A systematic approach to classify and characterize genomic islands driven by conjugative mobility using protein signatures

Bioteau Audrey, Nicolas Cellier, Frédérique White, Pierre-Étienne Jacques, Vincent Burrus

Nucleic Acids Research, 2023 **1**
https://doi.org/10.1093/nar/gkad644

A systematic approach to classify and characterize genomic islands driven by conjugative mobility using protein signatures

Bioteau Audrey [1], Nicolas Cellier [2], Frédérique White[1], Pierre-Étienne Jacques [1,*] and Vincent Burrus [1,*]

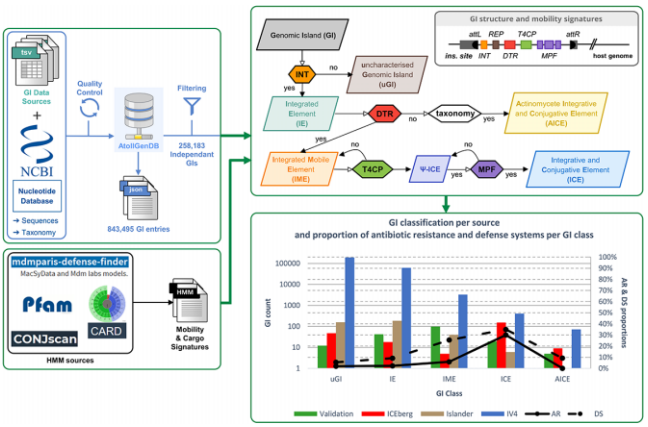
[1]Département de biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada and [2]CGI IMT Mines Albi, Albi, France

Received November 14, 2022; Revised July 17, 2023; Editorial Decision July 18, 2023; Accepted July 21, 2023

ABSTRACT

Genomic islands (GIs) play a crucial role in the spread of antibiotic resistance, virulence factors and antiviral defense systems in a broad range of bacterial species. However, the characterization and classification of GIs are challenging due to their relatively small size and considerable genetic diversity. Predicting their intercellular mobility is of utmost importance in the context of the emerging crisis of multidrug resistance. Here, we propose a large-scale classification method to categorize GIs according to their mobility profile and, subsequently, analyze their gene cargo. We based our classification decision scheme on a collection of mobility protein motif definitions available in publicly accessible databases. Our results show that the size distribution of GI classes correlates with their respective structure and complexity. Self-transmissible GIs are usually the largest, except in Bacillota and Actinomycetota, accumulate antibiotic and phage resistance genes, and favour the use of a tyrosine recombinase to insert into a host's replicon. Non-mobilizable GIs tend to use a DDE transposase instead. Finally, although tRNA genes are more frequently targeted as insertion sites by GIs encoding a tyrosine recombinase, most GIs insert in a protein-encoding gene. This study is a stepping stone toward a better characterization of mobile GIs in bacterial genomes and their mechanism of mobility.

GRAPHICAL ABSTRACT



INTRODUCTION

Genomic islands (GIs) are large, discrete, and unstable portions of chromosomal DNA that usually encode mobility functions enabling their intra- and/or intercellular mobility, and adaptive functions enhancing the bacterial host's fitness and survival. These functions include multi-drug and heavy metal resistances, pathogenicity and colonization, toxins, alternative metabolic pathways, or anti-phage defense systems (1,2). The term 'genomic island' encompasses diverse types of mobile genetic elements that exhibit various structures and gene contents, including prophages, transposons, integrated plasmids, integrative and mobilizable elements (IMEs), and integrative and conjugative elements (ICEs) (3,4). Although mobility is usually considered an essential component in the GI class depiction, flexible clusters of syntenic genes simply coding for a specific biological function, such as Gram-negative bacteria O-polysaccharide chain synthesis, and illegitimate recombination products also fall under that category (5,6).

*To whom correspondence should be addressed. Tel: +1 819 821 8000 (Ext 65914); Fax: +1 819 821 8049; Email: pierre-etienne.jacques@usherbrooke.ca
Correspondence may also be addressed to Vincent Burrus. Tel: +1 819 821 8000 (Ext 65223); Fax: +1 819 821 8049; Email: vincent.burrus@usherbrooke.ca
Present address: Bioteau Audrey, INRAE Centre Île-de-France, Jouy-en-Josas, France.

When known, a key feature distinguishing each GI subclass is their mechanism of intracellular and intercellular mobility. Intracellular mobility, i.e. the ability of a GI to move from one chromosomal location to another, is often mediated by dedicated DDE transposases or integrases belonging to the family of site-specific serine or tyrosine recombinases (7). Intercellular mobility, the hallmark of horizontal gene transfer, relies on natural transformation, transduction, or conjugation. Transformation is the capture and incorporation of free DNA available in the surrounding of a competent cell and resulting from DNA secretion or DNA release upon the death and lysis of a donor cell. GIs can be occasional riders during this process and passively transmitted between hosts of the same or different species (8). Likewise, general transduction mediates the passive dissemination of randomly encapsidated GIs within viral particles encoded by a transducing phage. A subset of GIs, the phage-inducible chromosomal islands, act as parasites of transducing phages to spread within bacterial populations (9,10). Finally, self-transmissible GIs include prophages, ICEs, and Actinomycete ICEs (AICEs). ICEs disseminate by conjugation, a mechanism involving the secretion of DNA from the donor cell into a recipient cell (11). For ICEs, the DNA is translocated between mating cells in direct contact by a type IV secretion system (T4SS), a multiprotein complex that spans the donor cell wall. For AICEs, the DNA is translocated as a double-stranded molecule by a DNA translocase of the FtsK/SpoIIIE family (12,13). Integrative and mobilizable elements (IMEs), another subset of GIs, spread via the conjugative apparatus encoded by a helper ICE or conjugative plasmid (3). This process is referred to as mobilization (14,15).

Detecting GIs in bacterial genomes can be performed by seeking local disparities within the genome sequence, such as differences in the GC content, codon usage, gene composition and organization, or the presence of short repeats (4,16). Following the idea of distinct gene composition and organization, a mechanism of GI mobility can be proposed by searching for a specific composition of genes involved in its intra- and intercellular mobility. Examples of mobility markers include integrase genes or clusters of genes involved in a single function (module), such as the assembly of the T4SS or replication process. Mobility-related genes can be searched for homology using a chosen reference against either whole protein sequences or specific protein motifs. MacSyFinder (17) uses this approach in addition to comparative genomics to detect the presence of essential mating pore formation genes and subsequently predict ICEs. Its pipeline considers the presence of specific genes through the CONJScan signature library, their relative position, and multiple conserved host genes.

Wet lab detection of GIs is often a moment of pure serendipity consecutive to the observation of transmission of a selectable or screenable phenotype such as antibiotic resistance. The discovery and study of new elements harbouring a cargo of adaptive genes can be tedious, time- and resource-consuming. However, it has been and remains a keystone upon which *de novo* prediction of GIs in bacterial genomes can be built. Experimental validation is also necessary to confirm bioinformatic predictions. After curation, GI data can be stored in and retrieved from many dedicated databases, such as ICEberg (ICE database) (18,19), Islander (mobile genomic island database) (20), or the current most up-to-date and comprehensive IslandViewer4 (IV4) database (21). However, identifying the subclass to which a specific GI belongs and predicting its mobility remains highly challenging. There is no easy way to infer *de novo* the precise extremities of GIs and identify their insertion site with single-nucleotide accuracy without relying on comparison with closely related reference genomes. A short target sequence duplication usually flanks GIs integrated by DDE transposases and serine integrases (large unidirectional type). Tyrosine recombinases often lead to a longer yet imperfect target sequence duplication, limiting their usefulness for accurate insertion site prediction (22).

This work aims to solidify the classification and characterization of GIs found in two curated datasets and three publicly accessible databases using mobility protein motif definitions. We focused on integration to better understand the genetic context of GI integration into a specific site at a systematic and large-scale level. We also assessed the types of gene cargo carried by GIs and more complex elements that help their host mitigate selective pressure exerted by the environment, such as phage infection and antibiotics.

