# 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] and Michael R. Gillings[1,2]

*Corresponding author: timothy.ghaly@mq.edu.au

[1]School of Natural Sciences, Macquarie University, New South Wales, 2109

[2]ARC Centre of Excellence in Synthetic Biology, Macquarie University, New South Wales, 2109

Running title: Integrons in Archaea

Keywords: Horizontal gene transfer; prokaryotes; metagenome-assembled genomes; evolution; site-specific DNA recombination

## Summary

Horizontal gene transfer between different domains of life is increasingly being recognised as an important driver of evolution, with the potential to provide the recipient with new gene functionality and assist niche adaptation[1-3]. However, the molecular mechanisms underlying the integration of exogenous genes from foreign domains are mostly unknown. Integrons are a family of genetic elements that facilitate this process within Bacteria via site-specific DNA recombination[4-7]. Integrons, however, have not been reported outside Bacteria, and thus their potential role in cross-domain gene transfer has not been investigated. Here we show that integrons are also present among diverse phyla within the domain Archaea. Further, we provide experimental evidence that integron-mediated recombination can facilitate the recruitment of archaeal genes by bacteria. Our findings establish a new mechanism that can facilitate horizontal gene transfer between the two domains of prokaryotes, which has important implications for prokaryotic evolution in both clinical and environmental contexts.

## Main

Horizontal gene transfer between different domains of life can be a major driver in species evolution[8]. There are now numerous examples of genes that have been transferred between Archaea, Bacteria and Eukarya[3,9-13]. Among the consequences of such gene transfers are the gain of novel biochemical functions and the ability to colonise specific environmental niches[1-3]. However, the molecular mechanisms for most of these transfer events are unknown.

Integrons are genetic elements known to facilitate horizontal gene transfer within Bacteria[4-7]. Integrons can capture exogenous genes, known as gene cassettes, by site-specific recombination. Gene cassette capture is mediated by an integron integrase (IntI), which catalyses 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 200+ sequential cassettes[4,6]. Integrons are mostly known for their role in driving the global antibiotic resistance crisis by disseminating diverse resistance determinants among bacterial pathogens[14,15]. However, it is now clear that integrons play a much broader role in bacterial evolution and niche adaptation[16]. The functions encoded by integron gene cassettes are extraordinarily diverse and extend far beyond those of clinical relevance[7,17,18].

To date, integrons have only been found within bacterial genomes, where they have been detected within diverse phyla[19]. However, gene cassette amplicon sequencing has yielded cassette-encoded proteins that share homology with archaeal proteins[20,21]. Without broader genomic context, however, the taxonomic residence of such gene cassettes is unknown.

Here, we screened all publicly available archaeal genomes to show for the first time that integrons are not limited to Bacteria, but are also present in Archaea. Archaeal integrons exhibit the same characteristics and functional components as bacterial integrons. Further, 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 potentially facilitate a cross-domain highway of gene transfer between Archaea and Bacteria, with important implications for prokaryotic evolution.

## Discovery of integrons in Archaea

Here, we report the discovery of integrons in the domain Archaea. We screened 6,718 archaeal genomes for integrons using the standard criteria applied to integron surveys in

Bacteria[19,22,23]. These include the presence of integron integrase 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 9 phyla (Fig. 1 and Supplementary Table 1). It is not surprising 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 Methods for details). Additionally, we found that ~7% of integron-bearing MAGs had at least one archaeal phylogenetic marker gene on the same contig as an integron (Supplementary Table 2), confirming these to be located on archaeal chromosomes. No integron was ever co-located with a bacterial marker gene. The markers used for this analysis consisted of a comprehensive set of 122 archaeal and 120 bacterial proteins identified as suitable for phylogenetic inference[24].

Among the 75 archaeal genomes, we detected six IntIs and 539 *attC* sites (excluding all singleton *attC*s). We found that archaeal *attC*s and IntIs are largely restricted to one clade of Archaea (Fig. 1), with some outliers, suggesting that integron diversification, for the most part, has likely occurred within one archaeal clade, with occasional horizontal movements to other archaeal phyla. In particular, integrons were significantly enriched in the phylum Asgardarchaeota ($\chi^2$ 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 3-4% of genomes from the phyla Hadarchaeota and Hydrothermoarchaeota (Fig. 1), although these comprised few available genomes (n<50). A skewed phylogenetic distribution of integrons has similarly been observed among Bacteria[19].

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[19,22].

*Genetic structure of archaeal integrons*

We found that archaeal integrons exhibit the same structure and functional components as bacterial integron cassette arrays (Extended Data Fig. 1). That is, tandem arrays of short open reading frames (ORFs), 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[25-31]. We also note that archaeal IntIs exhibit the defining characteristics of bacterial IntIs, being tyrosine recombinases that possess a unique IntI-specific additional domain surrounding the patch III motif region necessary for integron-mediated recombination[32]. We found examples of 'complete' integrons, these being cassette arrays adjacent to a detectable *intI* gene (Extended Data Fig. 1). 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 (Extended Data Fig. 2a,b), and exhibited the same canonical insertion point as all known bacterial *attI*s (Extended Data Fig. 2c).

Most archaeal integrons that we identified were CALINs (clusters of *attC*s lacking integron integrases; Supplementary Table 3). This is not surprising given the fragmented nature of MAGs, and the high prevalence of CALINs also found in bacterial genomes. Indeed, among Bacteria, CALINs are more abundant than complete integrons that possess an *intI* gene, and exhibit a much wider taxonomic distribution[19]. Two In0 elements were also detected among Archaea. These are integrons that have an *intI* gene without an adjacent *attC* site (Extended Data Fig. 1). 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 (over 18 contigs) within a single MAG (Supplementary Table 1). The number of *attC*s within a single MAG ranged from 2 to 107, with an average of 7 *attC*s.

**Platforms for cross-domain gene transfer**

Archaeal gene cassettes with *attC*s from diverse phyla can be recognised and recruited by Bacteria (Fig. 2). We demonstrate that cassette insertion (*attC* x *attI* recombination) can occur following the conjugation of circular DNA molecules with archaeal *attC*s into an *Escherichia coli* recipient harbouring a bacterial class 1 integron (Fig. 2a). Insertion events were confirmed with Sanger sequencing of the PCR-amplified *attC*/*attI* recombination junctions (Fig. 2a, Extended Data Fig. 3). 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, Extended Data Table 1). We observed an average recombination frequency of $2.5x10^{-1}$ between *attI1* and $attC_{aadA7}$. Comparable frequencies (ranging from $1.9x10^{-4}$ – $3.2x10^{-1}$, with an average of $5.1x10^{-2}$) were observed for eight out of nine archaeal *attC*s (Kruskal-Wallis test, p=0.488), which were selected from multiple archaeal phyla. Further, we confirmed that cassette recruitment was mediated by IntI1 activity, since no *attC* x *attI* recombination events were detected when *intI1* was absent or when its expression was suppressed (Extended Data Table 1). We therefore show that integron-mediated gene transfer can occur between the two domains of prokaryotes.

Importantly, 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 due to their association with diverse mobile genetic elements, facilitating their spread into at least 100 bacterial species[7]. They collectively carry more than 130 different resistance genes[14], most of which are of unknown taxonomic origin[22]. 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 95-100% nucleotide identity to *attC*s within sequenced bacterial integrons (Supplementary Table 4). The archaeal *attC*s were from three phyla: Nanoarchaeota, Thermoproteota and Hadarchaeota. The homologous *attC*s in Bacteria were found in 26 genomes from 5 phyla: Proteobacteria, Spirochaetota, Myxococcota, Nitrospirota and Desulfobacterota. One of these *attC* sites was associated with a class 1 integron gene cassette, encoding an NADPH-dependent oxidoreductase found on five different Enterobacteriaceae plasmids (Supplementary Table 4). In Archaea, however, this *attC* site was part of a cassette that encoded a ligand-binding protein of unknown function. Nevertheless, since strong *attC* homology is a characteristic of cassettes that share the same taxonomic origin[22,33,34], it is possible that some clinically relevant gene cassettes now found on class 1 integrons might be of archaeal origin.

## Diversity of archaeal integrons

### *Diversity of integron integrases*

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 (Extended Data Fig. 4). We found that archaeal IntIs form their own monophyletic clade separate from known bacterial IntIs[22]. This strongly suggests that IntI radiation has occurred within Archaea and that their distribution, at least among the archaeal genomes in our dataset, is not likely to be the result of multiple IntI acquisitions 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). Intriguingly, these two IntIs are phylogenetically distinct from ‘typical’ Spirochaetota IntIs, which are generally in reverse orientation[5,35]. Further, the two Spirochaetota that harboured 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[36]. Thus, these atypical Spirochaetota IntIs might have been horizontally acquired from Archaea that share the same extreme environments.

