Author for correspondence:
Nobuto Takeuchi
e-mail: nobuto.takeuchi@auckland.ac.nz

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

Nobuto Takeuchi[1,2], Sophia Hamada-Zhu[1] and Haruo Suzuki[3,4]

[1]School of Biological Sciences, the University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[2]Universal Biology Institute, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata, Japan
[4]Faculty of Environment and Information Studies, Keio University, Fujisawa, Japan

NT, 0000-0003-4949-6476; HS, 0000-0003-1447-6109

Mobile genetic elements (MGEs), such as phages and plasmids, often possess accessory genes encoding bacterial functions, facilitating bacterial evolution. Are there 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, relative to genomic backgrounds. In *Escherichia coli*, where this prophage–plasmid disparity is 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 disparity is not detected, ARGs and VFGs are barely found in prophages and plasmids. These results indicate that MGEs can differentiate in the types of accessory genes they carry depending on their infection strategies, suggesting a rule governing horizontal gene transfer mediated by MGEs.

1. 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 can be thought of as 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 hypothesized 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 suffer substantial disadvantage even if they lack 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, and are hence likely to hinge critically on 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 specialized transduction is the most similar to HGT mediated by plasmids [24] (see Discussion for the other mechanisms). In specialized transduction, phages transfer genes carried in their genomes. Therefore, specialized 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 specialized transduction [25]. However, the specialized 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 specialized 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 in subsets of the examined species. However, such biases were not detected in many species, where both ARGs and VFGs are hardly encoded in prophages and plasmids. Moreover, we found that prophage-borne VFGs are more functionally specific than plasmid-borne VFGs in *Escherichia coli*. Taken together, these results indicate that prophages and plasmids can differ in the types of accessory genes they carry.

2. Methods

Our method is sketched in figure S1 in the electronic supplementary material.

(a) 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 (table 1; the complete list of 7175 genomes analysed in this study is in electronic supplementary material, table S1).

(b) Prophage prediction

To predict prophages in the bacterial genomes, VIBRANT (v. 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 stand-alone 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.

(c) 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

Table 1. Numbers of predicted prophages per genome (s.d., standard deviation), numbers of orthologous groups of antibiotic resistance genes (OGARGs), and those of virulence factor genes (OGVFGs) in genomes (total), predicted plasmidic prophages (pl.) and predicted chromosomal prophages (ch.). Numbers in brackets are for OGARGs in prophages containing at least one phage-structure gene.

| species (no. of genomes considered) | prophage (s.d.) | OGARG | | | OGVFG | | |
| | | total | prophage | | total | prophage | |
			pl.	ch.		pl.	ch.
A. baumannii (275)	4.3 (1.7)	455	0	4 (4)	48	0	0
Bacillus anthracis (99)	4.1 (0.6)	17	0	1 (1)	39	0	0
Bordetella pertussis (562)	1.9 (0.3)	0	0	0	95	0	0
Brucella melitensis (64)	1.0 (0.3)	1	0	0	183	0	0
Burkholderia pseudomallei (126)	2.2 (1.3)	3	0	0	192	0	0
Campylobacter jejuni (220)	1.1 (1.2)	65	0	0	480	0	0
Enterococcus faecalis (64)	3.8 (1.8)	179	2 (0)	0	114	1	0
E. coli + Shigella spp. (1444)	7.1 (3.8)	3041	9 (7)	9 (9)	18 019	6	1590
Haemophilus influenzae (92)	3.0 (1.4)	7	0	0	161	0	1
Helicobacter pylori (225)	0.2 (0.5)	6	0	0	285	0	0
K. pneumoniae (873)	4.5 (2.1)	1473	7 (2)	7 (7)	372	0	0
Legionella pneumophila (102)	0.1 (0.3)	5	0	0	773	0	0
Listeria monocytogenes (263)	2.5 (1.3)	5	0	0	172	0	0
Mycobacterium tuberculosis (285)	0 (0.1)	4	0	0	997	0	0
Neisseria meningitidis (122)	1.9 (0.6)	4	0	0	123	0	1
P. aeruginosa (320)	4.5 (2.7)	543	0	0	526	0	2
Salmonella enterica (996)	3.9 (1.8)	826	2 (0)	0	1844	2	64
Staphylococcus aureus (618)	3.5 (1.3)	423	0	6 (2)	217	0	28
Streptococcus agalactiae (91)	2.0 (1.0)	46	0	1 (1)	71	0	0
Streptococcus pyogenes (235)	3.1 (1.5)	28	0	3 (3)	103	0	48
Vibrio cholerae (99)	1.5 (1.2)	101	1 (0)	2 (1)	251	1	5

'Extrachromosomal Element' were ignored (3, 1 or 5 contigs, respectively).

(d) ARG prediction

To predict ARGs in the bacterial genomes, AMRFinderPlus (v. 3.10.5) was run against translated coding sequences with the core subset of the database (v. 2021-09-11.1) and, if possible, the organism option enabled [45]. The genes predicted by AMR-FinderPlus 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).

(e) VFG prediction

To predict VFGs in the bacterial genomes, every entry in the combined VFG database (VFG database entry, for short) was queried against the translated coding sequences of every bacterial genome with BLASTP with an E-value threshold of 1×10^{-9} [46]. A gene in a bacterial genome (bacterial gene, for short) could match multiple VFG database entries, in which case the VFG database entry with the highest bit-score was selected as the best match. A bacterial gene was considered as encoding VF if it met the following additional criteria: (i) it was not annotated as a pseudo-gene in RefSeq [43]; (ii) the BLASTP alignment between the bacterial gene and its best-match VFG database entry, if any,

had at least 80% sequence identity and covers at least 80% of both the bacterial gene and the best-match VFG database entry; (iii) the species of the genome in which the bacterial gene resides was identical to the species in which experimental evidence for the best-match VFG database entry is available [38–40].

