MICROBIAL GENOMICS

RESEARCH ARTICLE
Kupczok *et al.*, *Microbial Genomics* 2022;8:000915
DOI 10.1099/mgen.0.000915

Co-transfer of functionally interdependent genes contributes to genome mosaicism in lambdoid phages

Anne Kupczok[1,*], Zachary M. Bailey[2], Dominik Refardt[3] and Carolin C. Wendling[2]

Abstract

Lambdoid (or Lambda-like) phages are a group of related temperate phages that can infect *Escherichia coli* and other gut bacteria. A key characteristic of these phages is their mosaic genome structure, which served as the basis for the 'modular genome hypothesis'. Accordingly, lambdoid phages evolve by transferring genomic regions, each of which constitutes a functional unit. Nevertheless, it is unknown which genes are preferentially transferred together and what drives such co-transfer events. Here we aim to characterize genome modularity by studying co-transfer of genes among 95 distantly related lambdoid (pro-)phages. Based on gene content, we observed that the genomes cluster into 12 groups, which are characterized by a highly similar gene content within the groups and highly divergent gene content across groups. Highly similar proteins can occur in genomes of different groups, indicating that they have been transferred. About 26% of homologous protein clusters in the four known operons (i.e. the early left, early right, immunity and late operon) engage in gene transfer, which affects all operons to a similar extent. We identified pairs of genes that are frequently co-transferred and observed that these pairs tend to be near one another on the genome. We find that frequently co-transferred genes are involved in related functions and highlight interesting examples involving structural proteins, the cI repressor and Cro regulator, proteins interacting with DNA, and membrane-interacting proteins. We conclude that epistatic effects, where the functioning of one protein depends on the presence of another, play an important role in the evolution of the modular structure of these genomes.

DATA SUMMARY

The genomes used in this research are publicly available (Table S1, available in the online version of this article) and can be retrieved from GenBank (https://www.ncbi.nlm.nih.gov/genbank/) or RefSeq (https://www.ncbi.nlm.nih.gov/refseq/). All supporting data are available in supplementary tables. Source code and documentation to calculate wGRR is available under GPLv2 (https://github.com/annecmg/GRRpair).

INTRODUCTION

(Bacterio-)phages, i.e. bacterial viruses, are considered the most abundant biological entity on earth and play a fundamental role in bacterial ecology and evolution. Phages can either be virulent or temperate. Virulent phages follow the lytic cycle, where phages replicate and lyse their host. Temperate phages can choose between the lytic cycle or a lysogenic cycle during which phage DNA is integrated into the host genome as a prophage that is replicated with the host cell. Prophages, highly relevant for bacterial fitness and evolution [1, 2], are very common in bacterial genomes. It is estimated that about 75% of bacteria are lysogens, i.e. they contain one or more prophages in their genome, which can, in extreme cases, make up to 20% of the bacterial genome content [3–6].

Phages have an astonishing level of genetic diversity [7, 8]. Even phages that infect the same bacterial host often share no sequence similarity [9–11]. This diversity is created and maintained by a high rate of *de novo* mutations and the acquisition of genetic material through

Received 02 July 2022; Accepted 19 October 2022; Published 24 November 2022
Author affiliations: [1]Bioinformatics Group, Wageningen University & Research, Wageningen, Netherlands; [2]ETH Zürich, Institute of Integrative Biology, Universitätstrasse 16, Zürich, Switzerland; [3]Institute of Natural Resource Sciences, Zürich University of Applied Sciences, Campus Grüental, Wädenswil, Switzerland.
***Correspondence:** Anne Kupczok, anne.kupczok@wur.nl
Keywords: co-transfer; horizontal gene transfer; phage evolution; phage lambda.
Abbreviations: ANI, average nucleotide identity; API, application programming interface; HGT, horizontal gene transfer; PC, protein cluster; PCO, Protein cluster Co-Occurrence; PCT, Protein cluster Co-Transfer; wGRR, weighted gene repertoire relatedness.
Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Six supplementary figures and seven supplementary tables are available with the online version of this article.

Impact Statement

Temperate phages, viruses that can integrate their own genetic material into bacterial genomes, are pervasive mobile genetic elements that can influence bacterial fitness in manifold ways. The *Escherichia coli* phage Lambda has been a model phage of molecular biology for decades. Lambdoid phages are highly prevalent in *Enterobacteria* such as *E. coli* and *Salmonella*, have a mosaic-like genome, the same genome architecture as Lambda and can recombine with phage Lambda. Nevertheless, these phages can be very distinct, and no lambdoid core genome exists. Although lambdoid phage genomes have been studied for decades, we know relatively little about how they evolve. Early observations led to the modular genome hypothesis, according to which phages are assemblages of genetic modules. But what determines the structure of these modules and which genes preferentially occur together in modules? In this study, we provide answers to these questions using a novel computational approach that allows us to infer gene transfer events between distantly related phages despite the absence of a core genome. We find that co-transfer of functionally related genes is frequent during the evolution of lambdoid phages. This suggests epistatic interactions among these genes, i.e. the co-transferred genes probably need to function together to ensure a viable phage. A prime example is the co-transfer of structural genes, such as genes encoding parts of the phage capsid or the phage tail. Additionally, we also find co-transfer of known interacting regulatory genes and co-transfer between functionally related genes that have so far been unknown to interact. Together, our analysis provides novel insights into the evolution of temperate phages. Moreover, our approach, which allows us to identify gene transfer in the absence of a core phylogeny, might be valuable for studying the evolution of other fast-evolving genomes, including viruses of other hosts.

horizontal gene transfer (HGT) events [12–14]. Recent genomic evidence revealed that gene flow is more prevalent in temperate phages, where it probably occurs between infecting phages and resident prophages or between active and defective prophages [15, 16]. Additionally, gene flow tends to occur among phages enriched for recombinases and transposases, and that are capable of non-homologous end joining [14]. This suggests that homologous and illegitimate recombination contribute to gene transfer in phages [14, 17, 18].

The 'modular theory of prophage evolution' [19, 20] states that these extensive gene transfers have created pervasive mosaicism. Accordingly, individual phages are composed of assemblages of shared modules [21–23], where as in a patchwork pattern, almost identical regions can alternate with highly divergent regions. This theory was further supported by several studies suggesting frequent gene flow among these phages [24–30].

Genome mosaicism was first described for lambdoid phages [31], a group of temperate tailed dsDNA enterobacteria phages that share a common genetic architecture and can recombine to form viable hybrids [32]. The family of lambdoid phages is sometimes regarded as a single biological species that draws functional genetic modules from a shared gene pool, and includes phages, prophages and defective prophages [24–30]. Since these modules evolve independently, it has even been suggested that they constitute minimal autonomously functional units, such as groups of genes that must function together and are independent of the other modules [24, 33]. The genomes of lambdoid phages constitute four different operons: the late operon comprises the morphogenetic proteins, the early left and the early right operons include phage genes transcribed early in the lytic cycle, and the immunity operon includes the repressor *c*I and the Rex system, known to abort lytic growth of phages [34].