### *Diversity of attC recombination sites*

Archaeal *attC*s exhibit broad sequence and structural diversity (Fig. 4a). We find that some archaeal phyla possess *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[7,22]. 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[19,34]. 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[22,33]. Thus, the ecological and evolutionary forces that promote and/or constrain *attC* diversity[7] are likely to be similar for both Archaea and Bacteria.

There is a clear overlap in the sequence and structural diversity of *attC*s from Archaea and Bacteria (Fig. 4a). This provides additional evidence that the mechanistic overlap between archaeal and bacterial *attC*s is extensive, and thus, cross-domain transfer of cassettes could be common in shared environments. 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[37]). The broad distribution of integrons among the two domains suggest 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 COG category (Extended Data Fig. 5). In contrast, 47.4% of all proteins from the 75 integron-bearing archaeal genomes could be assigned a known COG category. This underrepresentation ($\chi^2$ test, $p < 0.00001$) of known COGs among cassette proteins has previously been reported for bacterial integrons[4,5,33]. To gain further insight into possible cassette functions, eggNOG 5.0[38] and Pfam[39] database searches were performed, assigning putative functions to 228 (41.5%) of the archaeal cassette-encoded proteins. Out 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 (Supplementary Table 5). These are the functions most commonly reported for gene cassettes in Bacteria[5,7,33,34,40]. TA gene cassettes are particularly common in bacterial integrons, where they can stabilise very large cassette arrays[41,42]. 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[43,44].

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%, $\chi^2$ 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[33], and are hypothesised to help facilitate interactions with their broader environment[7].

Indeed, 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 occurs between archaeal phyla co-located in the same environment.

## Conclusion

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. 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 is widespread and plays an important role in prokaryotic evolution.

## Methods

### *Data acquisition and quality filtering*

All available archaeal genomes were downloaded from the NCBI Assembly Database (n=8,160; last accessed 2021-Oct-5). Of these, ~ 95% were metagenome-assembled genomes (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[45] 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[46]; '*clade-markers*', which finds contaminating contigs using a database of 855,764 clade-specific prokaryotic marker genes (MetaPhlAn2 database[47]); '*tetra-freq*', which employs principal component analysis (PCA) to identify contaminating contigs with outlier tetra-nucleotide 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[48], which uses single-copy lineage-specific marker 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 ≥50% and contamination <10%[24,45,49-52]. In this context, however, we were more concerned with the level of contamination than completeness, and thus removed all genomes with an estimated contamination ≥5%. The completeness of the remaining genomes ranged from 15% – 100%, with a median of 81%. The estimated contamination ranged from 0% – 4.98%, with a median of 0.93%.

Archaeal genomes were assigned taxonomic classifications based on the Genome Taxonomy Database (GTDB)[49-51] using GTDB-Tk v1.6.0[53] with release 06-RS202 of the GTDB. We employed the *classify_wf* command with default settings. This workflow identifies and aligns 120 bacterial and 122 archaeal phylogenetic marker genes[24]. 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 6,718 archaeal genomes retained for further analysis.

To infer the phylum-level phylogeny of Archaea, the highest quality representative genome from each phylum was selected based on its genome quality score (defined by Parks et al.[24] as the estimated completeness of a genome minus five times its estimated contamination). From representative genomes, a concatenated multiple protein sequence alignment of the 122 archaeal phylogenetic markers was generated using GTDB-Tk v1.6.0[53]. A maximum-likelihood tree was generated from the alignment using IQ-TREE v1.6.12[54] with the best-suited protein model as determined by ModelFinder[55] and 1,000 bootstrap replicates [parameters: -m MFP -bb 1000].

*Integron detection*

Due to faster processing speeds of large datasets, we initially screened all filtered genomes for *attC* recombination sites using *attC*-screening.sh[37] (https://github.com/timghaly/integron-filtering) with default parameters. This script uses the HattCI[56] + Infernal[57] pipeline (first described by Pereira *et al.*[23]) 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[19] [parameters: --local-max --cpu 24 --gbk], which searches for integron integrases and gene cassette arrays. Only IntIs, *attCs* 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[53]. These consisted of 122 archaeal and 120 bacterial proteins identified as suitable phylogenetic markers[24]. We found a total of nine prokaryotic marker genes among seven integron-bearing contigs. All nine markers were confirmed to be archaeal via a BLASTP search of the NCBI nr database (Supplementary Table 2).

*Analysis of integron integrases, attC sites and cassette-encoded proteins*

IntegronFinder identifies IntIs using the overlap of two protein hidden Markov model (HMM) 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[32]. Identified archaeal IntIs, with matches to both protein profiles, were placed in a phylogeny alongside a set of previously identified bacterial IntIs[22]. IntIs shorter than 200 amino acids or those that did not span the complete IntI-specific domain were removed from phylogenetic analysis. The remaining IntIs were aligned using MAFFT v7.271[58] [parameters: --localpair --maxiterate 1000] and trimmed using trimAl v1.2rev59 [parameters: -automated1]. A maximum-likelihood tree was generated from the alignment using IQ-TREE v1.6.12[54] with the best-suited protein model as determined by ModelFinder[55] and 1,000 bootstrap replicates [parameters: -m MFP -bb 1000].

The sequence and structural diversity of *attC*s was assessed using RNAclust v1.3[59] as previously described[22]. RNAclust uses LocARNA[60,61] 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 WPGMA analysis. All archaeal *attC*s along with a set of previously identified *attC*s from representative bacterial taxa[22] were clustered using RNAclust's default parameters.

Cassette-encoded proteins identified by IntegronFinder were functionally annotated using InterProScan v5.44-79.0[62], with default parameters against the Pfam[39] database, and eggNOG-mapper v2.0.1b[63,64], executed in DIAMOND[65] mode against the eggNOG 5.0 database[38]. To identify cassettes that encode transmembrane and secreted proteins, we searched protein sequences for prokaryotic signal peptides using SignalP 5.0[66] with default parameters. The correlation analysis of cassette functions was performed as described in Penesyan et al[67]. Briefly, Pearson's correlations, based on co-occurrences between Pfam functions, specific environments and archaeal phyla were calculated using the Hmisc v4.5-0 R package[68]. The network was generated from all positive correlations with p-values <0.05 using the ForceAtlas2 layout algorithm[69] within the Gephi software[70]. Specific correlations and the description of Pfam functions are listed in Supplementary Table 6.

*Bacterial strains and plasmids for attC recombination assays*

The bacterial strains and plasmids used in this study are listed in Supplementary Table 7. 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 acid (DAP) to culture the auxotrophic *E. coli* WM3064 λpir strain[71].

*Construction of attC donor strains*

Nine archaeal *attC*s, selected from diverse archaeal phyla (Supplementary Table 8) 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 mobilisable suicide vector pJP5603[72,73]. 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 Supplementary Table 9.

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[74] that carried the *intI1* gene and the *attI1* recombination site residing on the pBAD24[75] and pACYC184[76] backbones, respectively. The *intI1* gene of the R388 plasmid[77] was PCR amplified using the primer pair *intI1_Eco*RI-F/*intI1_Hin*dIII-R (Supplementary Table 9). 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, United States). 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 co-transformed with pBAD::*intI1* and pACYC184::*attI1* to generate the *E. coli* UB5201 + pBAD::*intI1* + pACYC184::*attI1* recipient strain for *attC* x *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 x 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* x *attI* suicide conjugation methods[25,29,31,78,79] 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, as well as on LB agar containing Cb. This method 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 co-integration (Fig. 2a). The recombination frequency was determined as the ratio of the colony forming units (CFU) for Km-resistant recombinants to the CFU for the total number of Cb-resistant recipients after two 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 co-integrates, 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 (Extended Data Fig. 3). Sanger sequencing of PCR products was performed for four recombinant colonies per conjugation set.

## Data availability

All genome sequences were downloaded from the NCBI Assembly Database (https://www.ncbi.nlm.nih.gov/assembly; last accessed 2021-Oct-5).

## Code availability

All code and software used in this study are described within the manuscript.

## References

1 Husnik, F. & McCutcheon, J. P. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* **16**, 67-79, doi:10.1038/nrmicro.2017.137 (2018).

2 Metcalf, J. A., Funkhouser-Jones, L. J., Brileya, K., Reysenbach, A.-L. & Bordenstein, S. R. Antibacterial gene transfer across the tree of life. *eLife* **3**, e04266 (2014).

3 Schönknecht, G. *et al.* Gene transfer from Bacteria and Archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207-1210, doi:doi:10.1126/science.1231707 (2013).

4 Mazel, D. Integrons: agents of bacterial evolution. *Nature Reviews Microbiology* **4**, 608-620, doi:10.1038/nrmicro1462 (2006).

5 Boucher, Y., Labbate, M., Koenig, J. E. & Stokes, H. W. Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends in Microbiology* **15**, 301-309, doi:https://doi.org/10.1016/j.tim.2007.05.004 (2007).

