Automatic and accurate identification of integrons and cassette arrays in bacterial genomes reveals unexpected patterns

Jean Cury[1,2,*], Thomas Jové[3], Marie Touchon[1,2], Bertrand Néron[4], Eduardo PC Rocha[1,2]

[1] Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France
[2] CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France
[3] Univ. Limoges, INSERM, CHU Limoges, UMR_S 1092, F-87000 Limoges, France.
[4] Centre d'Informatique pour la Biologie, C3BI, Institut Pasteur, Paris, France

* To whom correspondence should be addressed. Tel: +33 1 40 61 36 37; Email: jean.cury@normalesup.org

Abstract

Integrons recombine gene arrays and favor the spread of antibiotic resistance. However, their broader roles in bacterial adaptation remain mysterious, partly due to lack of computational tools. We made a program – IntegronFinder – and used it to identify integrons in bacterial genomes with high accuracy and sensitivity. Some key taxa, such as α-Proteobacteria, lacked integrons suggesting they constitute deleterious genetic backgrounds for these elements. Integrons were much more frequent in intermediate size genomes, suggesting selection for compact gene acquisition. We used comparative genomics to quantify the differences between mobile and persistent integrons. The use of a covariance model to identify and align *attC* sites showed higher intrinsic variability in mobile integrons and a correlation between *attC* sites homogeneity and the number of integron cassettes. Surprisingly, numerous arrays of *attC* sites lacked nearby integrases (or pseudogenes of integrases), included many novel cassettes, and exhibited very diverse *attC* sites. These attC0 elements might provide incoming mobile integrons with a large unexplored pool of novel cassettes in genomes that currently lack integrons. They might also represent an intermediate step of the process of horizontal gene transfer following integron-capture and preceding definitive stabilization by loss of genetic mobility.

Introduction

Bacterial genomes can integrate exogenous genes at high rates (1). This ability is driven by the action of molecular systems facilitating the spread of genetic information (2,3). Integrons are gene-capturing devices playing a major role in the spread of antibiotic resistance genes (reviewed in (4-7)). Complete integrons have two main components (Figure 1). The first includes an integrase (*intI*) and its promoter (P_{intI}), an integration site named *attI* (attachment site of the integron) and a constitutive promoter (Pc) for the gene cassettes integrated at the *attI* site. The second is an array of gene cassettes, varying in size, from 0 to around 200 cassettes (8). A gene cassette is defined by an open reading frame (ORF) surrounded by its recombination sites named *attC* (attachment site of the cassette). Integrons lacking cassettes are named In0 elements (9). By analogy, we will name the arrays of *attC* sites without nearby integrases as "attC0" elements. Recombination between two adjacent *attC* sites by the action of the integrase leads to the excision of a circular DNA fragment composed of an ORF and an *attC* site. The recombination of the *attC* site of this circular DNA fragment with an *attI* site leads to integration of the circular DNA fragment at the *attI* site (10,11). This mechanism allows integrons to capture cassettes from other integrons or to rearrange the order of their cassettes (12). The constitutive promoter of the integron drives the expression of the downstream genes (13). Cassettes may also carry their own promoters leading to the expression of genes that may be distant from the P_C (14). The mechanism of the creation of new cassettes is unknown.

The key features of integrons are thus the integrase (IntI) and the arrays of *attC* sites (Figure 1). Integrases are site-specific tyrosine recombinases closely related to Xer proteins (15). Contrary to most other tyrosine recombinases, IntI can recombine nucleotide sequences of low similarity (16,17). The structure of the *attC* site, not its sequence, is essential for recombination by IntI (18). The sequence of IntI is distinguished from the other tyrosine recombinases by the presence of an additional ~35 residue domain near the patch III region that is involved in the recombination reaction of single stranded DNA (19). The *attC* sites have well-characterized traits reflecting a secondary structure that is essential for recombination (20). The *attC* site is split upon integration at *attI*, producing chimeric *attI*/*attC* sites on one side and

chimeric *attC*/*attC* sites on the other side of the cassette. Hence, integrons include arrays of chimeric *attC* sites arising from recombination between sequences that can be very different, because their palindromic structure is similar.

Previous literature often separates integrons between chromosomal (or super-) integrons carrying many cassettes, and five classes of so-called mobile integrons carrying few cassettes and associated with transposons (21 ,22,23). The IntI sequences within classes of mobile integrons show little genetic diversity, suggesting that each class of mobile integrons has emerged recently from a much larger and diverse pool of integrons (24-27). These events were probably caused by ancestral genetic rearrangements that placed the integrons on mobile elements. The latter drive the characteristically high frequency of horizontal transfer of mobile integrons (28). Accordingly, prototypical class 1 integrons are often found on the chromosomes of non-pathogenic soil and freshwater β-Proteobacteria (29). Yet, some integrons are both mobile and chromosomal (30 ,31), *e.g.*, encoded in integrative conjugative elements (32 ,33), and some chromosomal integrons have intermediate sizes (34). This challenges the dichotomy between short mobile and long stable chromosomal integrons (6,25). In fact, it might be more pertinent, from an evolutionary point of view, to split integrons in terms of the frequency with which they are lost and gained in bacterial lineages. This evolution-based classification might also facilitate the study of the molecular mechanisms leading to the generation of new cassettes, because it has been proposed that cassettes might originate in persistent integrons to be later spread by mobile integrons (35). This hypothesis was spurred by the presence of similar cassettes in *Vibrio cholerae* super-integrons and in mobile integrons suggesting recent transfer between the two. Furthermore, *attC* sites within the large integrons of *Vibrio* spp. showed higher similarity than between mobile integrons. It was thus suggested that cassettes were created by large chromosomal integrons and then recruited by mobile integrons (35).