Despite the gene synteny between lambdoid phage genomes, the sequence diversity of lambdoid phages is so high that there exists no lambdoid core genome [26]. The genome mosaicism that has been observed in pairwise genome comparisons suggests that some genes are preferentially transferred together. Nevertheless, it is unknown which genes are frequently co-transferred and which functions they encode. This study is based on 26 temperate lambdoid phages with the known ability to perform the lytic and lysogenic cycle (denoted focus phages). To restrict the analysis to active temperate phages, we only included (pro-)phages from databases with a high genome-wide similarity to the focus phages. Due to the high diversity of lambdoid phages and the absence of a core genome, it is impossible to infer gene transfer with phylogenetic approaches that require a fully resolved phylogeny. Instead, we here build on a previously established two-step approach [14]. First, we estimated similarity between two genomes based on protein identities of the homologous protein pairs. Second, we detected gene transfer by identifying highly similar proteins encoded on highly dissimilar genomes. Finally, we inferred co-transfer among lambdoid phages by detecting proteins that are frequently transferred together between the same pairs of genomes.

METHODS

Genome similarity calculations

For distantly related phages that do not share a core genome, weighted gene repertoire relatedness (wGRR) has recently been suggested as a measure to calculate pairwise genome similarity [14]. To calculate wGRR between a pair of phages, first, homologous proteins are detected as best bidirectional Blastp hits using Blast+ v2.10 [35]. Second, global identities (%) for these homologous protein pairs were determined using powerneedle, a modification of needle from the EMBOSS package for multiple input pairs [36]. wGRR represents the sum of all global protein identities divided by the number of proteins in the smaller genome and varies between 0 and 100 %. The script to calculate wGRR for a pair of genomes is available at https://github.com/annecmg/GRRpair.

Average nucleotide identity (ANI) was calculated with FastANI using a fragment length of 300 [37].

Data

Focus phages: This study is based on 26 temperate lambdoid phages (Table S1A). Sequences of eight of these phages (HK022, HK97, HK620, Φ80, Gifsy, P22, N15 and λ *PaPa*) were obtained from GenBank. The remaining focus phages were obtained from two sources and subsequently sequenced: phages with the prefix mEp (Mexican *Escherichia coli* phage) were isolated by Kameyama *et al.* [38] from human faecal samples in Mexico and were provided to us by Ing-Nang Wang (SUNY, Albany, NY, USA). Two phages could not be matched to their original designation and were renamed mEpX1 and mEpX2. A clear plaque mutant of phage mEp043 was used for DNA extraction because its lysate yielded a higher titre. All other phages are likely to be identical to those described in earlier publications [38–40]. Phages with the prefix HK were isolated by Elvera and Tarlochan Dhillon in the 1970s and early 1980s from different sources in Hong Kong (some of which are referenced in [41]), and were provided to us by Rodney King (Western Kentucky University, KY, USA).

High-titre lysates of all phages were prepared by confluent lysis [42] and purified (Lambda Midi Kit; Qiagen). DNA content of samples was quantified (Quant-iT DNA assay kit; Invitrogen) and 5 µg of genomic DNA per phage was used to prepare a shotgun library with multiplex identifiers, which was sequenced (Roche 454 Genome Sequencer FLX system with a Titanium kit; Roche Diagnostics). Sequencing was done at the Functional Genomics Centre Zürich, Switzerland.

Sequences were assembled using Newbler (Roche Diagnostics) and the assembly software of Geneious 5.4. As coverage was very high, this yielded in all but one case a single contig of the complete genome sequence. We obtained two contigs for phage mEp460, which were then combined using Sanger sequencing. Phage genomes were annotated using RAST searching the Virus domain and using genetic code 11 [43].

The ability to perform both the lytic and lysogenic cycle *in E. coli* is confirmed for the 26 focus phages (Table S1A [44–47]). The prophages can be induced by mitomycin C and (if they are no clear plaque mutants as indicated with the suffix c-1) are able to integrate into the genome of *E. coli* MG1655 (Fig. S1) [48]. We extend this list of 26 focus phages by including (i) temperate phage genomes from NCBI and (ii) predicted prophages from *E. coli* genomes.

Temperate phage genomes from NCBI: We searched for virus sequences with host bacteria from NCBI genomes (https://www.ncbi.nlm.nih.gov/genome/browse#!/viruses/) on 16 April 2020. We downloaded the corresponding GenBank files using the Entrez package in biopython. From this list, we retained those genomes that had an ANI of at least 95% to a focus phage resulting in 34 genomes. To discard incomplete prophages, one genome was filtered due to length (length <27 kb), resulting in 33 phages which we included in the analysis.

Predicted prophages from E. coli: We used the search term 'Escherichia coli[organism] AND genbank[Filter] AND complete genome[Title] NOT phage[Title]' to find all accession numbers of deposited *E. coli* genomes submitted to NCBI as of 18 May 2020 and downloaded them using the NCBI E-utilities toolkit from the NCBI nucleotide database [49], which resulted in a total of 1115 *E. coli* genomes. Subsequently, we used the downloaded accession numbers on the PHASTER database utilizing their API (application programming interface) to find the corresponding PHASTER prophage output of each *E. coli* genome [50]. PHASTER predicted a total of 4919 'intact' prophages from these 1115 genomes. Predicted prophage regions were extracted as a FASTA file and annotated using RASTtk (from PATRIC tools) utilizing the Virus domain and genetic code 11 [43]. Next, we compared these PHASTER-derived prophage sequences to the 26 focus phage sequences using an ANI cutoff of 95% [37]. We found 558 prophages with similarity to at least one of the 26 focus phages. Of these, 14 phages were filtered due to length (length <27 kb or length >5 Mb), resulting in a final list of 544 prophages.

Phage deduplication: The total set of 603 phages (26 focus phages, 33 NCBI phages and 544 *E. coli* prophages) were then deduplicated by grouping highly similar phages by both single linkage and an ANI cutoff of 99.8%. When choosing a representative, we prioritized (i) the focus phages, (ii) the phages from NCBI and (iii) the phage with the smallest name, i.e. the one originating from the earliest sequencing. We found that one prophage from CP010240.1 contained an insertion element and was thus excluded from the analysis. This resulted in 300 deduplicated phages (26 focus phages, ten phages from NCBI, 264 prophages).

Genome groups

We assigned the genomes of NCBI phages and prophages to the same group if they have a wGRR of at least 60% to one of the focus phages.

Protein clusters

First, we performed an all-against-all Blastp with Blast+ v2.10 [35]. Significant pairs (e-value <0.001) were aligned with powerneedle, and pairs with a global identity of at least 40% were clustered with mcl based on the global identities [51].

Gene transfer detection

We say that genes from different genome groups were *transferred* if their encoded proteins have a global identity of at least 80%. We calculated PCT (Protein cluster Co-Transfer) between two protein clusters as PCT=GI/(G1+G2-GI), where G1 (G2) is the number of different genome pairs between which protein cluster 1 (2) is transferred and GI is the number of genome pairs between which both clusters are transferred. We define two clusters to be *frequently co-transferred* when their PCT is greater than 0.5.

