Defense systems are pervasive across chromosomally integrated mobile genetic elements and are inversely correlated to virulence and antimicrobial resistance

João Botelho*

Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Madrid, Spain

*To whom correspondence should be addressed: botelho@evolbio.mpg.de

Abstract

Mobile genetic elements (MGEs) are key promoters of bacterial and archaeal genome evolution. These elements can be located extrachromosomally (such as plasmids) or integrated within the chromosome. Well-known examples of chromosomally integrated MGEs (ciMGEs) are the integrative and conjugative/mobilizable elements (ICEs and IMEs), and most studies so far have focused on the biological mechanisms that shape the lifestyle of these elements. It is crucial to illustrate the overall diversity and understand the distribution of circulating ciMGEs in the microbial community as the number of genome sequences increases exponentially. Herein, I scanned a publicly available collection of more than 20000 bacterial and archaeal non-redundant genomes and found more than 13000 ciMGEs across multiple phyla, representing a massive increase in the number of ciMGEs currently available in public databases (<1000). Although ICEs are the most important ciMGEs for the accretion of defensive systems, virulence genes, and antimicrobial resistance (AMR) genes, IMEs outnumbered ICEs. Moreover, I discovered that defense systems, AMR, and virulence genes are negatively correlated in both ICEs and IMEs. Multiple representatives of these ciMGEs form heterogeneous communities and challenge interphylum barriers. Finally, I observed that the functional landscape of ICEs is populated by uncharacterized proteins. Taken together, this study provides a comprehensive catalog compiling the nucleotide sequence and associated metadata for ciMGEs from 34 distinct phyla across the bacterial and archaeal domains.

Introduction

Bacterial and archaeal genomes have evolved diverse defense systems to target endoparasites, such as different types of mobile genetic elements (MGEs) that can be in conflict with their host cells (1). In fact, more than one class of defense systems can be found in the same cell, nonrandomly clustered in so-called defense islands (2). These systems are expected to be prone to loss (3, 4), and in the long-term their co-occurrence with genes involved in mobilization is crucial for persistence. For example, two of the most well-known defense systems (toxin-antitoxin and restriction-modification) have a distinct evolutionary history from their prokaryotic hosts, and exploit MGEs as vectors for their spread (5). From an evolutionary viewpoint, it is tempting to propose that the presence of multiple defense systems may be more involved in the maintenance of MGEs than of the cell (1, 6). Besides, MGEs can use additional strategies to offset the short-term deleterious effects of their cargo genes, such as increased rates of conjugation (7, 8). The ability for horizontal transmission contributes to the distinct patterns of diversity and ecological distribution of MGEs. Indeed, horizontal gene transfer (HGT) and gene loss have crucial roles in the patchy distribution of defense mechanisms and antimicrobial resistance (AMR) genes at the species and strain level (9). Different defense systems can work together or independently to defend the host against foreign DNA (10, 11). Besides, a recent study has found that AMR genes and defense systems can cluster together in mobile elements (12). Worryingly, this study found that phage infection itself spur the excision of integrative and conjugative elements (ICEs) and subsequent transfer by conjugation, leading to the spread of AMR genes.

Pathogenicity islands are a group of chromosomally integrated MGEs (ciMGEs), including different types of bona fide MGEs such as ICEs, integrative and mobilizable elements (IMEs), and actinomycete ICEs (AICEs), among others. ICEs are transferred as linear single-stranded DNA and use a type-IV secretion systems for conjugation (similar to the mechanism used by conjugative plasmids) (13). In Actinobacteria, however, some ICEs are delivered as double-stranded DNA following a distinct process (14). As so, these ICEs have been categorized as AICEs. While AICEs and ICEs encode an intact conjugation apparatus, and are thus self-transmissible, IMEs encode their own excision and integration module but lack a fully functional conjugative apparatus for autonomous transfer. Still, the latter elements can pirate conjugative plasmids and ICEs to promote their own dissemination (15). Both elements are a driving force for bacterial adaptation and evolution, at least in part due to the different cargo genes that may provide the host a selective advantage in specific environments (16).

Since phage predation selects for multiple defense systems that frequently cluster on mobilizable defense islands, I explore here if different ciMGE types (i.e., ICEs, IMEs, and

AICEs) are important hotspots for the accretion of defense systems, helping to protect the host from superinfection by other MGEs. Considering the known evolutionary relationship MGEs have with defense systems (2, 11, 17, 18), AMR (12, 19, 20) and virulence genes (21), I also explore if these three functional groups are positively or negatively correlated across ciMGEs from multiple phyla. I found that i) IMEs and ICEs are widespread across different phyla; ii) ICEs are the most important ciMGEs for the accumulation of defense systems, AMR, and virulence genes; iii) these three functional groups are negatively correlated across ICEs and IMEs; and iv) ICEs and IMEs from multiple phyla share high genetic similarity and challenge interphylum barriers. Finally, this work provides a comprehensive resource compiling 13274 ciMGEs from Bacteria and Archaea with associated metadata extracted from 34 distinct phyla.

Material and methods

Genome collection

Bacterial genomes sequenced at the complete level (i.e., all replicons included inside the genomes, such as the chromosomes and extra-chromosomal elements, are fully assembled with gaps not exceeding ten ambiguous bases) were downloaded from NCBI RefSeq (08-07-2021) using ncbi-genome-download v0.3.0 (https://github.com/kblin/ncbi-genome-download). In parallel, archaeal genomes were downloaded using the same approach (28-06-2022). To examine the presence of redundant genomes, I used Assembly Dereplicator v0.1.0 (https://github.com/rrwick/Assembly-Dereplicator) with a Mash distance clustering threshold of 0.001 and a batch size of 25000. The classify workflow from GTDB-Tk v2.0.0 (22) was used to correct the taxonomy classification of the downloaded bacterial and archaeal genomes. Two genomes (RefSeq assembly accession numbers GCF_900660555.1 and GCF_002158865.1) were excluded, as these were flagged by the align module in GTDB-Tk as insufficient number of amino acids in MSA.

Identification of ciMGEs

I then used a custom python script to split all genome files into 45482 individual replicon files (i.e, chromosomal and extra-chromosomal replicons that are part of the genomes). A total of 22269 replicons with the word 'plasmid' in the fasta-headers were removed. The remaining 23213 chromosomal replicon files were used as input in ICEfinder (23) to look for ciMGEs (i.e., ICEs, IMEs, and AICEs). The ciMGEs were then extracted from the chromosomes and reannotated with Prokka v1.14.6 (24).

Functional annotation and network analysis

The resulting proteins and gff files from Prokka annotation were used as input to search for defense systems with PADLOC v1.1.0 (DB v1.4.0) (25). AMRFinder v3.10.30 (26) was used to search for the presence of AMR genes across the ciMGEs. To look for virulence genes, I used the virulence factor database VFDB (27) (4327 genes, 27-06-2022) with abricate v1.0.1 (https://github.com/tseemann/abricate). The ciMGE dataset was then dereplicated with MMseqs2 v13.45111 (28) using 90% sequence identity and 80% coverage. Cluster of orthologous groups (COG) categories were searched with eggNOG-mapper v2.1.9 (DB v5.0.2) (29). To estimate the pairwise distances between all ciMGE types (i.e., ICEs, IMEs, and AICEs), I reduced the dereplicated ciMGEs into sketches and compared the Jaccard index (JI) and mutation distances between pairs of ciMGEs using BinDash v 0.2.1 (30). Each ciMGE nucleotide sequence was converted to a set of 21-bp k-mers. The mutation distances were used as weights to plot the undirected and weighted network and the communities were detected with Infomap using 0.75 Markov time (31, 32) in Cytoscape v3.9.1 (https://cytoscape.org/) under the prefuse force directed layout. I then used the Analyzer function in Cytoscape to calculate a comprehensive set of topological parameters, such as the network density, the clustering coefficient, the centralization, and the heterogeneity.