Many mobile integrons carry antibiotic resistance genes, whereas the *Vibrio* spp. super-integrons encode very diverse functions, including virulence factors, secreted proteins, and toxin-antitoxin modules (24). Metagenomic data shows that there is a vast pool of poorly sampled cassettes in microbial communities (36). Although antibiotic-resistance integrons are abundant in human-associated environments such as sewage (37-39), most cassettes in environmental datasets encode other functions

or genes of unknown function (40 ,41). Hence, mobile integrons could facilitate the spread of a very large panel of potentially adaptive functions. Yet, the study of these functions has been hindered by the difficulty in identifying integron cassettes due to poor sequence similarity between *attC* sites. Approaches based on the analysis of sequence conservation have previously been used to identify these sites. The program XXR identifies *attC* sites in *Vibrio* super-integrons using pattern-matching techniques (24). The programs ACID (8) (no longer available) and ATTACCA (42) (now a part of RAC, available under private login) search mostly for class 1 to class 3 mobile integrons. Since the structure, not the sequence, of *attC* sites is important for function, the classical motif detection tools based on sequence conservation identify *attC* sites only within restricted classes of integrons. They are inadequate to identify or align distantly related *attC* sites.

Here we built a program named IntegronFinder (Figure 2) to detect integrons and their main components: the integrase with the use of HMM profiles and the *attC* sites with the use of a covariance model (Figure 3). Covariance models use stochastic context-free grammars to model the constraints imposed by sequence pairing to form secondary structures. Such models have been previously used to detect structured motifs, such as tRNAs (43). They provide a good balance between sensitivity, the ability to identify true elements even if very diverse in sequence, and specificity, the ability to exclude false elements (44). They are ideally suited to model elements with high conservation of structure and poor conservation of sequence, such as *attC* sites. IntegronFinder also annotates known *attI* sites, P_{intI} and P_C, and any pre-defined type of protein coding genes in the cassettes (*e.g.*, antibiotic resistance genes). IntegronFinder was built to accurately identify integrases and *attC* sites of any generic integron. Importantly, we have made the program available through a webserver that is free, requires no login, and has a long track record of stability (45). We also provide a standalone application for large-scale genomics and metagenomics projects. We used IntegronFinder to identify integrons in bacterial genomes. This allowed the characterization of integron distribution and diversity. Finally, we assessed the within integron diversity of *attC* sites to understand their evolution.

Material and Methods

Data. The sequences and annotations of complete genomes were downloaded from NCBI RefSeq (last accessed in November 2013, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Our analysis included 2484 bacterial genomes. We used the classification of replicons in plasmids and chromosomes as provided in the GenBank files. Our dataset included 2626 replicons labeled as chromosomes and 2007 as plasmids. The *attC* sites used to build the covariance model and the accession numbers of the replicons manually curated for the presence or absence of *attC* sites were retrieved from INTEGRALL, the reference database of integron sequences (http://integrall.bio.ua.pt/) (46). We used a set of 291 *attC* sites to build and test the model. We used a set of 346 sequences with expert annotation of 596 *attC* sites to analyze the quality of the predictions of the program.

Protein profiles. We built a protein profile for the region specific to the integron tyrosine recombinase. For this, we retrieved the 402 IntI homologues from the Supplementary file 11 of Cambray et al (47). These proteins were clustered using uclust 3.0.617 (48) with a threshold of 90% identity to remove very closely related proteins (the largest homologs were kept in each case). The remaining 79 proteins were used to make a multiple alignment using MAFFT (49) (--globalpair --maxiterate 1000). The position of the specific region of the integron integrase in *V. cholerae* was mapped on the multiple alignments using the coordinates of the specific region taken from (19). We recovered this section of the multiple alignment to produce a protein profile with hmmbuild from the HMMer suite version 3.1b1 (50). This profile was named intI_Cterm.

We used 119 protein profiles of the Resfams database (core version, last accessed on January 20, 2015 v1.1), to search for genes conferring resistance to antibiotics (http://www.dantaslab.org/resfams, (51)). We retrieved from PFAM the generic protein profile for the tyrosine recombinases (PF00589, phage_integrase, http://pfam.xfam.org/, (52)). All the protein profiles were searched using hmmsearch from the HMMer suite version 3.1b1. Hits with e-value smaller than 0.001 and coverage of at least 50% of the profile were regarded as significant.

Construction and analysis of *attC* models. We built a covariance model for the *attC* sites. These models score a combination of sequence and secondary structure

consensus (43) (with the limitation that these are DNA not RNA structures). To produce the *attC* models, 96 *attC* sites (33%) were chosen randomly from 291 known *attC* (see Data). The alignments were manually curated to keep the known conserved regions of the R and L boxes aligned in blocks. The unpaired central spacers (UCS) and the variable terminal structure (VTS) were not aligned because they are poorly conserved in sequence and length. Gaps were inserted in the middle of the VTS sequence as needed to keep the blocks of R and L boxes aligned. The consensus secondary structure was written in WUSS format beneath the aligned sequences (Supplementary file 1). The model was then built with INFERNAL 1.1 (44) using *cmbuild* with the option "--hand". This option allows the user to set the columns of the alignment that are actual matches (consensus). This is crucial for the quality of the model, because most of the columns in the R and L boxes would otherwise be automatically assigned as inserts due to the lack of sequence conservation. The R-UCS-L sections of the alignment were chosen as the consensus region, and the VTS was designed as a gap region. We used *cmcalibrate* from INFERNAL 1.1 to fit the exponential tail of the covariance model e-values, with default options. The model was used to identify *attC* sites using INFERNAL with two alternative modes. The default mode uses heuristics to reduce the sequence space of the search. The Inside algorithm is more accurate, but computationally much more expensive (typically 10^4 times slower) (44).

