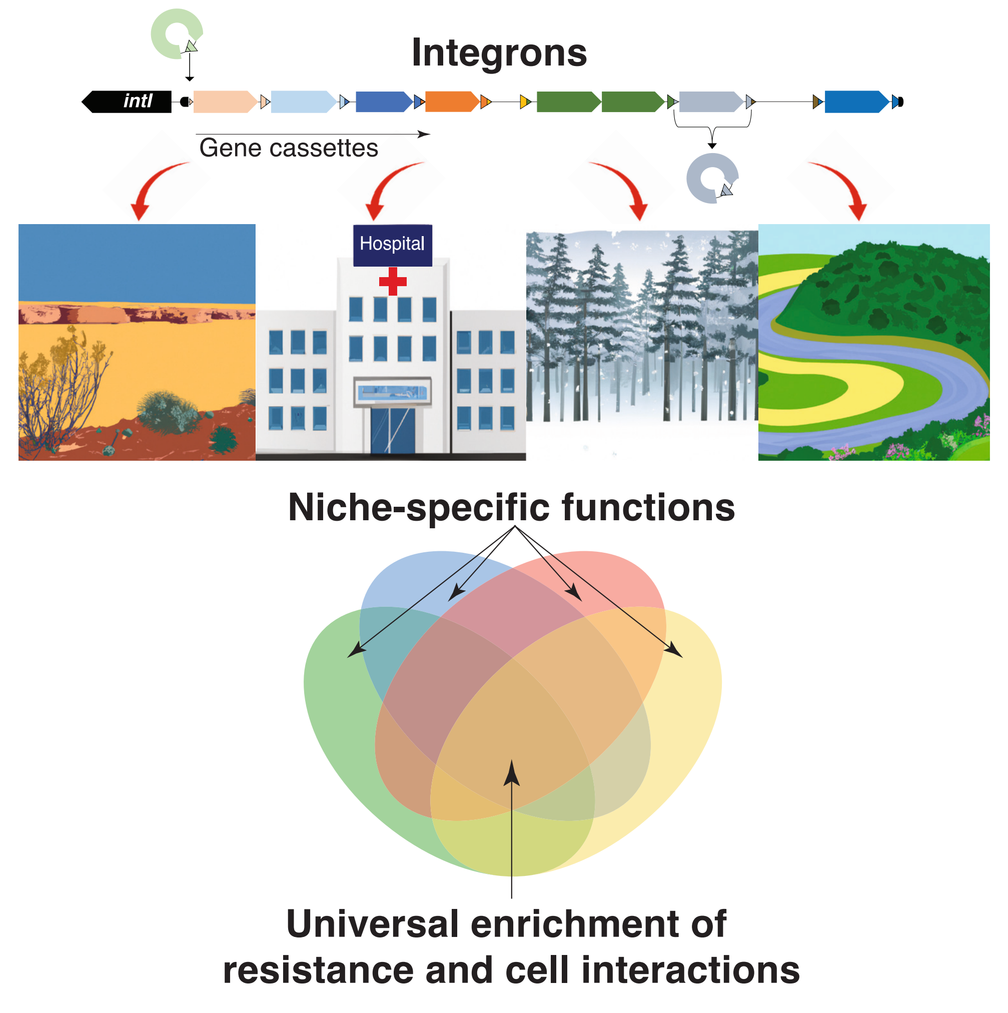
iScience

CellPress
OPEN ACCESS

Article

Functional enrichment of integrons: Facilitators of antimicrobial resistance and niche adaptation



Timothy M. Ghaly, Vaheesan Rajabal, Anahit Penesyan, Nicholas V. Coleman, Ian T. Paulsen, Michael R. Gillings, Sasha G. Tetu

timothy.ghaly@mq.edu.au (T.M.G.)
sasha.tetu@mq.edu.au (S.G.T.)

Highlights

Functional profiles of integron gene cassettes differ from their wider metagenomes

Diverse and novel antibiotic resistance, and anti-phage gene cassettes discovered

Both universal and niche-specific gene cassette functions identified

Universal functions are largely involved in biotic/abiotic interactions

Ghaly et al., iScience *26*, 108301
November 17, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.108301

Article

Functional enrichment of integrons: Facilitators of antimicrobial resistance and niche adaptation

Timothy M. Ghaly,[1,3,4,*] Vaheesan Rajabal,[1,2,3] Anahit Penesyan,[1,2] Nicholas V. Coleman,[1] Ian T. Paulsen,[1,2] Michael R. Gillings,[1,2] and Sasha G. Tetu[1,2,*]

SUMMARY

Integrons are genetic elements, found among diverse bacteria and archaea, that capture and rearrange gene cassettes to rapidly generate genetic diversity and drive adaptation. Despite their broad taxonomic and geographic prevalence, and their role in microbial adaptation, the functions of gene cassettes remain poorly characterized. Here, using a combination of bioinformatic and experimental analyses, we examined the functional diversity of gene cassettes from different environments. We find that cassettes encode diverse antimicrobial resistance (AMR) determinants, including those conferring resistance to antibiotics currently in the developmental pipeline. Further, we find a subset of cassette functions is universally enriched relative to their broader metagenomes. These are largely involved in (a)biotic interactions, including AMR, phage defense, virulence, biodegradation, and stress tolerance. The remainder of functions are sample-specific, suggesting that they confer localised functions relevant to their micro-environment. Together, they comprise functional profiles different from bulk metagenomes, representing niche-adaptive components of the prokaryotic pangenome.

INTRODUCTION

Integrons are genetic elements that act as gene capture and expression systems.[1,2] They are hotspots of genetic diversity and can facilitate the rapid adaptation of host cells. Integrons capture and express small mobile elements, known as gene cassettes, which generally consist of an open reading frame and a cassette recombination site, *attC*.[3] Capture of gene cassettes is mediated by the integron integrase, IntI, and involves recombination between the *attC* site of the inserting cassette and the endogenous integron recombination site, *attI*.[4–7] Multiple gene cassettes can be inserted at the *attI* site to form an integron cassette array, which, in some bacterial chromosomes, can consist of more than 200 consecutive cassettes.[1] A promoter, Pc, located within *intI* in many characterised integrons, facilitates expression of integrated gene cassettes at the start of the array.[8] These cassette arrays are dynamic, as IntI activity can lead to the excision of any gene cassette within the array. Excised cassettes can either be lost or re-inserted at the *attI* site, where expression might be maximised due to the proximity to the promoter. Integrons are mostly known for the capture and dissemination of antimicrobial resistance gene cassettes among diverse bacteria, however, they are extremely versatile systems that can generate genomic complexity and facilitate rapid adaptation "on demand."

Integrons were originally thought to only exist in Bacteria, but have recently been discovered in Archaea.[9] Consequently, integrons can facilitate the capture of archaeal gene cassettes by bacteria, driving cross-domain gene transfer between Archaea and Bacteria.[9] Thus, integron activity has important implications for the ecology and evolution of prokaryotes. They are also extremely prevalent elements that can be detected in every environment surveyed.[10] However, despite their broad taxonomic distribution, geographical prevalence, and evolutionary potential, the functions of gene cassettes remain poorly characterised.[3] We propose that such investigations should be considered high priority as they can help us understand the functional and adaptive potential of integron-carrying archaea and bacteria, and uncover traits that could be readily shared between these two domains.

