Genome diversification via genetic exchanges between temperate and virulent bacteriophages

Jorge A. Moura de Sousa[1,*], Eugen Pfeifer[1], Marie Touchon[1] & Eduardo P.C. Rocha[1,*]

[1]Microbial Evolutionary Genomics, Institut Pasteur, CNRS, UMR3525, Paris 75015, France

* corresponding authors: jorge-andre.sousa@pasteur.fr, erocha@pasteur.fr

ABSTRACT

Bacteriophages (henceforth phages) evolve by mutation and recombination. Temperate phages are known to evolve rapidly by genetic exchanges. In contrast, recombination events between virulent and between temperate and virulent phages have rarely been reported. A gene flow barrier between these two large distinct groups of phages could affect the ability of phages to acquire novel functions. Here, we show that genomes of temperate and virulent phages are very distinct but often have a few almost identical genes. These cases are due to recent genetic exchanges of both phage-like and bacterial-like genes. These exchanges were probably mediated by phage recombinases with low homology requirements and mechanisms of non-homologous end joining, since both were over-represented in recombinant phages and their hosts. When assessing the impact of gene flow across the two populations of phages we realized that temperate phages have narrower host ranges than virulent phages. This suggests, and we find some evidence, that gene flow between temperate phages infecting distant bacterial hosts might be mediated by recombination with the broader host virulent phages. Hence, gene flow across phages with distinct lifestyles could drastically increase the gene repertoire available for phage evolution, including the transfer of functional innovations across taxa. These results also have implications for bacterial evolution because of the impact of phage predation in bacterial population dynamics and the contribution of temperate phages to the evolution of bacterial gene repertoires.

SIGNIFICANCE STATEMENT

Many microbial communities are intensely predated by temperate and virulent bacteriophages. The former also contribute to bacterial gene repertoires. Gene transfers between bacteriophages favor their rapid adaptation, but are thought to occur rarely between virulent and temperate bacteriophages because their genomes are very different in terms of gene repertoires and genetic organization. We found that genetic exchanges occur frequently

between these types of bacteriophages, involve different phage and bacterial functions, and are likely due to multiple DNA repair processes. Since we show that virulent bacteriophages have broader host ranges, they can shuttle genes between temperate bacteriophages present in taxa that are beyond the typical range of the latter. This enhances phages diversification and has multiple impacts on bacterial evolution.

Introduction

Bacterial viruses (bacteriophages or phages) are ubiquitous across environments, where they are often the most abundant biological entities and shape bacterial population dynamics. Most known phages have double stranded DNA genomes and can be either virulent or temperate (1). Virulent phages are restricted to lytic infections, where rapid viral replication ends in progeny release and bacterial death. Temperate phages can follow either a lytic or a lysogenic cycle. In the latter, viral DNA maintains itself in the host as a prophage. Most prophage genes are silent until a signal activates their lytic cycle. Yet, some genes are expressed and they may provide adaptive phenotypes to their hosts (2, 3). Half of the bacterial genomes have at least one prophage and some have up to 20 prophages (4), showing the important contribution of temperate phages to bacterial gene repertoires. The relative frequency of temperate and virulent phages vary across bacterial taxa and environments. In the mammalian gut, for instance, temperate phages seem to be more prevalent (5), while most intracellular and many environmental bacteria lack prophages (6), and are thus presumably rarely infected by temperate phages. Phages also drive horizontal gene transfer among bacteria by transduction (3), which may disseminate virulence factors (7) and antibiotic resistance (8).

Phage genomes vary considerable in size, from a few dozen to hundreds of genes (9). The largest genomes tend to be found in virulent phages and seem to have gene repertoire dynamics like those of bacteria, where a core genome is complemented by very diverse sets of accessory genes (10, 11). There is very little information on the mechanisms of gene accretion in virulent phages, even if some families of virulent phages have remarkable diversity in their genomes' size, suggesting the existence of mechanisms of gene exchange (12). The genomic plasticity of temperate lambdoid phages has been much more extensively studied (9, 13, 14). Their analyses reveal that pairs of phages tend to have patches of regions of very similar sequences within genomes that are often very dissimilar. This genome mosaicism is facilitated by the modular organization of phage genomes, where groups of

72 functionally related genes are encoded, and potentially exchanged, together (15). It is also
73 facilitated by the key role of recombination in the production of phage genome concatemers
74 to be packaged into the virion (16). The molecular mechanisms of recombination underlying
75 the mosaicism are not completely understood. They may involve phage-encoded
76 recombinases, which are more permissive to differences between sequences than bacterial
77 RecA-mediated recombination (17). They may also involve homology-free mechanisms such
78 as non-homologous end joining or some type of illegitimate recombination (18, 19). In any
79 case, past studies revealed pairs of (typically temperate) phages with a few almost identical
80 genes within extremely divergent genomes.
81
82 The evolutionary dynamics of phage genomes reflect their distinct lifestyles. A recent study
83 showed that temperate and virulent phages have different "evolutionary modes" representing
84 the relative importance of genetic exchanges in their evolution (20). Indeed, there are many
85 reports of extensive mosaicism in temperate phages. These exchanges can occur between
86 phages during co-infections, between prophages in the genome, or between prophages and
87 infecting temperate phages (14, 17). Given the aforementioned prevalence of prophages in
88 bacteria, and that prophages tend to reside in bacterial genomes for many generations, this
89 may explain the very high rates of recombination of temperate phages. Virulent phages tend
90 to have more conserved core genomes and only a few highly variable regions (21, 22). They
91 might not exchange DNA outside restricted phage taxonomic groups (23, 24). Recombination
92 between virulent phages requires co-infection by two phages, which is probably rare.
93 Nevertheless, some cases have been studied in detail. Evolution of virulent phages of
94 *Lactococcus lactis* from a dairy production line showed that recombination brings more
95 polymorphism than point mutations to their genomes (25), and studies on T4 and T7 phages
96 revealed events of recombination and gene accretion (26, 27). Much less is known regarding
97 genetic exchanges between temperate and virulent phages, even if they are assumed to be
98 rare. For instance, a recent study found no evidence of recent exchanges between a set of 84
99 virulent and temperate phages of *Eschericha coli* (17), and a broader analysis suggested that

100 temperate phages rarely have genes in common with virulent ones (20). However, there are
101 some reports of virulent phages acquiring genes from temperate phages or their prophages
102 (28, 29). To the best of our knowledge, no study has systematically traced genetic exchanges
103 between temperate and virulent phages, even if their existence could have important
104 consequences. From a fundamental point of view, gene exchanges between the two types of
105 phages vastly increase the gene repertoire available for the evolution of each type of phage.
106 This could be particularly important for temperate phages infecting bacterial clades where they
107 are rare and for virulent phages if they are segregated in distinct families with little gene flux
108 between them (as previously proposed (23, 24)). Furthermore, if virulent and temperate
109 phages have different host ranges, gene exchanges can expand the taxonomic range of gene
110 flux in the most restricted types of phages (via recombination with phages with broader host
111 ranges). These exchanges could also affect the adaptation of bacteria, because phage
112 predation shapes microbial communities and temperate phages contributes to bacterial gene
113 repertoires. These exchanges could also have consequences for the safety of phage therapy.
114
115 In this work, we searched for the existence, prevalence and mechanisms underlying genetic
116 exchanges between temperate and virulent phages. Genetic exchanges between cellular
117 organisms are usually detected from variations in the density of polymorphism in genomes or
118 the phylogenetic congruence of core genes (30). This is impossible when analyzing divergent
119 phages because they have no core genes that could be used to build reliable phylogenies and
120 many genes lack homologs in other phages. Furthermore, the observed extensive mosaicism
121 of temperate phages implies that most genes have different phylogenies. Therefore, we
122 searched instead for strong mosaicism, i.e. highly similar genes within highly dissimilar
123 genomes. We were able to identify recent events of recombination between virulent and
124 temperate phages, although our conservative approach will miss very ancient recombination
125 events, or those between similar genomes. These findings impact our understanding of how
126 virulent and temperate phages evolve, especially because our results suggest that virulent
127 phages can facilitate the transfer of adaptive traits across more distant taxa.

128 # Results

129 ## The network of similarities between temperate and virulent phages
130

131 We analysed a dataset of 2487 complete bacteriophage (phage) sequences available from
132 NCBI (as of May of 2019). We used PHACTS (31) to predict their lifestyle and identified 1166
133 virulent and 1336 temperate phages. This method classifies some phages with high
134 confidence and others with less confidence (see Methods). The analysis presented in the main
135 text of this article focuses on this complete dataset, but restricting the study to the ca. 54% of
136 phages classed by PHACTS with high confidence provides similar qualitative results
137 (Supplementary Data). We searched for protein similarity across all phages in the dataset with
138 at least 15 predicted proteins (2387 phages, ~96% of the dataset), identified reciprocal best
139 hits between pairs of phages and computed the corresponding fraction of homologs weighted
140 by sequence identity (weighted Genome Repertoire Relatedness, wGRR, see Methods). This
141 resulted in 2,847,691 pairwise wGRR values, among which ca. 91% were null because there
142 was no detectable homology using our procedure. The remaining 9% of non-null pairwise
143 wGRR values involve almost all phages included in the analysis (2386 phages, >99%). The
144 histograms of non-null wGRR values (Fig 1A) revealed higher values for the comparisons
145 between pairs of phages with similar lifestyles than for the comparisons involving virulent and
146 temperate phages. Very few (6%) of the comparisons between phages with different lifestyles
147 show some level of homology, and even these tend to have lower wGRR than the ones
148 observed for pairs of phages of similar lifestyles (i.e., temperate-temperate or virulent-virulent).
149 The analysis of the phages classed with high confidence in terms of lifestyle showed an even
150 greater difference in the distributions between pairs of phages of similar or opposite lifestyle
151 (Fig S1A). This may indicate that the few high wGRR values between temperate and virulent
152 phages correspond to erroneous assignments of lifestyles, or that confident lifestyle
153 assignment becomes difficult when there are homologous genes across temperate and
154 virulent phages. In any case, this analysis suggests the existence of relations of homology

155 between many temperate and virulent phages, even if these tend to result in small wGRR
156 values.
157
158 Network based representations are frequently used to describe evolutionary relationships
159 between phages (1, 24). We used the gene repertoire relatedness to build a graph where
160 phages are nodes and edges represent wGRR. This results in a network that is highly
161 connected, given that 99% of the phages have some level of homology with at least one other
162 phage (Fig S2). We used the Louvain method for community detection to cluster phages by
163 their wGRR, which resulted in 34 different clusters with at least 3 phages, 5 with 2 phages,
164 and 3 singletons (Fig S3). In order to compare the effects of the phage lifestyle, family (the
165 characterized virion morphology, see Methods) and host phyla in the separation of phages in
166 different clusters of wGRR, we used Shannon information entropy to quantify how
167 homogeneous each cluster is, regarding each of these three traits (Fig 1B, Fig S1B). As a
168 control, we randomly re-assigned all the labels (host phyla, family and lifestyle) to the phage
169 set. Although randomized clusters are more heterogeneous than any of the three traits being
170 tested, there was a clear and significant difference in terms of cluster homogeneity between
171 the lifestyle and the other two traits. Clusters are very homogeneous in terms of host phyla
172 and phage family, as revealed by average entropies close to zero. In contrast, they are much
173 less homogeneous in terms of lifestyle, where almost all clusters included both temperate and
174 virulent phages. These results show that there is genetic similarity between many temperate
175 and virulent phages independently of phage or bacterial taxonomic considerations.
176

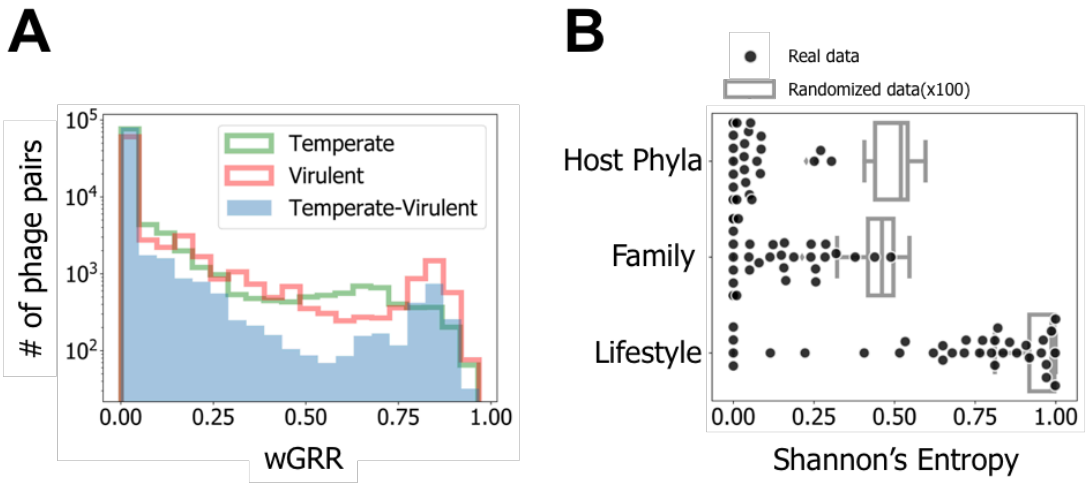


Fig 1. Results of the analysis of similarity between phages. A) Histograms of the wGRR values (with wGRR>0). B) Shannon's Entropy values for each cluster identified with the Louvain community detection method in the wGRR matrix. Results are given for the three phage traits (N=34 for each trait, one per cluster). Boxplots represent the distribution of the concatenation of 100 repetitions of randomized cluster label re-assignments (N=3400 for each trait). All distributions are significantly less heterogeneous than their random counterparts (all $p<0.0001$, 2-sample Kolmogorov Smirnov test). The observed clusters are significantly more heterogeneous in phage lifestyle than in the other variables (both $p<0.001$, Tukey Honest Significant Difference (HSD) Test).

