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
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

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 that are not essential to MGEs. What evolutionary rules govern the arsenal of accessory genes MGEs carry? Such rules, if they exist, might reflect the different infection strategies of MGEs, such as whether MGEs lyse host cells when they spread to other cells. Here we investigate whether different MGEs carry different types of accessory genes by comprehensively analysing public databases. Specifically, we compare prophages and plasmids with respect to the frequencies at which they carry antibiotic resistance genes (ARGs) and virulence factor genes (VFGs) in 21 pathogenic species of bacteria spanning three phyla (Actinobacteria, Firmicute, and Proteobacteria). The results indicate that prophages and plasmids display opposite trends: prophages are biased towards carrying VFGs more frequently than ARGs in two species of Gammaproteobacteria (*Escherichia coli* and *Salmonella enterica*) and one species of Firmicute (*Staphylococcus aureus*), whereas plasmids are biased towards carrying ARGs more frequently than VFGs in

six species of Gammaproteobacteria (*Acinetobacter baumannii*, *E. coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *S. enterica*, and *Vibrio cholerae*), one species of Epsilonproteobacteria (*Campylobacter jejuni*), and two species of Firmicutes (*Enterococcus faecalis* and *S. aureus*). In the other species, prophage-borne or plasmid-borne ARGs and VFGs are barely detected. Taken together, our results suggest that MGEs differentiate in terms of the types of accessory gene they carry depending on their strategies of infection, providing an insight into the rules of evolution 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 apparently integral to the replication and transmission of MGEs. For example, plasmids frequently possess various antibiotic resistance genes (ARGs) [1-3]. Phages and plasmids also possess virulence factor genes (VFGs), the genes required for bacterial pathogenicity [4-8]. Mediating the horizontal transfer of accessory genes between bacteria, MGEs play important roles for 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 in order to transmit 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. To understand the rules governing HGT mediated by MGEs, it is thus beneficial to investigate whether different MGEs carry different types of accessory genes.

To address the above question, we focus on 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 [18]. Phages mediate HGT through multiple mechanisms, among which specialised transduction is the most similar to HGT mediated by plasmid [19]. In specialised transduction, phages transfer genes carried in their genomes. Therefore, specialised transduction is strictly coupled with the infectious transmission of phages, the coupling that is also entailed in plasmid conjugation [19]. Laboratory experiments have demonstrated that phages are capable of transferring ARGs to bacteria through specialised transduction [20]. However, the specialised transduction of ARGs in nature has been

scarcely documented [3, 21]. While metagenomic studies have detected ARGs in viral fractions of environmental DNA samples [22-25], other studies provide evidence suggesting that the detection of ARGs was due to the contamination of bacterial DNA in the viral fractions [26, 27]. 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* [28], *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* [29] [see also [30]]. Also, a previous study has isolated 29 phages from wastewater, of which 15 carry ARGs, suggesting that phages frequently possess ARGs [31]. 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 [32]. 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 [28-30]. 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 overestimation in the number of prophage-borne ARGs. Contrariwise, a true prophage can be missed, which leads to underestimation in 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 depend on the types of gene. 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). Finally, to investigate a potential differentiation between prophages and plasmids, we also analyse the distributions of ARGs and VFGs in plasmids. The 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.

4. Methods

4.1 Data acquisition

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

[33-35]. The VFGs in Victors were refined by removing those carried by non-bacterial pathogens or lacking NCBI protein GIs (4575 genes remained). Some VFGs in Victors were missing protein sequences, which were downloaded from Genbank based on their protein GIs [36]. All VFGs were pooled and clustered to remove redundancy with CD-HIT with the protein sequence identity threshold of 1.0 [37], resulting in a combined database of 7218 VFGs.

The genome assemblies with the 'Complete' status were downloaded from RefSeq in September 2021 [38]. The following criteria were used to select species: a species had at least 60 complete genomes in RefSeq and at least 70 VFGs in the combined VFG database. These criteria resulted in the following 21 species of bacterial pathogens (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), *Escherichia coli* (1444), *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). These species span three phyla: Actinobacteria, Firmicute, and Proteobacteria. The complete list of the bacterial genomes analysed in this study is shown in Table S1.