6 Gillings, M. R. Integrons: past, present, and future. *Microbiology and Molecular Biology Reviews* **78**, 257-277 (2014).

7 Ghaly, T. M. *et al.* The natural history of integrons. *Microorganisms* **9**, 2212 (2021).

8 Bruto, M. *et al.* in *Evolutionary Biology: Exobiology and Evolutionary Mechanisms* (ed Pierre Pontarotti) 165-179 (Springer Berlin Heidelberg, 2013).

9 Sutherland, K. M., Ward, L. M., Colombero, C.-R. & Johnston, D. T. Inter-domain horizontal gene transfer of nickel-binding superoxide dismutase. *Geobiology* **19**, 450-459, doi:https://doi.org/10.1111/gbi.12448 (2021).

10 Frigaard, N.-U., Martinez, A., Mincer, T. J. & DeLong, E. F. Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**, 847-850, doi:10.1038/nature04435 (2006).

11 Dunning Hotopp, J. C. Horizontal gene transfer between bacteria and animals. *Trends in Genetics* **27**, 157-163, doi:https://doi.org/10.1016/j.tig.2011.01.005 (2011).

12 Bock, R. The give-and-take of DNA: horizontal gene transfer in plants. *Trends in Plant Science* **15**, 11-22, doi:https://doi.org/10.1016/j.tplants.2009.10.001 (2010).

13 Nelson, K. E. *et al.* Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323-329 (1999).

14 Partridge, S. R., Tsafnat, G., Coiera, E. & Iredell, J. R. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiology Reviews* **33**, 757-784, doi:10.1111/j.1574-6976.2009.00175.x (2009).

15 Zhu, Y.-G. *et al.* Continental-scale pollution of estuaries with antibiotic resistance genes. *Nature Microbiology* **2**, 16270, doi:10.1038/nmicrobiol.2016.270 (2017).

16 Escudero, J. A., Loot, C., Nivina, A. & Mazel, D. The integron: adaptation on demand. *Microbiology Spectrum* **3**, 3.2.10, doi:doi:10.1128/microbiolspec.MDNA3-0019-2014 (2015).

17 Ghaly, T. M., Geoghegan, J. L., Tetu, S. G. & Gillings, M. R. The peril and promise of integrons: beyond antibiotic resistance. *Trends in Microbiology* **28**, 455-464, doi:https://doi.org/10.1016/j.tim.2019.12.002 (2020).

18 Ghaly, T. M., Geoghegan, J. L., Alroy, J. & Gillings, M. R. High diversity and rapid spatial turnover of integron gene cassettes in soil. *Environmental Microbiology* **21**, 1567-1574, doi:https://doi.org/10.1111/1462-2920.14551 (2019).

19 Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Research* **44**, 4539-4550, doi:10.1093/nar/gkw319 (2016).

20 Elsaied, H., Stokes, H. W., Yoshioka, H., Mitani, Y. & Maruyama, A. Novel integrons and gene cassettes from a Cascadian submarine gas-hydrate-bearing core. *FEMS Microbiology Ecology* **87**, 343-356, doi:10.1111/1574-6941.12227 (2014).

21 Koenig, J. E. *et al.* Integron gene cassettes and degradation of compounds associated with industrial waste: The case of the Sydney Tar Ponds. *PLoS One* **4**, e5276, doi:10.1371/journal.pone.0005276 (2009).

22 Ghaly, T. M., Tetu, S. G. & Gillings, M. R. Predicting the taxonomic and environmental sources of integron gene cassettes using structural and sequence homology of *attC* sites. *Communications Biology* **4**, 946, doi:10.1038/s42003-021-02489-0 (2021).

23 Pereira, M. B. *et al.* A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* **21**, 495, doi:10.1186/s12864-020-06830-5 (2020).

24 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542, doi:10.1038/s41564-017-0012-7 (2017).

25 Bouvier, M., Demarre, G. & Mazel, D. Integron cassette insertion: a recombination process involving a folded single strand substrate. *The EMBO Journal* **24**, 4356-4367, doi:https://doi.org/10.1038/sj.emboj.7600898 (2005).

26 MacDonald, D., Demarre, G., Bouvier, M., Mazel, D. & Gopaul, D. N. Structural basis for broad DNA-specificity in integron recombination. *Nature* **440**, 1157-1162, doi:10.1038/nature04643 (2006).

27 Bouvier, M., Ducos-Galand, M., Loot, C., Bikard, D. & Mazel, D. Structural features of single-stranded integron cassette *attC* sites and their role in strand selection. *PLoS Genetics* **5**, e1000632 (2009).

28 Demarre, G., Frumerie, C., Gopaul, D. N. & Mazel, D. Identification of key structural determinants of the IntI1 integron integrase that influence *attC*×*attI1* recombination efficiency. *Nucleic Acids Research* **35**, 6475-6489, doi:10.1093/nar/gkm709 (2007).

29 Nivina, A., Escudero, J. A., Vit, C., Mazel, D. & Loot, C. Efficiency of integron cassette insertion in correct orientation is ensured by the interplay of the three unpaired features of *attC* recombination sites. *Nucleic Acids Research* **44**, 7792-7803, doi:10.1093/nar/gkw646 (2016).

30 Mukhortava, A. *et al.* Structural heterogeneity of *attC* integron recombination sites revealed by optical tweezers. *Nucleic Acids Research* **47**, 1861-1870, doi:10.1093/nar/gky1258 (2018).

31 Nivina, A. *et al.* Structure-specific DNA recombination sites: Design, validation, and machine learning–based refinement. *Science Advances* **6**, eaay2922 (2020).

32 Messier, N. & Roy, P. H. Integron integrases possess a unique additional domain necessary for activity. *Journal of Bacteriology* **183**, 6699-6706 (2001).

33 Rowe-Magnus, D. A., Guerout, A.-M., Biskri, L., Bouige, P. & Mazel, D. Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Research* **13**, 428-442 (2003).

34 Rowe-Magnus, D. A. *et al.* The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proceedings of the National Academy of Sciences* **98**, 652-657 (2001).

35 Wu, Y.-W., Doak, T. G. & Ye, Y. The gain and loss of chromosomal integron systems in the *Treponema* species. *BMC Evolutionary Biology* **13**, 16, doi:10.1186/1471-2148-13-16 (2013).

36 Rampelotto, P. H. Extremophiles and extreme environments. *Life* **3**, 482-485 (2013).

37 Ghaly, T. M. *et al.* Methods for the targeted sequencing and analysis of integrons and their gene cassettes from complex microbial communities. *bioRxiv*, 2021.2009.2008.459516, doi:10.1101/2021.09.08.459516 (2021).

38 Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309-D314 (2019).

39 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427-D432 (2019).

40 Cambray, G., Guerout, A.-M. & Mazel, D. Integrons. *Annual Review of Genetics* **44**, 141-166 (2010).

41 Iqbal, N., Guérout, A.-M., Krin, E., Le Roux, F. & Mazel, D. Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin systems substantiates their role in stabilizing the superintegron. *Journal of Bacteriology* **197**, 2150-2159 (2015).

42 Szekeres, S., Dauti, M., Wilde, C., Mazel, D. & Rowe-Magnus, D. A. Chromosomal toxin–antitoxin loci can diminish large-scale genome reductions in the absence of selection. *Molecular Microbiology* **63**, 1588-1605 (2007).

43 Wilbaux, M., Mine, N., Guérout, A.-M., Mazel, D. & Van Melderen, L. Functional interactions between coexisting toxin-antitoxin systems of the *ccd* family in *Escherichia coli* O157: H7. *Journal of Bacteriology* **189**, 2712-2719 (2007).

44 Guérout, A.-M. *et al.* Characterization of the *phd-doc* and *ccd* toxin-antitoxin cassettes from *Vibrio* superintegrons. *Journal of Bacteriology* **195**, 2270-2283 (2013).

45 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505-510, doi:10.1038/s41586-019-1058-x (2019).

46 Wu, D., Jospin, G. & Eisen, J. A. 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).

47 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**, 902-903 (2015).

48 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

49 Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996-1004, doi:10.1038/nbt.4229 (2018).

50 Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **38**, 1079-1086, doi:10.1038/s41587-020-0501-8 (2020).

51 Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nature Microbiology* **6**, 946-959, doi:10.1038/s41564-021-00918-8 (2021).

52 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725-731, doi:10.1038/nbt.3893 (2017).

53 Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927, doi:10.1093/bioinformatics/btz848 (2019).

54 Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268-274 (2015).

55 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**, 587-589 (2017).

56 Pereira, M. B., Wallroth, M., Kristiansson, E. & Axelson-Fisk, M. HattCI: fast and accurate *attC* site identification using hidden Markov models. *Journal of Computational Biology* **23**, 891-902 (2016).

57 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

58 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).

59 Engelhardt, J., Heyne, S., Will, S. & Reiche, K. *RNAclust Documentation*, <http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAclust/manual.pdf> (2010).