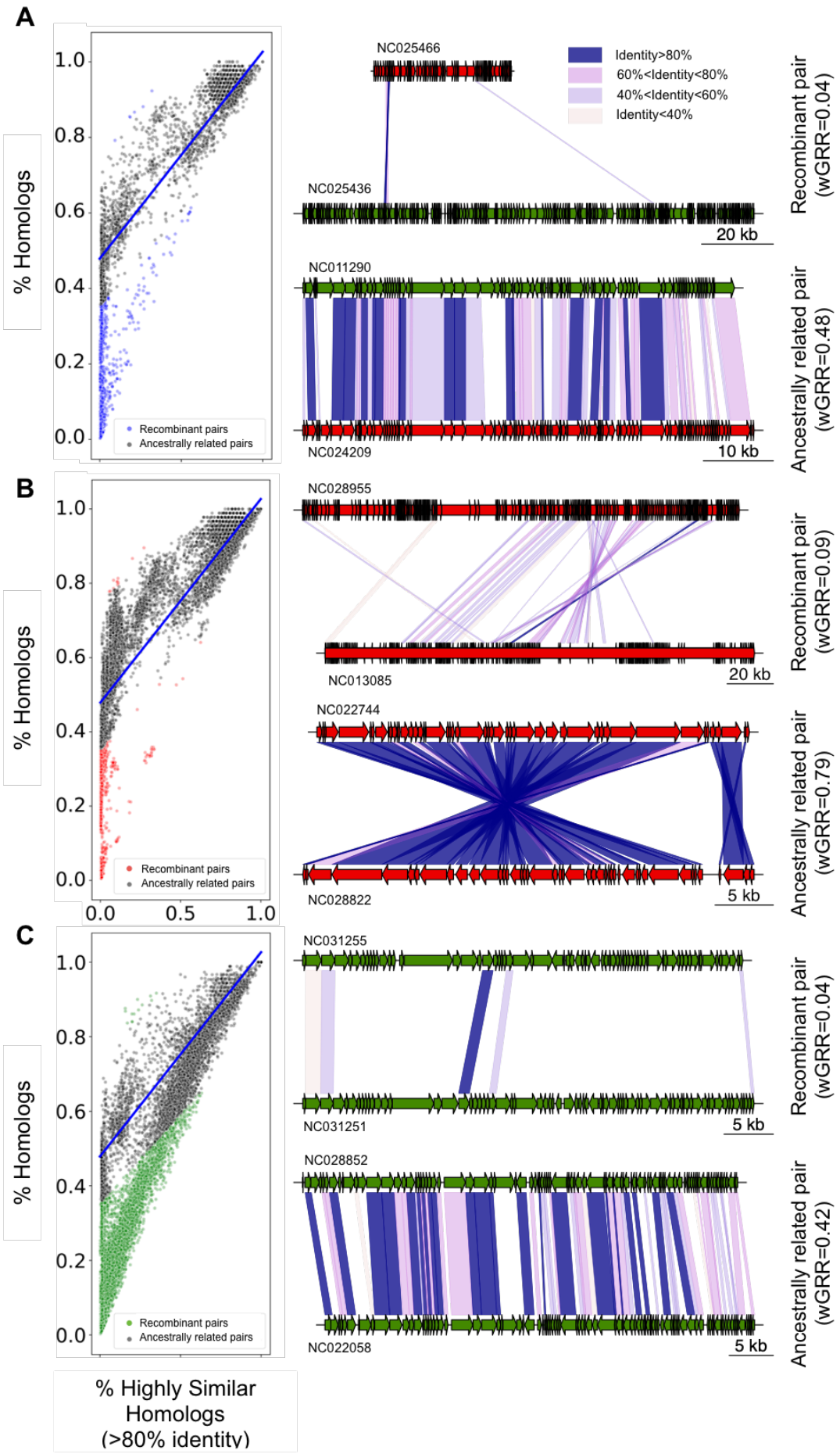
Recombination drives genetic relatedness between temperate and virulent phage

The observed similarities between temperate and virulent phages could be explained by common ancestry as well as by recent gene flux. When the wGRR is small, the two processes translate into different patterns: common ancestry results in many homologs of low similarity along the genomes, whereas recent genetic exchanges result in a small number of highly similar homologs within very dissimilar genomes. To distinguish between these two scenarios, one can use the contrast between the fraction of homologous genes (above a low minimal similarity threshold of 35% identity) and the fraction of genes that are very similar (more than 80% identity). Distant common ancestry results in a sizeable fraction of the pairs of genomes with homologs but no single very similar gene, whereas recent gene flow between distantly related phages results in a few highly similar homologs and a low wGRR. This procedure does not allow to identify very ancient events of recombination that result in a few homologs of low

200 sequence similarity. It also misses events of recombination between very closely related
201 genomes or when a large fraction of the genome is exchanged. Since we showed above that
202 the most confident virulent and temperate genomes have low wGRR values, these two cases
203 should be rare regarding comparisons between temperate and virulent phages (Fig 1A).
204 Genomes that engage rarely in recombination, as is thought to be the case between virulent
205 and temperate phages, should show a rapid decrease of the fraction of very similar genes in
206 function of the fraction of detectable homologs. Indeed, comparisons between temperate and
207 virulent (Fig 2A) and between virulent phages (Fig 2B) show a few nearly identical genes
208 (>80% identity) within genomes otherwise lacking homologues. These cases are best
209 explained by recent genetic exchanges. In contrast, when genomes exchange genes very
210 frequently, as is the case of many temperate phages, the comparisons show mosaics of highly
211 similar and highly dissimilar (eventually non-homologous) genes. This results in a linear
212 distribution where almost identical genes co-exist in very different genomes (Fig 2C).
213
214 The comparisons between temperate and virulent phages suggest a frequency of
215 recombination that is lower between the two populations of phages than between temperate
216 phages. This is in agreement with the network-based analyses above, where relations of
217 homology between temperate and virulent phages were frequent but limited to a small number
218 of genes (low wGRR). We then built a linear model based on the relationship between the
219 fraction of very similar proteins (>80% identity) and the fraction of homologous proteins (see
220 Methods) between temperate and virulent genomes (Fig 2A, blue line). We limited this dataset
221 to pairs of phages with at least 50% of homologous proteins, where the influence of outliers
222 that are associated with recombination is expected to be weaker. The linear model fitted to
223 this dataset represents a conservative model of the null hypothesis that the relationship
224 between these two parameters is mostly due to ancestry. It fits well the major group of
225 comparisons between virulent and temperate phages across almost all the range of the
226 regression (Fig 2A). We computed the negative residuals of the linear model to identify
227 significant negative deviations to the main trend. These represent cases where genomes have

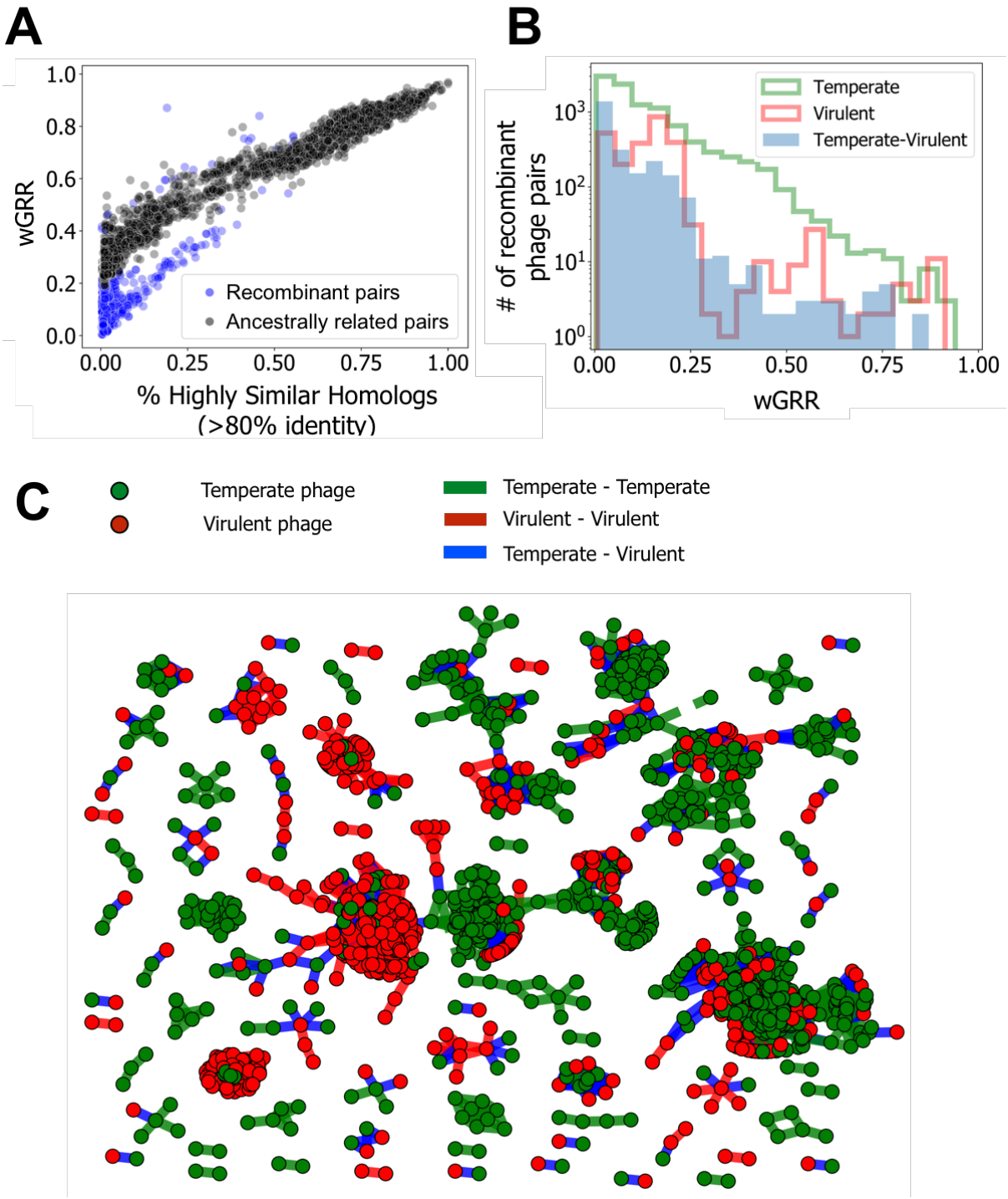
228 an unexpectedly high number of very similar genes given the overall level of homology. This
229 threshold on the value of negative residuals was used across all datasets to identify putative
230 recent events of recombination (Fig 2, see Methods). For the case of exchanges between
231 temperate and virulent or between virulent phages, the overwhelming majority of exchanges
232 were found in very dissimilar phages (wGRR below 0.2, Fig 3A and Fig 3B). However, the
233 comparisons implicating pairs of temperate phages have a notably different distribution of
234 genetic similarity (Fig 3B), with a large fraction showing intermediate (>0.25) wGRR values.
235 This is again suggestive of a higher gene flux in temperate phages, in agreement with other
236 works showing their high mosaicism (1, 9).
237
238 The analysis of sequence similarity between phages connected in the wGRR network shows
239 two distinct patterns: genomes that are more likely to be ancestrally related tend to have large
240 regions of low sequence similarity, whereas the others show homology limited to one or a few
241 very similar genes (Fig 2, Fig S4). This suggests that our method discriminates recent
242 recombination from shared ancestry (see examples of pairs of phages more likely to be
243 ancestrally related in Fig 2 and Fig S4). 48% (528) of the virulent phages and 66% (852) of
244 temperate phages in our dataset were classified as recombinant. Considering only
245 recombination between phages of different lifestyle, 36% of the virulent phages were inferred
246 to recombine with temperate phages, whilst 37% of temperate phages were found to
247 recombine with their virulent counterparts.
248
249 The network restricted to pairs of recombinant phages (Fig 3C) shows once again clusters
250 that are less homogeneous in terms of lifestyle than in terms of host phyla or phage family
251 (Fig S5A). This means that there is more recombination inferred between phages of different
252 lifestyle than between phages of different families or between those that infect different
253 bacterial clades. Furthermore, it suggests that the coincidence of virulent and temperate
254 phages in the same clusters of the graph is partly driven by recombination. We found
255 qualitatively similar conclusions when restricting the analysis to phages with confidently

256 assigned lifestyles, albeit in this case there was a non-significant difference in heterogeneity
257 between phage lifestyle and the phage family (Fig S5B). Overall, these results show that
258 recombination between temperate and virulent phages is frequent in our dataset.