4.2 Prophage prediction

Prophages were predicted with VIBRANT (version 1.2.1), which was selected for three reasons. First, VIBRANT has comparatively high performance as reported by a recent benchmark [39]. Second, it is a standalone tool, which can be run on local computers. Third, its algorithm is based on the similarity search of known phage proteins, which is suitable for the present study because our aim was to analyse known prophages rather than discover novel prophages. VIBRANT was run against the genomic nucleotide files (*_genomic.fna) with default parameters.

4.3 Plasmids

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

ARGs were predicted with AMRFinderPlus version 3.10.5 and the core subset of the database version 2021-09-11.1 with the organism option if possible [40]. AMRFinderPlus was run against translated coding sequences. The genes predicted by AMRFinderPlus as ARGs (i.e., "element subtype AMR") and not annotated as pseudo-genes in RefSeq were considered as ARGs (the genes predicted as "element subtype POINT", which are genes containing known point mutation associated with AR, were excluded).

169

170 ## 4.5 VFG prediction

171 Every VFG in the combined VFG database was queried against every bacterial genome with
172 BLASTP with E-value threshold of 1e-9 [41]. A gene in a bacterial genome (bacterial gene, for
173 short) could match multiple VFGs in the combined VFG database, in which case the VFG with
174 the highest bit-score was selected as the best match. A bacterial gene was considered as
175 encoding VF if it met the following additional criteria: (i) it was not annotated as a pseudo-
176 gene in RefSeq [38]; (ii) the BLASTP alignment between the bacterial gene and its best
177 match VFG, if any, had at least 80% sequence identity and covers at least 80% of both the
178 bacterial gene and the best match VFG; (iii) the species of the genome in which the bacterial
179 gene resides was identical to the species in which the best match VFG resides [33-35].
180

181 ## 4.6 Orthology prediction

182 Genes annotated in RefSeq were clustered into putative orthologous groups based on
183 protein sequence similarity and synteny conservation, as follows. First, preliminary
184 orthologous pairs of genes were identified between every pair of genomes within each
185 species through all-against-all sequence similarity searches using ProteinOrtho version
186 6.0.25 (with DIAMOND ver. 2.0.6 [42]; E-value cut-off of 1e-5; minimum coverage of best
187 alignments of 75%; minimum per-cent identity of best alignments of 25%; minimum
188 reciprocal similarity of 0.95) [43]. ProteinOrtho defines a preliminary orthologous pair of
189 genes as a reciprocal nearly-best hit (RNBH), as follows. A nearly-best hit (NBH) of a gene
190 queried against a target genome is defined as a hit (alignment) whose bit-score is not
191 smaller by a factor f than that of the best hit (e.g., if f=1, NBHs would include only the best
192 hit). The value of f used in this study was 0.95, which is the default value of ProteinOrtho. If
193 two genes are reciprocally (i.e., mutually) NBH of each other, they form RNBH [43].
194

195 RNBHs obtained with ProteinOrtho were pruned based on synteny conservation with an in-
196 house script, as follows. Let x and y be a pair of genes forming RNBH, and let X and Y be the
197 genomic neighbours of x and y, respectively, where the genomic neighbour of a gene
198 (denoted by z) is defined as a set of 21 genes consisting of ten genes upstream of z, ten
199 genes downstream of z, and z itself (all contigs were assumed to be circular, and the
200 orientation of genes were ignored). Let N_x and N_y be the number of genes in X and Y that
201 form RNBHs with at least one gene in Y and X, respectively (note that a single gene in one
202 genome can form RNBHs with multiple genes in another genome owing to tandem
203 duplication). If both N_x and N_y are greater than ten (i.e., a majority of the genes in X form
204 RNBHs with the genes in Y, and *vice versa*), the RNBH formed by x and y was kept;
205 otherwise, it was discarded [44].
206

207 Finally, the pruned RNBHs were clustered into putative orthologous groups with the spectral
208 clustering algorithm implemented in ProteinOrtho version 6.0.25 (minimum algebraic
209 connectivity of 0.1; exact step 3; minimum number of species of 0; purity of 1e-7) [43].
210

4.7 Classification of orthologous gene groups

A gene (VFG or ARG) was considered as encoded in a prophage if the entire gene is included within a genomic region predicted as a prophage.