To compare the PCT statistic to co-occurrence of clusters, we also calculated the PCO (Protein cluster Co-Occurrence) as PCO=PI/(P1+P2-PI), where P1 (P2) is the number of different genomes that contain protein cluster 1 (2) and PI is the number of genome pairs that contain both protein clusters.

Functional annotation

We manually curated the operon structures by annotating nine model phages based on literature (HK022, HK225, HK620, HK629, Lambda, mEp043-c1, mEp460, P22, Phi80) [52, 53]. For each annotated protein in a protein cluster, all proteins in that cluster were assigned to the same operon. We observed that the annotation within protein clusters was generally consistent, i.e. cluster members were assigned to the same operon. Only four protein clusters formed exceptions by containing proteins belonging to the early left and to the early right operon; they were not assigned to any operon. Next, we manually curated the operon structure of all genomes by including genes into an operon, if they were in the same orientation and ≤10 nt from a neighbouring annotated gene or if they were in the same orientation, ≤ 200 nt from a neighbouring annotated gene and the annotation of the other flanking genes does not conflict with the transferred annotation. The resulting annotation is listed in Table S2.

Plotting was done in R using ggplot2 [54], and networks were visualized with Cytoscape [55].

RESULTS AND DISCUSSION

To identify frequently co-transferred genes and to characterize their function, we estimated gene transfer and gene co-transfer among temperate lambdoid phages (Fig. 1). To distinguish horizontal transfer from vertical inheritance, we aimed to find proteins

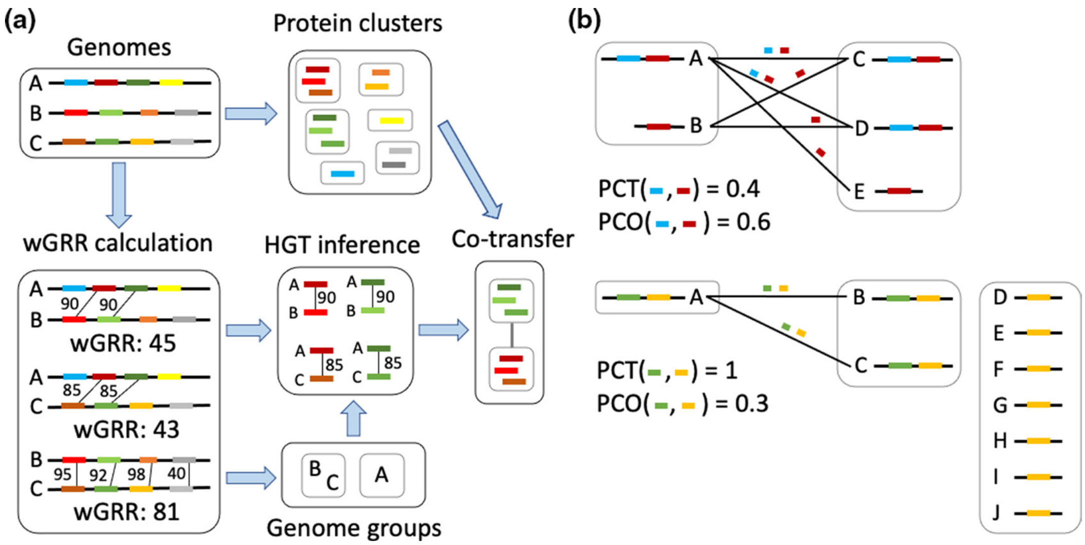
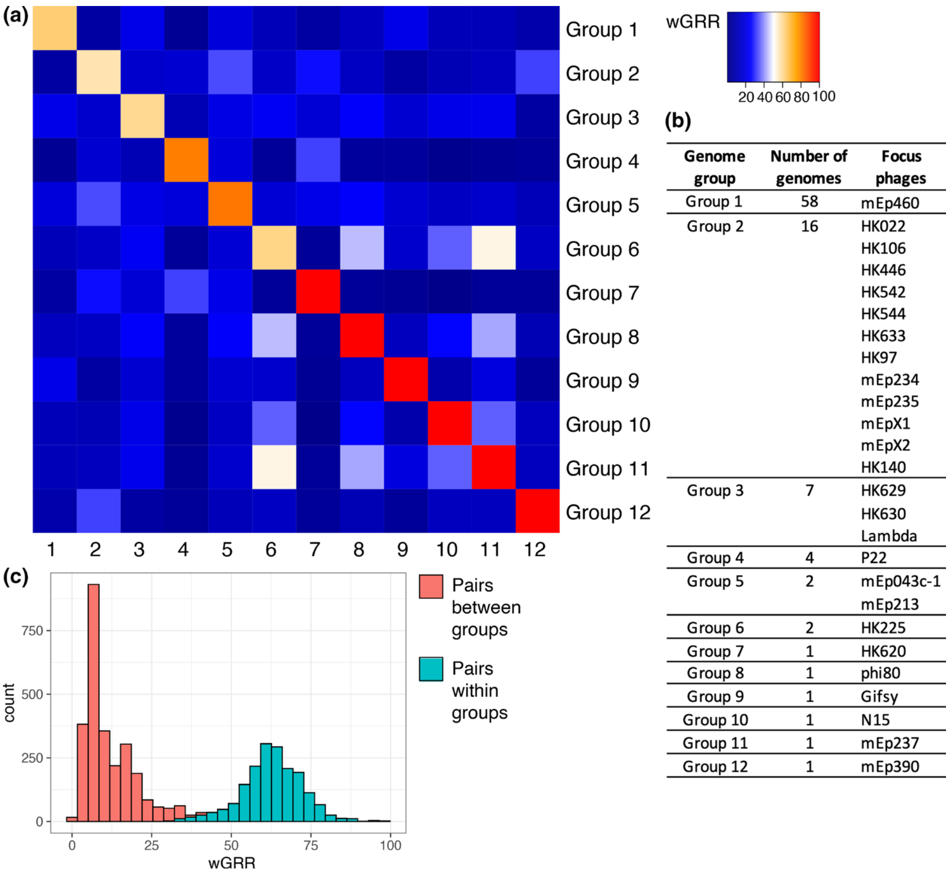


Fig. 1. Overview of the approach to detect horizontal transfer and co-transfer. (a) For a set of genomes, pairwise wGRR is calculated based on protein identities of homologous proteins. Genomes with high wGRR are grouped into genome groups. Next, horizontal gene transfer is inferred by detecting highly similar proteins encoded on genomes that belong to different groups. Lastly, co-transfer between protein clusters is inferred when proteins in the cluster are frequently transferred together. (b) Example calculations for PCT (protein cluster co-transfer) and PCO (protein cluster co-occurrence). Membership to genome groups is indicated by grey boxes. The blue and red genes are those co-occurring frequently (PCO>0.5) although they are not frequently co-transferred, since PCT<0.5. In contrast, the green and yellow genes are those frequently co-transferred (PCT>0.5) but do not co-occur frequently (PCO<0.5).