A
B
C
% Homologs
% Highly Similar Homologs (>80% identity)
Recombinant pairs
Ancestrally related pairs
NC025466
NC025436
Identity>80%
60%<Identity<80%
40%<Identity<60%
Identity<40%
20 kb
NC011290
NC024209
10 kb
NC028955
NC013085
20 kb
NC022744
NC028822
5 kb
NC031255
NC031251
5 kb
NC028852
NC022058
5 kb
Recombinant pair (wGRR=0.04)
Ancestrally related pair (wGRR=0.48)
Recombinant pair (wGRR=0.09)
Ancestrally related pair (wGRR=0.79)
Recombinant pair (wGRR=0.04)
Ancestrally related pair (wGRR=0.42)

259

Fig 2. Identification of putative recombinant phage pairs. A-C) Scatterplots of pairs of phage genomes with recombinant proteins indicated as colored points (otherwise grey). The linear regression model (blue lines) was inferred for the temperate-virulent dataset in A and applied to all datasets. The genomic maps in front of each panel show representative examples of recombinant (top) or non-recombinant (ancestrally related) phage pairs (bottom). Color codes for points and genes: virulent (red), temperate (green), virulent-temperate (blue). Colors in the blocks linking the phages indicate sequence similarity between homologs.



268 **Fig 3. Identification of putative recombinant phage pairs in the wGRR network.** A) Scatterplot of pairs of temperate-virulent
269 phages in terms of wGRR and the fraction of high sequence-identity homologous genes. B) Histogram of wGRR values for the
270 subset of recombinant phage pairs. C) Network of recombinant phages. Each node represents a phage genome and each edge
271 a relationship of homology (with wGRR below 0.5).

272 Functional classification of genes recombined between temperate
273 and virulent phage
274

275 We investigated the genes exchanged between phages with different lifestyles to understand
276 if recombination affects specific functions. These genes were the ones encoding highly similar
277 proteins (at least 80% identity) present in dissimilar phage genomes (wGRR lower than 0.5).
278 We found an average of 6 recombinant genes per phage genome. These genes were only
279 slightly smaller than the other genes (191 vs 207 amino acids, Fig S6). To inquire on their role,
280 we clustered the Prokaryotic Virus Orthologous Groups (pVOGs) profiles database into
281 functional classes and used it to annotate phage-related functions (see Methods). Some
282 functions are over (e.g., proteins involved in packaging, injection and assembly) or under (e.g.,
283 structural proteins) represented in the proteins exchanged between temperate and virulent
284 phages. Yet, these genes matched a very large number of different protein profiles (>100),
285 showing that genes identified as recombinant are not restricted to a few functional categories
286 of slow evolving proteins that might spuriously be identified as recombinant. Hence,
287 recombination between phages with different lifestyles can result in the exchange of a large
288 range of functions (Table 1, Table S1, File S2).
289

290 Temperate phages sometimes carry genes that were acquired from bacterial genomes. We
291 thus searched if recombinant genes were associated with some specific bacterial functions.
292 This would suggest that recombination between temperate and virulent phages disseminates
293 bacterial traits. We removed from the bacteria Non-supervised Orthologous Groups
294 (bactNOG) profiles database the functions associated with phages by removing profiles
295 matching pVOG profiles with a VQ higher than 80%. We then used the remaining bactNOG
296 profiles to annotate the recombinant genes (see Methods). The vast majority of these genes

297 have no defined function. Some of the remaining might be typical phages genes (like putative
298 structural components) that were missed by the filter using the pVOG profiles. Yet, other genes
299 matched profiles whose functions are related to carbohydrate or nucleotide transport and
300 metabolism, DNA related functions, or energy production and conservation (Table 2, Table
301 S2, File S2). The category of energy production corresponds to the majority of genes of known
302 function. They are found in phages infecting marine bacteria and correspond to proteins
303 implicated in photosynthetic activity. Genes with these types of functions were previously
304 found to be beneficial in virulent phages that infect cyanobacteria (32). Other notable functions
305 include methylases, transporters and single-strand binding proteins. Hence, our results
306 suggest that a multiplicity of phage and bacterial functions can be exchanged between
307 temperate and virulent phages by recombination.
308

309 **Table 1. Bacteriophage-related functions of recombinant proteins resulting from genetic exchanges between temperate**
310 **and virulent phages.** pVOG profiles matching the phage protein dataset (first column). Number of recombinant proteins
311 (between temperate and virulent phages, second column). Enrichment (+) or depletion (-) of the functional category in
312 recombinant proteins relative to all phage proteins (Fisher-exact test, significant differences for each functional category shown
313 as asterisks (*) and adjusted for multiple comparisons using the Benjamin-Hochberg method).

pVOG Category	Matched profiles	Matched recombinant proteins in temperate-virulent phages	P-value (sign)
Unknown	201	821	**** (+)
Structure	26	96	**** (-)
DNA metabolism, Regulation & Recombination	25	73	**** (-)
Packaging, Injection & Assembly	21	115	* (+)
Others	3	5	**** (-)
Lysis	5	10	* (-)
Structure & Lysis	6	21	*** (+)
Other combinations	5	14	n.s.

314

315

316

317 **Table 2. Bacterial-related functions of recombinant proteins resulting from genetic exchanges between temperate and**
318 **virulent phages.** Profiles matching the phage protein dataset and number of phage recombinant proteins matched by them.

bactNOG Category	Matched profiles	Matched recombinant proteins in temperate-virulent phages
Function unknown	87	212
Replication, recombination and repair	7	27
Transcription	6	19
Energy production and conversion	3	68
Carbohydrate transport and metabolism	1	2
Inorganic ion transport and metabolism	1	6
Cell wall/membrane/envelope biogenesis	1	1
Nucleotide transport and metabolism	1	11
Transcription + Signal transduction mechanisms	1	1

319

320

321 Recombinant phages are enriched for recombinases

322

323 Recombination between distantly related phages cannot usually be achieved by the cellular
324 recombinases, because these tolerate few mismatches (33). However, some phage
325 recombinases, e.g. lambda Red, allow recombination between more dissimilar sequences
326 (17). To test if these recombinases could facilitate recombination between temperate and
327 virulent phages, we searched for the Erf, Sak, Sak4, RecT, RecA and Gp2.5 families in the

328 genomes of phages (see Methods). We found that ca. 42% of the phages with recombinant
329 genes encode at least one of these recombinases, while only 26% of the remaining phages
330 encode them (Tables S3-4). This suggests that these recombinases may play an important
331 role in the recombination between distant phages.
332
333 There is no known mechanism that allows for genetic exchanges between virions. Instead,
334 the DNA of different phages can only recombine within bacterial hosts. It is therefore possible
335 that phage-like recombinases encoded in the genomes of bacteria play a role in the genetic
336 exchanges between phages. We searched for these recombinases in the bacterial hosts (see
337 Methods). For this, we removed RecA recombinases from the analysis, since they are
338 encoded by the vast majority of bacterial genomes (34). We found that hosts of phages with
339 recombinant genes are 29% more likely to encode recombinases in their genomes than the
340 hosts of the remaining phages. These host recombinases are concentrated (ca. 90% of the
341 total) in prophages and absent from the rest of the bacterial genomes (Tables S5-6). This
342 suggests that recombinases from prophages facilitate recombination between other phages
343 within bacterial cells. In agreement with this view, the genomes of bacterial species that host
344 phages with recombinant proteins tend to have more prophages (Fig S7-8).
345
346 Since it has been proposed that recombination between phages could also result from non-
347 homologous recombination (18), we searched the genomes for the gene encoding the Ku
348 protein. Ku is a key marker of the non-homologous end-joining (NHEJ) pathway (35). This
349 analysis showed that genomes of bacterial hosts of phages with recombinant proteins are
350 significantly more likely to encode NHEJ than the others (Fig S9, see Methods). In the Mu
351 phage, a Ku homolog has been shown to promote NHEJ using a host ligase (36). Interestingly,
352 we found that 10 phages in our dataset (6 temperate and 4 virulent) also code for the Ku
353 protein. All of these phages have recombinant proteins. Even if their number is small, the over-
354 representation of phages carrying Ku within those with recombinant genes is significant

355 (p=0.003, Fisher exact test). Either encoded in phages or bacteria, such a mechanism could
356 facilitate genetic exchanges between phages lacking homologs.
357

358 Broader host range of virulent phages facilitates gene flux across
359 clades of temperate phages
360

361 Defining the host range of phages is a notoriously difficult problem. However, if recombinant
362 genes are observed between phages that have distinct bacterial hosts, it is reasonable to
363 assume that either their host ranges overlap or that their ancestors must have at some point
364 infected the same bacteria. Because the recombined genes are very similar, this
365 recombination event must not have been too ancient (although translation of sequence
366 divergence into coalescence time is impossible with this data). Consequently, recombination
367 events can contribute to the characterisation of the phages' host range. To assess this
368 hypothesis, we computed both the patristic distance (based on the 16S rDNA, Fig 4A) and the
369 differences in 3-mer genomic signatures between the bacterial hosts (identified at the species
370 level, see Methods) of all pairs of phages with recombinant genes (Fig 4B, Fig S10-11). In all
371 datasets, most cases of recombination occurred within very closely related bacteria. However,
372 all three datasets (recombination between temperate, between virulent and between
373 temperate and virulent) significantly differ in their host range (as measured by this approach).
374 Importantly, pairs of virulent phages with recombinant genes are more frequently associated
375 with distant (or dissimilar) hosts, compared to those of temperate phages. This suggests that
376 virulent phages are able to infect (and undergo recombination in) a wider range of hosts
377 species than temperate phages. Interestingly, recombination between temperate and virulent
378 phages is associated with intermediate distances between bacterial hosts, which likely reflects
379 the contribution of broader host (virulent) and narrower host (temperate) phages.
380

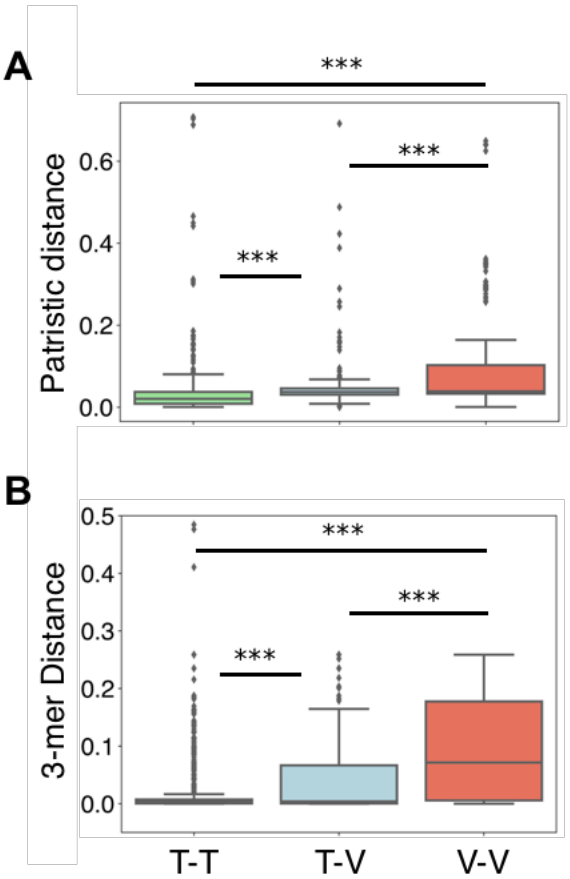


Fig 4. Recombination between phages indicates a broader host range of virulent phages. A) Boxplots of the non-null patristic distances (computed from the 16S rDNA gene tree) between the hosts' species of each recombinant phage pair. **B)** Distributions of the differences in 3-mer genomic signatures between the hosts's species of each phage pair with recombinant genes. *** p=0.001, Tukey HSD for all pairs. T-T: temperate-temperate, T-V: temperate-virulent, V-V: virulent-virulent.