(f) Orthology prediction

To avoid double-counting orthologous genes in closely related strains, all genes within a species were clustered into putative orthologous groups based on protein sequence similarity and synteny conservation, as follows. First, preliminary orthologous pairs of genes were identified between every pair of genomes within each species through all-against-all sequence similarity searches using ProteinOrtho v. 6.0.25 (with DIAMOND v. 2.0.6 [47]; E-value cut-off of 11×10^{-5}; minimum coverage of best alignments of 75%; minimum per cent identity of best alignments of 25%; minimum reciprocal similarity of 0.95 [48]. ProteinOrtho defines a preliminary orthologous pair of genes as a reciprocal nearly best hit (RNBH), as follows. A nearly best hit (NBH) of a gene queried against a target genome is defined as a hit whose bit-score is not smaller by a factor f than that of the best hit. The value of f was 0.95, which is the default value of ProteinOrtho. If two genes are mutually NBH of each other, they form RNBH [48].

RNBHs obtained with ProteinOrtho were pruned based on synteny conservation with an in-house script (electronic

supplementary material), as follows. Let x and y be a pair of genes forming RNBH, and let X and Y be the genomic neighbours of x and y, respectively, where X is defined as a set of 21 genes consisting of ten genes upstream of x, 10 genes downstream of x, and x itself, and Y is likewise defined in terms of y (all contigs were assumed to be circular, and the orientation of genes were ignored). Let N_x and N_y be the number of genes in X and Y that form RNBHs with at least one gene in Y and X, respectively (note that a single gene in one genome can form RNBHs with multiple genes in another genome owing to tandem duplication). If both N_x and N_y are greater than 10 (i.e. a majority of the genes in X form RNBHs with the genes in Y, and *vice versa*), the RNBH formed by x and y was kept; otherwise, it was discarded [49].

Finally, the pruned RNBHs were clustered into putative orthologous groups with the spectral clustering algorithm implemented in ProteinOrtho v. 6.0.25 (minimum algebraic connectivity of 0.1; exact step 3; minimum number of species of 0; purity of 1×10^{-7}) [48].

(g) 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 contig. Moreover, genes in prophages residing in plasmids (plasmidic prophages, for short) were 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.

(h) Functional classification of prophage- and plasmid-borne virulence factor genes

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 'VFG prediction' section of Methods). 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] (electronic supplementary material, table S2). Consequently, 856 out of the 862 VFG database entries were assigned gene symbols (electronic supplementary material, 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 (electronic supplementary material, table S2). Consequently, prophage-borne and plasmid-borne OGVFGs were collectively assigned 232 gene symbols (electronic supplementary material, 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 categorized by their functions based on information obtained from VFDB, NCBI CDD and Uniprot [38,51,53] (electronic supplementary material, table S3). While most of the functional categories were adopted from VFDB [38], a new category, 'phage-related', was added, and the two categories, 'exoenzymes' and 'enzymes', were merged into 'enzymes'. Six gene symbols were annotated to have multiple functions and were therefore placed under multiple categories.

3. Results

(a) Prophage prediction

To examine the distribution of ARGs in prophages, we computationally predicted prophages using VIBRANT [54] and ARGs using AMRFinderPlus [45] in the genomes of 21 pathogenic bacterial species downloaded from the RefSeq database [43] (Methods; electronic supplementary material, tables S1 and S4). To avoid double-counting orthologous ARGs in closely related strains, we clustered all genes within a species 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 [50]). The result shows that one to nine prophage-borne OGARGs were detected in ten of the 21 examined species (table 1; electronic supplementary material, table S5).

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 probe the precision of the prophage prediction. The result of the examination shows that 24 out of 29 (i.e. 83% of) chromosomal prophages carrying ARGs contained phage-structure genes, whereas seven out of 18 (i.e. 39% of) plasmidic prophages carrying ARGs contained phage-structure genes (electronic supplementary material, table S6). This result means that 28 out of

33 (i.e. 85% of) OGARGs in chromosomal prophages are in prophages carrying phage-structure genes, whereas nine out of 21 (i.e. 43% of) OGARGs in plasmidic prophages are in prophages carrying phage-structure genes (table 1; note that the number of prophages carrying ARGs is smaller than that of prophage-borne ARGs because one prophage can carry multiple ARGs). This result suggests that the precision of VIBRANT is acceptable for chromosomal prophages.

During the manual examination, we noted that a subset of plasmidic prophages carrying ARGs also carried genes encoding integron integrases according to the RefSeq annotation (electronic supplementary material, table S6). Moreover, these prophages have a higher frequency of lacking phage-structure genes than those without integron integrase genes (IIGs, for short). Specifically, seven out of ten (i.e. 70% of) prophages carrying both ARGs and IIGs lack phage-structure genes, whereas nine out of 37 (i.e. 24% of) prophages carrying ARGs and lacking IIGs lack phage-structure genes. VIBRANT predicted that the genes annotated by RefSeq as IIGs matched 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]).

(b) Prophages carry significantly more virulence factor genes than antibiotic resistance genes relative to genomic backgrounds in three 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 (table 1; electronic supplementary material, table 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 electronic supplementary material, figure S1 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, unless otherwise stated [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 1): *E. coli* (Gammaproteobacteria), *S. enterica* (Gammaproteobacteria) and *S. aureus* (Firmicute). In all these species, chromosomal prophages carry VFGs more frequently than ARGs relative to genomic backgrounds. 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* and *L. monocytogenes*); and the three species where prophages were rarely or hardly detected (*H. pylori*, *L. pneumophila* and *M. tuberculosis*) (table 1 and figure 1). 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 in the three species in which they carry a sufficient number of ARGs or VFGs. However, in many bacterial species, prophages either are hardly present or carry little ARGs and VFGs, resulting in sample sizes that are too small to detect any biases. One possible exception is *S. pyogenes*, where prophages carry many more VFGs than ARGs, although the bias is not statistically significant.

Additionally, we tested the robustness of our results by repeating the aforementioned analysis for *E. coli* using PHASTER [58] as described in electronic supplementary material, text S1.