To further investigate the functional diversity of gene cassettes, we employed cassette-targeted amplicon sequencing and whole metagenome sequencing from diverse environmental samples. Samples, which consisted of soil, sediment and biofilm material, were obtained from sites spanning terrestrial and aquatic environments to maximize the functional diversity of gene cassettes recovered in the present study. We identified functions that are enriched among gene cassettes relative to whole metagenomes from the same samples. We find that a subset of functions is universally enriched, largely comprising those that mediate biotic and abiotic interactions. The remainder of cassettes are generally sample-specific, often encoding rare, uncharacterized proteins, and generating distinct functional profiles relative to whole

[1] School of Natural Sciences, Macquarie University, New South Wales 2109, Australia
[2] ARC Centre of Excellence in Synthetic Biology, Macquarie University, New South Wales 2109, Australia
[3] These authors contributed equally
[4] Lead contact
*Correspondence: timothy.ghaly@mq.edu.au (T.M.G.), sasha.tetu@mq.edu.au (S.G.T.)
https://doi.org/10.1016/j.isci.2023.108301

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

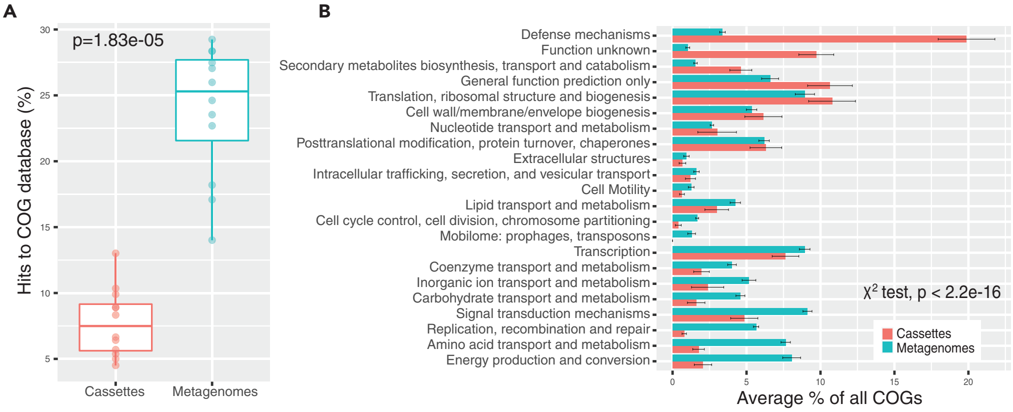


Figure 1. COG functional profiles of gene cassettes and metagenomes

(A) Percentage of proteins within a sample with hits to domain profiles within the COG database.

(B) Average percentage per sample (±1 SEM) of each COG functional category ranked according to the relative difference in abundance the gene cassette pool and the associated metagenome.

metagenomes. Gene cassettes are thus enriched in niche functions from the prokaryote pangenome, which we hypothesize to have specific relevance to their microenvironment.

RESULTS AND DISCUSSION

Here, we sequenced integron gene cassette amplicons and whole metagenomes from twelve environmental samples collected from a range of terrestrial and aquatic sites across Australia and Antarctica. The average assembled metagenome size from each sample was 48.9 megabases (Mb), ranging from 3.6Mb to 154.9Mb. Together, the assembled metagenomes encoded a total of 817,555 proteins (μ = 68,130; 4,867–217,132 per sample). From the same samples, amplicon sequencing of integron gene cassettes yielded a total of 45,399 cassette-encoded proteins (μ = 3,783; 1,193–6,989 per sample). See Tables S3 and S4 for a summary of all sequencing and assembly information.

Enriched functions of cassette-encoded proteins

Cassette-encoded proteins exhibited distinct functional profiles relative to their broader metagenomes. Functional predictions were made by protein homology searches against the Clusters of Orthologous Genes (COG)[11] and Pfam 35.0[12] databases. First, we note that gene cassettes, relative to the metagenomes, were significantly enriched with regard to proteins that have no matches to COG domain profiles (Wilcoxon test, p = 1.83e-05; Figure 1A). Similarly, there was a significantly lower proportion of cassettes with Pfam hits (μ = 24%) relative to the metagenomes (μ = 40%; T-test, p = 4.2e-06). This supports the idea that gene cassettes are an untapped resource of genetic novelty that is not captured by standard genomic and metagenomic sequencing approaches, which provide diminishing returns for protein discovery.[13]

We found that gene cassettes with domain hits exhibit distinct profiles in terms of COG functional categories relative to the metagenomes (χ^2 test, p < 2.2e-16; Figure 1B). In particular, of the categories with known biological functions, cassettes were significantly enriched in the categories "Defense mechanism" and "Secondary metabolite biosynthesis, transport and catabolism," as determined by Pearson's residuals (Figure 2A). Further, we examined the 30 most overrepresented Pfams with known biological functions, also determined by Pearson's residuals (Figure S1). These comprised detoxification and stress tolerance proteins, including acetyltransferases, MAPEG-family proteins, glutathione-dependent formaldehyde-activating enzymes, vicinal oxygen chelate (VOC) enzymes and a freeze-responsive protein; there were also diverse toxin-antitoxin system proteins, polyketide antibiotic synthesis proteins, different membrane-associated proteins, a peptidase, and a DNA-binding protein. These dominant Pfams were generally found across all samples, regardless of environment type (Figure S1), suggesting that they could provide significant advantages across broad environmental contexts.

Anti-phage and defense functions

The "Defense mechanism" COGs largely included toxin-antitoxin systems; anti-phage functions (comprising restriction endonucleases and CRISPR-associated proteins); and functions related to defense against antimicrobials and reactive metabolites (Figure 2B). Cassettes encoding the CRISPR-associated protein Csa3 (COG3415) were universally enriched across all samples (Figure 2B). Csa3 is a transcriptional regulator involved in type I CRISPR adaptation and type III CRISPR RNA interference, indicating the potential for the functional coupling of integrons and CRISPR-Cas systems. Additionally, restriction-modification (RM) systems also appear to be commonly enriched among the gene cassette pool (Figure 2B). Such systems, including the *Xba*I RM system, have been previously observed within integron gene cassettes.[14] These findings add to the growing evidence of integrons mediating anti-phage defense, highlighting potential concerns that integron gene cassettes, detrimental to the success of phage therapy, will be selected in clinical settings, similar to what has previously occurred with antibiotic

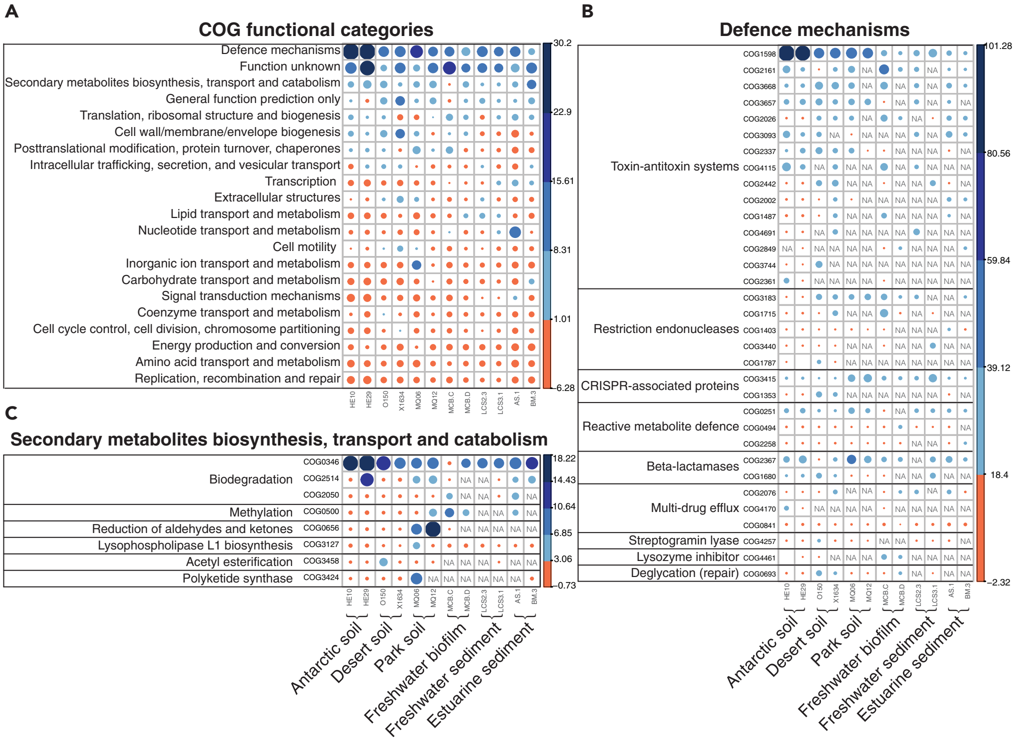


Figure 2. Enriched COG functions among gene cassettes relative to metagenomes

(A) Enrichment of COG categories among cassette-encoded proteins within each sample relative to the broader metagenome, as determined by Pearson's residuals. Colored scale bar indicates range of Pearson's residuals, where positive values (blue) are overrepresented among cassettes, and negative values (red) are underrepresented. Size of circles indicates the degree to which each category is over- or under-represented among cassette-encoded proteins.

(B) Enrichment of individual COGs within the "Defense mechanism" category.