MATERIALS AND METHODS

Genomic island sources and processing

A set of 112 GI entries gathered from the Burrus laboratory's resources (hereafter referred to as 'A. Bioteau') was used for prototyping and to serve as a control group. Next, entries were gathered from the databases Islander, ICEberg, IV4, and the 'J. Lao' *Streptococcus salivarius* dataset (Supplementary Table 1) (18–21,23). IslandViewer as a tool to generate IV4's dataset integrates three distinct GI detection methods and pre-computed results. SIGI-HMM, which measures codon usage; IslandPath-DIMOB, which measures dinucleotide bias, the presence of 8 + distinct ORFs and at least one mobility gene; and IslandPick, an automated comparative genomics-based method.

These heterogeneous GIs data were parsed and harmonized into a single database named AtollGenDB. NCBI accession identifier of the host and relative start/end coordinates of GIs were the necessary information required for incorporation into AtollGenDB and subsequent analyses. The fasta sequence and the DocSum description files of each unique host were fetched from NCBI through the Entrez API to extract the corresponding GI sequences and their immediate genomic environment (up to 2 kb upstream and downstream). The taxonomy ID and various complementary metadata were also extracted from the DocSum.

A size filter was first applied to remove spurious GIs and artifacts (<5 kb or >1 Mb), removing ∼0.2% of entries. Then, identical GI entries from AtollGenDB (same accession/start/end from different sources and detection methods) were combined into single unique entries to remove duplicates. Finally, a filter on the GI genomic environment availability (ability to fetch the GI host sequence) was applied, resulting in a total of 536 233 valid entries for analyses.

Considering at that stage that about half of the GIs were overlapping, two different data groups were produced. To

Table 1. Mobility categories. Description of each of the five mobility modules

Mobility Category	Sub-category	Description
INT	INT_Tyr	Integration module
	INT_Ser	CDS matching this category's protein signatures allow their island's insertion in the host genome. These are looked for in the 1st and last 5000bp of the island sequence
	DDE	
DTR	DTR	DNA linking module
	MPF_DTR_like	Also referring to the relaxosome, allows linking to and nicking of the DNA sequence at the origin of transfer (*oriT*)
	INT_DTR_like	
T4CP	T4CP	Relaxosome coupling module
		Connects the relaxosome to the secretion system for ulterior dissemination
MPF	MPF	Mating pore formation module
		Proteins harbouring these signatures may be part of the secretion system
REP	REP	Replication module
		Protein signatures related to the polymerase's activities and DNA replication

avoid redundancy when analyzing the content of the GIs, the group of 'Independent' GIs, therefore not overlapping any other GI from AtollGenDB, was created and used in the main figures. To confirm that this group remains representative of the complete collection, a group composed of all unique filtered GIs was also created and named 'Overlapping' (results presented in supplementary figures). Supplementary files containing information on the complete collection of unique filtered GIs as well as the analysis outputs are available on figshare (https://doi.org/10.6084/m9.figshare.21440952).

Classification of GIs based on mobility protein signatures

The classification of GIs was based on the presence of mobility proteins selected from a collection of Hidden Markov Model (HMM) signatures. They originated from the PFAM database (PFAM-A, v.34.0) (24) and CONJScan (25), and are organized into five different modules (Table 1). The integrase (INT) module contains HMM signatures for tyrosine (INT_Tyr) and serine (large unidirectional type, INT_Ser) recombinases, as well as DDE transposases (DDE). The DNA transfer and replication (DTR) module is represented by the general DTR-related HMM signatures and two specific signatures of DTR-like proteins, which are the INT HMM signature PF12835 (INT_DTR_like) and the MPF HMM signature PF01580 (FtsK/SpoIIIE, MPF_DTR_like). The three other modules are type IV coupling protein (T4CP), mating pore formation (MPF), and replication (REP). The collection of signatures used for the classification is available in Supplementary Tables 2 and 3.

The classification workflow was developed in Python v.3.8, using snakemake for local use, and click/slurm to use on parallelized clusters. Briefly, Prodigal v.2.6.3 (26) was first used on the nucleic acid sequence of each GI and its environment to extract the most probable coding sequences (CDSs) as amino acid sequences. HMMER3 v.3.3.2 (27) was then used with the default parameters to identify protein motifs matching the provided HMM signatures of the five mobility modules. To be considered in the classification process, the e-value threshold of the signatures detected must be under 1e-2, and the coverage factor greater than or equal to 0.6. Each GI was then classified based on the rules presented in Figure 1, developed using the A. Bioteau curated dataset. This dataset was carefully examined, showing that the predicted class was correct in all cases, the only discrepancies being caused by frameshifts or partial sequence data in input GI sequences (e.g. GIs spread over multiple contigs), or HMM signatures missing from our set (Supplementary Table 4).

The integrative mobile elements (IMEs) and ICE class definitions were mostly tailored around mobile genetic elements studied in V. Burrus's lab, including the particular case of the *Salmonella* genomic island 1 (SGI1). SGI1 uses an atypical relaxase (*mpsA*) belonging to the tyrosine recombinase family (INT_DTR) as well as 3 MPF genes (*traG*, *traN* and *traH*) (28,29). The Ψ-ICE is a necessary 'in-between" class mainly lacking the *virB4* gene, mandatory for the mating pore formation and ultimately for the element transfer capacity. Unfortunately, the AICE class definition is incomplete as several HMM signatures to identify them all are currently missing (e.g. RepPP (30)).

Additional annotations and analyses

Additional annotations were added to each GI entry. tRNAScan-SE v.2.0.9 (31) was used to identify tRNA genes in the GI environment. We extended the tRNA search within 300 bp on both GI extremities to consider possible cases where borders include the insertion site. Moreover, the CDS identified by Prodigal in each GI were used by custom adaptations of the RGI v.6.0.2 module of CARD v.3.2.6 (32) and Defense-Finder v.1.1.2 (17,33) tools to extract the antibiotic resistance determinants and phage defense systems, respectively. The complete annotations of the GIs were exported as JSON files integrated within AtollGenDB. Considering that antibiotic resistance-related CDSs annotated by CARD are often associated with multiple drug classes and their related resistance mechanisms, we decided to use the resistance mechanisms representation.

The insertion site of a given GI is defined by the closest CDS or tRNA found either from the start or end coordinate of the island's sequence. An insertion within the first or last 50 nt of a CDS is labeled as a '5′/3′ CDS' insertion. As GI boundaries are often imprecise, an additional 20 nt upstream and downstream of the GI are also included for this label. Any insertions anywhere else within a CDS are labeled 'Disrupted CDS', likely impairing the function. 'tRNA' refers to insertion within a tRNA gene and 'Intergenic' at >20 nt from any gene. 'Ambiguous' refers to insertion within two distinct annotated CDSs or tRNAs. 'ND' (not determined) corresponds to either a well-defined GI provided without its flanking sequences or to a GI existing as a contig.

