MICROBIOLOGY

Discovery of integrons in Archaea: Platforms for cross-domain gene transfer

Timothy M. Ghaly[1]*, Sasha G. Tetu[1,2], Anahit Penesyan[1,2], Qin Qi[1], Vaheesan Rajabal[1,2], Michael R. Gillings[1,2]

Horizontal gene transfer between different domains of life is increasingly being recognized as an important evolutionary driver, with the potential to increase the pace of biochemical innovation and environmental adaptation. However, the mechanisms underlying the recruitment of exogenous genes from foreign domains are mostly unknown. Integrons are a family of genetic elements that facilitate this process within Bacteria. However, they have not been reported outside Bacteria, and thus their potential role in cross-domain gene transfer has not been investigated. Here, we discover that integrons are also present in 75 archaeal metagenome-assembled genomes from nine phyla, and are particularly enriched among Asgard archaea. Furthermore, we provide experimental evidence that integrons can facilitate the recruitment of archaeal genes by bacteria. Our findings establish a previously unknown mechanism of cross-domain gene transfer whereby bacteria can incorporate archaeal genes from their surrounding environment via integron activity. These findings have important implications for prokaryotic ecology and evolution.

INTRODUCTION

Horizontal gene transfer between different domains of life can be a major driver in species evolution (*1*). There are now numerous examples of genes that have been transferred among Archaea, Bacteria, and Eukarya (*2*–*7*). Among the consequences of these gene transfers are the gain of novel biochemical functions and the ability to colonize specific environmental niches (*3*, *8*, *9*). However, the molecular mechanisms for most of these transfer events are unknown.

Integrons are genetic elements known to facilitate horizontal gene transfer within Bacteria (*10*–*13*). Integrons can capture exogenous genes, known as gene cassettes, by site-specific recombination. Gene cassette capture is mediated by an integron integrase (IntI), which catalyzes the recombination between the recombination site of the inserting cassette (*attC*) and the endogenous integron attachment site (*attI*), immediately adjacent to the *intI* gene. Multiple gene cassettes can be inserted within a single integron, forming cassette arrays that range from 1 to more than 200 sequential cassettes (*10*, *12*). Integrons are mostly known for their role in driving the global antibiotic resistance crisis by disseminating diverse resistance determinants among bacterial pathogens (*14*–*16*). However, it is now clear that integrons play a much broader role in bacterial evolution and niche adaptation (*17*). The functions encoded by integron gene cassettes are extraordinarily diverse and extend far beyond those of clinical relevance (*13*, *18*, *19*).

To date, integrons have only been found within bacterial genomes, where they have been detected within diverse phyla (*20*). However, gene cassette amplicon sequencing has yielded cassette-encoded proteins that share homology with archaeal proteins (*21*, *22*). Without broader genomic context, however, the taxonomic residence of these gene cassettes is unknown.

Here, we screened all publicly available archaeal genomes to show that integrons are not only limited to Bacteria but also present in Archaea. Archaeal integrons exhibit the same characteristics and functional components as bacterial integrons. Furthermore, we demonstrate experimentally that diverse archaeal gene cassettes can be successfully recruited by a bacterial host, facilitated by integron-mediated recombination. Such a mechanism can thus permit bacteria to recruit archaeal gene cassettes present in their surrounding environment, with important implications for prokaryotic evolution.

RESULTS AND DISCUSSION

Discovery of integrons in Archaea

Here, we report the discovery of integrons in the domain Archaea. We screened 6718 archaeal genomes for integrons using the standard criteria applied to integron surveys in Bacteria (*20*, *23*, *24*). These include the presence of IntI genes and/or clusters of gene cassette *attC*s (defined as at least two *attC*s with less than 4 kb between each). We identified integrons in 75 archaeal metagenome-assembled genomes (MAGs) from nine phyla (Fig. 1 and data S1). It is not unexpected that integrons were detected only in MAGs, given that they constituted ~95% of all available archaeal genomes. However, to ensure that these integrons did not arise from contaminating bacterial contigs, incorrectly binned with archaeal MAGs, we applied stringent MAG refinement and quality filtering (see Materials and Methods for details). In addition, we found that ~7% of integron-bearing MAGs had at least one archaeal phylogenetic marker gene on the same contig as an integron (data S2), confirming these to be located on archaeal chromosomes. No integron was ever colocated with a bacterial marker gene. The markers used for this analysis consisted of a comprehensive set of 233 marker proteins identified as suitable for phylogenetic inference (*25*).

Among the 75 archaeal genomes, we detected six IntIs and 539 *attC* sites (excluding all singleton *attC*s). We found that archaeal integrons have a patchy distribution with varying prevalence across the phylogeny of Archaea (Fig. 1). In particular, integrons were

[1]School of Natural Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [2]ARC Centre of Excellence in Synthetic Biology, Macquarie University, Sydney, New South Wales 2109, Australia.
*Corresponding author. Email: timothy.ghaly@mq.edu.au