60 Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* **18**, 900-914 (2012).

61 Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F. & Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology* **3**, e65 (2007).

62 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

63 Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115-2122 (2017).

64 Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, doi:10.1093/molbev/msab293 (2021).

65 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60 (2015).

66 Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**, 420-423, doi:10.1038/s41587-019-0036-z (2019).

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

68 Harrell, F. E. & Dupont, C. Hmisc: harrell miscellaneous. R package version 4.5-0. https://CRAN.R-project.org/package=Hmisc. (2021).

69 Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS One* **9**, e98679 (2014).

70 Bastian, M., Heymann, S. & Jacomy, M. in *Third International AAAI Conference on Weblogs and Social Media.*

71 Dehio, C. & Meyer, M. 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*. *Journal of Bacteriology* **179**, 538-540, doi:doi:10.1128/jb.179.2.538-540.1997 (1997).

72 Penfold, R. J. & Pemberton, J. M. An improved suicide vector for construction of chromosomal insertion mutations in bacteria. *Gene* **118**, 145-146, doi:https://doi.org/10.1016/0378-1119(92)90263-O (1992).

73 Riedel, T., Rohlfs, M., Buchholz, I., Wagner-Döbler, I. & Reck, M. Complete sequence of the suicide vector pJP5603. *Plasmid* **69**, 104-107, doi:10.1016/j.plasmid.2012.07.005 (2013).

74 Sanchez, J., Bennett, P. M. & Richmond, M. H. Expression of *elt*-B, the gene encoding the B subunit of the heat-labile enterotoxin of *Escherichia coli*, when cloned in pACYC184. *FEMS Microbiology Letters* **14**, 1-5, doi:10.1111/j.1574-6968.1982.tb08623.x (1982).

75 Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *Journal of Bacteriology* **177**, 4121-4130, doi:10.1128/jb.177.14.4121-4130.1995 (1995).

76 Rose, R. E. The nucleotide sequence of pACYC184. *Nucleic Acids Research* **16**, 355, doi:10.1093/nar/16.1.355 (1988).

77 Avila, P. & de la Cruz, F. Physical and genetic map of the IncW plasmid R388. *Plasmid* **20**, 155-157, doi:10.1016/0147-619x(88)90019-4 (1988).

78 Vit, C., Loot, C., Escudero, J. A., Nivina, A. & Mazel, D. in *Horizontal Gene Transfer: Methods and Protocols* (ed Fernando de la Cruz) 189-208 (Springer US, 2020).

79 Vit, C. *et al.* Cassette recruitment in the chromosomal Integron of *Vibrio cholerae*. *Nucleic Acids Research* **49**, 5654-5670, doi:10.1093/nar/gkab412 (2021).

80 Moura, A. *et al.* INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics* **25**, 1096-1098, doi:10.1093/bioinformatics/btp105 (2009).

## Acknowledgements

The authors would like to thank Ian Paulsen for comments on earlier versions of the manuscript. This work was funded by the Australian Research Council (Discovery Project DP200101874).

## Author contributions

TMG contributed to the conception of the study, performed all data analyses, wrote the original draft of the paper, and contributed to the final editing of the paper. SGT contributed to the conception of the study and the final editing of the paper. AP performed the correlation analysis of cassette functions, and contributed to the final editing of the paper. QQ was involved with the design and implementation of the experimental work, and contributed to the final editing of the paper. VR was involved with the design and implementation of the experimental work, and contributed to the final editing of the paper. MRG contributed to the conception of the study and the final editing and revision of the paper. All authors contributed to the article and approved the final submitted version.

## Competing interests

The authors declare no competing interests.

## Materials & Correspondence

Correspondence and material requests should be addressed to TMG (timothy.ghaly@mq.edu.au).

## Figures

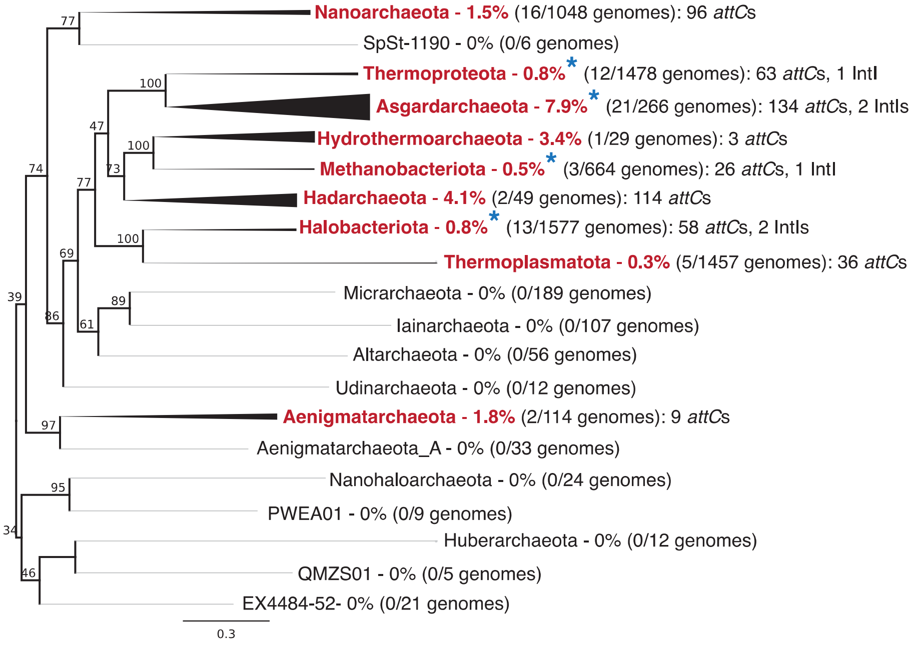

Nanoarchaeota - 1.5% (16/1048 genomes): 96 attCs
SpSt-1190 - 0% (0/6 genomes)
Thermoproteota - 0.8%* (12/1478 genomes): 63 attCs, 1 IntI
Asgardarchaeota - 7.9%* (21/266 genomes): 134 attCs, 2 IntIs
Hydrothermoarchaeota - 3.4% (1/29 genomes): 3 attCs
Methanobacteriota - 0.5%* (3/664 genomes): 26 attCs, 1 IntI
Hadarchaeota - 4.1% (2/49 genomes): 114 attCs
Halobacteriota - 0.8%* (13/1577 genomes): 58 attCs, 2 IntIs
Thermoplasmatota - 0.3% (5/1457 genomes): 36 attCs
Micrarchaeota - 0% (0/189 genomes)
Iainarchaeota - 0% (0/107 genomes)
Altarchaeota - 0% (0/56 genomes)
Udinarchaeota - 0% (0/12 genomes)
Aenigmatarchaeota - 1.8% (2/114 genomes): 9 attCs
Aenigmatarchaeota_A - 0% (0/33 genomes)
Nanohaloarchaeota - 0% (0/24 genomes)
PWEA01 - 0% (0/9 genomes)
Huberarchaeota - 0% (0/12 genomes)
QMZS01 - 0% (0/5 genomes)
EX4484-52- 0% (0/21 genomes)
77
100
47
100
73
77
100
74
69
39
86
89
61
97
95
34
46
0.3

**Fig. 1: Phylogenetic distribution of integrons among Archaea**. Archaeal phyla found to carry integrons are labelled in red, and those found to have an integron integrase gene (*intI*) are denoted with blue asterisks. Branch thickness indicates the proportion of genomes with integrons for each phylum.

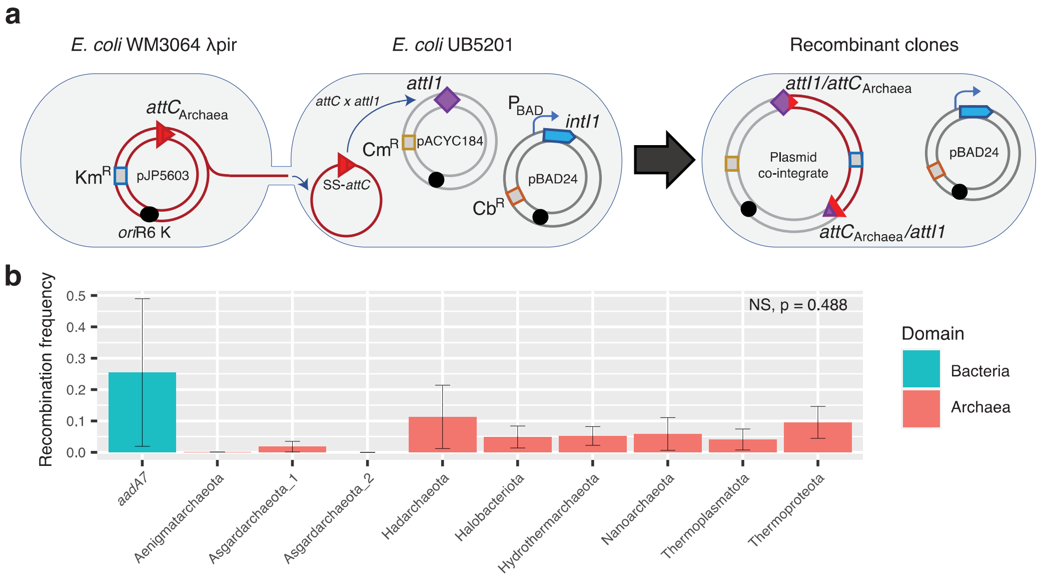


**Fig. 2: Cassette recruitment (*attC* x *attI* recombination) assays. a,** schematic outlining the experimental setup of the cassette insertion assays. The suicide vector pJP5603 with an *attC* site is delivered into the recipient *E. coli* UB5201 strain via conjugation. The recipient strain carries an *intI1* gene, expressed from the inducible P<sub>BAD</sub> promoter, and an *attI1* site, residing on the pBAD24 and pACYC184 backbones, respectively. The donor suicide vector cannot replicate within the recipient host, and thus, can only persist following *attC* x *attI* recombination to form a plasmid co-integrate. **b,** average recombination frequencies (±1 S.E.) between *attI1* and nine archaeal *attC*s (with phyla of origin labelled 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 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 Extended Data Table 1 for *attC* top strand recombination frequencies.