An orthologous group of genes was considered as encoded in either a prophage residing in a chromosome (chromosomal plasmid, for short), a prophage residing in a plasmid (plasmidic plasmid), or a plasmid if the majority of the genes belonging to the group were in chromosomal prophages, plasmidic prophages, or plasmids, respectively (the cases of ties were ignored). Orthologous groups in plasmidic prophages were classified separately from those in chromosomal prophages or plasmids, for two reasons. First, it was ambiguous whether they should be regarded as encoded by plasmids, prophages, or both. Second, plasmidic prophages potentially represent a distinct class of mobile genetic elements called phage-plasmids [45].

An orthologous group of genes was considered as 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.

5. Results

5.1 Prophage prediction: VIBRANT is more precise than PhiSpy

To examine the distribution of ARGs in prophages, we computationally predicted prophages using VIBRANT [46] and ARGs using AMRFinderPlus [40] in the genomes of 21 pathogenic bacterial species downloaded from the RefSeq database [38] (Methods; Tables S1 and S2). To avoid double-counting orthologous ARGs in different genomes, we clustered all genes into putative orthologous groups based on sequence similarity and synteny conservation (Methods). We then counted the number of orthologous groups of ARGs (OGARGs) encoded in the predicted prophages, distinguishing between prophages residing in bacterial chromosomes and prophages residing in plasmids (chromosomal prophage and plasmidic prophage, respectively, for short). This distinction was made because it was ambiguous whether ARGs in plasmidic prophages should be regarded as encoded in prophages, plasmids, or both [plasmidic prophages potentially represent phage-plasmids, which are a separate class of MGEs from typical phages and plasmids [45]]. The result shows that a few to several prophage-borne OGARGs were detected in ten of the 21 examined species (Tables 1 and S3).

To probe the precision of the above prediction, we manually examined whether predicted prophages carrying ARG contained at least one phage-structure gene (e.g., phage baseplate, capsid, portal, tail, tail fibre, tail sheath, tail assembly, head-tail connector, and tail tape measure) according to the RefSeq annotation. Although this criterion cannot perfectly distinguish true prophages from false ones, it allows us to split the predicted prophages into those enriched with true prophages and those enriched with false prophages, allowing us to

255 probe the precision of the prophage prediction. The result of the examination shows that 24
256 out of 29 (i.e., 83% of) chromosomal prophages carrying ARGs contained phage-structure
257 genes, whereas seven out of 18 (i.e., 39% of) plasmidic prophages carrying ARGs contained
258 phage-structure genes (Table S4). This result means that 28 out of 33 (i.e., 85% of) OGARGs
259 in chromosomal prophages are in prophages carrying phage-structure genes, whereas nine
260 out of 21 (i.e., 43% of) OGARGs in plasmidic prophages are in prophages carrying phage-
261 structure genes (Table 1; note that the number of prophages carrying ARGs is smaller than
262 that of prophage-borne ARGs because one prophage can carry multiple ARGs). This result
263 suggests that the precision of VIBRANT is acceptable for chromosomal prophages.
264
265 During the manual examination, we noticed that a subset of plasmidic prophages carrying
266 ARGs also carried genes encoding integron integrases according to the RefSeq annotation
267 (Tables S4). Moreover, these prophages have a higher frequency of lacking phage-structure
268 genes than those without integron integrase genes (IIGs, for short). Specifically, seven out of
269 ten (i.e., 70% of) prophages carrying both ARGs and IIGs lack phage-structure genes,
270 whereas nine out of 37 (i.e., 24% of) prophages carrying ARGs and lacking IIGs lack phage-
271 structure genes. VIBRANT predicted that the genes annotated by RefSeq as IIGs matched
272 phage integrase genes, which are typical components of phage genomes. However, the
273 RefSeq annotation implies that proteins encoded by these genes are more similar to
274 integron integrases than phage integrases because the RefSeq annotation considers a
275 broader set of protein families than does VIBRANT [38, 46]. Therefore, the above result
276 suggests that about a half of the prophages carrying ARGs and lacking phage-structure
277 genes might have arisen from the misidentification of integrons as prophages [see also
278 [47]].
279