(c) Plasmids carry more antibiotic resistance genes than virulence factor genes relative to genomic backgrounds in nine species

To compare prophages to plasmids, we counted the number of OGARGs and OGVFGs encoded in plasmids (table 2). We then performed similar binomial tests under the null hypothesis that the relative frequencies of OGARGs and OGVFGs in the plasmids of each species are the same as those of all OGARGs and OGVFGs in the genomes of the respective species. The results of the tests indicate that the relative frequencies of OGARGs and OGVFGs in plasmids are significantly different from those of the genomic background in nine out of the 21 examined species (figure 2). In all nine species, plasmids are biased towards carrying ARGs more frequently than VFGs. These species include six species of Gammaproteobacteria (*A. baumannii*, *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *S. enterica* and *V. cholerae*), one species of Epsilonproteobacteria (*C. jejuni*) and two species of Firmicutes (*E. faecalis* and *S. aureus*). The remaining 12 species, where significant biases were not detected, can be grouped into three categories: the four species where plasmids are hardly present (*B. pertussis*, *B. melitensis*, *M. tuberculosis* and *N. meningitidis*), the seven
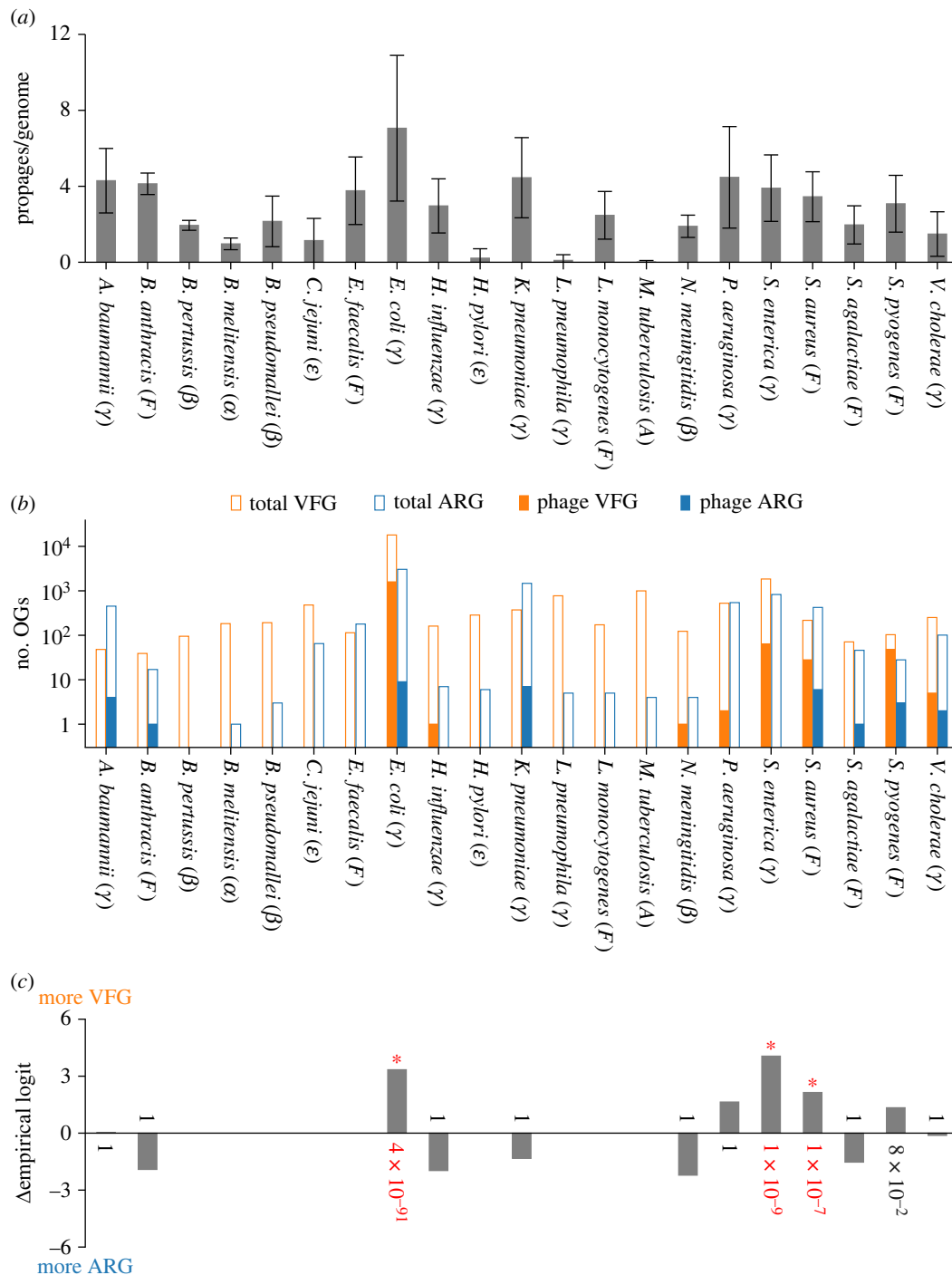
Table 2. Numbers of genetically distinct plasmids per genome (s.d., standard deviation), and numbers of orthologous groups of antibiotic resistance genes (OGARGs) and those of virulence factor genes (OGVFGs) in genomes (total) and plasmids.

	plasmid (s.d.)	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	18 019	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

Additionally, we examined the impact of relaxing the detection criteria for VFGs as described in electronic supplementary material, text S2, and figures S2 and S3.

(d) Prophage-borne virulence factor genes are functionally more restricted than plasmid-borne VFGs

We next asked whether chromosomal prophages and plasmids differ in the functional categories of VFGs they carry. To address this question, we mapped prophage-borne and plasmid-borne OGVFGs in E. coli to gene symbols (i.e. abbreviated gene names) and classified their functions using VFDB, NCBI CDD and UniProt [38,51,53] (see 'Functional classification of prophage- and plasmid-borne VFGs' section of Methods). The result shows that prophage-borne VFGs are mapped to 47 unique gene symbols, whereas plasmid-borne VFGs are mapped to 186 unique gene symbols (electronic supplementary material, table S8). This result suggests that the range of functions encoded by prophage-borne VFGs is less diverse than that encoded by plasmid-borne VFGs, as described in more detail below.