We manually constructed the taxonomic groups from the NCBI Taxonomy browser (Supplementary Table 5). Briefly,

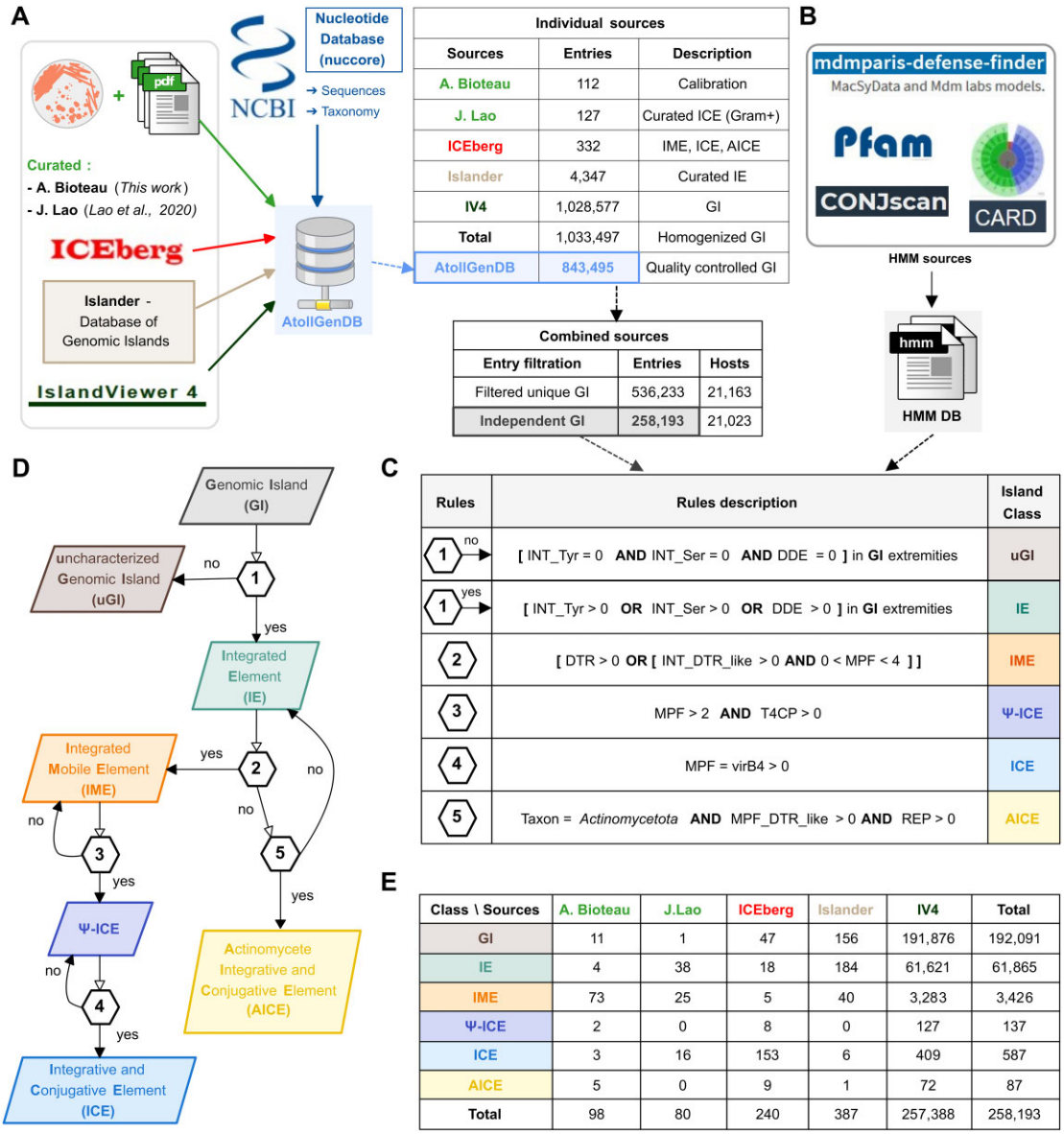


Figure 1. GI classification from AtollGenDB. (**A**) Number of raw entries for each of the five GI sources centralized in AtollGenDB. Quality control is performed on the total number of GIs to obtain AtollGenDB data to be analyzed, completed by sequence and taxonomy information from the NCBI Nucleotide database (nuccore). The number of unique and independent islands are then extracted. (**B**) HMM signature sources used for the classification and characterization steps. (**C**) Description of the five classification rules. (**D**) Description of the classification decision tree. (**E**) Classification results of the 258193 independent GIs by source.

we collected taxonomic information through each entry's DocSum file via their taxonomy identifier (taxID). Taxonomy counts were made according to the entries' taxa rank. Considering the amount of data in the Pseudomonadota's Gamma-proteobacteria class, its phylogenetic distribution and diversity, we chose to detail it into its most frequently represented orders. All other classes were attributed to the 'Pseudomonadota others' category. Statistical tests were conducted to assess the average variability of GI sequence length distribution across the 8 most represented taxa (top-8 taxa) (Supplementary Table 5 & Supplementary Figures 1 and 2). As data generally followed a log-normal distribution, a two-sided Welch's *t*-test was used on log-transformed data (no assumption made on standard deviation), relative to the Bacillota phylum.

RESULTS AND DISCUSSION

Classification of GIs from public databases

AtollGenDB was created to gather and harmonize a heterogeneous dataset of more than 1 million GIs from five different sources (Figure 1A). More than half of these GIs were retained after the combination and filtering steps (see Materials and Methods for details). However, 258 193 GIs from 21 023 different hosts did not overlap with any other GI, and for this reason, were labeled as independent and used in subsequent analyses to avoid data redundancy. The 536 233 unique GIs, of which many had sequence overlap with other GIs, were also analyzed for comparison purposes, resulting in the same downstream conclusions. This measure would be evitable if the boundaries of all the GIs were accurately

known, curated and reliable, which is rarely the case when GIs are predicted from bacterial genomes.

The GIs were then classified based on the presence of a combination of mobility protein signatures represented by HMM signatures from various sources (Figure 1B). The classification rules and the companion decision tree (Figure 1C, D, Table 1) allowed the attribution of each GI to one of the following classes: uncharacterized GIs (uGIs), integrated elements (IEs), integrative and mobilizable elements (IMEs), integrative and conjugative elements (ICEs), pseudo-ICE or putative degenerated ICE (Ψ-ICEs), and Actinomycete ICE (AICEs). An IE possesses at least one gene coding for an integrase or a transposase at one extremity (Integration module). An IME also includes a gene coding for at least one mobilization protein that can be a relaxase or an *oriT*-binding protein (DTR module). Adding to the complexity, an ICE also carries genes coding for a type IV secretion system (T4SS, MPF module) and a type IV coupling protein (T4CP) that confer self-transmissible properties. A Ψ-ICE is either an ICE in a decaying state or an incomplete ICE due to missing or spurious sequences. Finally, an AICE differs from an ICE by the absence of T4SS genes, replaced by a gene coding for a replication protein (Rep) and a gene coding for a double-stranded DNA translocase (MPF_DTR_like) to ensure self-transmissibility (34,35).

Application of these rules correctly classified the GIs from the curated A. Bioteau dataset, providing high confidence for the classification of the GIs from IV4 (Figure 1E). Indeed, all observed differences were easily explained by truncated signature sequences and NCBI records updates and GI sequence spanning several contigs (Supp. Tab. 4). We also compared our classification to the 19 ICE sequences from the ICEberg database selected by Cury *et al.* (36) on the basis of experimental validation, and observed coherent results where the few classification differences were due to missing signatures or low e-value/coverage, such as a missing transposase signature resulting in a uGI classification and an undetected VirB4 signature due to low coverage resulting in a Ψ-ICE classification (Supplementary Table 6). Interestingly, comparing the impact of IV4's detection methods on our classification results, we observed that the most restrictive detection method (DIMOB) resulted in a more precise classification as about half of its predictions (55% on average) are categorized as uGI compared to the three other IV4 methods (Supplementary Table 7). The ambiguity of the Ψ-ICE definition precluding any further interpretation regarding their biological significance, this class was not considered further in our downstream analyses, though related data are available in the supplementary tables.