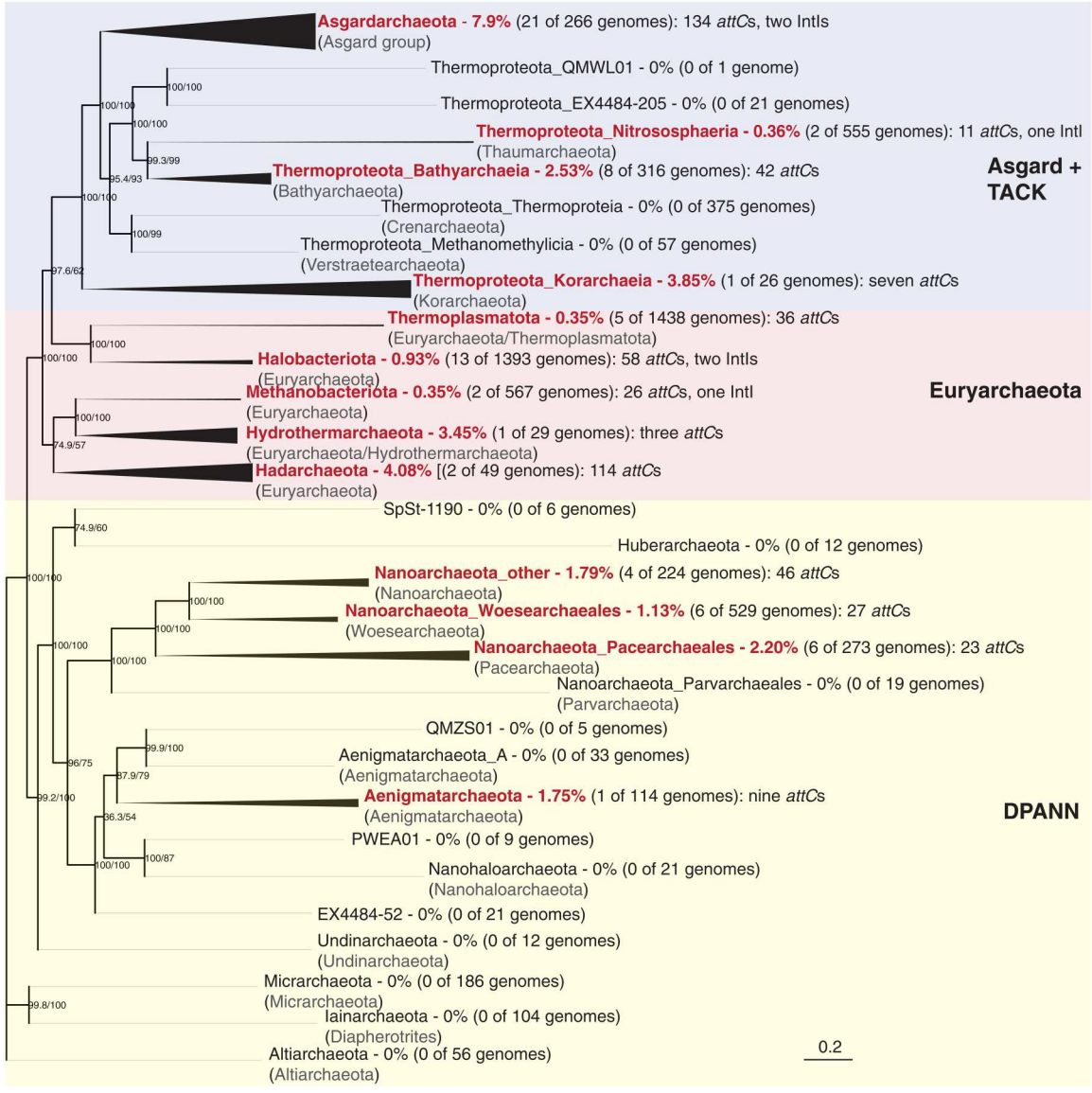


Fig. 1. Phylogenetic distribution of integrons among Archaea. Archaeal taxa found to carry integrons are labeled in red. Branch thickness indicates the proportion of genomes with integrons for each taxon. The phylogeny is based on a concatenated alignment of 53 top-ranked marker proteins recently identified to better recover monophyletic archaeal lineages (*55*). A maximum likelihood phylogenetic tree was inferred with the LG+C60+F+R model with an ultrafast bootstrap approximation (left node label) and Shimodaira–Hasegawa-like approximate likelihood ratio test (right node label), each run with 1000 replicates. The tree was artificially rooted with Altiarchaeota to agree with recent rooted archaeal phylogenies (*55*). Tips are labeled using the Genome Taxonomy Database (GTDB; top) and National Center for Biotechnology Information (NCBI; bottom) taxonomic nomenclature. To show major NCBI phyla collapsed under GTDB taxonomy, the GTDB phyla Thermoproteota (TACK group) and Nanoarchaeota (DPANN group) were split into class-level and order-level groupings, respectively. The scale bar indicates the average number of substitutions per site.

significantly enriched in the phylum Asgardarchaeota (χ^2 test, $P < 0.00001$) (Fig. 1), being detected in almost 8% of available Asgard genomes. Asgardarchaeota contributed the most genomes with detectable integrons (28%) and the greatest number of gene cassettes (24.9%), despite having relatively few genomes among the dataset (comprising 4% of available archaeal genomes). We also detected integrons in 3 to 4% of genomes from Hadarchaeota, Hydrothermoarchaeota, and Korarchaeia (Fig. 1), although these comprised few available genomes ($n < 50$). A patchy phylogenetic distribution of integrons has similarly been observed among Bacteria (*20*). For example, in the phylum Proteobacteria, integrons are enriched within the class Gammaproteobacteria (20% of genomes) while being entirely absent from its sister class Alphaproteobacteria. This is intriguing given that integrons have been detected at widely varying prevalence in more distantly related bacterial phyla such as Cyanobacteria, Spirochaetota, Planctomycetota, Chloroflexota, Bacteroidota, and Desulfobacterota (*20*, *23*).

Genetic structure of archaeal integrons

We found that archaeal integrons exhibit the same structure and functional components as bacterial integron cassette arrays (fig. S1). That is, tandem arrays of short open reading frames, generally

in the same orientation, interspersed by *attC* recombination sites. Archaeal *attC*s exhibit the same single-stranded folding structure as bacterial *attC*s, which is essential for them to act as structure-specific DNA recombination sites (*26*–*32*). We also note that archaeal IntIs exhibit the defining characteristics of bacterial IntIs, being tyrosine recombinases that have a unique IntI-specific additional domain surrounding the patch III motif region necessary for integron-mediated recombination (*33*). We found examples of "complete" integrons, these being cassette arrays adjacent to a detectable *intI* gene (fig. S1). We also found examples of putative *attI* sites, which act as insertion points for incoming gene cassettes. These *attI*s were immediately downstream of the *intI* gene, semi-conserved across distinct archaeal phyla (fig. S2, A and B), and exhibited the same canonical insertion point as all known bacterial *attI*s (fig. S2C).

Most archaeal integrons that we identified were CALINs (clusters of *attC*s lacking IntIs; data S3). This is not unexpected given the fragmented nature of MAGs and the high prevalence of CALINs also found in bacterial genomes. Among Bacteria, CALINs are more abundant than integrons that have an *intI* gene and exhibit a much wider taxonomic distribution (*20*). Note that cassettes within CALINs can still be excised and/or captured by exogenous IntIs, either in trans or following the uptake of DNA from a lysed cell. Thus, even without IntI genes, CALINs are still ecologically and evolutionarily important. Two so-called 'In0' elements were also detected among Archaea. These are integrons that have an *intI* gene without an adjacent *attC* site (fig. S1). However, both archaeal genomes with an In0 also had clusters of *attC* sites on other contigs. Among our dataset, the longest array of *attC*s on the same contig was 12; however, we found as many as 107 *attC*s (more than 18 contigs) within a single MAG (data S1). The number of *attC*s within a single MAG ranged from 2 to 107, with an average of seven *attC*s.

Platforms for cross-domain gene transfer

Archaeal gene cassettes with *attC*s from diverse phyla can be recognized and recruited by Bacteria (Fig. 2). We demonstrate that cassette insertion (*attC* × *attI* recombination) can occur following the conjugation of circular DNA molecules with archaeal *attC*s into an *Escherichia coli* recipient harboring a bacterial class 1 integron (Fig. 2A). Insertion events were confirmed with Sanger sequencing of the polymerase chain reaction (PCR)–amplified *attC*/*attI* recombination junctions (Fig. 2A and fig. S3). We found that recruitment of cassettes with archaeal *attC*s occurred at similar frequencies to that of the paradigmatic bacterial *attC* site, $attC_{aadA7}$, which we used as a positive control (Fig. 2B and table S1). We observed an average recombination frequency of 2.5×10^{-1} between *attI1* and $attC_{aadA7}$. Comparable frequencies (ranging from 1.9×10^{-4} to 3.2×10^{-1}, with an average of 5.1×10^{-2}) were observed for eight of the nine archaeal *attC*s (Kruskal-Wallis test, $P = 0.488$), which were selected from multiple archaeal phyla. Furthermore, we confirmed that cassette recruitment was mediated by IntI1 activity, because no *attC* × *attI* recombination events were detected when *intI1* was absent or when its expression was suppressed (table S1). This is of significant ecological and evolutionary importance, as the uptake of gene cassettes released into the environment from a lysed archaeal cell can be readily incorporated into bacterial genomes via integron-mediated recombination. We therefore show that integrons can facilitate gene transfer between the two domains of prokaryotes.

We find that the most clinically significant class of integrons (class 1) can recruit archaeal cassettes as efficiently as bacterial cassettes. Class 1 integrons are highly promiscuous because of their association with diverse mobile genetic elements, facilitating their spread into at least 104 bacterial species from 44 genera (*13*). They collectively carry more than 130 different resistance genes (*14*), most of which are of unknown taxonomic origin (*23*). Our findings open the possibility that Archaea could be an unexplored source of class 1 integron gene cassettes. Regardless, our findings indicate that any bacterial strain with a class 1 integron has the capacity to incorporate exogenous genes from diverse archaeal phyla, greatly expanding the genetic pool that they have access to.

The cross-domain transfer of integron gene cassettes is possibly widespread. For example, we detected 23 *attC*s from six archaeal genomes that exhibited 95 to 100% nucleotide identity to *attC*s within sequenced bacterial integrons (data S4). The archaeal *attC*s were from three phyla: Nanoarchaeota, Thermoproteota, and Hadarchaeota. The homologous *attC*s in Bacteria were found in 26 genomes from five phyla: Proteobacteria, Spirochaetota, Myxococcota, Nitrospirota, and Desulfobacterota. One of these *attC* sites was associated with a class 1 integron gene cassette, encoding a reduced form of nicotinamide adenine dinucleotide phosphate–dependent oxidoreductase found on five different Enterobacteriaceae plasmids (data S4). In Archaea, however, this *attC* site was part of a cassette that encoded a ligand-binding protein of unknown function. Nevertheless, because strong *attC* homology is a characteristic of cassettes that share the same taxonomic origin (*23*, *34*, *35*), it is possible that some clinically relevant gene cassettes now found on class 1 integrons might be of archaeal origin. We also find that an archaeal cassette (Methanobacteriota; accession: GCA_020055905.1), encoding a putative addiction module, shares a 92% nucleotide homology with a bacterial cassette (Planctomycetota; accession: AP021856), suggesting that a recent common ancestral gene has transferred between the two domains.