Statistics

Comparisons between i) the number of defense systems, virulence, and AMR genes normalized to ciMGE size per phylum; ii) GC content deviation (GC host genome – GC ciMGE) across the different ciMGE types; iii) the number of cargo genes per ciMGE per phylum; iv) the COG counts normalized to ciMGE size were performed using the Kruskal-Wallis test, and the p-values adjusted using the Holm–Bonferroni method. Comparisons between the normalized number of COG categories between ICEs and IMEs were performed using the Wilcoxon test, and the p-values adjusted using the Holm–Bonferroni method. Correlation between defense systems, AMR, and virulence genes was assessed by the Pearson method. This method was also used to compute the correlation coefficient between the GC content and sequence length of ciMGEs and their host genome. Values above 0.05 were considered as non-significant (ns). We used the following convention for symbols indicating statistical significance: * for p <= 0.05, ** for p <= 0.01, *** for p <= 0.001, and **** for p <= 0.0001.

Results

IMEs and ICEs are widespread across multiple phyla

A total of 22334 complete genomes from 57 phyla (21897 genomes in 48 bacterial phyla and 437 genomes in 9 archaeal phyla, **Supplementary Table 1**) were used as input to search for

ciMGEs. More than 80% of the downloaded genomes belong to three bacterial phyla: Proteobacteria, Firmicutes, and Actinobacteriota (n=11843, 4706, and 2226, respectively, **Figure S1**). The search resulted in the identification of 13274 ciMGEs, including 6331 IMEs, 5869 ICEs, 1061 AICEs, and 13 elements annotated as putative conjugative regions. These ciMGEs extracted from a total of 34 phyla (13258 ciMGEs in 8000 bacterial genomes and 16 ciMGEs in 16 archaeal genomes, **Supplementary Table 2**). This means that 36.5% of bacterial genomes (8000/21897) carry at least one ciMGE, while only 3.7% of archaeal genomes (16/437) carry at least one ciMGE. As expected, the absolute number of ICEs and ICEs was higher across the most dominant phyla (Proteobacteria, Firmicutes, Actinobacteriota, Bacteroidota, Campylobacterota, and Firmicutes_A), each carrying more than 100 ICEs and IMEs (**Supplementary Tables 1** and **2**). IMEs were the most frequently identified ciMGE across our dataset (**Figures 1A and 1B)**, both in bacteria and archaea (n=6319 and 12, respectively, across a total of 4749 genomes), and the one with the widest dispersion across different phyla (32 out of the 57, **Supplementary Table 2**). ICEs were also widely disseminated in bacterial and archaeal genomes (n=5865 and 4, respectively, across a total of 4328 genomes), and across 25 phyla. While these ciMGEs were found in both domains, AICEs were only found in bacteria (n=1061 in 519 genomes). In fact, AICEs were constrained to particular phyla (mostly in Actinobacteriota, and rarely in Firmicutes and Proteobacteria). When comparing with the number of ciMGEs deposited in ICEberg with nucleotide sequences available (n=718 ICEs, 111 IMEs, and 50 AICEs), this analysis considerably expands the repertoire of these elements.

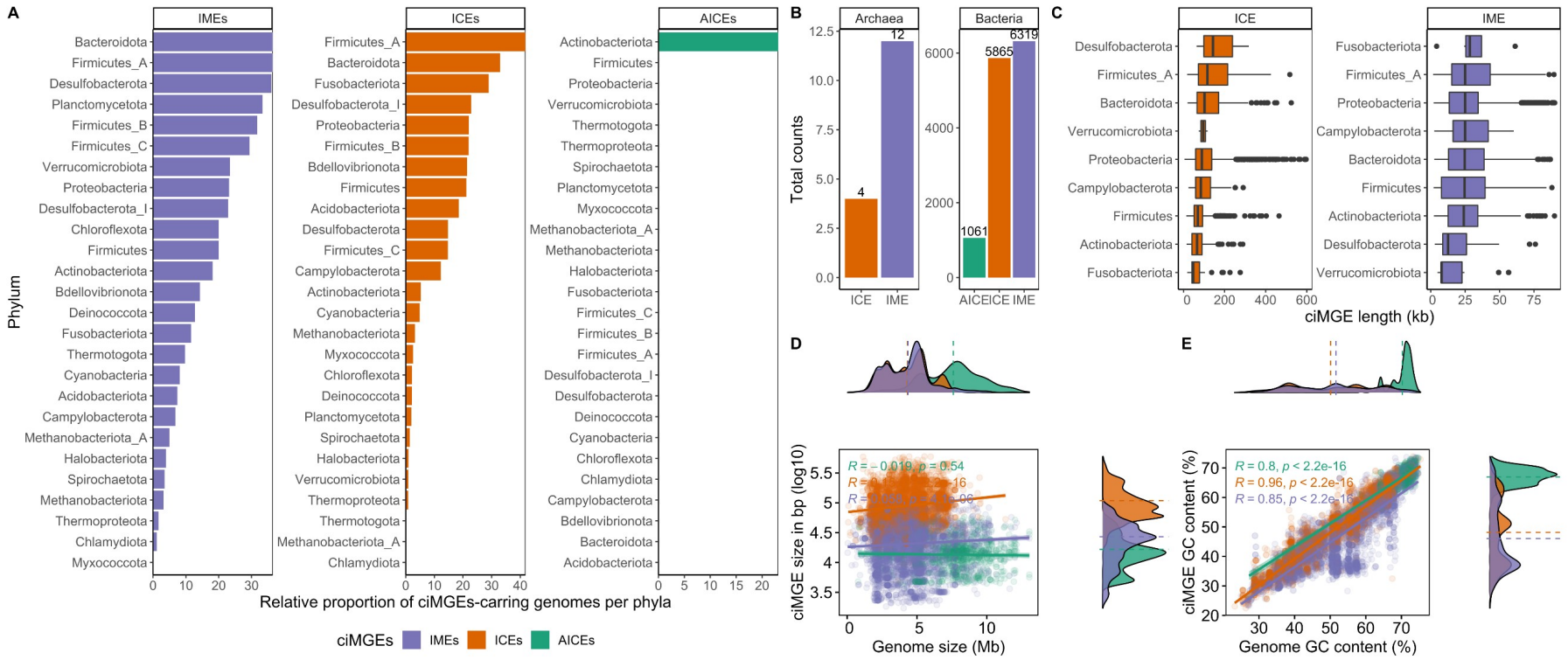

Figure 1. Distribution of ciMGEs across Bacteria and Archaea. A) Relative proportion of IME-, ICE-, and AICE-carrying genomes per phyla. Phylum in the y-axis are listed in descending order. **B)** Total counts of ICEs, IMEs, and AICEs across bacterial and archaeal genomes. **C)** Boxplots showing the variation in ICEs and IMEs' size across multiple phyla. Phyla in the y-axis are listed in descending order of the median value for each boxplot. Only phyla with more than 30 ciMGEs are shown. **D)** Scatter plot with marginal density plots comparing the ciMGE size in bp and log10 scaled and the size of the host genome in Mb. Mean values for each ciMGE type are represented by dashed lines in each density plot. **E)** Scatter plot with marginal density plots comparing the ciMGE GC content (%) and the GC content of the host genome. Mean values for each ciMGE type are represented by dashed lines in each density plot. Bars, boxplots, density curves, regression lines, dashed lines, and correlation coefficients are coloured according to the ciMGE type.