If virulent phages have broader host ranges and recombine with temperate phages in phylogenetically distant bacteria, they can shuttle genes between groups of temperate phages that are sexually isolated (because they infect distantly related bacterial clades). To systematize these observations, we built a simplified network with only recombinant temperate-virulent phage pairs where each phage infects a different host genus (Fig 5A). This network reveals several exchanges between virulent and temperate phages from distinct genuses. In some cases, these exchanges concern distinct genes (Fig 5B), whereas in others they concern the exact same gene (Fig 5C). In these examples, temperate phages share little or no homology beyond genes homologous to the virulent phage (Fig S12), suggesting that

these exchanges were mediated by the virulent phage. These results suggest that recombination between temperate and virulent phages, in tandem with the broader host range of virulent phages, increases the frequency of genomic exchanges between temperate phages in phylogenetically distant hosts.

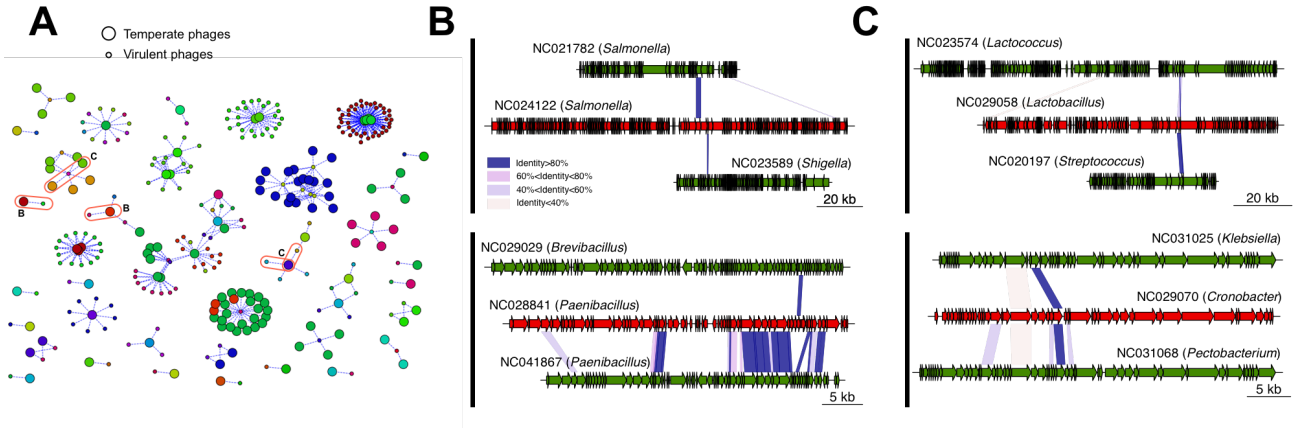


Fig 5. Genetic exchanges between virulent and temperate phages with different host genus. A) Simplified network restricted to pairs of temperate-virulent phages with distinct bacterial host genuses (i.e., edges shown are only those that link phages with different hosts' genuses). Each node color represents a different bacterial host genus, and the node sizes identify either temperate (large nodes) or virulent (small nodes) phages. Examples from panels B and C are highlighted in ellipses. B) Examples of virulent phages (red) having recombined with different recombinant proteins exchanged with temperate phages infecting (green) of distinct bacterial genuses (green). C) Analogous to B, but representing genes that were exchanged with the virulent phage sharing the exact same protein with both temperate phages infecting distinct bacterial genuses**.** Colors in the blocks linking the phages indicate sequence similarity between homologs. The genomic maps between each of the two temperate phages for each of the cases are shown in Fig S12.

413 ## Discussion
414

415 Our study provides a first systematic quantification of genetic exchanges between temperate
416 and virulent phages, highlighting their potential to drive phage diversification. We had to
417 develop a specific method to identify these exchanges, since the lack of species phylogenies
418 and core genes in phages precludes the use of standard approaches. A recent
419 characterization of the evolution of phage genomes suggested that there are very few
420 similarities between temperate and virulent phages (20). Interestingly, the phages in these few
421 cases showed a pattern of rapid gene acquisition and loss. Moreover, this study found that
422 these genomes have very few homologous genes which is consistent with the exchange of a
423 small number of genes between temperate and virulent phages. Our method focuses on these
424 types of cases, by analysing the mosaicism in individual (pairwise) genomes and by
425 contrasting low global genome similarity, based on wGRR, with high individual protein
426 similarity. However, our approach does have some caveats. First, incorrect classification of
427 lifestyle (virulent or temperate) leads to spurious inference of inter-lifestyle recombination
428 events. We controlled for this uncertainty by redoing the key analyses using the subset of
429 phages with a lifestyle assigned with high confidence, which revealed similar qualitative
430 patterns, often with a stronger statistical signal (see Supplementary Data). Second, the
431 method is incapable of identifying certain types of recombination events: the ancient ones
432 (proteins that are now less than 80% identical), those covering a significant fraction of the
433 phage genome (raising the value of wGRR due to a high global level of similarity, as in (37)),
434 and those between closely related phages (where wGRR is high). Even if this last limitation is
435 unlikely to be important, since virulent and temperate phages are typically quite distinct, this
436 means that we have probably underestimated the number of gene exchanges between
437 phages with different lifestyles. Third, the distinction between recombination and ancestry is
438 particularly challenging in the analysis of temperate phages because of their pervasive
439 mosaicism. This leads to less confident identification of specific recombination events between

440 temperate phages. Nevertheless, and even with these caveats, a large fraction of the pairs of
441 phages with recombinant genes show clear traces of recent genetic exchanges.
442
443 When does recombination between temperate and virulent phages occur? We envisage three
444 major opportunities for such events. First, temperate and virulent phages may recombine
445 when co-infecting a bacterial host or when one of them is already present in the cell in a state
446 of pseudo-lysogeny. We do not know the frequency of such co-infections, but the existence of
447 recombination between virulent phages suggests that co-infections occur in nature (except if
448 all these occur indirectly through temperate phages). Second, virulent phages may recombine
449 with prophages present in the infected bacteria. Given the prevalence of bacteria with active
450 or defective prophages, such events are probably much more likely than those of co-infection.
451 Accordingly, the few previous descriptions of recombination between virulent phages and
452 temperate phages involved recombination with prophages (28, 38, 39). Virulent phages could
453 acquire genes from prophages by restricting chromosome DNA, as found in T2 and T4 phages
454 decades ago (40, 41). Such exchanges could also occur if prophages are induced by the
455 infection of virulent phage. Finally, prophages could acquire genes from virulent phages when
456 the genomes of the latter are cleaved by the hosts' defence systems, such as restriction
457 modification systems or CRISPR-Cas systems, which produce recombinogenic linear double
458 stranded DNA (42). A quantification of the relevance of each of these situations requires more
459 genome sequences, and experimental data, than the one available at this stage. Indeed, our
460 approach does not allow to identify the direction of the gene exchanges (from temperate to
461 virulent or vice versa), and this is key to know the relevance of the above scenarios.
462
463 Which mechanisms mediate gene flux between temperate and virulent phages? Given the
464 differences between these types of phages, genetic exchanges between them are likely to
465 occur through mechanisms with less stringent homology requirements than bacterial RecA
466 mediated homologous recombination. Two schools of thought have proposed that
467 recombination between very divergent temperate lambdoid phages could result from either

468 relaxed homologous recombination or illegitimate recombination (9, 17). These mechanisms
469 are not mutually exclusive, and both were proposed to act on the observed recombination
470 between virulent phages and prophages in *Lactococcus* (43). We observe that phages with
471 recombinant genes tend to encode recombinases known to tolerate more mismatches than
472 bacterial RecA-mediated recombination. Similarly, many bacterial genomes encode such
473 recombinases in their prophages. These recombinases might facilitate homologous
474 recombination between divergent phages. It is possible that alternative mechanisms not
475 involving homologous recombination also favour gene flux across temperate and virulent
476 phages. Accordingly, we observed that phages with recombinant genes are more likely to be
477 hosted by bacteria encoding Ku, the key protein of the NHEJ pathway. This mechanism results
478 in the ligation of two linear double stranded DNA molecules even if they may lack sequence
479 similarity (44). Recent works have shown that NHEJ-like activity could occur even in the
480 absence of Ku, *e.g.* it was found to recombine non-homologous DNA in *E. coli* (45) and the
481 T4 ligase performs NHEJ-like functions with high efficiency in the absence of Ku (46, 47). The
482 poor understanding of these Ku-independent processes precludes assessing their role in
483 phage recombination, but they could contribute to explain gene flux between unrelated
484 phages. In summary, the presence of prophages in many bacterial genomes provides ample
485 opportunity for recombination with infecting virulent phages, and these genetic exchanges
486 could occur by a multitude of molecular processes.
487
488 Recombination between temperate and virulent phages allows transfer of functional
489 innovations from one group to the other. This has the potential to facilitate phage adaptation
490 to novel challenges. When investigating the functions of exchanged genes, we found proteins
491 involved in the production of virions, but also bacteriocins or RNA polymerases. Some
492 exchanged proteins are recombinases and their transfer could disseminate the potential for
493 further genetic exchanges. Moreover, a recent work found evidence that anti-CRISPR
494 proteins, used by phage to evade their hosts' CRISPR defences, can be exchanged between
495 temperate and virulent phages (48).

496

497 Our findings suggest that the impact of recombination in phage functional and morphological
498 diversification can be enhanced by the differences in host range between temperate and
499 virulent phages. We observed that virulent phages have broader host ranges, which suggests
500 that temperate phages could be more constrained in their host range by the additional genetic
501 interactions they establish with their hosts when deciding between lysis and lysogeny and
502 during the lysogenic cycle. These processes require some integration of the phage genome
503 on the cell's genetic network, e.g. the development of genetic regulation for induction under
504 the control of the host SOS response. This integration adds constraints to the genome of
505 temperate phages and may contribute to their narrower host ranges. There is some published
506 evidence on differences in host ranges between temperate and virulent phages. For example,
507 virulent coliphages isolated from the faeces of toddlers infect a broader range of gut strains
508 than temperate ones (49). The implications of our findings are that groups of temperate
509 phages that might be sexually isolated (because they infect phylogenetically distant hosts)
510 can exchange genes indirectly through recombination with broader host range virulent
511 phages. Conversely, it was observed that virulent-to-virulent genetic exchanges are rare
512 outside restricted taxonomic groups (23, 24). The genetic flux between virulent phages, and
513 also between virulent and temperate phages, might then increase when they infect hosts
514 whose genomes have many prophages.

515

516 The expression of prophage genes can change bacterial phenotypes. This so-called lysogenic
517 conversion has been shown to have an important impact on pathogenicity, growth and stress
518 responses (50, 51). We analysed the functions of genes exchanged between temperate and
519 virulent phages that are not typically found in phages and tend to be associated with bacteria.
520 Although the permanence of genes with bacterial functions might be, for most cases, transient
521 in virulent genomes, there are examples of such genes that have been acquired and
522 repurposed by those phages. For instance, bacterial photosynthetic genes encoded by
523 genomes of virulent phages were shown to increase viral progeny (32, 52). These genes are

524 often referred to as phage auxiliary metabolic genes (53). Amongst other genes with bacterial
525 functions, they were identified as being exchanged between temperate and virulent phages in
526 our data, which opens the possibility that they were acquired by the latter through this
527 mechanism. Interestingly, photosynthetic genes were previously shown to be exchanged
528 between *Prochlorococcus* and *Synechococcus* through viral intermediates (54), suggesting
529 that gene flow with and between phages plays an active role in the evolution of cyanobacterial
530 photosynthesis. Thus, in addition to the acquisition of genes from bacterial genomes (55) and
531 other prophages (56), exchanges between virulent and temperate phages could enhance the
532 genetic repertoire of prophages with functions adaptive to the bacterial host.
533
534 Given the pervasiveness of prophages in bacterial genomes, recombination between
535 temperate and virulent phages might be more frequent than expected based on their low
536 global sequence similarity, and perhaps even more than what we report here. This could have
537 implications for the use of phage therapy, because it suggests that there may be exchanges
538 between temperate and virulent phages during therapy and the former may favour transfer to
539 and between bacterial genomes. The assessment of safety in therapies should account for
540 this factor. At this stage, the rates of exchange between the two populations of phages seem
541 low to dramatically change the flux of genetic information between bacteria at the very short
542 time scale of an infection. Furthermore, bacteria themselves can exchange DNA through many
543 other means (transduction, conjugation, transformation, etc). In any case, the frequency of
544 recombination events between temperate and virulent phages, coupled with the ability of
545 virulent phages to shuttle genes between temperate phages that are segregated in distinct
546 clades, can contribute to the high rates of diversification of phage genomes, as well as to the
547 expansion of gene repertoires of lysogenic bacteria.