Genome group	Number of genomes	Focus phages
Group 1	58	mEp460
Group 2	16	HK022 HK106 HK446 HK542 HK544 HK633 HK97 mEp234 mEp235 mEpX1 mEpX2 HK140
Group 3	7	HK629 HK630 Lambda
Group 4	4	P22
Group 5	2	mEp043c-1 mEp213
Group 6	2	HK225
Group 7	1	HK620
Group 8	1	phi80
Group 9	1	Gifsy
Group 10	1	N15
Group 11	1	mEp237
Group 12	1	mEp390

Fig. 2. Genome groups. (a) Average wGRR over genome pairs. Individual wGRR values are shown in Fig. S2. (b) Number of genomes and focus phages per genome group. (c) Histogram of wGRR between all genome pairs that fall within (green) or between (red) groups.

with high sequence similarity despite low genome-wide similarity of the phage genomes, where they are encoded. Co-transfer is then estimated between protein clusters that are frequently transferred together, i.e. between the same genomes.

Lambdoid phages form distinct groups

To study the genome-wide similarities of lambdoid phages, we analysed 26 focus phages that were shown experimentally to be active (Fig. S1). We supplemented this data set with lambdoid phages retrieved from NCBI and with prophages predicted from *E. coli* genomes. We included (pro-)phages having an ANI of at least 95% to one of the 26 focus phages, resulting in ten temperate phage genomes from NCBI and 264 *E. coli*-derived prophages after deduplication. Next, we grouped these phages, where phages were assigned to the same group if they have a wGRR of at least 60% to one of the focus phages. NCBI phages and prophages with a wGRR <60% to any focus phage were not included in subsequent analyses. We found that five NCBI phages and 64 prophages met the threshold, resulting in a total of 95 phages that were analysed in the present study (Table S1). By merging overlapping groups, we found that the 95 genomes fall into 12 groups, of which six groups are singletons, two groups contain two genomes and four groups contain at least four genomes (Fig. 2, Fig. S2A). Notably, the groups were defined by wGRR to the focus phages; thus, groups can also contain pairs of phages with a wGRR less than 60%. Nevertheless, by looking at all pairwise wGRR values, we found a good distinction between groups, where a high wGRR (i.e. generally >40%) within groups and a low wGRR (i.e. generally <40%) across groups is observed (Fig. 2).

Here, we include publicly available phage genomes by first having an ANI of at least 95% and second an wGRR of at least 60% to one of the focus phages. Notably, ANI is calculated as the average nucleotide identity of the shared regions; thus, phages that share highly similar sequences but also contain non-homologous regions can reach high ANI values. In contrast, wGRR is based on the average protein identity of homologous proteins normalized by the number of proteins in the smaller genome. Thus, pairs of phages can have high ANI but low wGRR, if they share some highly similar genes but do also contain highly different or even non-homologous proteins. The high degree of mosaicism observed in pairwise phage genome comparisons might explain that many phages have high ANI values but low wGRR and were thus not included in the subsequent analysis. Here we used a high wGRR cut-off of 60%, which allowed us to filter out distantly related phages. This is important as it allows us to define clearly separated groups, which ensures that we reliably infer genes that are affected by gene transfer. Nevertheless, additional mosaic phage genomes might exist that could not be reliably grouped with this approach.

Protein clustering shows that there is no core genome of lambdoid phages

The 6210 proteins from all 95 genomes were clustered into 1145 protein clusters, of which 608 are singletons (Fig. S3A). We confirmed that lambdoid phages do not contain any core genes, i.e. no gene is present in all 95 phages. Instead, the largest protein cluster contains 83 proteins from 83 different phages belonging to eight different genome groups. Furthermore, only 31 clusters occur in at least 48 (50%) of the phage genomes and no protein cluster contains phages from more than eight genome groups.

Next, we assigned proteins to operons, where 640 (56%) protein clusters could be assigned to a particular operon. We found that the late operon contains larger protein clusters, whereas the early left operon contains smaller protein clusters (Fig. S3B).

Our analysis confirmed the absence of a core genome of lambdoid phages as previously suggested [26]. Thus, a core genome-based phylogeny and traditional phylogenetic methods to infer HGT are not feasible. Instead, we present an approach to infer HGT based on genome groups.

Gene transfer between distantly related lambdoid phage genomes is frequent and affects all operons to a similar extent

Next, we inferred gene transfer events between phages that belong to different genome groups. Since phages from different genome groups are highly dissimilar (wGRR<40%, Fig. 2c), it is very unlikely that they contain highly similar proteins that have been vertically inherited. Thus, to detect gene transfer events among phages, we searched for pairs of highly similar homologous proteins (proteins in the same protein cluster with an identity of at least 80%) that occur in two different genome groups (Fig. S4A and B). In total, 5238 protein pairs met this condition (Fig. S4B, Table S3). We confirmed that these transferred proteins originate from distantly related genomes characterized by low wGRR values that are generally below 40% (Fig. S4C). We found at least one gene transfer in each genome group (Table 1). Gene transfer is more prevalent in some genome groups compared to others; for example, for groups 3, 5, 7, and 11, more than 50% of the protein clusters are involved in transfers (Table 1). The average proportion of transferred genes per genome also varies between genome groups, ranging between 2% (Group 9) and 77% (Group 7). Note that a very high proportion of genes in a genome can be involved in gene transfer. These gene transfers usually involve multiple partner genomes, consistent with the observations that genome-wide similarities as estimated by wGRR are still low between genomes connected by gene transfer.

The transferred proteins belong to 180 different protein clusters, where transfers with the focus phages are included in 168 (93%) of these protein clusters. The majority of the 180 transferred clusters fall into known operons, while only 15 (8.3%) were assigned to an unknown operon (Fig. 3a, Table S4). We found that 26% of all protein clusters in operons are involved in gene transfer and that gene transfer affects all operons to a similar extent (Chi-square test, χ^2=0.038, df=3, *P*>0.05, Fig. 3a, Table S4).

We reconstructed HGT among lambdoid phages by identifying pairs of highly similar proteins that are encoded on highly dissimilar genomes. A similar approach has been applied to phages before [14]. Here, we extended this approach by focusing on protein clusters instead of protein pairs, which goes beyond previous pairwise approaches and allows us to investigate co-transfer events.

Co-transfer of genes is particularly frequent in the immunity and late operons

Next, we inferred pairs of protein clusters that are frequently co-transferred (PCT>0.5), i.e. they are transferred together between more than 50% of the genome pairs where at least one of them is transferred. This cutoff was motivated by the observation that there was a sudden increase in the distribution at PCT=0.5 (Fig. S5). Many of the pairs with PCT=0.5 are only involved in two co-transfers in total [37 pairs with PCT=0.5, of which 24 (i.e. 68%) with only two co-transfers] and should thus not be considered as frequently co-transferred.