280 ## 5.2 Prophages carry significantly more VFGs than ARGs in multiple species

281
282 To compare the frequencies of ARGs to those of VFGs in chromosomal prophages, we used
283 BLASTP [41] to search bacterial genomes for VFGs collected from VFDB [33], Victors [34] and
284 PATRIC_VF [35] (Methods). We counted the number of orthologous groups of VFGs
285 (OGVFGs, for short) encoded in the predicted prophages (Tables 1 and S5). The number of
286 OGVFGs cannot directly be compared to that of OGARGs because we have no *a priori* reason
287 to expect that bacteria possess an equal number of VFGs and ARGs. Thus, we instead
288 compared the relative frequencies of OGVFGs and OGARGs against the genomic
289 background. Specifically, we performed binomial tests under the null hypothesis that the
290 relative frequencies of OGARGs and OGVFGs in the chromosomal prophages of each species
291 are the same as those of all OGARGs and OGVFGs in the genomes of the respective species.
292 In this test, we included prophages lacking phage-structure genes for fairness because
293 prophages carrying VFGs were too numerous to be manually examined (however, we found
294 that none of the prophages carrying VFGs carried IIG, which suggests that the prophages
295 carrying VFGs do not contain false positives arising from integron misidentification). Also,
296 we corrected P values using the Holm–Bonferroni method to control the family-wise error
297 rate of all the statistical tests conducted in this study [48, 49]. The results of the tests
298 indicate that the relative frequencies of OGARGs and OGVFGs in chromosomal prophages
299 are significantly different from those of respective genomic backgrounds in the following
300 three species (Figure 1): *E. coli* (Gammaproteobacteria), *S. enterica* (Gammaproteobacteria),

301 and *S. aureus* (Firmicute). In all these species, chromosomal prophages carry VFGs more
302 frequently than ARGs. The remaining 18 species, where significant biases were not
303 detected, can be grouped into three categories: the one species where a large number of
304 prophage-borne VFGs were detected, but a bias was not significant (*S. pyogenes*), the eight
305 species where only a small number of prophage-borne ARGs or VFGs were detected (*A.*
306 *baumannii*, *B. anthracis*, *H. influenzae*, *K. pneumoniae*, *N. meningitidis*, *P. aeruginosa*, *S.*
307 *agalactiae*, and *V. cholerae*), the seven species where no prophage-borne ARG and VFG
308 were detected (*B. pertussis*, *B. melitensis*, *B. pseudomallei*, *C. jejuni*, *E. faecalis*, *H. pylori*, *L.*
309 *pneumophila*), and the two species where prophages were hardly detected (*L.*
310 *monocytogenes*, and *M. tuberculosis*). The paucity or absence of prophage-borne ARGs and
311 VFGs in the second and third categories could be due to the limited sensitivity of the
312 prophage prediction tool. However, a large number of prophages were predicted in these
313 species. Therefore, if the prophage prediction tool missed prophages carrying ARGs or VFGs
314 in these species, those prophages are likely to be distinct from the currently known phages.
315 Taken together, the above results suggest that phages tend to carry ARGs less frequently
316 than VFGs if they carry a sufficient number of ARGs or VFGs; however, in many bacterial
317 species, they carry little or no ARGs and VFGs.
318