Methods

Data

We retrieved the complete genomes of 13513 bacteria and 2502 phages from NCBI non-redundant RefSeq database (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/, last accessed in May 2019). The lifestyle of these phages was predicted using PHACTS v.0.3 (31). Predictions were considered as confident if the average probability score of the predicted lifestyle was at least two standard deviations (SD) away from the average probability score of the other lifestyle, which leads a precision rate in lifestyle identification of 99% (31). Using these criteria, we classified as confident 54% of the phages into 571 virulent and 780 temperate phages.

We retrieved information (when available) on the phage hosts from the Genbank files of the phages or from the Virus-Host DB (57) (https://www.genome.jp/virushostdb/). A few phages (~2%) have no identified host, since they were collected from environmental (soil, sea, etc.) or faeces samples. In this dataset, the phage hosts belong to 332 species and 145 genera. Most of these species (69%) have at least one complete genome sequenced and available in NCBI RefSeq. In the absence of a complete genome of the same species (31%, indicated as "manually annotated" in File S1), we used another genome from the same genus as a proxy for the species. Some analysis are repeated by excluding these phages with manually annotated (non-confident) host species, in order to provide a conservative control for these assignments.

Prophages (integrated temperate phages) were predicted using VirSorter v.1.0.3 (58) with the RefSeqABVir database in these bacterial genomes (corresponding to the host species/genus of the phages). The least confident predictions, i.e., categories 3 and 6, which may be prophage remnants or erroneous assignments, were excluded from the analyses. We also retrieved the phage family from the Genbank files of each phage. Most of them are Caudovirales (94%) and belong to the 3 main phage families, i.e., Siphoviridae (1291), Myoviridae (632) and Pododviridae (383). The complete dataset of phage genomes can be found in File S1.

577

578 **Protein similarity and weighted gene repertoire between bacteriophage genomes**

579 We searched for sequence similarity between all proteins of all phages using mmseqs2 (59)
580 (Nature Biotechnology release, August 2017) with the sensitivity parameter set at 7.5. The
581 results were converted to the blast format for analysis and we kept for analysis the hits
582 respecting the following thresholds: e-value lower than 0.0001, at least 35% identity, and a
583 coverage of at least 50% of the proteins. The hits were used to compute the bi-directional best
584 hits between pair of phages, which were used to compute a score of gene repertoire
585 relatedness weighted by sequence identity:

586 $$wGRR = \frac{\sum_i^p id(A_i, B_i)}{\min(A, B)}$$

587 where A_i and B_i is the pair *i* of homologous proteins present in *A* and *B*, id(A_i,B_i) is the percent
588 sequence identity of their alignment, and min(*A*,*B*) is the number of proteins of the genome
589 encoding the fewest proteins (*A* or *B*). wGRR is the fraction of bi-directional best hits between
590 two genomes weighted by the sequence identity of the homologs. It varies between zero (no
591 bi-directional best hits) and one (all genes of the smallest genome have an identical homolog
592 in the largest genome). wGRR integrates information on the frequency of homologs and
593 sequence identity. Hence, when the smallest genome is 100 proteins long, a wGRR of 0.03
594 can result from three pairs of identical homologs or six pairs of homologs with 50% identity.

595

596 **Similarity networks, community detection and calculation of entropy**

597 The phage network was built based on the filtered wGRR values, using the *networkx* and
598 *graphviz* Phyton (v2.7) packages, and the *neato* algorithm. The Louvain community clusters
599 (60) were calculated using the *best_partition* function from the *community* package in Python
600 (v2.7). For each cluster (considering only those with at least 3 nodes), we first calculated their
601 total number of nodes, and then the number of nodes corresponding to each category: nodes
602 in either one of two lifestyles (temperate or virulent), nodes with a given phage family and
603 nodes with a given described bacterial host phyla. The Shannon entropy of a cluster (X) with

604 nodes classed according to a given variable that takes t different values ($y_{1,X}$… $y_{t,X}$) was
605 calculated as:

606 $$Entropy(X) = -\sum_{i\in\{1,..,t\}} f(y_{i,X}) * \log_t f(y_{i,X})$$

607 Where $f(y_{i,X})$ is the relative frequency of the nodes classed *i* in the cluster. As an example,
608 consider a cluster composed of 20 nodes, where 10 of them have a temperate lifestyle and
609 10 have a virulent lifestyle. Because there are two labels (temperate or virulent), $t = 2$. If the
610 frequencies of the two types of phages were identical, close to the expectation under a random
611 distribution, the two frequencies would be 0.5 and the entropy would amount to 1, which is the
612 maximum and corresponds to a highly heterogeneous cluster. When all phages are from one
613 single type, the entropy is equal to 0, corresponding to an homogeneous cluster. To be able
614 to compare each trait to the expectation of random cluster compositions, we randomly re-
615 assigned the labels (host phyla, phage family and lifestyle) to the phages and computed the
616 Shannon entropy as described above. The results (shown as boxplots in the figures)
617 summarise the distribution of the 100 random experiments.
618

619 **Identification of pairs of phage genomes with recombinant genes**

620 In order to separate events of recombination and from homology due to shared ancestry, we
621 used the relationship between the fraction of homologous proteins (over the average number
622 of proteins for each phage pair) and the fraction of homologous proteins with very high identity
623 (at least 80% identity, also over the average number of proteins for each phage pair). We used
624 the *ols* function from the *statsmodels* package in Python (v2.7) to generate a linear model
625 describing the expectation of ancestry between phage genomes. To characterize common
626 ancestry as robustly as possible, we restricted the dataset to temperate-virulent pairs of
627 phages, as this dataset is expected to contain the lowest frequency of recombination events.
628 Further, and in order to reduce the influence of the outliers associated with recombination (and
629 thus maximize the fit of the linear model to cases of common ancestry), we analysed only the
630 pairs of phage with at least 50% of the proteins in their (average) genomes classified as

631 homologous (meaning at least 35% identity and 50% coverage) and a minimum of one highly
632 similar homologous protein (at least 80% identity). Using this dataset, we applied residual
633 analysis (with the *fit* and *outlier_test* function from *statsmodels*) to identify the first significant
634 residual value. This was defined as the least distant residual from the linear model that is
635 classed as outside the confidence interval, with an adjusted pvalue<0.05 using the
636 Benjamin/Hochberg method – *fdr_bh*. The value of this residual is assumed as the significance
637 threshold beyond which a residual is classified as a putative recombination event (since it
638 significantly departs from the expectation of common ancestry defined by the linear model).
639 Thus, the linear model and the residual threshold are applied to the entire range of the data
640 across the three datasets (temperate-virulent, temperate-temperate and virulent-virulent
641 pairs). The pairs of phages whose residuals are larger than the minimum significant residual
642 threshold (and whose wGRR are below 0.5 to avoid high similarity caused by recent
643 divergence of the lineages) are classed as pairs of putatively recombined phages. The
644 recombinant genes are identified as the proteins with at least 80% identity between the pairs
645 of phages showing evidence of recombination.
646
647 **Functional annotation of proteins**
648 We used HMMER v3.1b2 (61) (default parameters) to search for genes matching the
649 prokaryotic Virus Orthologous Groups (pVOGS) (Version May 2016 (62)) database of hmm
650 profiles (filters used were e-value<1e-5 and profile coverage>60%). Only pVOGs with a viral
651 quotient (VQ) above 0.8 were used (7751 out of 9518 pVOG profiles in total). The pVOG
652 profiles were classed into seven functional categories: a) structure, b) lysis, c) packaging,
653 maturation/Assembly and DNA injection, d) DNA metabolism, recombination and regulation,
654 e) others and f) unknown by two approaches. First a profile-profile comparison was done using
655 the HHsuite 2.0.9 (63) with phage-specific profiles from the PFAM (64) and TIGRFAM (65)
656 database (taken from (66)). Applying a threshold of p-value $<10^{-5}$ resulted in 711 profiles that
657 cluster in 261 groups using the Louvain algorithm. The annotations of the PFAM and

658 TIGFRAM profiles were used to assign one of the functional categories to a group. The
659 remaining profiles were manually assigned considering the piled-up annotations of the pVOGs
660 (File S3). The identification of bacterial functions was performed based the EggNOG database
661 of hmm profiles (bactNOG Version (67)). In order to minimize the number of bactNOG entries
662 that derive from prophages in bacterial genomes, we used hhsearch (HHSuite version 3.2
663 (68)) to remove from bactNOG the profiles matching pVOG profiles. The bactNOG profiles
664 with matches in pVOGs with VQ above 0.8 (p-value <0.0001) were discarded. This resulted
665 in a reduction of 34% of the bactNOG profiles' dataset. The remaining 135814 bactNOG
666 profiles were used to class the dataset of putative recombinant genes (filters used were e-
667 value<1e-5 and profile coverage>60%) in broad functional categories.
668

669 **Detection of phage-like recombinases in phage and bacterial genomes**

670 The families of recombinases of phage were described in (69), for which we built profiles or
671 recovered them from PFAM. To build the profiles, we retrieved the homologs given in the
672 reference above and aligned them using default options with MAFFT (v7.407) (70). The
673 alignments were used to build the profiles using hmmbuild from HMMER (default options).
674 Our novel profiles are given in supplementary material. We used HMMER to search for
675 homologs of Sak (PF04098, --cut_ga), Sak4 (Sak4 from phage T7, minimum score = 20), Erf
676 (PF04404, --cut_ga), RecT (PF03837, --cut_ga) and gp2.5 (gp2.5 from phage T7, minimum
677 score = 20). An additional profile, RecA (PF00154, --cut_ga) was searched only in phage
678 genomes. Hosts were retrieved from the GenBank file of each phage, and all the genomes in
679 the database belonging to that host's species were used for the calculation. E.g., if the host
680 species of a given phage is described as *E. coli*, we calculated the mean number of
681 recombinases in all *E. coli* genomes. The values shown on the table are the mean of these
682 means for phages with recombinant genes and the remaining phages. If recombinases were
683 found within the coordinates of a prophage, they were associated with prophage regions. Note
684 that a particular phage host can be included multiple times in the distribution, if subsequent

685 phages have a similar host, and can even result in the host being included in the dataset of
686 recombinant and non-recombinant phages simultaneously. Although this results in repeated
687 sampling, it represents the likelihood that hosts of recombinant or non-recombinant phages
688 encode for recombinases. Moreover, this is unbiased for either class of phages, since the
689 hosts of both recombinant and non-recombinant phages are subject to this repeated sampling
690 process.
691

692 **Detection of NHEJ in bacterial genomes**

693 We used HMMER to search for homologs of the Ku protein (PF02735.16) in the dataset of
694 bacterial genomes, using the -cut_ga parameter. We did not require the presence of a
695 neighboring ligase to consider the system complete, because such a ligase is often absent
696 (71). The set of bacterial genomes assigned as host of a given phage was asserted as
697 described in the section above. The fraction of hosts that encode for NHEJ was calculated as
698 the number of genomes that where at least one homolog of Ku was found, divided by the total
699 number of genomes considered for a given phage. The use of the presence/absence if Ku as
700 the key variable, and not the number of Ku genes, is due to three reasons: 1) the numbers
701 were always low (most often 0 or 1), 2) some genomes encoding multiple NHEJ systems
702 express them during different times, 3) some bacteria seem to encode heterodimeric Ku where
703 two genes are necessary for the process (72).
704

705 **Calculation of the patristic distance between bacterial hosts**

706 We used the 16S rRNA of the bacterial genomes identified in the RefSeq annotations,
707 corresponding to the species level of the identified phage hosts. We selected the first entry of
708 each genome and aligned them using the secondary structure models with the program
709 SSU_Align version 0.1.1 (73). Poorly aligned positions were eliminated with the ssu-mask.
710 The alignment was trimmed with trimAl version 1.2 (74) using the option –noallgaps to delete
711 only the gap proteins but not the regions that are poorly conserved. The 16S rRNA