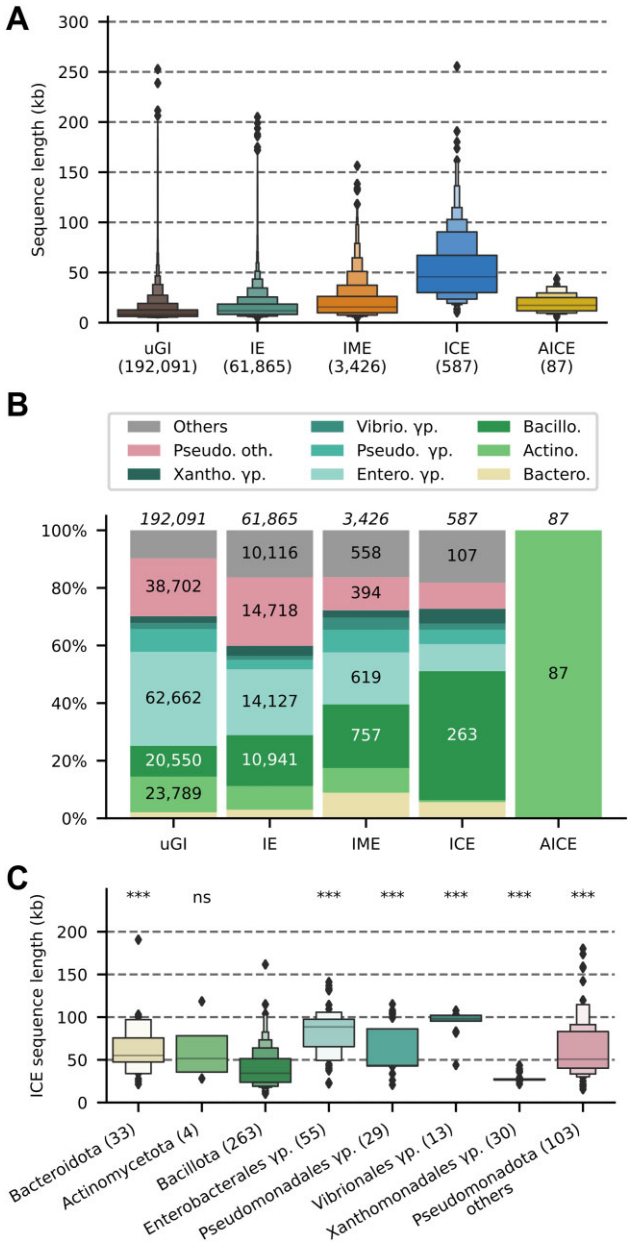


Figure 2. Sequence length and inter-taxa distributions. (**A**) Sequence length distribution of independent GIs across classes (total numbers in parenthesis). Two data points from the GI class exceeding 300 kb are not shown. (**B**) Distribution of top-8 taxa per GI class. (**C**) Sequence length distribution of ICEs across top-8 taxa. Statistical significance was calculated using Welch's *t*-test and the Bacillota phylum data as the reference (ns = not significant; ***$P < 0.001$).

Sequence length distribution supports GI class complexity and inter-taxa variability

As an initial assessment of the data resulting from the classification, we measured and compared the sequence lengths of GIs in each class and top-8 taxa (Figure 2A and Supplementary Figure 3). Overall, the sequence length distribution of every GI class was in agreement with what one can expect from their respective structure, from the typically smaller non-self-transmissible GIs (which may include pathogenicity or symbiotic islands, transposons) to the more complex self-transmissible ICEs. AICEs are known to be significantly smaller than ICEs in part due to their minimalistic DNA translocation apparatus (13). Next, we selected the top-8 represented taxa in our datasets and examined the proportion of each GI class they comprise. Interestingly, GIs and IEs were overrepresented in Enterobacterales and Pseudomonadota in general, whereas nearly 50% of ICEs were found in Bacillota (Figure 2B). More than half

of all IMEs originated from these three taxa. Due to the upstream Actinomycetota filtering, AICEs were exclusively found in Actinomycetales, as further supported by a similar recent study from Botelho (37). We observed similar trends using the overlapped dataset (available as supplementary material). As ICEs stood out from other GI classes in average length and complexity, we assessed their distribution across the top-8 taxa presented above (Figure 2C). ICEs found in Bacillota tend to be much smaller than those found in the other taxa, particularly the Enterobacterales and Vibrionales. This trend, already suggested by the small sizes of the prototypical ICEs Tn*916* from *Enterococcus faecalis* (18 kb) or ICE*Bs*1 from *Bacillus subtilis* (∼20.5 kb) relative to the larger sizes of SXT from *Vibrio cholerae* (99 kb), ICEclc from *Pseudomonas knackmussii* (103 kb) or CTnDOT from *Bacteroides* spp. (65 kb) (38–42), is here confirmed on a larger dataset. This difference may result from the additional T4SS subunits needed to span the outer membrane of Gram-negative bacteria. Other evolutionary innovations such as the separation of MPF and DTR genes over multiple operons under the control of a unique transcriptional activator may also have provided the flexibility for extra cargo genes acquisition without impairing the ICE's transmissibility. For instance, MPF and DTR genes of SXT/R391 ICEs are part of six distinct transcriptional units whose expression is activated by SetCD (43). In contrast, MPF and DTR genes organized as a single long transcriptional unit, as exemplified by Tn*916* and related ICEs (44), reduce the likelihood of cargo gene acquisition.

Comparison of integration modules and insertion sites between GI classes

The insertion of integrative elements (e.g. temperate phages and ICEs) is mediated by DNA recombinases (integrase or DDE transposase). The gene coding for these enzymes usually lies at or near one of the two extremities of the integrated element. We addressed the composition of the integration module for each GI class using dedicated protein signatures. Our analyses revealed that site-specific tyrosine integrases (INT_Tyr) are prevalent in IMEs, ICEs and AICEs (Figure 3A and Supplementary Figure 4). Serine integrases (INT_Ser) have a slightly higher prevalence in AICEs but remain below 25%. This observation could result from a sample bias as identified AICE representatives are rare in databases ($n = 87$). Strikingly, 60% of IEs use a DDE transposase to integrate, an enzyme rarely used by other GI classes. Combinations of integrases of different types, i.e. INT_Tyr and INT_Ser protein signatures found in the same CDS, were not explored further as they may be artifacts. Nearly all tyrosine integrases harboured the C-terminal catalytic PF00589 domain, combined with other signatures in their N-terminal half (e.g. PF02899, PF13102, or PF13495) (Supplementary Table 8). These domains are typically associated with binding to arm-type DNA sites within the *attL*, *attR*, and *attP* attachment sites (45). The arm-type sites also contain binding sites for other proteins, including recombination directionality factors (RDF). RDFs are small fast-evolving DNA-binding proteins encoded by a gene located within the integration module. They help displace the recombination reaction toward the excision of the integrated element (46). Since RDFs are unreliable markers for detecting integrative elements due to their small size and vast diversity, they were not sought in this study.

We next looked for associations between the type of insertion sites (based on the distance to the closest CDS or tRNA) and the integration module of the GIs (Figure 3B). DDE transposase remained the most represented category across all insertion sites, being the most versatile, except for tRNA genes and ND for which tyrosine integrases are by far the most prevalent (Figure 3C). The overwhelming majority of GIs integrate into an intergenic region or at the 5′ or 3′ end of a protein-coding gene, while a minority disrupt a CDS (Figure 3D). All GIs disrupting CDSs might not get referenced as such however, as CDS themselves may be miss-annotated by Prodigal due to sequence breakage by GI insertion. Surprisingly, tRNA-encoding genes are not the most frequent target sites (6,20,47). Most GIs seem to favour intergenic regions or the 3′ end of a gene coding for a protein or its 5′ end, though more rarely. No specific trend emerged when comparing the insertion sites between GI classes, except for the depletion of tRNA gene insertion for IEs. A sensitivity analysis on the detection methods identifies as expected a preferential insertion of IE in tRNA genes for Islander (20) and an over-representation in ND for ICEberg considering the lack of genomic environment, while an unexplained preferential insertion of IE and IME into CDS is observed for Islandpick (Supplementary Table 9). Insertions in ambiguous loci (two distinct genes interrupted by a GI sequence) highlights the critical need for well-defined GI coordinates. To further consolidate GI sequence borders, their systematic and accurate definition as well as experimental validation inventory of the direct repeats resulting from the integration events are urgently needed.