319 ### 5.3 Plasmids carry more ARGs than VFGs in many of species

320 To compare prophages to plasmids, we counted the number of OGARGs and OGVFGs
321 encoded in plasmids (Table 2). We then performed binomial tests under the null hypothesis
322 that the relative frequencies of OGARGs and OGVFGs in the plasmids of each species are the
323 same as those of all OGARGs and OGVFGs in the genomes of the respective species. The
324 results of the tests indicate that the relative frequencies of OGARGs and OGVFGs in plasmids
325 are significantly different from those of the genomic background in nine out of the 21
326 examined species (Figure 2). In all these species, plasmids are biased towards carrying ARGs
327 more frequently than VFGs. These species include six species of Gammaproteobacteria (*A.*
328 *baumannii*, *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *S. enterica*, and *V. cholerae*), one species
329 of Epsilonproteobacteria (*C. jejuni*), and two species of Firmicutes (*E. faecalis* and *S. aureus*).
330 The remaining 12 species, where significant biases were not detected, can be grouped into
331 three categories: the four species where plasmids are hardly present (*B. pertussis*, *B.*
332 *melitensis*, *M. tuberculosis*, and *N. meningitidis*), the seven species where plasmids are
333 present, but plasmid-borne ARGs and VFGs were barely detected (*B. pseudomallei*, *H.*
334 *influenzae*, *H. pylori*, *L. pneumophila*, *L. monocytogenes*, *S. agalactiae*, and S. pyogenes), and
335 the one species where plasmids carry more VFGs than ARGs, but a bias was not significant
336 (*B. anthracis*). Taken together, the above results suggest that plasmids tend to carry ARGs
337 more frequently than VFGs if they carry a sufficient number of VFGs or ARGs, with a possible
338 exception of *B. anthracis*; however, in several species, plasmids barely carry VFGs and ARGs.
339

340 ## 6. Discussion

341 The results presented above indicate that prophages tend to carry VFGs more frequently
342 than ARGs in three of the 21 examined species. In most of the other species, prophage-
343 borne ARGs and VFGs were hardly detected. In contrast, plasmids carry ARGs more
344 frequently than VFGs in nine of the 21 examined species. In most of the other species,