(C) Enrichment of individual COGs within the "Secondary metabolite biosynthesis, transport and catabolism" category.

therapy.[3] Together, the abundance and prevalence of defence-related functions encoded by integrons likely provide a means for host cells to rapidly acquire and change the genetic cargo of cassette arrays to defend against a variety of biotic and abiotic threats.

The ubiquitous and diverse vicinal oxygen chelate family cassettes include novel resistance genes

The enrichment of the category "Secondary metabolite biosynthesis, transport and catabolism" was largely driven by a single COG (COG0346), belonging to the vicinal oxygen chelate (VOC) family of enzymes (Figure 2C). VOC enzymes catalyze a wide range of chemical reactions, including the degradation of aromatic hydrocarbons (catechol 2,3-dioxygenase), highly toxic methylglyoxal (glyoxalase I), antimicrobial compounds, fosfomycin and bleomycin, and the phytotoxin, toxoflavin.[15] We found that cassette-encoded VOC enzymes were present in every sample, and were significantly enriched in cassette pools relative to the wider metagenomes in all environment types (Figure 2C).

Phylogenetic analysis of these VOC enzymes suggest that cassettes collectively encode many of the characterised functions among this diverse superfamily (Figure 3A). To experimentally validate the activity of some of these cassette-encoded proteins, we synthesised and expressed VOC family gene cassettes in *E. coli*. We found that two different VOC cassettes could confer increased resistance to fosfomycin (2- to 4-fold MIC increase) and bleomycin (2-fold MIC increase), with one of these cassettes providing simultaneous resistance to both antibiotics (Figures 3B and 3C). This represents the first instance of a characterised integron gene cassette conferring resistance to both of these compounds. It is important to note that these cassettes were expressed from a multi-copy vector (~40 copies/cell), and, therefore, their observed AMR activity may be context dependent. However, since we do not know the genomic context of these cassettes, it is conceivable that they might indeed exist at high-copy numbers, given AMR cassettes are routinely detected on multi-copy plasmids in clinical contexts. This suggests that, at least in clinical settings, these cassettes are likely to confer antibiotic resistance. Interestingly, bleomycin resistance VOC proteins are functionally distinct from other members of the superfamily, having lost the metal-binding sites characteristic of VOC enzymes in place of a hydrophobic cavity for bleomycin adhesion.[16] Thus, a VOC gene cassette that confers resistance to both bleomycin and fosfomycin (Figures 3B and 3C) reveals the capacity for substrate flexibility among this superfamily. This finding, together with their broad phylogenetic diversity, suggest that VOC cassettes can collectively degrade a wide range of compounds, which would explain their prevalence and enrichment across diverse environments.

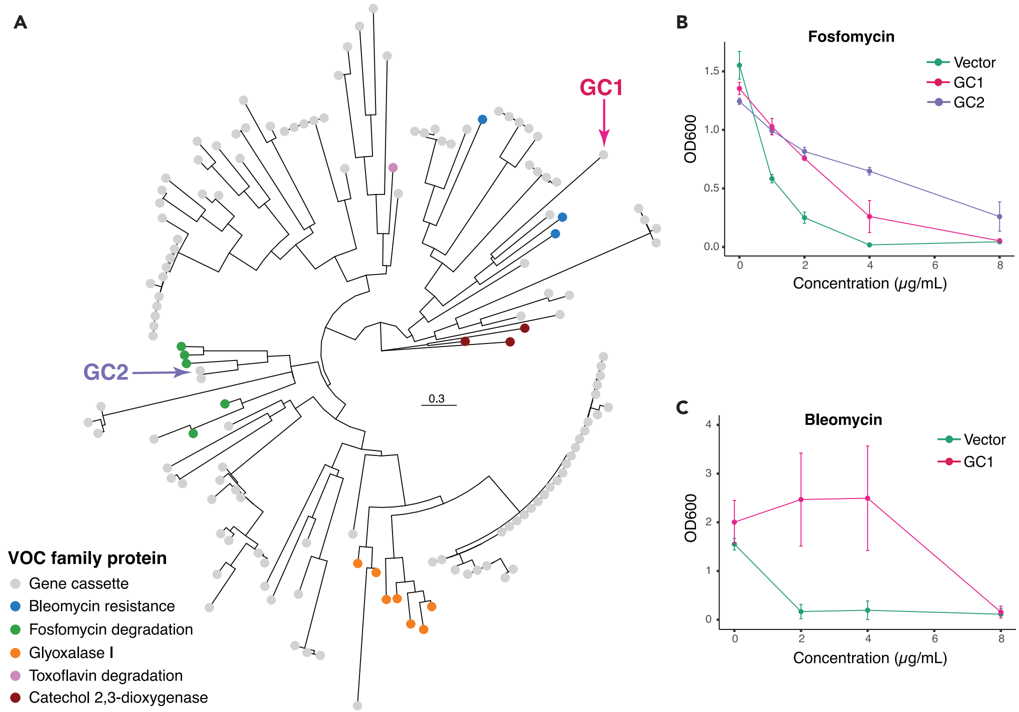


Figure 3. Functional diversity of vicinal oxygen chelate (VOC) family gene cassettes

(A) Maximum-likelihood phylogeny of VOC cassette-encoded proteins. Included in the phylogeny are experimentally characterised VOC members involved in bleomycin resistance (WP_001242578.1, AAA73391.1, AKJ21157.1), fosfomycin degradation (WP_014714131.1, WP_215810875.1, B9IY29.1, A6QJH4.1, AYW42209.1), glyoxalase I activity (NP_310387.1, ZP_01887743.1, CAA74673.1, AAG06912.1, AAG04099.1, AAG08496.1, AAN69360.1), catechol 2,3-dioxygenases (AAR90133.1, AAD02148.1, AAC79918.1), and taxoflavin degradation (ANS71543.1). The functions of gene cassettes GC1 and GC2 (black arrows) have been experimentally confirmed via fosfomycin, B, and bleomycin, C, resistance assays.

(B and C) Growth of *E. coli* BL21(DE3) in the presence of fosfomycin and bleomycin, respectively, while expressing gene cassettes GC1 (pink) and GC2 (purple), as well as with the empty expression vector pET29b (green). Results indicate average OD600 measurements (±1 SEM) for n = 4 (fosfomycin), or n = 3 (bleomycin) replicates.

Cassette functional profiles are niche-specific and distinct from their broader metagenomes

Cassettes are modular, interchangeable units that can be rapidly incorporated, rearranged or discarded. Their insertion and excision are driven by the SOS response, a global gene regulatory system that can be triggered by DNA-damaging stress.[17,18] For this reason, they might encode diverse functions that are conditionally beneficial to their host cell depending on prevailing conditions. One exception to this general rule is toxin-antitoxin (TA) genes, which are selfish, parasitic entities. This explains the abundance of TA genes among the gene cassette pool, as they cannot be lost from an integron without killing the host cell. However, under specific circumstances, these TA cassettes may also be advantageous to their host cell. In addition to stabilising large cassette arrays, the antitoxin product of TA cassettes can inactivate the toxin from homologous TA systems of phages and plasmids, thus providing defense from newly invading mobile elements.[19,20] Beyond TA cassettes, however, the benefit, and thus, the stability of any gene cassette will be highly dependent on environmental conditions, which will vary over time and space. Indeed, cassette content within soils rapidly change at centimetre scales,[10] suggesting that cassettes have functions directly relevant to their particular microenvironments.

Indeed, we found that beyond the subset of cassette functions that were universal, the remaining cassettes were largely sample-specific (Figure 4). When examining significant correlations among Pfams, cassette-encoded functions clustered according to sample, and samples further clustered into environment type (Figure 4). This suggests that while there are cassettes whose functions are broadly relevant in all environments (Figures 1 and 2), a majority of cassettes encode functions that are sample- and environment-specific. This strongly suggest that most functions of cassette-encoded genes are context-specific. Consequently, cassette populations and functions turn over at small spatial scales, reflecting the microenvironments occupied by their host cells.

Further, we found that the Pfam profiles of cassettes were distinct from those of the broader metagenome (Figure S2). Thus, cassettes generally encode functions that are distinct from the overall pool of functions present in metagenomes, and enrich specific, niche functions from the prokaryotic pangenome. It further suggests that the examination of cassette-encoded proteins will recover functions unlikely to be found via the examination of metagenomes.

