Prophages and plasmids display opposite trends in the types of accessory genes they carry

1.1 Author names

Nobuto Takeuchi, ORCiD 0000-0003-4949-6476
Sophia Hamada-Zhu
Haruo Suzuki, ORCiD 0000-0003-1447-6109

1.2 Affiliation

Nobuto Takeuchi: School of Biological Sciences, the University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Nobuto Takeuchi: Universal Biology Institute, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Sophia Hamada-Zhu: School of Biological Sciences, the University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Haruo Suzuki: Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan

Haruo Suzuki: Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan

1.3 Corresponding author

Nobuto Takeuchi (nobuto.takeuchi@auckland.ac.nz)

1.4 Keyword

virus, bacteriophage, selfish genetic elements, evolution of pathogenicity, evolution of antimicrobial resistance, lateral gene transfer

2. Abstract

Mobile genetic elements (MGEs), such as phages and plasmids, often possess accessory genes encoding bacterial functions, facilitating bacterial evolution. Are there evolutionary rules governing the arsenal of accessory genes MGEs carry? If such rules exist, they might be reflected in the types of accessory genes different MGEs carry. To test this hypothesis, we compare prophages and plasmids with respect to the frequencies at which they carry antibiotic resistance genes (ARGs) and virulence factor genes (VFGs) in the genomes of 21 pathogenic bacterial species using public databases. Our results indicate that prophages tend to carry VFGs more frequently than ARGs in three species, whereas plasmids tend to carry ARGs more frequently than VFGs in nine species. In *Escherichia coli*, where these trends are detected, prophage-borne VFGs encode a much narrower range of functions than do plasmid-borne VFGs, typically involved in damaging host cells or modulating host

immunity. In the species where the above trends were not detected, ARGs and VFGs are barely found in prophages and plasmids. These results indicate that MGEs differentiate in the types of accessory genes they carry depending on their infection strategies, suggesting an evolutionary rule governing horizontal gene transfer mediated by MGEs.

3. Introduction

Mobile genetic elements (MGEs), such as phages and plasmids, often possess accessory genes encoding bacterial phenotypes that are not necessarily integral to the replication or transmission of MGEs. For example, plasmids frequently possess various antibiotic resistance genes (ARGs) [1-3]. Also, phages and plasmids possess virulence factor genes (VFGs) required for bacterial pathogenicity [4-8]. Mediating the horizontal transfer of accessory genes, MGEs play important roles in the evolution of bacterial genomes and phenotypes [9, 10].

MGEs are parasites of bacteria. Thus, horizontal gene transfer (HGT) mediated by MGEs can be regarded as the genetic manipulation of hosts by parasites [11]. Given that MGEs are self-interested evolving entities, MGEs are expected to possess accessory genes that advantage themselves [12, 13]. For example, plasmids are considered to gain selective advantages from ARGs by improving the survival of their bacterial hosts in heterogeneous environments [13-16]. It has also been hypothesised that phages gain selective advantages from VFGs by modifying environments in which their bacterial hosts live [17].

What evolutionary rules govern the arsenal of accessory genes carried by MGEs [12, 13]? Such rules, if they exist, might reflect the different infection strategies of MGEs. For example, phages typically lyse host cells to transmit to other cells, whereas plasmids do not. Consequently, phages might not gain much of an advantage by carrying genes that improve the survival of bacteria, such as ARGs, because if their hosts are in danger, they can enter a lytic replication cycle to abandon their hosts and seek new ones [18-22]. By contrast, plasmids cannot typically use such evacuative strategies, hence likely to benefit from genes improving host survival. Thus, to understand the rules governing HGT mediated by MGEs, it is beneficial to investigate whether different MGEs carry different types of accessory genes.

To address the above question, we consider an ongoing debate about phage-borne ARGs. While it is well established that plasmids frequently carry ARGs [1-3], how frequently phages carry ARGs is controversial [23]. Phages mediate HGT through multiple mechanisms, among which specialised transduction is the most similar to HGT mediated by plasmid [24] (see Discussion for the other mechanisms). In specialised transduction, phages transfer genes carried in their genomes. Therefore, specialised transduction is strictly coupled with the infectious transmission of phage genomes, the coupling that is also entailed in plasmid conjugation [24]. Laboratory experiments have demonstrated that phages are capable of transferring ARGs to bacteria through specialised transduction [25]. However, the specialised transduction of ARGs in nature has been scarcely documented [3, 26]. While metagenomic studies have detected ARGs in viral fractions of environmental DNA samples [27-30], other studies provide evidence suggesting that the detection of ARGs was due to the contamination of bacterial DNA in the viral fractions [31, 32]. Genomics studies have

predicted a number of prophages—i.e., phage genomes inserted into bacterial chromosomes as a consequence of specialised transduction—carrying ARGs in the genomes of *Acinetobacter baumannii* [33], *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* [34] (see also [35]). Also, a previous study has isolated 29 phages from wastewater, of which 15 carry ARGs, suggesting that phages frequently possess ARGs [36]. However, these results appear at odds with a recent comprehensive analysis of phage genomes in public databases, which shows that ARGs are carried by only 0.3% of phages [37]. Taken together, the existing studies present mixed messages about the frequency at which phages carry ARGs.

To investigate how frequently phages carry ARGs, here we compare the distributions of ARGs and VFGs between the prophages and plasmids of pathogenic bacteria by comprehensively analysing public databases. We consider prophages instead of phages to compare different MGEs belonging to the same bacterial genomes. Our approach is designed to mitigate two issues we consider to be involved in the computational analyses of ARGs encoded in prophages, which are not taken into account in previous studies [33-35]. First, the misidentification of prophages can cause systematic biases in the number of prophage-borne ARGs. For example, non-prophage regions can be misidentified as prophages, causing an overestimation in the number of prophage-borne ARGs. Contrariwise, a true prophage can be missed, which leads to an underestimation of the number of prophage-borne ARGs. To avoid these biases due to prophage prediction, we compare the number of prophage-borne ARGs to that of prophage-borne VFGs, where both numbers are expected to be biased by common factors so that the biases can be cancelled out. The second issue involved in the analysis of prophage-borne ARGs is a sampling bias in bacterial genomes, which can cause overestimation in the numbers of ARGs and VFGs owing to the double-counting of orthologous genes. The degree to which this bias occurs can differ between ARGs and VFGs. To correct this bias, we cluster all genes into putative orthologous groups based on sequence similarity and synteny conservation and count the numbers of putative orthologous groups of ARGs and VFGs (OGARGs and OGVFGs, respectively). Furthermore, to investigate a potential differentiation between prophages and plasmids, we analyse the distributions of ARGs and VFGs in plasmids. Finally, we examine whether prophages and plasmids also differ in the functional categories of VFGs they carry. Our results suggest that prophages are biased towards carrying VFGs, whereas plasmids are biased towards carrying ARGs. However, in many species, both ARGs and VFGs are hardly detected in prophages and plasmids. Moreover, we found that prophage-borne VFGs are less functionally diverse than plasmid-borne VFGs in *Escherichia coli*. Taken together, these results indicate that prophages and plasmids differ in the types of accessory genes they carry.

4. Methods

Our method is sketched in Figure 1.

4.1 Data acquisition

Three VFG databases, namely, VFDB (3685 genes in set A), Victors (5085 genes), and PATRIC_VF (1293 genes), were downloaded from the respective websites in December 2020

[38-40]. The VFG entries in Victors were refined by removing those carried by non-bacterial pathogens or lacking NCBI protein GIs (4575 out of 5085 remained). Some VFG entries in Victors were missing protein sequences, which were downloaded from Genbank based on their protein GIs [41]. All VFG entries in the three databases were pooled and clustered to remove redundancy with CD-HIT with the protein sequence identity threshold of 1.0 [42], resulting in a combined VFG database of 7218 entries.