Our classification method cannot take into account composite or aggregated elements, such as tandem IMEs/ICEs (34,48) or tripartite ICEs, such as ICE*Mc*Sym1271 (49). Currently, a composite element would either be classified as one element or tagged as 'overlapping'. The β and γ segments of ICE*Mc*Sym1271 would likely be classified as IE or IMEs, while the α region would fall in the uGI class as it contains no integrase genes.

GI cargo diversity

To better evaluate the ecological importance of GIs for their hosts, we investigated the presence and distribution of two types of cargo genes, those involved in antibiotic resistance and defense systems against bacteriophage infection.

Antibiotic resistance genes. Multidrug resistance is a pressing global issue in animal husbandry and healthcare systems. Since the mid-1970s, extensive efforts have been deployed to understand the emergence and dissemination of multidrug-resistant bacteria, focusing on mobile genetic elements bearing antibiotic resistance gene cargo. Consequently, these elements and their hosts are likely to be over-represented in databases. Indeed, the top-8 taxa represented in our datasets mostly contain pathogenic species affecting animals and humans (Supplementary Table 5). Surprisingly, we found that a tiny fraction of GIs and IEs harbour

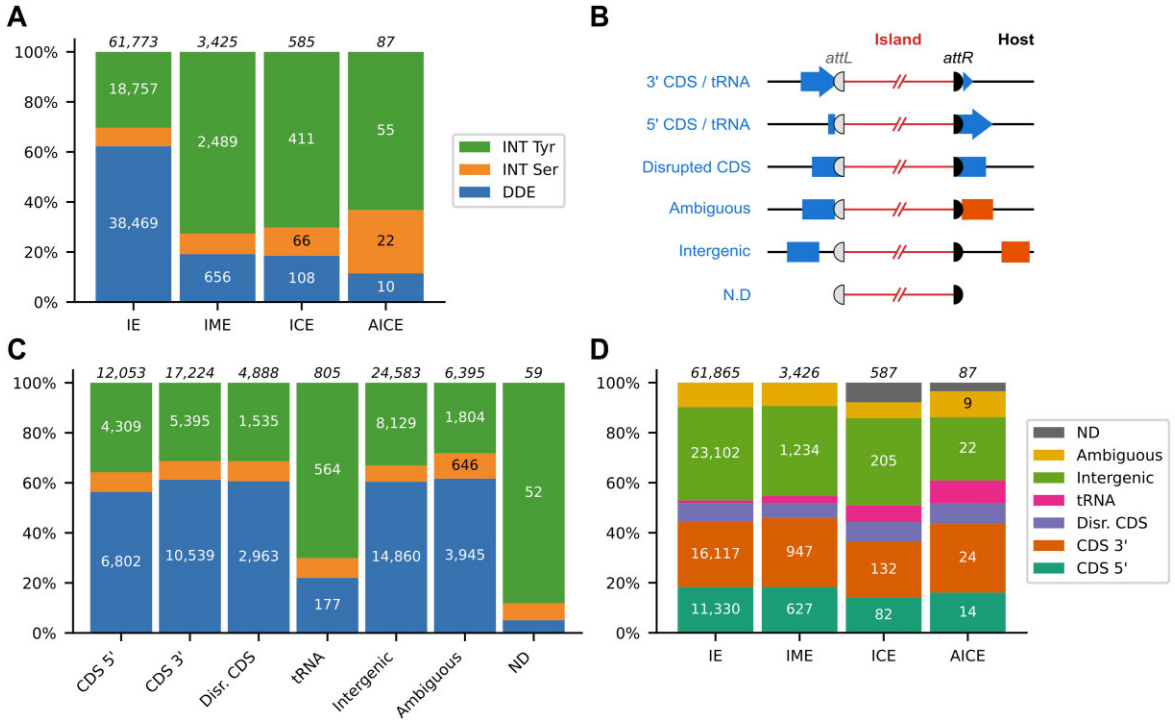


Figure 3. Integration modules and insertion sites characterization. (**A**) Distribution of integrase types across the various GI classes. The few GIs (95) containing more than one integrase type in the same CDS are not shown in the bars. (**B**) Illustration of the insertion site categories (see Material and Methods for definitions). Black hemispheres correspond to GI boundaries. (**C**) Distribution of integrase types across insertion sites. (**D**) Distribution of insertion sites across GI classes.

antibiotic resistance genes referenced in the CARD database (Figure 4A and Supplementary Figure 5). However, nearly 6% of IMEs and 30% of ICEs carry at least one antibiotic resistance determinant, with 5.5% of ICEs bearing five or more. Antibiotic resistance is a rare occurrence in AICEs, though the small sample size could explain this observation (Figure 4A and C). Soil-dwelling Actinobacteria such as *Streptomyces* are known to be prolific producers of antibiotics and are themselves highly-resistant microorganisms (50). In producers, resistance can result from the absence of the target (e.g. lack of dihydrofolate reductase targeted by trimethoprim), antibiotic modification or efflux, or target alteration. Non-mobile resistance genes might be more favourable to producers than those located on AICEs due to improved stability in a context of intense selective pressure exerted by the myriad of antibiotics produced in the soil and reduced odds of sharing them with competitors. The bulk of GIs harbouring an antibiotic resistance gene cargo is found primarily in the top-8 taxa (Figure 4B). Drug efflux is the most common resistance mechanism found across all taxa and prevails in the Enterobacterales. Target protection accounts for nearly 62% of the resistance mechanisms in the Bacillota. GIs encode the bulk of drug efflux, which represents ∼80% of all resistance mechanisms in this class. Drug efflux represents 61% of the resistance mechanisms across all classes, followed by antibiotic inactivation (19.4%, mainly in IEs). Integrative elements (IEs, IMEs and ICEs) exhibit comparable resistance mechanism distributions that are strikingly different from those observed for GIs. However, ICEs encode target protection mechanisms more frequently than other integrative elements (Figure 4C, D, proportions conserved with Overlaps data). Based on our observations, general drug efflux mechanisms seem to be more frequently associated with chromosomal structures lacking apparent mobility features, suggesting a more stable and ancient association with the host. The accumulation and diversity of drug-resistance genes linked to self-transmissible elements (ICEs) highlight a clear advantage for their survival and spread within bacterial populations subjected to antibiotic selective pressure (11,51).

Defense systems. Besides antibiotic resistance genes conferring a selective advantage in stressful conditions, GIs often harbour other selectable traits such as defense systems against bacteriophage predation. Phages outnumber bacterial cells in the environment, exerting considerable selective pressure for the acquisition, evolution and dissemination of GIs encoding defense mechanisms such as restriction-modification (RM), abortive infection (Abi), and CRISPR-Cas systems (52). To assess the diversity and prevalence of anti-phage systems associated with GIs belonging to different classes, we probed our datasets using Defense-Finder (17,33). As observed above with antibiotic resistance genes, most GIs are devoid of known anti-phage functions (Figure 5A and Supplementary Figure 6). However, roughly 20% of IMEs and 30% of ICEs harbour at least one anti-phage system. Our results confirm that a bacterial host or GI can accumulate multiple defense systems (25). A few GIs encoded more than three anti-phage mechanisms each. The distribution of the top-10 anti-phage systems across the top-8 taxa was relatively homogeneous, with no specific trends (Figure 5B). The five types of restriction-modification systems