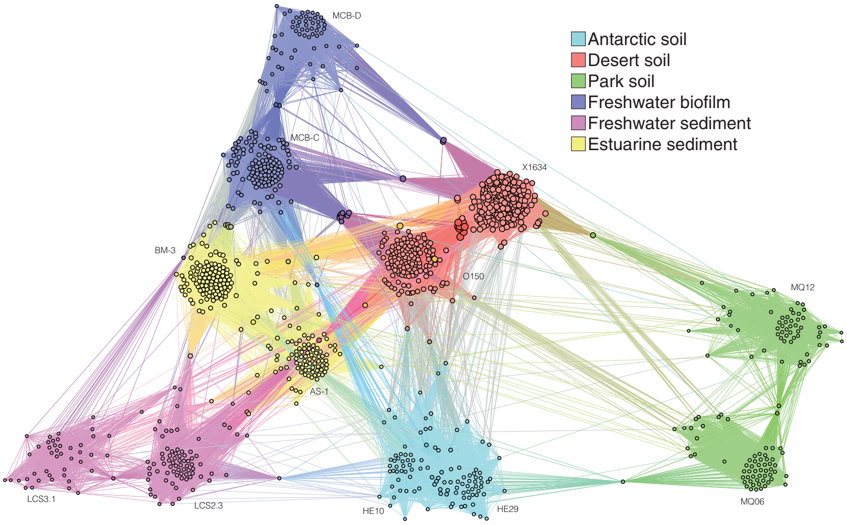


Figure 4. Correlation network of cassette-encoded Pfam functions

The network depicts significant correlations ($p < 0.05$) between Pfam functions (nodes) encoded by gene cassettes. The network is comprised of 1,158 protein families, comprising ~6% of all families within the Pfam r35.0 database. Labels indicate individual samples. Colors indicate the environments from which the samples were collected. Pfams linked to more than one environment are shown in overlapping colors. The size of the node is relative to the node authority (based on degree of correlation). Edges (connecting lines) represent correlations between Pfams. Edges are colored based on the overlapping color of the two nodes that they connect.

Facilitators of cell interactions: Cassette-encoded transmembrane and secretory proteins

Cassette-encoded proteins were significantly enriched with signal peptides (16.1%) relative to the broader metagenome (6.8%; Wilcoxon test, p = 1.83e-05; Figure 5A). Signal peptides are tags that direct proteins to be transported into or across the cell membrane. Such proteins mediate interactions with other cells and with their surrounding environment. Indeed, functional predictions of these cassette-encoded transmembrane and secretory proteins suggest that they facilitate local biotic and abiotic interactions. We found that the cassette proteins with signal peptides could be grouped into three main types of interactions. These were "Microbe-microbe," "Microbe-environment," and "Microbe-host" interactions (Figure 5B).

The dominant cassette proteins facilitating "microbe-microbe" interactions were involved with the inactivation and production of antimicrobial compounds. Such functions are known to be key regulators of the structure and function of microbiomes, suggesting that integrons could influence community dynamics. Proteins associated with "microbe-environment" interactions were mainly involved with detoxification, stress tolerance, environment-mediated cell signaling, and energy transduction. These cassettes likely help their host cell to adapt and thrive under changing environmental conditions. Potential virulence gene cassettes involved in cell invasion and immune evasion comprised the "microbe-host" interaction category. These latter cassettes highlight the ongoing potential for integrons to mediate negative impacts on plant and animal health.[21] The dynamic nature of these different types of interaction makes integrons a key resource for the microbes involved, by providing a "plug-and-play" system that responds to environmental change.

Cassette-encoded antimicrobial resistance

A diverse set of putative antimicrobial resistance (AMR) determinants were encoded by gene cassettes. Although these constituted a little less than 1% of cassette-encoded proteins, this still represents almost an order of magnitude increase in relative abundance in the cassette pool compared to the broader metagenomes (Figure 6A). Gene cassettes were predicted to encode resistance to a diverse range of antibiotic classes (Figures 6A and 6B). Within each sample, cassettes encoded putative resistance to between two to five different antibiotic classes (Figure 6A). The most abundant AMR cassettes were those predicted to confer resistance to β-lactam antibiotics, followed by aminoglycosides (Figure 6B). Despite the observed diversity of putative AMR cassettes, it is evident that their full diversity was not captured, since the AMR gene accumulation curve shows no sign of saturation (Figure 6C).

Most predicted AMR cassettes appear to be isolated with no other AMR cassette on the same contig. Although, as a caveat, it must be considered that due to the use of primers targeting locations within *attC* sites, a single integron might be assembled into multiple contigs due to the amplification of different parts of the cassette array that do not overlap. However, we note that AMR gene cassettes that do co-occur on the same integron are often within the same class of resistance family (Figure S3). This shows that integrons can accumulate different genes that confer resistance to similar sets of antimicrobial compounds, potentially providing higher levels of resistance to a given class of antibiotic.

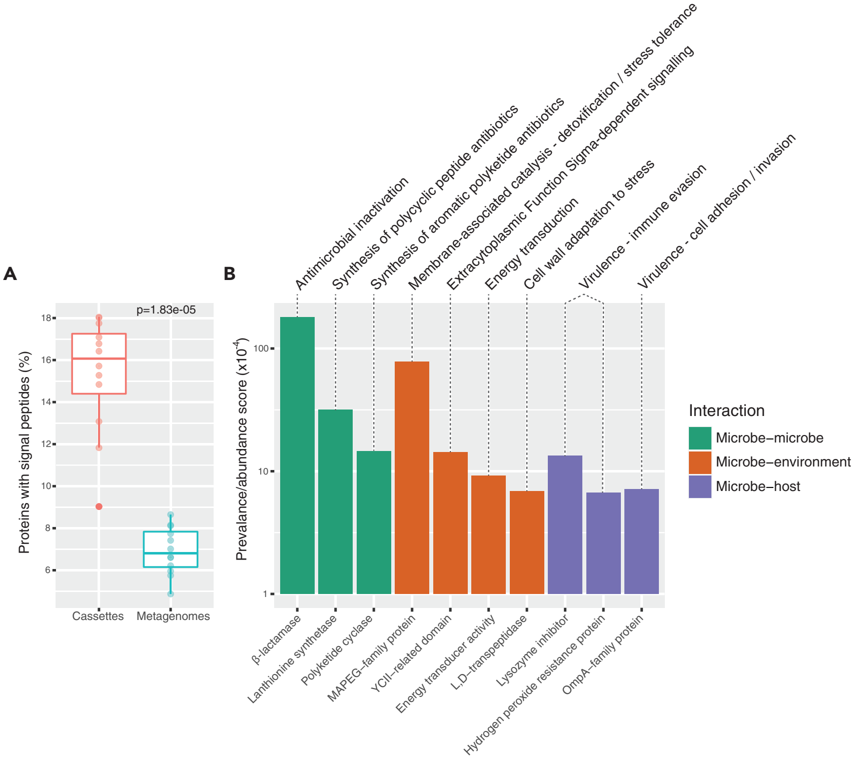


Figure 5. Functions of cassette-encoded transmembrane and secretory proteins

(A) Percentage of proteins with prokaryotic signal peptides, which target proteins into, or across, cell membranes.

(B) Dominant functions of cassette proteins with signal peptides. The y axis ($\log_{10}$ scale) denotes a prevalence/abundance score calculated as the product of the average percentage of a functional annotation across all samples and the number of samples it occurred in. This metric was used to consider both abundance and prevalence of each functional annotation. The x axis denotes each functional annotations, while their broader functional roles are annotated above each bar. Bars are colored based on the interaction type as per the legend panel.

The prevalence of integron-mediated resistance around the globe[22–24] is an example of the power of selection imposed by human activities. Although we found that AMR genes make up only ~1% of environmental gene cassettes, they constitute more than one-third of cassettes in clinical integrons.[25] This is particularly concerning for AMR cassettes that are yet to make their way into clinical settings. We found that 108 (36.4%) putative AMR cassettes, which encode 41 different resistance determinants, were not present in the global integron database, INTEGRALL[26] (Table S5). In particular, we detected gene cassettes that potentially confer resistance to elfamycin and pleuromutilin. Resistance to these two classes of antibiotics have, to the best of our knowledge, never been associated with integrons. Due to the rapid rise of resistance to approved antibiotics and the slowing of novel antibiotic discovery, both elfamycin and pleuromutilin antibiotics are now being developed for human therapeutic use.[27,28] The pre-existence of cassettes that putatively confer resistance to these alternate antibiotics, coupled with the capacity of integrons to rapidly spread and increase the abundance of resistance upon selection raises serious concerns for their long-term success as therapeutic agents.