The genome assemblies with the 'Complete' status were downloaded from RefSeq in September 2021 with the following criteria [43]: a species had at least 60 complete genomes in RefSeq and at least 70 VFGs in the combined VFG database. These criteria resulted in 21 species of bacterial pathogens spanning three phyla, Actinobacteria, Firmicute, and Proteobacteria, as follows (numbers in brackets indicate the number of genomes examined in this study): *Acinetobacter baumannii* (275), *Bacillus anthracis* (99), *Bordetella pertussis* (562), *Brucella melitensis* (64), *Burkholderia pseudomallei* (126), *Campylobacter jejuni* (220), *Enterococcus faecalis* (64), *E. coli* (1444; *E. coli* was considered to include *Shigella boydii* (17), *Shigella dysenteriae* (21), *Shigella flexneri* (35), and *Shigella sonnei* (1)), *Haemophilus influenzae* (92), *Helicobacter pylori* (225), *Klebsiella pneumoniae* (873), *Legionella pneumophila* (102), *Listeria monocytogenes* (263), *Mycobacterium tuberculosis* (285), *Neisseria meningitidis* (122), *Pseudomonas aeruginosa* (320), *Salmonella enterica* (996), *Staphylococcus aureus* (618), *Streptococcus agalactiae* (91), *Streptococcus pyogenes* (235), and *Vibrio cholerae* (99). The complete list of 7175 genomes analysed in this study is in Table S1.

4.2 Prophage prediction

To predict prophages in the bacterial genomes, VIBRANT (version 1.2.1) was used for the following reasons. First, VIBRANT has comparatively high performance, as reported by a recent benchmark [44]. Second, it is a standalone tool, so it can be run on local computers. Third, its algorithm is based on the similarity search of known phage proteins rather than predicting prophages based on nucleotide signatures. VIBRANT was run against the genomic sequences with default parameters.

4.3 Plasmid prediction

Sequences (contigs) were considered as plasmids or chromosomes if they were annotated as such in the RefSeq assembly report files (in total, there were 9279 plasmid and 6607 chromosome contigs). Contigs annotated as 'Segment', 'Genome Segment', or 'Extrachromosomal Element' were ignored (3, 1, or 5 contigs, respectively).

4.4 ARG prediction

To predict ARGs in the bacterial genomes, AMRFinderPlus (version 3.10.5) was run against translated coding sequences with the core subset of the database (version 2021-09-11.1) and, if possible, the organism option enabled [45]. The genes predicted by AMRFinderPlus as ARGs (i.e., 'element subtype AMR') and not annotated as pseudo-genes in RefSeq were considered as ARGs (those predicted as 'element subtype POINT', which contain point mutations associated with AR, were excluded).

174

175 ## 4.5 VFG prediction

176 To predict VFGs in the bacterial genomes, every entry in the combined VFG database (VFG
177 database entry, for short) was queried against the translated coding sequences of every
178 bacterial genome with BLASTP with an E-value threshold of 1e-9 [46]. A gene in a bacterial
179 genome (bacterial gene, for short) could match multiple VFG database entries, in which case
180 the VFG database entry with the highest bit-score was selected as the best match. A
181 bacterial gene was considered as encoding VF if it met the following additional criteria: (i) it
182 was not annotated as a pseudo-gene in RefSeq [43]; (ii) the BLASTP alignment between the
183 bacterial gene and its best-match VFG database entry, if any, had at least 80% sequence
184 identity and covers at least 80% of both the bacterial gene and the best-match VFG
185 database entry; (iii) the species of the genome in which the bacterial gene resides was
186 identical to the species in which the best-match VFG database entry resides [38-40].
187

188 ## 4.6 Orthology prediction

189 To avoid double-counting orthologous genes in closely related strains, all genes within a
190 species were clustered into putative orthologous groups based on protein sequence
191 similarity and synteny conservation, as follows. First, preliminary orthologous pairs of genes
192 were identified between every pair of genomes within each species through all-against-all
193 sequence similarity searches using ProteinOrtho version 6.0.25 (with DIAMOND ver. 2.0.6
194 [47]; E-value cut-off of 1e-5; minimum coverage of best alignments of 75%; minimum per-
195 cent identity of best alignments of 25%; minimum reciprocal similarity of 0.95) [48].
196 ProteinOrtho defines a preliminary orthologous pair of genes as a reciprocal nearly-best hit
197 (RNBH), as follows. A nearly-best hit (NBH) of a gene queried against a target genome is
198 defined as a hit whose bit-score is not smaller by a factor *f* than that of the best hit. The
199 value of *f* was 0.95, which is the default value of ProteinOrtho. If two genes are mutually
200 NBH of each other, they form RNBH [48].
201
202 RNBHs obtained with ProteinOrtho were pruned based on synteny conservation with an in-
203 house script, as follows. Let *x* and *y* be a pair of genes forming RNBH, and let *X* and *Y* be the
204 genomic neighbours of *x* and *y*, respectively, where *X* is defined as a set of 21 genes
205 consisting of ten genes upstream of *x*, ten genes downstream of *x*, and *x* itself, and *Y* is
206 likewise defined in terms of *y* (all contigs were assumed to be circular, and the orientation of
207 genes were ignored). Let N_x and N_y be the number of genes in *X* and *Y* that form RNBHs with
208 at least one gene in *Y* and *X*, respectively (note that a single gene in one genome can form
209 RNBHs with multiple genes in another genome owing to tandem duplication). If both N_x and
210 N_y are greater than ten (i.e., a majority of the genes in *X* form RNBHs with the genes in *Y*,
211 and *vice versa*), the RNBH formed by *x* and *y* was kept; otherwise, it was discarded [49].
212
213 Finally, the pruned RNBHs were clustered into putative orthologous groups with the spectral
214 clustering algorithm implemented in ProteinOrtho version 6.0.25 (minimum algebraic
215 connectivity of 0.1; exact step 3; minimum number of species of 0; purity of 1e-7) [48].
216

4.7 Classification of orthologous gene groups

A gene (VFG or ARG) was considered to be encoded in a prophage if the entire gene is included within a genomic region predicted as a prophage. Similarly, a gene was considered to be encoded in a plasmid if the gene resides in a plasmid contigue. Moreover, genes in prophages residing in plasmids (plasmidic prophages, for short) are distinguished from those in prophages residing in chromosomes (chromosomal prophages) for two reasons. First, it was ambiguous whether genes in plasmidic prophages should be regarded as encoded by plasmids, prophages, or both. Second, plasmidic prophages potentially represent a distinct class of MGEs called phage-plasmids [50].

An orthologous group of genes was considered to be encoded in chromosomal prophages, plasmidic prophages, or plasmids if the majority of the genes belonging to the group were encoded in the respective genomic contexts (the cases of ties were ignored).

An orthologous group of genes was considered an ARG or VFG if the majority of the genes belonging to the group were predicted as ARGs or VFGs, respectively. The majority rule was used because a subset of genes in OGARG or OGVFG could be predicted as non-ARGs or non-VFGs, respectively, owing to sequence divergence. However, for most orthologous groups of ARGs and VFGs, all genes in a group were predicted as either ARGs or VFGs. Moreover, no orthologous group contained both ARGs and VFGs.

4.8 Functional classification of prophage- and plasmid-borne VFGs

To classify the functions of VFGs, gene symbols (i.e., abbreviated gene names) were assigned to prophage-borne and plasmid-borne OGVFGs in *E. coli*, as follows. All OGVFGs in *E. coli* genomes were collectively associated with 862 best-match VFG database entries (see Method under 'VFG prediction'). Each of these entries was looked up in the respective VFG database (VFDB, Victors, or PATRIC_VF) to find a gene symbol assigned to it [38-40]. When a gene symbol was missing, a gene symbol was manually determined by querying the protein sequence of a VFG database entry against VFDB, the NCBI Conserved Domain Database (CDD), and the NCBI NR database [38, 51, 52] (Table S2). Consequently, 856 out of the 862 VFG database entries were assigned gene symbols (Table S2). The remaining six entries, which are annotated as hypothetical proteins in the VFG databases, were discarded from the subsequent analyses together with the OGVFGs associated with them. Next, prophage-borne and plasmid-borne OGVFGs were assigned gene symbols based on what VFG database entries they were associated with. If an OGVFG was associated with VFG database entries having an identical gene symbol, that gene symbol was assigned. If an OGVFG was associated with multiple VFG database entries with conflicting gene symbols (only 17 out of 3922 OGVFGs were of this type), the conflict was manually resolved by querying these entries against VFDB, the NCBI Conserved Domain Database, and the NCBI NR database (Table S2). Consequently, prophage-borne and plasmid-borne OGVFGs were collectively assigned 232 gene symbols (Table S2). Subsequently, these gene symbols were also assigned to OGVFGs that are neither prophage-borne nor plasmid-borne.

The 232 gene symbols were manually categorised by their functions based on information obtained from VFDB, NCBI CDD, and Uniprot [38, 51, 53] (Table S3). While most of the

262 functional categories were adopted from VFDB [38], a new category, 'phage-related', was
263 added, and the two categories, 'exoenzymes' and 'enzymes', were merged into 'enzymes.'
264 Six gene symbols were annotated to have multiple functions and, therefore, placed under
265 multiple categories.