Identification of promoters and *attI* sites. The sequences of the Pc promoters for the expression of the cassette genes, of the P_{intI} promoters, and of the *attI* site were retrieved from INTEGRALL for the integron of class 1, 2 and 3 when available (see Table S1). We searched for exact matches of these sequences using pattern matching as implemented in Biopython v1.65 (53).

Overview of IntegronFinder: a program for the identification of *attC* sites, *intI* genes, integrons, and attC0 elements. IntegronFinder receives as input a sequence of DNA in FASTA format (Figure 2). It first annotates the CDSs in the sequence using Prodigal v2.6.2 (54) using the default mode for replicons larger than 200kb and the metagenomic mode for smaller replicons. In the present work, we omitted this part and used the NCBI RefSeq annotations because they are curated. The annotation step is particularly useful to study newly acquired sequences or poorly annotated ones.

The program searches for the two protein profiles of the integrase using hmmsearch from HMMER suite version 3.1b1 and for the *attC* sites with the default mode of cmsearch from INFERNAL 1.1 (Figure 1). Two *attC* sites are put in the same array if they are less than 4 kb apart on the same strand. The arrays are built by transitivity: an *attC* site less than 4 kb from any *attC* site of an array is integrated in that array. Arrays are merged when localized less than 4 kb apart. The threshold of 4kb was determined empirically as a compromise between sensitivity (large values decrease the probability of missing cassettes) and specificity (small values are less likely to put together two independent integrons). More precisely, the threshold is twice the size of the largest known cassettes (~2 kb (8)). The results of these searches are integrated to class the loci in three categories (Figure 1 - B, C, D). (1) *Complete integrons* have *intI* and at least one *attC* site. (2) *In0* have *intI* but not *attC*. One should note that we do not strictly follow the original definition of In0, which also includes the presence of an *attI* (9). We do not use this constraint because the sequence of *attI* is not known for most integrons. (3) The *attC0* have at least two *attC* sites and lack nearby *intI*.

To obtain a better compromise between accuracy and running time, IntegronFinder re-runs INFERNAL with the Inside algorithm ("--max" option in INFERNAL), but only around elements previously identified ("--local_max" option in IntegronFinder). More precisely, if a locus contains an integrase and *attC* sites (complete integron), the search is constrained to the strand encoding *attC* sites between the end of the integrase and 4kb after its most distant *attC*. If other *attC* sites are found after this one, the search is extended by 4 kb in that direction until no more new sites are found. If the element contains only *attC* sites (attC0), the search is performed on the same strand on both directions. If the integron is In0, the search is done on 4kb of both strands around the integrase. For complete and In0 integrons, the program then searches for nearby promoters and *attI* sites. Finally, the program can annotate the cassettes of the integron (defined in the program as the CDS found between intI and 200 bp after the last *attC* site, or 200 bp before the first and 200 bp after the last *attC* site if there is no integrase) with a database of protein profiles with the option "--func_annot". For example, in the present study we used the ResFams database to search for antibiotic resistance genes. One can use any hmmer-compatible profile databases with the program.

The program outputs tabular and GenBank files listing all the identified genetic elements associated with an integron. The program also produces a figure in pdf format representing each complete integron. For an interactive view of all the hits, one can use the GenBank file as input in specific programs such as Geneious (55).

Phylogenetic analyses. We have made two phylogenetic analyses. One analysis encompasses the set of all tyrosine recombinases and the other focuses on IntI. The phylogenetic tree of tyrosine recombinases (Figure S1) was built using 204 proteins, including: 21 integrases adjacent to *attC* sites and matching the PF00589 profile but lacking the intI_Cterm domain, seven proteins identified by both profiles and representative of the diversity of IntI, and 176 known tyrosine recombinases from phages and from the literature (15). We aligned the protein sequences with Muscle v3.8.31 with default options (56). We curated the alignment with BMGE using default options (57). The tree was then built with IQ-TREE multicore version 1.2.3 with the model LG+I+G4, which was identified as the best using the "--m TEST" option. This option computes the log-likelihood of different molecular evolution models, and chooses the model that minimizes the Bayesian Information Criterion (BIC). We made 10000 ultra fast bootstraps to evaluate node support (Figure S1, Tree S1).

The phylogenetic analysis of IntI was done using the sequences from complete integrons or In0 elements (*i.e.*, integrases identified by both HMM profiles). We added to this dataset some of the known integrases of class 1, 2, 3, 4 and 5 retrieved from INTEGRALL. Given the previous phylogenetic analysis we used known XerC and XerD proteins to root the tree. Alignment and phylogenetic reconstruction were done using the same procedure; except that we built ten trees independently, and picked the one with best log-likelihood for the analysis (as recommended by the IQ-TREE authors (58)). The robustness of the branches was assessed using 1000 bootstraps (Figure S2, Tree S2).

Pan-genomes. We built pan-genomes for 12 species having at least 4 complete genomes available in Genbank RefSeq and encoding at least one IntI. It represents 40% of complete integrons. Pan-genomes are the full complement of genes in the species and were built by clustering homologous proteins into families for each of the species (as previously described in (59)). We did not build a pan-genome for *Xanthomonas oryzae* because it contained too many rearrangements and repeated elements(60) making the analysis of positional orthologs inaccurate. Briefly, we

determined the lists of putative homologs between pairs of genomes with BLASTP and used the e-values ($<10^{-4}$) to cluster them using SILIX (61). SILIX parameters were set such that a protein was homologous to another in a given family if the aligned part had at least 80% identity and if it included more than 80% of the smallest protein. IntI proteins were regarded as persistent if they were present in at least 60% of the genomes of the species.

Pseudo-genes detection. To detect *IntI* pseudo-genes around attC0 elements, we translated the 6 frames of the region containing the attC0 element plus 10kb before and after the element. Then we ran hmmsearch from HMMER suite v3.1b1 with the profile intI_Cterm and the profile PF00589. We recovered hits with e-values lower than 10^{-3} and whose alignments covered more than 50% of the profiles.