Nearly half (22 out of 47) of the gene symbols assigned to prophage-borne VFGs (prophage-borne VFG symbols, for short) are categorized as the effectors of type III secretion system (T3SS), the proteins secreted by the T3SS apparatus (figure 3). The T3SS effectors are involved in the destruction of host cells or the modulation of host immune systems [59]. The other prophage-borne VFG symbols include those encoding exotoxins (e.g. Shiga toxins), immune modulation and phage-related components (e.g. phage tails) (figure 3). Phage-related components are represented by the largest number of prophage-borne OGVFGs (930 out of 1580) (electronic supplementary material, table S8). Despite their abundance, the specific function of these genes is largely unknown, although some may be involved in intestinal colonization [60,61]. In addition, two prophage-borne VFG symbols are involved in adherence to host cells (figure 3). However, only one of them (porcine attaching-effacing associated protein, paa) is frequently located in prophages, whereas the other (E. coli common pilus chaperon, yagV/ecpE) is rarely located in prophages (electronic supplementary material, 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
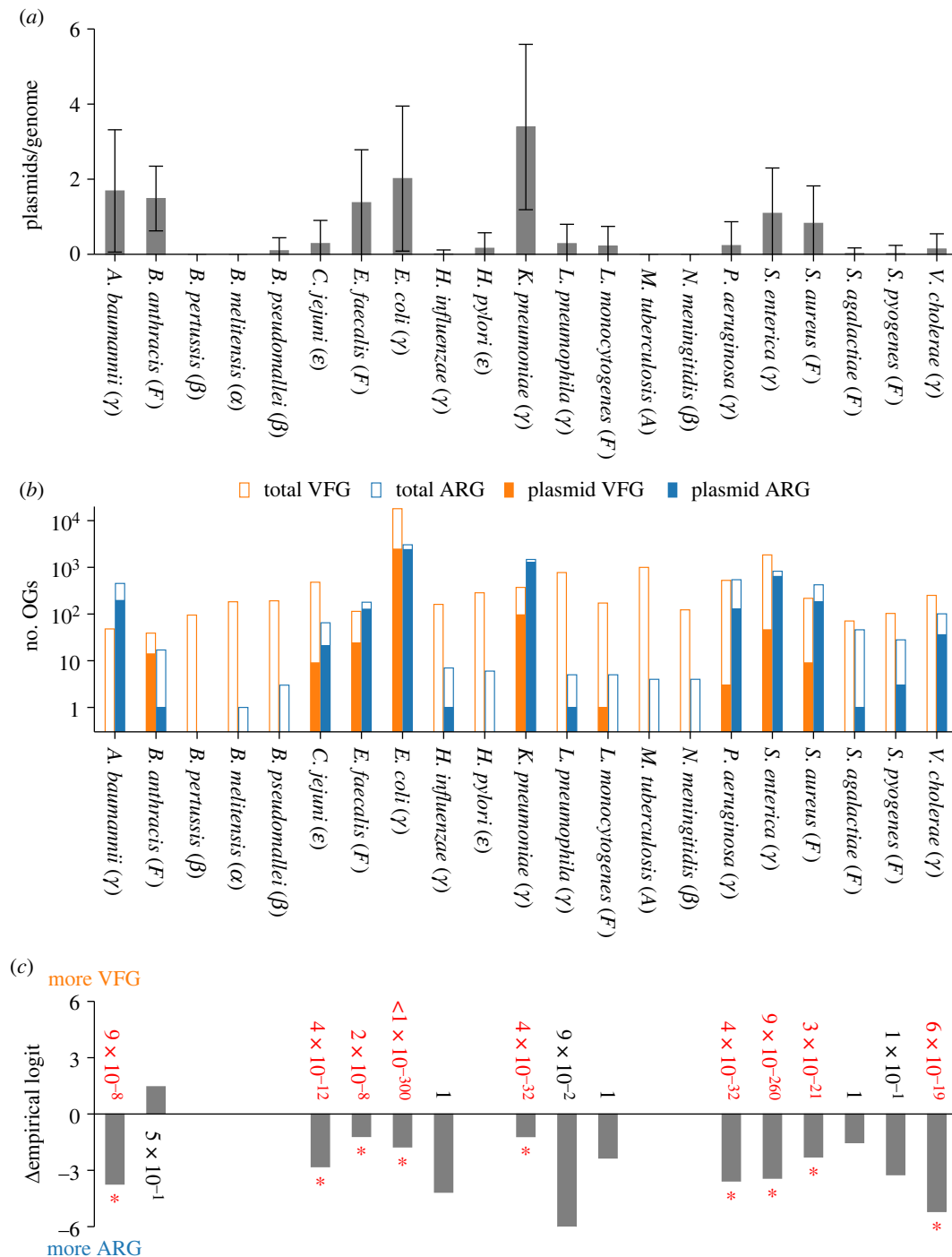
Figure 2. (*a*) Average number of genetically distinct plasmids per genome and standard deviation. (*b*) Numbers of orthologous groups of virulence factor genes (OGVFGs) and antibiotic resistance genes (OGARGs) in plasmids (filled bars) and genomes (open bars). Symbols in brackets indicate taxonomic 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). When $a = b = 0$, Δempirical logit is not plotted. 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.

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; electronic supplementary material, table S8). Moreover, the total

number of prophage-borne VFGs without clustering of orthologous genes is greater than that of plasmid-borne VFGs (10 965 versus 8020 VFGs; electronic supplementary material, table S8).