712 phylogenetic tree was inferred using maximum likelihood with IQTREE version 1.6.5 (75)
713 under the best-fit model automatically selected by ModelFinder (76), and with the options –
714 wbtl (to conserve all optimal trees and their branch lengths) and –bb 1000 to run the ultrafast
715 bootstrap option with 1000 replicates. The patristic distances amongst the taxa in the 16S
716 trees were calculated from the tree (weighted by the edge distances) with the *dendropy*
717 package in Python (77), using the functions *phylogenetic_distance_matrix* and
718 *patristic_distance*, with the default parameters.
719

720 **Calculation of the genetic distance between bacterial hosts**

721 We calculated the trinucleotide composition, or 3-mer genomic signature, of each bacterial
722 genome from the phage host species, using the relative abundance value of each of the 64
723 possible trinucleotides, as in (78). This is defined as the observed trinucleotide frequency
724 divided by the expected trinucleotide frequency (the product of the mononucleotide
725 frequencies), or

726 $$x_{ijk} = \frac{f_{ijk}}{f_i f_j f_k}$$

727 where f_i, f_j and f_k represent the frequency of the nucleotides *i*, *j* and *k*, respectively (with *i*, *j*,
728 *k* ∈ A, C, T, G). This allows to quantify the deviation of the observed frequency of trinucleotides
729 to the one expected given the nucleotide composition of the genomes, which are known to
730 differ between phages and bacteria (79). The genetic distance between two hosts was then
731 calculated using the average absolute difference between the 3-mer genomic composition of
732 each, as

733 $$\delta(a,b) = \frac{1}{64} \sum_{i \in \{ACGT\}} \sum_{j \in \{ACGT\}} \sum_{k \in \{ACGT\}} \left| x_{ijk,a} - x_{ijk,b} \right|$$

734 where $x_{ijk,a}$ and $x_{ijk,b}$ are the relative abundances of the trinucleotide *ijk* in each of the
735 bacterial hosts. Note that multiple genomes might be available for each host. To use all the
736 available information when multiple genomes are available for a species, we used the grand
737 mean of all pairwise comparisons. As an example, if one host is *Escherichia coli* and the

another is *Staphylococcus aureus*, we calculated the differences in genomic signature between each pairwise combinations of all genomes of *Escherichia coli* and all genomes of *Staphylococcus aureus*. We then calculate the average of all the pairwise calculations of the genetic distance as δ_T, with

$$\delta_T(a,b) = \frac{1}{N}\sum_a \sum_b \delta(a,b)$$

where *a* and *b* are individual strain genomes of the first and second host, respectively, and *N* is the total pairwise calculations between the strains of the first and second hosts.

ACKNOWLEDGEMENTS

We acknowledge support by the PRESTIGE programme (PRESTIGE-2017-1-0012), the ANR grants (ANR-16-CE15-0022 and ANR-16-CE16-0029), the Fondation pour la Recherche Médicale (Equipe FRM EQU201903007835), and the Laboratoire d'Excellence IBEID (ANR-10-LABX-62-IBEID). We thank Mireille Ansaldi, Marta Lourenço and the members of the Microbial Evolutionary Genomics laboratory by discussions and comments on earlier versions of this manuscript. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

AUTHOR CONTRIBUTIONS

Designed research: JMS, EPCR. Analyzed the data: JMS. Visualization: JMS. Provided materials and methods: EP, JMS, MT. Drafted the paper: JMS, EPCR. Revised the manuscript: all authors.

764 REFERENCES

765

766 1. Dion MB, Oechslin F, Moineau S (2020) Phage diversity, genomics and phylogeny.
767 *Nat Rev Microbiol* 437:356–14.

768 2. Harrison E, Brockhurst MA (2017) Ecological and Evolutionary Benefits of Temperate
769 Phage: What Does or Doesn't Kill You Makes You Stronger. *Bioessays*
770 197(12):1700112.

771 3. Touchon M, Moura de Sousa JA, Rocha EP (2017) Embracing the enemy: the
772 diversification of microbial gene repertoires by phage-mediated horizontal gene
773 transfer. *Curr Opin Microbiol* 38:66–73.

774 4. Touchon M, Bernheim A, Rocha EP (2016) Genetic and life-history traits associated
775 with the distribution of prophages in bacteria. *ISME J* 10(11):2744–2754.

776 5. Sausset R, Petit MA, Gaboriau-Routhiau V, De Paepe M (2020) New insights into
777 intestinal phages. *Mucosal Immunol* 16(11):760–11.

778 6. Canchaya C, Fournous G, Brüssow H (2004) The impact of prophages on bacterial
779 chromosomes. *Mol Microbiol* 53(1):9–18.

780 7. Penadés JR, Chen J, Quiles-Puchalt N, Carpena N, Novick RP (2015) Bacteriophage-
781 mediated spread of bacterial virulence genes. *Curr Opin Microbiol* 23:171–178.

782 8. Fillol-Salom A, et al. (2019) Bacteriophages benefit from generalized transduction.
783 *PLoS Pathog* 15(7):e1007888.

784 9. Hatfull GF, Hendrix RW (2011) Bacteriophages and their genomes. *Curr Opin Virol*
785 1(4):298–303.

786 10. Arbiol C, Comeau AM, Kutateladze M, Adamia R, Krisch HM (2010) Mobile
787 Regulatory Cassettes Mediate Modular Shuffling in T4-Type Phage Genomes.
788 *Genome Biol Evol* 2:140–152.

789 11. Devoto AE, et al. (2019) Megaphages infect Prevotella and variants are widespread in
790 gut microbiomes. *Nat Microbiol* 4(4):1–11.

791 12. Yuan Y, Gao M (2017) Jumbo Bacteriophages: An Overview. *Front Microbiol* 8:403.

792 13. Botstein D (1980) A theory of modular evolution for bacteriophages. *Ann N Y Acad*
793 *Sci* 354:484–490.

794 14. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S (2000) The origins and ongoing
795 evolution of viruses. *Trends Microbiol* 8(11):504–508.

796 15. Campbell A, Botstein D (1983) *Evolution of the lambdoid phages* (Cold Spring Harbor
797 Laboratory, Cold Spring Harbor, NY).

798 16. Smith GR (1983) *General Recombination* (Cold Spring Harbor, NY: Cold Spring
799 Harbor Laboratory).

800 17. De Paepe M, et al. (2014) Temperate Phages Acquire DNA from Defective
801 Prophages by Relaxed Homologous Recombination: The Role of Rad52-Like

802 Recombinases. *PLoS Genet* 10(3):e1004181–15.

803 18. Morris P, Marinelli LJ, Jacobs-Sera D, Hendrix RW, Hatfull GF (2008) Genomic
804 characterization of mycobacteriophage Giles: evidence for phage acquisition of host
805 DNA by illegitimate recombination. *J Bacteriol* 190(6):2172–2182.

806 19. d'Adda di Fagagna F, Weller GR, Doherty AJ, Jackson SP (2003) The Gam protein of
807 bacteriophage Mu is an orthologue of eukaryotic Ku. *EMBO Rep* 4(1):47–52.

808 20. Mavrich TN, Hatfull GF (2017) Bacteriophage evolution differs by host, lifestyle and
809 genome. *Nat Microbiol* 2:1–9.

810 21. Filée J, Tétart F, Suttle CA, Krisch HM (2005) Marine T4-type bacteriophages, a
811 ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci USA*
812 102(35):12471–12476.

813 22. Ignacio-Espinoza JC, Sullivan MB (2012) Phylogenomics of T4 cyanophages: lateral
814 gene transfer in the "core" and origins of host genes. *Environ Microbiol* 14(8):2113–
815 2126.

816 23. Chopin A, Bolotin A, Sorokin A, Ehrlich SD, Chopin M (2001) Analysis of six
817 prophages in Lactococcus lactis IL1403: different genetic structure of temperate and
818 virulent phage populations. *Nucleic Acids Res* 29(3):644–651.

819 24. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R (2008) Reticulate
820 Representation of Evolutionary and Functional Relationships between Phage
821 Genomes. 25(4):762–777.

822 25. Kupczok A, et al. (2018) Rates of Mutation and Recombination in Siphoviridae Phage
823 Genome Evolution over Three Decades. *Mol Biol Evol* 35(5):1147–1159.

824 26. Filée J, Bapteste E, Susko E, Krisch HM (2006) A selective barrier to horizontal gene
825 transfer in the T4-type bacteriophages that has preserved a core genome with the
826 viral replication and structural genes. *Mol Biol Evol* 23(9):1688–1696.

827 27. Dekel-Bird NP, et al. (2013) Diversity and evolutionary relationships of T7-like
828 podoviruses infecting marine cyanobacteria. *Environ Microbiol* 15(5):1476–1491.

829 28. Bouchard JD, Moineau S (2000) Homologous Recombination between a Lactococcal
830 Bacteriophage and the Chromosome of Its Host Strain. *Virology* 270(1):65–75.

831 29. Garneau JE, Tremblay DM, Moineau S (2008) Characterization of 1706, a virulent
832 phage from Lactococcus lactis with similarities to prophages from other Firmicutes.
833 *Virology* 373(2):298–309.

834 30. Martin DP, Lemey P, Posada D (2011) Analysing recombination in nucleotide
835 sequences. *Mol Ecol Resour* 11(6):943–955.

836 31. McNair K, Bailey BA, Edwards RA (2012) PHACTS, a computational approach to
837 classifying the lifestyle of phages. *Bioinformatics* 28(5):614–618.

838 32. Mann NH, Cook A, Millard A, Bailey S, Clokie M (2003) Bacterial photosynthesis
839 genes in a virus. *Nature* 424(6950):741–741.

840 33. Shen P, Huang HV (1986) Homologous recombination in Escherichia coli:
841 dependence on substrate length and homology. *Genetics* 112(3):441–457.

842 34. Rocha EPC, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the
843 bacterial homologous recombination systems. *PLoS Genet* 1(2):e15.

844 35. Aravind L, Koonin EV (2001) Prokaryotic homologs of the eukaryotic DNA-end-
845 binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic
846 double-strand break repair system. *Genome Res* 11(8):1365–1374.

847 36. Bhattacharyya S, et al. (2018) Phage Mu Gam protein promotes NHEJ in concert with
848 Escherichia coli ligase. *Proc Natl Acad Sci USA* 115(50):E11614–E11622.

849 37. Latino L, Essoh C, Blouin Y, Vu Thien H, Pourcel C (2014) A novel Pseudomonas
850 aeruginosa Bacteriophage, Ab31, a Chimera Formed from Temperate Phage PAJU2
851 and P. putida Lytic Phage AF: Characteristics and Mechanism of Bacterial
852 Resistance. *PLoS ONE* 9(4):e93777–16.

853 38. Durmaz E, Klaenhammer TR (2000) Genetic analysis of chromosomal regions of
854 Lactococcus lactis acquired by recombinant lytic phages. *Appl Environ Microbiol*
855 66(3):895–903.

856 39. Schuch R, Fischetti VA (2006) Detailed genomic analysis of the Wbeta and gamma
857 phages infecting Bacillus anthracis: implications for evolution of environmental fitness
858 and antibiotic resistance. *J Bacteriol* 188(8):3037–3051.

859 40. Hershey AD (1953) Nucleic acid economy in bacteria infected with bacteriophage T2.
860 *J Gen Physiol* 37(1):1–23.

861 41. Wiberg JS (1966) Mutants of bacteriophage T4 unable to cause breakdown of host
862 DNA. *Proc Natl Acad Sci USA* 55(3):614–621.

863 42. Arber W (2000) Genetic variation: molecular mechanisms and impact on microbial
864 evolution. *FEMS Microbiol Rev* 24(1):1–7.

865 43. Labrie SJ, Moineau S (2007) Abortive Infection Mechanisms and Prophage
866 Sequences Significantly Influence the Genetic Makeup of Emerging Lytic Lactococcal
867 Phages. *J Bacteriol* 189(4):1482–1487.

868 44. Shuman S, Glickman MS (2007) Bacterial DNA repair by non-homologous end
869 joining. *Nat Rev Microbiol* 5(11):852–861.

870 45. Chayot R, Montagne B, Mazel D, Ricchetti M (2010) An end-joining repair mechanism
871 in Escherichia coli. *Proc Natl Acad Sci USA* 107(5):2141–2146.