Conclusion

Integrons provide a source of on-demand genetic diversity for their host cell.[29] Investigations into the functional diversity of integron-borne genes has important implications for our understanding of the ecology and evolution of Archaea and Bacteria. Here, we examined the functional diversity of integron gene cassettes with a particular focus on those functions enriched relative to the broader metagenomes. The PCR-based approach used here offers a significant advantage in analysing environmental integrons due to the capacity to sequence a significantly larger proportion of the gene cassette pool compared to screening bulk metagenomic data. For example, bioinformatic screening of metagenomes yields an average of only 0.003 unique cassette ORFs per megabase of assembled data, compared to our approach, with an average rate of 22 cassette ORFs per megabase of assembled data.

The subset of dominant integron-encoded functions that appear to be universal in their distribution are largely involved with biotic and abiotic interactions, including antibiotic biosynthesis and inactivation, detoxification, stress tolerance, biodegradation, phage defense, environment-mediated cell signaling, energy transduction and virulence. Integrons can thus provide traits that facilitate host cells to survive, adapt, and interact with their surrounding environment, and provide the means to exchange these traits to best suit prevailing conditions.

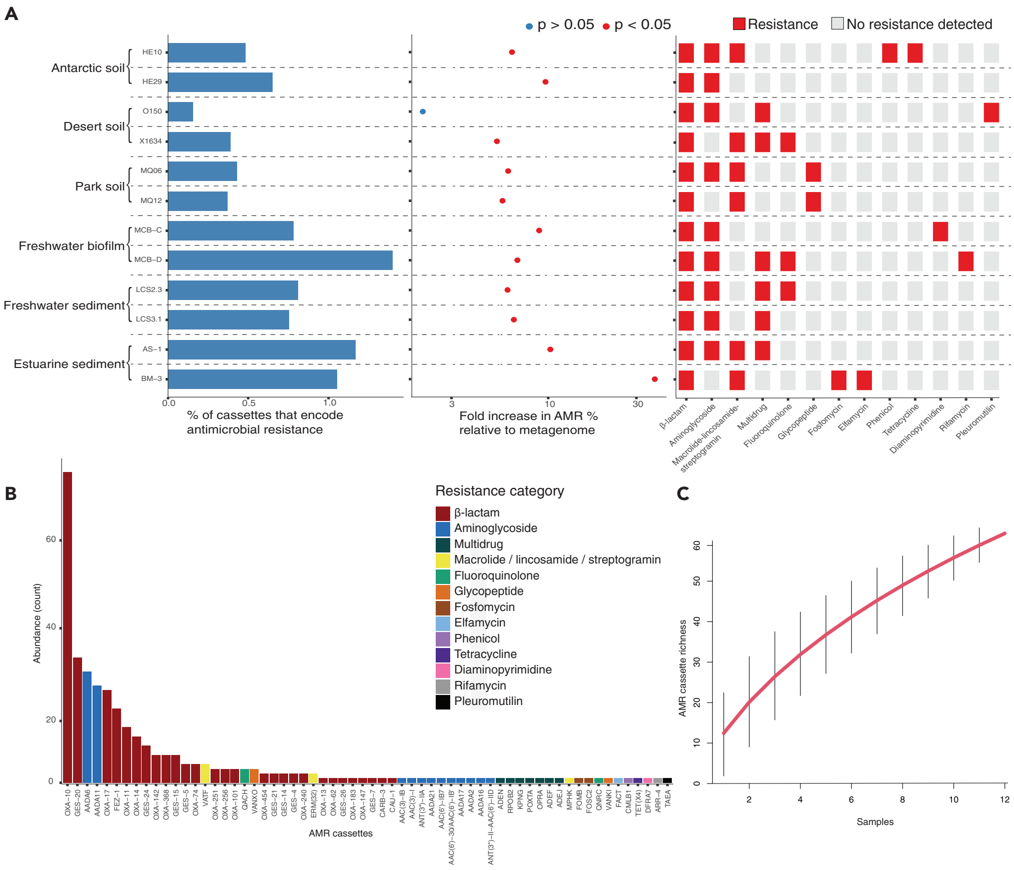


Figure 6. Abundance and diversity of antimicrobial resistance (AMR) gene cassettes

(A) From left to right: the percentage of gene cassettes that encode AMR determinants; the fold-increase in AMR % relative to the broader metagenomes, and whether this increase is significant (red) or not (blue), as determined by the χ^2 test for proportions; and presence/absence of resistance to different antibiotic classes detected within each sample.

(B) Abundance (count) across all samples of each AMR cassette. Bars are colored according to the class of antibiotic to which they confer resistance.

(C) AMR cassette accumulation curve along with 95% confidence intervals (denoted by the black vertical lines). Note, that the curve shows no sign of saturation, indicating that the full diversity of AMR cassettes was not sampled.

Importantly, integron gene cassettes and the traits that they encode can be shared between integron-carrying archaea and bacteria, highlighting integrons as important genetic elements for the ecology and evolution of prokaryotes.

While the dominant functions are widely distributed, most functions contained within the gene cassette pool were distinct to each sample, indicating that specific microenvironments likely select for bespoke sets of cassettes encoding relevant adaptive functions. Notably, the total functional reservoir of gene cassettes differs from the broader metagenome of each environment, likely representing important, niche-adaptive components of the prokaryotic pangenome.

The genetic novelty and diversity of gene cassettes suggest that they might also provide a fruitful avenue for protein discovery. In general, most protein functions remain unknown, encoded by small, rare protein families.[30] Their characterisation remains a key goal for various fields, including environmental sciences, medicine, and synthetic biology.[31] Protein discovery improves our understanding of microbial community function and microbial interactions, adding to our catalog of known biochemical functions that can be drawn upon for therapeutic or synthetic biology applications. New proteins that lack sequence or structural homology to known protein families are key candidates for activity-based approaches that might help speed up the rate at which new functions are discovered. The genetic cargo of integrons, which are significantly enriched in genetic novelty, and likely represent single-gene/single-trait entities, are ideal for such functional screening.[3,21]

Limitations of the study

Limitations of this study are the variation in the assembled size and eukaryotic content of the metagenomes. Such variation has the potential to confound functional enrichment patterns. Some of the metagenomic assemblies were relatively small (Table S1), which might limit the functional diversity captured within a sample. Similarly, given that integron gene cassettes are specific to prokaryotes, an abundance of eukaryotic ORFs in the metagenomes might also have an influence on observed enrichment patterns. Taxonomic analysis of all metagenomic ORFs revealed a small, albeit variable, proportion of eukaryotic ORFs among the metagenomes (Table S6). Although, in total, these made up a minor proportion, with only 0.89% ORFs across all samples (Median: 1.12%; Range: 0%–10% per sample) derived from eukaryotic microorganisms. Given that the dominant cassette-encoded functions are universally enriched across all environment types, regardless of the size or eukaryotic content of the metagenomes, it is unlikely that these limitations are having a genuine effect on the observed enrichment patterns.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Sample collection, DNA extraction and PCR amplification
  - Nanopore sequencing of cassette amplicons and whole metagenomes
  - Sequence processing and quality control
  - Functional analyses
  - Phylogenetic analysis of vicinal oxygen chelate (VOC) gene cassettes
  - Functional assays of VOC family proteins
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108301.

ACKNOWLEDGMENTS

This research was supported by the Australian Research Council Discovery Grant DP200101874. TMG would like to thank Mary, Saoirse, and Maebh Ghaly for their loving support.