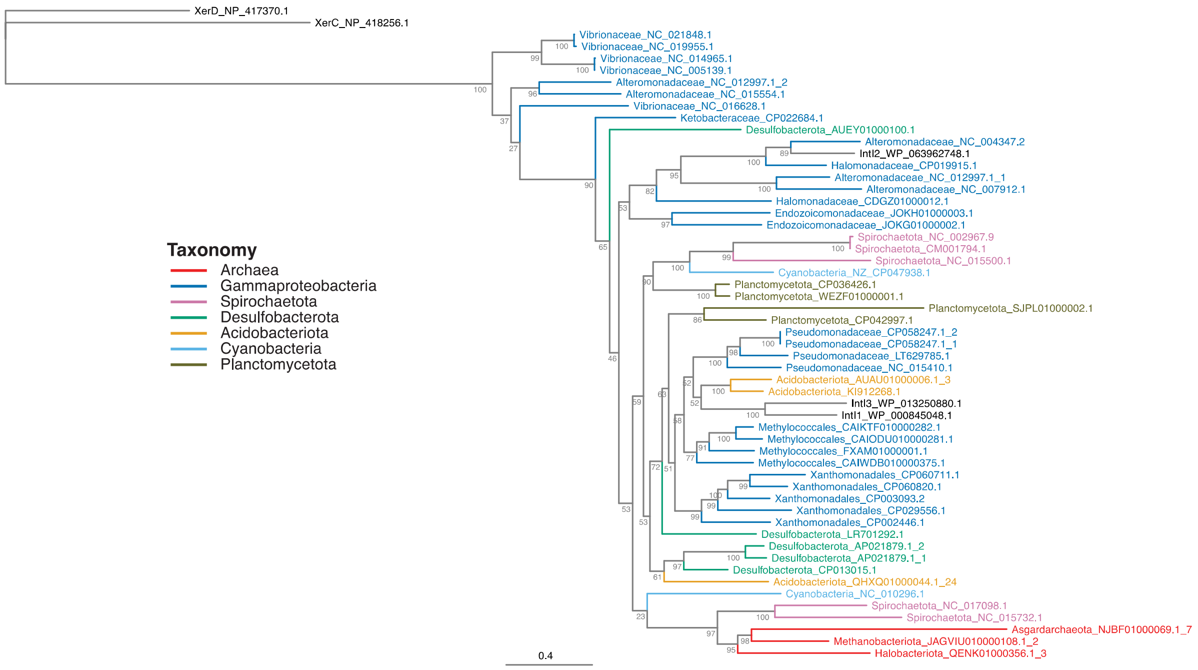

XerD_NP_417370.1
XerC_NP_418256.1
Vibrionaceae_NC_021848.1
Vibrionaceae_NC_019955.1
Vibrionaceae_NC_014965.1
Vibrionaceae_NC_005139.1
Alteromonadaceae_NC_012997.1_2
Alteromonadaceae_NC_015554.1
Vibrionaceae_NC_016628.1
Ketobacteraceae_CP022684.1
Desulfobacterota_AUEY01000100.1
Alteromonadaceae_NC_004347.2
IntI2_WP_063962748.1
Halomonadaceae_CP019915.1
Alteromonadaceae_NC_012997.1_1
Alteromonadaceae_NC_007912.1
Halomonadaceae_CDGZ01000012.1
Endozoicomonadaceae_JOKH01000003.1
Endozoicomonadaceae_JOKG01000002.1
Spirochaetota_NC_002967.9
Spirochaetota_CM001794.1
Spirochaetota_NC_015500.1
Cyanobacteria_NZ_CP047938.1
Planctomycetota_CP036426.1
Planctomycetota_WEZF01000001.1
Planctomycetota_SJPL01000002.1
Planctomycetota_CP042997.1
Pseudomonadaceae_CP058247.1_2
Pseudomonadaceae_CP058247.1_1
Pseudomonadaceae_LT629785.1
Pseudomonadaceae_NC_015410.1
Acidobacteriota_AUAU01000006.1_3
Acidobacteriota_KI912268.1
IntI3_WP_013250880.1
IntI1_WP_000845048.1
Methylococcales_CAIKTF010000282.1
Methylococcales_CAIODU010000281.1
Methylococcales_FXAM01000001.1
Methylococcales_CAIWDB010000375.1
Xanthomonadales_CP060711.1
Xanthomonadales_CP060820.1
Xanthomonadales_CP003093.2
Xanthomonadales_CP029556.1
Xanthomonadales_CP002446.1
Desulfobacterota_LR701292.1
Desulfobacterota_AP021879.1_2
Desulfobacterota_AP021879.1_1
Desulfobacterota_CP013015.1
Acidobacteriota_QHXQ01000044.1_24
Cyanobacteria_NC_010296.1
Spirochaetota_NC_017098.1
Spirochaetota_NC_015732.1
Asgardarchaeota_NJBF01000069.1_7
Methanobacteriota_JAGVIU010000108.1_2
Halobacteriota_QENK01000356.1_3
Taxonomy
Archaea
Gammaproteobacteria
Spirochaetota
Desulfobacterota
Acidobacteriota
Cyanobacteria
Planctomycetota
0.4

**Fig. 3: Phylogeny of integron integrases from Archaea and Bacteria.** To root the tree, the tyrosine recombinases XerC and XerD from *Escherichia coli* were used as outgroups. Integron integrases (IntIs) are coloured according to their taxonomy.