We found 181 frequently co-transferred pairs of protein clusters (Table S5). These frequent co-transfers involve 102 different protein clusters (58% of all the transferred clusters). Clusters involved in frequent co-transfers occurred more frequently in the immunity and late operons (Chi-square test, χ^2=9.1, df=3, *P*=0.028, Fig. 3b, Table S4). The immunity operon is involved in the switch from the lysogenic to the lytic cycle and, in phage Lambda, it contains genes coding for the lysogenic conversion proteins RexA and RexB and the repressor protein cI [53]. We find that most of the co-transferred genes in the immunity operon encode cI (see Example 2 below), which leads to the overrepresentation of co-transferred clusters in this small operon. The late operon is transcribed late in the phage lytic cycle and it encodes the structural proteins that make up the phage virion. These structural proteins need to successfully form complexes in virion assembly, suggesting that only specific combinations of proteins work well together [53, 56], which explains the frequent co-transfer of these genes.

Table 1. Numbers of transferred proteins between genome groups

Left: numbers of unique protein clusters that are transferred between each pair of genome groups (upper triangle) and the proportion of these numbers among all different protein clusters that occur in any genome of the two pairs (lower triangle). Right: total numbers of protein clusters and transferred clusters.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**No. of different protein clusters**	**No. of transferred clusters**	**Percentage**	**Avg. proportion of transferred genes per genome**
1		7	43	2	2	6	4	5	0	5	6	1	563	47	8%	17%
2	0.9%		43	20	42	14	31	6	0	2	10	19	260	106	41%	60%
3	6.3%	11.1%		13	17	6	29	6	0	5	7	0	176	91	52%	64%
4	0.3%	6.2%	5.3%		10	0	14	0	0	0	0	0	91	27	30%	31%
5	0.3%	14.6%	7.4%	6.5%		3	12	11	0	0	3	1	79	51	65%	66%
6	0.9%	4.1%	2.3%	0.0%	1.8%		0	23	0	19	32	2	103	45	44%	53%
7	0.7%	10.8%	14.3%	10.8%	9.6%	0.0%		0	0	0	0	0	60	46	77%	77%
8	0.8%	1.9%	2.7%	0.0%	8.7%	17.6%	0.0%		0	6	18	1	79	34	43%	51%
9	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%		0	1	0	62	1	2%	2%
10	0.8%	0.6%	2.3%	0.0%	0.0%	13.4%	0.0%	5.5%	0.0%		18	4	62	22	35%	35%
11	1.0%	3.2%	3.1%	0.0%	2.2%	24.2%	0.0%	17.0%	0.8%	16.5%		2	70	40	57%	56%
12	0.2%	6.5%	0.0%	0.0%	0.8%	1.3%	0.0%	0.8%	0.0%	3.5%	1.6%		61	23	38%	38%

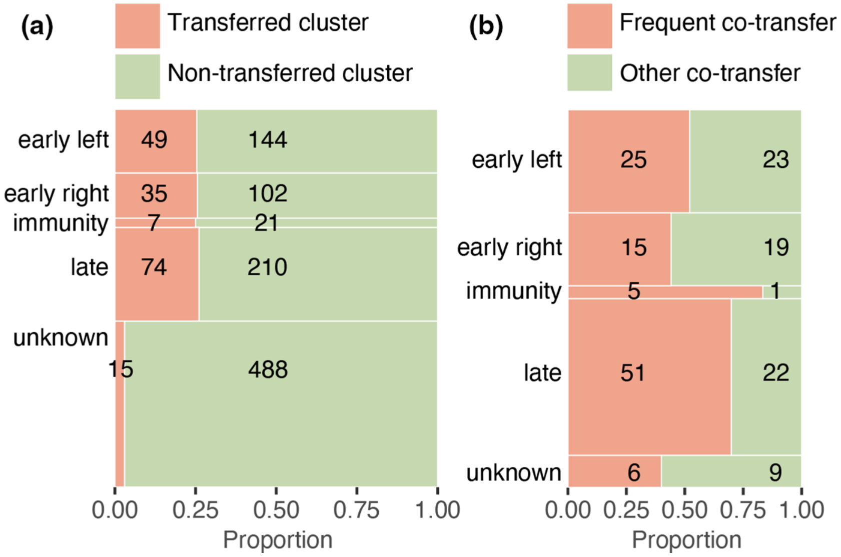


Fig. 3. Gene transfers per operon. (a) Mosaic plot of protein clusters per operon that are either transferred (orange) or not (green). The proportions between the operons are not statistically different (Chi-square test excluding 'unknown', χ^2=0.038, df=3, *P*>0.05). (b) Mosaic plot of the number of protein clusters that are involved in frequent co-transfers (orange; PCT>0.5) or other co-transfers (green; PCT≤0.5) per operon. The proportions are significantly different between the operons (Chi-square test excluding 'unknown', χ^2=9.1, df=3, *P*=0.028). Note that only protein clusters that are involved in at least one co-transfer are included.

Despite the even distribution of HGT across the different operons, we find an uneven distribution of co-transfer events across operons. The higher occurrence of co-transfer in the immunity and late operons suggests that particular genes in these operons are preferentially transferred together with other genes.

Frequently co-transferred genes fall into modules of consecutive genes on the genome

To analyse if co-transferred genes might potentially be transferred together by one horizontal transfer event of a DNA segment with multiple genes, we calculated if they are physically close to each other on the genome. We found that the proportion of frequent co-transfer events decreases with physical distance (Fig. 4). The proportion of co-transfers is particularly high for pairs that are directly next to each other or that have at most five genes between them. The high co-transfer rate for genes up to five genes away from each other on the genomes suggests that long segments with more than two genes can get transferred in one event.

By linking all 181 pairs of co-transferred protein clusters, we reconstructed a network of 28 distinct modules (i.e. connected components; Fig. 5, Table S6). Of these 28 modules, ten contain two protein clusters, seven contain three protein clusters and 11 contain more than three protein clusters. In many cases, the co-transferred genes are adjacent on the genome, i.e. the distance in number of genes between them is zero (Table S6). Thus, these modules of consecutive genes lead to small pairwise distances between all pairs of co-transferred genes (Fig. 4).

Here we find modules of co-transferred genes. Note that the term co-transfer suggests that these genes have been transferred together in one transfer event, where one phage recombinant is formed by taking up both genes either in one stretch of DNA or in multiple recombination events that occurred in the same infection cycle. Nevertheless, our approach could also pick up sequential transfers, where genes have been transferred after each other during subsequent infections if the respective intermediate stages were viable. Our results suggest that most co-transfers involve adjacent genes on the genome, which makes it likely that they have indeed been transferred together on a single stretch of DNA. Nevertheless, we also observe co-transfer between distant genes, which might be the result of multiple recombination events.