IS detection: We identified insertion sequences (IS) by searching for sequence similarity between the genes present 4kb around or within each genetic element and a database of IS from ISFinder (62). Details can be found in (63).

Detection of cassettes in INTEGRALL: We searched for sequence similarity between all the CDS of attC0 elements and the INTEGRALL database using blastn from BLAST 2.2.30+. Cassettes were considered homologous to those of INTEGRALL if they had at least 40% identity in the blastn alignment.

Results

Models for *attC* sites

We selected a manually curated set of 291 *attC* sites representative of the diversity of sequences available in INTEGRALL (see Methods). From these, we selected 96 (33%) to build a covariance model of the *attC* site and set aside the others for testing. The characteristics of these sequences were studied in detail (Figure 3A), notably concerning the R and L boxes, the UCS, and the EHB (18). The AAC and the complementary GTT sequences from the aptly named Conserved Triplet were indeed highly conserved. Other positions were less conserved, but nevertheless informative for the model (Figure 3B and 3D). The VTS length was highly variable, between 20 and 100 nucleotides long, as previously observed (4). We then used the covariance model to search for *attC* sites on 2484 complete bacterial genomes. The genomic *attC* sites had some differences relative to those used to build the model. They showed stronger consensus sequences and more homogeneous VTS lengths (Figure 3C). The analyses of sensitivity in the next paragraph show that our model missed very few sites. Hence, the differences between the initial and the genomic *attC* sites might be due to our explicit option of using diverse sequences to build the model (to maximize diversity). They may also reflect differences between mobile integrons (very abundant in INTEGRALL) and integrons in sequenced bacterial genomes (where a sizeable fraction of cassettes were identified in *Vibrio* spp.).

We tested the covariance model in two ways. Firstly, by searching for the 195 remaining *attC* sites that were not used to build the model. We randomized bacterial genomes with varying G+C content (Table S2), where we integrated five *attC* sites at 2 kb intervals. We searched for the *attC* sites and found very few false positives (~0.02 FP/Mb, Figure 4B), independently of the run mode (see Methods for details). The proportion of true *attC* actually identified (sensitivity), was 60 % for the default mode and 83% for the most accurate mode (with option local_max). We identified at least two of the *attC* sites in 98% of the arrays (with the most accurate mode). Hence, the array is identified even when some *attC* sites are not detected. The sensitivity of the model showed very little dependency on genome G+C composition in all cases (Figure 4). Secondly, we searched *attC* sites in 346 DNA sequences containing 596 *attC* sites annotated in INTEGRALL (Table S4). We found 571 *attC*

sites with the most accurate mode (96% of sensitivity). We missed 25 *attC* sites, among which 14 were on the integron edges, and were probably missed because of the absence of R' box on the 3' side. All the 57 sequences with integrons annotated as In0 in INTEGRALL also lacked *attC* site in our analysis. We found 246 *attC* sites missing in the annotations of INTEGRALL (0.71 FP/Mb). Importantly, more than 89% of these were found in arrays of two *attC* sites or more. If isolated *attC* sites were all false positives (and the only ones), then the false positive rate would be 0.072 FP/Mb, *i.e.*, less than one false *attC* site per genome. These two analyses indicated a rate of false positives between 0.02 FP/Mb and 0.71 FP/Mb. The probability of having arrays of two or more false *attC* sites (within 4 kb) given this density of false positives is between 4.10^{-6} and 3.10^{-9} depending on the false positive rate (assuming a Poisson process). Hence, the arrays of *attC* sites given by our model are extremely unlikely to be false positives.

Identification of integrases

Tyrosine recombinases can be identified using the PFAM PF00589 protein profile. To distinguish IntI from the other tyrosine recombinases, we built an additional protein profile corresponding to the IntI specific region near the patch III domain (19) (henceforth named intI_Cterm, see Methods). Within all complete genomes we found 215 proteins matching both profiles. Only six genes matched intI_Cterm but not PF00589, and among the more than 19,000 occurrences of PF00589 not matching intI_Cterm, 47 co-localized with an *attC* site. Among the latter, 26 were in genomes that encoded IntI elsewhere in the replicon (Figure S3). The remaining 21 integrases were scattered in the phylogenetic tree of tyrosine recombinases, and only four of them were placed in an intermediate position between IntI and Xer (Figure S1). These four sequences resembled typical phage integrases at the region of the patch III domain characteristic of IntI. Furthermore, they co-localized with very few *attC* sites, (always less than three). This analysis strongly suggests that tyrosine recombinases lacking the intI_Cterm domain are most likely not IntI.

Most *intI* genes identified in bacterial genomes co-localized with *attC* sites (73%, Figure S3). It is difficult to assess if the remaining *intI* genes are true or false, since In0 elements have often been described in the literature (9,64). We were able to identify IntI in the integrons of class 1 to class 5, as well as in well-known chromosomal integrons (*e.g.*, in *Vibrio* super-integrons). We also identified all In0

elements in the Integrall dataset mentioned above. Overall, these results show that IntI could be identified accurately using the intersection of both protein profiles.

We built a phylogenetic tree of the 215 IntI proteins identified in genomes (Figure S2). Together with the analysis of the broader phylogenetic tree of Tyrosine recombinases (Figure S1), this extends previous analyses (4,25,27,65): 1) The XerC and XerD sequences are close outgroups. 2) The IntI are monophyletic. 3) Within IntI, there are early splits, first for a clade including class 5 integrons, and then for *Vibrio* super-integrons. 4) A major split occurs between a clade including the classed 4 and 2 on one side and most other integrons and especially classes 1 and 3 on the other. 5) A group of integrons displaying an integrase in the opposite direction (inverted integrase group) was previously described as a monophyletic class (25), but in our analysis it was clearly paraphyletic (Figure S2). Notably, a Class 1 integron present in the genome of *Acinetobacter baumannii* 1656-2 had an inverted integrase.