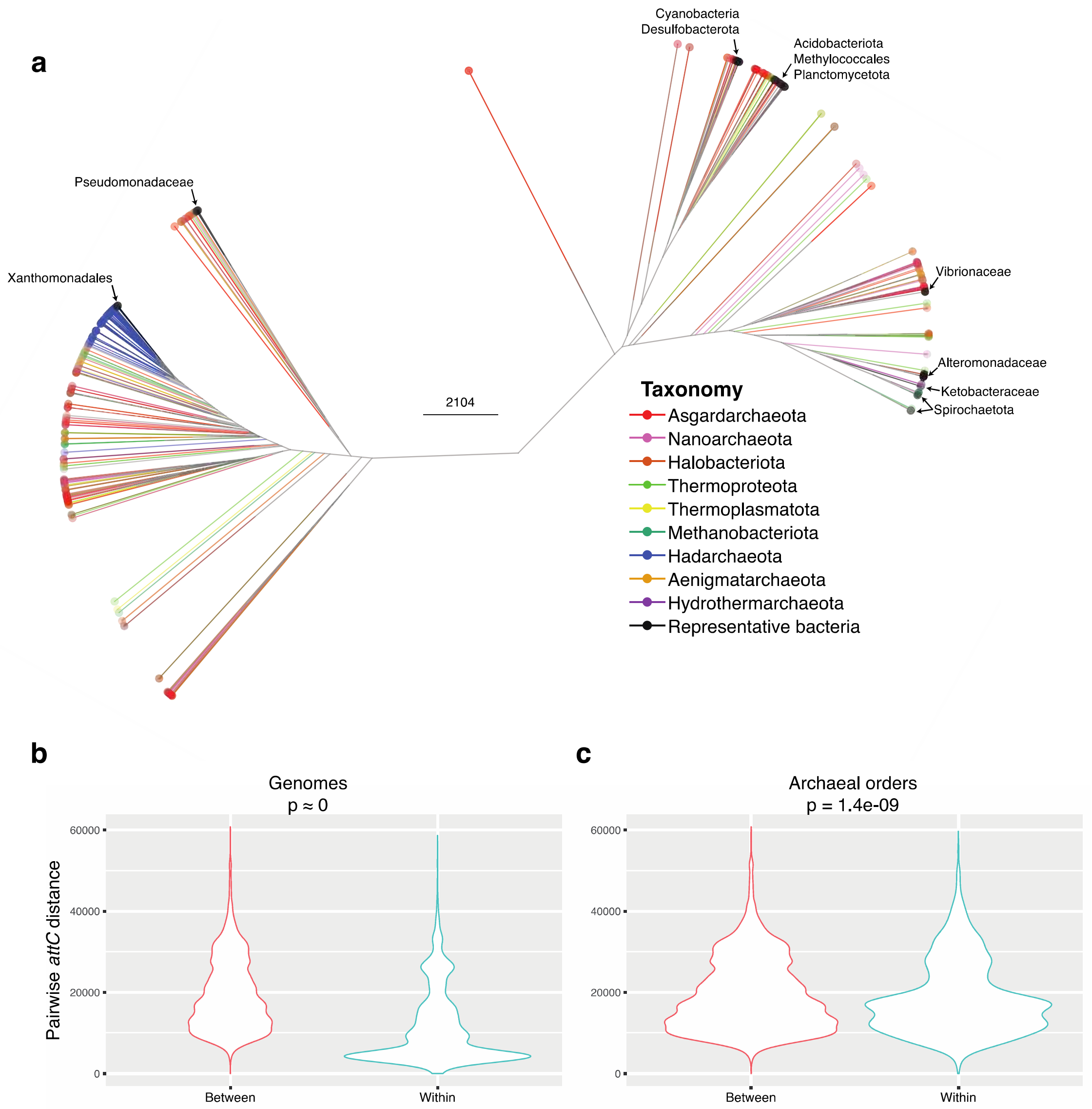
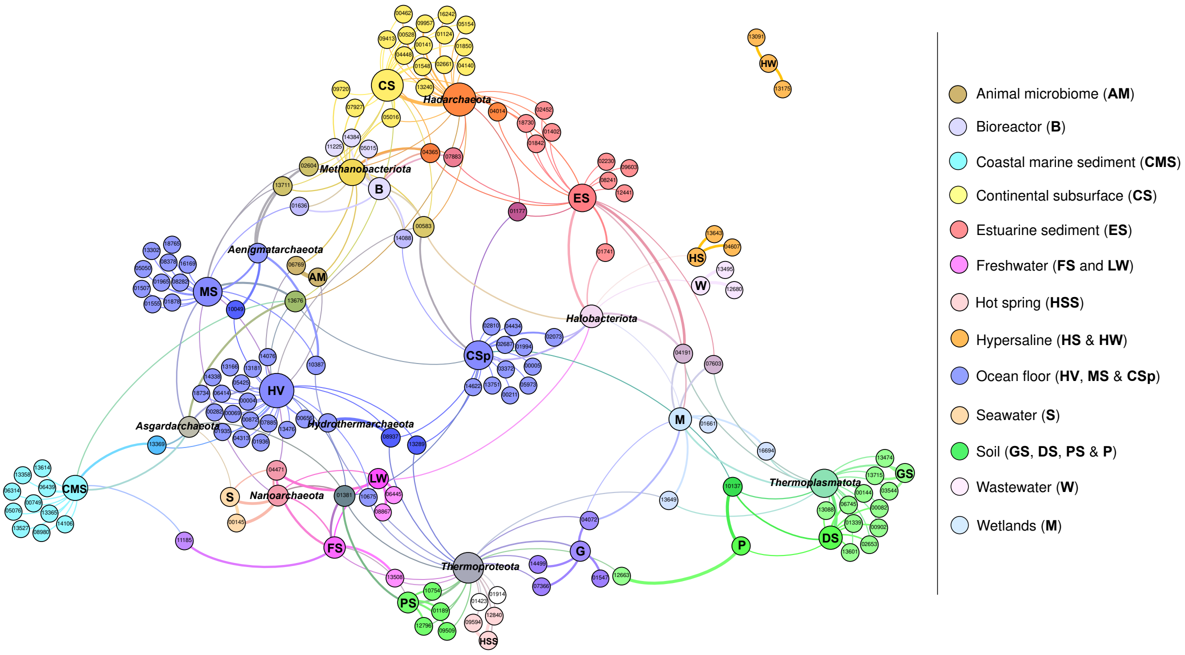


**Fig. 4: Structural and sequence diversity of archaeal *attC* recombination sites. a,** structure-based clustering of all archaeal and representative bacterial *attC*s. Branches and tips are coloured according to archaeal phylum. The taxa of bacterial *attC*s are labelled with arrows. **b**, distribution 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.



**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.

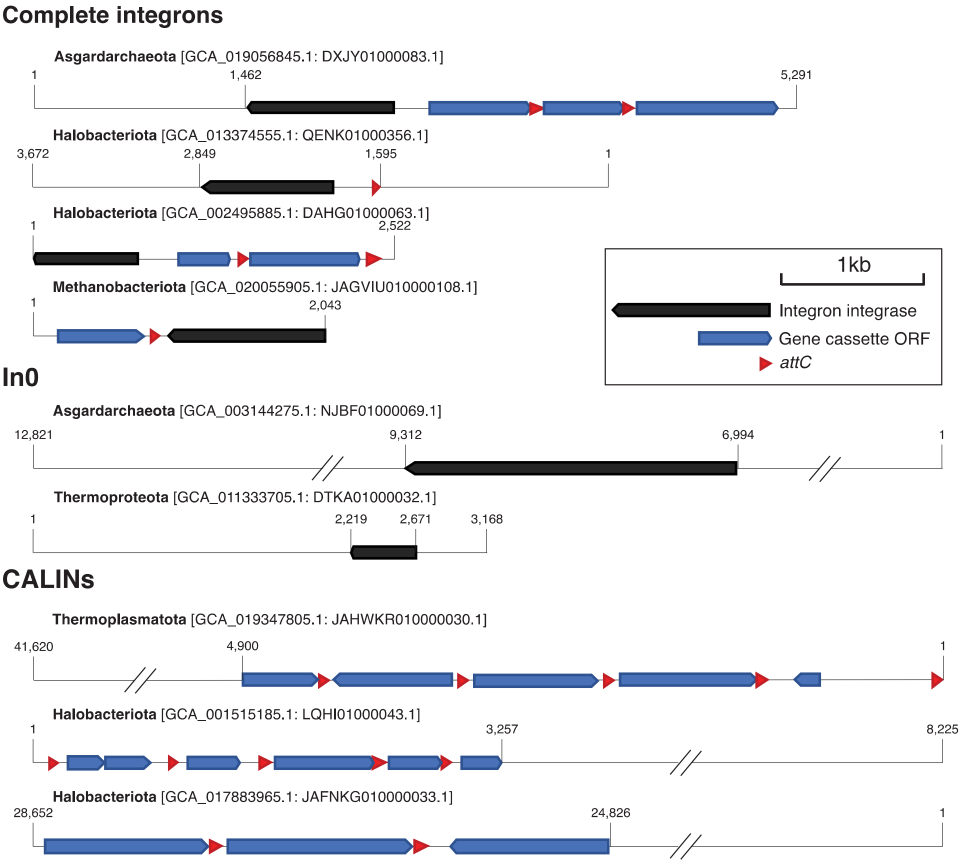
**Fig. 5: A network linking Pfam functions of archaeal integron gene cassettes with their taxonomic and environmental contexts.** The force-directed representation of the network is constructed based on co-occurrence patterns and correlations ($p < 0.05$) between Pfam functions, taxonomic groups, and specific environments from which the organisms were sampled. Nodes that represent taxonomic groups and specific environments are labelled accordingly. All other nodes denote Pfam functions and are labelled with a Pfam number preceded by ‘PF’. Specific environments are grouped into broader environment types, each of which is coloured as per the panel. Pfams directly linked to specific environment types are coloured in corresponding colours. Pfams linked to more than one environment type are coloured in overlapping colours. 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 colour denotes the overlapping colour 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 Supplementary Table 6.