345 plasmid-borne ARGs and VFGs were barely detected. Taken together, these results indicate
346 that prophages and plasmids display opposite trends if they carry a sufficient number of
347 VFGs or ARGs: prophages are biased towards carrying VFGs, whereas plasmids toward
348 carrying ARGs. This difference between phages and plasmids led us to formulate the
349 following hypothesis to test for future: Temperate phages do not gain much benefit from
350 carrying genes that defend their hosts against antibiotics because if their hosts are in
351 danger, they can abandon their hosts and seek new ones.
352
353 The results described above, however, do not mean that plasmids rarely carry VFGs or that
354 prophages always carry VFGs. For example, in *K. pneumoniae*, plasmids carry ARGs more
355 frequently than VFGs; nevertheless, plasmids carry a large number of VFGs (viz., 96 out of
356 372 orthologous groups), whereas prophages carry none (Tables 1 and 2). A similar, yet less
357 striking, pattern is seen in *E. faecalis*. Given that many prophages are predicted in these
358 species (Table 1), the absence of prophage-borne VFGs is unlikely to be merely due to the
359 limited sensitivity of the prophage detection tool. By contrast, in *S. aureus* and *S. pyogenes*,
360 prophages carry greater numbers of VFGs than those of plasmids (Tables 1 and 2). This
361 heterogeneity in the distribution of VFGs suggests that plasmids and prophages play
362 variable roles in the pathogenicity of different bacterial species.
363
364 We should add that the absolute numbers of prophage-borne ARGs reported in this study
365 need to be interpreted with caution because of the limited sensitivity and precision of the
366 prophage prediction tool. In particular, these numbers do not necessary indicate that
367 prophages are devoid of ARGs in many bacterial species since prophages carrying ARGs
368 could have been missed. However, the conclusion of this study does not directly depend on
369 the absolute numbers of prophage-borne ARGs because it is based on comparison between
370 the relative frequencies of prophage-borne ARGs and VFGs.
371
372 Regarding the limitations of prophage prediction tools, it is pertinent to discuss a
373 discrepancy between our result and the result of Kondo et al. (2021) with respect to *P.*
374 *aeruginosa* [29]. While we found no prophage-borne ARGs in *P. aeruginosa*, Kondo et al.
375 (2021) reports that more than 10% of *P. aeruginosa* genomes possess prophage-borne ARGs
376 [29]. The important difference between Kondo et al. (2021) and our study is that they used
377 different tools for prophage prediction: Kondo et al. (2021) uses PHASTER [50], whereas we
378 used VIBRANT [46]. To investigate the cause of the above discrepancy, we manually
379 examined the 11 prophages described in Kondo et al. (2021) that carry ARGs and are
380 predicted as "intact" by PHASTER in *P. aeruginosa* [29]. We found that the examined
381 prophages could be grouped into two categories (Table S6). In the first group (five
382 prophages), both VIBRANT and PHASTER predicted prophages in almost the same genomic
383 locations. However, PHASTER predicts longer genomic regions including ARGs as prophages,
384 whereas VIBRANT predicted shorter regions excluding ARGs. We do not know which
385 prophage boundaries are more accurate. In the second group (6 prophages), prophages
386 were predicted only by PHASTER. These prophages, however, contained no phage-structure
387 genes. Although they contained phage-related genes, such as integrase, transposase, and
388 protease, these genes are not exclusively associated with phages. Moreover, PHASTER
389 annotated tellurium resistance proteins, TerD, as virion structural proteins in three
390 prophages in the second group, which is likely to be erroneous. 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-exiting 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 [32]. 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 because they are based on comparison between the relative frequencies 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 [19, 51, 52]. In specialised transduction (the focus of this study), a transferred gene constitutes a part of a phage genome [19]. 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 [19, 51, 52]. Thus, phages can mediate the horizontal transfer of ARGs even if their genomes do not contain ARGs.

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

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 NeSI to the results of this research.

434

435 ## 8. Tables and figures

436 **Table 1**. Numbers of predicted prophages per genome (SD, standard deviation), and number
437 of orthologous groups of antibiotic resistance genes (OGARGs) predicted by AMRFinderPlus
438 in genomes (total), predicted plasmidic prophages, and predicted chromosomal prophages.
439 Numbers in brackets are for OGARGs in prophages containing at least one phage-structure
440 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

441

442 **Table 2**. Numbers of plasmids per genome (SD, standard deviation), and numbers of
443 orthologous groups of virulence factor genes (OGVFGs) and those of antibiotic resistance
444 genes (OGARGs) 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