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 (electronic supplementary material, table S3). This result means that prophages and plasmids carry distinct
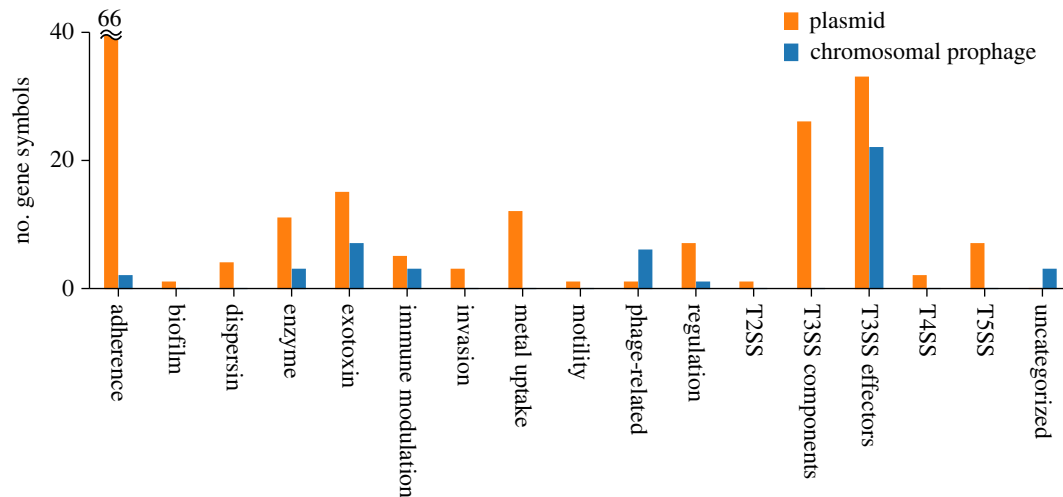
Figure 3. Numbers of virulence factor gene symbols detected in plasmids and chromosomal prophages in *E. coli*.
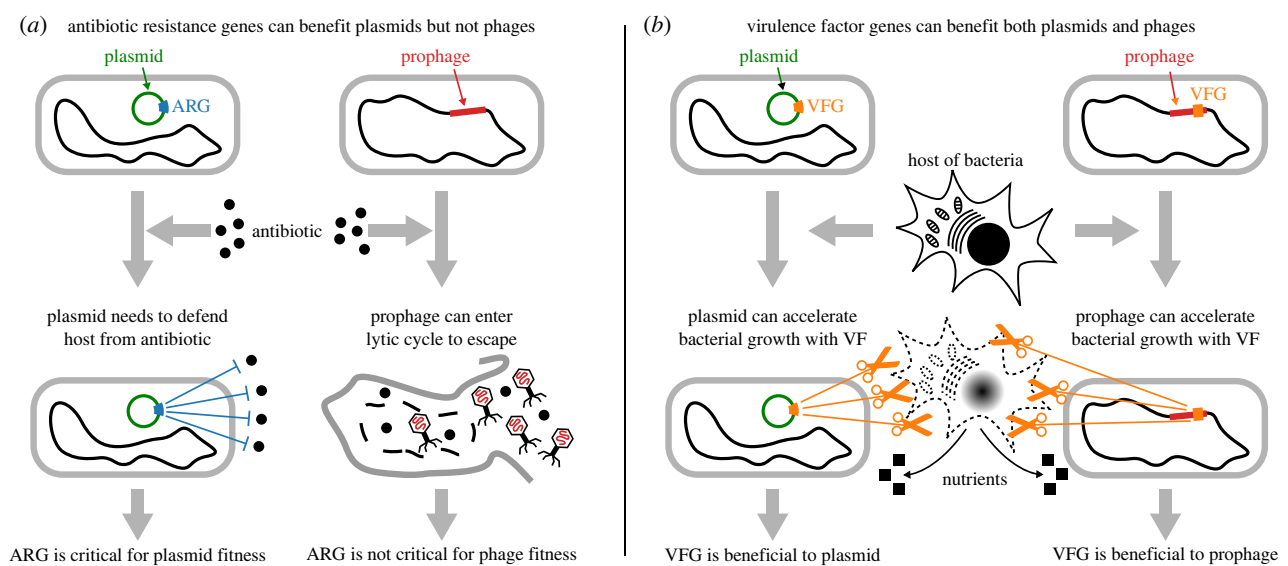

Figure 4. Schematic drawing of hypothesis formulated based on our results. (*a*) Plasmids substantially benefit from carrying ARG because they cannot abandon their hosts and thus need to minimize death of bacteria. By contrast, temperate phages do not suffer substantial disadvantage even if they lack ARGs (i.e. do not critically depend on ARGs) because if their hosts are in danger, they can abandon their hosts and seek new hosts by entering lytic cycles. (*b*) Both plasmids and phages benefit from VFGs because VFGs can enhance growth of bacteria by causing bacteria to exploit their hosts more aggressively.

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 (electronic supplementary material, 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.

4. Discussion

The results presented above indicate that in the species where 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, relative to the genomic backgrounds. In the

other species, plasmids or prophages either barely carry ARGs and VFGs or are nearly absent. Lastly, 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, these findings indicate that prophages and plasmids can carry different types of bacterial accessory genes.

Based on the results presented above, we formulate the following hypothesis to test for the future (figure 4). Temperate phages do not gain sufficient selective advantage from carrying ARGs because if their hosts are in danger, they can abandon their hosts and seek new hosts by undergoing lytic replication [18–22]. By contrast, both phages and plasmids can benefit from VFGs because VFGs can accelerate bacterial replication by making bacteria exploit their hosts more aggressively [17]. In addition, our results suggest that plasmids benefit from a wider range of virulence functions than that from which phages benefit in *E. coli* (figure 3; electronic supplementary material, table S8).