An analysis of cassette ribosome-binding sites (RBSs) suggests that the gene cassettes observed in archaeal genomes can be expressed by taxa from both prokaryotic domains. All RBS motifs associated with archaeal integrons are present among complete archaeal and bacterial genomes, respectively (data S5). This suggests that, following cross-domain transfer, it is possible for gene cassettes to be expressed within their new archaeal or bacterial host.

Diversity of IntIs

Archaeal IntIs are phylogenetically distinct from bacterial IntIs (Fig. 3). We detected six IntIs from four archaeal phyla (Fig. 1); however, three of these were excluded from further phylogenetic analysis based on either short sequence length (<200 amino acids) or partial coverage of the IntI-specific domain (fig. S4). We found that archaeal IntIs form their own monophyletic clade separate from known bacterial IntIs (*23*). This strongly suggests that IntI radiation has occurred within Archaea after a single ancient acquisition event from Bacteria. Regardless, we show that IntIs from distinct archaeal phyla, isolated from different environments, are more closely related to each other than they are to any bacterial IntI.

The closest sister clade to the archaeal IntIs comprises two Spirochaetota IntIs (Fig. 3). These two IntIs are phylogenetically distinct from "typical" Spirochaetota IntIs, which are generally in

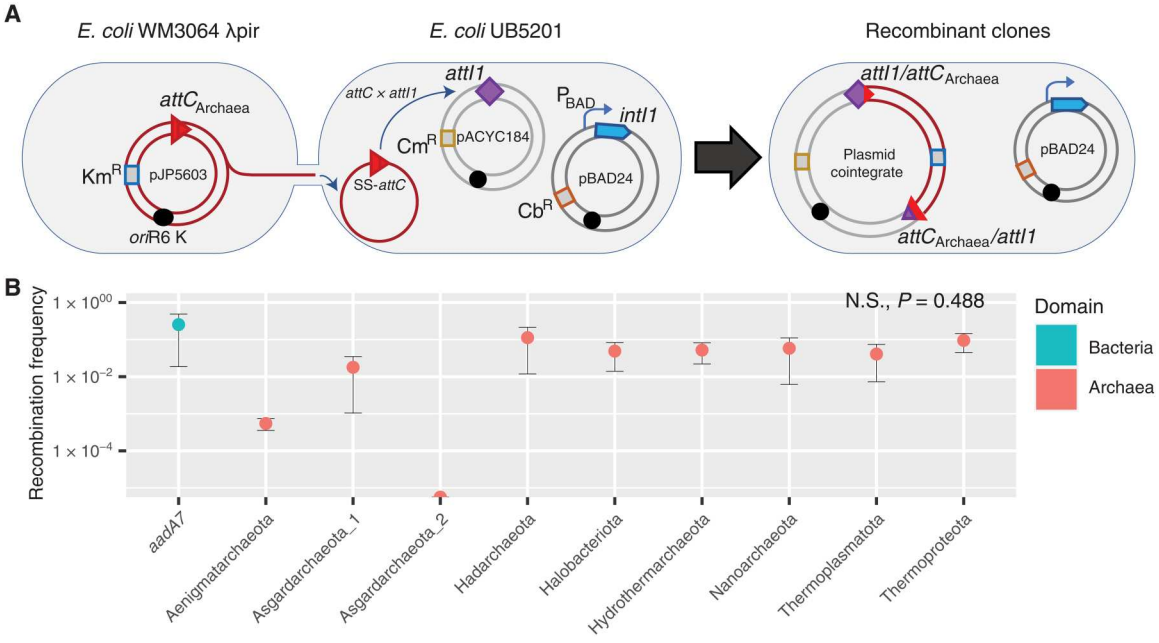


Fig. 2. Cassette recruitment (*attC* × *attI* recombination) assays. (**A**) Schematic outlining the experimental setup of the cassette insertion assays. The kanamycin resistance (Km^R) suicide vector pJP5603 with an *attC* site is delivered into the recipient *E. coli* UB5201 strain via conjugation. The recipient strain carries an *intI1* gene, expressed from the inducible P_{BAD} promoter, and an *attI1* site, residing on the carbenicillin resistance (Cb^R) pBAD24 and chloramphenicol resistance (Cm^R) pACYC184 backbones, respectively. The donor suicide vector cannot replicate within the recipient host and thus can only persist following *attC* × *attI* recombination to form a plasmid cointegrate. (**B**) Average recombination frequencies (log_{10} scale, ±1 SE) between *attI1* and nine archaeal *attC*s (with phyla of origin labeled along the *x* axis) and the paradigmatic bacterial *attC* site ($attC_{aadA7}$), used as positive control. Average frequencies were calculated following three independent cassette insertion assays (see Materials and Methods for details). No statistically significant difference in recombination frequencies were detected among the tested *attC*s (Kruskal-Wallis test, $n = 27$; df = 8, $P = 0.488$). Recombination frequencies are shown for *attC* bottom strands only. See table S1 for *attC* top strand recombination frequencies. N.S., not significant.

reverse orientation (*11*, *36*). Furthermore, the two Spirochaetota that harbored atypical IntIs were isolated from extreme environments: a brine layer within an alkaline lake and a hot spring, respectively; environments known to have a relatively high abundance of Archaea (*37*). Thus, these atypical Spirochaetota IntIs might have been horizontally acquired from Archaea that share the same extreme environments, although the direction of gene transfer cannot be determined with certainty.

Diversity of *attC* recombination sites

Archaeal *attC*s exhibit broad sequence and structural diversity (Fig. 4A). We find that some archaeal phyla have *attC*s with a restricted diversity (e.g., Hadarchaeota and Aenigmatarchaeota), while other phyla have extremely variable *attC*s distributed throughout the *attC* diversity space (e.g., Asgardarchaeota, Nanoarchaeota, and Thermoproteota). This distribution could indicate that different taxa have different propensities for horizontal exchange of gene cassettes (*13*, *23*). We show that archaeal *attC*s are significantly more similar within a genome than between genomes (Fig. 4B). This characteristic is also a hallmark of chromosomal bacterial integrons (*20*, *35*). We also show that *attC*s are more similar between different genomes from the same archaeal order than they are between genomes from different orders (Fig. 4C). This order-level *attC* homology is also seen within Bacteria (*23*, *34*). Thus, the ecological and evolutionary forces that promote and/or constrain *attC* diversity (*13*) are likely to be similar for both Archaea and Bacteria.

There is an extensive overlap in the sequence and structural diversity of *attC*s from Archaea and Bacteria (Fig. 4A), which are regularly placed in the same clade. This provides evidence that there is a mechanistic overlap between archaeal and bacterial *attC*s and suggests that gene cassette transfer between the two domains does occur. It also suggests that the recruitment of extra domain gene cassettes can be facilitated by diverse classes of integrons, of which there are thousands [based on IntI amino acid homology (*38*)]. The broad distribution of integrons among the two domains suggests that integron-mediated transfer plays an important role in prokaryotic evolution.