AUTHOR CONTRIBUTIONS

TMG was involved in the study design, experimental procedures, data analysis, and wrote the original draft of the article. VR was involved in the study design and experimental procedures. AP was involved in the study design and data analysis. NVC was involved in the study design. ITP, MRG and SGT were involved in the study design and funding acquisition. All authors contributed to the final editing of the article.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: July 18, 2023
Revised: October 11, 2023
Accepted: October 19, 2023
Published: October 24, 2023

REFERENCES

1. Mazel, D. (2006). Integrons: agents of bacterial evolution. Nat. Rev. Microbiol. *4*, 608–620. https://doi.org/10.1038/nrmicro1462.
2. Gillings, M.R. (2014). Integrons: past, present, and future. Microbiol. Mol. Biol. Rev. *78*, 257–277.
3. Ghaly, T.M., Gillings, M.R., Penesyan, A., Qi, Q., Rajabal, V., and Tetu, S.G. (2021). The natural history of integrons. Microorganisms *9*, 2212.
4. Bouvier, M., Demarre, G., and Mazel, D. (2005). Integron cassette insertion: a recombination process involving a folded single strand substrate. EMBO J. *24*, 4356–4367. https://doi.org/10.1038/sj.emboj.7600898.
5. Bouvier, M., Ducos-Galand, M., Loot, C., Bikard, D., and Mazel, D. (2009). Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. PLoS Genet. *5*, e1000632.
6. Nivina, A., Escudero, J.A., Vit, C., Mazel, D., and Loot, C. (2016). Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of *attC* recombination sites. Nucleic Acids Res. *44*, 7792–7803. https://doi.org/10.1093/nar/gkw646.
7. Mukhortava, A., Pöge, M., Grieb, M.S., Nivina, A., Loot, C., Mazel, D., and Schlierf, M. (2019). Structural heterogeneity of *attC* integron recombination sites revealed by optical tweezers. Nucleic Acids Res. *47*, 1861–1870. https://doi.org/10.1093/nar/gky1258.
8. Hall, R.M., and Collis, C.M. (1995). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol. Microbiol. *15*, 593–600.
9. Ghaly, T.M., Tetu, S.G., Penesyan, A., Qi, Q., Rajabal, V., and Gillings, M.R. (2022). Discovery of integrons in Archaea: Platforms for cross-domain gene transfer. Sci. Adv. *8*,

eabq6376. https://doi.org/10.1126/sciadv.abq6376.

10. Ghaly, T.M., Geoghegan, J.L., Alroy, J., and Gillings, M.R. (2019). High diversity and rapid spatial turnover of integron gene cassettes in soil. Environ. Microbiol. *21*, 1567–1574.
11. Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. *28*, 33–36.
12. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432.
13. Chubb, D., Jefferys, B.R., Sternberg, M.J.E., and Kelley, L.A. (2010). Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. Bioinformatics *26*, 2664–2671.
14. Rowe-Magnus, D.A., Guerout, A.M., Ploncard, P., Dychinco, B., Davies, J., and Mazel, D. (2001). The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. Proc. Natl. Acad. Sci. USA *98*, 652–657. https://doi.org/10.1073/pnas.98.2.652.
15. Choi, J.-E., Nguyen, C.M., Lee, B., Park, J.H., Oh, J.Y., Choi, J.S., Kim, J.-C., and Song, J.K. (2018). Isolation and characterization of a novel metagenomic enzyme capable of degrading bacterial phytotoxin toxoflavin. PLoS One *13*, e0183893.
16. Armstrong, R.N. (2000). Mechanistic diversity in a metalloenzyme superfamily. Biochemistry *39*, 13625–13632. https://doi.org/10.1021/bi001814v.
17. Guerin, É., Cambray, G., Sanchez-Alberola, N., Campoy, S., Erill, I., Da Re, S., Gonzalez-Zorn, B., Barbé, J., Ploy, M.-C., and Mazel, D. (2009). The SOS response controls integron recombination. Science *324*, 1034.
18. Cambray, G., Sanchez-Alberola, N., Campoy, S., Guerin, É., Da Re, S., González-Zorn, B., Ploy, M.-C., Barbé, J., Mazel, D., and Erill, I. (2011). Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. Mobile DNA *2*, 6. https://doi.org/10.1186/1759-8753-2-6.
19. Wilbaux, M., Mine, N., Guérout, A.M., Mazel, D., and Van Melderen, L. (2007). Functional interactions between coexisting toxin-antitoxin systems of the *ccd* family in *Escherichia coli* O157: H7. J. Bacteriol. *189*, 2712–2719.
20. Guérout, A.M., Iqbal, N., Mine, N., Ducos-Galand, M., Van Melderen, L., and Mazel, D. (2013). Characterization of the *phd-doc* and *ccd* toxin-antitoxin cassettes from *Vibrio* superintegrons. J. Bacteriol. *195*, 2270–2283.
21. Ghaly, T.M., Geoghegan, J.L., Tetu, S.G., and Gillings, M.R. (2020). The peril and promise of integrons: beyond antibiotic resistance. Trends Microbiol. *28*, 455–464. https://doi.org/10.1016/j.tim.2019.12.002.
22. Partridge, S.R., Tsafnat, G., Coiera, E., and Iredell, J.R. (2009). Gene cassettes and cassette arrays in mobile resistance integrons. FEMS Microbiol. Rev. *33*, 757–784. https://doi.org/10.1111/j.1574-6976.2009.00175.x.
23. Zhu, Y.-G., Zhao, Y., Li, B., Huang, C.-L., Zhang, S.-Y., Yu, S., Chen, Y.-S., Zhang, T., Gillings, M.R., and Su, J.-Q. (2017). Continental-scale pollution of estuaries with antibiotic resistance genes. Nat. Microbiol. *2*, 16270. https://doi.org/10.1038/nmicrobiol.2016.270.
24. Ghaly, T.M., Paulsen, I.T., Sajjad, A., Tetu, S.G., and Gillings, M.R. (2020). A novel family of *Acinetobacter* mega-plasmids are disseminating multi-drug resistance across the globe while acquiring location-specific accessory genes. Front. Microbiol. *11*, 605952. https://doi.org/10.3389/fmicb.2020.605952.
25. Ghaly, T.M., Penesyan, A., Pritchard, A., Qi, Q., Rajabal, V., Tetu, S.G., and Gillings, M.R. (2022). Methods for the targeted sequencing and analysis of integrons and their gene cassettes from complex microbial communities. Microb. Genom. *8*, 000788. https://doi.org/10.1099/mgen.0.000788.
26. Moura, A., Soares, M., Pereira, C., Leitão, N., Henriques, I., and Correia, A. (2009). INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. Bioinformatics *25*, 1096–1098.
27. Prezioso, S.M., Brown, N.E., and Goldberg, J.B. (2017). Elfamycins: inhibitors of elongation factor-Tu. Mol. Microbiol. *106*, 22–34. https://doi.org/10.1111/mmi.13750.
28. Chahine, E.B., and Sucher, A.J. (2020). Lefamulin: the first systemic pleuromutilin antibiotic. Ann. Pharmacother. *54*, 1203–1214. https://doi.org/10.1177/1060028020932521.
29. Escudero, J.A., Loot, C., Nivina, A., and Mazel, D. (2015). The integron: adaptation on demand. Microbiol. Spectr. *3*. https://doi.org/10.1128/microbiolspec.MDNA3-0019-2014.
30. Harrington, E.D., Singh, A.H., Doerks, T., Letunic, I., von Mering, C., Jensen, L.J., Raes, J., and Bork, P. (2007). Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc. Natl. Acad. Sci. USA *104*, 13913–13918. https://doi.org/10.1073/pnas.0702636104.
31. Ufarté, L., Potocki-Veronese, G., and Laville, É. (2015). Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. Front. Microbiol. *6*, 563. https://doi.org/10.3389/fmicb.2015.00563.
32. Stokes, H.W., Holmes, A.J., Nield, B.S., Holley, M.P., Nevalainen, K.M., Mabbutt, B.C., and Gillings, M.R. (2001). Gene cassette PCR: sequence-independent recovery of entire genes from environmental DNA. Appl. Environ. Microbiol. *67*, 5240–5246.
33. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
34. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736. https://doi.org/10.1101/gr.215087.116.
35. Bushnell, B. (2014). BBTools Software Package.
36. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.
37. Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.
38. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol. *37*, 540–546.
39. 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. https://doi.org/10.1186/1471-2105-11-119.
40. Eddy, S.R. (2018). HMMER 3.2: Biosequence analysis using profile hidden Markov models. http://hmmer.org/.
41. Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. *37*, 420–423. https://doi.org/10.1038/s41587-019-0036-z.
42. de Nies, L., Lopes, S., Busi, S.B., Galata, V., Heintz-Buschart, A., Laczny, C.C., May, P., and Wilmes, P. (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. Microbiome *9*, 49.
43. Madden, T. (2013). The BLAST sequence analysis tool. In The NCBI Handbook [Internet], 2nd edition, D.B.J. Beck, J. Coleman, M. Hoeppner, M. Johnson, and D. Maglott, eds. (National Center for Biotechnology Information (US)).
44. Jari, O., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R., Simpson, G.L., and Solymos, P. (2019). vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project.org/package=vegan.
45. von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol. *20*, 217–314.
46. Harrell, F.E., and Dupont, C. (2021). Hmisc: Harrell Miscellaneous. R Package Version 4.5-0. https://CRAN.R-project.org/package=Hmisc.
47. Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.
48. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
49. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.
50. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.
51. Yeates, C., Gillings, M.R., Davison, A.D., Altavilla, N., and Veal, D.A. (1998). Methods for microbial DNA extraction from soil for PCR amplification. Biol. Proced. Online *1*, 40–47.
52. ONT Guppy v4.3.4: Local accelerated basecalling for Nanopore data. https://community.nanoporetech.com/downloads.
53. Kolmogorov, M., Bickhart, D.M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S.B., Kuhn, K., Yuan, J., Polevikov, E., Smith, T.P.L., and Pevzner, P.A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat. Methods *17*, 1103–1110.

