushman,F. (2001) Lateral DNA Transfer: Mechanisms and Consequences CSHL Press, New York, NY.

7. Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Umayam,L., *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature*, **406**, 477–83.

8. McShan,W.M. and Nguyen,S.V. (2016) The Bacteriophages of Streptococcus pyogenes. In Ferretti,J.J., Stevens,D.L., Fischetti,V.A. (eds), *Streptococcus pyogenes: Basic Biology to Clinical Manifestations*. University of Oklahoma Health Sciences Center, Oklahoma City (OK).

9. Canchaya,C., Fournous,G., Chibani-Chennoufi,S., Dillmann,M.L. and Brüssow,H. (2003) Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.*, **6**, 417–424.

10. Modi,S.R., Lee,H.H., Spina,C.S. and Collins,J.J. (2013) Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature*, **499**, 219–222.

11. Mann,N.H., Cook,A., Millard,A., Bailey,S. and Clokie,M. (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. *Nature*, **424**, 741–741.

12. Knowles,B., Silveira,C.B., Bailey,B.A., Barott,K., Cantu,V.A., Cobián-Güemes,A.G., Coutinho,F.H., Dinsdale,E.A., Felts,B., Furby,K.A., *et al.* (2016) Lytic to temperate switching of viral communities. *Nature*, **531**, 466–470.

13. Hatfull,G.F. (2008) Bacteriophage genomics. *Curr. Opin. Microbiol.*, **11**, 447–453.

14. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M., *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

15. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M., *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206-214.

16. Zhou,Y., Liang,Y., Lynch,K.H., Dennis,J.J. and Wishart,D.S. (2011) PHAST: A Fast Phage Search Tool. *Nucleic Acids Res.*, **39**, W347–W352.

17. Bose,M. and Barber,R.D. (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *Silico Biol*, **6**, 223–7.

18. Fouts,D.E. (2006) Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*, **34**, 5839–51.

19. Lima-Mendez,G., Van Helden,J., Toussaint,A. and Leplae,R. (2008) Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinforma. Oxf. Engl.*, **24**, 863–865.

20. Roux,S., Enault,F., Hurwitz,B.L. and Sullivan,M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ*, **3**, e985.

21. Akhter,S., Aziz,R.K. and Edwards,R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.*, **40**, e126–e126.

22. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, **25**, 955–64.

23. Davis,J.J. and Olsen,G.J. (2010) Modal Codon Usage: Assessing the Typical Codon Usage of a Genome. *Mol. Biol. Evol.*, **27**, 800–810.

24. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–2.

25. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

26. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695-699.

27. Holmes,E.C. and Rambaut,A. (2004) Viral evolution and the emergence of SARS coronavirus. *Philos. Trans. R. Soc. B Biol. Sci.*, **359**, 1059–1065.

28. NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7-19.

29. Sanchez,S.E., Cuevas,D.A., Rostron,J.E., Liang,T.Y., Pivaroff,C.G., Haynes,M.R., Nulton,J., Felts,B., Bailey,B.A., Salamon,P., *et al.* (2015) Phage Phenomics: Physiological Approaches to Characterize Novel Viral Proteins. *J. Vis. Exp.*, 10.3791/52854.

30. Seguritan,V., Feng,I.-W., Rohwer,F., Swift,M. and Segall,A.M. (2003) Genome sequences of two closely related Vibrio parahaemolyticus phages, VP16T and VP16C. *J. Bacteriol.*, **185**, 6434–6447.

31. Rohwer,F., Segall,A., Steward,G., Seguritan,V., Breitbart,M., Wolven,F. and Azam,F. (2000) The complete genomic sequence of the marine phage Roseophage SIOI shares homology with nonmarine phages. *Limnol. Oceanogr.*, **45**, 408–418.

32. Hendrix,R.W., Smith,M.C.M., Burns,R.N., Ford,M.E. and Hatfull,G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci.*, **96**, 2192–2197.

33. Luque,A. and Reguera,D. (2013) Theoretical Studies on Assembly, Physical Stability and Dynamics of Viruses. In Mateu,M.G. (ed), *Structure and Physics of Viruses*, Subcellular Biochemistry. Springer Netherlands, pp. 553–595.

34. Roux,S., Hallam,S.J., Woyke,T. and Sullivan,M.B. (2015) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, **4**, e08490.

35. Edwards,R.A., McNair,K., Faust,K., Raes,J. and Dutilh,B.E. (2015) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.*, 10.1093/femsre/fuv048.

36. Rajeev,L., Malanowska,K. and Gardner,J.F. (2009) Challenging a paradigm: the role of DNA homology in tyrosine recombinase reactions. *Microbiol. Mol. Biol. Rev. MMBR*, **73**, 300–309.

37. Hirano,N., Muroi,T., Takahashi,H. and Haruki,M. (2011) Site-specific recombinases as tools for heterologous gene integration. *Appl. Microbiol. Biotechnol.*, **92**, 227–239.