Co-transferred genes are physically closer on the genome than co-occurring genes

In bacteria, the co-occurrence of genes across genomes has been used as a signal to detect if genes are functionally associated [57]. To compare our measure of co-transfer to co-occurrence, we compared the frequently co-transferred clusters (PCT>0.5) to the frequently co-occurring clusters (PCO>0.5). We observed that co-transfer results in a lower number of pairs compared to co-occurrence (Fig. S6). Furthermore, pairs that are frequently co-transferred have a lower physical distance compared to the frequently co-occurring pairs (Fig. S6).

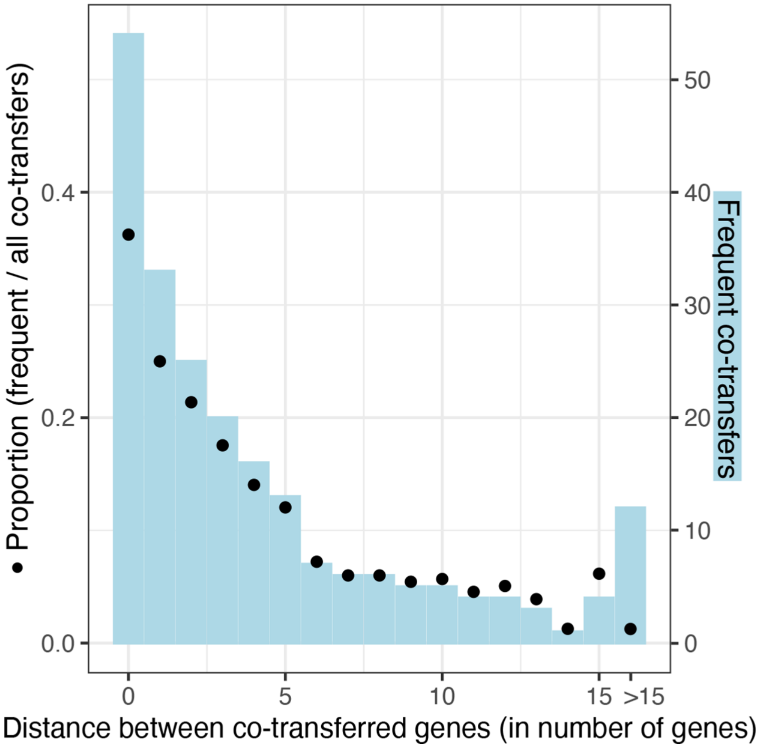


Fig. 4. Distance (in number of genes) between co-transferred pairs. The number of frequent co-transfers (PCT>0.5) per genetic distance is plotted in light blue (right axis) and the proportion of frequent co-transfers among all the co-transfers per genetic distance is shown by dots (left axis).

When comparing co-transfer to co-occurrence, we find that different pairs of proteins are detected with both approaches. With the co-transfer method, fewer genes were found, and these also had a lower physical distance. We thus conclude that the co-transfer method is more specific whereas the co-occurrence method might include more spurious associations. To estimate co-occurrence, we here use a simplistic approach that is based on the number of genomes containing two protein families. Further methods have been developed that need a phylogeny to estimate co-occurrence (e.g. [57]); however, such methods are not applicable to this data set as no core genome exists and thus no core phylogeny can be inferred. The co-transfer approach that we present here circumvents the issue of a missing core phylogeny by detecting co-transferred genes. These gene transfers have been detected as highly similar proteins present on distantly related phage genomes. Instead of estimating a fully resolved core phylogeny, we inferred distinct genome groups and detected gene transfers as highly similar proteins that occur in genomes of different groups. Thus, our co-transfer approach can be applied to genomes where no core genes and thus no core phylogeny exists, which makes it particularly suitable for viruses.

Frequently co-transferred genes are functionally related

Lastly, we analysed the functions of co-transferred genes by zooming into the functions of specific modules. We found that the large modules only involve genes that belong to the late operon (modules 1–5 in Fig. 5). Some modules from the late operon include proteins involved in similar functions, suggesting co-transfers of functionally related genes. For instance, module 3 contains only phage tail proteins whereas module 4 contains only phage head proteins. This finding is consistent with the previous observation that head and tail genes belong to separate modules within the genomes of lambdoid phages [52]. In the following, we describe three additional examples of co-transfers of functionally related genes.

Example 1: the mEp phages have been selected to contain a variety of immunity groups [38]. Thus, our data set contains seven gene clusters of the repressor *cI* (in the immunity operon) and eight gene clusters of the transcriptional regulator *cro* (in the early right operon). Among these, we detected three independent cases of frequent co-transfers that involve cI and Cro (modules 13, 20 and 21, Fig. 5, Table S7A). There are some similarities between the proteins in the different clusters. The different Cro proteins that are involved in co-transfer share no sequence similarity, whereas the cI proteins have low pairwise protein identity and were thus clustered separately (Table S7B). Although the co-transfer of this neighbouring gene pair can be explained by one transfer event affecting the whole region, it is remarkable that at least three independent co-transfers affected this well-known functional

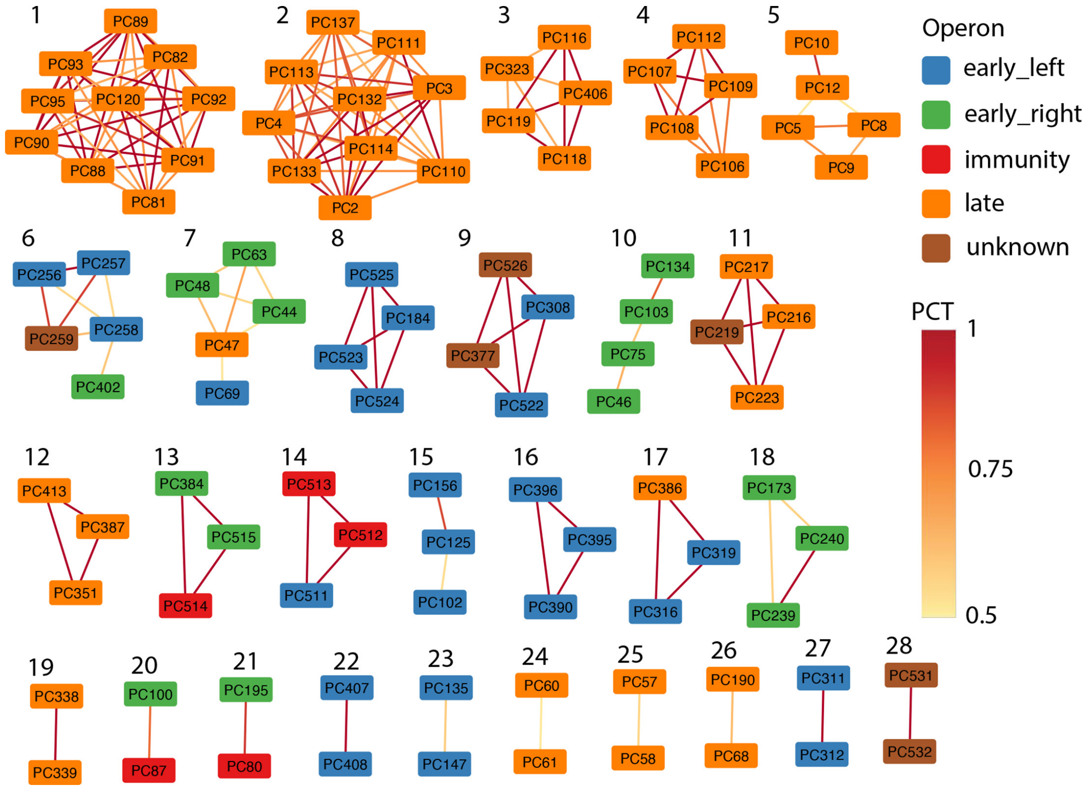


Fig. 5. Network of frequently co-transferred clusters, colour-coded by operon. The annotation of the protein clusters is given in Table S6. PCT (protein cluster co-transfer) denotes the proportion of how often the genes are transferred together.

unit, whereas other neighbouring genes are not transferred as often. This suggests that different cI and Cro proteins depend on each other.