I then explored the distribution of the sequence length and GC content from the ciMGEs identified in this study. The size of ICEs and IMEs was split by phyla, and it was possible to observe that the largest ICEs are found across Desulfobacterota, while the largest IMEs are found in Fusobacteriota (**Figure 1C**). Overall, IMEs follow a normal distribution for the sequence length, with a mean size of 27kb, while AICEs have asymmetrical distributions, with mean sizes of 17kb and 109kb, respectively (**Figure 1D**). Unsurprisingly, given the presence of a full conjugative type IV secretion system on ICEs, these elements tend to be larger than IMEs and AICEs. A weak positive correlation was found between the ICEs/IMEs and host genome size ($R = 0.15$ and p-value < 2.2e-16, $R = 0.058$ and p-value = 4.1e-06, respectively).When it comes to the GC content, while AICEs again have asymmetrical distributions and a mean GC content of 67%, IMEs follow a bimodal distribution and ICEs a multimodal distribution, with mean GC contents of 46% and 48%, respectively (**Figure 1E**). Genome GC content strongly correlates to ICEs, IMEs, and AICEs' GC content, following a trend similar to plasmids (33) ($R \geq 0.8$, p-value < 2.2e-16). A weak positive correlation was also observed between the ICEs/IMEs size and GC content ($R = 0.17$ and p-value < 2.2e-16, $R = 0.035$ and p-value = 0.0052, respectively), while a weak negative correlation was found between the size and GC content of AICEs ($R = -0.15$ and p-value = 5.8e-07, **Figure S2A**). When comparing the ciMGEs' GC content with that of the host genome, IMEs' GC deviation from that of their hosts was significantly higher than the difference observed for ICEs and AICEs (**Figure S2B,** p-value < 2.2e-16). Altogether, these results show that IMEs outnumber ICEs, their GC content is strongly correlated to that of their host, and both elements are widespread across multiple phyla.

Defense systems, AMR genes, and virulence genes are preferentially located in ICEs

Known defense systems are pervasive across bacterial ciMGEs (**Supplementary Table 2**). There are 4588 ciMGEs with at least one defense system gene: a total of 26876 hits in 4586 bacterial ciMGEs, and only 32 hits in 2 archaeal ciMGEs. These hits are dispersed across 23 phyla. Even though the total number of ICEs found in this study is smaller than that of IMEs (5869 vs 6331), these elements are the most important hotspots for the accretion of defense systems across ciMGEs. There are 2519 ICEs with a total of 15897 defense system genes, 1965 IMEs and 103 AICEs with a total of 10619 and 388 genes, respectively (**Figure 2A**). Nearly half of ICEs carry at least one defense system (42,9%, 2519/5869), while nearly a third of IMEs carry at least one of these systems (31,0%, 1965/6331). Given that ICEs and IMEs have a diverse range of sizes across different phyla (**Figure 1D**), I corrected the number of defense system genes to the size of the carrying ciMGE. Proteobacterial ICEs typically carry a large number of defense systems, while IMEs are important vectors for the accretion of defense system genes across Campylobacterota (**Figure 2B**). Restriction-modification systems, abortive

infection, and CBASS are among the 46 defense systems most frequently found across ICEs, while restriction-modification systems, CBASS, and Thoeris represent a large share of the 45 defense systems found on IMEs (**Figure 2C** and **Supplementary Table 3**). Multiple DNA modification systems (DMS) were found across these elements. The DMS model was developed by PADLOC to select potential defense systems where there are two or more genes from modification-based systems such as restriction-modification, BREX, and DISARM. Hits found by the DMS model will require manual curation. A total of 21 defense systems is found on AICEs from Actinobacteriota, including restriction-modification systems, ietAS, and BREX.

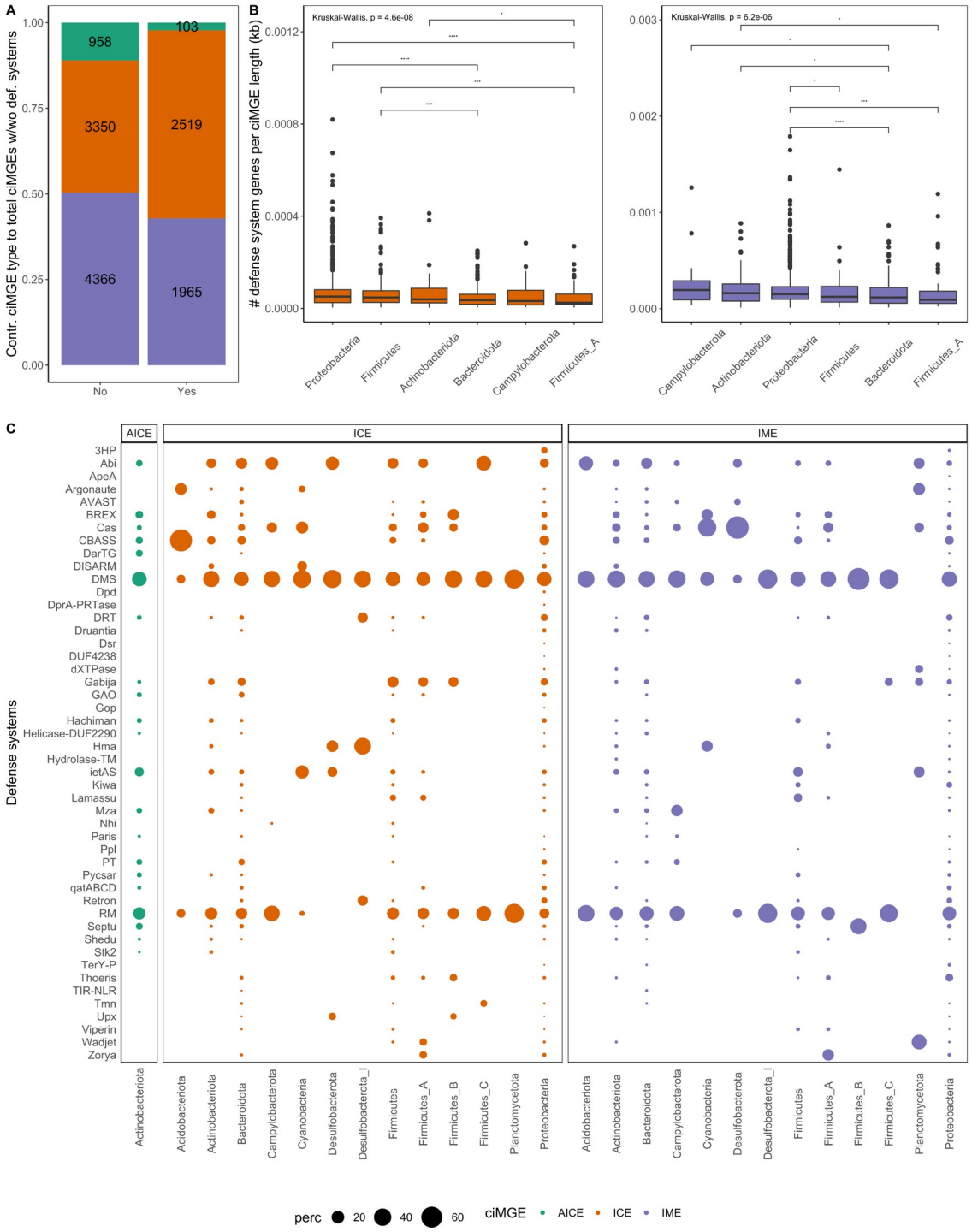

Figure 2. Defense systems are widespread across multiple phyla. A) Relative contribution of each ciMGE type to the total number of ciMGEs with and without defense systems. Absolute counts are shown inside the bars. **B)** Boxplots showing the variation between the number of defense systems normalized by ICEs (left) and IMEs' (right) size (kb). Phyla in the x-axis are listed from left to right in descending order of the median value of each boxplot. Only phyla with more than 10 ciMGEs carrying defense systems are shown. Only statistically significant comparisons are shown above the boxplots. The following convention was used for symbols

indicating statistical significance: * for p <= 0.05, ** for p <= 0.01, *** for p <= 0.001, and **** for p <= 0.0001. **C)** Distribution of defense system genes across multiple phyla. Phyla in the x-axis are listed from left to right in alphabetical order. Only phyla with more than 5 different defense systems are shown. The size of the circles is proportional to the percentage of particular defense systems per ciMGE type per phylum. Bars, boxplots, and circles are coloured according to the ciMGE type.

