Phage-plasmid hybrids are found throughout diverse environments and encode niche-specific functional traits

Mullet, J.[1,3], Zhang, L.[2], Pruden, A.,[1*] Brown, C.L.[1*]
[1]Department of Civil and Environmental Engineering, Virginia Tech
[2]Department of Computer Science, Virginia Tech
[3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology
[*]co-corresponding authors

ABSTRACT

Phage-plasmids are unique mobile genetic elements that function as plasmids and temperate phages. While it has been observed that such elements often encode antibiotic resistance genes and defense system genes, little else is known about other functional traits they encode. Further, no study to date has documented their environmental distribution and prevalence. Here, we performed genome sequence mining of public databases of phages and plasmids utilizing a random forest classifier to identify phage-plasmids. We recovered 5,742 unique phage-plasmid genomes from a remarkable array of disparate environments, including human, animal, plant, fungi, soil, sediment, freshwater, wastewater, and saltwater environments. The resulting genomes were used in a comparative sequence analysis, revealing functional traits/accessory genes associated with specific environments. Host-associated elements contained the most defense systems (including CRISPR and anti-CRISPR systems) as well as antibiotic resistance genes, while other environments, such as freshwater and saltwater systems, tended to encode components of various biosynthetic pathways. Interestingly, we identified genes encoding for certain functional traits, including anti-CRISPR systems and specific antibiotic resistance genes, that were enriched in phage-plasmids relative to both plasmids and phages. Our results highlight that phage-plasmids are found across a wide-array of environments and likely play a role in shaping microbial ecology in a multitude of niches.

IMPORTANCE

Phage-plasmids are a novel, hybrid class of mobile genetic element which retain aspects of both phages and plasmids. However, whether phage-plasmids represent merely a rarity or are instead important players in horizontal gene transfer and other important ecological processes has remained a mystery. Here, we document that these hybrids are encountered across a broad range of distinct environments and encode niche-specific functional traits, including the carriage of

antibiotic biosynthesis genes and both CRISPR and anti-CRISPR defense systems. These findings highlight phage-plasmids as an important class of mobile genetic element with diverse roles in multiple distinct ecological niches.

INTRODUCTION

Vehicles of horizontal gene transfer (HGT), such as plasmids and phages, are key drivers of prokaryotic adaptation and evolution (1, 2). In this regard, their role in the mobility of accessory genes, i.e., genes that are not required for the basic life cycle of a mobile genetic element (MGE), is of particular interest (2, 3). MGEs can carry accessory genes encoding diverse traits that may be advantageous to their hosts, including antibiotic resistance genes (ARGs), virulence factors, defense systems such as CRISPR-Cas, metal resistance genes (MRGs), and toxin-antitoxin systems, among many others (3). Such genes can provide hosts with resiliency in the face of changing selective pressures. While MGEs are typically categorized as independent classes (2,4), there is an emerging awareness of inter-element conflicts that can occur between MGEs within individual bacterial hosts (5,6,7,8). For example, some phages, plasmids, and integrative and conjugative elements carry genes encoding defense systems that interfere with the function of co-infecting MGEs (9). Prokaryotic defense systems like these are hypothesized to be acquired through selective bacteriophage predation and have been demonstrated to cluster with and potentially increase the spread of ARGs (10, 11). The carriage of defense systems by MGEs can result in complex ecological and evolutionary dynamics within their host and can significantly alter the community dynamics of microbial populations (9, 10).

Phage-plasmids (P-Ps) are a newly characterized hybrid class of MGE that occupy a unique place in the landscape of prokaryotic genomic elements. These elements can be generally described as temperate (i.e., integrated) phages that retain the ability to replicate in a plasmid-like manner as extra-chromosomal DNA as part of their host life cycle (12). A small set of P-Ps have been shown experimentally to employ a unique combinatorial replication strategy, leveraging both phage lysis and reinfection and the multi-copy number potential of plasmids (12,14). Additionally, P-Ps have been shown to transfer ARGs, certain defense systems, and additional accessory genes from both phages and plasmids (13, 14, 15). With supporting research indicating that P-Ps are significant promoters of genetic exchange between phages and plasmids, the composition and diversity of their accessory genomes remains a key knowledge gap (15). Their unique biology makes the question of their accessory genome particularly intriguing, with the potential for distinct infection and HGT strategies. P-Ps thus represent a distinct class of MGE, a poorly understood dimension of microbial community dynamics, and, hypothetically, a new mechanism of transfer for accessory genes such as AMR or CRISPR-Cas systems.

However, to date, the environmental distribution of P-Ps has not been determined. Indeed, whether P-Ps are common features of microbial communities or merely rare oddities that emerge in specific niches has not yet been ascertained. This limited examination into P-P biological diversity becomes critical to understand as phages and plasmids independently possess unique functional variation across different environments (16, 17). Understanding the diversity of P-Ps across these environments can provide improved insights into the impacts and potential interactions these elements have in the genetic exchange of accessory genes between phages and plasmids.

Here, we recovered a unique dataset of 5,742 P-Ps across four public databases of plasmids and phages and found that they are remarkably prolific across a diverse array of environments. Examination of P-P accessory genome contents suggests a strong linkage to niche-specific ecological dynamics. Compilation and annotation of P-P genomes herein expands knowledge of their genomic diversity and provides new insight into their unique biological and ecological function.

RESULTS

P-P hybrids are prolific in public databases of phages and plasmids

We analyzed 1,179,858 genomes from databases of plasmids and phages (PLSDB (19), GPD (20), MGV (21), and IMG/VR (22)) for putative P-Ps using a random forest classifier. The features of the model included the number of hallmark protein hits to each class of MGE (bacteriophage, plasmid, integrative elements, insertion sequences, and multiple), the associated mobileOG-db major categories for each protein, and the number of total proteins and open reading frames for each genome (Supplementary Methods; Table S1) and were trained on (10,289 genomes from [Pfiefer et al.])(12, 18). The final classifier demonstrated an accuracy/FPR/FNR of 95.4%, 2.9%, 19.7%, respectively and was found to outperform manual assignment based on proportions of phage and plasmid-associated gene content (Supplementary Methods; Table S1). This model was employed to generate a conservative, high-confidence set of P-Ps, which was especially relevant because of our usage of IMG/VR v4.0, a database of phage genomes derived primarily from environmental metagenomes (22).

The final P-P dataset examined in this study was composed of 5,742 dereplicated genomes with 137 from GPD, 13 from MGV, 4,425 from IMG/VR, and 1,167 from PLSDB (Fig 2A) (19, 20, 21, 22). PLSDB was predicted to contain several phage genomes (0.8% of PLSDB sequences) and phage databases, such as IMG/VR, were found to harbor many plasmid sequences (Fig 2A) (19, 22). This is not necessarily surprising, as accuracy of plasmid and phage identification can be affected by both low-quality annotated databases and the inherent bias of tools and datasets that specifically classify only one type of MGE. Prior studies have shown that plasmid classification tools can be prone to misidentifying phages as plasmids and, likewise, phage identification tools sometimes misidentify plasmids as phages (18, 28). These inherent biases of analyses targeting a single class of MGE highlight the value of predicting multiple MGE classes simultaneously.

Metadata across the P-P set was harmonized to group P-Ps according to the environment from which the original sample was sourced: terrestrial (n = 689); aquatic (n = 1,868); host-associated (n = 2,105); and unclassified (n = 1,080) (Supplementary Methods; Table S1). Comparative analysis of mobileOGs (i.e., MGE hallmark genes) highlighted distinct profiles of gene content across phages, plasmids, and P-Ps (Fig. 2B). These profiles were consistent with expectations in that P-Ps encoded more phage genes than plasmids (median 88 genes vs. 8 genes; $p < 0.001$); more plasmid genes than phages (55 genes vs. 0 genes; $p < 0.001$); and more total genes than both phages and plasmids (179 P-P genes vs. 46 phage genes vs 19 plasmid genes; $p < 0.001$) (Fig. 2B). In addition, P-Ps were found to have larger average genome sizes than either phages or plasmids, as has been reported previously in studies that examined a smaller dataset of P-Ps (Supplementary Methods; Fig. S3, 12).

P-Ps are associated with disparate hosts and ecological niches

Examining the putative hosts of P-Ps can provide insight into the ecology of P-Ps across distinct environmental niches. A compilation of source database metadata was used in tandem with sequence analysis to identify predicted host taxonomy, plasmid incompatibility groups, phage morphology, and the source environment of the P-P genomes. Only 820 (14.3%) genomes were able to be placed within archived plasmid incompatibility groups, likely due to underrepresentation of incompatibility groups beyond *Enterobacteriaceae* in reference databases (30) (Supplementary

Methods; Table S1). A putative viral taxonomic classification was obtained for 5,182 genomes classified into viral taxonomic families using geNOMAD (19). The bacterial host taxonomy was obtained with 58.8% of P-Ps (n=3,371) receiving a phylum-level classification (Supplementary Methods; Table S1).