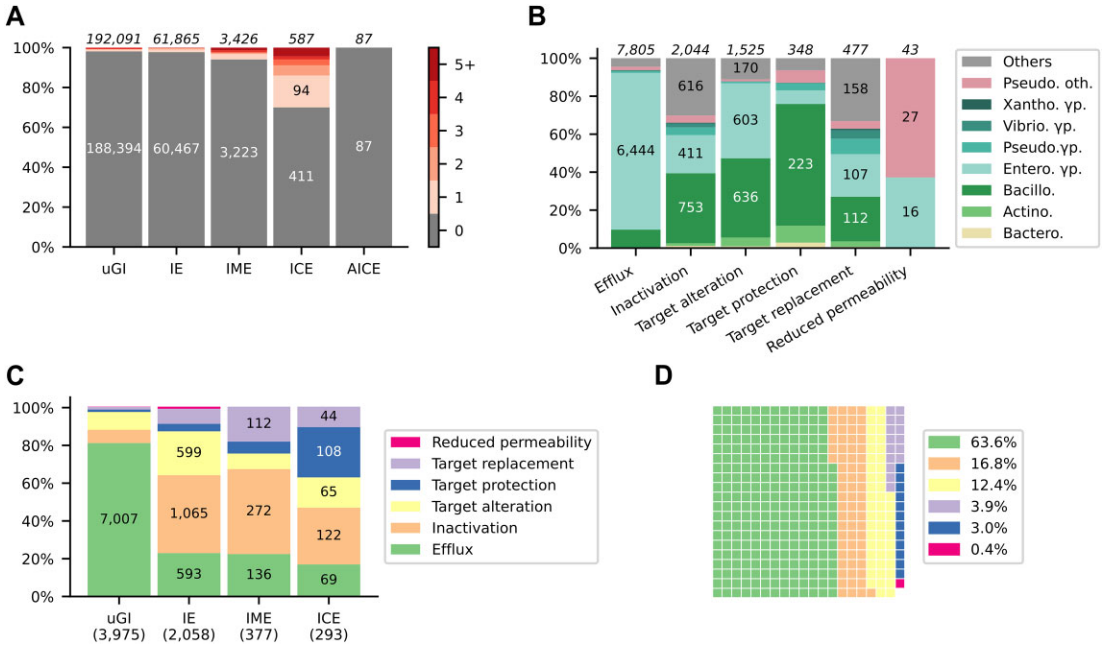


Figure 4. Diversity of antibiotic resistance mechanisms in cargo. (**A**) Distribution of GIs harbouring antibiotic resistance-related CDS per GI class. (**B**) Distribution of top-8 taxa per antibiotic resistance mechanism. (**C**) Distribution of antibiotic resistance mechanisms CDS count across GI classes. GIs containing multiple mechanisms are counted more than once. (**D**) Waffle chart of antibiotic resistance mechanisms distribution across all GIs bearing antibiotic resistance.

(RM type I, II, IIG, III and IV) correspond to 40.9% of all anti-phage systems found, followed by Cas (11.8%) and Abi (2 and Eii, 11.1%) systems (Figure 5C, D). Surprisingly, RM systems are more frequent in uGIs, IEs, and IMEs than ICEs. Retron, Lamassu, and Gabija emerged when looking at the top-represented systems in uGIs, IMEs, and ICEs, respectively. We observed a broad diversity of combinations of anti-phage systems in a single GI sequence (Supplementary Figures 7 and 8). We could infer patterns defining specific GI classes, but this trend was inconsistent when comparing overlap data.

The study of defense systems is a fast-developing field with new systems being discovered regularly and several mechanisms of resistance to phage infection remaining to be characterized. For this reason, we expected to miss annotations, but also observed a vast proportion of GIs displaying such defense mechanisms.

Combination of antibiotic resistance and defense system cargos. Combining the data on antibiotic resistance and anti-phage systems puts into perspective the defensive cargo carried by GIs (Figure 6A and Supplementary Figure 9). Once again, the ICE class shows the most extensive defensive cargo as more than half of ICEs carry either one or both (37,48,53). Bacillota, Enterobacterales and Pseudomonadales are the most examined taxa for antibiotic resistance since they contain the most common pathogenic species for humans and animals. GIs in these three taxa harbour the highest proportions of antibiotic-resistance genes, whereas anti-phage systems dominate the other taxa (Figure 6B).

ICEs are large complex elements that likely impose a metabolic burden on their host and lower their fitness. Therefore, the accumulation of functions that enhance cell survival is critical to enhancing the persistence of ICEs in the genome of their host. Our antibiotic resistance and defense system cargo study confirms the trends observed in recent large-scale studies (33,37,52,54).

CONCLUSION

This study demonstrated that mobility protein features can be exploited for identifying and classifying prokaryotic GIs. Using curated datasets, we designed the classification rules around these mobility protein signatures to assign a high number of sequences to specific GI classes. We found that increased GI complexity, which culminates with the ICE class, correlates with a larger cargo of adaptive traits.

ICEs, as self-transmissible GIs, require a complete set of mobility genes ensuring the integration into and excision from the chromosome, T4SS assembly, and DNA processing (relaxosome components and T4CP). Also self-transmissible, AICEs have a simpler DNA translocation machinery and carry a dedicated replication module. The integration module, whose characterization in this study lacked directionality factor-related protein signatures, plays a critical role in selecting the insertion site. A comprehensive inventory associating integrases with their cognate or preferred insertion site(s) (relaxed sequences such as AT-rich region, e.g. Tn*916* integrase (55)) is lacking, though needed to facilitate the *in silico* prediction of the boundaries of GIs. According to the dataset, tyrosine integrases mediate integration more frequently at the 5′ or 3′ of CDS than tRNA genes. Proportionally, larger GIs such as ICEs will display a variety of antibiotic resistance mechanisms and defense

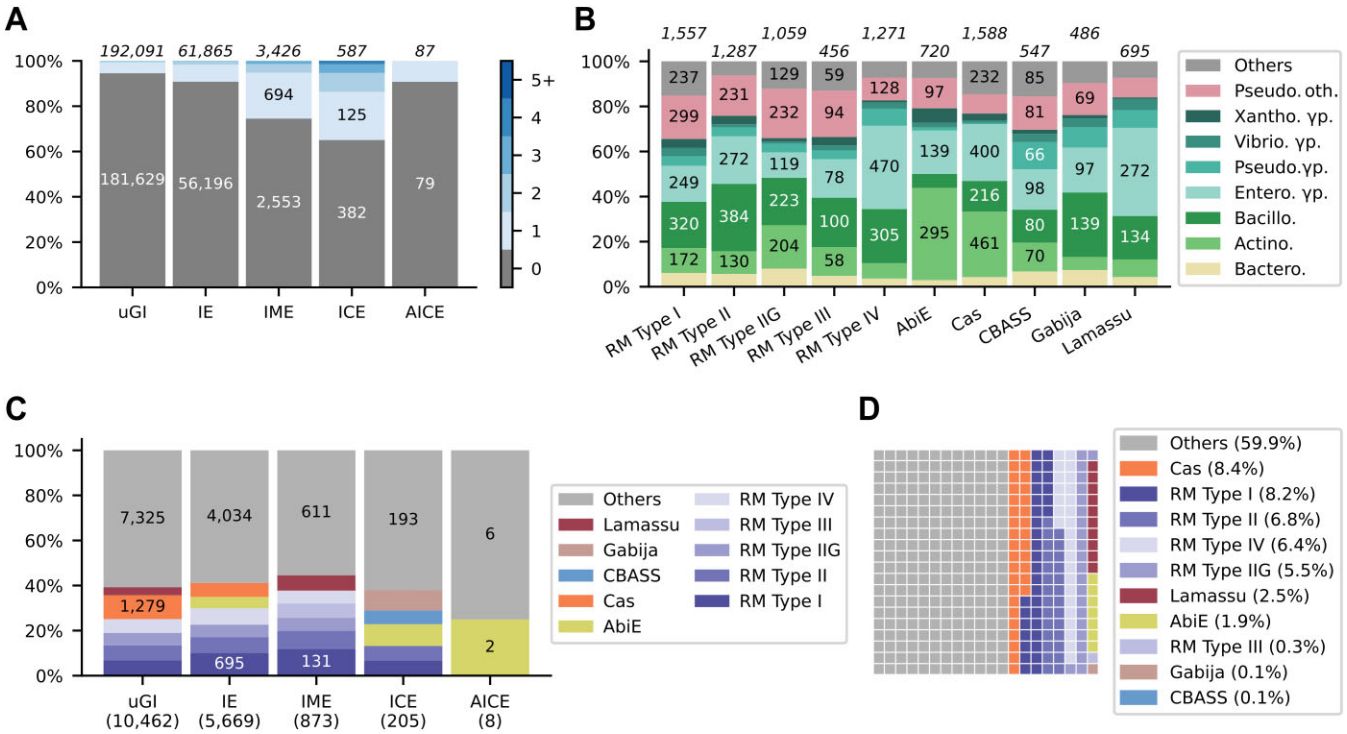


Figure 5. Diversity of defense systems in cargo. (**A**) Distribution of distinct and complete defense systems found per GI class. (**B**) Distribution of top-8 taxa across the top-10 defense systems found in data. (**C**) Distribution of the top-6 defense systems found in each GI class. (**D**) Waffle chart of the top-10 defense system distribution.