Known virulence genes are often found across bacterial ciMGEs and are absent from archaeal ciMGEs. ICEs are by far the main vectors for the accretion of these genes across ciMGEs. I found a total of 9243 virulence genes in 1490 bacterial ciMGEs. Of these 1490 ciMGEs, 1218 correspond to ICEs, 267 IMEs, and only 5 AICEs (**Figure 3A**). As expected, the total number of virulence genes in ICEs also far outnumbers that of IMEs – 8695 and 542, respectively. After correcting the number of virulence genes to the size of the carrying ciMGE, I observed that proteobacterial ICEs typically carry a large number of virulence genes normalized to the ICEs' size (**Figure S3A**), while in IMEs a higher number of these genes was found in Campylobacterota (**Figure S3B**). I discovered a total of 13 virulence categories across ICEs, dominated by genes related to nutritional/metabolic factors, exotoxin, immune modulation, and effector delivery system, while out of the 11 virulence categories found on IMEs, the most frequently identified are mostly related to immune modulation, exotoxin, and effector delivery system (**Figure 3B** and **Supplementary Table 4**). The 6 virulence genes found on the 5 AICEs from Actinobacteriota are devoted to effector delivery system and as nutritional/metabolic factors.

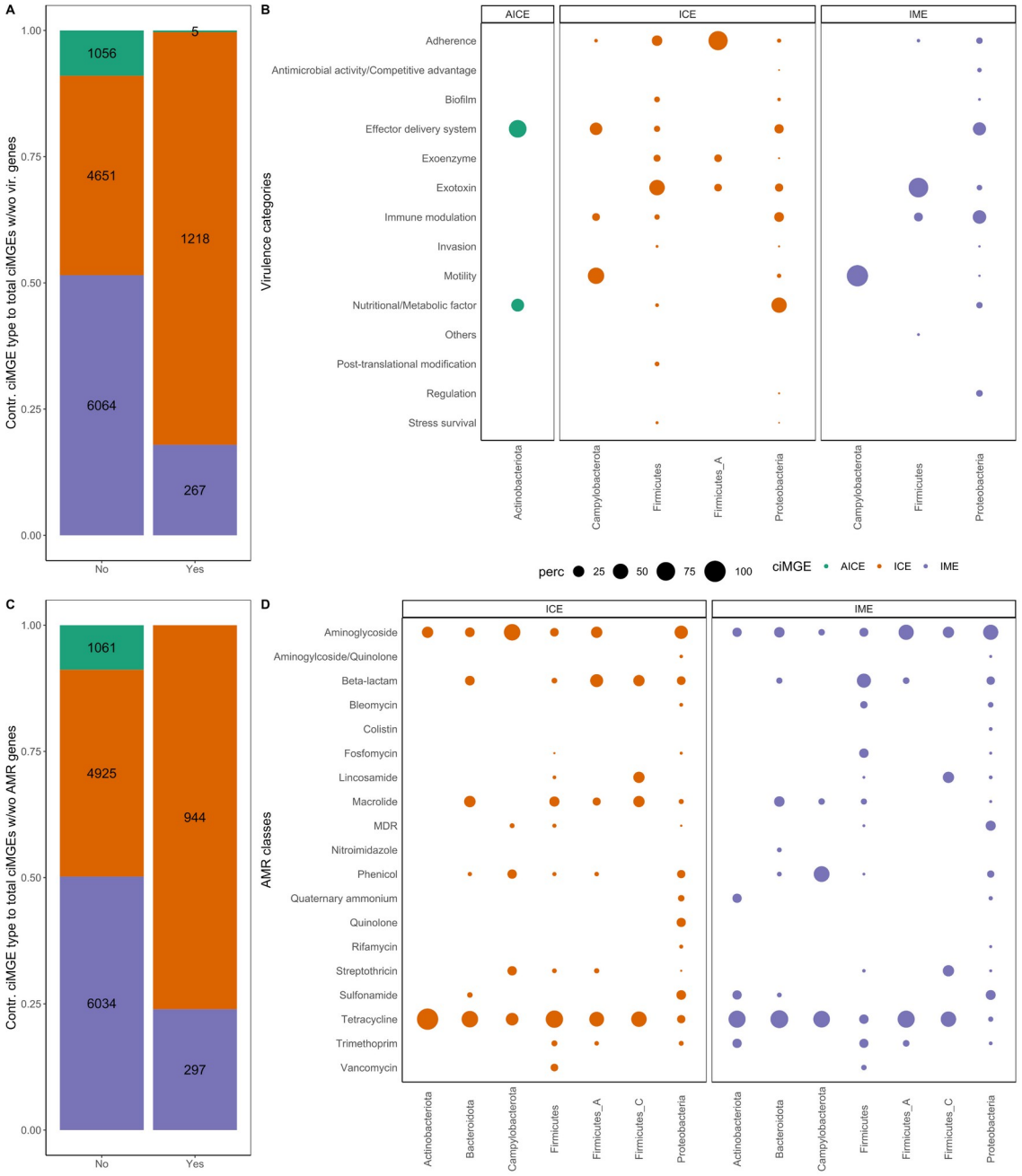

Figure 3. Virulence and AMR genes are more frequently found across ICEs. A) Relative contribution of each ciMGE type to the total number of ciMGEs with and without virulence genes. Absolute counts are shown inside the bars. **B)** Distribution of virulence categories across multiple phyla. Phyla in the x-axis are listed from left to right in alphabetical order. Phyla with ciMGEs carrying genes associated with only one virulence category were excluded. **C)** Relative contribution of each ciMGE type to the total number of ciMGEs with and without AMR genes. Absolute counts are shown inside the bars. **D)** Distribution of AMR classes across multiple phyla. Phyla in the x-axis are listed from left to right in alphabetical order. Only phyla with more than 5 different AMR classes are shown. The size of the circles is proportional to the

percentage of particular virulence categories or AMR classes per ciMGE type per phylum. Bars and circles are coloured according to the ciMGE type.