We next investigated the prokaryotic hosts of P-Ps across different environments. The most commonly predicted bacterial host phyla across all environments were *Pseudomonadota* and Firmicutes. The aquatic P-Ps possessed the highest diversity in predicted prokaryotic host phyla, including several phyla (*Verrucomicrobia*, *Crenarchaeota*, and *Euryarchaeota*) exclusively associated with aquatic P-Ps (Fig 3). Further examination revealed differences in the class-level taxonomy of the P-P bacteria. Within the *Pseudomonadota* phylum, *Gammaproteobacteria* was the most common predicted bacteria class, particularly in host-associated P-Ps (95.1% host-associated P-Ps, 76.3% terrestrial P-Ps, and 56.5% of aquatic P-Ps). *Alphaproteobacteria* and *Betaproteobacteria* classes were associated with more terrestrial and aquatic P-Ps (4.8% host-associated P-Ps, 23.6% terrestrial P-Ps, and 46.2% of aquatic P-Ps). The terrestrial P-Ps in *Gammaproteobacteria* were primarily from the *Pseudomonadales* order, while aquatic P-Ps were affiliated with a broader array of bacterial carriers (Fig 3). Host-associated P-Ps were predominately carried by *Enterobacteriaceae* (71.2% of *Pseudomonadota* -associated hosts), a family that includes many enteric Gram negatives of clinical relevance, such as *Escherichia*, *Salmonella*, and *Shigella* (Fig. 3)(31). The *Enterobacteriaceae* bearing P-Ps were more frequently found among the host-associated P-Ps compared to both the aquatic (22.9%) and terrestrial (25.0%) *Pseudomonadota* bearing P-Ps (Fig 3).

We next examined taxonomy of the P-Ps themselves. An analysis of the updated ICTV family classifications and plasmid incompatibility groups was performed using geNOMAD and PlasmidFinder, respectively (28, 30, 32). *Caudoviricetes* represented the dominant viral order across all environments, with 1.6% of aquatic P-Ps assigned classifications from *Megaviricetes* – an order containing giant viruses (33). At the family-level, few P-Ps could be classified using geNOMAD, but it was noted that the most frequently detected viral family was *Kyanoviridae* (n=70), which was only found among aquatic P-Ps (28). The plasmid incompatibility groups were similar across different environments, with IncFIB, IncY, and p0111 being the most common classifications.

P-Ps encode diverse and niche-specific accessory functions

The broad distribution of P-Ps across disparate environments led us to question what functional traits P-Ps might carry across a correspondingly wide variety of ecological niches. We next investigated the accessory genome of P-Ps, including ARGs, metabolism-related genes, metal resistance genes, defense systems, toxin-antitoxin systems, anti-CRISPR systems, and virulence factors.

Accessory gene content of P-Ps was relatively unchanged within each of the distinct environments from which the P-Ps originated (Fig. 4). When comparing P-Ps to phage and plasmid accessory genes, P-P accessory gene profiles were most similar to those of plasmids (Kruskal-Wallis and post hoc Dunn test; p=5.30 x 10^{-1}), with very few accessory genes found among phages relative to plasmids and P-Ps (Kruskal-Wallis and post hoc Dunn test; p= 1.15x 10^{-9}) (Fig. 4). However, it was noted that the P-Ps had enriched anti-CRISPR genes compared to phages and plasmids (Fischer exact test; 240 P-P genes vs. 5 phage genes vs 1 plasmid genes; $p < 0.001$). While most ARGs, MRGs, and virulence factors likely predominately originated from plasmid sources, it is also possible that phages still contribute to certain metabolism and defense system accessory genes among P-Ps.

Diversity within the unique accessory genomes of phage-plasmids

We sought to further characterize the diversity among the unique P-P accessory genomes and to assess additional differentiating features and trends among their profiles. First, we analyzed the differences between P-P and plasmid ARG gene distributions. Similar to prior research, it was noted that P-Ps possess ARGs less frequently than plasmids. However, some ARGs, including cpxA, EcoI_emrE, and CTX-M-142, were enriched in P-Ps compared to plasmids (Fischer exact test; $p < 0.01$) (Supplementary Methods; Fig. S6). We found that several of the most common ARGs are associated with Class I integrons, including sul1, aadA2, and qacEdelta1 (Fig. 5a). While approximately 5% of the host-associated P-Ps contained ARGs, the aquatic and terrestrial environment phage-plasmids appeared to be more depleted in the number of ARGs (38). It was noted that the P-Ps associated with wastewater environments contained a few ARGs possessing the *blaCTX-M-15* gene, which is one of the most common extended-spectrum beta-lactamase (ESBL)-encoding ARGs found to be associated with infections that are resistant to third-generation cephalosporins (37). Through the visualization of genetic contexts surrounding CTX-M-15, we found a conserved region that was encountered in P-Ps encountered across several examined source environments (Fig. 5c).

Because of the hybrid status of P-Ps as having both phage and plasmid type genes, an intriguing question is whether they utilize distinct defensive and offensive systems for interelement competition. We assessed the diversity of defense system genes including both CRISPR and anti-CRISPR systems. From this examination, P-Ps were found to possess more anti-CRISPR systems compared to CRISPR-Cas systems (Fig. 6a) across all environments. We determined that only one P-P (NZ_CP063966.1) possessed both a CRISPR-Cas system and anti-CRISPR system (Supplementary Methods; Fig S7). This indicates that P-Ps typically utilize only one of these defensive or offensive strategies for limiting additional MGE co-infection (39, 40). The reduced variation of both systems in P-Ps was also noted. CRISPR-Cas systems genes were only found in Class I (n=132), Class III (n=32), and Class IV (n=36) among the five major categories. Interestingly, the predominant anti-CRISPR system genes detected was the AcrIIA7 (n=233), one of the most abundant anti-defenses CRISPR-associated inhibitors (41).

CRISPR-Cas defense systems were frequently found in host-associated and terrestrial P-Ps, with lower abundance among the aquatic P-Ps (Fig. 6). Most environments were characterized by even abundance of both classes of defense systems. However, some samples only recorded examples of one defense system class, such as animal host-associated P-Ps that possessed only CRISPR-Cas systems and fungi and plant P-P genomes that carried anti-CRISPR systems. These results demonstrate that P-Ps can carry CRISPR-Cas and anti-CRISPR systems in various environmental sources, however, these defense systems appear to be most commonly encountered in host-associated P-Ps.

After examining the unique contributions of ARGs, CRISPR-Cas systems, and anti-CRIPSR systems to phage-plasmid accessory genes, it appeared that many of these genes are relatively consistent in their distribution across all environments from which the P-Ps were derived. However, ARGs and certain defense systems were more abundant among host-associated P-Ps. The metabolic accessory genes were then examined to further investigate how this trend could impact other accessory gene functions. It was noted that the host-associated P-Ps possessed higher abundances of ARGs, CRISPR-Cas systems, anti-CRISPR systems and specific metabolic pathways such as pyrimidine metabolism, drug resistance, and cofactor and vitamin biosynthesis (Fischer Exact Test with a Benjamini-Hochberg correction; $p < 0.001$) (Supplementary Methods;

Fig. S9). The freshwater and saltwater P-Ps contained enriched macrolide biosynthesis, photosynthetic genes and unique nucleotide metabolic pathways such as polyketide sugar biosynthetic pathways (Fischer Exact Test with a Benjamini-Hochberg correction; $p < 0.001$) (Supplementary Methods; Fig. S9).

To further examine the diversity of the metabolic accessory genes found on P-Ps, we considered the dTDP-6-deoxy-α-D-allose biosynthesis pathway. This pathway is a critical for the formation of mycinose, as dTDP-6-deoxy-α-D-allose is the last free intermediate in this biosynthesis pathway (46). Mycinose is an important biomolecule that assists in forming several macrolide antibiotics (46). The P-Ps noted to possess this metabolic pathway were exclusively aquatic P-Ps and they all contained identical KEGG Modules (M00794) (45). In particular, these aquatic P-Ps contained three of the four enzymes in this pathway, including dTDP glucose 4,6-dehydratase, an enzyme that assists in forming all 6-deoxy sugar biosynthesis (Fig. 7) (47). These P-Ps contained genes encoding two enzymes (dTDP-4-dehydro-6-deoxy-D-glucose-3-epimerase and dTDP-4-dehydro-6-deoxy-α-D-gulose-4-ketoreducatase) that are essential to the dTDP-6-deoxy-α-D-allose biosynthesis pathway (42). The presence of the intermediate steps of the nucleotide sugar pathways (e.g., Fig. 7) suggests that P-Ps could stimulate auxiliary metabolite production from host-derived inputs of glucose 1-phosphate, dTTPs, and thymidyltransferase. Many polyketide sugars are frequently associated precursors for bacterial-produced antibiotic pathways, and these were exclusively found in aquatic P-Ps.

DISCUSSION

Here, we investigated the functional repertoire of accessory genes and the ecological diversity of P-Ps. P-Ps were found to inhabit a wide range of environments and exhibited notable genetic variation, with evidence suggesting that most accessory genes are derived from plasmids (12, 13, 15). P-P encoded accessory genes included a diverse arsenal of ARGs, CRISPR-Cas systems, virulence factors, and metabolism genes. While prior research primarily demonstrates that P-Ps possess most accessory genes at rates intermediate to both phages and plasmids, we found evidence that some accessory gene elements are disproportionately associated with P-Ps (12, 13, 15). Specifically, we found that anti-CRISPR systems and some ARGs [cpxA, EcoI_emrE, and CTX-M-142] were enriched in these elements (Supplementary Methods; Fig. S6, Fig. 6). With our developing understanding of MGE competition (e.g., plasmids containing CRISPR-Cas systems that may target bacteriophages), it raises questions about the role of P-Ps in such interactions (9, 48). Prior work has shown that some phages bearing anti-CRISPR systems have density-dependent protection from CRISPR-Cas, suggesting a role for cooperation and/or co-infection in the defense mechanism (49). Furthermore, it has been observed that P-Ps can exploit the replication machinery of plasmids to achieve a plasmids' relatively high copy number potential (14, 49). This replication strategy could allow for higher phage densities, thus potentiating anti-CRISPR systems.