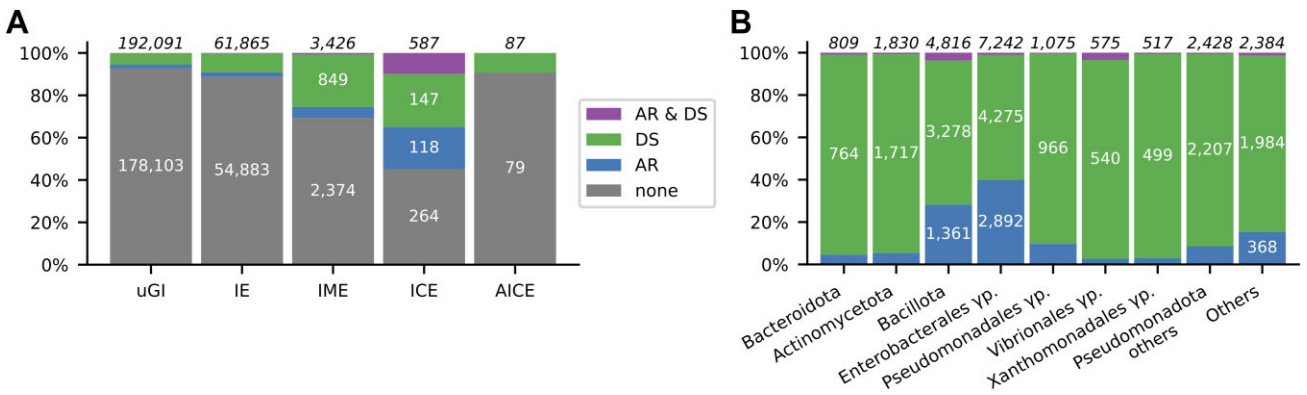


Figure 6. Coexisting antibiotic resistance mechanisms and defense systems in cargo. (**A**) Distribution of the absence or presence of at least one antibiotic resistance (AR) or defense system (DS) gene per GI class. (**B**) Distribution of AR, DS or both for each of the top-8 taxa.

systems (mostly restriction-modification systems) to maintain themselves in their host. The prevalence of RM systems could result from their activity as innate immunity systems targeting improperly methylated invasive DNA to ward off phages and plasmids and as addictive systems to enhance their maintenance and stability in their host (56).

These observations must be taken with caution, however, as the dataset is strongly skewed towards clinical isolates, influencing the cargo study and the types of elements found in the top-represented taxa. To improve the classification rules and mobility models, features like DTR, REP or RDF signatures, as well as *oriT* sequence, should be added. Additional features could also integrate prophages and integrons. The limits of GIs should be defined accurately (through *attL* and *attR* direct repeats detection, for example), as it dictates the quality of the resulting classification.

Composite and aggregated elements classification should also be integrated into a future version of the AtollGen classification method, following the example of tools like MacSyFinder (17) or ICEScreen (57) that use anchors and protein families to delimit genomic regions belonging to a particular element.

There are still gaps in our understanding of GIs and what makes them mobilizable or self-transmissible. They play a significant role in disseminating survival tools to a wide range of pathogenic hosts, which makes their study critical to health research efforts.

DATA AVAILABILITY

AtollGen is an open-source collaborative initiative available in the GitHub repository (https://gitlab.com/atollgen), including:

Python library / command-line utility for running annotation analyses on one or multiple GIs (https://gitlab.com/atollgen/atollgen-cli)

The pipeline used to produce results for this work (https://gitlab.com/atollgen/atollgen-pipeline)

Input data (https://doi.org/10.5281/zenodo.7866495)

Output data with overlapping and independent groups results (https://doi.org/10.6084/m9.figshare.21440952)

Complete documentation (https://atollgen.gitlab.io/docs/)

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are grateful to Nicolas Rivard for his critical reading of the manuscript. We thank all members of the Jacques and Burrus labs for fruitful discussions. This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada (alliancecan.ca).

FUNDING

Discovery Grant [RGPIN-2016-04365; RGPIN-2021-02814] from the Natural Sciences and Engineering Research Council of Canada (NSERC); Project Grant [PJT-153071, PJT-186081] from the Canadian Institutes of Health Research (CIHR to V.B.); Université de Sherbrooke (to P.É.J. and V.B.); A.B. is the recipient of a Fonds de recherche du Québec – Nature et Technologies (FRQNT) doctoral fellowship; P.É.J a Senior Research Scholar from Fonds de recherche du Québec – Santé (FRQS). Funding for open access charge: CIHR [PJT-186081].
Conflict of interest statement. None declared.