38. Das,B., Martínez,E., Midonet,C. and Barre,F.-X. (2013) Integrative mobile elements exploiting Xer recombination. *Trends Microbiol.*, **21**, 23–30.

39. Askora,A., Abdel-Haliem,M.E.F. and Yamada,T. (2012) Site-specific recombination systems in filamentous phages. *Mol. Genet. Genomics*, **287**, 525–530.

40. Krupovic,M. and Forterre,P. (2015) Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Ann. N. Y. Acad. Sci.*, **1341**, 41–53.

41. Jayaram,M., Ma,C.-H., Kachroo,A.H., Rowley,P.A., Guga,P., Fan,H.-F. and Voziyanov,Y. (2015) An Overview of Tyrosine Site-specific Recombination: From an Flp Perspective. *Microbiol. Spectr.*, **3**.

42. Smith,M.C.M., Brown,W.R.A., McEwan,A.R. and Rowley,P.A. (2010) Site-specific recombination by phiC31 integrase and other large serine recombinases. *Biochem. Soc. Trans.*, **38**, 388–394.

43. Ghosh,P., Kim,A.I. and Hatfull,G.F. (2003) The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of attP and attB. *Mol. Cell*, **12**, 1101–1111.

44. Argos,P., Landy,A., Abremski,K., Egan,J.B., Haggard-Ljungquist,E., Hoess,R.H., Kahn,M.L., Kalionis,B., Narayana,S.V. and Pierson,L.S. (1986) The integrase family of site-specific recombinases: regional similarities and global diversity. *EMBO J.*, **5**, 433–440.

45. Esposito,D. and Scocca,J.J. (1997) The integrase family of tyrosine recombinases: evolution of a conserved active site domain. *Nucleic Acids Res.*, **25**, 3605–3614.

46. Nunes-Düby,S.E., Kwon,H.J., Tirumalai,R.S., Ellenberger,T. and Landy,A. (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res.*, **26**, 391–406.

47. Wilson,J.H. (1973) Function of the bacteriophage T4 transfer RNA's. *J. Mol. Biol.*, **74**, 753–757.

48. Weiss,S.B., Hsu,W.T., Foft,J.W. and Scherberg,N.H. (1968) Transfer RNA coded by the T4 bacteriophage genome. *Proc. Natl. Acad. Sci. U. S. A.*, **61**, 114–121.

49. Kunisawa,T. (1992) Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host Escherichia coli. *J. Theor. Biol.*, **159**, 287–298.

50. Bailly-Bechet,M., Vergassola,M. and Rocha,E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res.*, **17**, 1486–1495.

51. Canchaya,C., Fournous,G. and Brüssow,H. (2004) The impact of prophages on bacterial chromosomes. *Mol. Microbiol.*, **53**, 9–18.

52. Casjens,S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol*, **49**, 277–300.

53. Frank,J.A., Lorimer,D., Youle,M., Witte,P., Craig,T., Abendroth,J., Rohwer,F., Edwards,R.A., Segall,A.M. and Burgin,A.B. (2013) Structure and function of a cyanophage-encoded peptide deformylase. *ISME J.*, **7**, 1150–1160.

54. Sharon,I., Battchikova,N., Aro,E.-M., Giglione,C., Meinnel,T., Glaser,F., Pinter,R.Y., Breitbart,M., Rohwer,F. and Béjà,O. (2011) Comparative metagenomics of microbial traits within oceanic viral communities. *ISME J.*, **5**, 1178–1190.

55. Muniesa,M., Recktenwald,J., Bielaszewska,M., Karch,H. and Schmidt,H. (2000) Characterization of a shiga toxin 2e-converting bacteriophage from an Escherichia coli strain of human origin. *Infect. Immun.*, **68**, 4850–4855.

56. Pope,W.H., Bowman,C.A., Russell,D.A., Jacobs-Sera,D., Asai,D.J., Cresawn,S.G., Jr,W.R.J., Hendrix,R.W., Lawrence,J.G., Hatfull,G.F., *et al.* (2015) Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife*, **4**, e06416.

57. Miller,E.S., Kutter,E., Mosig,G., Arisaka,F., Kunisawa,T. and Rüger,W. (2003) Bacteriophage T4 Genome. *Microbiol. Mol. Biol. Rev.*, **67**, 86–156.

58. Marrs,C.F. and Howe,M.M. (1990) Kinetics and regulation of transcription of bacteriophage Mu. *Virology*, **174**, 192–203.

59. Pero,J., Hannett,N.M. and Talkington,C. (1979) Restriction cleavage map of SP01 DNA: general location of early, middle, and late genes. *J. Virol.*, **31**, 156–171.

60. Ventura,M. and Brüssow,H. (2004) Temporal Transcription Map of the Virulent Streptococcus thermophilus Bacteriophage Sfi19. *Appl. Environ. Microbiol.*, **70**, 5041–5046.