Example 2: we identified three independent cases of frequent co-transfer that involve the early left protein Kil, an FtsZ inhibitor known to function in host killing (PC184, module 8; PC102, module 15; PC319, module 17, Fig. 5) [58]. There is a low pairwise protein identity between PC102 and PC184, whereas the other Kil proteins do not show sequence similarity (Table S7). Interestingly, *kil* is always co-transferred with genes known to encode proteins that are involved in DNA binding (PC527) or in recombination (PC316 – *recT* and PC125 – *erf*). Kil is not essential but becomes critical during conditions of high recombination frequency [58], suggesting that the observed co-transfers are functionally important.

Example 3: although most modules involve physically close genes, we detected one frequent co-transfer between genes that are distant on the genome (module 5): PC12 encodes a holin protein and is located at the beginning of the late operon whereas PC10 encodes the Lambda outer membrane protein Lom and is located at the end of the late operon. Remarkably, both genes interact with the host membrane, where holin is involved in lysis of the inner membrane at the end of the lytic cycle and Lom is integrated in the outer membrane. Lom has so far only been described as a lysogenic conversion protein that belongs to a family of virulence proteins (Pfam accession PF06316) and is involved in adhesion to the eukaryotic host cell [59]. Nevertheless, Lom is also strongly expressed during the lytic cycle [60]. The co-transfer of *lom* and the holin-encoding gene suggests that Lom might also interact with holin and be involved in bacterial lysis during the lytic cycle.

CONCLUSIONS

Here we show that co-transfer of functionally related genes is frequent during the evolution of lambdoid phages. From this we conclude that the co-transferred genes probably need to function together, suggesting epistatic interactions among gene presences in phage genomes. The importance of epistasis in phages is debated, where recombination has been shown to result in fewer epistatic interactions [61]. Nevertheless, the functionally associated frequent co-transfers described here suggest that epistasis is abundant among the frequently recombining temperate phages. We also observe that the co-transferred genes are physically close on the genome. This is expected for recombining organisms, where simulations showed that sexually reproducing organisms evolve modular genomes, where related functions tend to be physically close [62–64]. Our results confirm this scenario for lambdoid phages. Functionally related genes evolved to occur physically close on the genomes, and gene transfers of functionally related genes maintain epistatic interactions despite frequent gene transfer. Thus, abundant gene transfer leads to the evolution of a highly modular genome architecture in lambdoid phages, which is, for example, also evident by the physical proximity of essential genes and of non-essential genes [53].

Many temperate phage groups with high genome mosaicism have been identified for various hosts, such as for mycobacteria, *Gordonia* or *Pseudomonas* [22, 23, 65]. The evolutionary framework described here for lambdoid phages can probably be extended to other temperate phages, where the particular functional groups that are involved in co-transfer might be group-specific. In conclusion, the interplay of epistasis and gene transfer can explain genome mosaicism among temperate phage genomes.

Funding information
This project received funding from the Swiss National Science Foundation (grant no. PZ00P3_179743) given to C.C.W. and the Competence Centre Environment and Sustainability of the ETH (grant no. 0-21248-08) given to D.R.

Acknowledgements
We would like to thank Mumun Gençoglu and Gabriela Klecak for help in the laboratory and Rodney King and Ing-Nang Wang for kindly sharing their phage isolates. We thank Franz Baumdicker and Ing-Nang Wang for valuable discussions on the manuscript.

Author contribution
Conceptualization – A.K., C.C.W.; Sequencing – D.R.; Methodology – A.K.; Formal Analysis – A.K., Z.B.; Data Curation – A.K., C.C.W.; Writing – Original Draft Preparation – A.K., C.C.W.; Writing – Review and Editing – A.K., C.C.W., Z.B., D.R.

Conflicts of interest
The author(s) declare that there are no conflicts of interest

References

1. Harrison E, Brockhurst MA. Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. *Bioessays* 2017;39:12.
2. Wendling CC, Refardt D, Hall AR. Fitness benefits to bacteria of carrying prophages and prophage-encoded antibiotic-resistance genes peak in different environments. *Evolution* 2021;75:515–528.
3. Touchon M, Bernheim A, Rocha EP. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J* 2016;10:2744–2754.
4. Kim MS, Bae JW. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J* 2018;12:1127–1141.
5. Casjens S. Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* 2003;49:277–300.
6. López-Leal G, Camelo-Valera LC, Hurtado-Ramírez JM, Verleyen J, Castillo-Ramírez S, *et al*. Mining of thousands of prokaryotic genomes reveals high abundance of prophages with a strictly narrow host range. *mSystems* 2022;7:e0032622.
7. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, *et al*. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* 2019;177:1109–1123.
8. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, *et al*. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* 2021;6:960–970.
9. Hatfull GF. Bacteriophage genomics. *Curr Opin Microbiol* 2008;11:447–453.
10. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, *et al*. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* 2014;513:242–245.
11. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, *et al*. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 2016;17:930.
12. Kupczok A, Neve H, Huang KD, Hoeppner MP, Heller KJ, *et al*. Rates of mutation and recombination in siphoviridae phage genome evolution over three decades. *Mol Biol Evol* 2018;35:1147–1159.
13. Kupczok A, Dagan T. Rates of molecular evolution in a marine *Synechococcus* phage lineage. *Viruses* 2019;11:E720.
14. Moura de Sousa JA, Pfeifer E, Touchon M, Rocha EPC. Causes and consequences of bacteriophage diversification via genetic exchanges across lifestyles and bacterial taxa. *Mol Biol Evol* 2021;38:2497–2512.
15. De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, *et al*. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. *PLOS Genet* 2014;10:e1004181.
16. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol* 2017;2:17112.
17. Martinsohn JT, Radman M, Petit MA. The lambda red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. *PLoS Genet* 2008;4:e1000065.
18. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev* 2014;38:865–891.
19. Susskind MM, Botstein D. Molecular genetics of bacteriophage P22. *Microbiol Rev* 1978;42:385–413.
20. Botstein D. A theory of modular evolution for bacteriophages. *Ann N Y Acad Sci* 1980;354:484–491.
21. Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, *et al*. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* 2015;4:e06416.
22. Pope WH, Mavrich TN, Garlena RA, Guerrero-Bustamante CA, Jacobs-Sera D, *et al*. Bacteriophages of gordonia spp. Display a spectrum of diversity and genetic relationships. *mBio* 2017;8(4):17.
23. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C. Origins of highly mosaic mycobacteriophage genomes. *Cell* 2003;113:171–182.
24. Campbell A, Botstein D. Evolution of the lambdoid phages. In: *Lambda II. Cold Spring Harbor Laboratory Press*. 1983. pp. 365–380.
25. Campbell A. Phage evolution and speciation. In: Calendar R (eds). *The Bacteriophages [Internet]*. Boston, MA: Springer US; 1988. pp. 1–14.
26. Campbell A. Comparative molecular biology of lambdoid phages. *Annu Rev Microbiol* 1994;48:193–222.
27. Monod C, Repoila F, Kutateladze M, Tétart F, Krisch HM. The genome of the pseudo T-even bacteriophages, a diverse group that resembles T4. *J Mol Biol* 1997;267:237–249.
28. Ford ME, Sarkis GJ, Belanger AE, Hendrix RW, Hatfull GF. Genome structure of mycobacteriophage D29: implications for phage evolution. *J Mol Biol* 1998;279:143–164.