Integrons in bacterial genomes

We build a program – IntegronFinder – to identify integrons in DNA sequences. This program searches for *intI* genes, *attC* sites, clusters them in function of their co-localization, and then annotates cassettes and other accessory genetic elements (see Figure 2 and Methods). The use of this program led to the identification of 215 IntI and 4543 *attC* sites in complete bacterial genomes. The combination of this data resulted in a dataset of 157 complete integrons, 58 In0 and 272 attC0 elements (see Figure 1 for their description). The frequency of complete integrons is compatible with previous data (25). While most genomes encoded a single integrase, we found 36 genomes encoding more than one, suggesting that multiple integrons are relatively frequent (20% of genomes encoding integrons). Interestingly, while many of the integrons reported in the literature were encoded in plasmids, this is not the case in the dataset of complete genomes, where only one plasmid was found to encode an In0 integron (apart from those of class 1), in a set of 24 plasmids.

The taxonomic distribution of integrons was very heterogeneous. Some clades contained many elements, *e.g.*, 19% of the γ-Proteobacteria encoded at least one complete integron (Figure 5 and S4). This is almost 4 times as much as expected by chance alone (χ^2 test in a contingency table, $P < 0.001$), since only 5% of the genomes were found to encode complete integrons. The β-Proteobacteria also encoded numerous integrons (~10% of the genomes). However, all the genomes of

Firmicutes, Tenericutes, and Actinobacteria lacked complete integrons. Furthermore, all genomes of α-Proteobacteria, the sister-clade of γ and β-Proteobacteria, were devoid of complete integrons, In0, and attC0 elements (no occurrence among 243 genomes). Interestingly, much more distantly related bacteria such as Spirochaetes, Chlorobi, Chloroflexi, Verrucomicrobia, and Cyanobacteria encoded integrons (Figure 5 and Figure S4). The complete lack of integrons in one large phylum of Proteobacteria is thus very intriguing.

Which are the traits associated with the presence of integrons? In the literature integrons are often associated with antibiotic resistance. To quantify this association, we searched for determinants of antibiotic resistance in integron cassettes (see Methods). We identified resistance genes in 105 cassettes, *i.e.*, in 3% of all cassettes from complete integrons (3470 cassettes). These cassettes were mostly found in class 1 to 5 integrons (90% of them), even though these classes of integrons accounted for only 4% of complete integrons' cassettes. This fits previous observations that integrons carrying antibiotic resistance determinants are rare in natural populations (29,36).

The association between genome size and the frequency of integrons has not been studied before. We binned the genomes in terms of their size and analyzed the frequency of complete integrons, In0, and attC0. This showed a clearly non-monotonic trend (Figure 6). The same result was observed in a complementary analysis using only integrons from Gamma-Proteobacteria (Figure S5). Very small genomes lack complete integrons, intermediate size genomes accumulate most of the integrons, and the largest genomes encode few. Importantly, the same trends were observed for In0 and attC0. Hence, the frequency of integrons is maximal for intermediate genome sizes.

Unexpected abundance of attC0 elements

The number of observed *attC* sites lacking nearby integrases is unexpectedly high and to the best of our knowledge has not been reported before. We found 432 occurrences of isolated single *attC* sites among the total of 1649 identified *attC* sites without a nearby integrase. We decided to discard them for further analysis on attC0 elements and keep only the 272 attC0 elements with two or more *attC* sites. If isolated single *attC* sites were all false, and were the only false ones, then the observed rate of false positives can be estimated at 0.047 FP/Mb. This is within the

range of the rates of false positives observed in the sensitivity analysis (between 0.02 FP/Mb and 0.71 FP/Mb). We showed above that with these rates the probability that attC0 elements are false positives is exceedingly small. The attC0 resemble mobile integrons in terms of the number of cassettes: 86% had fewer than six *attC* sites and only 6.6% had more than 10 (Figure S6). On the other hand, they were remarkably different from mobile integrons in terms of the actual known genes encoded in cassettes: only 117 cassettes out of the 1649 cassettes were homologous to cassettes reported in INTEGRALL. Accordingly, we only found antibiotic resistance genes in 32 cassettes (2%) of the attC0 (to be compared with 70% among class 1 to class 5 integrons and to 0.3% among complete integrons not from class 1 to 5). Hence, attC0 are relatively small and have mostly unknown gene cassettes.

The attC0 elements might have arisen from the loss of the integrase in a previously complete integron. Therefore, we searched for pseudogenes matching the specific IntI_Cterm domain less than 10kb away from attC0. We found such pseudo-genes near 22 out of 272 attC0 elements. Among the 22 hits, three matched integron integrases. These three cases correspond to IntI encoded more than 4kb away from the closest *attC* site (which is why they were missed previously). It is worth noting that out of the 22 hits, 15 pseudo-genes are also matched with the PF00589 profile, which is consistent with the idea that they previously encoded intI. Overall, our analysis showed that most attC0 (90%) are not close to recognizable IntI pseudogenes.

Chromosomal rearrangements may split integrons and separate some cassettes from the neighborhood of the integrase, thus producing attC0 elements. Under these circumstances, an *intI* and the attC0 would be present in the same genome. To identify these cases, we searched for *intI* in the genomes with attC0. We found that half of the genomes with attC0 also had an *intI*. In some few cases IntI was actually encoded in another replicon (3.5% of attC0 elements). Hence, half of the attC0 elements are in genomes lacking any *intI*.