P-P accessory gene content differed across environments. We examined various functional genes to investigate whether P-Ps confer traits that assist their prokaryotic hosts in adapting to their local environments. While these are not an exhaustive list of potential accessory genes, they are among the most important in understanding the ecology of P-Ps and their relevance to human health. We found that the distributions of these accessory genes varied significantly across environments. Host-associated P-Ps were enriched with defense systems and ARGs compared to aquatic P-Ps with increased abundances of intermediate secondary metabolic pathway genes. Many of these accessory genes appeared to be conserved, but the frequency depended on the environments from which these elements were sourced (e.g., Fig. 5,6,7). The overall trends of accessory genes appear similar to prior studies investigating plasmid gene diversity, although future works should investigate the differences between plasmid and phage-plasmid accessory genes (38, 50). The variability in accessory gene content among P-Ps suggests that these elements might occupy unique niches within microbial communities depending on their environments.

P-P genomic variation has the potential to alter microbial communities. Through the diversity of accessory gene content in host-associated, aquatic, and terrestrial-sourced P-Ps, we found a wide array of biologically-relevant accessory genes. These elements are prone to recombination and genetic exchange with other MGEs, making them of particular interest when considering their accessory genomes (15). These unique biological features with the diverse array of accessory genes highlight the importance of further study into these elements (12, 13, 14, 15, 50). Our results suggest that P-Ps offer notable genetic diversity and complexity that may impact MGE and bacterial evolution. The inherent variability of their hosts, viral genes, plasmid components, and functional genes these elements possess can play a significant role in shaping the recombination and HGT events in microbial populations. Understanding and potentially monitoring P-P populations offers potential benefits to mechanistic understanding of the recombination and transmission of accessory genes such as ARGs, MRGs, and virulence factors, contributing to their overall spread. The P-P accessory genome should be studied further to fully understand how these elements spread this diverse assortment of accessory genes.

METHODS:

Data Acquisition and Processing

The complete genomes of 33,595 plasmids were retrieved from PLSDB, 19,510 genomes from GPD, 52,958 genomes from MGV, and 1,416,547 genome and associated fragments from IMG/VR databases (19, 20, 21, 22). An additional 8,248 plasmids, 2,256 phages, and 780 P-Ps were obtained from Pfeifer et al. for training the random forest classifier (12). We removed genomes smaller than 10 kb to remove potentially fragmented genomes and genomes larger than 300 kb to avoid megaplasmids and chromatids. The information regarding the appropriate virus taxonomy, sampling source location, and additional information was collected from the metadata from PLSDB, GPD, MGV, and IMG/VR sources (19, 20, 21, 22). All analyses were conducted in Python (https://www.python.org/) unless otherwise stated.

Annotation of Protein Sequences

The genomes from PLSDB, GPD, MGV, IMG/VR, and Pfeifer et al. were processed with Prodigal (v2.6.3) using the (-p) meta setting to generate open reading frames (12, 19, 20, 21, 22, 51). The open reading frames were aligned to predicted protein sequences using diamond blastp (v4.6.8) using a minimum identity of 40%, minimum query coverage of 50%, maximum e-score of 1 x 10^{-5}, and k value of 15 (51). The less stringent settings allowed for the acquisition of more diverse phage species to ensure a high identification of all MGEs using mobileOG-db (Beatrix v1.6) (18). This database provides an inclusive and diverse distribution of MGE protein sequences, which allows for a robust analysis of MGEs.

Identification of Phage-Plasmids (P-Ps)

A random forest classifier was trained using the outputted results from the protein alignments using mobileOG-db (18). The features utilized in the classifier included the number of protein hits to bacteriophages, integrative elements, insertion sequences, plasmids, and multiple MGE class proteins. In addition, the associated mobileOG-db major categories of the proteins (phage, integration/excision, replication/recombination/repair, transfer, and stability/transfer/defense) were included with the total number of proteins and ORFs found in each genome (18). The Pfeifer et al. paper used several classification techniques including identifying

P-Ps from literature sources, plasmid HMMs found in phages, and plasmids with identified phage-specific profiles for classifying P-Ps due to the limited known P-Ps prior to their work (12). This paper utilizes the prior data obtained to train this classifier with the now known quantity of P-Ps. The model's training began by performing ten randomized training sets using approximately 20% of samples as test data and 80% as training data. The random forest classifier had a max decision depth of 8 and used entropy as the criteria measurement. The performance results from the ten randomized trials were averaged to examine the effectiveness of the random forest classifier. The classifier achieved an average accuracy of 95.4% and a false positive rate of 2.9%. The testing data consisted of approximately 160 P-Ps, 500 phages, and 1350 plasmids, while the training data contained 620 P-Ps, 2,000 phages, and 5,400 plasmids (12). The PLSDB, MGV, GPD, and IMG/VR genomes were then classified using the trained random forest classifier to identify whether each element was a plasmid, phage-plasmid, or bacteriophage (19, 20, 21, 22). CD-HIT-EST v4.6.8 was utilized to cluster the sequences and remove sequences with < 97% sequence similarity (53). The P-Ps were then examined using CompareM to compare the average nucleotide identity between the samples to compare the sequence similarity after clustering (54).

Manual Curation of Source Location

The classified phage-plasmid genomes were cross-referenced with the source database metadata to determine additional information regarding the source locations and taxonomy for additional analysis. The P-Ps were then categorized by environmental source location into the following categories: aquatic, terrestrial, host-associated, and unclassified environments. These categories were separated into more unique categories according to the exact location of the genomes, including the subcategories of saltwater, freshwater, wastewater, other aquatic genomes, soil, sediment, human facilities, other terrestrial, human, fungi, animal, plant, other host-associated genomes, and unclassified genomes. Genomes designated as others had designated source locations but were too generalized to classify the genomes further correctly. Phage-plasmids with undocumented source locations were cross-referenced with NCBI BioSample to classify the elements further, but genomes that still could not be classified were designated as other. All genomes with no metadata source locations or metadata with ambiguous locations were removed from source location analysis.

Data Analysis

The taxonomy of the P-Ps was classified using the associated source metadata from the respective databases. Due to the limited phage and plasmid taxonomy, the associated incompatibility groups of the P-Ps were further classified using PlasmidFinder (v2.1.6) with default parameters and the viral taxonomic classifications were classified using geNOMAD (v1.5.2) (19, 30). The study further identified the key accessory genes of the phage-plasmids, including ARGs, defense systems, toxin-antitoxin systems, metabolism genes, metal resistance genes, and virulence factors. The defense systems were identified using PADLOC (v.1.1.0) classification tool (55). The phage-plasmid genomes were processed through GhostKoala to extract the KEGGs from the Reconstruction Mapper function for identifying the metabolic genes (45, 56). Microbe Annotator (light-v2.0.5) was used to identify complete or partially complete KEGG Module pathways from the specific P-Ps using the blast settings (45, 57). These pathways were classified if the P-Ps contain 50% of the required genes for a specific biosynthesis pathway.

The virulence factors, metal resistance genes, anti-CRISPR genes, and the toxin-antitoxin systems were classified by processing the phage-plasmids using Diamond blastp (v4.6.8) against the VFDB genes from set A, the BacMet2 Predicted dataset, Anti-CRISPRdb (v2.2) database, and the TADB (v.2.0) database with query coverage of 80%, percent identity of 90%, and e-score of 1x10-5 (34, 35, 52, 58, 59). The classified phage, plasmid, and P-P genomes were queried against CARD (v3.0.7) with a minimum identity of 80% and an e-value<10^{-10} (26). The phage-plasmid genomes were then processed through EggNOG-Mapper (v2) to get the associated PFAMs, and COGs for the additional P-P analysis (60, 61, 62). A random selection of 500 phages and 500 plasmids were isolated from the prior classified phage and plasmids classifications. These genomes were processed utilizing the same tools as the phage-plasmids to determine the accessory genes found in these genomes. The graphical analysis was performed using R (https://www.r-project.org/), draw.io (http://draw.io/), and bioicons (https://bioicons.com/).

Data Availability:

All available data can be download from the databases analyzed in this study with all associated accession identification numbers located in the supplemental tables. The supplementary data can be found at the manuscript FigShare repository located at: https://figshare.com/s/b0ffbc71c0bf43e251df . Scripts used in data mining, processing the data, and generating the scripts can be found at https://github.com/jamesm224/phage-plasmid-classification.

Acknowledgements:

We would like to express our gratitude and appreciate for all of the assistance and help that went into designing this study. The authors acknowledge the Advanced Research Computing at Virginia Tech for providing the computational resources used in this study. This study was supported by NSF CI4-WARS (PI Zhang) Award 2004751, the NSF NRT-HDR training (PI Pruden) Award 2125798, and the Fralin Life Sciences Institute.

The authors report no conflict of interest.

REFERENCES

1. Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements, and why? Heredity (Edinb) 106:1–10.

2. Hall JPJ, Harrison E, Baltrus DA. 2022. Introduction: the secret lives of microbial mobile genetic elements. Philos Trans R Soc B Biol Sci 377.

3. Segerman B. 2012. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. Front Cell Infect Microbiol 2:116.

4. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat Rev Microbiol 2021 196 19:347–359.

5. Takeuchi N, Hamada-Zhu S, Suzuki H. 2023. Prophages and plasmids can display opposite trends in the types of accessory genes they carry. Proc R Soc B Biol Sci 290.

6. Saunders JR, Allison H, James CE, McCarthy AJ, Sharp R. 2001. Phage-mediated transfer of virulence genes. J Chem Technol Biotechnol 76:662–666.

7. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe JA, Chisholm SW. 2011. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A 108:E757–E764.

8. Kieft K, Zhou Z, Anderson RE, Buchan A, Campbell BJ, Hallam SJ, Hess M, Sullivan MB, Walsh DA, Roux S, Anantharaman K. 2021. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun 2021 121 12:1–16.

9. Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements, and why? Heredity (Edinb) 106:1–10.

10. LeGault KN, Hays SG, Angermeyer A, McKitterick AC, Johura FT, Sultana M, Ahmed T, Alam M, Seed KD. 2021. Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. Science 373.

11. Botelho J. 2023. Defense systems are pervasive across chromosomally integrated mobile genetic elements and are inversely correlated to virulence and antimicrobial resistance. Nucleic Acids Res 51:4385–4397.

12. 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.

13. Pfeifer E, Bonnin RA, Rocha EPC. 2022. Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. MBio 13.

14. Shan X, Szabo RE, Cordero OX. 2023. Mutation-induced infections of phage-plasmids. Nat Commun 2023 141 14:1–10.

15. Pfeifer, E., Rocha, E.P.C. Phage-plasmids promote recombination and emergence of phages and plasmids. Nat Commun 15, 1545 (2024). https://doi.org/10.1038/s41467-024-45757-3

16. Finks SS, Martiny JBH. 2023. Plasmid-Encoded Traits Vary across Environments. MBio 14.

17. Parmar K, Dafale N, Pal R, Tikariha H, Purohit H. 2018. An Insight into Phage Diversity at Environmental Habitats using Comparative Metagenomics Approach. Curr Microbiol 75:132–141.

18. Brown CL, Mullet J, Hindi F, Stoll JE, Gupta S, Choi M, Keenum I, Vikesland P, Pruden A, Zhang L. 2022. mobileOG-db: a Manually Curated Database of Protein Families Mediating the Life Cycle of Bacterial Mobile Genetic Elements. Appl Environ Microbiol 88.

19. Schmartz GP, Hartung A, Hirsch P, Kern F, Fehlmann T, Müller R, Keller A. 2022. PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res 50:D273–D278.

20. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098-1109.e9.

21. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, Kyrpides NC. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol 2021 67 6:960–970.

22. Camargo AP, Nayfach S, Chen IMA, Palaniappan K, Ratner A, Chu K, Ritter SJ, Reddy TBK, Mukherjee S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC, Roux S. 2023. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res 51:D733–D743.

23. Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen CY, Graham M, Van Domselaar G, Stothard P. 2023. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res 51:W484–W492.

24. Starikova E V., Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilina EN, Govorun VM. 2020. Phigaro: high-throughput prophage sequence annotation. Bioinformatics 36:3882–3884.

25. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.

26. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen AL V., Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran HK, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar G V., McArthur AG. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48:D517–D525.

27. Arakawa K, Tomita M. 2007. The GC Skew Index: A Measure of Genomic Compositional Asymmetry and the Degree of Replicational Selection. Evol Bioinform Online 3:159.

28. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. 2023. Identification of mobile genetic elements with geNomad. Nat Biotechnol 2023 1–10.

29. SankeyMATIC: Make Beautiful Flow Diagrams. https://sankeymatic.com/. Retrieved 10 December 2023.

30. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, Aarestrup FM, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903.

31. Guentzel MN. 1996. Escherichia, Klebsiella, Enterobacter, Serratia, Citrobacter, and Proteus. Med Microbiol.

32. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, Dempsey DM, Dutilh BE, García ML, Curtis Hendrickson R, Junglen S, Krupovic M, Kuhn JH, Lambert AJ, Łobocka M, Oksanen HM, Orton RJ, Robertson DL, Rubino L, Sabanadzovic S, Simmonds P, Smith DB, Suzuki N, Van Doorslaer K, Vandamme AM, Varsani A, Zerbini FM. 2022. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). Arch Virol 167:2429–2440.

33. Rigou S, Santini S, Abergel C, Claverie JM, Legendre M. 2022. Past and present giant viruses diversity explored through permafrost metagenomics. Nat Commun 2022 131 13:1–13.

34. Xie Y, Wei Y, Shen Y, Li X, Zhou H, Tai C, Deng Z, Ou HY. 2018. TADB 2.0: an updated database of bacterial type II toxin–antitoxin loci. Nucleic Acids Res 46:D749.

35. Dong C, Wang X, Ma C, Zeng Z, Pu DK, Liu S, Wu CS, Chen S, Deng Z, Guo FB. 2022. Anti-CRISPRdb v2.2: an online repository of anti-CRISPR proteins including information on inhibitory mechanisms, activities and neighbors of curated anti-CRISPR proteins. Database 2022:1–9.

36.Wilkins D. gggenes: Draw Gene Arrow Maps in 'ggplot2'_. R package version 0.5.1, https://CRAN.R-project.org/package=gggenes.

37. Ogbolu DO, Terry Alli OA, Webber MA, Oluremi AS, Oloyede OM. 2018. CTX-M-15 is Established in Most Multidrug-Resistant Uropathogenic Enterobacteriaceae and Pseudomonaceae from Hospitals in Nigeria. Eur J Microbiol Immunol (Bp) 8:20–24.

38. Anthony WE, Burnham CAD, Dantas G, Kwon JH. 2021. The Gut Microbiome as a Reservoir for Antimicrobial Resistance. J Infect Dis 223:S209.

39. Pawluk A, Amrani N, Zhang Y, Garcia B, Hidalgo-Reyes Y, Lee J, Edraki A, Shah M, Sontheimer EJ, Maxwell KL, Davidson AR. 2016. Naturally Occurring Off-Switches for CRISPR-Cas9. Cell 167:1829-1838.e9.

40. Pinilla-Redondo R, Russel J, Mayo-Muñoz D, Shah SA, Garrett RA, Nesme J, Madsen JS, Fineran PC, Sørensen SJ. 2022. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. Nucleic Acids Res 50:4315–4328.

41. Uribe R V., van der Helm E, Misiakou MA, Lee SW, Kol S, Sommer MOA. 2019. Discovery and Characterization of Cas9 Inhibitors Disseminated across Seven Bacterial Phyla. Cell Host Microbe 25:233-241.e5.

42. Thuy TTT, Liou K, Oh TJ, Kim DH, Nam DH, Yoo JC, Sohng JK. 2007. Biosynthesis of dTDP-6-deoxy-β-d-allose, biochemical characterization of dTDP-4-keto-6-deoxyglucose reductase (GerKI) from Streptomyces sp. KCTC 0041BP. Glycobiology 17:119–126.

43. MetaCyc: Metabolic Pathways From all Domains of Life. https://metacyc.org/. Retrieved 3 December 2023.

44. Bate N, Cundliffe E. 1999. The mycinose-biosynthetic genes of Streptomyces fradiae, producer of tylosin. J Ind Microbiol Biotechnol 23:118–122.

45. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457.

46. Li S, Anzai Y, Kinoshita K, Kato F, Sherman DH. 2009. Functional Analysis of MycE and MycF, Two O-Methyltransferases Involved in Biosynthesis of Mycinamicin Macrolide Antibiotics. Chembiochem 10:1297.

47. Beis K, Allard STM, Hegeman AD, Murshudov G, Philp D, Naismith JH. 2003. The structure of NADH in the enzyme dTDP-D-glucose dehydratase (RmlB). J Am Chem Soc 125:11872–11878.

48. Siedentop B, Rüegg D, Bonhoeffer S, Chabas H. 2024. My host's enemy is my enemy: plasmids carrying CRISPR-Cas as a defence against phages. Proc R Soc B 291.

49. Landsberger M, Gandon S, Meaden S, Rollie C, Chevallereau A, Chabas H, Buckling A, Westra ER, van Houte S. 2018. Anti-CRISPR Phages Cooperate to Overcome CRISPR-Cas Immunity. Cell 174:908.

50. Finks SS, Martiny JBH. 2023. Plasmid-Encoded Traits Vary across Environments. MBio 14.

51. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:1–11.

52. Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2014 121 12:59–60.

53. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

54. donovan-h-parks/CompareM: A toolbox for comparative genomics. https://github.com/donovan-h-parks/CompareM. Retrieved 22 October 2023.

55. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, Jackson SA. 2022. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. Nucleic Acids Res 50:W541–W550.

56. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 428:726–731.

57. Ruiz-Perez CA, Conrad RE, Konstantinidis KT. 2021. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics 22:1–

58. Liu B, Zheng D, Zhou S, Chen L, Yang J. 2022. VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic Acids Res 50:D912.

59. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DGJ. 2014. BacMet: antibacterial biocide and metal resistance genes database. Nucleic Acids Res 42:D737.

60. Cantalapiedra CP, Herņandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829.

61. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412–D419.

62. Tatusov RL, Galperin MY, Natale DA, Koonin E V. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36.

63. Couvin D, Bernheim A, Toffano-Nioche C, Touchon M, Michalik J, Néron B, Rocha EPC, Vergnaud G, Gautheret D, Pourcel C. 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 46:W246–W251.

FIGURE LEGENDS

Figure 1. Filtering and identification of phage-plasmids from publicly-available phage and plasmid genomes. The genomes from the three phage databases (n=1,155,953) and one plasmid database (n=23,905) were processed against mobileOG-db to identify MGE-related hallmark genes (18). The genomes were then reclassified into phages (n=1,031,108), plasmid (n=140,367), and phage-plasmid (n=8,383) using a random forest classifier that identifies P-Ps using phage and plasmid hallmark proteins. The phage plasmids were then clustered (n=5,742) to remove identical genomes and manually curated by the associated source location of the classified genomes.