29. Lucchini S, Desiere F, Brüssow H. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. *J Virol* 1999;73:8647–8656.

30. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, *et al*. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* 2000;299:27–51.

31. Westmoreland BC, Szybalski W, Ris H. Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. *Science* 1969;163:1343–1348.

32. Campbell A, Schneider SJ, Song B. Lambdoid phages as elements of bacterial genomes (integrase/phage21/*Escherichia coli* K-12/icd gene). *Genetica* 1992;86:259–267.

33. Casjens SR, Thuman-Commike PA. Evolution of mosaically related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly. *Virology* 2011;411:393–415.

34. Parma DH, Snyder M, Sobolevski S, Nawroz M, Brody E, *et al*. The Rex system of bacteriophage lambda: tolerance and altruistic cell death. *Genes Dev* 1992;6:497–510.

35. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:421.

36. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–277.

37. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.

38. Kameyama L, Fernández L, Calderón J, Ortiz-Rojas A, Patterson TA. Characterization of wild lambdoid bacteriophages: detection of a wide distribution of phage immunity groups and identification of a nus-dependent, nonlambdoid phage group. *Virology* 1999;263:100–111.

39. Uc-Mass A, Loeza EJ, de la Garza M, Guarneros G, Hernández-Sánchez J, *et al*. An orthologue of the cor gene is involved in the exclusion of temperate lambdoid phages. Evidence that Cor inactivates FhuA receptor functions. *Virology* 2004;329:425–433.

40. Hernández-Sánchez J, Bautista-Santos A, Fernández L, Bermúdez-Cruz RM, Uc-Mass A, *et al*. Analysis of some phenotypic traits of feces-borne temperate lambdoid bacteriophages from different immunity groups: a high incidence of cor+, FhuA-dependent phages. *Arch Virol* 2008;153:1271–1280.

41. Dhillon EK, Dhillon TS, Lam YY, Tsang AH. Temperate coliphages: classification and correlation with habitats. *Appl Environ Microbiol* 1980;39:1046–1053.

42. Arber W, Enquist L, Hohn B, Murray N, Murray K. Experimental methods for use with Lambda. In: Hendrix RW, Stahl FW and Weisberg RA (eds). *Lambda II*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1983. pp. 433–466.

43. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 2008;9:75.

44. Menouni R, Champ S, Espinosa L, Boudvillain M, Ansaldi M. Transcription termination controls prophage maintenance in *Escherichia coli* genomes. *Proc Natl Acad Sci* 2013;110:14414–14419.

45. Mardanov AV, Ravin NV. The antirepressor needed for induction of linear plasmid-prophage N15 belongs to the SOS regulon. *J Bacteriol* 2007;189:6333–6338.

46. Garcia-Russell N, Elrod B, Dominguez K. Stress-induced prophage DNA replication in *Salmonella enterica* serovar *Typhimurium*. *Infect Genet Evol* 2009;9:889–895.

47. Refardt D, Rainey PB. Tuning a genetic switch: experimental evolution and natural variation of prophage induction. *Evolution* 2010;64:1086–1097.

48. Refardt D. Within-host competition determines reproductive success of temperate bacteriophages. *ISME J* 2011;5:1451–1460.

49. Sayers E. The E-utilities In-depth: parameters, syntax and more. In: *In: Entrez Programming Utilities Help [Internet] [Internet]*. Bethesda (MD): National Center for Biotechnology Information (US), 2009. https://www.ncbi.nlm.nih.gov/books/NBK25499/

50. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–21.

51. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 2002;30:1575–1584.

52. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. The origins and ongoing evolution of viruses. *Trends Microbiol* 2000;8:504–508.

53. Casjens SR, Hendrix RW. Bacteriophage lambda: early pioneer and still relevant. *Virology* 2015;479–480:310–330.

54. Wickham H. ggplot2. In: *Ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer-Verlag, 2016.

55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13:2498–2504.

56. Rajagopala SV, Casjens S, Uetz P. The protein interaction map of bacteriophage lambda. *BMC Microbiol* 2011;11:213.

57. Whelan FJ, Rusilowicz M, McInerney JO. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb Genom* 2020;6.

58. Haeusser DP, Hoashi M, Weaver A, Brown N, Pan J, *et al*. The Kil peptide of bacteriophage λ blocks *Escherichia coli* cytokinesis via ZipA-dependent inhibition of FtsZ assembly. *PLoS Genet* 2014;10:e1004217.

59. Vica Pacheco S, García González O, Paniagua Contreras GL. The lom gene of bacteriophage lambda is involved in *Escherichia coli* K12 adhesion to human buccal epithelial cells. *FEMS Microbiol Lett* 1997;156:129–132.

60. Liu X, Jiang H, Gu Z, Roberts JW. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc Natl Acad Sci* 2013;110:11928–11933.

61. Malmberg RL. The evolution of epistasis and the advantage of recombination in populations of bacteriophage T4. *Genetics* 1977;86:607–621.

62. Misevic D, Ofria C, Lenski RE. Sexual reproduction reshapes the genetic architecture of digital organisms. *Proc Biol Sci* 2006;273:457–464.

63. Yang YF, Cao W, Wu S, Qian W. Genetic interaction network as an important determinant of gene order in genome evolution. *Mol Biol Evol* 2017;34:3254–3266.

64. Singhal S, Gomez SM, Burch CL. Recombination drives the evolution of mutational robustness. *Curr Opin Syst Biol* 2019;13:142–149.

65. Johnson G, Banerjee S, Putonti C. Diversity of *Pseudomonas aeruginosa* temperate phages. *mSphere* 2022;7:e0101521.