266 5. Results

267 5.1 Prophage prediction

268 To examine the distribution of ARGs in prophages, we computationally predicted prophages
269 using VIBRANT [54] and ARGs using AMRFinderPlus [45] in the genomes of 21 pathogenic
270 bacterial species downloaded from the RefSeq database [43] (Methods; Tables S1 and S4).
271 To avoid double-counting orthologous ARGs in closely related strains, we clustered all genes
272 within a species into putative orthologous groups based on sequence similarity and synteny
273 conservation (Methods). We then counted the number of orthologous groups of ARGs
274 (OGARGs) encoded in the predicted prophages, distinguishing between prophages residing
275 in bacterial chromosomes and prophages residing in plasmids (chromosomal prophage and
276 plasmidic prophage, respectively, for short). This distinction was made because it was
277 ambiguous whether ARGs in plasmidic prophages should be regarded as encoded in
278 prophages, plasmids, or both (plasmidic prophages potentially represent phage-plasmids,
279 which are a separate class of MGEs from typical phages and plasmids [50]). The result shows
280 that a few to several prophage-borne OGARGs were detected in ten of the 21 examined
281 species (Tables 1 and S5).
282
283 To probe the precision of the above prediction, we manually examined whether predicted
284 prophages carrying ARG contained at least one phage-structure gene (e.g., phage baseplate,
285 capsid, portal, tail, tail fibre, tail sheath, tail assembly, head-tail connector, and tail tape
286 measure) according to the RefSeq annotation. Although this criterion cannot perfectly
287 distinguish true prophages from false ones, it allows us to split the predicted prophages into
288 those enriched with true prophages and those enriched with false prophages, allowing us to
289 probe the precision of the prophage prediction. The result of the examination shows that 24
290 out of 29 (i.e., 83% of) chromosomal prophages carrying ARGs contained phage-structure
291 genes, whereas seven out of 18 (i.e., 39% of) plasmidic prophages carrying ARGs contained
292 phage-structure genes (Table S6). This result means that 28 out of 33 (i.e., 85% of) OGARGs
293 in chromosomal prophages are in prophages carrying phage-structure genes, whereas nine
294 out of 21 (i.e., 43% of) OGARGs in plasmidic prophages are in prophages carrying phage-
295 structure genes (Table 1; note that the number of prophages carrying ARGs is smaller than
296 that of prophage-borne ARGs because one prophage can carry multiple ARGs). This result
297 suggests that the precision of VIBRANT is acceptable for chromosomal prophages.
298
299 During the manual examination, we noticed that a subset of plasmidic prophages carrying
300 ARGs also carried genes encoding integron integrases according to the RefSeq annotation
301 (Tables S6). Moreover, these prophages have a higher frequency of lacking phage-structure
302 genes than those without integron integrase genes (IIGs, for short). Specifically, seven out of
303 ten (i.e., 70% of) prophages carrying both ARGs and IIGs lack phage-structure genes,
304 whereas nine out of 37 (i.e., 24% of) prophages carrying ARGs and lacking IIGs lack phage-
305 structure genes. VIBRANT predicted that the genes annotated by RefSeq as IIGs matched
306 phage integrase genes, which are a typical component of phage genomes. However, the

RefSeq annotation implies that proteins encoded by these genes are more similar to integron integrases than phage integrases because the RefSeq annotation considers a broader set of protein families than does VIBRANT [43, 54]. Therefore, the above result suggests that about half of the prophages carrying ARGs and lacking phage-structure genes might have arisen from the misidentification of integrons as prophages (see also [55]).

5.2 Prophages carry significantly more VFGs than ARGs in multiple species

To compare the frequencies of ARGs to those of VFGs in chromosomal prophages, we used BLASTP [46] to search bacterial genomes for VFGs collected from VFDB [38], Victors [39] and PATRIC_VF [40] (Methods). We then counted the number of orthologous groups of VFGs (OGVFGs, for short) encoded in the predicted prophages (Tables 1 and S7). The number of OGVFGs cannot directly be compared to that of OGARGs because we have no *a priori* reason to expect that bacteria possess an equal number of VFGs and ARGs. Thus, we instead compared the relative frequencies of OGVFGs and OGARGs against the genomic background (see Figure 1 for the illustration of what we did). Specifically, we performed binomial tests under the null hypothesis that the relative frequencies of OGARGs and OGVFGs in the chromosomal prophages of each species are the same as those of all OGARGs and OGVFGs in the genomes of the respective species. In this test, we included prophages carrying ARGs and lacking phage-structure genes for fairness because prophages carrying VFGs were too numerous to be manually examined (however, we found that none of the prophages carrying VFGs carried IIG, which suggests that the prophages carrying VFGs do not contain false positives arising from integron misidentification). Also, we corrected P values using the Holm–Bonferroni method to control the family-wise error rate of all the statistical tests conducted in this study [56, 57]. The results of the tests indicate that the relative frequencies of OGARGs and OGVFGs in chromosomal prophages are significantly different from those of respective genomic backgrounds in the following three species (Figure 2): *E. coli* (Gammaproteobacteria), *S. enterica* (Gammaproteobacteria), and *S. aureus* (Firmicute). In all these species, chromosomal prophages carry VFGs more frequently than ARGs. The remaining 18 species, where significant biases were not detected, can be grouped into four categories: the one species where many prophage-borne VFGs and no prophage-borne ARGs were detected, but a bias towards VFGs was not statistically significant (*S. pyogenes*), the eight species where only a small number of prophage-borne ARGs or VFGs were detected (*A. baumannii*, *B. anthracis*, *H. influenzae*, *K. pneumoniae*, *N. meningitidis*, *P. aeruginosa*, *S. agalactiae*, and *V. cholerae*), the six species where no prophage-borne ARG and VFG were detected (*B. pertussis*, *B. melitensis*, *B. pseudomallei*, *C. jejuni*, *E. faecalis*, *L. monocytogenes*), and the three species where prophages were rarely or hardly detected (*H. pylori*, *L. pneumophila*, and *M. tuberculosis*) (Table 1 and Figure 2). The paucity or absence of prophage-borne ARGs and VFGs could be due to the limited sensitivity of the prophage prediction tool. However, a large number of prophages were predicted in all species except for *H. pylori*, *L. pneumophila*, and *M. tuberculosis*. Therefore, if the prophage prediction tool missed prophages carrying ARGs or VFGs, those prophages are likely to be distinct from the currently known phages. Taken together, the above results suggest that prophages tend to carry ARGs less frequently than VFGs if they carry a sufficient number of ARGs or VFGs; however, in many bacterial species, they carry little or no ARGs and VFGs.

353 ### 5.3 Plasmids carry more ARGs than VFGs in many species

354 To compare prophages to plasmids, we counted the number of OGARGs and OGVFGs
355 encoded in plasmids (Table 2). We then performed similar binomial tests under the null
356 hypothesis that the relative frequencies of OGARGs and OGVFGs in the plasmids of each
357 species are the same as those of all OGARGs and OGVFGs in the genomes of the respective
358 species. The results of the tests indicate that the relative frequencies of OGARGs and
359 OGVFGs in plasmids are significantly different from those of the genomic background in nine
360 out of the 21 examined species (Figure 3). In all nine species, plasmids are biased towards
361 carrying ARGs more frequently than VFGs. These species include six species of
362 Gammaproteobacteria (*A. baumannii*, *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *S. enterica*, and
363 *V. cholerae*), one species of Epsilonproteobacteria (*C. jejuni*), and two species of Firmicutes
364 (*E. faecalis* and *S. aureus*). The remaining 12 species, where significant biases were not
365 detected, can be grouped into three categories: the four species where plasmids are hardly
366 present (*B. pertussis*, *B. melitensis*, *M. tuberculosis*, and *N. meningitidis*), the seven species
367 where plasmids are rare, and plasmid-borne ARGs and VFGs were barely detected (*B.*
368 *pseudomallei*, *H. influenzae*, *H. pylori*, *L. pneumophila*, *L. monocytogenes*, *S. agalactiae*, and
369 *S. pyogenes*), and the one species where plasmids carry more VFGs than ARGs, but a bias
370 was not significant (*B. anthracis*). Taken together, the above results suggest that plasmids
371 tend to carry ARGs more frequently than VFGs if they carry a sufficient number of VFGs or
372 ARGs, with a possible exception of *B. anthracis*; however, in several species, plasmids barely
373 carry VFGs and ARGs.
374