REFERENCES

1. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Micro.*, **3**, 722–732.
2. Hentschel,U. and Hacker,J. (2001) Pathogenicity islands: the tip of the iceberg. *Microbes Infect.*, **3**, 545–548.
3. Bellanger,X., Bellanger,X., Payot,S., Leblond-Bourget,N. and Guédon,G. (2014) Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.*, **38**, 720–760.
4. Langille,M.G.I., Hsiao,W.W.L. and Brinkman,F.S.L. (2010) Detecting genomic islands using bioinformatics approaches. *Nat Rev. Micro.*, **8**, 373–382.
5. Rodriguez-Valera,F., Martin-Cuadrado,A.-B. and López-Pérez,M. (2016) Flexible genomic islands as drivers of genome evolution. *Curr. Opin. Microbiol.*, **31**, 154–160.
6. Dobrindt,U., Hochhut,B., Hentschel,U. and Hacker,J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Micro.*, **2**, 414–424.
7. Hickman,A.B. and Dyda,F. (2015) Mechanisms of DNA Transposition. *Microbiol. Spectr.*, **3**, 531–553.
8. Godeux,A.-S., Svedholm,E., Barreto,S., Potron,A., Venner,S., Charpentier,X. and Laaberki,M.-H. (2022) Interbacterial transfer of carbapenem resistance and large antibiotic resistance islands by natural transformation in pathogenic acinetobacter. *mBio*, **13**, e02631-21.
9. Fillol-Salom,A., Martínez-Rubio,R., Abdulrahman,R.F., Chen,J., Chen,J., Davies,R.L. and Penadés,J.R. (2018) Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. *ISME J.*, **12**, 2114–2128.
10. McKitterick,A.C., Hays,S.G., Hays,S.G., Johura,F.-T., Alam,M. and Seed,K.D. (2019) Viral satellites exploit phage proteins to escape degradation of the bacterial host chromosome. *Cell Host Microbe*, **26**, 504–514.
11. Botelho,J. and Schulenburg,H. (2021) The role of integrative and conjugative elements in antibiotic resistance evolution. *Trends Microbiol.*, **29**, 8–18.
12. Chan,H., Mohamed,A.M.T., Grainge,I. and Rodrigues,C.D.A. (2022) FtsK and SpoIIIE, coordinators of chromosome segregation and envelope remodeling in bacteria. *Trends Microbiol.*, **30**, 480–494.
13. Thoma,L. and Muth,G. (2015) The conjugative DNA-transfer apparatus of *Streptomyces*. *Int. J. Med. Microbiol.*, **305**, 224–229.
14. Gogarten,J.P. and Townsend,J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Micro.*, **3**, 679–687.
15. Thomas,C.M. and Nielsen,K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Micro.*, **3**, 711–721.
16. Lu,B. and Leong,H.W. (2016) Computational methods for predicting genomic islands in microbial genomes. *Comput. Struct. Biotechnol. J.*, **14**, 200–206.
17. Abby,S.S., Néron,B., Ménager,H., Touchon,M. and Rocha,E.P.C. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
18. Bi,D., Xu,Z., Harrison,E.M., Tai,C., Wei,Y., He,X., Jia,S., Deng,Z., Rajakumar,K. and Ou,H.-Y. (2012) ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res.*, **40**, D621–D626.
19. Liu,M., Li,X., Xie,Y., Bi,D., Sun,J., Li,J., Li,J., Li,J., Tai,C., Deng,Z. *et al.* (2019) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.*, **47**, D660–D665.
20. Hudson,C.M., Lau,B.Y. and Williams,K.P. (2015) Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res.*, **43**, D48–D53.
21. Bertelli,C., Laird,M.R., Williams,K.P., Lau,B.Y., Hoad,G., Winsor,G.L. and Brinkman,F.S. (2017) IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.*, **45**, W30–W35.
22. Hirano,N., Muroi,T., Takahashi,H. and Haruki,M. (2011) Site-specific recombinases as tools for heterologous gene integration. *Appl. Microbiol. Biotechnol.*, **92**, 227–239.
23. Lao,J., Guédon,G., Lacroix,T., Charron-Bourgoin,F., Libante,V., Loux,V., Chiapello,H., Payot,S. and Leblond-Bourget,N. (2020) Abundance, diversity and role of ICEs and IMEs in the adaptation of *Streptococcus salivarius* to the environment. *Genes*, **11**, 999.
24. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
25. Abby,S.S. and Rocha,E.P.C. (2017) Identification of protein secretion systems in bacterial genomes using MacSyFinder. *Methods Mol. Biol.*, **1615**, 1–21.
26. Hyatt,D., Chen,G.-L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.*, **11**, 119.
27. Finn,R.D., Clements,J., Arndt,W., Miller,B.L., Wheeler,T.J., Schreiber,F., Bateman,A. and Eddy,S.R. (2015) HMMER web server: 2015 update. *Nucleic Acids Res.*, **43**, W30–W38.
28. Carraro,N., Durand,R., Rivard,N., Anquetil,C., Barrette,C., Humbert,M. and Burrus,V. (2017) *Salmonella* genomic island 1 (SGI1) reshapes the mating apparatus of IncC conjugative plasmids to promote self-propagation. *PLoS Genet.*, **13**, e1006705.
29. Kiss,J., Kiss,J., Szabó,M., Szabó,M., Hegyi,A., Hegyi,A., Douard,G., Praud,K., Nagy,I., Nagy,I. *et al.* (2019) Identification and characterization of *oriT* and two mobilization genes required for conjugative transfer of *Salmonella* genomic Island 1. *Frontiers in Microbiology*, **10**, 457–457.

30. Ghinet,M.G., Bordeleau,E., Beaudin,J., Brzezinski,R., Roy,S. and Burrus,V. (2011) Uncovering the prevalence and diversity of integrating conjugative elements in actinobacteria. *PLoS One*, **6**, e27846.
31. Chan,P.P. and Lowe,T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.*, **1962**, 1–14.
32. Alcock,B.P., Raphenya,A.R., Lau,T.T.Y., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.-L.V., Cheng,A.A., Liu,S. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.
33. 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.*, **13**, 2561.
34. 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*, **8**, 337.
35. Durand,R., Deschênes,F. and Burrus,V. (2021) Genomic islands targeting *dusA* in *Vibrio* species are distantly related to *Salmonella* Genomic Island 1 and mobilizable by IncC conjugative plasmids. *PLos Genet.*, **17**, e1009669.
36. Cury,J., Touchon,M. and Rocha,E.P.C. (2017) Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.*, **45**, 8943–8956.
37. 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.
38. Bose,B., Auchtung,J.M., Lee,C.A. and Grossman,A.D. (2008) A conserved anti-repressor controls horizontal gene transfer by proteolysis. *Mol. Microbiol.*, **70**, 570–582.
39. Flannagan,S.E., Zitzow,L.A., Su,Y.A. and Clewell,D.B. (1994) Nucleotide sequence of the 18-kb conjugative transposon Tn*916* from *Enterococcus faecalis*. *Plasmid*, **32**, 350–354.
40. Beaber,J.W., Hochhut,B. and Waldor,M.K. (2002) Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from *Vibrio cholerae*. *J. Bacteriol.*, **184**, 4259–4269.
41. Sentchilo,V., Czechowska,K., Pradervand,N., Minoia,M., Miyazaki,R. and van der Meer,J.R. (2009) Intracellular excision and reintegration dynamics of the ICEclc genomic island of *Pseudomonas knackmussii* sp. strain B13. *Mol. Microbiol.*, **72**, 1293–1306.
42. Waters,J.L., Wang,G.-R. and Salyers,A.A. (2013) Tetracycline-related transcriptional regulation of the CTnDOT mobilization region. *J. Bacteriol.*, **195**, 5431–5438.
43. Poulin-Laprade,D., Matteau,D., Jacques,P.-É., Rodrigue,S. and Burrus,V. (2015) Transfer activation of SXT/R391 integrative and conjugative elements: unraveling the SetCD regulon. *Nucleic Acids Res.*, **43**, 2045–2056.
44. Celli,J. and Trieu-Cuot,P. (2002) Circularization of Tn*916* is required for expression of the transposon-encoded transfer functions: characterization of long tetracycline-inducible transcripts reading through the attachment site. *Mol. Microbiol.*, **28**, 103–117.
45. Smyshlyaev,G. and Bateman,A.. Barabas (2021) Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. *Mol. Syst. Biol.*, **17**, e9880.
46. Lewis,J.A. and Hatfull,G.F. (2001) Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic Acids Res.*, **29**, 2205–2216.
47. Ramsay,J.P., Hynes,M.F., Sullivan,J.T. and Ronson,C.W. (2017) Symbiosis Islands. In: *Reference Module in Life Sciences*. Elsevier, pp. 598–600.
48. Johnson,C.M., 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.
49. Haskett,T.L., Terpolilli,J.J., Bekuma,A., O'Hara,G.W., Sullivan,J.T., Wang,P., Ronson,C.W. and Ramsay,J.P. (2016) Assembly and transfer of tripartite integrative and conjugative genetic elements. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 12268–12273.
50. D'Costa,V.M., McGrann,K.M., Hughes,D.W. and Wright,G.D. (2006) Sampling the antibiotic resistome. *Science*, **311**, 374–377.
51. LeGault,K.N., Hays,S.G., Angermeyer,A., McKitterick,A.C., Johura,F.-T., Sultana,M., Ahmed,T., Alam,M. and Seed,K.D. (2021) Temporal shifts in antibiotic resistance elements govern phage-pathogen conflicts. *Science*, **373**, eabg2166.
52. Bernheim,A. and Sorek,R. (2020) The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.*, **18**, 113–119.
53. Rocha,E.P.C. and Bikard,D. (2022) Microbial defenses against mobile genetic elements and viruses: who defends whom from what?*PLoS Biol.*, **20**, e3001514.
54. Hussain,F.A., Dubert,J., Elsherbini,J., Murphy,M., VanInsberghe,D., Arevalo,P., Kauffman,K.M., Rodiño-Janeiro,B.K., Gavin,H., Gomez,A. *et al.* (2021) Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*, **374**, 488–492.
55. Lu,F. and Churchward,G. (1995) Tn*916* target DNA sequences bind the C-terminal domain of integrase protein with different affinities that correlate with transposon insertion frequency. *J. Bacteriol.*, **177**, 1938–1946.
56. Mruk,I. and Kobayashi,I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.*, **42**, 70–86.
57. Lao,J., Lacroix,T., Guédon,G., Coluzzi,C., Payot,S., Leblond-Bourget,N. and Chiapello,H. (2022) ICEscreen: a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures. *NAR Genomics Bioinformatics*, **4**, lqac079.