Besides the above hypothesis, the difference between prophages and plasmids might also be explained by the

interplay between the scope of natural selection and the host range of MGEs (an idea suggested by an anonymous reviewer). Antibiotic resistance might be beneficial to all bacteria in a community, causing community-wide selection. By contrast, virulence factors might be beneficial only to a subset of bacteria depending on their life-history traits [62], resulting in strain-specific selection. Interestingly, plasmids typically have broader host ranges than those of phages [63,64]. This alignment between the scope of selection and the host range of MGEs might contribute to the observed difference between prophages and plasmids in the distribution of VFGs and ARGs.

The results described in this study, however, do not necessarily mean that plasmids rarely carry VFGs or that prophages always carry VFGs. For example, in *K. pneumoniae*, plasmids carry ARGs more frequently than VFGs relative to the genomic background; nevertheless, plasmids carry a large number of VFGs (96 out of 372 orthologous groups), whereas prophages carry none (tables 1 and 2). A similar, yet less striking, pattern is also seen in *E. faecalis*. Given that many prophages are predicted in these species (table 1 and figure 1), the absence of prophage-borne VFGs is unlikely to be merely due to the limited sensitivity of the prophage detection tool.

Heterogeneity in the distributions of VFGs suggests that plasmids and prophages play variable roles in the pathogenicity of different bacterial species. For example, prophages carry many distinct VFGs in *E. coli* (1590 OGVFGs), *S. enterica* (64 OGVFGs), *S. aureus* (28 OGVFGs) and *S. pyogenes* (48 OGVFGs), whereas they hardly carry any VFGs in *A. baumannii* (0 OGVFGs), *B. anthracis* (0 OGVFGs), *E. faecalis* (0 OGVFGs) and *K. pneumoniae* (0 OGVFGs), even though prophages are frequently found in the genomes of all these species (figure 1 and table 1). Similar heterogeneity is observed for plasmids: plasmids carry many distinct VFGs in *B. anthracis* (14 OGVFGs), *E. faecalis* (24 OGVFGs), *E. coli* (2452 OGVFGs), *K. pneumoniae* (96 OGVFGs) and *S. enterica* (46 OGVFGs), whereas they hardly carry any VFGs in *A. baumannii* (0 OGVFGs), even though plasmids are frequently found in the genomes of all these species (figure 2 and table 2). This heterogeneity might be related to species' different life-history traits [62], although further research is needed to explore this relationship. The heterogeneity of VFG distributions is in contrast with the situation for ARGs. In the species where many distinct ARGs are detected (e.g. *A. baumannii*, *E. coli*, *K. pneumoniae*, *P. aeruginosa*, *S. enterica* and *S. aureus*), sizable fractions of these ARGs (greater than 20%, or greater than 40% if *P. aeruginosa* is excluded) are carried by plasmids (figure 2 and table 2), a result that corroborates the notion that plasmids play a major role in the dissemination of ARGs [1–3].

The absolute numbers of prophage-borne ARGs reported in this study need to be interpreted with caution because the prophage prediction tool has limited sensitivity and precision. In particular, limited sensitivity implies that prophages carrying ARGs might have been missed, so the results do not necessarily indicate that prophages are truly devoid of ARGs in many bacterial species. However, the conclusion of this study does not directly depend on the absolute numbers of prophage-borne ARGs because it is based on the comparison between the relative frequencies of prophage-borne ARGs and VFGs and those of plasmid-borne ARGs and VFGs. This comparison hinges on the assumption that the prediction of prophages carrying ARGs and VFGs are biased by common factors so that the biases are cancelled out.

To consider the question investigated in this study from another angle, we additionally examined the densities of non-clustered VFGs and ARGs per nucleotide in prophages, plasmids and genomes as described in electronic supplementary material, text S3, and figures S4 and S5.

Regarding the limitations of prophage prediction tools, it is pertinent to discuss a discrepancy between our result and the result of Kondo *et al.* [34] with respect to *P. aeruginosa* [34]. While we found no prophage-borne ARGs in *P. aeruginosa*, Kondo *et al.* [34] report that more than 10% of *P. aeruginosa* genomes possess prophage-borne ARGs [34]. The important difference between Kondo *et al.* [34] and our study is that they used different tools for prophage prediction: Kondo *et al.* [34] uses PHASTER [58], whereas we used VIBRANT [54]. To investigate the cause of the above discrepancy, we manually examined the 11 prophages described in Kondo *et al.* [34] that carry ARGs and are predicted as 'intact' by PHASTER in *P. aeruginosa* [34]. We found that the examined prophages could be grouped into two categories (electronic supplementary material, table S9). In the first group (five prophages), both VIBRANT and PHASTER predicted prophages in almost the same genomic locations. However, PHASTER predicted longer genomic regions that included ARGs as prophages, whereas VIBRANT predicted shorter regions excluding ARGs. We do not know 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.* [34] 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: specialized, generalized and lateral

transduction [24,65,66]. In specialized transduction (the focus of this study), a transferred gene constitutes a part of a phage genome [24]. By contrast, in generalized and lateral transduction, a transferred gene is originally encoded in bacterial DNA, which is encapsulated into phage particles and subsequently transferred to other cells [24,65,66]. 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 can differ in the functional categories of accessory genes they carry depending on their strategies of infection.

References

1. Bennett PM. 2009 Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. *Br. J. Pharmacol.* **153**(S1), S347–S357. (doi:10.1038/sj.bjp.0707607)

2. Von Wintersdorff CJH, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, Savelkoul PHM, Wolffs PFG. 2016 Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 173. (doi:10.3389/fmicb.2016.00173)

3. Partridge SR, Kwong SM, Firth N, Jensen SO. 2018 Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **31**, e00088–17. (doi:10.1128/CMR.00088-17)

4. Wagner PL, Waldor MK. 2002 Bacteriophage control of bacterial virulence. *Infect. Immun.* **70**, 3985–3993. (doi:10.1128/IAI.70.8.3985-3993.2002)

5. Gyles C, Boerlin P. 2014 Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet. Pathol.* **51**, 328–340. (doi:10.1177/0300985813511131)

6. O'Neill J. 2016 *Tackling drug-resistant infections globally: final report and recommendations*. Review on Antimicrobial Resistance. See https://wellcomecollection.org/works/thvwsuba/items.