Functional diversity of gene cassettes

We detected 549 cassette-encoded proteins among Archaea. Only 23.1% of these could be classified into a known Clusters of Orthologous Groups of proteins (COG) category (fig. S5). In contrast, 47.4% of all proteins from the 75 integron-bearing archaeal genomes could be assigned a known COG category. This underrepresentation (χ^2 test, $P < 0.00001$) of known COGs among cassette proteins has previously been reported for bacterial integrons (*10*, *11*, *34*). To gain further insight into possible cassette functions, eggNOG 5.0 (*39*) and Pfam (*40*) database searches were performed, assigning putative functions to 228 (41.5%) of the archaeal cassette–encoded proteins. Of those with functional predictions, proteins involved in toxin-antitoxin (TA) systems (10.5%), phage resistance proteins via DNA methylation or restriction endonuclease activities (8.3%), and acetyltransferases (4.4%) were particularly prevalent

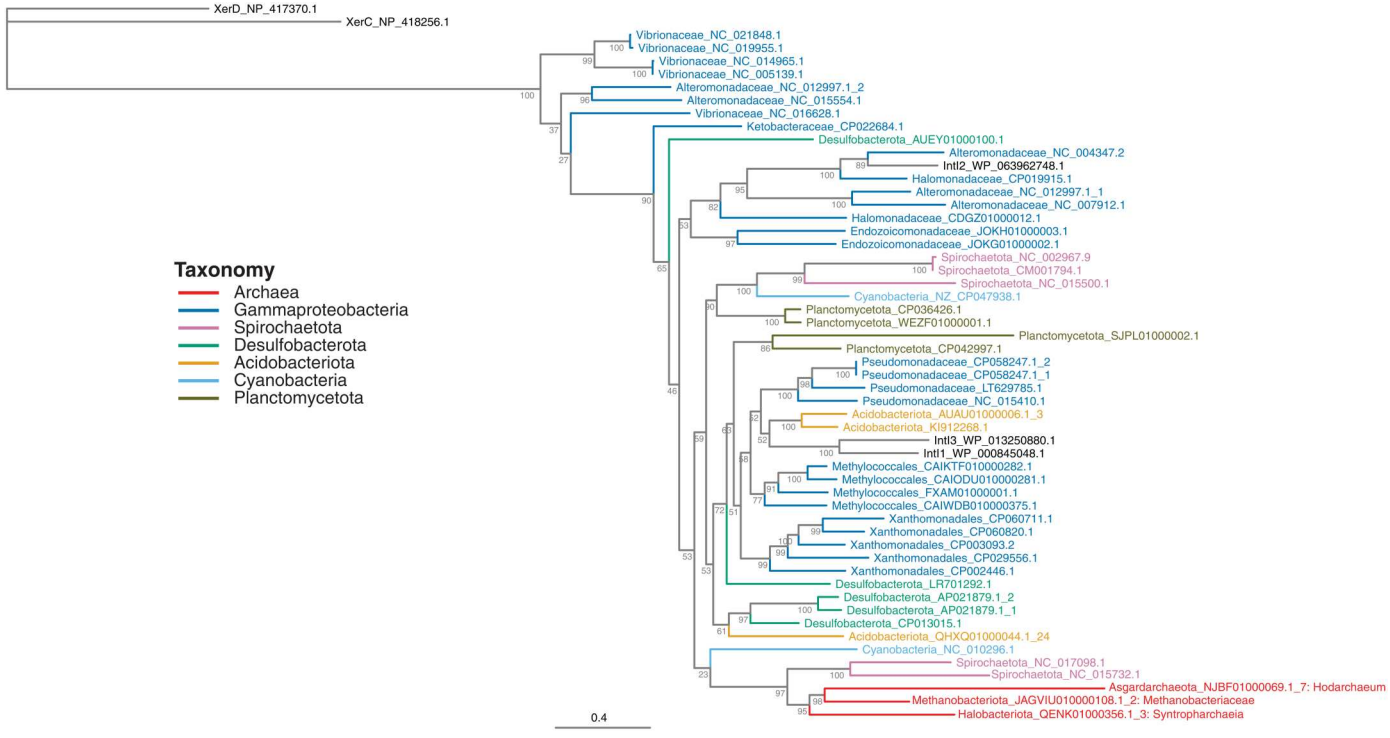


Fig. 3. Phylogeny of IntIs from Archaea and Bacteria. To root the tree, the tyrosine recombinases XerC and XerD from *E. coli* were used as out-groups. Integron integrases (IntIs) are colored according to their taxonomy. Archaeal IntIs are also labeled with their lowest taxonomic classification. The scale bar indicates the average number of substitutions per site.

(data S6). These are the functions most commonly reported for gene cassettes in Bacteria (*11*, *13*, *34*, *35*, *41*). TA gene cassettes are particularly common in bacterial integrons, where they can stabilize very large cassette arrays (*42*, *43*). The antitoxin modules of TA cassettes can also counteract the toxins of homologous systems found on plasmids and phage, thus potentially protecting their host from invading mobile elements (*44*, *45*).

In addition, 13.2% of archaeal cassette–encoded proteins had signal peptides, which represents a significant enrichment relative to their broader genomic contexts (6.9%; χ^2 test, $P < 0.00001$). Signal peptides are short amino acid tag sequences that target proteins into, or across, membranes. Again, transmembrane and secreted proteins are commonly encoded by gene cassettes in Bacteria (*34*) and are hypothesized to help facilitate interactions with their broader environment (*13*).

We find that functions of archaeal cassettes are associated with their environment (Fig. 5). Functional families cluster according to their specific environment, and these environmental clusters, in turn, group according to their broader environmental type (Fig. 5). This environmentally explicit clustering might be the result of local ecological and evolutionary forces. That is, gene cassettes in Archaea confer niche-specific functional traits and/or horizontal transfer of cassettes that occurs between archaeal phyla colocated in the same environment.

Here, we present the first evidence of integrons in the domain Archaea. We demonstrate that they have the same functional characteristics as bacterial integrons. We also present experimental evidence that bacteria can successfully recruit archaeal gene cassettes, facilitated by integron-mediated DNA recombination. Our results thus establish a novel mechanism for cross-domain gene transfer between Archaea and Bacteria, whereby integron-carrying bacteria can incorporate archaeal gene cassettes present in their surrounding environment. We also find that, although archaeal IntIs are phylogenetically distinct from bacterial IntIs, their associated *attC* recombination sites are shared with Bacteria. This suggests that integron-mediated cross-domain gene transfer might be common and could play an important role in prokaryotic evolution.

MATERIALS AND METHODS

Data acquisition and quality filtering

All available archaeal genomes were downloaded from the National Center for Biotechnology Information (NCBI) Assembly Database (n = 8160; last accessed 5 October 2021). Of these, ~95% were MAGs. We applied stringent filtering criteria to remove low-quality MAGs. First, to improve MAG quality, we identified and removed contaminating contigs from each MAG using MAGpurify v2.1.2 (*46*) with the following modules: "phylo-markers," which finds taxonomically discordant contigs using 100 archaeal and 88 bacterial single-copy taxonomic marker genes from the PhyEco database (*47*); "clade-markers," which finds contaminating contigs using a database of 855,764 clade-specific prokaryotic marker genes [MetaPhlAn2 database (*48*)]; "tetra-freq," which uses principal components analysis (PCA) to identify contaminating contigs with outlier tetranucleotide frequency; and "gc-content," which uses PCA to identify contaminating contigs with outlier GC content.

After refinement, the quality of the genomes was assessed using CheckM v1.1.3 (*49*), which uses single-copy lineage-specific marker