Known AMR genes are found across bacterial ICEs and IMEs, and absent from archaeal ciMGEs. I discovered a total of 3087 AMR genes in 1241 bacterial ciMGEs. Of these 1241 ciMGEs, 944 correspond to ICEs and 297 to IMEs (**Figure 3C**). This means AMR genes are also more often found in ICEs - 16,1% ICEs carry at least one AMR gene (944/5865), while only 4,7% IMEs carry at least one of these genes (297/6319). The total number of AMR genes was also higher in ICEs than in IMEs - 2394 vs 693, respectively. Curiously, even though the total number of virulence genes in ciMGEs far exceeds that of AMR genes, AMR genes are spread across a wider range of phyla than virulence genes do (12 vs 5, **Figures 3B** and **3D** and **Supplementary Table 2**). After correcting the number of AMR genes to the size of the carrying ciMGE, I observed that ICEs and IMEs in Campylobacterota typically carry a large number of these genes (**Figures S3C** and **S3D**). The AMR genes found across ICEs and IMEs encode resistance to a total of 17 and 18 AMR classes, respectively, and most confer resistance to tetracyclines, aminoglycosides, and beta-lactams (**Figure 3D** and **Supplementary Table 5**). Finally, the distribution of defense systems, AMR, and virulence genes was compared between ICEs and IMEs from different phyla. The number of defense system genes is significantly higher than that of AMR and virulence genes across most ICEs and IMEs from multiple phyla (**Figure S4**). Taken together, these results underline the role of ICEs as important hotspots for the accumulation of defense systems, virulence, and AMR genes across ciMGEs from multiple phyla.

Defense systems, AMR, and virulence genes are negatively correlated across ICEs and IMEs

I then explored to what extent the prevalence of defense systems, AMR classes, and virulence categories is correlated across ICEs/IMEs. Since AICEs carry no AMR genes, the analysis was focused exclusively on ICEs and IMEs. Given that the distribution for each class are not normal, the non-parametric Spearman correlation coefficient was used. These three functional groups are inversely correlated across the ICEs and IMEs identified in this study (**Figure 4**). Genes encoding resistance to multiple antibiotics were positive correlated with GAO, BREX, and Abi defense systems across proteobacterial ICEs (**Figure S5A**). Positive correlations were also found between virulence genes associated with nutritional/metabolic factors and retrons and DRT defense systems. Also, negative correlations were found between virulence genes acting as effector delivery systems and multiple defense systems such as restriction-modification systems

and CBASS. Curiously, different associations were found across ICEs from Firmicutes (**Figure S5B**). For example, positive correlations were found between tetracyclines and viperins, as well as between virulence genes involved in post-translation modification and Lamassu defense systems. Additionally, genes encoding resistance to multiple antibiotics were positively correlated with Kiwa defense systems. Abi and restriction-modification systems were negatively correlated with multiple virulence categories, including exoenzymes and genes involved in adherence functions. When looking into proteobacterial IMEs (**Figure S5C**), genes encoding resistance to distinct antibiotic classes (e.g., beta-lactams, aminoglycosides, and sulphonamides) were positively correlated, consistent with the previous observations that these genes tend to be co-localized in genetic blocks named integrons (34). Negative correlations were found between multiple defense systems and virulence genes involved in immune modulation. Finally, Lamassu defense systems were positively correlated with genes encoding resistance to aminoglycosides and virulence genes involved in immune modulation across IMEs from Firmicutes (**Figure S5D**). Additionally, negative correlations were found between beta-lactams and exotoxins and restriction-modification systems. These results show that while defense systems, AMR, and virulence genes are inversely correlated across ICEs and IMEs, positive and negative correlations between specific genes belonging to these functional groups can be observed in both ciMGEs.

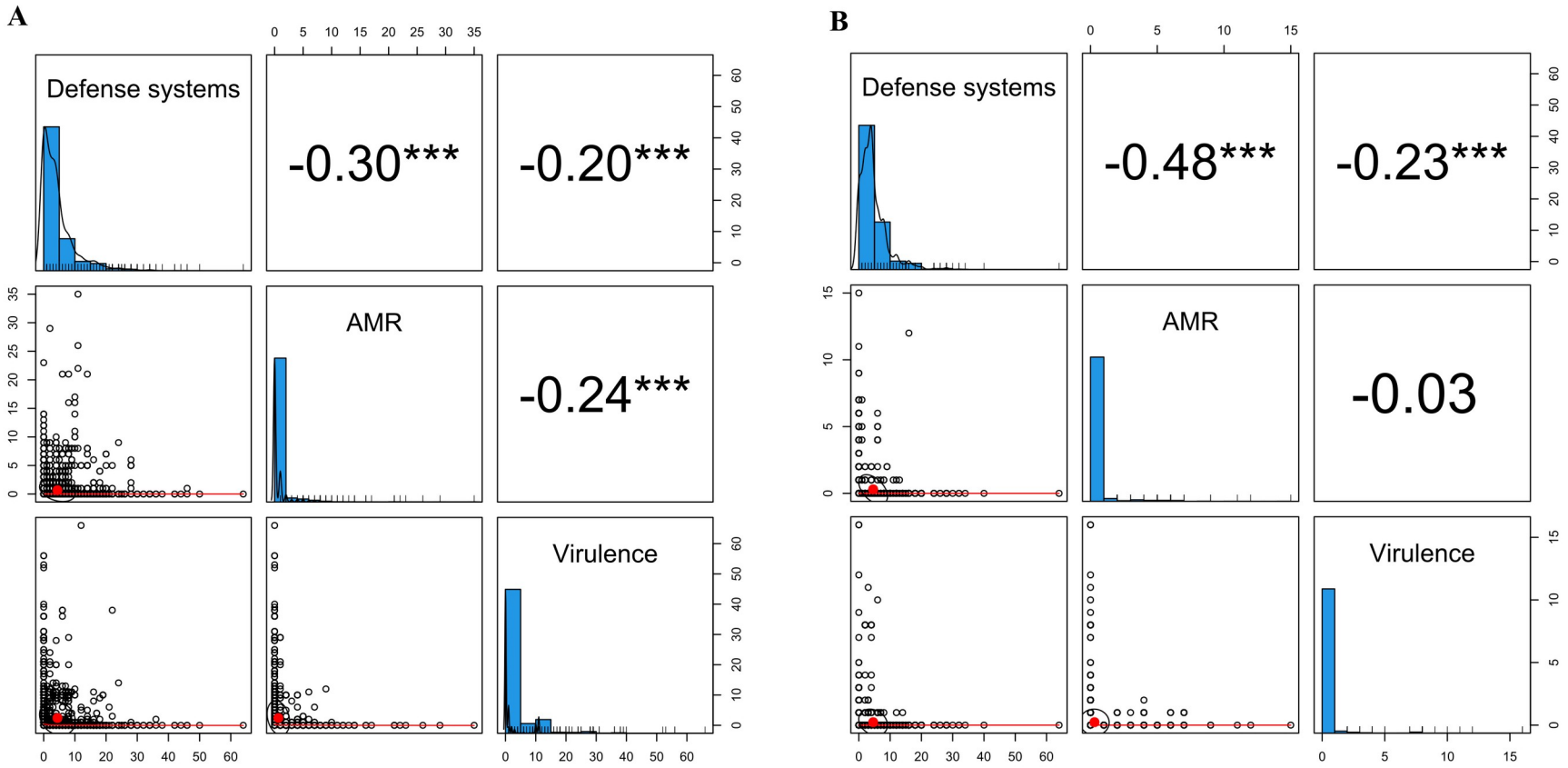

Figure 4. Defense systems, virulence, and AMR genes are negatively correlated. A) Correlations across ICEs. **B)** Correlations across IMEs. The scatter plot of matrices shows bivariate scatter plots below the diagonal, histograms on the diagonal, and the Spearman correlations above the diagonal. Values above 0.05 were considered as non-significant and no asterisks are shown. The following convention was used for symbols indicating statistical significance: * for p <= 0.05, ** for p <= 0.01, *** for p <= 0.001, and **** for p <= 0.0001.