Insertion sequences (IS) may be responsible for the creation of some of the abovementioned attC0 elements by promoting chromosomal rearrangements in a previously complete single integron. We searched for IS near attC0 elements, In0 and complete integrons (see Methods). We found that 16% of attC0 and 26% of the complete integrons encoded at least one IS within their cassettes. Upon IS-mediated rearrangements, the attC0 elements should be close to an IS. Indeed, 44% of the

attC0 had a neighboring IS. Furthermore, attC0 in genomes with distant *intI* were more likely to be close to an IS than attC0 in genomes lacking *intI* ($P < 0.001$, χ^2 contingency table). Hence, even if half of the attC0 elements are in genomes lacking *intI*, these results are consistent with the hypothesis that IS contribute to disrupt integrons and create attC0 elements.

Divergence of *attC* sites

The covariance model that we defined for the *attC* sites allowed their alignment. This opened the possibility of assessing rigorously the sequence heterogeneity of *attC* sites in function of integron mobility to test the hypothesis that *attC* sites are less heterogeneous in persistent integrons. We made a preliminary classification of complete integrons and In0 using the species' pan-genomes (see Methods). Integrons from classes 1 to 5, those carried in plasmids, and those present in less than 60% of the genomes of a species were all classed as mobile. As expected, mobile integrons had few cassettes: only two have more than six *attC* sites (respectively nine and fifteen cassettes) (Figure 7). The integrons present in more than 60% of the strains of a species carried many more cassettes, with some notable exceptions (*e.g.*, in *Xanthomonas campestris* they had between 0 and 22 cassettes). Since large arrays of cassettes were not found in known mobile integrons, we built a dataset of persistent integrons including all the abovementioned integrons present in more than 60% of the strains of a species and those with more than 20 *attC* sites. This resulted in a set of 27 persistent integrons. With these criteria, around 67% of the chromosomal integrons were classed as mobile in the species with pan-genomes, showing that the traditional separation between chromosomal and mobile integrons may be misleading. Both mobile and persistent integrons were found in the major clades of the IntI phylogeny (Figure S2).

We then tested the hypothesis that persistent integrons have less diverse *attC* sites than mobile integrons. For this, we analyzed the identity between the R-UCS-L box of *attC* sites of mobile and persistent integrons while controlling for the effect of intra and inter-integron comparisons. Since *attC* sites are poorly conserved in sequence, we aligned them using the covariance model. As expected, *attC* sites were more similar within than between integrons, and more similar within persistent integrons than within mobile integrons (Figure 8A). Since some integrons have many more cassettes than others, we made a complementary analysis between mobile and

persistent integrons comparing the average value of *attC* similarity per integron. This analysis allowed to control for integron size and produced similar results (both $P < 0.05$). Interestingly *attC* sites within attC0 elements were even less similar than those of mobile integrons ($P<0.001$). The length of the VTS sequences showed similar patterns: higher similarity between *attC* sites of the same integron, and especially within persistent integrons (Figure 8B). The difference remains significant when the analysis was done using the average variation in VTS per integron ($P < 0.001$). We then quantified the relationship between the number of *attC* sites in an integron and the average intra-integron sequence dissimilarity in *attC* sites. The sequence dissimilarity diminishes exponentially with an increasing number of *attC* sites (Figure 9), *i.e.*, the integrons carrying the longest arrays of cassettes are those with most homogeneous *attC* sites.

Discussion

IntegronFinder, limitations and perspectives

IntegronFinder identifies the vast majority of known *attC* sites and *intI* genes. With an observed accuracy of ~83% in the identification of individual *attC* sites, the probability of missing all elements in an array of four *attC* sites is less than 0.08%. The high accuracy of IntegronFinder allowed the identification of attC0 elements and many novel cassettes. Importantly, IntegronFinder is relatively insensitive to genomic G+C content. Hence, compositional biases are unlikely to affect the identification of integrons.

In some circumstances, it may be necessary to interpret with care the results of IntegronFinder. For example, genome rearrangements resulting in the split of integrons will lead to the separate identification of an attC0 and an integron (eventually an In0 if the rearrangement takes place near the *attI* site). IntegronFinder accurately identifies these two genetic elements. However, these elements may remain functionally linked because cassettes from the attC0 may be excised by the integrase and re-inserted in the integron at its *attI* site. Overall, this reflects the dynamics of the integrons, but it is unclear if the two elements should be regarded as independent, as it is done by default, or as a single integron. One should note that such cases might be difficult to distinguish from alternative evolutionary *scenarii* involving the loss of the integrase in one of multiple integrons of a genome.

Our analyses show that IntegronFinder detects few false positives. In this study we only analyzed the *attC* sites co-occurring in arrays and those neighboring *intI*. Given the specificity in the identification of *attC* sites these arrays are unlikely to be false positives. The function of the other 432 single *attC* sites observed in genomes is less clear. Many of these sites are likely to be false positives because their frequency in genomes is close to that observed for false positives in our validation procedure. However, one cannot exclude the hypothesis that they are associated with other functions or result from the genetic degradation of integron cassettes.

IntegronFinder can be used to analyze diverse types of data. Our study was restricted to complete bacterial genomes to avoid the inference of the poor quality of some genome assemblies with the assessment of the program accuracy. However, IntegronFinder can be used to analyze draft genomes or metagenomes as long as

one is aware of the limitations of the procedure in such data. The edges of contigs of draft genomes often coincide with transposable elements. We have shown that a significant number of integrons have cassettes with IS. These elements induce contig breaks in the sequence assemblies. Hence, integrons with IS are split in the assembly process leading to the loss of the information on genetic linkage between their components. Under these circumstances, IntegronFinder will identify several genetic elements even if the genome actually encodes one single complete integron. Metagenomic data is even more challenging because it includes numerous small contigs where it is difficult to identify complete integrons. Yet, we showed that the models for *attC* sites and IntI are very accurate. They can thus be used independently to identify the occurrence of cassettes and integrases in assembled metagenomes. This might dramatically increase our ability to identify novel gene cassettes in environmental data.