375 ### 5.4 Prophage-borne VFGs are functionally more restricted than plasmid-borne VFGs

376 We next asked whether chromosomal prophages and plasmids differ in the functional
377 categories of VFGs they carry. To address this question, we mapped prophage-borne and
378 plasmid-borne OGVFGs in *E. coli* to gene symbols (i.e., abbreviated gene names) and
379 classified their functions using VFDB, NCBI CDD, and UniProt [38, 51, 53] (Methods under
380 "Functional classification of prophage- and plasmid-borne VFGs"). The result shows that
381 prophage-borne VFGs are mapped to 47 unique gene symbols, whereas plasmid-borne VFGs
382 are mapped to 186 unique gene symbols (Table 3). This result suggests that the range of
383 functions encoded by prophage-borne VFGs is less diverse than that encoded by plasmid-
384 borne VFGs, as described in more detail below.
385

386 Nearly half (viz., 22 out of 47) of the gene symbols assigned to prophage-borne VFGs
387 (prophage-borne VFG symbols, for short) are categorised as the effectors of type III
388 secretion system (T3SS), the proteins secreted by the T3SS apparatus (Table 3). The T3SS
389 effectors are involved in the destruction of host cells or the modulation of host immune
390 systems [58]. The other prophage-borne VFG symbols include those encoding exotoxins
391 (e.g., Shiga toxins), immune modulation, and phage-related components (e.g., phage tails)
392 (Table 3). Phage-related components are represented by the largest number of prophage-
393 borne OGVFGs (viz., 930 out of 1580) (Table 3). Despite their abundance, the specific
394 function of these genes is largely unknown, although some may be involved in intestinal
395 colonisation [59, 60]. In addition, two prophage-borne VFG symbols are involved in
396 adherence to host cells (Table 3). However, only one of them (porcine attaching-effacing
397 associated protein, *paa*) is frequently located in prophages, whereas the other (*E. coli*

common pilus chaperon, *yagV*/*ecpE*) is rarely located in prophages (Table S3). Taken together, the above results indicate that many prophage-borne VFGs are involved in causing direct damage to host cells or suppressing host immune systems, and a relatively small number of them are involved in host attachment.

By contrast, the gene symbols assigned to plasmid-borne VFGs (plasmid-borne VFG symbols) include a much wider range of functions, including T3SS effectors and apparatus, exotoxins, immune modulation, adherence and invasion to host cells, Type V secretion systems, enzymes (including exoenzymes), metal uptake, and gene regulation, anti-aggregation (dispersin), biofilm formation, motility, and so forth. The above results indicate that plasmid-borne VFGs encode a much wider range of functions than those encoded by prophage-borne VFGs.

Notwithstanding their comparatively small functional diversity, prophage-borne VFGs appear to play quantitatively no less important roles than those played by plasmid-borne VFGs in the virulence of *E. coli*. This is because the total number of prophage-borne OGVFGs is comparable to that of plasmid-borne OGVFGs (1580 versus 2342 OGVFGs; Table 3). Moreover, the total number of prophage-borne VFGs without clustering of orthologous genes is greater than that of plasmid-borne VFGs (10965 versus 8020 VFGs; Table 3).

Another notable pattern observed in the above analysis is that almost all gene symbols (220 out of 232) are associated with either prophage-borne or plasmid-borne VFGs, but not both (Table S3). This result means that prophages and plasmids carry distinct types of VFG even within the same functional categories. For example, within T3SS effectors, 22 and 32 gene symbols are associated with prophage-borne and plasmid-borne VFGs, respectively; however, only three gene symbols are associated with both prophage-borne and plasmid-borne VFGs (Table S3). This result suggests that prophages and plasmids have separate gene pools of virulence factors.

Taken together, the above results indicate that prophages and plasmids differ, not only in the functional diversity of VFGs they carry, but also in the specific types of VFG they carry within the same functional categories.

6. Discussion

The results presented above indicate that prophages tend to carry VFGs more frequently than ARGs in three of the 21 examined species. In most of the other species, prophage-borne ARGs and VFGs were hardly detected. In contrast, plasmids carry ARGs more frequently than VFGs in nine of the 21 examined species. In most of the other species, plasmid-borne ARGs and VFGs were barely detected. Taken together, these results indicate that if prophages and plasmids carry a sufficient number of VFGs or ARGs, they display opposite trends: prophages are biased towards carrying VFGs, whereas plasmids are biased toward carrying ARGs. Moreover, the comparison between prophage-borne and plasmid-borne VFGs shows that prophage-borne VFGs are functionally less diverse than plasmid-borne VFGs in *E. coli*. Taken together, the above results indicate that prophages and plasmids carry different types of bacterial accessory genes.

443
444 Based on the results presented above, we formulate the following hypothesis to test for the
445 future (Figure 4). Temperate phages do not benefit much from carrying ARGs because if
446 their hosts are in danger, they can abandon their hosts and seek new hosts by undergoing
447 lytic replication [18-22]. In contrast, both phages and plasmids can benefit from VFGs
448 because VFGs can accelerate bacterial replication by making bacteria exploit their hosts
449 more aggressively [17]. In addition, our results suggest that plasmids benefit from a wider
450 range of virulence functions than that from which phages benefit in *E. coli* (Table 3).
451
452 The results described above, however, do not necessarily mean that plasmids rarely carry
453 VFGs or that prophages always carry VFGs. For example, in *K. pneumoniae*, plasmids carry
454 ARGs more frequently than VFGs; nevertheless, plasmids carry a large number of VFGs (viz.,
455 96 out of 372 orthologous groups), whereas prophages carry none (Tables 1 and 2). A
456 similar, yet less striking, pattern is also seen in *E. faecalis*. Given that many prophages are
457 predicted in these species (Table 1, Figure 2), the absence of prophage-borne VFGs is
458 unlikely to be merely due to the limited sensitivity of the prophage detection tool. By
459 contrast, in *S. aureus* and *S. pyogenes*, prophages carry greater numbers of VFGs than do
460 plasmids (Tables 1 and 2). This heterogeneity in the distribution of VFGs suggests that
461 plasmids and prophages play variable roles in the pathogenicity of different bacterial
462 species.
463
464 The absolute numbers of prophage-borne ARGs reported in this study need to be
465 interpreted with caution because the prophage prediction tool has limited sensitivity and
466 precision. In particular, limited sensitivity implies that prophages carrying ARGs might have
467 been missed, so the results do not necessarily indicate that prophages are truly devoid of
468 ARGs in many bacterial species. However, the conclusion of this study does not directly
469 depend on the absolute numbers of prophage-borne ARGs because it is based on the
470 comparison between the relative frequencies of prophage-borne ARGs and VFGs and those
471 of plasmid-borne ARGs and VFGs. This comparison hinges on the assumption that the
472 prediction of prophages carrying ARGs and VFGs are biased by common factors so that the
473 biases are cancelled out.
474
475 Regarding the limitations of prophage prediction tools, it is pertinent to discuss a
476 discrepancy between our result and the result of Kondo et al. (2021) with respect to *P.*
477 *aeruginosa* [34]. While we found no prophage-borne ARGs in *P. aeruginosa*, Kondo et al.
478 (2021) report that more than 10% of *P. aeruginosa* genomes possess prophage-borne ARGs
479 [34]. The important difference between Kondo et al. (2021) and our study is that they used
480 different tools for prophage prediction: Kondo et al. (2021) uses PHASTER [61], whereas we
481 used VIBRANT [54]. To investigate the cause of the above discrepancy, we manually
482 examined the 11 prophages described in Kondo et al. (2021) that carry ARGs and are
483 predicted as 'intact' by PHASTER in *P. aeruginosa* [34]. We found that the examined
484 prophages could be grouped into two categories (Table S8). In the first group (five
485 prophages), both VIBRANT and PHASTER predicted prophages in almost the same genomic
486 locations. However, PHASTER predicted longer genomic regions that included ARGs as
487 prophages, whereas VIBRANT predicted shorter regions excluding ARGs. We do not know
488 which prophage boundaries are more accurate. In the second group (six prophages),

prophages were predicted only by PHASTER. These prophages, however, contained no phage-structure genes. Although they contained phage-related genes, such as integrase, transposase, and protease, these genes are not exclusively associated with phages. Moreover, PHASTER annotated tellurium resistance proteins, TerD, as virion structural proteins in three prophages in the second group, which is likely to be an error. These findings suggest that the second group of the prophages could be false positives. Prophages predicted as 'incomplete' or 'questionable' by PHASTER are less likely to be true than those predicted as 'intact.' Taken together, the above results suggest that the frequency of prophage-borne ARGs in *P. aeruginosa* is potentially underestimated in our study and overestimated in Kondo et al. (2021) owing to the limitations of the prophage prediction tools.