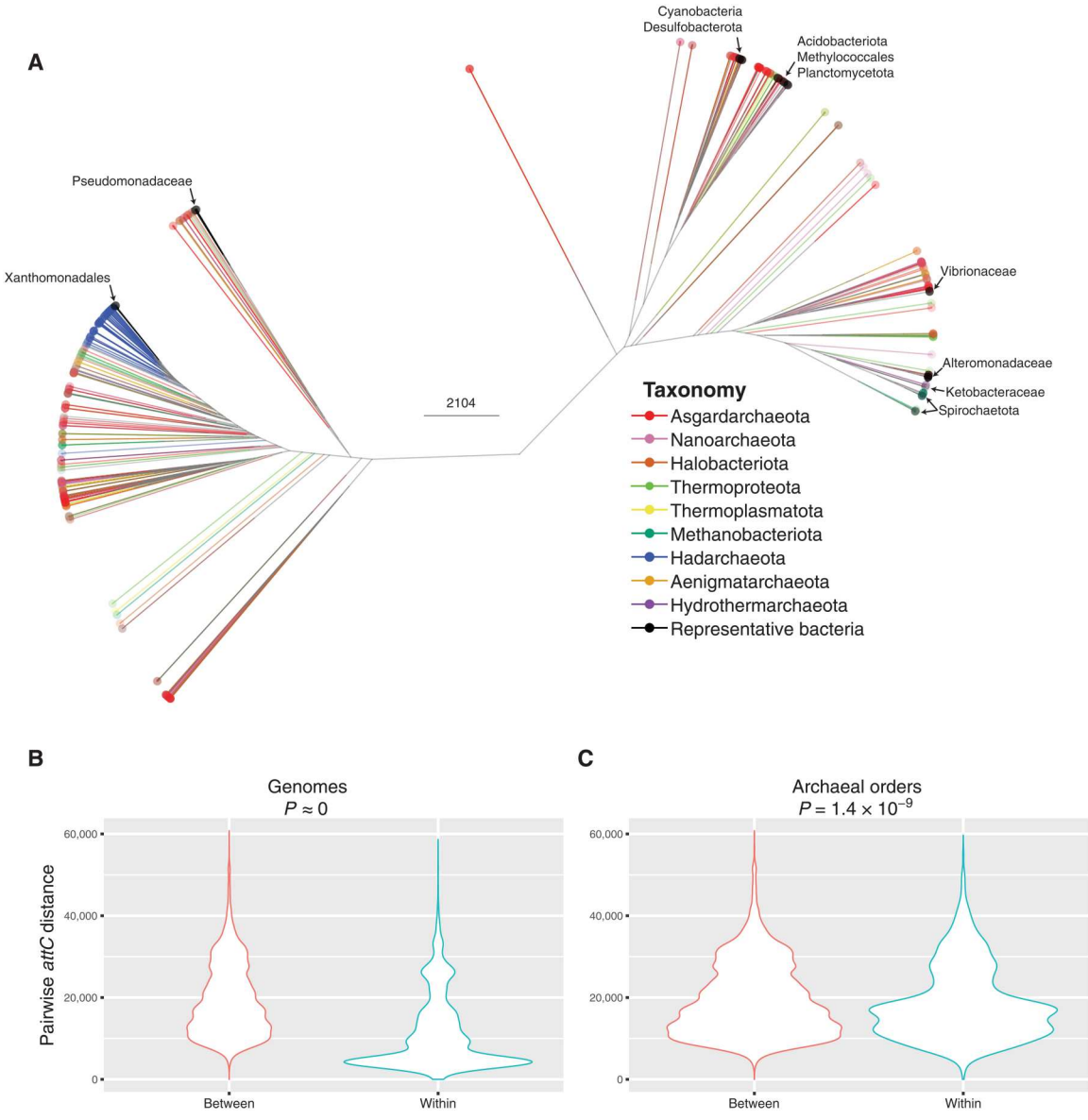
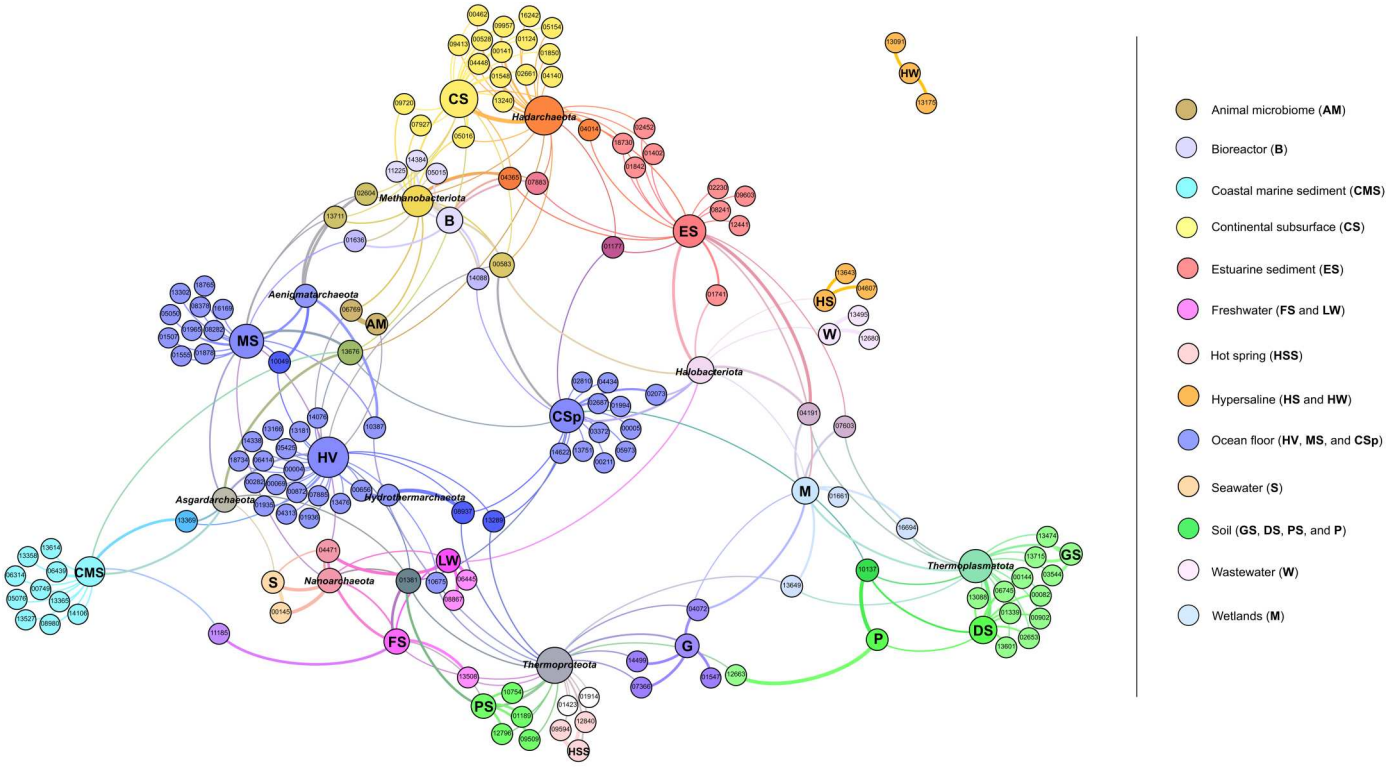


Fig. 4. Structural and sequence diversity of archaeal *attC* recombination sites. (**A**) Structure-based clustering of all archaeal and representative bacterial *attC*s. Branches and tips are colored according to archaeal phylum. The taxa of bacterial *attC*s are labeled with arrows. The scale bar indicates pairwise *attC* alignment distances as determined by RNAclust (*61*) and LocARNA (*62, 63*). (**B**) Distribution of the sequence and structural distances calculated for all pairwise comparisons of *attC*s within and between genomes. (**C**) Distribution of distances for all pairwise comparisons of *attC*s from different genomes that are either from the same or different archaeal orders.

genes to estimate genome completeness and contamination. There is strong community consensus that high-quality MAGs are those that are more than 90% complete and have less than 5% contamination, while medium-quality MAGs have a completeness of ≥50% and contamination of <10% (*25*, *46*, *50*–*53*). In this context, however, we were more concerned with the level of contamination than completeness and thus removed all genomes with an estimated contamination of ≥5%. The completeness of the remaining genomes ranged from 15 to 100%, with a median of 81%. The estimated contamination ranged from 0 to 4.98%, with a median of 0.93%.

Archaeal genomes were assigned taxonomic classifications based on the Genome Taxonomy Database (GTDB) (*50*–*52*) using GTDB-Tk v1.6.0 (*54*) with release 06-RS202 of the GTDB. We used the classify_wf command with default settings. This workflow identifies and aligns 120 bacterial and 122 archaeal phylogenetic marker genes (*25*). GTDB-Tk then classifies each genome based on its placement into domain-specific reference trees (built from 47,899 prokaryote genomes), its relative evolutionary divergence, and average nucleotide identity to reference genomes in the GTDB. Any genomes not classified within the domain Archaea were removed. This resulted in a final set of 6718 archaeal genomes retained for further analysis.