54. ONT Medaka v1.4.3. https://github.com/nanoporetech/medaka.

55. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

56. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P., and Zhang, L. (2018). DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome *6*, 23.

57. 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.

58. Penesyan, A., Nagy, S.S., Kjelleberg, S., Gillings, M.R., and Paulsen, I.T. (2019). Rapid microevolution of biofilm cells in response to antibiotics. NPJ Biofilms Microbiomes *5*, 34. https://doi.org/10.1038/s41522-019-0108-3.

59. Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. Math. J. *10*, 37–71.

60. Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

61. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods *14*, 587–589.

62. Wiegand, I., Hilpert, K., and Hancock, R.E.W. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. Nat. Protoc. *3*, 163–175.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Escherichia coli BL21(DE3)	Pharmacia Biotech	*E. coli* BL21(DE3)
Chemicals, peptides, and recombinant proteins		
Fosfomycin disodium salt	Sigma-Aldrich	P5396; CAS: 26016-99-9
Bleomycin sulfate	Sigma-Aldrich	BP971; CAS: 9041-93-4
Kanamycin sulfate	Sigma-Aldrich	60615; CAS: 70560-51-9
IPTG, Isopropyl β-D-thiogalactoside	Sigma-Aldrich	I6758; CAS: 367-93-1
Ligation Sequencing Kit	Oxford Nanopore Technologies	Cat#SQK-LSK109
Native Barcoding Expansion Kits	Oxford Nanopore Technologies	Cat#EXP-NBD104, Cat#EXP-NBD114
Flow Cell Wash Kit	Oxford Nanopore Technologies	Cat#EXP-WSH004
Deposited data		
Raw sequence data of the cassette amplicons	This paper	SAMN21354384 to SAMN21354431
Raw sequence data of the whole metagenomes	This paper	SAMN27966069 to SAMN27966080
Oligonucleotides		
Primer HS286 GGGATCC TCSGCTKGARCGAMTTGTTAGVC	Stokes et al.,[32] 2001	N/A
Primer HS287 GGGATCC GCSGCTKANCTCVRRCGTTAGSC	Stokes et al.,[32] 2001	N/A
Recombinant DNA		
pET29b(+)	Novagen	Cat#69872
pET29b(+)-GC1	This paper	N/A
pET29b(+)-GC2	This paper	N/A
Software and algorithms		
Guppy v.4.3.4	Oxford Nanopore Technologies	https://nanoporetech.com/
NanoFilt v2.8	De Coster et al.[33]	https://github.com/wdecoster/nanofilt
Canu v2.0	Koren et al.[34]	https://github.com/marbl/canu
BBTools v35	Bushnell[35]	https://sourceforge.net/projects/bbmap/
Minimap2 v2.22-r1101	Li[36]	https://github.com/lh3/minimap2
Racon v1.4.20	Vaser et al.[37]	https://github.com/isovic/racon
attC-screening pipeline	Ghaly et al.[25]	https://github.com/timghaly/integron-filtering
Flye v2.9-b1768	Kolmogorov et al.[38]	https://github.com/fenderglass/Flye
ONT_polish pipeline	This paper	https://github.com/timghaly/ONT_polish
Prodigal v2.6.3	Hyatt et al.[39]	https://github.com/hyattpd/Prodigal
RPS-BLAST v2.12.0	NCBI	https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml
hmmer v3.2	Eddy[40]	http://hmmerorg/
SignalP v5.0	Almagro et al.[41]	https://services.healthtech.dtu.dk/services/SignalP-5.0/
eggNOG-mapper v2.0.1b	Cantalapiedra et al. 2021	https://github.com/eggnogdb/eggnog-mapper
PathoFact v1.0	De Nies et al.[42]	https://git-r3lab.uni.lu/laura.denies/PathoFact

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BLAST v2.71	Madden[43]	https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
vegan 2.5–6 R package	Jari et al.[44]	https://CRAN.R-project.org/package=vegan
CAT v5.2.3	Von Meijenfeldt et al.[45]	https://github.com/dutilh/CAT
Hmisc v4.5-0 R package	Harrell & Dupont[46]	https://CRAN.R-project.org/package=Hmisc
Gephi	Bastian et al.[47]	https://gephi.org/
MAFFT v7.508	Katoh & Standley[48]	https://mafft.cbrc.jp/alignment/software/
trimAl v1.4.rev15	Capella-Gutiérrez[49]	http://trimal.cgenomics.org/trimal
IQ-TREE v2.2.0.3	Minh et al.[50]	http://www.iqtree.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Timothy Ghaly (timothy.ghaly@mq.edu.au).

Materials availability

The plasmid vector, pET29b(+), used in this study is available from Twist Bioscience (USA). DNA sequences of cloned gene cassettes are available as Table S2.

Data and code availability

- Raw sequence data have been deposited in the NCBI SRA database and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. Cassette amplicons are available under NCBI BioProject PRJNA761546 (BioSample accessions: SAMN21354384 to SAMN21354431). Whole metagenomes are available from under NCBI BioProject PRJNA833246 (BioSample accessions: SAMN27966069 to SAMN27966080).
- The code used for correcting the Nanopore assemblies is available at https://github.com/timghaly/ONT_polish, and the code used for filtering sequences to ensure that they represent amplicons from genuine integrons is available at https://github.com/timghaly/integron-filtering. All bioinformatic software and input parameters are described under the ' method details' heading. All other code used for data analysis is available at https://github.com/timghaly/Gene_cassette_functional_enrichment.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The microbial strain, *Escherichia coli* BL21(DE3), was used in this study. After, electroporating the different vectors (see 'method details'), each strain was maintained in Muller-Hinton (MH) medium supplemented with kanamycin (50 μg/mL) at 37°C.

METHOD DETAILS

Sample collection, DNA extraction and PCR amplification

All samples were collected as previously described.[25] These consisted of duplicate samples collected from six different sites: three terrestrial and three aquatic. Terrestrial samples consisted of soils collected from Herring Island, Antarctica, Sturt National Park, New South Wales (NSW), Australia, and Macquarie University campus, NSW.[10] Aquatic samples were comprised of freshwater biofilms from Mars Creek, NSW, freshwater sediment from Lane Cove River, NSW, and estuarine sediment from Parramatta River, NSW.[25] Samples were collected across a wide spatiotemporal range, from centimetres to kilometres scales within sampling sites, spanning 22 years of collection (Table S1). DNA was extracted from 0.3g of sample material using standard bead-beating methods.[51]

Integron gene cassettes were amplified from extracted DNA using the PCR primers HS287 and HS286.[32] These primers anneal to the semi-conserved flanking regions of cassette *attC* sites in outward-facing orientations to amplify the intervening gene cassettes(s). Amplification was carried out using the long-range Phusion Hot Start II DNA Polymerase (ThermoFisher Scientific, Waltham, MA, USA). All PCRs conditions and reagent concentrations were as previously described.[25]

Nanopore sequencing of cassette amplicons and whole metagenomes

All cassette amplicons and whole metagenomes were sequenced using Oxford Nanopore Technologies (ONT). PCR products were multiplexed into a single sequencing library using the ONT Ligation Sequencing Kit (SQK-LSK109) and the ONT Native Barcoding Expansion Kits (EXP-NBD104 and EXP-NBD114). The DNA library was sequenced using a MinION MK 1B on an R10.3 flow cell for 24 h. Basecalling of the resulting nanopore signal data was carried out with Guppy v.4.3.4[52] using the high accuracy basecalling model.