# Extended data

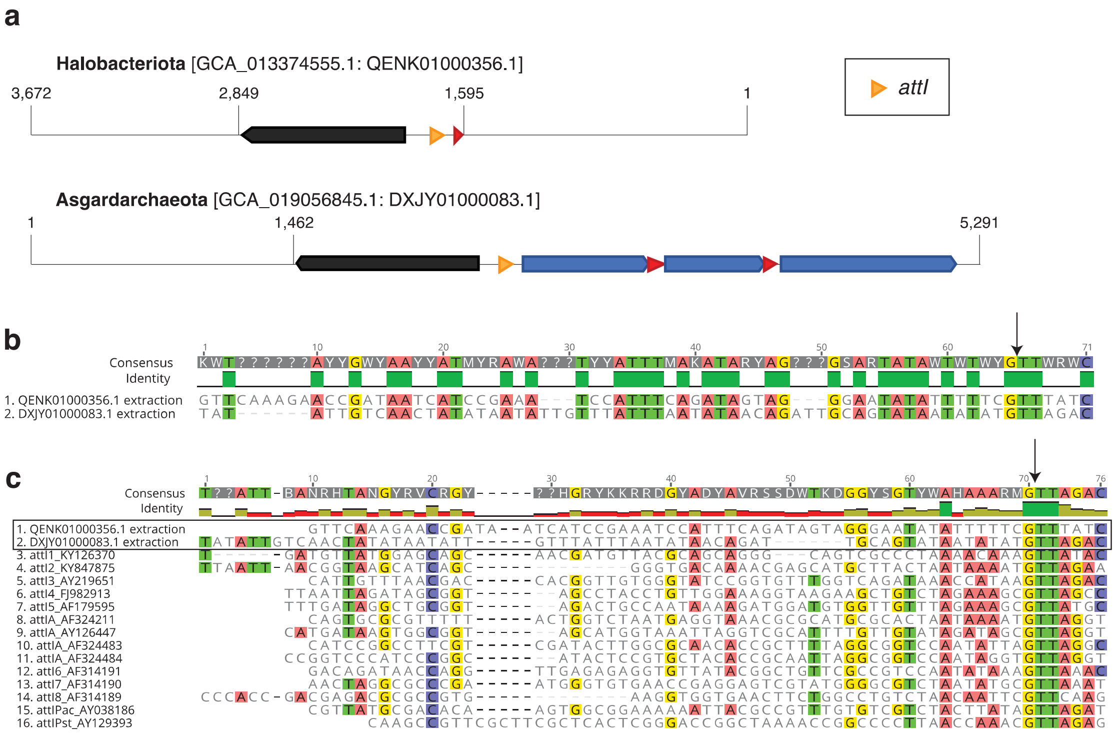
**Extended Data Table 1. Average recombination frequencies for the *attC* x *attI* suicide conjugation assays.**

| | ***attC* bottom strand (*intI1* induced[*])** | ***attC* bottom strand (*intI1* suppressed[†])** | ***attC* top strand (*intI1* induced)** | ***attC* top strand (*intI1* suppressed)** |
|---|---|---|---|---|
| *attC*$_{aadA7}$ | 2.54x$10^{-1}$ | ND[††] | 2.48x$10^{-3}$ | ND |
| *attC*$_{\mathrm{Aenigmatarchaeota}}$ | 5.46x$10^{-4}$ | ND | 8.66x$10^{-7}$ | ND |
| *attC*$_{\mathrm{Asgardarchaeota\_1}}$ | 1.79x$10^{-2}$ | ND | 6.74x$10^{-4}$ | ND |
| *attC*$_{\mathrm{Asgardarchaeota\_2}}$ | ND | ND | ND | ND |
| *attC*$_{\mathrm{Hadarchaeota}}$ | 1.13x$10^{-1}$ | ND | 1.18x$10^{-3}$ | ND |
| *attC*$_{\mathrm{Halobacteriota}}$ | 4.88x$10^{-2}$ | ND | 4.33x$10^{-4}$ | ND |
| *attC*$_{\mathrm{Hydrothermarchaeota}}$ | 5.21x$10^{-2}$ | ND | 6.92x$10^{-3}$ | ND |
| *attC*$_{\mathrm{Nanoarchaeota}}$ | 5.84x$10^{-2}$ | ND | 1.55x$10^{-3}$ | ND |
| *attC*$_{\mathrm{Thermoplasmatota}}$ | 4.08x$10^{-2}$ | ND | 2.28x$10^{-3}$ | ND |
| *attC*$_{\mathrm{Thermoproteota}}$ | 9.54x$10^{-2}$ | ND | 1.80x$10^{-3}$ | ND |

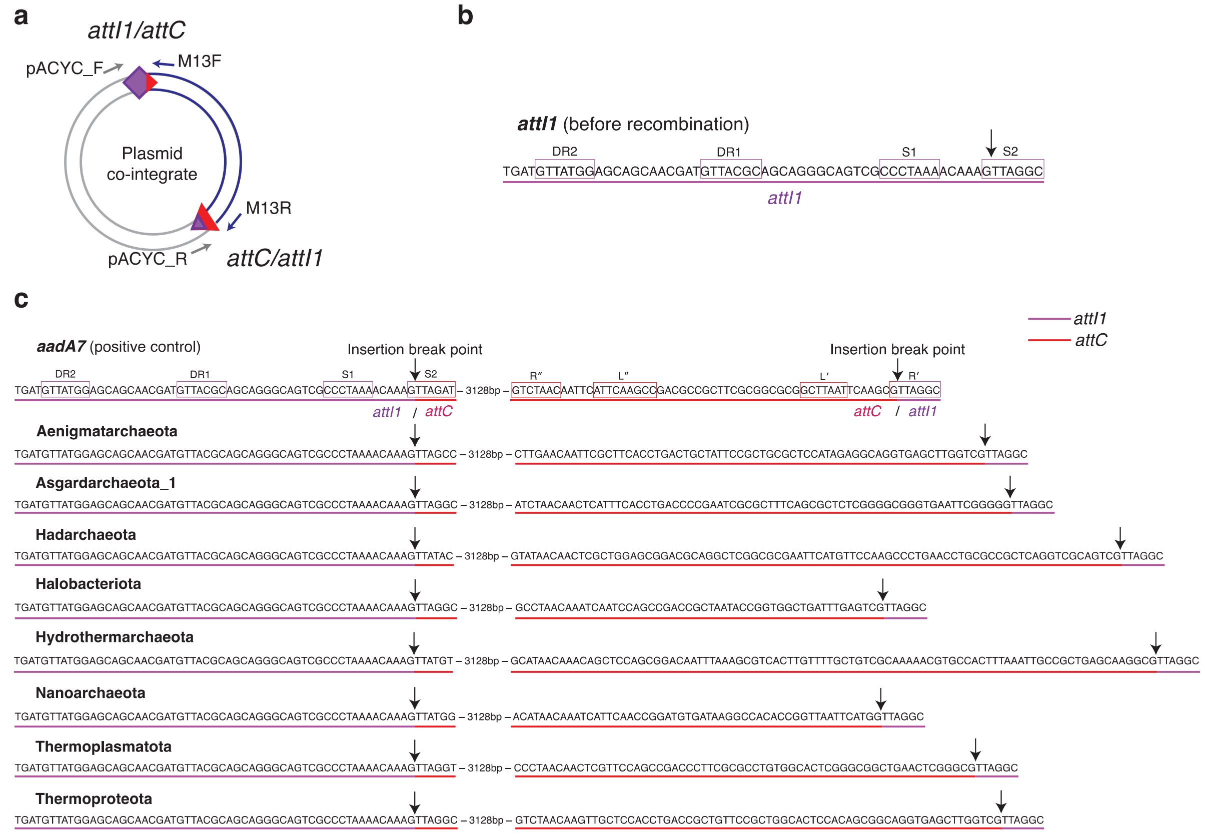
*induced using L-arabinose; †suppressed using D-glucose; ††ND = Not detected


Complete integrons
Asgardarchaeota [GCA_019056845.1: DXJY01000083.1]
1
1,462
5,291
Halobacteriota [GCA_013374555.1: QENK01000356.1]
3,672
2,849
1,595
1
Halobacteriota [GCA_002495885.1: DAHG01000063.1]
1
2,522
Methanobacteriota [GCA_020055905.1: JAGVIU010000108.1]
1
2,043
1kb
Integron integrase
Gene cassette ORF
attC
In0
Asgardarchaeota [GCA_003144275.1: NJBF01000069.1]
12,821
9,312
6,994
1
Thermoproteota [GCA_011333705.1: DTKA01000032.1]
1
2,219
2,671
3,168
CALINs
Thermoplasmatota [GCA_019347805.1: JAHWKR010000030.1]
41,620
4,900
1
Halobacteriota [GCA_001515185.1: LQHI01000043.1]
1
3,257
8,225
Halobacteriota [GCA_017883965.1: JAFNKG010000033.1]
28,652
24,826
1

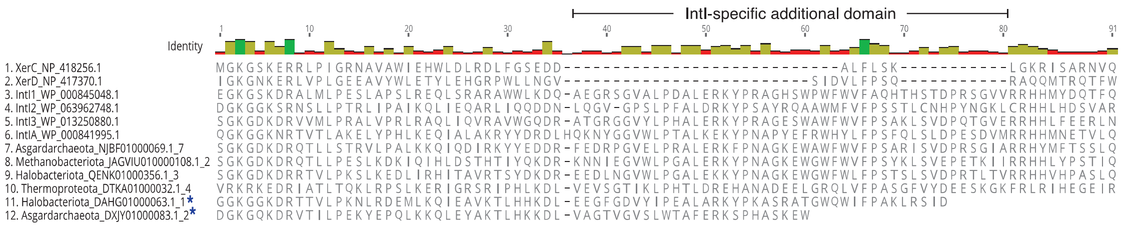
**Extended Data Fig. 1: Example structure of archaeal integrons.** Maps of all ‘complete integrons’, which are those that comprise an integron integrase gene (*intI*) and at least one gene cassette recombination site (*attC*); all ‘In0’ elements, which are those with *intI* but no detectable *attC* site; and three examples of ‘CALINs’ (clusters of *attC*s lacking integron integrases).