In interpreting the results obtained in this study, we assumed that ARGs and VFGs found within prophages were carried by phage genomes. However, these genes could have been inserted into pre-existing inactivated prophages (i.e., inserted after lysogeny). Although this possibility cannot be completely excluded, the following evidence suggests that not all ARGs and VFGs are inserted into pre-existing inactivated prophages. A previous study has shown that ARGs and VFGs are found in the genomes of temperate phages (which are thus not prophages) and that these genes are hardly found in the genomes of virulent phages [37]. This result would not be expected if all ARGs and VFGs were inserted into pre-existing inactivated prophages. More important, we do not have an *a priori* expectation that VFGs are more likely to be inserted into pre-existing prophages than ARGs. In the absence of such an expectation, our results are likely to be robust to post-hoc insertions of ARGs and VFGs.

That phages do not possess ARGs does not necessarily mean that phages do not mediate the horizontal transfer of ARGs because they can mediate HGT even if their genomes do not contain ARGs. Phages mediate HGT through three known mechanisms: specialised, generalised, and lateral transduction [24, 62, 63]. In specialised transduction (the focus of this study), a transferred gene constitutes a part of a phage genome [24]. By contrast, in generalised and lateral transduction, a transferred gene is originally encoded in bacterial DNA, which is encapsulated into phage particles and subsequently transferred to other cells [24, 62, 63]. Thus, phages can still mediate the horizontal transfer of ARGs even if their genomes do not contain ARGs.

In conclusion, the results presented above suggest that MGEs differ in the functional categories of accessory genes they carry depending on their strategies of infection.

7. Author statements

7.1 Conflicts of interest

The authors declare that there are no conflicts of interest.

7.2 Funding information

NT is supported by The University of Auckland Faculty Research Development Fund. HS is supported by Ohsumi Frontier Science Foundation.

7.3 Acknowledgements

The authors acknowledge the contribution of New Zealand eScience Infrastructure to the results of this research. The authors thank Bram van Dijk for discussion.

538

539 ## 8. Tables and figures

540 **Table 1**. Numbers of predicted prophages per genome (SD, standard deviation), numbers of
541 orthologous groups of antibiotic resistance genes (OGARGs), and those of virulence factor
542 genes (OGVFGs) in genomes (total), predicted plasmidic prophages, and predicted
543 chromosomal prophages. Numbers in brackets are for OGARGs in prophages containing at
544 least one phage-structure gene.

	prophage (SD)	OGARG			OGVFG		
		total	prophage		total	prophage	
			plasmidic	chromosomal		plasmidic	chromosomal
A. baumannii	4.3 (1.7)	455	0	4 (4)	48	0	0
B. anthracis	4.1 (0.6)	17	0	1 (1)	39	0	0
B. pertussis	1.9 (0.3)	0	0	0	95	0	0
B. melitensis	1.0 (0.3)	1	0	0	183	0	0
B. pseudomallei	2.2 (1.3)	3	0	0	192	0	0
C. jejuni	1.1 (1.2)	65	0	0	480	0	0
E. faecalis	3.8 (1.8)	179	2 (0)	0	114	1	0
E. coli	7.1 (3.8)	3041	9 (7)	9 (9)	18019	6	1590
H. influenzae	3.0 (1.4)	7	0	0	161	0	1
H. pylori	0.2 (0.5)	6	0	0	285	0	0
K. pneumoniae	4.5 (2.1)	1473	7 (2)	7 (7)	372	0	0
L. pneumophila	0.1 (0.3)	5	0	0	773	0	0
L. monocytogenes	2.5 (1.3)	5	0	0	172	0	0
M. tuberculosis	0 (0.1)	4	0	0	997	0	0
N. meningitidis	1.9 (0.6)	4	0	0	123	0	1
P. aeruginosa	4.5 (2.7)	543	0	0	526	0	2
S. enterica	3.9 (1.8)	826	2 (0)	0	1844	2	64
S. aureus	3.5 (1.3)	423	0	6 (2)	217	0	28
S. agalactiae	2.0 (1.0)	46	0	1 (1)	71	0	0
S. pyogenes	3.1 (1.5)	28	0	3 (3)	103	0	48
V. cholerae	1.5 (1.2)	101	1 (0)	2 (1)	251	1	5

545

546 **Table 2**. Numbers of plasmids per genome (SD, standard deviation), and numbers of
547 orthologous groups of antibiotic resistance genes (OGARGs) and those of virulence factor
548 genes (OGVFGs) in genomes (total) and plasmids.

	Plasmid (SD)	OGARG		OGVFG	
		total	plasmid	total	plasmid
A. baumannii	1.7 (1.6)	455	195	48	0
B. anthracis	1.5 (0.9)	17	1	39	14
B. pertussis	0.0 (0.0)	0	0	95	0
B. melitensis	0.0 (0.0)	1	0	183	0
B. pseudomallei	0.1 (0.3)	3	0	192	0
C. jejuni	0.3 (0.6)	65	21	480	9
E. faecalis	1.4 (1.4)	179	127	114	24
E. coli	2.0 (1.9)	3041	2397	18019	2452
H. influenzae	0.0 (0.1)	7	1	161	0
H. pylori	0.2 (0.4)	6	0	285	0
K. pneumoniae	3.4 (2.2)	1473	1274	372	96
L. pneumophila	0.3 (0.5)	5	1	773	0
L. monocytogenes	0.2 (0.5)	5	0	172	1
M. tuberculosis	0.0 (0.0)	4	0	997	0
N. meningitidis	0.0 (0.0)	4	0	123	0
P. aeruginosa	0.2 (0.6)	543	128	526	3
S. enterica	1.1 (1.2)	826	638	1844	46
S. aureus	0.8 (1.0)	423	182	217	9
S. agalactiae	0.0 (0.1)	46	1	71	0
S. pyogenes	0.0 (0.2)	28	3	103	0
V. cholerae	0.1 (0.4)	101	36	251	0

549

550 **Table 3**. Numbers of VFG symbols, orthologous groups of virulence factor genes (OGVFG),
551 and VFGs before clustering that are detected in plasmids, chromosomal prophages, and
552 chromosomes (which may include MGEs other than plasmids and prophages, such as
553 transposons) in *E. coli*. Columns under 'Chromosomes' do not include gene symbols,
554 OGVFGs, and VFGs that are found in neither plasmids nor chromosomal prophages. 'Unique
555 sum' refers to numbers of non-redundant items (some gene symbols are counted in
556 multiple rows owing to having multiple functions).

	Plasmids			Chromosomal prophages			Chromosomes		
Classification	symbols	OGVFGs	VFGs	symbols	OGVFGs	genes	symbols	OGVFGs	VFGs
Adherence	66	233	434	2	43	271	22	126	3553
Biofilm	1	12	35	0	0	0	0	0	0
Dispersin	4	15	75	0	0	0	0	0	0
Enzyme	11	158	409	3	9	156	11	331	7670
Exotoxin	15	236	1286	7	154	1020	7	37	255
Immune modulation	5	326	639	3	297	357	6	307	3075
Invasion	3	149	618	0	0	0	3	13	2881
Metal update	12	389	1523	0	0	0	12	775	2451
Motility	1	12	35	0	0	0	0	0	0
Phage-related	1	1	17	6	930	7183	6	209	390
Regulation	7	77	192	1	1	1	4	120	4318
T2SS	1	2	2	0	0	0	1	17	914
T3SS components	26	59	1111	0	0	0	0	0	0
T3SS effectors	33	659	1428	22	140	1969	21	474	2932
T4SS	2	4	87	0	0	0	0	0	0
T5SS	7	63	322	0	0	0	3	34	102
Uncategorised	0	0	0	3	6	8	3	1418	4557
Unique sum	186	2342	8020	47	1580	10965	97	3681	31705