872 46. Su T, et al. (2019) The phage T4 DNA ligase mediates bacterial chromosome DSBs
873 repair as single component non-homologous end joining. *Synth Syst Biotechnol*
874 4(2):107–112.

875 47. Wu DY, Wallace RB (1989) Specificity of the nick-closing activity of bacteriophage T4
876 DNA ligase. *Gene* 76(2):245–254.

877 48. Hynes AP, et al. (2018) Widespread anti-CRISPR proteins in virulent bacteriophages
878 inhibit a range of Cas9 proteins. *Nat Commun* 9(1):2919–10.

879 49. Mathieu A, et al. (2020) Virulent coliphages in 1-year-old children fecal samples are
880 fewer, but more infectious than temperate coliphages. *Nat Commun*:1–12.

881 50. Brüssow H, Canchaya C, Hardt W-D (2004) Phages and the evolution of bacterial

882 pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol*
883 *Rev* 68(3):560–602– table of contents.

884 51. Obeng N, Pratama AA, van Elsas JD (2016) The Significance of Mutualistic Phages
885 for Bacterial Ecology and Evolution. *Trends Microbiol* 24(6):1–10.

886 52. Lindell D, et al. (2004) Transfer of photosynthesis genes to and from Prochlorococcus
887 viruses. *Proc Natl Acad Sci USA* 101(30):11013–11018.

888 53. Thompson LR, et al. (2011) Phage auxiliary metabolic genes and the redirection of
889 cyanobacterial host carbon metabolism. *Proc Natl Acad Sci USA* 108(39):E757–64.

890 54. Zeidner G, et al. (2005) Potential photosynthesis gene recombination between
891 Prochlorococcus and Synechococcus via viral intermediates. *Environ Microbiol*
892 7(10):1505–1513.

893 55. Mata M, Delstanche M, Robert-Baudouy J (1978) Isolation of specialized transducing
894 bacteriophages carrying the structural genes of the hexuronate system in Escherichia
895 coli K-12: exu region. *J Bacteriol* 133(2):549–557.

896 56. Mirold S, Rabsch W, Tschäpe H, Hardt WD (2001) Transfer of the Salmonella type III
897 effector sopE between unrelated phage families. *J Mol Biol* 312(1):7–16.

898 57. Mihara T, et al. (2016) Linking Virus Genomes with Host Taxonomy. *Viruses* 8(3):66.

899 58. Roux S, Enault F, Hurwitz BL, Sullivan MB (2015) VirSorter: mining viral signal from
900 microbial genomic data. *PeerJ* 3(348):e985.

901 59. Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence
902 searching for the analysis of massive data sets. *Nat Biotechnol* 35(11):1026–1028.

903 60. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of
904 communities in large networks. *J Stat Mech* 2008(10):P10008.

905 61. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol*
906 7(10):e1002195.

907 62. Grazziotin AL, Koonin EV, Kristensen DM (2017) Prokaryotic Virus Orthologous
908 Groups (pVOGs): a resource for comparative genomics and protein family annotation.
909 *Nucleic Acids Res* 45(D1):D491–D498.

910 63. Söding J (2005) Protein homology detection by HMM-HMM comparison.
911 *Bioinformatics* 21(7):951–960.

912 64. El-Gebali S, et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids*
913 *Res* 47(D1):D427–D432.

914 65. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families.
915 *Nucleic Acids Res* 31(1):371–373.

916 66. Fouts DE (2006) Phage_Finder: automated identification and classification of
917 prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*
918 34(20):5839–5851.

919 67. Huerta-Cepas J, et al. (2019) eggNOG 5.0: a hierarchical, functionally and
920 phylogenetically annotated orthology resource based on 5090 organisms and 2502

921 viruses. *Nucleic Acids Res* 47(D1):D309–D314.

922 68. Steinegger M, et al. (2019) HH-suite3 for fast remote homology detection and deep
923 protein annotation. *BMC Bioinformatics* 20(1):473–15.

924 69. Lopes A, Amarir-Bouhram J, Faure G, Petit M-A, Guerois R (2010) Detection of novel
925 recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote
926 homologs. *Nucleic Acids Res* 38(12):3952–3962.

927 70. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version
928 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.

929 71. Bernheim A, Bikard D, Touchon M, Rocha EPC (2019) A matter of background: DNA
930 repair pathways as a possible cause for the sparse distribution of CRISPR-Cas
931 systems in bacteria. *Philos Trans R Soc Lond, B, Biol Sci* 374(1772):20180088.

932 72. Bertrand C, Thibessard A, Bruand C, Lecointe F, Leblond P (2019) Bacterial NHEJ: a
933 never ending story. *Mol Microbiol* 111(5):1139–1151.

934 73. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments.
935 *Bioinformatics* 25(10):1335–1337.

936 74. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for
937 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*
938 25(15):1972–1973.

939 75. Nguyen L-T, Schmidt HA, Haeseler von A, Minh BQ (2015) IQ-TREE: a fast and
940 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol*
941 *Evol* 32(1):268–274.

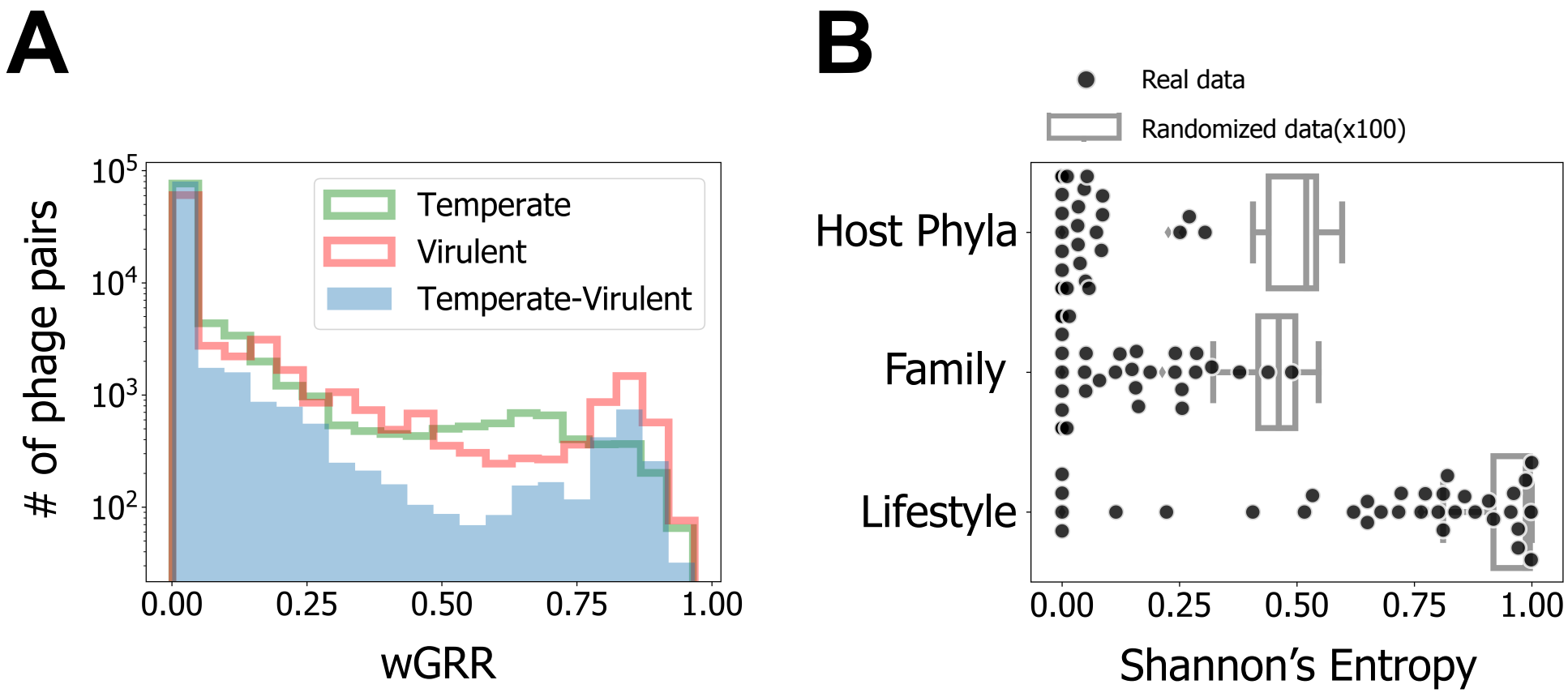
942 76. Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler von A, Jermiin LS (2017)
943 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*
944 14(6):587–589.

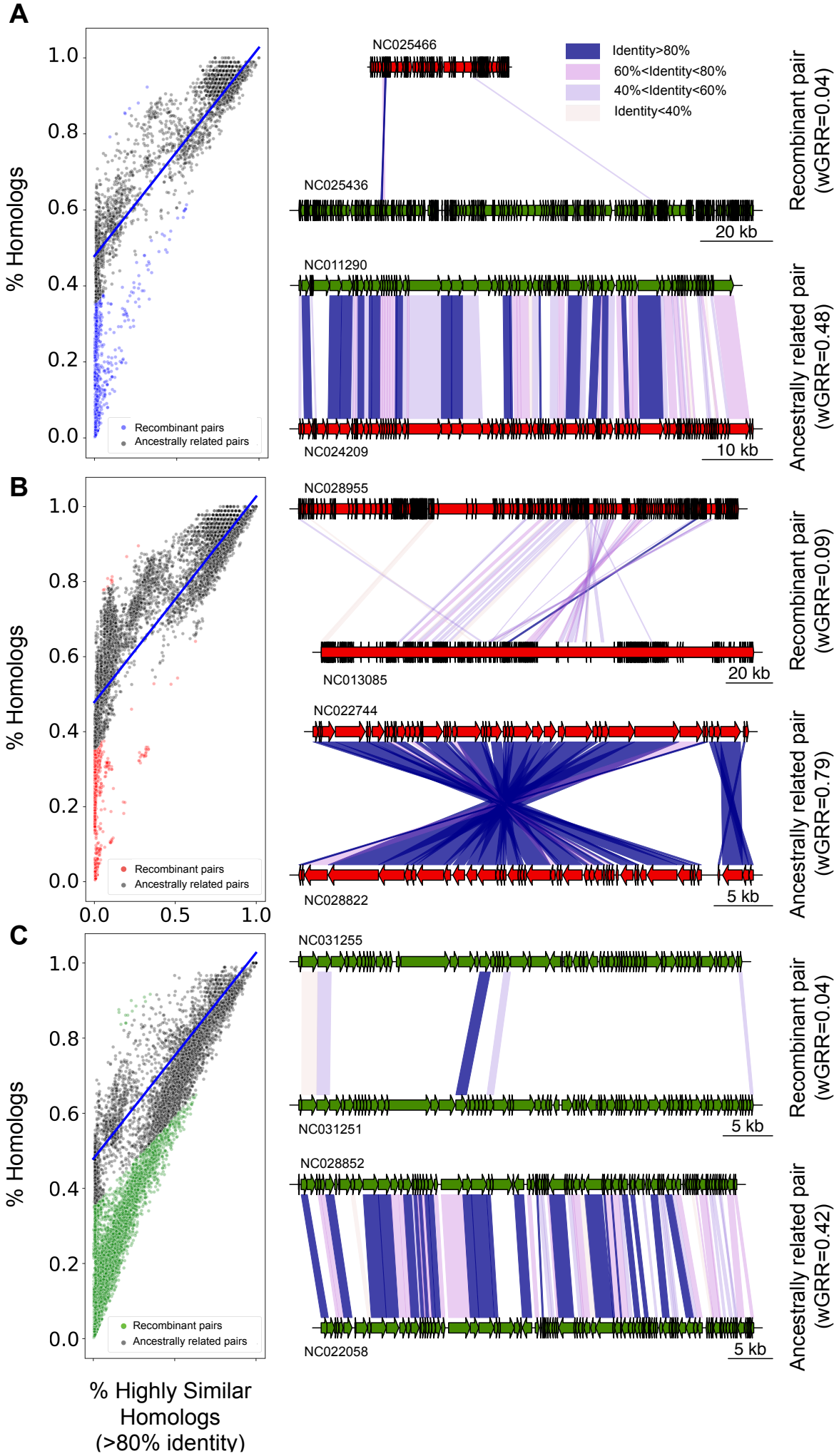
945 77. Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic
946 computing. *Bioinformatics* 26(12):1569–1571.