Determinants of integron distribution

Our analysis highlighted associations between the frequency of integrons and certain genetic traits. The frequency of attC0, complete integrons, and In0 is often highly correlated in relation to all of these traits, *e.g.*, all three types of elements show roughly similar distributions among bacterial phyla and in terms of genome size. This strong association between the three types of elements is most likely caused by their common evolutionary history.

Integrons have well-known roles in the spread of antibiotic resistance. Nevertheless, we identified very few known antibiotic resistance genes in complete integrons outside the class 1 to class 5 integrons. Interestingly, we also found few resistance genes in attC0 elements. This fosters previous suggestions that integrons carry a very diverse set of adaptive traits, beyond antibiotic resistance genes, in many natural populations (36).

We found an under-representation of integrons in both small and large bacterial genomes. Since integrons are gene-capturing devices, one would expect a positive association between the frequency of integrons and that of horizontal transfer. Under this hypothesis, the lack of integrons in small genomes is not surprising since many of these bacteria are under sexual isolation, and they typically have few or no transposable elements, plasmids, or phages (66-68). However, the largest genomes have few integrons, but many mobile elements and are thought to engage in very

frequent horizontal transfer (69,70). We can only offer a speculation to explain this puzzling result. Integrons have been regarded as compact platforms of genetic recombination, which potentially have applications in synthetic biology (71). Horizontal transfer is often brought by mobile genetic elements of which some can be very large and costly (72). If there were selection for compact acquisition of adaptive traits by horizontal gene transfer, then one would expect that its intensity should scale with the inverse of genome size. This is because constraints on the size of incoming genetic material are expected to be less important for larger genomes. Hence, integrons might be under less intense selection in larger genomes. The combined effect of the frequency of transfer (increasing with genome size) and selection for compactness (decreasing with genome size) could explain the high abundance of integrons in medium-sized genomes.

Most integrons available in INTEGRALL from known taxa are from γ-Proteobacteria (90%) (46). We found a more diverse set of phyla in our analyses. While most integrons were found in Proteobacteria, which represent around half of the available complete genomes, the frequency of integrons in β-Proteobacteria was almost as high as among γ-Proteobacteria. We also identified some complete integrons in clades distantly related from Proteobacteria, including one in Cyanobacteria. Surprisingly, the genomes from α-Proteobacteria had no integrons, even if they encoded many tyrosine recombinases involved in the integration of a variety of mobile genetic elements. The complete absence of integrons, In0, and attC0 in α-Proteobacteria is extremely puzzling. It cannot solely be ascribed to the frequency of small genomes in certain branches of α-Proteobacteria, since our dataset included 99 genomes larger than 4Mb in the clade. We also did not find complete integrons in Gram-positive bacteria. Transfer of genetic information between clades of Proteobacteria and between Proteobacteria and Gram-positive bacteria is well documented (73,74). Accordingly, many mobile genetic elements have spread among bacteria, *e.g.*, conjugative elements have adapted to the diverse cell envelopes of all major bacterial phyla (75). Importantly, integrons have occasionally been identified in Firmicutes and α-Proteobacteria (46,76), and we have found attC0 in Firmicutes and In0 in Actinobacteria. This shows that these elements are sometimes transferred to these bacteria. Their absence from the many complete genomes available for these phyla in our dataset suggests the existence of some

unknown mechanism hindering the stable establishment of integrons in these bacteria after transfer. It is well known that differences in the translation machinery hinder the expression of transferred genetic information from Proteobacteria to Firmicutes (*e.g.*, the protein S1 (77)). However, these differences cannot explain the lack of integrons in α-Proteobacteria.

The evolution of integrons

Our study sheds some new light on how integrons evolve, acquire new cassettes, and may eventually disappear. The use of the covariance model confirmed that *attC* sites are more similar within than between integrons. It also showed that mobile integrons have *attC* sites more heterogeneous in sequence and length than persistent integrons. A previous observation that *Vibrio* super-integrons had very homogenous *attC* sites spurred the hypothesis that they created the integron cassettes, which might be later spread by mobile integrons (35,78). Our results are compatible with this idea and with the more general view that persistent integrons are responsible for the creation of most cassettes whereas mobile integrons are responsible for their spread among bacteria. Importantly, we found a negative association between *attC* heterogeneity within an integron and the number of cassettes carried by the integron. This may indicate that the largest persistent integrons are those creating more cassettes.

On the other extreme, many arrays of *attC* are not even associated with an integrase. Several previous works have identified IntI pseudo-genes in bacterial genomes (27,34). Here, we have found a surprisingly high number of attC0 elements, nearly half of which are found in genomes lacking *intI* genes or nearby pseudogenes of *intI*. AttC0 elements may have arisen in several ways. 1) By the unknown mechanism creating novel cassettes if this mechanism does not depend on IntI. 2) By integration of cassettes at a non-specific site by an integron encoded elsewhere in the genome (as described before (10)). 3) By loss of *intI*, even if most attC0 lacked neighboring pseudogenes of *intI*. 4) By genome rearrangements splitting a group of cassettes from the neighborhood of *intI* (as observed in (79)). The two last mechanisms are consistent with the presence of IS in nearly a fourth of complete integrons and might explain why nearly half of the attC0 are in replicons encoding IntI.

Which integrons gave rise to attC0? The average number of *attC* sites in attC0 is close to that of mobile integrons. The within-element sequence identity of *attC* sites is

slightly lower in attC0 than in mobile integrons. These two observations might suggest that attC0 elements were derived from mobile elements having lost the integrase. Yet, this does not fit our observations that attC0 have very few cassettes homologous to those of mobile integrons and far fewer antibiotic resistant genes than mobile integrons. Hence, most attC0 do not derive from the well-known mobile integrons carrying antibiotic resistance genes. Instead, they may derive from other less well-characterized types of mobile integrons, or from partially deleted ancestral persistent integrons. Actually, some attC0 elements might be small and encode heterogeneous *attC* sites because they are not under selection for mobilization by an integron.