To infer the phylum-level phylogeny of Archaea, the highest quality representative genome from each phylum was selected on the basis of its genome quality score [defined by Parks *et al.* (*25*)



AM, animal microbiome; **B**, bioreactor; **CMS**, coastal marine sediment; **CSp**, cold seep; **CS**, continental subsurface; **DS**, desert soil; **ES**, estuarine sediment; **FS**, freshwater sediment; **G**, groundwater; **GS**, grassland soil; **HSS**, hot spring sediment; **HV**, hydrothermal vent; **HS**, hypersaline sediment; **HW**, hypersaline water; **MS**, marine sediment; **LW**, lake water; **M**, mangroves; **P**, peatland; **PS**, permafrost soil; **S**, seawater; **W**, wastewater

Fig. 5. A network linking Pfam functions of archaeal integron gene cassettes with their taxonomic and environmental contexts. The force-directed representation of the network is constructed on the basis of co-occurrence patterns and correlations ($P < 0.05$) among Pfam functions, taxonomic groups, and specific environments from which the organisms were sampled. Nodes that represent taxonomic groups and specific environments are labeled accordingly. All other nodes denote Pfam functions and are labeled with a Pfam number preceded by "PF." Specific environments are grouped into broader environment types, each of which is colored as per the panel. Pfams directly linked to specific environment types are colored in corresponding colors. Pfams linked to more than one environment type are colored in overlapping colors. The size of the node is relative to the node authority based on the degree of correlations. Edges (the lines connecting the nodes) represent correlations between nodes. Edge color denotes the overlapping color of the two nodes it connects. Edge thickness represents the strength of correlation. The full description of all correlations and Pfam functions is presented in data S7.

as the estimated completeness of a genome minus five times its estimated contamination]. We inferred the phylogeny of the representative genomes using a set of 53 top-ranked marker proteins recently identified to better recover monophyletic lineages in Archaea (*55*). We used the updated version of GTDB-Tk v2.0.0 (*54*) to generate a concatenated and trimmed multiple sequence alignment of the 53 marker proteins. A maximum likelihood phylogenetic tree was generated from the alignment using IQ-TREE v1.6.12 (*56*) with the LG+C60+F+R model (parameters: -m LG+C60+F+R -bb 1000 -alrt 1000).

Integron detection

Because of faster processing speeds of large datasets, we initially screened all filtered genomes for *attC* recombination sites using *attC*-screening.sh (*38*) (https://github.com/timghaly/integron-filtering) with default parameters. This script uses the HattCI (*57*) + Infernal (*58*) pipeline [first described by Pereira *et al.* (*24*)] to search for the conserved sequence and structure of *attC* sites. Genomes that had at least one detectable *attC* site were additionally screened using IntegronFinder v2.0rc6 (parameters: --local-max --cpu 24 --gbk) (*20*), which searches for IntIs and gene cassette arrays. Only IntIs, *attC*s, and cassette-encoded proteins identified by IntegronFinder were included in downstream analyses.

To ensure that these integrons were not from contaminating bacterial contigs that had been incorrectly binned with archaeal MAGs, we screened all contigs containing an integron for prokaryotic marker genes using GTDB-Tk v1.6.0 (*54*). These consisted of a comprehensive set of 233 proteins identified as suitable phylogenetic markers (*25*). We found a total of nine prokaryotic marker genes among seven integron-bearing contigs from five genomes. To identify the taxonomy of the marker genes, we searched for their homologs in the NCBI nr database via a BLASTP search and found that they best aligned with genes from a diverse set of archaeal genomes (excluding the specific genomes in which they derived from). None of these prokaryotic marker genes were ever identified as bacterial. See data S2 for a list of marker proteins and BLASTP hits.

Analysis of IntIs, *attC* sites, and cassette-encoded proteins

IntegronFinder identifies IntIs using the overlap of two protein hidden Markov model profiles. The first is the Pfam profile PF00589 to identify tyrosine recombinases, and the second is a protein profile built from the IntI-specific domain that separates

IntIs from other tyrosine recombinases (*33*). Identified archaeal IntIs, with matches to both protein profiles, were placed in a phylogeny alongside a set of previously identified bacterial IntIs (*23*). IntIs shorter than 200 amino acids or those that did not span the complete IntI-specific domain, needed to distinguish IntIs from other tyrosine recombinases, were removed from phylogenetic analysis. The remaining IntIs were aligned using MAFFT v7.271 (parameters: --localpair --maxiterate 1000) (*59*) and trimmed using trimAl v1.2rev59 (parameters: -automated1). A maximum likelihood tree was generated from the alignment using IQ-TREE v1.6.12 (*56*) with the best-suited protein model as determined by ModelFinder (*60*) and 1000 bootstrap replicates (parameters: -m MFP -bb 1000).

The sequence and structural diversity of *attC*s was assessed using RNAclust v1.3 (*61*) as previously described (*23*). RNAclust uses LocARNA (*62*, *63*) to generate pairwise structural alignments (based on both sequence and folding structure) of input sequences. RNAclust then calculates pairwise distances to create a hierarchical clustering tree from a Weighted Pair Group Method with Arithmetic Mean (WPGMA) analysis. All archaeal *attC*s along with a set of previously identified *attC*s from representative bacterial taxa (*23*) were clustered using RNAclust′s default parameters.

Cassette-encoded proteins identified by IntegronFinder were functionally annotated using InterProScan v5.44-79.0 (*64*), with default parameters against the Pfam (*40*) database, and eggNOG-mapper v2.0.1b (*39*, *65*, *66*), executed in DIAMOND (*67*) mode. To identify cassettes that encode transmembrane and secreted proteins, we searched protein sequences for prokaryotic signal peptides using SignalP 5.0 (*68*) with default parameters. The correlation analysis of cassette functions was performed as described in Penesyan *et al.* (*69*). Briefly, Pearson′s correlations based on co-occurrences among Pfam functions, specific environments, and archaeal phyla were calculated using the Hmisc v4.5-0 R package (*70*). The network was generated from all positive correlations with *P* values of <0.05 using the ForceAtlas2 layout algorithm (*71*) within the Gephi software (*72*). Specific correlations and the description of Pfam functions are listed in data S7.

RBS motifs associated with archaeal cassettes were detected using Prodigal v2.6.3 (*73*) with the implementation of a full RBS motif scan (parameters: -p meta -q -n). To compare RBS motifs against those detected among complete archaeal and bacterial genomes, we downloaded all RefSeq Archaeal genomes (n = 443 genomes, 1.2×10^6 genes; downloaded 11 July 2022) and one representative genome from every bacterial order in Reference Sequence (RefSeq) (n = 416 genomes, 1.3×10^6 genes; downloaded 11 July 2022). Using Prodigal with the above parameters, RBS motifs were predicted for all archaeal and bacterial genomes.

Bacterial strains and plasmids for *attC* recombination assays

The bacterial strains and plasmids used in this study are listed in data S8. LB medium (Lennox) was used to grow bacterial strains supplemented with appropriate antimicrobial agents. The final concentrations of antimicrobial agents used were kanamycin (Km; 50 μg/ml), carbenicillin (Cb; 75 μg/ml), and chloramphenicol (Cm; 20 μg/ml). LB medium was supplemented with 0.3 mM 2,6-diaminopimelic (DAP) acid to culture the auxotrophic *E. coli* WM3064 λpir strain (*74*).

Construction of *attC* donor strains

Nine archaeal *attC*s, selected from diverse archaeal phyla (data S9) along with one bacterial *attC* ($attC_{aadA7}$) were used for the recombination assays. Two donor strains were constructed for each *attC*, delivering either the *attC* top or bottom strands via conjugation. Overlapping forward and reverse primers were designed to generate each *attC* sequence flanked by *Xba*I and *Bam*HI overhangs, respectively (e.g., primer pair *attC-aadA7*-FW/REV for $attC_{aadA7}$). The annealed primer dimers were then ligated into the mobilizable suicide vector pJP5603 (*75*, *76*). The *attC* top strand donor strains were generated by transforming the ligation product into electrocompetent cells of the DAP auxotrophic *E. coli* strain WM3064 λpir. Using the same procedures, all *attC* top strand donor plasmids and strains were constructed using the pairs of long primers listed in data S10.

To deliver *attC* bottom strands, the pJP5603rev (pJPrev) vector was generated to invert *oriT* orientation relative to that of the pJP5603 parental vector. The multiple-cloning site and vector backbone of pJP5603 were PCR-amplified using the primer pairs pJP-MCS-FW/REV and pJP-Backbone-FW/REV, respectively (with *Xho*I and *Mlu*I restriction sites introduced), followed by restriction digest and ligation. The same primer pairs for generating the top strand donor plasmids were used to create the bottom strand donor plasmids and strains by cloning the same *attC* sequences into the *Xba*I/*Bam*HI sites of pJPrev.