ICEs and IMEs form heterogeneous communities composed of multiple phyla

Given the presence of highly similar MGEs in this dataset, the 13274 elements found here were dereplicated into a representative set of 9618 ciMGEs (nucleotide identity threshold of 90% and 80% coverage). Each ciMGE was then reduced to a set of *k*-mers and the Jaccard index (JI) was used as a measure of nucleotide sequence similarity between all ciMGE pairs. In line with the high diversity frequently observed across MGEs, the majority of ciMGE pairs shared little similarity, with JI values below 0.25 (**Figure S6**). Next, I used an alignment-free nucleotide sequence similarity comparison between the representative set of 9618 ciMGE pairs to infer an undirected and weighted network (**Figures 5A and 5B**). A total of 14 communities (i.e., set of nodes that are more densely connected with one another than expected by chance) were detected with Infomap. AICEs from Actinobacteria were mostly clustered in two homogeneous communities, while ICEs and IMEs typically formed heterogeneous communities dominated by either Proteobacteria or Firmicutes. Interestingly, ICEs and IMEs from Bacteroidota and Campylobacterota were also present in these heterogeneous communities. The absence of pairwise distance similarities with intermediate JI (**Figure S6**) helps to explain this clustering in discrete communities, instead of a continuous genetic structure. Altogether, these results underline the broad host range of ICEs and IMEs and their ability to challenge interphylum barriers and to form heterogeneous communities with elements from multiple phyla.

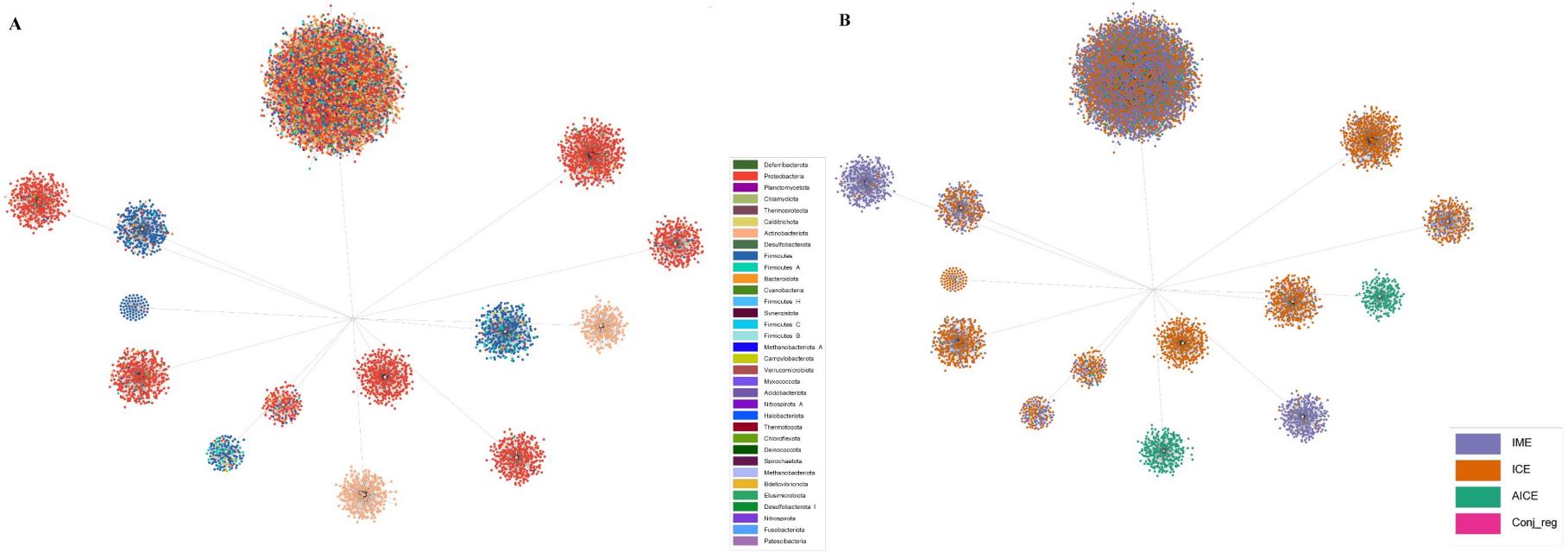

Figure 5. Network of ciMGE communities detected with Infomap. Nodes coloured by **A)** Phylum; and **B)** ciMGE type. Each ciMGE is represented by a node, connected by edges according to the mutation distances between all ciMGE pairs.

The functional landscape of ICEs is populated by uncharacterized proteins

The proteome of the dereplicated dataset of ciMGEs from both Bacteria and Archaea (total number of proteins = 629300) was then scanned for cluster or orthologous groups (COGs) categories. Given that ICEs are usually larger than IMEs (**Figure 1D**), the absolute counts of COG categories found on each ICE and IME (**Supplementary Table 6**) was corrected to the size of the corresponding ICE and IME, respectively. For the majority of COG categories, the normalized number of IME proteins with an assigned function was significantly higher than that of ICE proteins (p<2.2e-16, **Figure 6**). This difference was consistently observed across the phyla with a higher incidence of both ICEs and IMEs: Proteobacteria, Firmicutes, Actinobacteriota, Bacteroidota, Campylobacterota, and Firmicutes_A (**Figure S7**). Interestingly, the only COG category that was significantly more assigned on ICE proteins was the category S, which refers to unknown functions (p<2.2e-16, **Figure 6**).

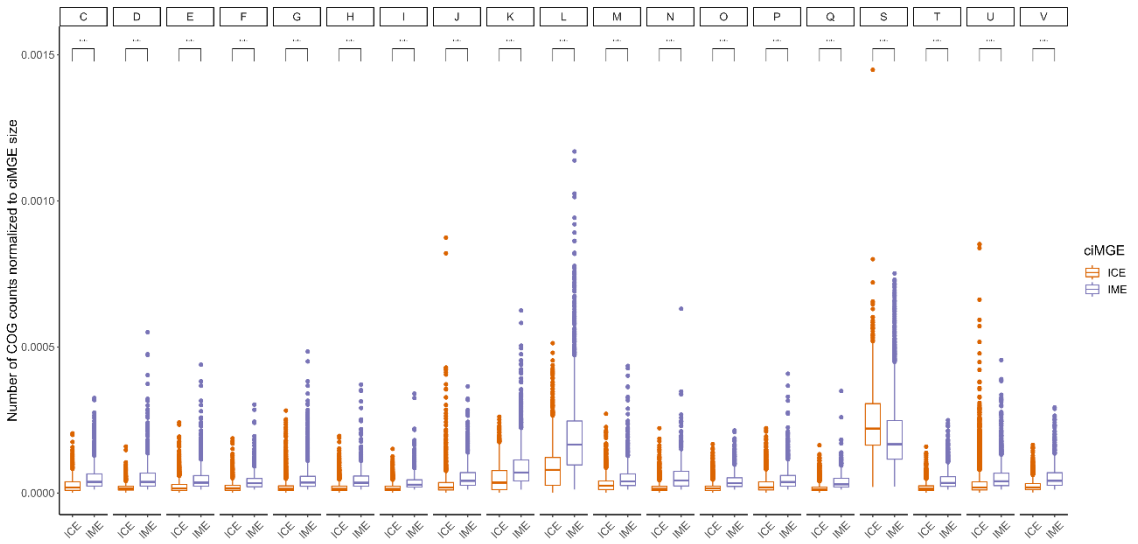

Figure 6. The normalized number of unknown functions is significantly higher across ICEs. Boxplots showing the number of COG categories found on each ICE or IME from the six major phyla found in this study and normalized to the ICEs and IMEs' size. Comparisons were performed using the Wilcoxon test, and the p-values adjusted with the Holm–Bonferroni method. The following convention was used for symbols indicating statistical significance: * for p <= 0.05, ** for p <= 0.01, *** for p <= 0.001, and **** for p <= 0.0001. Boxplots are coloured according to the ciMGE tpye. COG categories: C - Energy production and conversion; D - Cell cycle control, cell division, chromosome partitioning; E - Amino acid transport and metabolism; F - Nucleotide transport and metabolism; G - Carbohydrate transport and metabolism; H - Coenzyme transport and metabolism; I - Lipid transport and metabolism; J - Translation, ribosomal structure and biogenesis; K – Transcription; L - Replication,