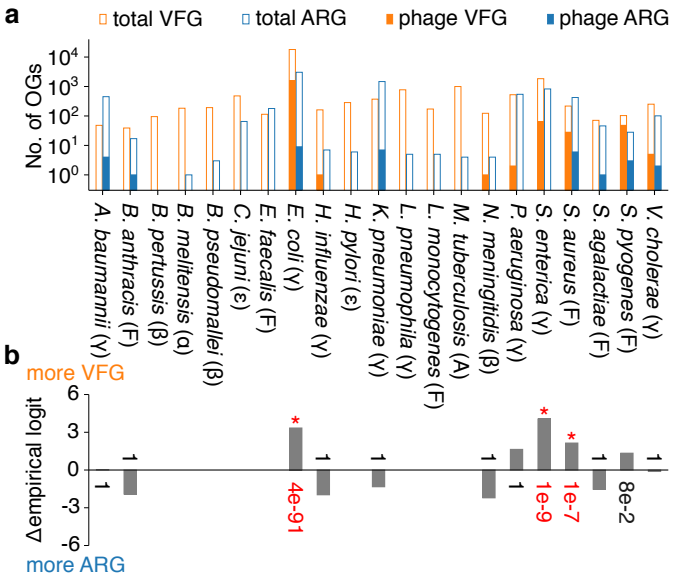

Figure 1. **a** Numbers of orthologous groups of virulence factor genes (OGVFGs) and antibiotic resistance genes (OGARGs) in prophages (filled bars) and genomes (open bars). Symbols in brackets indicate taxonomic groups: Actinobacteria (A), Firmicute (F), Alphaproteobacteria (α), Betaproteobacteria (β), Epsilonproteobacteria (ε), and Gammaproteobacteria (γ). **b** 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 a), and c and d are numbers of all OGVFGs and OGARGs in genomes, respectively (open bars in a). 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. Asterisks and red numbers indicate that P values are less than or equal to 0.05 (corrected by the Holm–Bonferroni method).

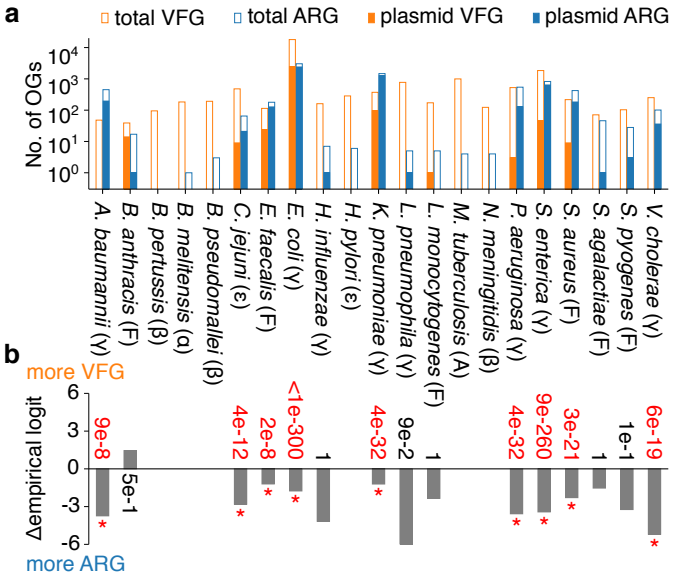

Figure 2. (a) **a** Numbers of orthologous groups of virulence factor genes (OGVFGs) and antibiotic resistance genes (OGARGs) in plasmids (filled bars) and genomes (open bars). Symbols in brackets indicate phyla: Actinobacteria (A), Firmicute (F), Alphaproteobacteria (α), Betaproteobacteria (β), Epsilonproteobacteria (ε), and Gammaproteobacteria (γ). **b** 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 a), and c and d are numbers of all OGVFGs and OGARGs in genomes, respectively (open bars in a). 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. Asterisks and red numbers indicate that P values are less than or equal to 0.05 (corrected by the Holm–Bonferroni method).

9. Supplementary information

Table S1: List of all genome assemblies used in this study.
Table S2: List of all prophages predicted by VIBRANT.
Table S3: List of all orthologous groups of antibiotic resistance genes.
Table S4: Information about all prophages manually examined.
Table S5: List of all orthologous groups of virulence factor genes.
Table S6: 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. Wallace, M.J., S.R.S. Fishbein, and G. Dantas, *Antimicrobial resistance in enteric bacteria: current state and next-generation solutions.* Gut Microbes, 2020. **12**(1).
19. Snyder, L., et al., *Molecular Genetics of Bacteria*. 4th ed ed. 2013, Washington, D.C.: ASM Press.
20. Villa, T.G., et al., *Horizontal Gene Transfer Between Bacteriophages and Bacteria: Antibiotic Resistances and Toxin Production*. 2019, Springer International Publishing. p. 97-142.
21. Giovanetti, E., et al., *Transduction of the Streptococcus pyogenes bacteriophage Φm46.1, carrying resistance genes mef(A) and tet(O), to other Streptococcus species.* Frontiers in Microbiology, 2015. **5**(DEC): p. 746.
22. Modi, S.R., et al., *Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome.* Nature, 2013. **499**(7457): p. 219-222.