Construction of the recipient strain

We generated a recipient strain using *E. coli* UB5201 (*77*) that carried the *intI1* gene and the *attI1* recombination site residing on the pBAD24 (*78*) and pACYC184 (*79*) backbones, respectively. The *intI1* gene of the R388 plasmid (*80*) was PCR-amplified using the primer pair *intI1_Eco*RI-F/*intI1_Hin*dIII-R (data S10). The L-arabinose inducible pBAD::*intI1* plasmid was generated by cloning *intI1* into the pBAD24 expression vector. The pACYC184::*attI1* recipient plasmid was created by assembling the *attI1* sequence (from R388) into the pACYC184 plasmid backbone using the NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs, USA). The PCR products required for the assembly were generated using the *attI1*_fw/*attI1*_rev and pACYC184_-backbone_F/pACYC184_backbone_R primer pairs. *E. coli* UB5201 strain was cotransformed with pBAD::*intI1* and pACYC184::*attI1* to generate the *E. coli* UB5201 + pBAD::*intI1* + pACYC184::*attI1* recipient strain for *attC* × *attI* suicide conjugation assays. *E. coli* UB5201 + pBAD24 + pACYC184::*attI1* was created as an *intI1*-negative control strain. All plasmid constructs were confirmed by Sanger sequencing and restriction enzyme digests.

attC × *attI* suicide conjugation assays

The frequencies of recombination between the archaeal *attC* sequences and the class 1 integron *attI1* site were quantified using previously established *attC* × *attI* suicide conjugation methods (*26*, *30*, *32*, *81*, *82*) with minor modifications. Briefly, the Cb-resistant UB5201 + pBAD::*intI1* + pACYC184::*attI1* recipient strain was filter-mated with Km-resistant WM3064 λpir *attC* donor strains in DAP-supplemented LB media. The expression of *intI1* was either induced using L-arabinose (2 mg/mL) or suppressed with D-glucose (10 mg/mL). After 6 hours of incubation at 37°C, the recovered conjugation mix was plated on DAP-free LB agar with Km and on LB agar containing Cb. This method was allowed for

negative selection of the donor strain, which cannot grow in the absence of DAP, and positive selection of the recombinant recipient clones, which become Km resistant following plasmid cointegration (Fig. 2A). The recombination frequency was determined as the ratio of the colony-forming units (CFUs) for Km-resistant recombinants to the CFU for the total number of Cb-resistant recipients after 2 days of incubation. All assays were performed in three biological replicates, and recombination frequencies were calculated as the mean of the three independent experiments. To confirm the cointegrates, colony PCR was performed on eight randomly chosen colonies per conjugation set for each biological replicate using the following primer pairs: pACYC_F/M13F and pACYC_R/M13R (fig. S3). Sanger sequencing of PCR products was performed for four recombinant colonies per conjugation set.

Supplementary Materials

This PDF file includes:

Table S1
Figs. S1 to S5

Other Supplementary Material for this manuscript includes the following:

Data S1 to S10

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. M. Bruto, C. Prigent-Combaret, P. Luis, G. Hoff, Y. Moënne-Loccoz, D. Muller, in *Evolutionary Biology: Exobiology and Evolutionary Mechanisms*, P. Pontarotti, Ed. (Springer Berlin Heidelberg, 2013), pp. 165–179.
2. K. M. Sutherland, L. M. Ward, C.-R. Colombero, D. T. Johnston, Inter-domain horizontal gene transfer of nickel-binding superoxide dismutase. *Geobiology* **19**, 450–459 (2021).
3. G. Schönknecht, W. H. Chen, C. M. Ternes, G. G. Barbier, R. P. Shrestha, M. Stanke, A. Bräutigam, B. J. Baker, J. F. Banfield, R. M. Garavito, K. Carr, C. Wilkerson, S. A. Rensing, D. Gagneul, N. E. Dickenson, C. Oesterhelt, M. J. Lercher, A. P. M. Weber, Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207–1210 (2013).
4. N.-U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**, 847–850 (2006).
5. J. C. Dunning Hotopp, Horizontal gene transfer between bacteria and animals. *Trends Genet.* **27**, 157–163 (2011).
6. R. Bock, The give-and-take of DNA: Horizontal gene transfer in plants. *Trends Plant Sci.* **15**, 11–22 (2010).
7. K. E. Nelson, R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, C. M. Fraser, Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
8. F. Husnik, J. P. McCutcheon, Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.* **16**, 67–79 (2018).
9. J. A. Metcalf, L. J. Funkhouser-Jones, K. Brileya, A.-L. Reysenbach, S. R. Bordenstein, Antibacterial gene transfer across the tree of life. *eLife* **3**, e04266 (2014).
10. D. Mazel, Integrons: Agents of bacterial evolution. *Nat. Rev. Microbiol.* **4**, 608–620 (2006).
11. Y. Boucher, M. Labbate, J. E. Koenig, H. W. Stokes, Integrons: Mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol.* **15**, 301–309 (2007).
12. M. R. Gillings, Integrons: Past, present, and future. *Microbiol. Mol. Biol. Rev.* **78**, 257–277 (2014).
13. T. M. Ghaly, M. R. Gillings, A. Penesyan, Q. Qi, V. Rajabal, S. G. Tetu, The natural history of integrons. *Microorganisms* **9**, 2212 (2021).
14. S. R. Partridge, G. Tsafnat, E. Coiera, J. R. Iredell, Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol. Rev.* **33**, 757–784 (2009).
15. Y.-G. Zhu, Y. Zhao, B. Li, C. L. Huang, S. Y. Zhang, S. Yu, Y. S. Chen, T. Zhang, M. R. Gillings, J. Q. Su, Continental-scale pollution of estuaries with antibiotic resistance genes. *Nat. Microbiol.* **2**, 16270 (2017).
16. T. M. Ghaly, I. T. Paulsen, A. Sajjad, S. G. Tetu, M. R. Gillings, 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 (2020).
17. J. A. Escudero, C. Loot, A. Nivina, D. Mazel, The integron: Adaptation on demand. *Microbiol. Spectr.* **3**, MDNA3-0019-2014 (2015).
18. T. M. Ghaly, J. L. Geoghegan, S. G. Tetu, M. R. Gillings, The peril and promise of integrons: Beyond antibiotic resistance. *Trends Microbiol.* **28**, 455–464 (2020).
19. T. M. Ghaly, J. L. Geoghegan, J. Alroy, M. R. Gillings, High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environ. Microbiol.* **21**, 1567–1574 (2019).
20. J. Cury, T. Jové, M. Touchon, B. Néron, E. P. Rocha, Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
21. H. Elsaied, H. W. Stokes, H. Yoshioka, Y. Mitani, A. Maruyama, Novel integrons and gene cassettes from a Cascadian submarine gas-hydrate-bearing core. *FEMS Microbiol. Ecol.* **87**, 343–356 (2014).
22. J. E. Koenig, C. Sharp, M. Dlutek, B. Curtis, M. Joss, Y. Boucher, W. F. Doolittle, Integron gene cassettes and degradation of compounds associated with industrial waste: The case of the Sydney Tar Ponds. *PLOS ONE* **4**, e5276 (2009).
23. T. M. Ghaly, S. G. Tetu, M. R. Gillings, Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of *attC* sites. *Commun. Biol.* **4**, 946 (2021).
24. M. B. Pereira, T. Österlund, K. M. Eriksson, T. Backhaus, M. Axelson-Fisk, E. Kristiansson, A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* **21**, 495 (2020).
25. D. H. Parks, C. Rinke, M. Chuvochina, P. A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, G. W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
26. M. Bouvier, G. Demarre, D. Mazel, Integron cassette insertion: A recombination process involving a folded single strand substrate. *EMBO J.* **24**, 4356–4367 (2005).
27. D. MacDonald, G. Demarre, M. Bouvier, D. Mazel, D. N. Gopaul, Structural basis for broad DNA-specificity in integron recombination. *Nature* **440**, 1157–1162 (2006).
28. M. Bouvier, M. Ducos-Galand, C. Loot, D. Bikard, D. Mazel, Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLOS Genet.* **5**, e1000632 (2009).
29. G. Demarre, C. Frumerie, D. N. Gopaul, D. Mazel, Identification of key structural determinants of the IntI1 integron integrase that influence *attC*×*attI1* recombination efficiency. *Nucleic Acids Res.* **35**, 6475–6489 (2007).
30. A. Nivina, J. A. Escudero, C. Vit, D. Mazel, C. Loot, 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 (2016).
31. A. Mukhortava, M. Pöge, M. S. Grieb, A. Nivina, C. Loot, D. Mazel, M. Schlierf, Structural heterogeneity of *attC* integron recombination sites revealed by optical tweezers. *Nucleic Acids Res.* **47**, 1861–1870 (2019).
32. A. Nivina, M. S. Grieb, C. Loot, D. Bikard, J. Cury, L. Shehata, J. Bernardes, D. Mazel, Structure-specific DNA recombination sites: Design, validation, and machine learning–based refinement. *Sci. Adv.* **6**, eaay2922 (2020).
33. N. Messier, P. H. Roy, Integron integrases possess a unique additional domain necessary for activity. *J. Bacteriol.* **183**, 6699–6706 (2001).
34. D. A. Rowe-Magnus, A.-M. Guerout, L. Biskri, P. Bouige, D. Mazel, Comparative analysis of superintegrons: Engineering extensive genetic diversity in the Vibrionaceae. *Genome Res.* **13**, 428–442 (2003).
35. D. A. Rowe-Magnus, A. M. Guerout, P. Ploncard, B. Dychinco, J. Davies, D. Mazel, The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 652–657 (2001).
36. Y.-W. Wu, T. G. Doak, Y. Ye, The gain and loss of chromosomal integron systems in the *Treponema* species. *BMC Evol. Biol.* **13**, 16 (2013).
37. P. H. Rampelotto, Extremophiles and extreme environments. *eLife* **3**, 482–485 (2013).
38. T. M. Ghaly, A. Penesyan, A. Pritchard, Q. Qi, V. Rajabal, S. G. Tetu, M. R. Gillings, Methods for the targeted sequencing and analysis of integrons and their gene cassettes from complex microbial communities. *Microb. Genomics* **8**, 000788 (2022).
39. J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, P. Bork, eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
40. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan,