7. Pilla G, Tang CM. 2018 Going around in circles: virulence plasmids in enteric pathogens. *Nat. Rev. Microbiol.* **16**, 484–495. (doi:10.1038/s41579-018-0031-2)

8. Murray CJL *et al.* 2022 Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655. (doi:10.1016/S0140-6736(21)02724-0)

9. Frost LS, Leplae R, Summers AO, Toussaint A. 2005 Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732. (doi:10.1038/nrmicro1235)

10. Hall JPJ, Brockhurst MA, Harrison E. 2017 Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Phil. Trans. R. Soc. B* **372**, 20160424. (doi:10.1098/rstb.2016.0424)

11. Ghaly TM, Gillings MR. 2018 Mobile DNAs as ecologically and evolutionarily independent units of life. *Trends Microbiol.* **26**, 904–912. (doi:10.1016/j.tim.2018.05.008)

12. Rankin DJ, Rocha EPC, Brown SP. 2011 What traits are carried on mobile genetic elements and why? *Heredity* **106**, 1–10. (doi:10.1038/hdy.2010.24)

13. Eberhard WG. 1990 Evolution in bacterial plasmids and levels of selection. *Q. Rev. Biol.* **65**, 3–22. (doi:10.1086/416582)

14. Bergstrom CT, Lipsitch M, Levin BR. 2000 Natural selection, infectious transfer and the existence conditions for bacterial plasmids. *Genetics* **155**, 1505–1519. (doi:10.1093/genetics/155.4.1505)

15. Svara F, Rankin DJ. 2011 The evolution of plasmid-carried antibiotic resistance. *BMC Evol. Biol.* **11**, 130–130. (doi:10.1186/1471-2148-11-130)

16. Lehtinen S, Huisman JS, Bonhoeffer S. 2021 Evolutionary mechanisms that determine which bacterial genes are carried on plasmids. *Evol. Lett.* **5**, 290–301. (doi:10.1002/evl3.226)

17. Abedon ST, Lejeune JT. 2005 Why bacteriophage encode exotoxins and other virulence factors. *Evol. Bioinform.* **1**, 117693430500100–117693430500100. (doi:10.1177/117693430500100001)

18. Matsushiro A, Sato K, Miyamoto H, Yamamura T, Honda T. 1999 Induction of prophages of enterohemorrhagic *Escherichia coli* O157:H7 with Norfloxacin. *J. Bacteriol.* **181**, 2257–2260. (doi:10.1128/JB.181.7.2257-2260.1999)

19. Goerke C, Köller J, Wolz C. 2006 Ciprofloxacin and trimethoprim cause phage induction and virulence modulation in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **50**, 171. (doi:10.1128/AAC.50.1.171-177.2006)

20. Allen HK, Looft T, Bayles DO, Humphrey S, Levine UY, Alt D, Stanton TB. 2011 Antibiotics in feed induce prophages in swine fecal microbiomes. *mBio* **2**, e00260-11. (doi:10.1128/mBio.00260-11)

21. López E, Domenech A, Ferrándiz MJ, Frias MJ, Ardanuy C, Ramirez M, García E, Liñares J, de la Campa AG. 2014 Induction of prophages by fluoroquinolones in *Streptococcus pneumoniae*: implications for emergence of resistance in genetically-related clones. *PLoS ONE* **9**, e94358–e94358. (doi:10.1371/journal.pone.0094358)

22. Sutcliffe SG, Shamash M, Hynes AP, Maurice CF. 2021 Common oral medications lead to prophage induction in bacterial isolates from the human gut. *Viruses* **13**, 455–455. (doi:10.3390/v13030455)

23. Wallace MJ, Fishbein SRS, Dantas G. 2020 Antimicrobial resistance in enteric bacteria: current state and next-generation solutions. *Gut Microbes* **12**, 1799654. (doi:10.1080/19490976.2020.1799654)

24. Snyder L, Champness W, Champness W. 2013 *Molecular genetics of bacteria*, 4th edn. Washington, DC: ASM Press.

25. Villa TG, Feijoo-Siota L, Rama JR, Sánchez-Pérez A, Viñas M. 2019 *Horizontal gene transfer between bacteriophages and bacteria: antibiotic resistances and toxin production*, pp. 97–142. Berlin, Germany: Springer International Publishing.

26. Giovanetti E, Brenciani A, Morroni G, Tiberi E, Pasquaroli S, Mingoia M, Varaldo PE. 2015 Transduction of the Streptococcus pyogenes bacteriophage Φm46.1, carrying resistance genes mef(A) and tet(O), to other Streptococcus species. *Front. Microbiol.* **5**(DEC), 746.

27. Modi SR, Lee HH, Spina CS, Collins JJ. 2013 Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222. (doi:10.1038/nature12212)

28. Subirats J, Sànchez-Melsió A, Borrego CM, Balcázar JL, Simonet P. 2016 Metagenomic analysis reveals that bacteriophages are reservoirs of antibiotic resistance genes. *Int. J. Antimicrob. Agents* **48**, 163–167. (doi:10.1016/j.ijantimicag.2016.04.028)

29. Debroas D, Siguret C. 2019 Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J.* **13**, 2856–2867. (doi:10.1038/s41396-019-0478-9)

30. Strange JE, Leekitcharoenphon P, Møller FD, Aarestrup FM. 2021 Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. *Sci. Rep.* **11**, 1600. (doi:10.1038/s41598-021-80990-6)

31. Muniesa M, García A, Miró E, Mirelis B, Prats G, Jofre Jp, Navarro F. 2004 Bacteriophages and diffusion of β-lactamase genes. *Emerg. Infect. Dis.* **10**, 1134–1137. (doi:10.3201/eid1006.030472)

32. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit M-A. 2017 Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247. (doi:10.1038/ismej.2016.90)

33. López-Leal G, Santamaria RI, Cevallos M, Gonzalez V, Castillo-Ramírez S. 2020 Prophages encode antibiotic resistance genes in *Acinetobacter baumannii. Microb. Drug Resist.* **26**, 1275–1277.