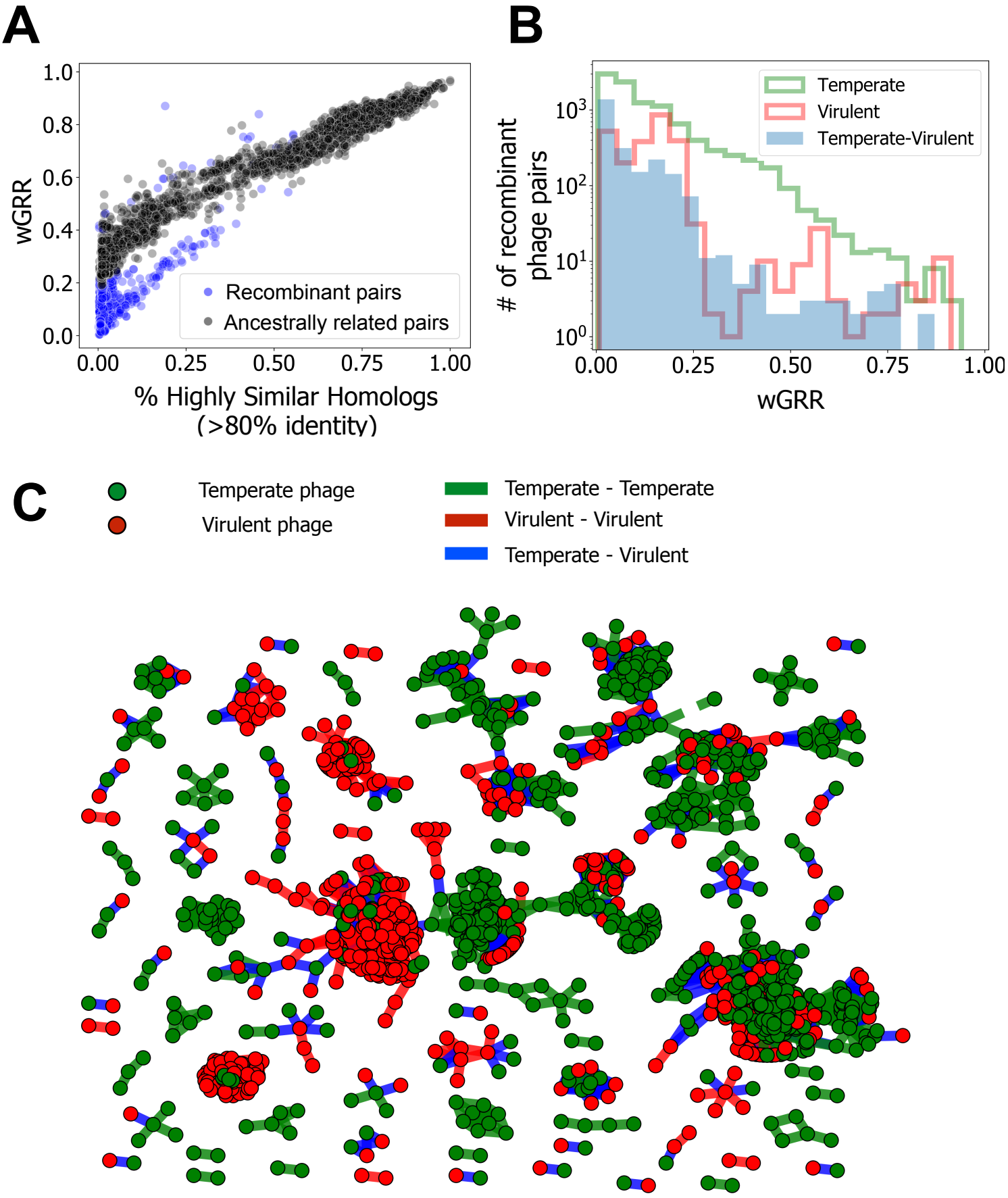
947 78. Suzuki H, Yano H, Brown CJ, Top EM (2010) Predicting plasmid promiscuity based
948 on genomic signature. *J Bacteriol* 192(22):6045–6055.

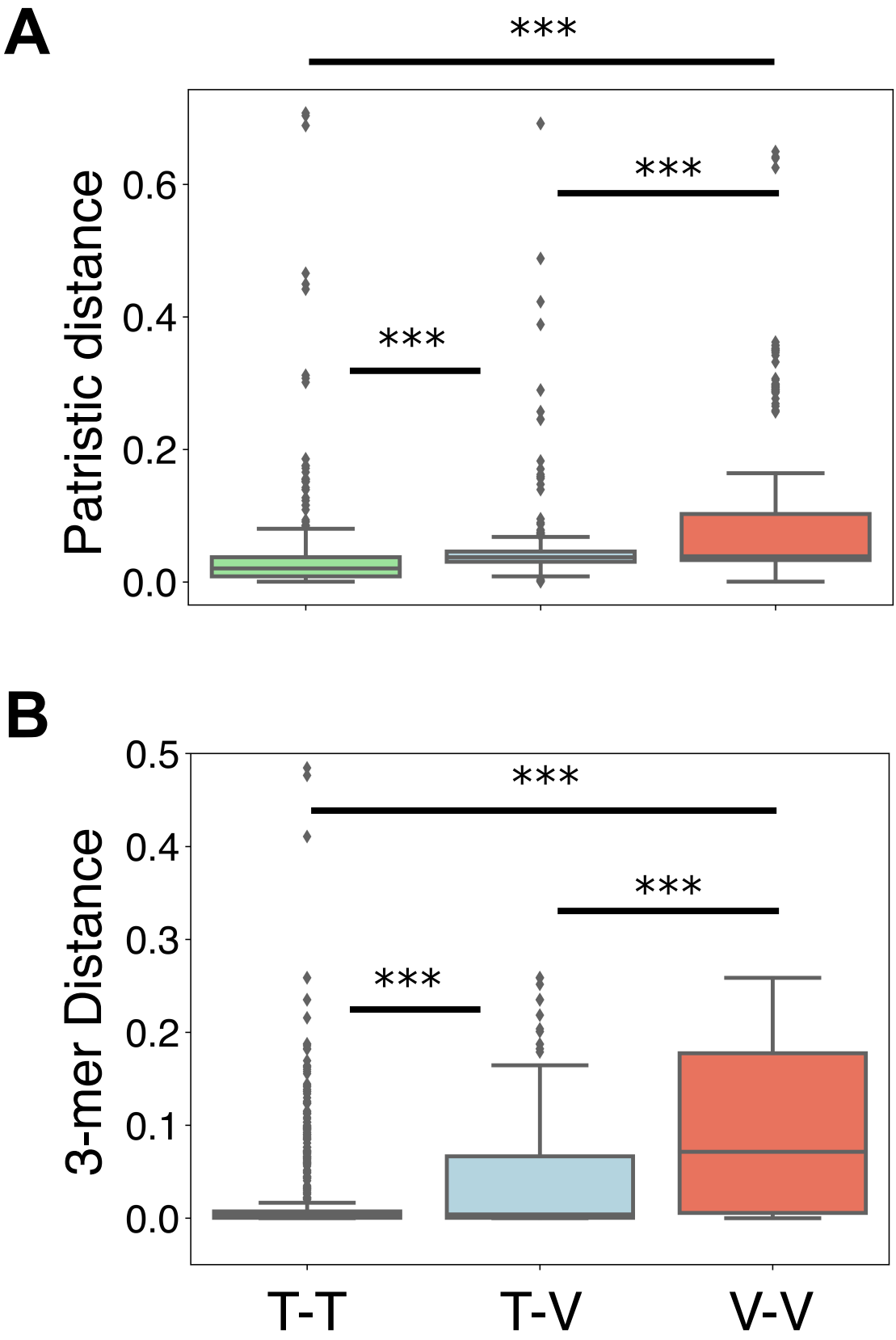
949 79. Rocha EPC, Danchin A (2002) Base composition bias might result from competition
950 for metabolic resources. *Trends Genet* 18(6):291–294.

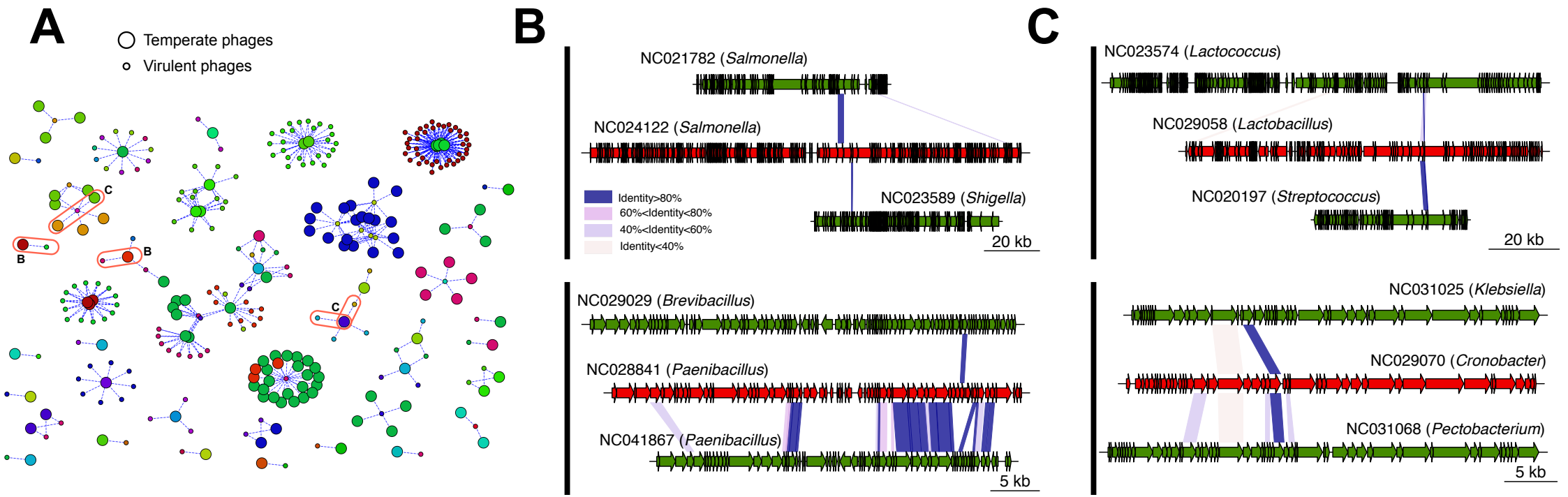
951


A
B
Temperate
Virulent
Temperate-Virulent
# of phage pairs
10^5
10^4
10^3
10^2
0.00
0.25
0.50
0.75
1.00
wGRR
Real data
Randomized data(x100)
Host Phyla
Family
Lifestyle
0.00
0.25
0.50
0.75
1.00
Shannon's Entropy


A
B
C
% Homologs
% Highly Similar Homologs (>80% identity)
Recombinant pairs
Ancestrally related pairs
NC025466
NC025436
NC011290
NC024209
NC028955
NC013085
NC022744
NC028822
NC031255
NC031251
NC028852
NC022058
Identity>80%
60%<Identity<80%
40%<Identity<60%
Identity<40%
20 kb
10 kb
5 kb
Recombinant pair (wGRR=0.04)
Ancestrally related pair (wGRR=0.48)
Recombinant pair (wGRR=0.09)
Ancestrally related pair (wGRR=0.79)
Recombinant pair (wGRR=0.04)
Ancestrally related pair (wGRR=0.42)


A
1.0
0.8
0.6
0.4
0.2
0.0
wGRR
Recombinant pairs
Ancestrally related pairs
0.00
0.25
0.50
0.75
1.00
% Highly Similar Homologs
(>80% identity)
B
Temperate
Virulent
Temperate-Virulent
# of recombinant
phage pairs
10^3
10^2
10^1
10^0
0.00
0.25
0.50
0.75
1.00
wGRR
C
Temperate phage
Virulent phage
Temperate - Temperate
Virulent - Virulent
Temperate - Virulent


A
***
***
***
Patristic distance
0.6
0.4
0.2
0.0
B
0.5
0.4
0.3
0.2
0.1
0.0
3-mer Distance
***
***
***
T-T
T-V
V-V


A
Temperate phages
Virulent phages
B
C
B
B
C
B
NC021782 (Salmonella)
NC024122 (Salmonella)
NC023589 (Shigella)
Identity>80%
60%<Identity<80%
40%<Identity<60%
Identity<40%
20 kb
NC029029 (Brevibacillus)
NC028841 (Paenibacillus)
NC041867 (Paenibacillus)
5 kb
C
NC023574 (Lactococcus)
NC029058 (Lactobacillus)
NC020197 (Streptococcus)
20 kb
NC031025 (Klebsiella)
NC029070 (Cronobacter)
NC031068 (Pectobacterium)
5 kb