Figure 2. Phage-plasmids (P-Ps) are prolific in databases of plasmids and phages. (A) Number of classified MGEs of each element class from the four respective databases before dereplication. (B) The hybrid nature of P-Ps are reflected in the patterns of mobileOGs. (C) Illustration of a phage-plasmid from PLSDB (id=NZ_CP025141.1) depicted using Proksee including Phigaro, Prokka, mobileOG-db, CARD, and GC Skew annotations (18, 19, 23, 24, 25, 26, 27). All unlabeled or unclassified proteins were removed from this figure.

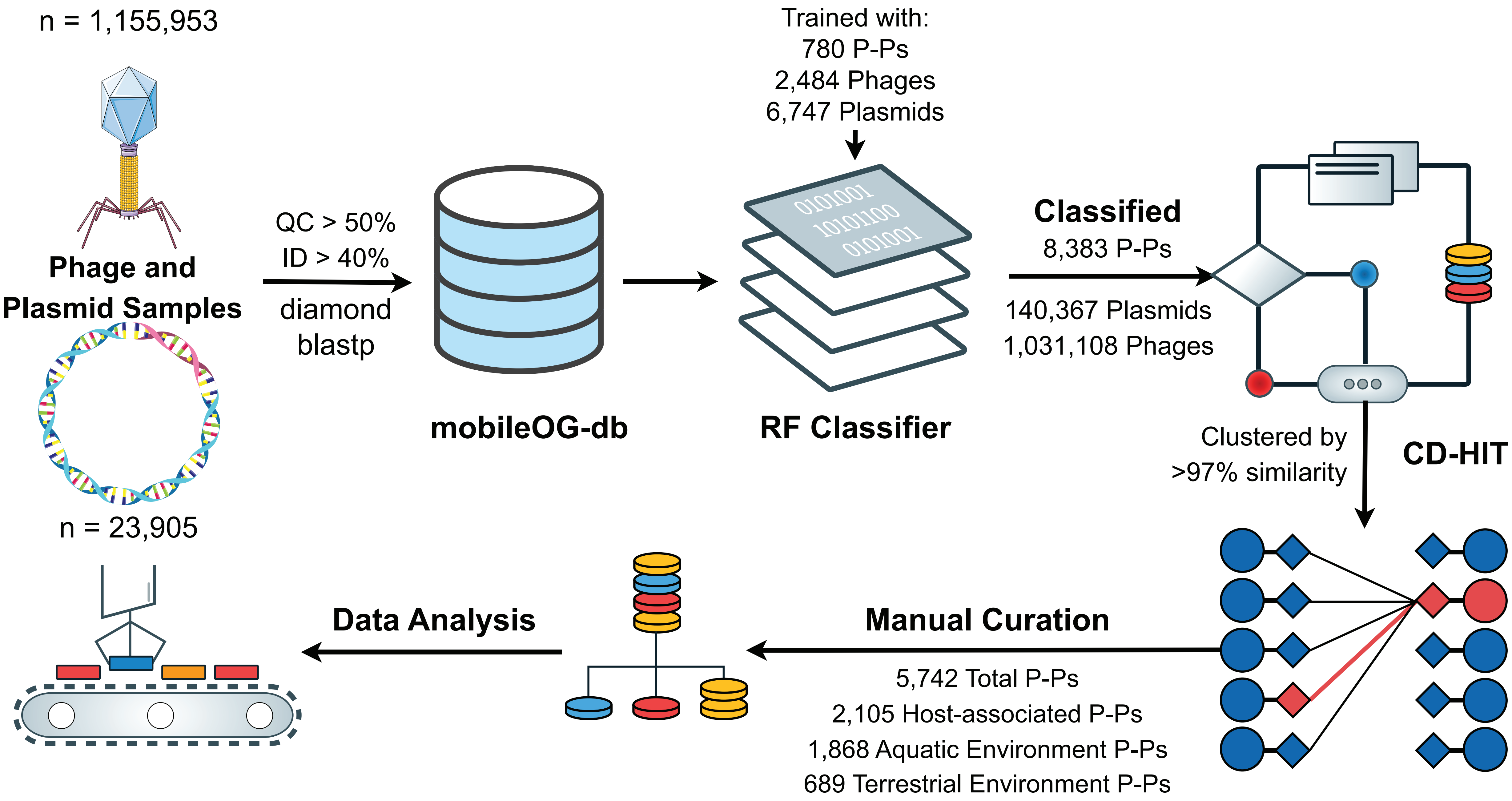
Figure 3. Classification and distribution of P-Ps according to the reported source organism and environment. Examination of the relative abundance of P-P host predicted taxa for aquatic (A), host-associated (B), and terrestrial (C) genomes. The predicted host phyla, class, order, and family of each respective source location are included in each subfigure (29). The predicted host taxa for any P-Ps without reported environmental source locations were excluded from this analysis. Any infrequent taxa that were <1% abundance in the respective environmental location were not included in this figure.

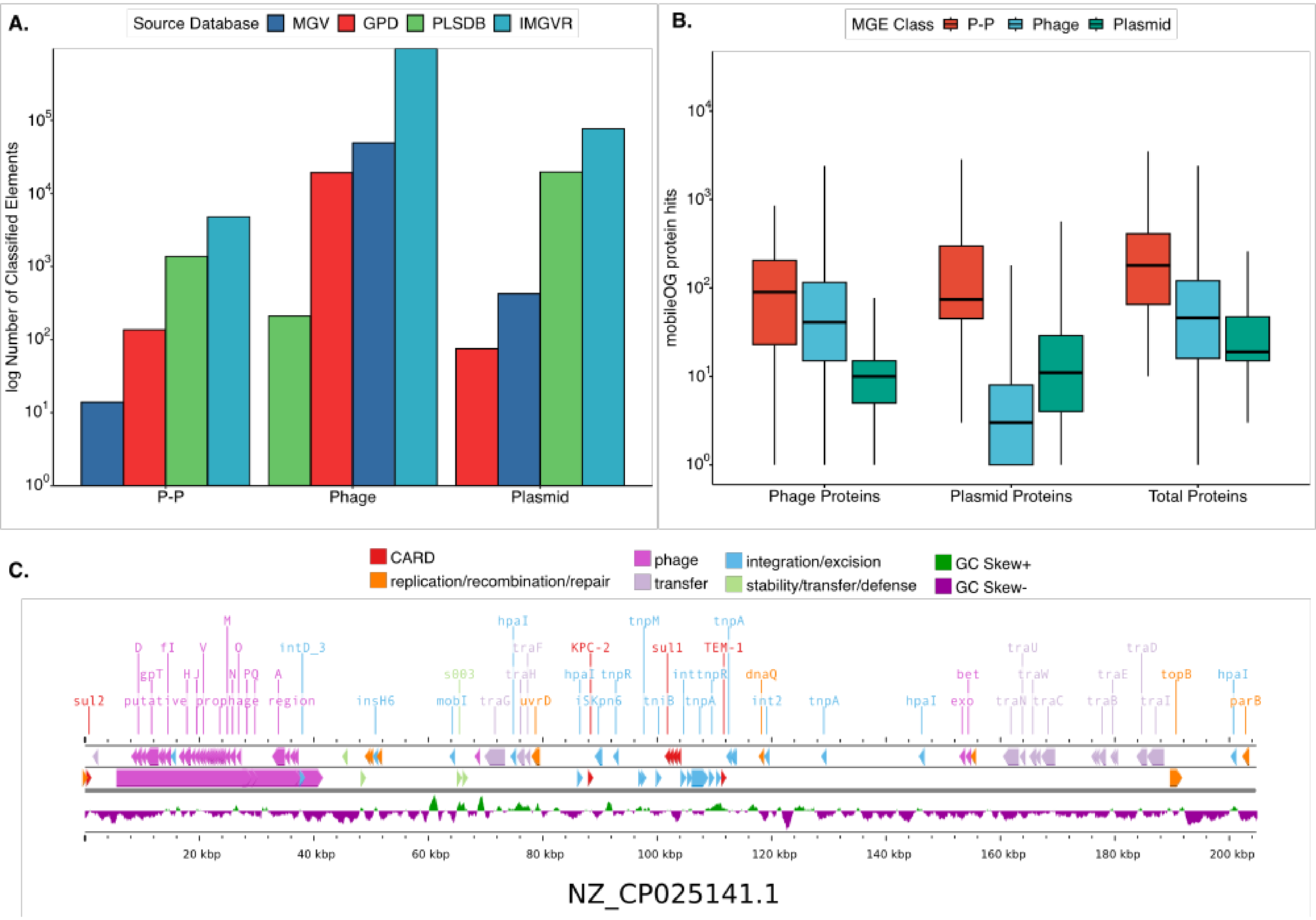
Figure 4. Comparative analysis of key accessory genes found to be carried by the P-Ps across a diverse range of source environments. The accessory genes were grouped into virulence factors, metal resistance, metabolism, defense systems, and antibiotic-resistance genes (ARGs). These genes were grouped into associated functional categories as shown in the supplementary tables. It was noted that both the toxin-antitoxin genes identified from TADB and the anti-CRISPR genes classified from Anti-CRISPRdb v2.2 were grouped with the defense system genes for visual purposes (34, 35). Only accessory gene categories with at least 25 hits were included in the figure above. The values were taken from the log10 of the relative frequency of the genes compared to the total number of accessory genes found in each element source location. The plasmid and phage categories comprise 500 random phages and plasmids, capturing differences between the various class of MGEs and acting as experimental baseline controls for comparing phages, plasmids, and P-Ps.