recombination and repair; M - Cell wall/membrane/envelope biogenesis; N - Cell motility; O - Posttranslational modification, protein turnover, chaperones; P - Inorganic ion transport and metabolism; Q - Secondary metabolites biosynthesis, transport and catabolism; S - Function unknown; T - Signal transduction mechanisms; U - Intracellular trafficking, secretion, and vesicular transport; V - Defense mechanisms. COG categories A (RNA processing and modification), B (Chromatin structure and dynamics), W (Extracellular structures), and Z (Cytoskeleton) were excluded due to low COG counts.

Discussion

In this study, I characterize a collection of more than 13000 ciMGEs, including more than 6000 IMEs, nearly 6000 ICEs, and more than 1000 AICEs. Comparing with the number of ciMGEs with nucleotide sequences available at ICEberg 2.0 (https://bioinfo-mml.sjtu.edu.cn/ICEberg2/download.html), this work represents a massive increase in the number of publicly available IMEs (6331 vs 111), ICEs (5869 vs 718), and AICEs (1061 vs 50). These large discrepancies are somehow anticipated, since the number of ciMGEs available at ICEberg was last updated in 2018, and the number of bacterial and archaeal genomes have been skyrocketing ever since (35). On top of that, ciMGEs characterized in this study were corrected for taxonomy of the host genome, based on the rank-normalized classification module from domain to species available at the Genome Database Taxonomy (36). The type of ciMGE (i.e., either IME, ICE, or AICE) strongly impacts the distribution of sequence length and GC content of these elements. The left-skewed distribution observed for AICE's GC content can be explained by the high GC content of the bacterial host. In fact, the vast majority of AICEs were identified in Actinobacteriota, which typically have high GC content (23). On the other hand, the multimodal distribution of ICEs/IMEs' GC content can be explained by the wide distribution of these elements across multiple phyla with variable GC content.

Recently, it was shown that mobile genes and defense systems are non-randomly clustered in genomic islands (2), but the type of ciMGEs involved in this process was not assessed. Here, I show that a wide repertoire of defense systems is accumulated across ICEs and IMEs from multiple phyla. Crucially, I found that defense systems, AMR genes, and virulence genes are inversely correlated across bacterial ICEs and IMEs, suggesting that carrying multiple cargo genes is detrimental to bacterial fitness. While defense systems and AMR genes were identified across ciMGEs from multiple phyla, virulence genes were limited to five phyla (**Figure 3B**). This can in part be explained by the inclusion of only 32 genera of pathogens with medical

importance in VFDB with full information available (27), meaning that virulence genes from multiple phyla were most likely missed from this analysis.

It was also found that a higher relative proportion of IME proteins have known functions across the most common bacterial phyla identified in this study (**Figures 6** and **S7**). On the other hand, ICEs have more uncharacterized proteins, which suggests the existence of a pool of unknown genes with functions yet to be discovered. Curiously, no hits for COG category X (mobilome: prophages, transposons) were found in this study. Since ICEs are also known as conjugative transposons and share genetic features with prophages (such as the presence of phage integrases), it would be expected to find protein hits for this COG category. This can be explained since COGs that are assigned to both L and X categories according to NCBI's Database of COGs (such as COG0582 – integrase/recombinase) were solely assigned to L category using eggNOG-mapper. In fact, COG0582 was the most commonly observed COG across category L, which helps to explain the high values for the normalization counts of this category across this dataset.

The network-based approach used in this work revealed that most ciMGEs form clusters of high nucleotide identity that are homogeneous to the host phylum, in agreement with the phylogenetic and biological barriers that shape HGT events (37). Still, distantly related phylum interactions were observed across multiple ciMGEs, for example between Campylobacterota, Actinobacteriota, and Firmicutes (**Figure 5A**), in line with the broad host range attributed to these elements (38). Interestingly, the dereplication approach used to build the ciMGE network (using a 90% nucleotide identity threshold) removed no elements from Archaea, uncovering that ciMGEs in this domain do not share high nucleotide identity. Indeed, only two elements (of the 16 ciMGEs found in archaeal genomes from this dataset) are plotted in the network and form a small cluster exclusively containing these ciMGEs (**Figure 5A**), meaning the JIs including most archaeal ciMGEs fall below the mean threshold for all pairwise comparisons between Bacteria and Archaea, and exposing how distantly related these elements are.

There are several bioinformatic tools available to search for extrachromosomal elements such as plasmids, however there are currently few alternatives for ciMGEs as ICEs and IMEs. Recently, ICEscreen was developed (39), with the purpose of detecting these elements across Firmicutes. I decided to use ICEfinder (23), since it is not restricted to a particular phylum. Even though this tool was designed to scan ciMGEs across bacterial genomes, it was able to identify 16 elements in archaeal genomes. Still, it is possible some elements may have been overlooked, since integrases, relaxases, and other signature proteins from bacterial ciMGEs may be too distantly related to those from Archaea. Conceptually, prophages would fit the idea of a chromosomally integrated MGE but were not considered in this study because multiple DNA sequences of these elements are already available in public databases. This study only focused on genomes that

were sequenced at the complete level, meaning the prevalence of ciMGEs across genomes sequenced at the scaffold and contig level was not assessed. Even though genomes sequenced at this level exceed the number of complete genomes by orders of magnitude (35), focusing on the latter was crucial to accurately delineate ICEs and IMEs across 'intact' chromosomes. Following the same rationale, metagenome-assembled genomes were also not included in this study. Still, studying the distribution of ICEs and IMEs in relevant reservoirs such as the human microbiome is crucial to better understand the eco-evolutionary dynamics that shape the acquisition of defense systems, AMR, and virulence genes in complex communities (40).

To conclude, this work represents a massive increase in the number of ciMGEs currently available in public databases (from <1000 to >13000). I found that IMEs outstrip ICEs, and both are prevalent across different phyla. ICEs represent the most important ciMGEs for the accumulation of defense systems, virulence genes, and antimicrobial resistance genes. Furthermore, I discovered that these genes are inversely correlated across ICEs and IMEs. Multiple representatives of these two elements share high genetic similarity and challenge phylogenetic barriers. Overall, I discovered that the functional landscape of ICEs is populated by proteins with unknown functions. Finally, this study offers a thorough catalog of ciMGEs from 34 different phyla in the bacterial and archaeal domains, including their nucleotide sequence and associated metadata.

Data availability

Analyses were made with a combination of shell and RStudio v2022.07.1 scripting. Code used to reproduce major analysis and figures is available at the Gitlab repository https://gitlab.gwdg.de/botelho/ices_imes. The nucleotide sequences and associated metadata of the ciMGEs identified in this study are available at the Figshare project https://doi.org/10.6084/m9.figshare.21583413.v1.