S. C. E. Tosatto, R. D. Finn, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
41. G. Cambray, A.-M. Guerout, D. Mazel, Integrons. *Annu. Rev. Genet.* **44**, 141–166 (2010).
42. N. Iqbal, A.-M. Guérout, E. Krin, F. Le Roux, D. Mazel, Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin systems substantiates their role in stabilizing the superintegron. *J. Bacteriol.* **197**, 2150–2159 (2015).
43. S. Szekeres, M. Dauti, C. Wilde, D. Mazel, D. A. Rowe-Magnus, Chromosomal toxin–Antitoxin loci can diminish large-scale genome reductions in the absence of selection. *Mol. Microbiol.* **63**, 1588–1605 (2007).
44. M. Wilbaux, N. Mine, A.-M. Guérout, D. Mazel, L. Van Melderen, Functional interactions between coexisting toxin-antitoxin systems of the *ccd* family in *Escherichia coli* O157: H7. *J. Bacteriol.* **189**, 2712–2719 (2007).
45. A.-M. Guérout, N. Iqbal, N. Mine, M. Ducos-Galand, L. van Melderen, D. Mazel, Characterization of the *phd-doc* and *ccd* toxin-antitoxin cassettes from *Vibrio* superintegrons. *J. Bacteriol.* **195**, 2270–2283 (2013).
46. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
47. D. Wu, G. Jospin, J. A. Eisen, Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLOS ONE* **8**, e77033 (2013).
48. D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, N. Segata, MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
49. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
50. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
51. D. H. Parks, M. Chuvochina, P. A. Chaumeil, C. Rinke, A. J. Mussig, P. Hugenholtz, A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
52. C. Rinke, M. Chuvochina, A. J. Mussig, P. A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the genome taxonomy database. *Nat. Microbiol.* **6**, 946–959 (2021).
53. R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke; Genome Standards Consortium, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
54. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).
55. N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J. H. Lee, B. Q. Minh, C. Rinke, A. Spang, Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).
56. L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
57. M. B. Pereira, M. Wallroth, E. Kristiansson, M. Axelson-Fisk, HattCI: Fast and accurate *attC* site identification using hidden Markov models. *J. Comput. Biol.* **23**, 891–902 (2016).
58. E. P. Nawrocki, S. R. Eddy, Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
59. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
60. S. Kalyaanamoorthy, B. Q. Minh, T. K. Wong, A. Von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
61. J. Engelhardt, S. Heyne, S. Will, K. Reiche, RNAclust: A tool for clustering of RNAs based on their secondary structures using LocARNA. http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAclust/?p=index.html (2010).
62. S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, R. Backofen, LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900–914 (2012).
63. S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, R. Backofen, Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput. Biol.* **3**, e65 (2007).
64. P. Jones, D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez, S. Hunter, InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
65. J. Huerta-Cepas, K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering, P. Bork, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
66. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
67. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
68. J. J. Almagro Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
69. A. Penesyan, S. S. Nagy, S. Kjelleberg, M. R. Gillings, I. T. Paulsen, Rapid microevolution of biofilm cells in response to antibiotics. *npj Biofilms Microbiomes* **5**, 34 (2019).
70. F. E. Harrell, C. Dupont, Hmisc: Harrell miscellaneous. R package version 4.5-0; https://CRAN.R-project.org/package=Hmisc(2021).
71. M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLOS ONE* **9**, e98679 (2014).
72. M. Bastian, S. Heymann, M. Jacomy, in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media* (AAAI, 2009).
73. D. Hyatt, G. L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
74. C. Dehio, M. Meyer, Maintenance of broad-host-range incompatibility group P and group Q plasmids and transposition of Tn*5* in *Bartonella henselae* following conjugal plasmid transfer from *Escherichia coli*. *J. Bacteriol.* **179**, 538–540 (1997).
75. R. J. Penfold, J. M. Pemberton, An improved suicide vector for construction of chromosomal insertion mutations in bacteria. *Gene* **118**, 145–146 (1992).
76. T. Riedel, M. Rohlfs, I. Buchholz, I. Wagner-Döbler, M. Reck, Complete sequence of the suicide vector pJP5603. *Plasmid* **69**, 104–107 (2013).
77. J. Sanchez, P. M. Bennett, M. H. Richmond, Expression of *elt*-B, the gene encoding the B subunit of the heat-labile enterotoxin of *Escherichia coli*, when cloned in pACYC184. *FEMS Microbiol. Lett.* **14**, 1–5 (1982).
78. L. M. Guzman, D. Belin, M. J. Carson, J. Beckwith, Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).
79. R. E. Rose, The nucleotide sequence of pACYC184. *Nucleic Acids Res.* **16**, 355 (1988).
80. P. Avila, F. de la Cruz, Physical and genetic map of the IncW plasmid R388. *Plasmid* **20**, 155–157 (1988).
81. C. Vit, C. Loot, J. A. Escudero, A. Nivina, D. Mazel, in *Horizontal Gene Transfer: Methods and Protocols*, F. de la Cruz, Ed. (Springer US, 2020), pp. 189–208.
82. C. Vit, E. Richard, F. Fournes, C. Whiteway, X. Eyer, D. Lapaillerie, V. Parissi, D. Mazel, C. Loot, Cassette recruitment in the chromosomal integron of *Vibrio cholerae*. *Nucleic Acids Res.* **49**, 5654–5670 (2021).

Acknowledgments: We are thankful to S. Petrovski for bacterial strains and plasmids and to M. Ghaly and I. Paulsen for comments on earlier versions of the manuscript. **Funding:** This work was supported by the Australian Research Council Discovery Project DP200101874 (to M.R.G. and S.G.T.). **Author contributions:** Conceptualization: T.M.G., S.G.T., and M.R.G. Methodology: T.M.G., V.R., and Q.Q. Investigation: T.M.G., V.R., and Q.Q. Visualization: T.M.G. and A.P. Funding acquisition: M.R.G. and S.G.T. Writing (original draft): T.M.G. Writing (review and editing): T.M.G., S.G.T., Q.Q., A.P., V.R., and M.R.G. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 21 April 2022
Accepted 19 October 2022
Published 16 November 2022
10.1126/sciadv.abq6376