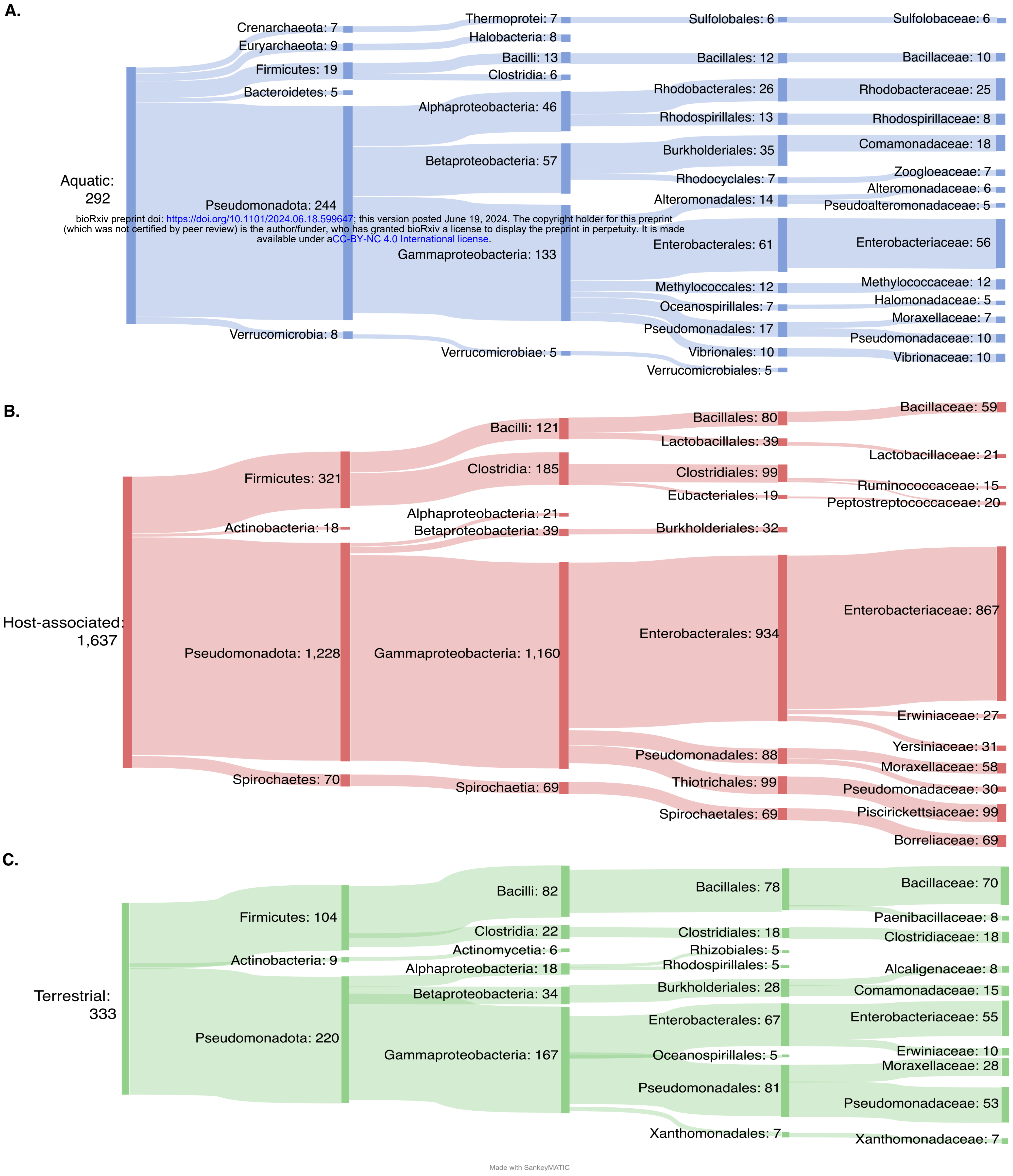
Figure 5. Diversity and distribution of ARGs among P-Ps of various origin. (A) Total number of all ARGs found in P-Ps originating from each source environment. (B) Frequency of common antibiotic resistance genes (ARGs) carried by the P-Ps relative to the total identified ARGs in each source environment. Only source environments possessing >8 unique ARGs were included in the figure. (C) Gene-to-gene alignment of the CTX-M-15 ARG grouped by the respective source environment of the phage-plasmids (36).

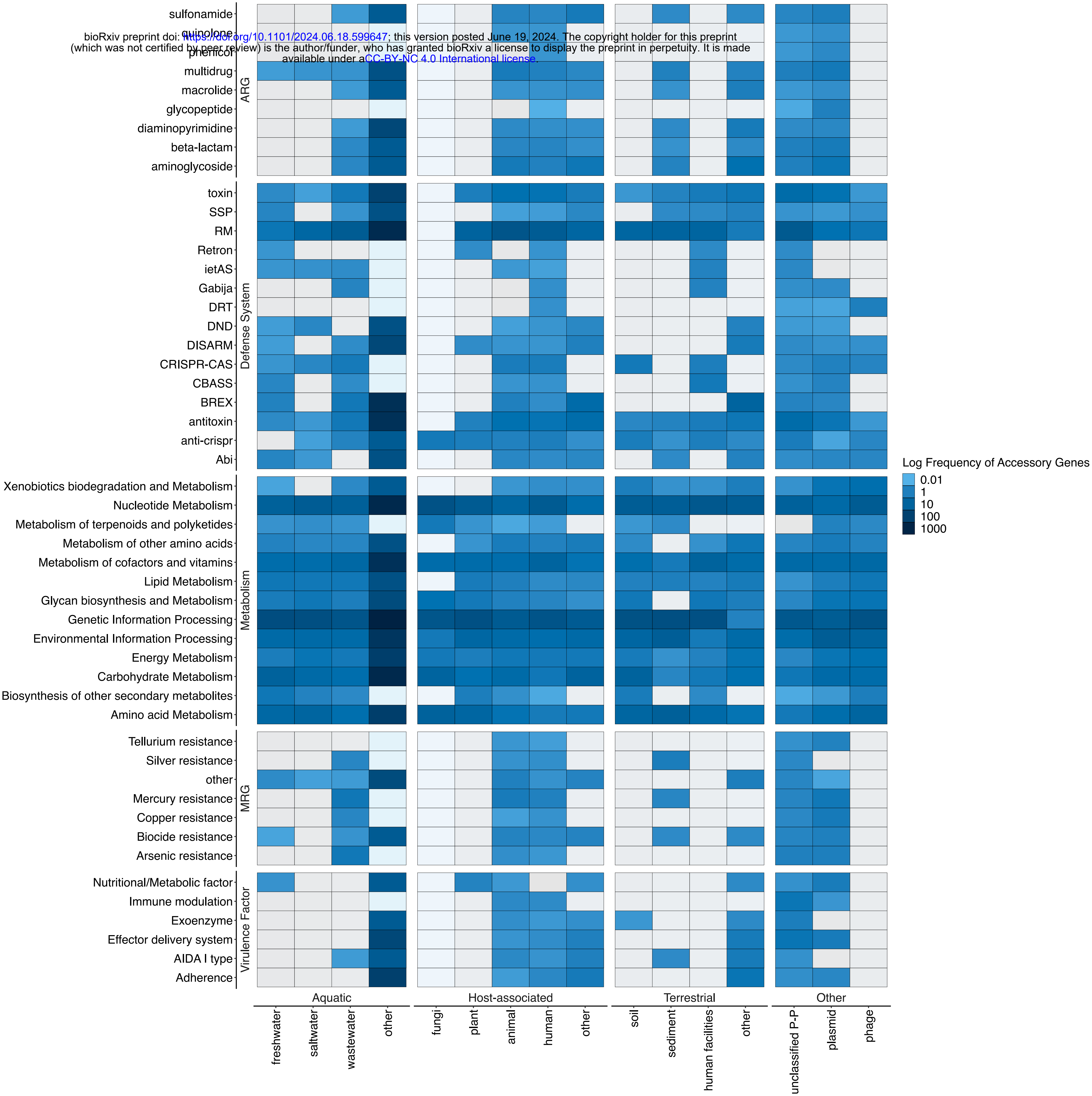
Figure 6. Analysis of the diversity and frequency of CRISPR-Cas and anti-CRISPR systems. (A) Relative abundance of anti-CRISPR and CRISPR-Cas systems encountered among unique P-P genomes reconstructed from each of the respective source environments (Relative Abundance = P-P genomes containing defense system from a respective source location / Total P-P genomes from respective source environments). (B) Distribution and occurrence of total CRISPR-Cas gene subtypes in P-Ps. “Other” includes systems that could not be classified into one single category or which were classified as a category other than the five primary classes of CRISPR-Cas systems. (C) Prevalence and abundance of anti-CRISPR system genes in P-Ps. Only subtypes found in the P-Ps are displayed in the figure.