Acknowledgments

I would like to thank Jaime Iranzo and Adrian Cazares for relevant discussions while preparing this manuscript.

Funding

João Botelho is supported by the Maria Zambrano grant (UP2021-035), and the Severo Ochoa Program for Centres of Excellence in R&D from the Agencia Estatal de Investigacion of Spain [CEX2020-000999-S (2022–2025)].

Conflict of interest

The author declares no competing interests.

References

1. Ghaly,T.M. and Gillings,M.R. (2018) Mobile DNAs as Ecologically and Evolutionarily Independent Units of Life. *Trends Microbiol.*, **26**, 904–912.

2. Makarova,K.S., Wolf,Y.I., Snir,S. and Koonin,E. V. (2011) Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *J. Bacteriol.*, **193**, 6039–6056.

3. Koonin,E. V. (2017) Evolution of RNA- and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol. Direct*, **12**.

4. Puigbò,P., Makarova,K.S., Kristensen,D.M., Wolf,Y.I. and Koonin,E. V. (2017) Reconstruction of the evolution of microbial defense systems. *BMC Evol. Biol.*, **17**.

5. Makarova,K.S., Wolf,Y.I. and Koonin,E. V. (2009) Comprehensive comparative-genomic analysis of Type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biol. Direct*, **4**, 1–38.

6. Rocha Id,E.P.C. and Id,D.B. (2022) Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLOS Biol.*, **20**, e3001514.

7. Igler,C., Schwyter,L., Gehrig,D. and Wendling,C.C. (2022) Conjugative plasmid transfer is limited by prophages but can be overcome by high conjugation rates. *Philos. Trans. R. Soc. B*, **377**.

8. Hall,J.P.J., Wood,A.J., Harrison,E. and Brockhurst,M.A. (2016) Source–sink plasmid transfer dynamics maintain gene mobility in soil bacterial communities. *Proc. Natl. Acad. Sci.*, **113**, 8260–8265.

9. Tesson,F., Hervé,A., Mordret,E., Touchon,M., d'Humières,C., Cury,J. and Bernheim,A. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun. 2022 131*, **13**, 1–10.

10. Jaskólska,M., Adams,D.W. and Blokesch,M. (2022) Two defence systems eliminate plasmids from seventh pandemic Vibrio cholerae. *Nat. 2022*, 10.1038/s41586-022-04546-y.

11. Hussain,F.A., Dubert,J., Elsherbini,J., Murphy,M., VanInsberghe,D., Arevalo,P., Kauffman,K., Rodino-Janeiro,B.K., Gavin,H., Gomez,A., *et al.* (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science (80-.).*, **374**, 488–492.

12. LeGault,K.N., Hays,S.G., Angermeyer,A., McKitterick,A.C., Johura,F., Sultana,M., Ahmed,T., Alam,M. and Seed,K.D. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science (80-.).*, **373**, eabg2166.

13. Johnson,C.M. and Grossman,A.D. (2015) Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annu. Rev. Genet.*, **49**, 577–601.

14. Thoma,L. and Muth,G. (2015) The conjugative DNA-transfer apparatus of Streptomyces. *Int. J. Med. Microbiol.*, **305**, 224–229.

15. Guédon,G., Libante,V., Coluzzi,C., Payot,S. and Leblond-Bourget,N. (2017) The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes (Basel).*, **8**, 337.

16. Botelho,J. and Schulenburg,H. (2021) The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. *Trends Microbiol.*, **29**, 8–18.

17. Pinilla-Redondo,R., Russel,J., Mayo-Muñoz,D., Shah,S.A., Garrett,R.A., Nesme,J., Madsen,J.S., Fineran,P.C. and Sørensen,S.J. (2021) CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Res.*, **1**, 13–14.

18. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science (80-.).*, **359**, eaar4120.

19. Ellabaan,M.M.H., Munck,C., Porse,A., Imamovic,L. and Sommer,M.O.A. (2021) Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. *Nat. Commun.*, **12**, 2435.

20. Jiang,X., Hall,A.B., Arthur,T.D., Plichta,D.R., Covington,C.T., Poyet,M., Crothers,J., Moses,P.L., Tolonen,A.C., Vlamakis,H., *et al.* (2019) Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*, **363**, 181–187.

21. He,J., Baldini,R.L., Déziel,E., Saucier,M., Zhang,Q., Liberati,N.T., Lee,D., Urbach,J., Goodman,H.M. and Rahme,L.G. (2004) The broad host range pathogen Pseudomonas aeruginosa strain PA14 carries two pathogenicity islands harboring plant and animal virulence genes. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 2530–5.

22. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2022) GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database. *Bioinformatics*, 10.1093/BIOINFORMATICS/BTAC672.

23. Liu,M., Li,X., Xie,Y., Bi,D., Sun,J., Li,J., Tai,C., Deng,Z. and Ou,H.-Y. (2018) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, 10.1093/nar/gky1123.

24. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

25. Payne,L.J., Todeschini,T.C., Wu,Y., Perry,B.J., Ronson,C.W., Fineran,P.C., Nobrega,F.L. and Jackson,S.A. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res.*, 10.1093/NAR/GKAB883.

26. Feldgarden,M., Brover,V., Haft,D.H., Prasad,A.B., Slotta,D.J., Tolstoy,I., Tyson,G.H., Zhao,S., Hsu,C.-H., McDermott,P.F., *et al.* (2019) Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob. Agents Chemother.*, **63**.

27. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.

28. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol. 2017 3511*, **35**, 1026–1028.

29. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.*, 10.1093/MOLBEV/MSAB293.

30. Zhao,X. (2019) BinDash, software for fast genome distance estimation on a typical personal laptop. *Bioinformatics*, **35**, 671–673.

31. Singhal,A., Cao,S., Churas,C., Pratt,D., Fortunato,S., Zheng,F. and Ideker,T. (2020) Multiscale community detection in Cytoscape. *PLOS Comput. Biol.*, **16**, e1008239.

32. Rosvall,M. and Bergstrom,C.T. (2008) Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 1118–1123.

33. Almpanis,A., Swain,M., Gatherer,D. and McEwan,N. (2018) Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb. genomics*, **4**.

34. Ghaly,T.M., Geoghegan,J.L., Tetu,S.G. and Gillings,M.R. (2020) The Peril and Promise of Integrons: Beyond Antibiotic Resistance. *Trends Microbiol.*, **28**, 455–464.

35. Koonin,E. V., Makarova,K.S. and Wolf,Y.I. (2021) Evolution of Microbial Genomics: Conceptual Shifts over a Quarter Century. *Trends Microbiol.*, 10.1016/j.tim.2021.01.005.

36. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.

37. Popa,O., Hazkani-Covo,E., Landan,G., Martin,W. and Dagan,T. (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res.*, **21**, 599–609.

38. Cury,J., Oliveira,P.H., de la Cruz,F. and Rocha,E.P.C. (2018) Host Range and Genetic Plasticity Explain the Coexistence of Integrative and Extrachromosomal Mobile Genetic Elements. *Mol. Biol. Evol.*, **35**, 2230–2239.

39. Lao,J., Lacroix,T., Gú Edon,G., Coluzzi,C., Payot,S., Leblond-Bourget,N. and Eì Ene Chiapello,H.′ (2022) ICEscreen: a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures. *NAR Genomics Bioinforma.*, **4**.

40. Sabino,Y.N.V., Santana,M.F., Oyama,L.B., Santos,F.G., Moreira,A.J.S., Huws,S.A. and Mantovani,H.C. (2019) Characterization of antibiotic resistance genes in the species of the rumen microbiota. *Nat. Commun.*, **10**, 5252.