a
Halobacteriota [GCA_013374555.1: QENK01000356.1]
3,672
2,849
1,595
1
attI
Asgardarchaeota [GCA_019056845.1: DXJY01000083.1]
1
1,462
5,291
b
Consensus
Identity
1
10
20
30
40
50
60
71
KWT??????AYYGWYAAYYATMYRAWA???TYYATTTMAKATARYAG???GSARTATAWTWTWYGTTWRWC
1. QENK01000356.1 extraction GTTCAAAGAACCGATAATCATCCGAAA---TCCATTTCAGATAGTAG---GGAATATATTTTTCGTTTATC
2. DXJY01000083.1 extraction TAT------ATTGTCAACTATATAATATTGTTTATTTAATATAACAGATTGCAGTATAATATATGTTAGAC
c
Consensus
Identity
1
10
20
30
40
50
60
70
76
T??ATT-BANRHTANGYRVCRGY-----??HGRYKKRRDGYADYAVRSSDWTKDGGYSGTYWAHAAARMGTTAGAC
1. QENK01000356.1 extraction GTTCAAAGAACCGATA--ATCATCCGAAATCCATTTCAGATAGTAGGGAATATATTTTTCGTTTATC
2. DXJY01000083.1 extraction TATATTGTCAACTATATAATATT-----GTTTATTTAATATAACAGAT------TGCAGTATAATATATGTTAGAC
3. attI1_KY126370 T------GATGTTATGGAGCAGC-----AACGATGTTACGCAGCAGGG---CAGTCGCCCTAAACAAAGTTATAC
4. attI2_KY847875 TTAATT-AACGGTAAGCATCAGC-----GGGTGACAAAACGAGCATGCTTACTAATAAAATGTTAGAA
5. attI3_AY219651 CATTGTTTAACGAC-----CACGGTTGTGGGTATCCGGTGTTTGGTCAGATAACCATAAGTTAGAC
6. attI4_FJ982913 TTAATTAGATAGCGGT-------AGCCTACCTGTTGGAAAGGTAAGAAGCTGTCTAGAAAGCGTTAGCC
7. attI5_AF179595 TTTGATAGGCTGCGGT-------AGACTGCCAATAAAAGATGGATGTGGTTGTTTAGAAAGCGTTATGC
8. attIA_AF324211 CAGTGCGCGTTTTT-----ACTGGTCTAATGAGGTAAACGCGCATGCGCACTAATAAAATGTTAGGT
9. attIA_AY126447 CATGATAAGTGGCGGT-------AGCATGGTAAATTAGGTCGCATTTTGTTGTATAGATAGCGTTAGGT
10. attIA_AF324483 CATCCGGCCTTCGT-----CGATACTTGGCGCAACACCGCTTTTAGGCGGTCCAATATTAGTTAGGC
11. attIA_AF324484 CCGGTCCCATCCCGGC-------ATACTCCGTGCTACACCGCAATTAGGCGGTCCAATAGGTGTTAGGT
12. attI6_AF314191 GACAGATAACCAGG-----TTGAGAGAGGTGTTATACGGCTGTTCGCCGTCCAATATAAAGTTAAAC
13. attI7_AF314190 AACTAGGCGCCCGA-----ATGGGTGTGAACCGAGGAGTCGTATGGGGCGTCTAATATGCGTTAAAT
14. attI8_AF314189 CCCACC-GACGAGACGCGCCGTG----------AAGGTGGTGAACTTCTGGCCTGTCTAACAATTCGTTCAAG
15. attIPac_AY038186 CGTTATGCGACACA-----AGTGGCGGAAAAAATTACGCCGTTTGTGTCGTCTACTTATAGTTAGAT
16. attIPst_AY129393 CAAGCCGTTCGCTTCGCTCACTCGGGACCGGCTAAAACCGGCCCCTTAACCAAACGTTAGAG

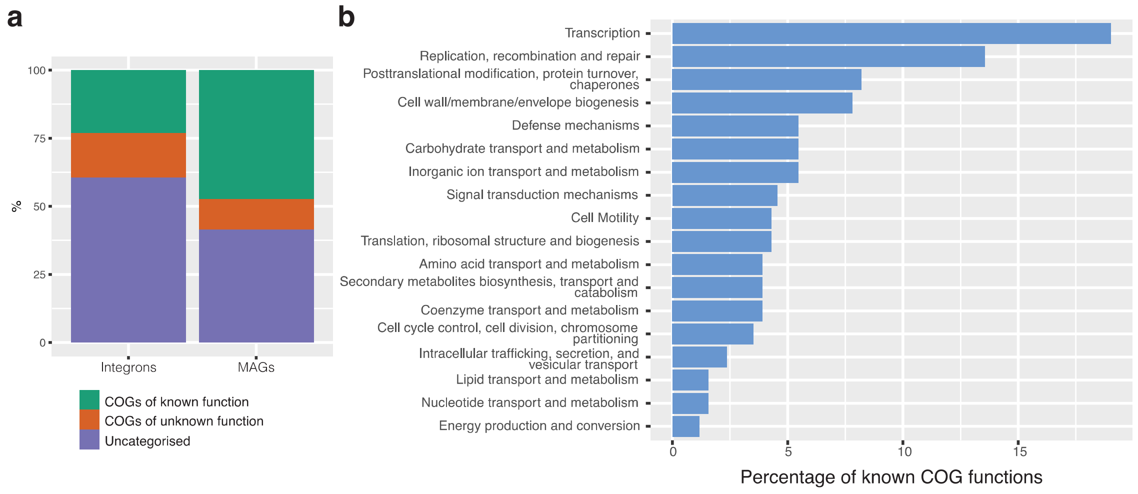
**Extended Data Fig. 2: Putative archaeal integron recombination sites, *attI*s. a**, maps showing the location of putative archaeal *attI*s. **b**, sequence alignment of the two putative archaeal *attI*s. **c**, multiple sequence alignment of the two archaeal *attI*s and all annotated bacterial *attI*s from the INTEGRALL database[80]. Nucleotides are coloured if they match with at least 50% of the sequences. Vertical arrows indicate the canonical insertion point of an inserting gene cassette.



**Extended Data Fig. 3: Sanger sequencing of *attI1* x *attC* recombination junctions. a,** schematic of PCR primer pairs (grey and blue arrows) that amplify the recombination junctions following cassette insertion (*attI1* x *attC* recombination). **b,** *attI1* sequence before recombination. Boxes denoted with S1 and S2 indicate the core IntI1 binding sites, and the direct repeats signified by DR1 and DR2, are additional strong and weak IntI1 binding sites, respectively. The black arrow indicates the insertion break point where cleavage takes place during recombination. **c,** Sanger sequence data of the recombinant clones following *attI1* recombination with the paradigmatic bacterial *attC* site (*attC*$_{aadA7}$), used as positive control, and eight archaeal *attC*s. Black arrows indicate the insertion break points following recombination. For *attC*$_{aadA7}$, the two sets of paired inverted repeats are boxed (R′ to R″ and L′ to L″).


IntI-specific additional domain
Identity
1
10
20
30
40
50
60
70
80
91
1. XerC_NP_418256.1
2. XerD_NP_417370.1
3. IntI1_WP_000845048.1
4. IntI2_WP_063962748.1
5. IntI3_WP_013250880.1
6. IntIA_WP_000841995.1
7. Asgardarchaeota_NJBF01000069.1_7
8. Methanobacteriota_JAGVIU010000108.1_2
9. Halobacteriota_QENK01000356.1_3
10. Thermoproteota_DTKA01000032.1_4
11. Halobacteriota_DAHG01000063.1_1*
12. Asgardarchaeota_DXJY01000083.1_2*

**Extended Data Fig. 4: A multiple protein sequence alignment of the additional domain unique to integron integrases.** Sequences (1) and (2) are tyrosine recombinases XerC and XerD that lack the IntI-specific domain. Sequences (3) to (6) are bacterial IntIs, and (7) to (12) are IntIs from Archaea. Blue asterisks indicate IntIs that did not span the full additional domain and were excluded from phylogenetic analysis.



**Extended Data Fig. 5: COG functional analysis of archaeal gene cassettes. a**, percentage of proteins assigned a COG category. 'Integrons' represent all cassette-encode proteins in Archaea, while 'MAGs' indicate all proteins from the 75 integron-bearing archaeal genomes. **b**, percentage of COGs with known functions assigned archaeal cassette-encoded proteins.