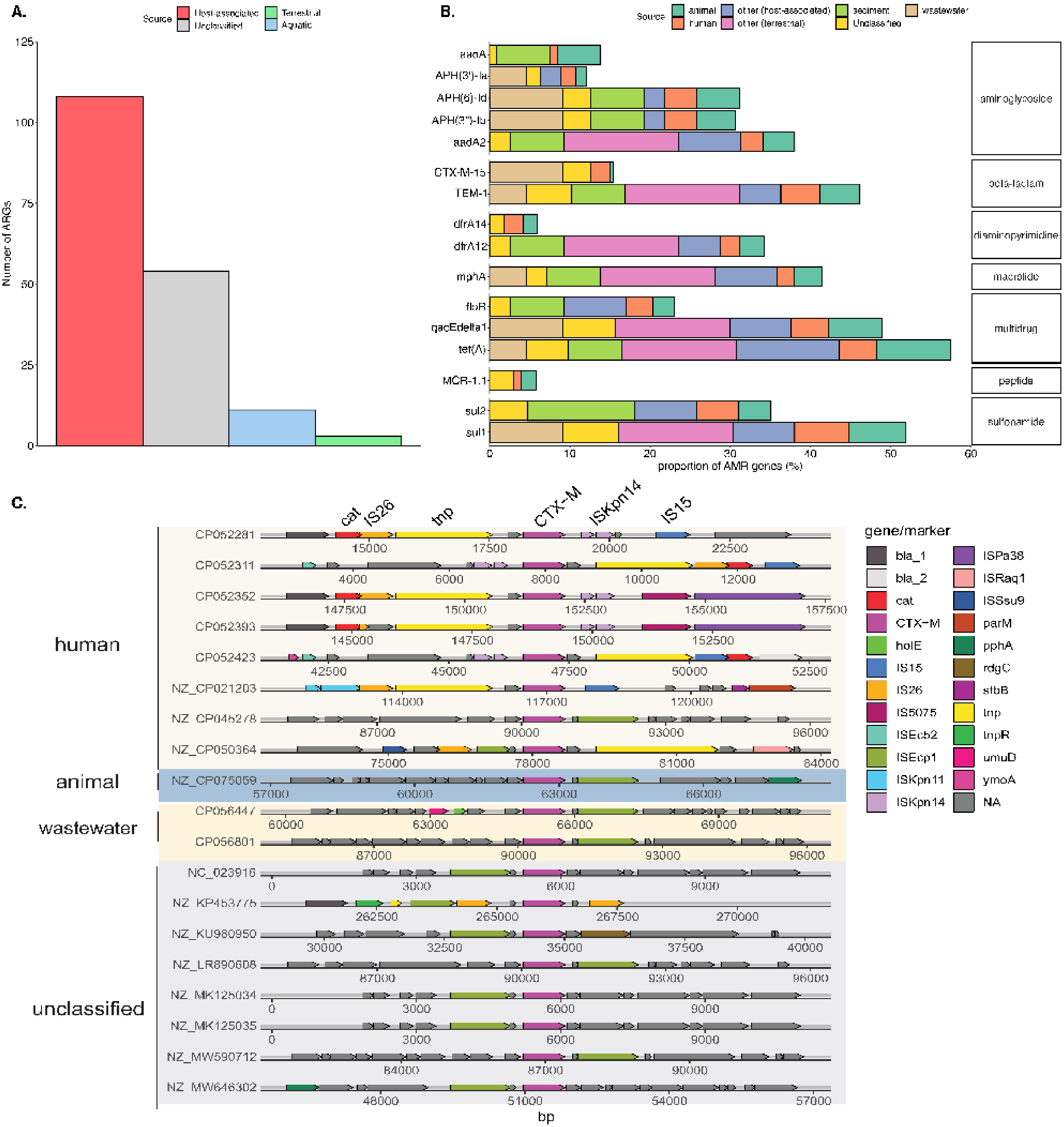
Figure 7. dTDP-6-deoxy-α-D-allose biosynthesis pathway found in some aquatic P-Ps (n=14/1,868) (42, 43, 44). The blue-outlined boxes indicate the portion of the associated pathway found in the P-Ps. Among the 14 P-Ps found to carry portions of this pathway, 13 were derived from freshwater and one from saltwater. The designated KEGG pathways align with the reaction products from these enzymes with the blue KEGG pathways indicating portions of the pathway that the P-P carried accessory genes (45).

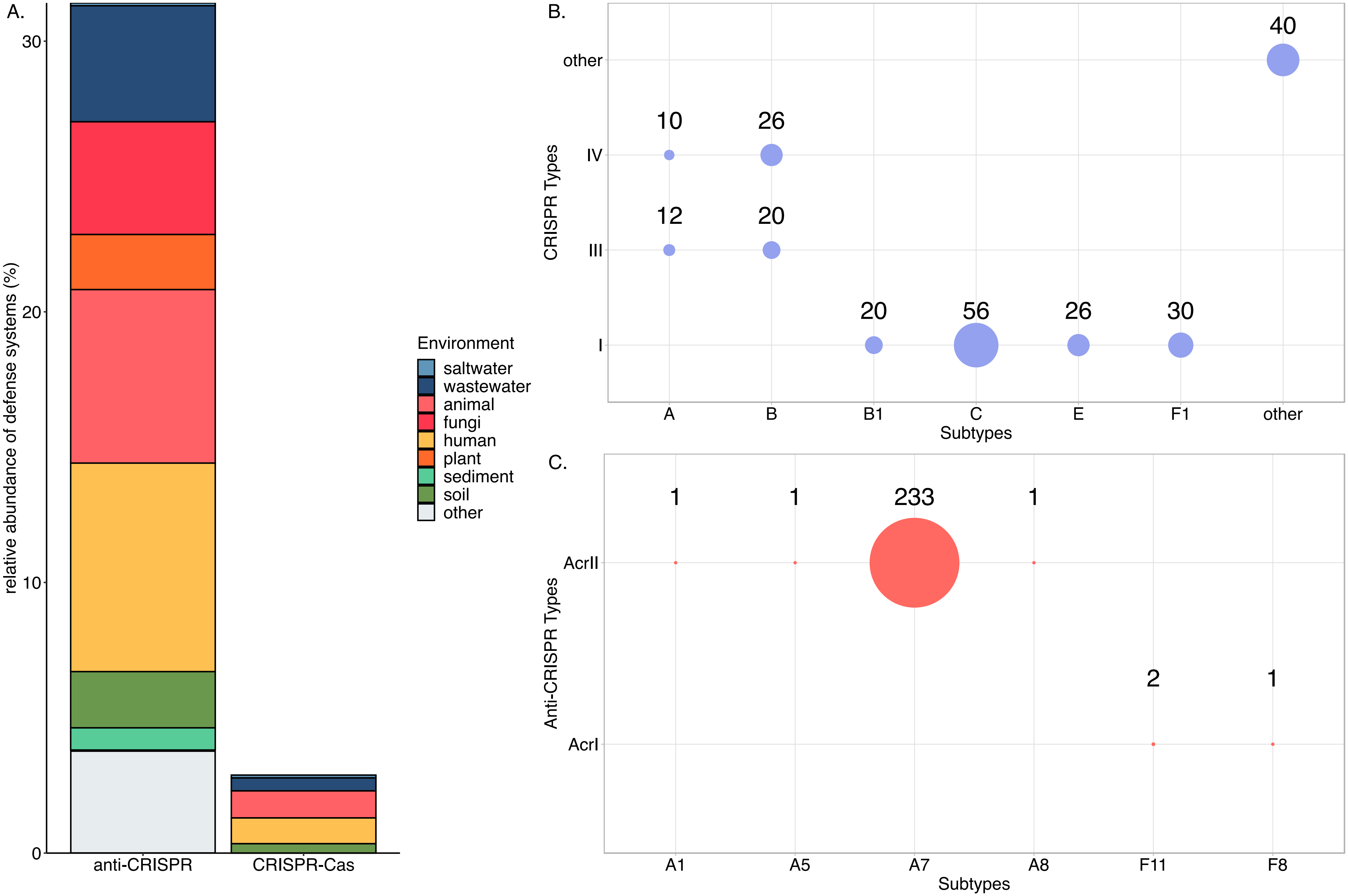

n = 1,155,953
Phage and Plasmid Samples
n = 23,905
QC > 50%
ID > 40%
diamond blastp
mobileOG-db
Trained with:
780 P-Ps
2,484 Phages
6,747 Plasmids
0101001
10101100
0101001
RF Classifier
Classified
8,383 P-Ps
140,367 Plasmids
1,031,108 Phages
Clustered by >97% similarity
CD-HIT
Manual Curation
5,742 Total P-Ps
2,105 Host-associated P-Ps
1,868 Aquatic Environment P-Ps
689 Terrestrial Environment P-Ps
Data Analysis


A.
Source Database
MGV
GPD
PLSDB
IMGVR
log Number of Classified Elements
10^0
10^1
10^2
10^3
10^4
10^5
P-P
Phage
Plasmid
B.
MGE Class
P-P
Phage
Plasmid
mobileOG protein hits
10^0
10^1
10^2
10^3
10^4
Phage Proteins
Plasmid Proteins
Total Proteins
C.
CARD
replication/recombination/repair
phage
transfer
integration/excision
stability/transfer/defense
GC Skew+
GC Skew-
M
hpaI
tnpM
tnpA
D
fI
V
O
intD_3
traF
KPC-2
sul1
TEM-1
traU
traD
gpT
H J
N
PQ
A
s883
traH
hpaI
tnpR
int
tnpR
dnaQ
bet
traW
traE
topB
hpaI
sul2
putative prophage region
insH6
mobI
traG
uvrD
iSKpn6
tniB
tnpA
int2
tnpA
hpaI
exo
traN
traC
traB
traI
parB
20 kbp
40 kbp
60 kbp
80 kbp
100 kbp
120 kbp
140 kbp
160 kbp
180 kbp
200 kbp
NZ_CP025141.1