34. Kondo K, Kawano M, Sugai M. 2021 Distribution of antimicrobial resistance and virulence genes within the prophage-associated regions in nosocomial pathogens. *mSphere* **6**, e00452-21. (doi:10.1128/mSphere.00452-21)

35. Kleinheinz KA, Joensen KG, Larsen MV. 2014 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* **4**, e27943–e27943. (doi:10.4161/bact.27943)

36. Gunathilaka GU, Tahlan V, Mafiz AI, Polur M, Zhang Y. 2017 Phages in urban wastewater have the potential to disseminate antibiotic resistance. *Int. J. Antimicrob. Agents* **50**, 678–683. (doi:10.1016/j.ijantimicag.2017.08.013)

37. Cook R et al. 2021 INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. *PHAGE* **2**, 214–223. (doi:10.1089/phage.2021.0007)

38. Liu B, Zheng D, Jin Q, Chen L, Yang J. 2018 VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**(D1), D687–D692. (doi:10.1093/nar/gky1080)

39. Sayers S et al. 2019 Victors: a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Res.* **47**(D1), D693–D700. (doi:10.1093/nar/gky999)

40. Davis JJ et al. 2020 The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**(D1), D606–D612.

41. Sayers EW et al. 2020 GenBank. *Nucleic Acids Res.* **48**(D1), D84–D86. (doi:10.1093/nar/gkz899)

42. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012 CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152. (doi:10.1093/bioinformatics/bts565)

43. Li W et al. 2021 RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.* **49**(D1), D1020–D1028. (doi:10.1093/nar/gkaa1105)

44. Roach MJ, Mcnair K, Giles SK, Inglis LK, Pargin E, Roux S, Decewicz P, Edwards RA. 2021 Philympics 2021: prophage predictions perplex programs. *F1000Research* **10**, 758–758. (doi:10.12688/f1000research.54449.1)

45. Feldgarden M et al. 2021 AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.* **11**, 12 728–12 728. (doi:10.1038/s41598-021-91456-0)

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410. (doi:10.1016/S0022-2836(05)80360-2)

47. Buchfink B, Xie C, Huson DH. 2015 Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60. (doi:10.1038/nmeth.3176)

48. Lechner M, Findeiäÿ S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011 Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinf.* **12**, 124–124. (doi:10.1186/1471-2105-12-124)

49. Novichkov PS, Ratnere I, Wolf YI, Koonin EV, Dubchak I. 2008 ATGC: a database of orthologous genes from closely related prokaryotic genomes and a research platform for microevolution of prokaryotes. *Nucleic Acids Res.* **37**(suppl_1), D448–D454.

50. Pfeifer E, Moura De Sousa JA, Touchon M, Rocha EPC. 2021 Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673. (doi:10.1093/nar/gkab064)

51. Lu S et al. 2020 CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**(D1), D265–D268. (doi:10.1093/nar/gkz991)

52. O'leary NA et al. 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**(D1), D733–D745. (doi:10.1093/nar/gkv1189)

53. Consortium TU. 2021 UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**(D1), D480–D489. (doi:10.1093/nar/gkaa1100)

54. Kieft K, Zhou Z, Anantharaman K. 2020 VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90–90. (doi:10.1186/s40168-020-00867-0)

55. Fouts DE. 2006 Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851. (doi:10.1093/nar/gkl732)

56. Holm S. 1979 A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70.

57. Seabold S, Perktold J. 2010 Statsmodels: econometric and statistical modeling with Python. In *9th Python in Science Conference. Austin, TX, 28 June to 3 July 2010*, pp. 57–61. (doi:10.25080/Majora-92bf1922-011)

58. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016 PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**(W1), W16–W21. (doi:10.1093/nar/gkw387)

59. Shenoy AR, Furniss RCD, Goddard PJ, Clements A. 2018 In *Modulation of host cell processes by T3SS effectors* (eds G Frankel, EZ Ron), pp. 73–115. Berlin, Germany: Springer International Publishing.

60. Van Diemen PM, Dziva F, Stevens MP, Wallis TS. 2005 Identification of Enterohemorrhagic *Escherichia coli* O26:H— genes required for intestinal colonization in calves. *Infect. Immun.* **73**, 1735–1743. (doi:10.1128/IAI.73.3.1735-1743.2005)

61. Dziva F, Van Diemen PM, Stevens MP, Smith AJ, Wallis TS. 2004 Identification of *Escherichia coli* O157: H7 genes influencing colonization of the bovine gastrointestinal tract using signature-tagged mutagenesis. *Microbiology* **150**, 3631–3645. (doi:10.1099/mic.0.27448-0)

62. Touchon M, Bernheim A, Rocha EPC. 2016 Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* **10**, 2744–2754. (doi:10.1038/ismej.2016.47)

63. Jain A, Srivastava P. 2013 Broad host range plasmids. *FEMS Microbiol. Lett.* **348**, 87–96. (doi:10.1111/1574-6968.12241)

64. De Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. 2019 Molecular and evolutionary determinants of bacteriophage host range. *Trends Microbiol.* **27**, 51–63. (doi:10.1016/j.tim.2018.08.006)

65. Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ, Ross Fitzgerald J, Penadés JR. 2018 Genome hypermobility by lateral transduction. *Science* **362**, 207–212. (doi:10.1126/science.aat5867)

66. Chiang YN, Penadés JR, Chen J. 2019 Genetic transduction by phages and chromosomal islands: the new and noncanonical. *PLoS Pathog.* **15**, e1007878. (doi:10.1371/journal.ppat.1007878)

67. Takeuchi N, Hamada-Zhu S, Suzuki H. 2023 Prophages and plasmids can display opposite trends in the types of accessory genes they carry. Figshare. (doi:10.6084/m9.figshare.c.6699176)