557

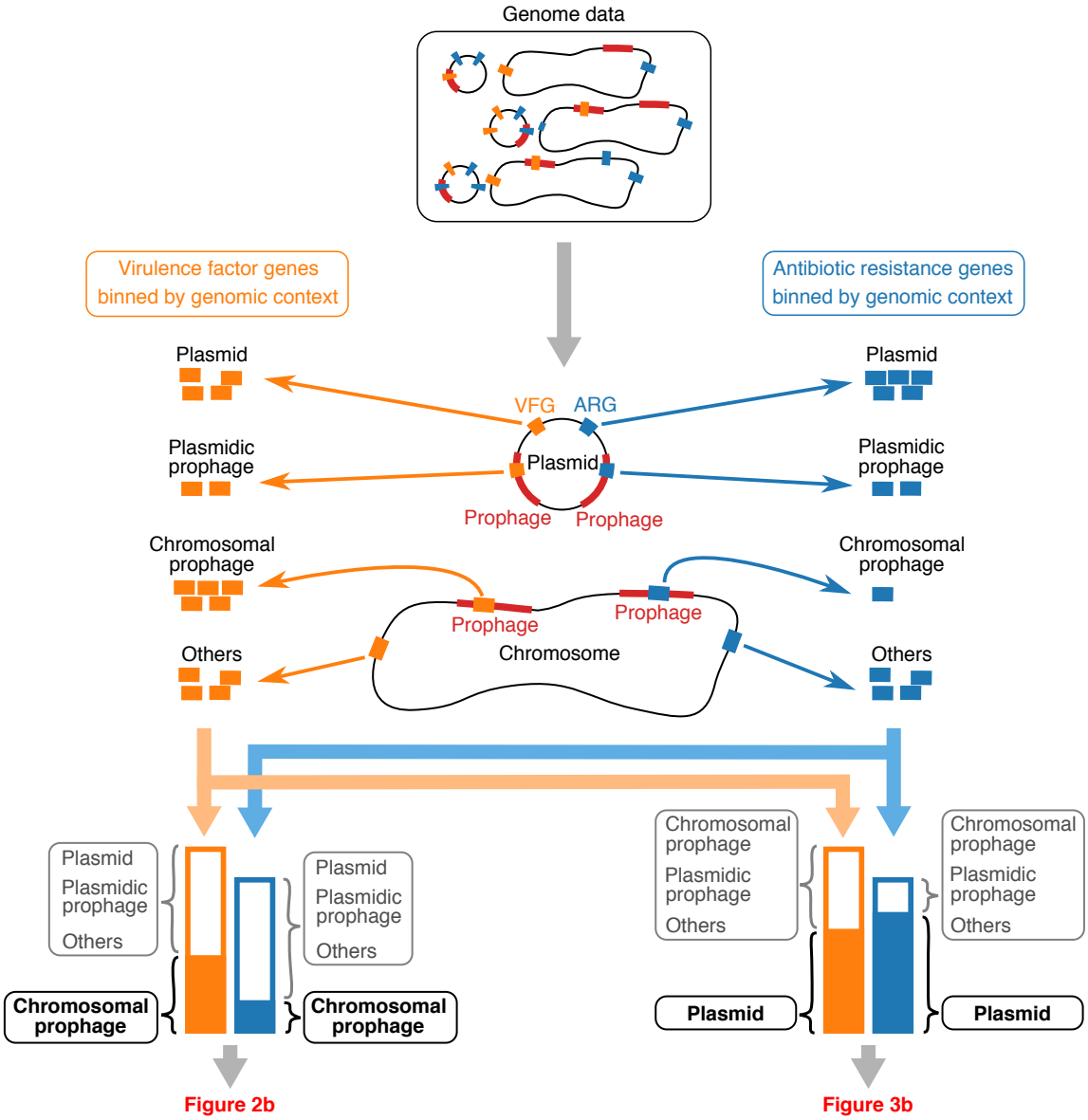

Figure 1. Schematic drawing of the method used in this study. For simplicity, clustering of orthologous genes is omitted from depiction (see Methods).

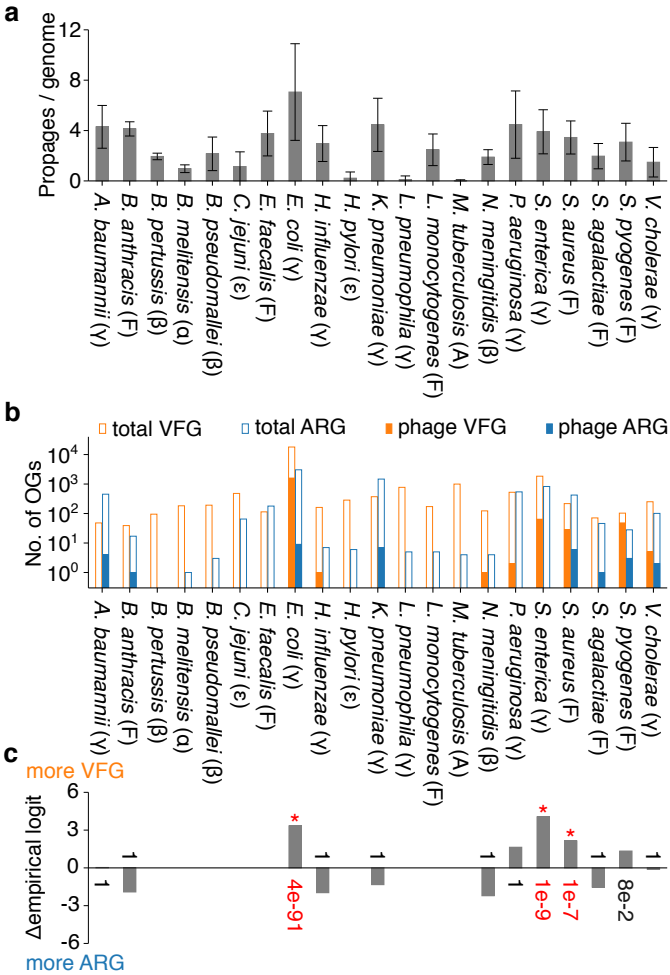

Figure 2. **a** Average number of prophages per genome and standard deviation. **b** Numbers of orthologous groups of virulence factor genes (OGVFGs) and antibiotic resistance genes (OGARGs) in prophages (filled bars) and genomes (open bars). Symbols in brackets indicate taxonomic groups: Actinobacteria (A), Firmicute (F), Alphaproteobacteria (α), Betaproteobacteria (β), Epsilonproteobacteria (ε), and Gammaproteobacteria (γ). **c** Degree of bias towards carrying VFGs more frequently than ARGs. Negative values indicate opposite bias (zero indicates no bias). Difference between empirical logits (denoted as Δempirical logit) is defined as $\log([a+0.5]/[b+0.5]) - \log([c+0.5]/[d+0.5])$, where a and b are numbers of OGVFGs and OGARGs in prophages, respectively (filled bars in b), and c and d are numbers of all OGVFGs and OGARGs in genomes, respectively (open bars in b). Numbers next to bars are P values of two-sided binomial test under null hypothesis that a and b are numbers drawn from binomial distribution with probabilities $c/(c+d)$ and $d/(c+d)$, respectively (corrected by Holm–Bonferroni method). Asterisks and red numbers indicate P values less than or equal to 0.05.