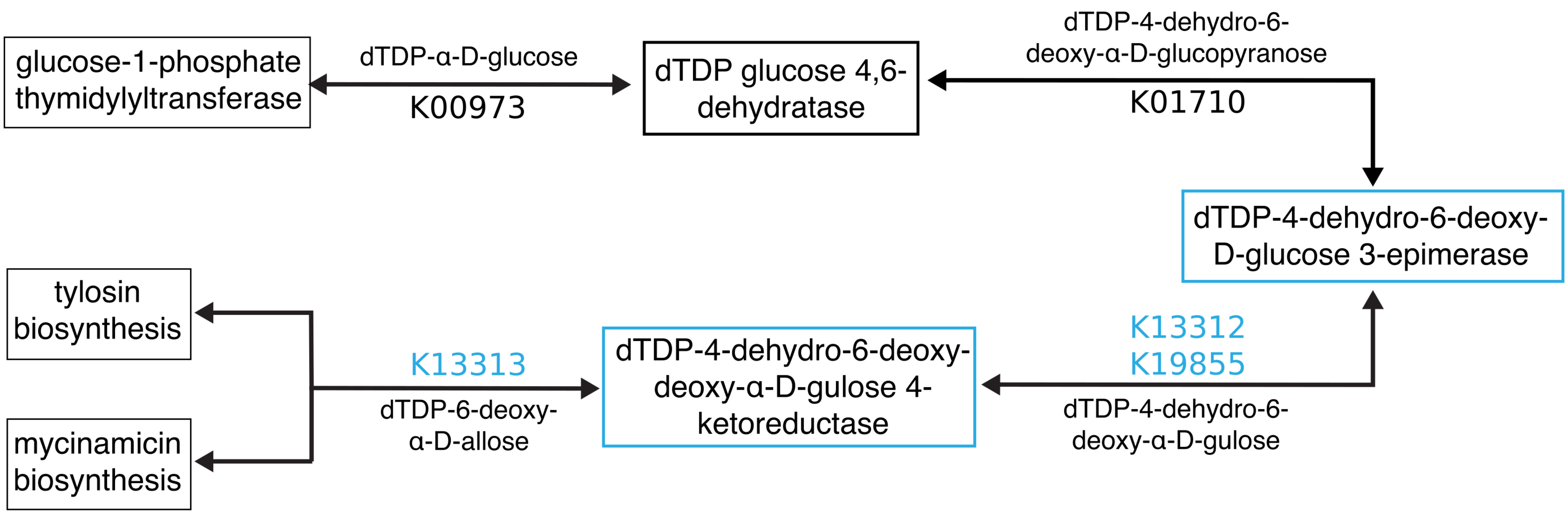

A.
Aquatic: 292
Crenarchaeota: 7
Euryarchaeota: 9
Firmicutes: 19
Bacteroidetes: 5
Pseudomonadota: 244
Verrucomicrobia: 8
Thermoprotei: 7
Halobacteria: 8
Bacilli: 13
Clostridia: 6
Alphaproteobacteria: 46
Betaproteobacteria: 57
Gammaproteobacteria: 133
Verrucomicrobiae: 5
Sulfolobales: 6
Bacillales: 12
Rhodobacterales: 26
Rhodospirillales: 13
Burkholderiales: 35
Rhodocyclales: 7
Alteromonadales: 14
Enterobacterales: 61
Methylococcales: 12
Oceanospirillales: 7
Pseudomonadales: 17
Vibrionales: 10
Verrucomicrobiales: 5
Sulfolobaceae: 6
Bacillaceae: 10
Rhodobacteraceae: 25
Rhodospirillaceae: 8
Comamonadaceae: 18
Zoogloeaceae: 7
Alteromonadaceae: 6
Pseudoalteromonadaceae: 5
Enterobacteriaceae: 56
Methylococcaceae: 12
Halomonadaceae: 5
Moraxellaceae: 7
Pseudomonadaceae: 10
Vibrionaceae: 10
B.
Host-associated: 1,637
Firmicutes: 321
Actinobacteria: 18
Pseudomonadota: 1,228
Spirochaetes: 70
Bacilli: 121
Clostridia: 185
Alphaproteobacteria: 21
Betaproteobacteria: 39
Gammaproteobacteria: 1,160
Spirochaetia: 69
Bacillales: 80
Lactobacillales: 39
Clostridiales: 99
Eubacteriales: 19
Burkholderiales: 32
Enterobacterales: 934
Pseudomonadales: 88
Thiotrichales: 99
Spirochaetales: 69
Bacillaceae: 59
Lactobacillaceae: 21
Ruminococcaceae: 15
Peptostreptococcaceae: 20
Enterobacteriaceae: 867
Erwiniaceae: 27
Yersiniaceae: 31
Moraxellaceae: 58
Pseudomonadaceae: 30
Piscirickettsiaceae: 99
Borreliaceae: 69
C.
Terrestrial: 333
Firmicutes: 104
Actinobacteria: 9
Pseudomonadota: 220
Bacilli: 82
Clostridia: 22
Actinomycetia: 6
Alphaproteobacteria: 18
Betaproteobacteria: 34
Gammaproteobacteria: 167
Bacillales: 78
Clostridiales: 18
Rhizobiales: 5
Rhodospirillales: 5
Burkholderiales: 28
Enterobacterales: 67
Oceanospirillales: 5
Pseudomonadales: 81
Xanthomonadales: 7
Bacillaceae: 70
Paenibacillaceae: 8
Clostridiaceae: 18
Alcaligenaceae: 8
Comamonadaceae: 15
Enterobacteriaceae: 55
Erwiniaceae: 10
Moraxellaceae: 28
Pseudomonadaceae: 53
Xanthomonadaceae: 7
Made with SankeyMATIC


ARG
sulfonamide
quinolone
phenicol
multidrug
macrolide
glycopeptide
diaminopyrimidine
beta-lactam
aminoglycoside
Defense System
toxin
SSP
RM
Retron
ietAS
Gabija
DRT
DND
DISARM
CRISPR-CAS
CBASS
BREX
antitoxin
anti-crispr
Abi
Metabolism
Xenobiotics biodegradation and Metabolism
Nucleotide Metabolism
Metabolism of terpenoids and polyketides
Metabolism of other amino acids
Metabolism of cofactors and vitamins
Lipid Metabolism
Glycan biosynthesis and Metabolism
Genetic Information Processing
Environmental Information Processing
Energy Metabolism
Carbohydrate Metabolism
Biosynthesis of other secondary metabolites
Amino acid Metabolism
MRG
Tellurium resistance
Silver resistance
other
Mercury resistance
Copper resistance
Biocide resistance
Arsenic resistance
Virulence Factor
Nutritional/Metabolic factor
Immune modulation
Exoenzyme
Effector delivery system
AIDA I type
Adherence
Aquatic
Host-associated
Terrestrial
Other
freshwater
saltwater
wastewater
other
fungi
plant
animal
human
other
soil
sediment
human facilities
other
unclassified P-P
plasmid
phage
Log Frequency of Accessory Genes
0.01
1
10
100
1000


A.
Source
Host-associated
Unclassified
Terrestrial
Aquatic
Number of ARGs
0
25
50
75
100
125
B.
Source
animal
human
other (host-associated)
other (terrestrial)
sediment
Unclassified
wastewater
aadA
APH(3')-Ia
APH(6)-Id
APH(3'')-Ib
aadA2
CTX-M-15
TEM-1
dfrA14
dfrA12
mphA
floR
qacEdelta1
tet(A)
MCR-1.1
sul2
sul1
aminoglycoside
beta-lactam
diaminopyrimidine
macrolide
multidrug
peptide
sulfonamide
0
10
20
30
40
50
60
proportion of AMR genes (%)
C.
cat
IS26
tnp
CTX-M
ISKpn14
IS15
human
animal
wastewater
unclassified
CP052281
CP052311
CP052352
CP052393
CP052423
NZ_CP021203
NZ_CP045278
NZ_CP050364
NZ_CP075059
CP056447
CP056801
NC_023916
NZ_KP453775
NZ_KU980950
NZ_LR890608
NZ_MK125034
NZ_MK125035
NZ_MW590712
NZ_MW646302
bp
gene/marker
bla_1
bla_2
cat
CTX−M
holE
IS15
IS26
IS5075
ISEcb2
ISEcp1
ISKpn11
ISKpn14
ISPa38
ISRaq1
ISSsu9
parM
pphA
rdgC
stbB
tnp
tnpR
umuD
ymoA
NA


A.
relative abundance of defense systems (%)
30
20
10
0
anti-CRISPR
CRISPR-Cas
Environment
saltwater
wastewater
animal
fungi
human
plant
sediment
soil
other
B.
CRISPR Types
other
IV
III
I
40
10
26
12
20
20
56
26
30
A
B
B1
C
E
F1
other
Subtypes
C.
Anti-CRISPR Types
AcrII
AcrI
1
1
233
1
2
1
A1
A5
A7
A8
F11
F8
Subtypes

dTDP-6-deoxy-α-D-allose Biosynthesis



glucose-1-phosphate thymidylyltransferase

dTDP-α-D-glucose

K00973

dTDP glucose 4,6-dehydratase

dTDP-4-dehydro-6-deoxy-α-D-glucopyranose

K01710

dTDP-4-dehydro-6-deoxy-D-glucose 3-epimerase

K13312

K19855

dTDP-4-dehydro-6-deoxy-α-D-gulose

dTDP-4-dehydro-6-deoxy-deoxy-α-D-gulose 4-ketoreductase

K13313

dTDP-6-deoxy-α-D-allose

tylosin biosynthesis

mycinamicin biosynthesis