Whole metagenomic DNA extracted from each of the twelve samples were multiplexed into three sequencing libraries using the ONT Ligation Sequencing Kit (SQK-LSK109) and the ONT Native Barcoding Expansion Kit (EXP-NBD104 and EXP-NBD114). Each DNA library was sequenced using a MinION MK 1B on an R9.4.1 flow cell for 48 h. After 24 h, each flow cell was washed using the ONT Flow Cell Wash Kit (EXP-WSH004) and the same DNA library was re-loaded and sequenced for a further 24 h. Basecalling was carried out with Guppy v.4.3.4[52] using the fast basecalling model.

Sequence processing and quality control

The sequence processing and quality control of sequenced cassette PCRs were as previously described.[25] Briefly, reads with an average quality score below 10 were removed using NanoFilt v2.8[33] [parameters: -q 10]. High-quality reads were assembled using Canu v2.0[34] [parameters: genomeSize = 5m minReadLength = 250 minOverlapLength = 200 corMinCoverage = 0 corOutCoverage = 20000 corMhapSensitivity = high maxInputCoverage = 20000 batMemory = 125 redMemory = 32 oeaMemory = 32 batThreads = 24 purgeOverlaps = aggressive]. Next, the tgStoreDump script within Canu was used to pool assembled contigs and unassembled reads together [parameters: -consensus -fasta]. The BBTools v35[35] dedupe script was used to remove any sequence redundancy, including reverse complement redundancies. The remaining sequences were corrected with four iterations of polishing with Minimap2 v2.22-r1101[36] [parameters: -x map-ont -t 24] and Racon v1.4.20 [parameters: -m 8 -x 6 -g −8 -w 500 -t 24].[37] To ensure sequences represented genuine gene cassettes, and not off-target PCR products, we applied our recently established cassette filtering pipeline, *attC*-screening (available: https://github.com/timghaly/integron-filtering).[25] This pipeline screens all sequences for gene cassette *attC* sites and discards those that lack a complete *attC* site. Thus, only high-confidence amplicons that contain multiple cassettes intervened by complete *attC* sites were retained for downstream analyses.

The metagenomic reads were assembled using Flye v2.9-b1768[38,53] [parameters: –nano-raw –genome-size 250m –meta -i 0 –threads 24]. The metagenome assemblies were then polished using the ONT_polish pipeline (https://github.com/timghaly/ONT_polish) [parameters: -m r941_min_fast_g303 -t 24], which applies four iterations of Minimap2[36] and Racon[37] correction, followed by Medaka[54] polishing.

Functional analyses

Open reading frames (ORFs) were predicted from the whole metagenome and cassette amplicon data using Prodigal v2.6.3[39] [parameters: -p meta -q]. We assigned putative functions to ORFs using both COG[11] and Pfam[12] databases. To assign COG functions, we employed a conserved domain search using RPS-BLAST v2.12.0 [command: rpsblast -query $i -db Cog -evalue 0.01 -outfmt 11 | rpsbproc -q -e 0.01 -m rep]. For Pfam annotations, we used hmmscan from the hmmer v3.2 package[40] against Pfam release 35.0. Only Pfam hits with an hmmscan e-value less than 0.001 were retained for downstream analysis.

To predict genes that encoded transmembrane or secreted proteins, we used SignalP v5.0[41] [parameters: -org arch|gram+|gram- -format short]. These SignalP parameters search for all prokaryotic signal peptide tag sequences, which target proteins into, or across, membranes. Additional functional analysis was carried out for proteins with detectable signal peptides using eggNOG-mapper v2.0.1b, based on eggNOG v5 orthology data, executed in DIAMOND[55] mode. We examined the most dominant of these functions based on a prevalence/abundance score above 6.5e-4, which we define here as the product of the average percentage of a functional annotation across all samples and the number of samples it occurred in. This metric was used to consider both abundance and prevalence of functions associated with cassette transmembrane and secretory proteins. All statistical analyses to determine functional enrichment were implemented using base R.

Genes encoding antimicrobial resistance (AMR) determinants were predicted using PathoFact v1.0[42] with default parameters. PathoFact is a pipeline that combines DeepARG[56] and RGI[57] to predict AMR genes. To determine the number of AMR cassettes found within the integron database, INTEGRALL[26] [downloaded: 2022-Apr-14], we performed a BlastP search using BLAST v2.71[43] with default parameters against the translated database. Hits with an amino acid identity >95% and a query cover >80%, were considered to be represented in the INTEGRALL database. The AMR cassette accumulation curve was generated using the *specaccum* function within the vegan 2.5–6 R package.[44] The Lomolino model was fitted to the exact accumulation using *fitspecaccum* function in vegan [parameters: model = "lomolino"].

Given that integron gene cassettes are specific to prokaryotes, we retrospectively assessed the potential confounding role that eukaryotic ORFs in the metagenomes might have on the observed functional enrichment patterns. To do this, we taxonomically classified all metagenomic ORFs using CAT v5.2.3[45] with default parameters. CAT first attempts to classify all metagenomic ORFs based on protein homology against the complete NCBI non-redundant (nr) database. It then classifies whole contigs using a voting algorithm based on the classifications of every ORF along a contig to provide robust taxonomic classifications. Consequently, it can classify ORFs with no homology matches by using the surrounding ORFs on the same contig. We assigned domain-level classifications to all ORFs based on their whole contig classifications.

The functional networks were generated as previously described.[9,58] Briefly, Pearson's correlations, based on correlations between Pfam functions, were calculated using the Hmisc v4.5-0 R package.[46] The networks were visualised from all correlations with p values less than 0.05 using the Yifan Hu force-directed layout algorithm[59] within the Gephi software.[47]

Phylogenetic analysis of vicinal oxygen chelate (VOC) gene cassettes

Gene cassettes annotated as vicinal oxygen chelate (VOC) family enzymes (COG0346) were selected for phylogenetic analysis. Their protein sequences, along with those of characterised VOC family proteins were aligned using MAFFT v7.508[48] [parameters: –localpair –maxiterate 100], and trimmed using trimAl v1.4.rev15[49] [parameters: -automated1]. Characterised VOC proteins consisted of bleomycin resistance proteins (WP_001242578.1, AAA73391.1, AKJ21157.1), fosfomycin degrading enzymes (WP_014714131.1, WP_215810875.1, B9IY29.1, A6QJH4.1, AYW42209.1), glyoxalase I (NP_310387.1, ZP_01887743.1, CAA74673.1, AAG06912.1, AAG04099.1, AAG08496.1, AAN69360.1), catechol 2,3-dioxygenase AAR90133.1, AAD02148.1, AAC79918.1), and a taxoflavin-degrading enzyme (ANS71543.1). A maximum-likelihood tree was generated from the alignment using IQ-TREE v2.2.0.3[50,60] with the best-suited protein model determined by ModelFinder[61] and 1,000 bootstrap replicates [parameters: -m MFP -bb 1000].

Functional assays of VOC family proteins

Eight gene cassettes predicted to encode VOC family enzymes were synthesised and cloned in-frame into *Nde*I and *Xho*I sites of the expression vector pET29b(+) (Twist Bioscience, USA). DNA sequences of the synthesised genes are available as Table S2. Constructs were transformed into electrocompetent *Escherichia coli* BL21(DE3) and maintained in Muller-Hinton (MH) medium supplemented with kanamycin (50 μg/mL) at 37°C. Minimum inhibitory concentrations (MIC) of fosfomycin sulfate (Sigma-Aldrich) and bleomycin disodium salt (Sigma-Aldrich) were determined via broth dilution method[62] in 96-well plates with a final volume of 200 μL. Overnight *E. coli* cultures carrying the constructs were subcultured at a 1:10 dilution, and gene insert expression was induced with 0.4 mM IPTG for 2 h at 37°C and 200 rpm. The cultures were then diluted to 5×10^5 CFU mL^{-1} and added to the microplates. MH media was supplemented with 0.4 mM ITPG and kanamycin (50 μg/mL) to maintain gene expression. Breathable films were used to seal the plates and incubated at 37°C with shaking for 24 h. OD600 measurements were taken every 15 min (PHERAstar FSX, BMG Labtech). All experiments were conducted with at least three biological replicates.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed using base R, unless otherwise specified. Specific statistical tests used can be found in the 'results and discussion' section. Significance was defined as $p < 0.05$.