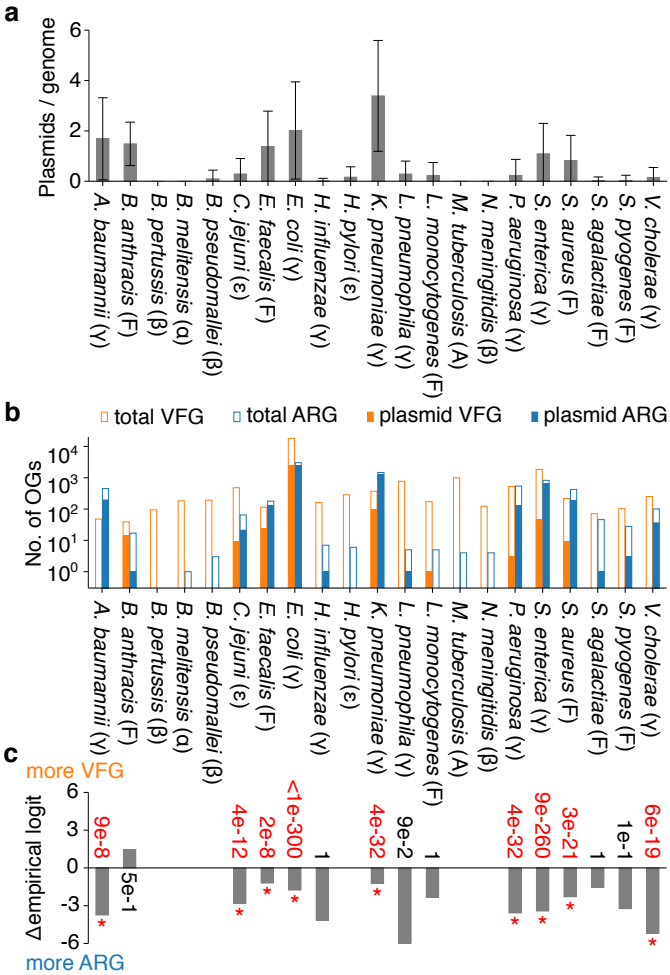

Figure 3. **a** Average number of plasmids per genome and standard deviation. **b** Numbers of orthologous groups of virulence factor genes (OGVFGs) and antibiotic resistance genes (OGARGs) in plasmids (filled bars) and genomes (open bars). Symbols in brackets indicate taxonomic groups: Actinobacteria (A), Firmicute (F), Alphaproteobacteria (α), Betaproteobacteria (β), Epsilonproteobacteria (ε), and Gammaproteobacteria (γ). **c** Degree of bias towards carrying VFGs more frequently than ARGs. Negative values indicate opposite bias (zero indicates no bias). Difference between empirical logits (denoted as Δempirical logit) is defined as $\log([a+0.5]/[b+0.5]) - \log([c+0.5]/[d+0.5])$, where a and b are numbers of OGVFGs and OGARGs in prophages, respectively (filled bars in b), and c and d are numbers of all OGVFGs and OGARGs in genomes, respectively (open bars in b). Numbers next to bars are P values of two-sided binomial test under null hypothesis that a and b are numbers drawn from binomial distribution with probabilities $c/(c+d)$ and $d/(c+d)$, respectively (corrected by Holm–Bonferroni method). Asterisks and red numbers indicate P values less than or equal to 0.05.

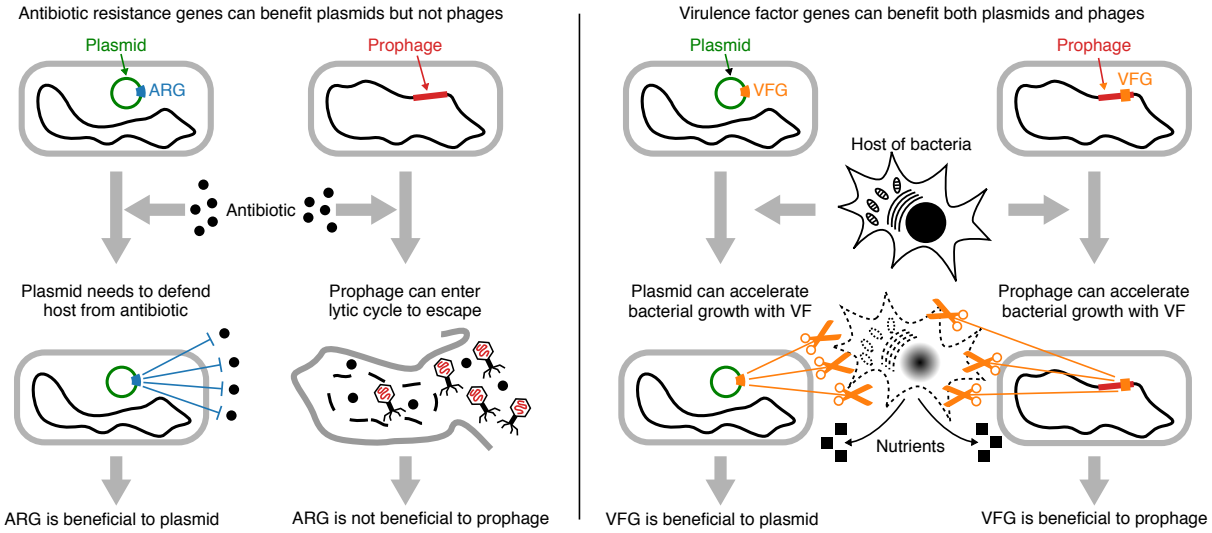

Figure 4. Schematic drawing of hypothesis formulated based on our results. Left: plasmids benefit from carrying ARG because they cannot abandon their hosts and thus need to minimise death of bacteria. In contrast, temperate phages do not benefit much from carrying ARGs because if their hosts are in danger, they can abandon their hosts and seek new hosts by entering lytic cycles. Right: both plasmids and phages benefit from VFGs because VFGs can enhance growth of bacteria by causing bacteria to exploit their hosts more aggressively.

9. Supplementary information

Supplementary files can be downloaded from https://doi.org/10.1101/2022.07.21.500938

Table S1: List of all genome assemblies used in this study.
Table S2: Information about the gene-symbol assignment for OGVFGs in *Escherichia coli*, consisting of four sheets: list of VFG database entries and their gene symbols as found in the three VFG databases; manual annotation of VFG database entries that lack gene symbols, plus virB and ipgB; finalised assignment of gene symbols to VFG database entries; conflict resolution for OGVFGs that are assigned multiple gene symbols.
Table S3: List of VFG symbols, the number of OGVFG, and their functional categories.
Table S4: List of all prophages predicted by VIBRANT.
Table S5: List of all orthologous groups of antibiotic resistance genes.
Table S6: Information about all prophages manually examined.
Table S7: List of all orthologous groups of virulence factor genes.
Table S8: List of prophages that are described in Kondo et al. (2021) and manually examined in this study.
In-house scripts

10. References

1. Bennett, P.M., *Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria.* British Journal of Pharmacology, 2009. **153**(S1): p. S347-S357.

2. von Wintersdorff, C.J.H., et al., *Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer.* Frontiers in Microbiology, 2016. **7**: p. 173.
3. Partridge, S.R., et al., *Mobile Genetic Elements Associated with Antimicrobial Resistance.* Clinical microbiology reviews, 2018. **31**(4): p. e00088-17.
4. Wagner, P.L. and M.K. Waldor, *Bacteriophage control of bacterial virulence.* Infection and Immunity, 2002. **70**(8): p. 3985-3993.
5. Gyles, C. and P. Boerlin, *Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease.* Veterinary Pathology, 2014. **51**(2): p. 328-340.
6. O'Neill, J., *Review on antimicrobial resistance. Tackling drug-resistant infections globally: Final report and recommendations*. 2016.
7. Pilla, G. and C.M. Tang, *Going around in circles: Virulence plasmids in enteric pathogens.* Nature Reviews Microbiology, 2018. **16**(8): p. 484-495.
8. Murray, C.J.L., et al., *Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis.* The Lancet, 2022. **399**(10325): p. 629-55.
9. Frost, L.S., et al., *Mobile genetic elements: the agents of open source evolution.* Nature Reviews Microbiology, 2005. **3**(9): p. 722-732.
10. Hall, J.P.J., M.A. Brockhurst, and E. Harrison, *Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria.* Philosophical Transactions of the Royal Society B: Biological Sciences, 2017. **372**(1735): p. 20160424.
11. Ghaly, T.M. and M.R. Gillings, *Mobile DNAs as ecologically and evolutionarily independent units of life.* Trends in Microbiology, 2018. **26**(11): p. 904-912.
12. Rankin, D.J., E.P.C. Rocha, and S.P. Brown, *What traits are carried on mobile genetic elements and why?* Heredity, 2011. **106**(1): p. 1-10.
13. Eberhard, W.G., *Evolution in bacterial plasmids and levels of selection.* Quarterly Review of Biology, 1990. **65**(1): p. 3-22.
14. Bergstrom, C.T., M. Lipsitch, and B.R. Levin, *Natural selection, infectious transfer and the existence conditions for bacterial plasmids.* Genetics, 2000. **155**(4): p. 1505-19.
15. Svara, F. and D.J. Rankin, *The evolution of plasmid-carried antibiotic resistance.* BMC Evolutionary Biology, 2011. **11**(1): p. 130-130.
16. Lehtinen, S., J.S. Huisman, and S. Bonhoeffer, *Evolutionary mechanisms that determine which bacterial genes are carried on plasmids.* Evolution Letters, 2021. **5**(3): p. 290-301.
17. Abedon, S.T. and J.T. LeJeune, *Why Bacteriophage Encode Exotoxins and other Virulence Factors.* Evolutionary Bioinformatics, 2005. **1**: p. 117693430500100-117693430500100.
18. Matsushiro, A., et al., *Induction of Prophages of Enterohemorrhagic Escherichia coli O157:H7 with Norfloxacin.* Journal of Bacteriology, 1999. **181**(7): p. 2257–2260-2257–2260.
19. Goerke, C., J. Köller, and C. Wolz, *Ciprofloxacin and Trimethoprim Cause Phage Induction and Virulence Modulation in Staphylococcus aureus.* Antimicrobial Agents and Chemotherapy, 2006. **50**(1): p. 171–177-171–177.
20. Allen, H.K., et al., *Antibiotics in feed induce prophages in swine fecal microbiomes.* mBio, 2011. **2**(6): p. e00260–11-e00260–11.
21. López, E., et al., *Induction of Prophages by Fluoroquinolones in Streptococcus pneumoniae: Implications for Emergence of Resistance in Genetically-Related Clones.* PLOS ONE, 2014. **9**(4): p. e94358-e94358.