23. Subirats, J., et al., *Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes.* International Journal of Antimicrobial Agents, 2016. **48**(2): p. 163-167.
24. Debroas, D. and C. Siguret, *Viruses as key reservoirs of antibiotic resistance genes in the environment.* The ISME Journal, 2019. **13**(11): p. 2856-2867.
25. Strange, J.E.S., et al., *Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage.* Scientific Reports, 2021. **11**(1): p. 1600.
26. Muniesa, M., et al., *Bacteriophages and diffusion of β-lactamase genes.* Emerging Infectious Diseases, 2004. **10**(6): p. 1134-1137.
27. Enault, F., et al., *Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses.* The ISME Journal, 2017. **11**(1): p. 237-247.
28. López-Leal, G., et al., *Prophages Encode Antibiotic Resistance Genes in Acinetobacter baumannii.* Microbial Drug Resistance, 2020.
29. Kondo, K., M. Kawano, and M. Sugai, *Distribution of Antimicrobial Resistance and Virulence Genes within the Prophage-Associated Regions in Nosocomial Pathogens.* mSphere, 2021.
30. Kleinheinz, K.A., K.G. Joensen, and M.V. Larsen, *Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences.* Bacteriophage, 2014. **4**(2): p. e27943-e27943.
31. Gunathilaka, G.U., et al., *Phages in urban wastewater have the potential to disseminate antibiotic resistance.* International Journal of Antimicrobial Agents, 2017. **50**(5): p. 678-683.
32. Cook, R., et al., *INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes.* PHAGE, 2021.
33. Liu, B., et al., *VFDB 2019: a comparative pathogenomic platform with an interactive web interface.* Nucleic Acids Research, 2018. **47**(D1): p. D687-D692.
34. Sayers, S., et al., *Victors: a web-based knowledge base of virulence factors in human and animal pathogens.* Nucleic Acids Research, 2019. **47**(D1): p. D693-D700.
35. Davis, J.J., et al., *The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities.* Nucleic Acids Research, 2020. **48**(D1): p. D606-D612.
36. Sayers, E.W., et al., *GenBank.* Nucleic Acids Research, 2020. **48**(D1): p. D84-D86.
37. Fu, L., et al., *CD-HIT: Accelerated for clustering the next-generation sequencing data.* Bioinformatics, 2012. **28**(23): p. 3150-3152.
38. 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.
39. Roach, M.J., et al., *Philympics 2021: Prophage Predictions Perplex Programs.* F1000Research, 2021. **10**: p. 758-758.
40. 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.
41. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
42. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature Methods, 2015. **12**(1): p. 59-60.
43. Lechner, M., et al., *Proteinortho: Detection of (Co-)orthologs in large-scale analysis.* BMC Bioinformatics, 2011. **12**(1): p. 124-124.

44. 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.
45. Pfeifer, E., et al., *Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires.* Nucleic Acids Research, 2021.
46. 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.
47. 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.
48. Holm, S., *A Simple Sequentially Rejective Multiple Test Procedure.* Scandinavian Journal of Statistics, 1979. **6**(2): p. 65–70-65–70.
49. Seabold, S. and J. Perktold. *statsmodels: Econometric and statistical modeling with Python*. in *9th Python in Science Conference*.
50. Arndt, D., et al., *PHASTER: a better, faster version of the PHAST phage search tool.* Nucleic Acids Research, 2016. **44**(W1): p. W16-W21.
51. Chen, J., et al., *Genome hypermobility by lateral transduction.* Science, 2018. **362**(6411): p. 207-212.
52. Chiang, Y.N., J.R. Penadés, and J. Chen, *Genetic transduction by phages and chromosomal islands: The new and noncanonical.* PLOS Pathogens, 2019. **15**(8): p. e1007878-e1007878.