677 22. Sutcliffe, S.G., et al., *Common Oral Medications Lead to Prophage Induction in*
678 *Bacterial Isolates from the Human Gut.* Viruses, 2021. **13**(3): p. 455-455.
679 23. Wallace, M.J., S.R.S. Fishbein, and G. Dantas, *Antimicrobial resistance in enteric*
680 *bacteria: current state and next-generation solutions.* Gut Microbes, 2020. **12**(1).
681 24. Snyder, L., et al., *Molecular Genetics of Bacteria*. 4th ed ed. 2013, Washington, D.C.:
682 ASM Press.
683 25. Villa, T.G., et al., *Horizontal Gene Transfer Between Bacteriophages and Bacteria:*
684 *Antibiotic Resistances and Toxin Production*. 2019, Springer International Publishing.
685 p. 97-142.
686 26. Giovanetti, E., et al., *Transduction of the Streptococcus pyogenes bacteriophage*
687 *Φm46.1, carrying resistance genes mef(A) and tet(O), to other Streptococcus species.*
688 Frontiers in Microbiology, 2015. **5**(DEC): p. 746.
689 27. Modi, S.R., et al., *Antibiotic treatment expands the resistance reservoir and*
690 *ecological network of the phage metagenome.* Nature, 2013. **499**(7457): p. 219-222.
691 28. Subirats, J., et al., *Metagenomic analysis reveals that bacteriophages are reservoirs of*
692 *antibiotic resistance genes.* International Journal of Antimicrobial Agents, 2016.
693 **48**(2): p. 163-167.
694 29. Debroas, D. and C. Siguret, *Viruses as key reservoirs of antibiotic resistance genes in*
695 *the environment.* The ISME Journal, 2019. **13**(11): p. 2856-2867.
696 30. Strange, J.E.S., et al., *Metagenomics analysis of bacteriophages and antimicrobial*
697 *resistance from global urban sewage.* Scientific Reports, 2021. **11**(1): p. 1600.
698 31. Muniesa, M., et al., *Bacteriophages and diffusion of β-lactamase genes.* Emerging
699 Infectious Diseases, 2004. **10**(6): p. 1134-1137.
700 32. Enault, F., et al., *Phages rarely encode antibiotic resistance genes: a cautionary tale*
701 *for virome analyses.* The ISME Journal, 2017. **11**(1): p. 237-247.
702 33. López-Leal, G., et al., *Prophages Encode Antibiotic Resistance Genes in*
703 *Acinetobacter baumannii.* Microbial Drug Resistance, 2020.
704 34. Kondo, K., M. Kawano, and M. Sugai, *Distribution of Antimicrobial Resistance and*
705 *Virulence Genes within the Prophage-Associated Regions in Nosocomial Pathogens.*
706 mSphere, 2021.
707 35. Kleinheinz, K.A., K.G. Joensen, and M.V. Larsen, *Applying the ResFinder and*
708 *VirulenceFinder web-services for easy identification of acquired antibiotic resistance*
709 *and E. coli virulence genes in bacteriophage and prophage nucleotide sequences.*
710 Bacteriophage, 2014. **4**(2): p. e27943-e27943.
711 36. Gunathilaka, G.U., et al., *Phages in urban wastewater have the potential to*
712 *disseminate antibiotic resistance.* International Journal of Antimicrobial Agents,
713 2017. **50**(5): p. 678-683.
714 37. Cook, R., et al., *INfrastructure for a PHAge REference Database: Identification of*
715 *Large-Scale Biases in the Current Collection of Cultured Phage Genomes.* PHAGE,
716 2021.
717 38. Liu, B., et al., *VFDB 2019: a comparative pathogenomic platform with an interactive*
718 *web interface.* Nucleic Acids Research, 2018. **47**(D1): p. D687-D692.
719 39. Sayers, S., et al., *Victors: a web-based knowledge base of virulence factors in human*
720 *and animal pathogens.* Nucleic Acids Research, 2019. **47**(D1): p. D693-D700.
721 40. Davis, J.J., et al., *The PATRIC Bioinformatics Resource Center: expanding data and*
722 *analysis capabilities.* Nucleic Acids Research, 2020. **48**(D1): p. D606-D612.
723 41. Sayers, E.W., et al., *GenBank.* Nucleic Acids Research, 2020. **48**(D1): p. D84-D86.
724 42. Fu, L., et al., *CD-HIT: Accelerated for clustering the next-generation sequencing*
725 *data.* Bioinformatics, 2012. **28**(23): p. 3150-3152.

43. Li, W., et al., *RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation.* Nucleic Acids Research, 2021. **49**(D1): p. D1020-D1028.
44. Roach, M.J., et al., *Philympics 2021: Prophage Predictions Perplex Programs.* F1000Research, 2021. **10**: p. 758-758.
45. Feldgarden, M., et al., *AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence.* Scientific Reports, 2021. **11**(1): p. 12728-12728.
46. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
47. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature Methods, 2015. **12**(1): p. 59-60.
48. Lechner, M., et al., *Proteinortho: Detection of (Co-)orthologs in large-scale analysis.* BMC Bioinformatics, 2011. **12**(1): p. 124-124.
49. Novichkov, P.S., et al., *ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes.* Nucleic Acids Research, 2008. **37**(suppl_1): p. D448-D454.
50. Pfeifer, E., et al., *Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires.* Nucleic Acids Research, 2021.
51. Lu, S., et al., *CDD/SPARCLE: the conserved domain database in 2020.* Nucleic Acids Research, 2020. **48**(D1): p. D265–D268-D265–D268.
52. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Research, 2016. **44**(D1): p. D733–D745-D733–D745.
53. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2021. **49**(D1): p. D480–D489-D480–D489.
54. Kieft, K., Z. Zhou, and K. Anantharaman, *VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences.* Microbiome, 2020. **8**(1): p. 90-90.
55. Fouts, D.E., *Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences.* Nucleic Acids Research, 2006. **34**(20): p. 5839–5851-5839–5851.
56. Holm, S., *A Simple Sequentially Rejective Multiple Test Procedure.* Scandinavian Journal of Statistics, 1979. **6**(2): p. 65–70-65–70.
57. Seabold, S. and J. Perktold. *statsmodels: Econometric and statistical modeling with Python*. in *9th Python in Science Conference*.
58. Shenoy, A.R., et al., *Modulation of Host Cell Processes by T3SS Effectors*, G. Frankel and E.Z. Ron, Editors. 2018, Springer International Publishing. p. 73–115-73–115.
59. van Diemen, P.M., et al., *Identification of Enterohemorrhagic Escherichia coli O26:H− Genes Required for Intestinal Colonization in Calves.* Infection and Immunity, 2005. **73**(3): p. 1735–1743-1735–1743.
60. Dziva, F., et al., *Identification of Escherichia coli O157 : H7 genes influencing colonisation of the bovine gastrointestinal tract using signature-tagged mutagenesis.* Microbiology, 2004. **150**(11): p. 3631–3645-3631–3645.
61. Arndt, D., et al., *PHASTER: a better, faster version of the PHAST phage search tool.* Nucleic Acids Research, 2016. **44**(W1): p. W16-W21.
62. Chen, J., et al., *Genome hypermobility by lateral transduction.* Science, 2018. **362**(6411): p. 207-212.

775 63. Chiang, Y.N., J.R. Penadés, and J. Chen, *Genetic transduction by phages and*
776 *chromosomal islands: The new and noncanonical.* PLOS Pathogens, 2019. **15**(8): p.
777 e1007878-e1007878.
778