The lack of an integrase in attC0 elements does not imply that these cassettes cannot be mobilized. We found many co-occurrences of complete integrons, In0, and attC0 elements in the same genomes. These co-occurrences facilitate the exchange of cassettes between elements. The integrases of mobile integrons have relaxed sequence similarity requirements to mediate recombination between divergent *attC* sites. It is thus tempting to speculate that integrons transferred into a genome encoding attC0 might be able to integrate attC0 cassettes in their own array of cassettes. If attC0 elements are frequently recruited by such incoming complete integrons, or In0 elements, then many genomes currently lacking integrons might be important reservoirs of novel cassettes.

Independently, of the origin and transfer of attC0 elements, the genes they encode might be expressed and have an adaptive value. In fact, these genes might often be adaptive, since otherwise they would have been inactivated as the result of the accumulation of mutations that resulted in divergent *attC* sites in attC0 elements. AttC0 with very degenerate *attC* sites might thus represent an intermediate step between the acquisition of a gene by an integron and its definitive stabilization in the genome by loss of the IntI-based cassette mobilizing activity.

Availability

The program was written in Python 2.7. It is freely available on a webserver (http://mobyle.pasteur.fr/cgi-bin/portal.py#forms::integron_finder). The standalone program is distributed under an open-source GPLv3 license and can be downloaded from Github (https://github.com/gem-pasteur/Integron_Finder/) to be run using the command line. Supplementary materials include tables containing all integrons found at different level (Tables S4, S5 and S6). It includes the list of the 596 *attC* sites with their annotated position (Table S7a), and the corresponding file with observed position (Table S7b). We provide the covariance model for the *attC* site (File S1).

Funding

This work was supported by an European Research Council grant [281605 to E.P.C.R.].

Acknowledgements

JC is a member of the "Ecole Doctorale Frontière du Vivant (FdV) – Programme Bettencourt". We thank Didier Mazel, Jose Escudero, Céline Loot, Aleksandra Nivina, Philippe Glaser, Claudine Médigue, Alexandra Moura, and Julian E Davies for fruitful discussions and comments on the manuscript.

Author contributions. Designed the study: JC EPCR. Made the analysis: JC. Wrote the software and webserver: JC and BN. Contributed with data: MT TJ. Drafted the manuscript: JC EPCR. All authors contributed to the final text of the manuscript.

References

1. Tettelin, H., Riley, D., Cattuto, C. and Medini, D. (2008) Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*, **11**, 472-477.
2. de la Cruz, F. and Davies, J. (2000) Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol*, **8**, 128-133.
3. Frost, L.S., Leplae, R., Summers, A.O. and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol*, **3**, 722-732.
4. Mazel, D. (2006) Integrons: agents of bacterial evolution. *Nat Rev Microbiol*, **4**, 608-620.
5. Partridge, S.R. (2011) Analysis of antibiotic resistance regions in Gram-negative bacteria. *FEMS Microbiology Reviews*, **35**, 820-855.
6. Gillings, M.R. (2014) Integrons: past, present, and future. *Microbiol Mol Biol Rev*, **78**, 257-277.
7. Escudero, J.A., Loot, C., Nivina, A. and Mazel, D. (2015) The Integron: Adaptation On Demand. *Microbiol Spectr*, **3**, MDNA3-0019-2014.
8. Joss, M.J., Koenig, J.E., Labbate, M., Polz, M.F., Gillings, M.R., Stokes, H.W., Doolittle, W.F. and Boucher, Y. (2009) ACID: annotation of cassette and integron data. *BMC Bioinformatics*, **10**, 118.
9. Bissonnette, L. and Roy, P.H. (1992) Characterization of In0 of Pseudomonas aeruginosa plasmid pVS1, an ancestor of integrons of multiresistance plasmids and transposons of gram-negative bacteria. *J Bacteriol*, **174**, 1248-1257.
10. Recchia, G.D., Stokes, H.W. and Hall, R.M. (1994) Characterisation of specific and secondary recombination sites recognised by the integron DNA integrase. *Nucleic Acids Res*, **22**, 2071-2078.
11. Recchia, G.D. and Hall, R.M. (1995) Plasmid evolution by acquisition of mobile gene cassettes: plasmid pIE723 contains the aadB gene cassette precisely inserted at a secondary site in the incQ plasmid RSF1010. *Mol Microbiol*, **15**, 179-187.
12. Hall, R.M., Brookes, D.E. and Stokes, H.W. (1991) Site-specific insertion of genes into integrons: role of the 59-base element and determination of the recombination cross-over point. *Mol Microbiol*, **5**, 1941-1959.
13. Collis, C.M. and Hall, R.M. (1995) Expression of antibiotic resistance genes in the integrated cassettes of integrons. *Antimicrob Agents Chemother*, **39**, 155-162.
14. Michael, C.A. and Labbate, M. (2010) Gene cassette transcription in a large integron-associated array. *BMC Genet*, **11**, 82.
15. Nunes-Duby, S.E., Kwon, H.J., Tirumalai, R.S., Ellenberger, T. and Landy, A. (1998) Similarities and differences among 105 members of the Int family of site-specific recombinases. *Nucleic Acids Res*, **26**, 391-406.
16. Collis, C.M., Recchia, G.D., Kim, M.J., Stokes, H.W. and Hall, R.M. (2001) Efficiency of recombination reactions catalyzed by class 1 integron integrase IntI1. *J Bacteriol*, **183**, 2535-2542.
17. MacDonald, D., Demarre, G., Bouvier, M., Mazel, D. and Gopaul, D.N. (2006) Structural basis for broad DNA-specificity in integron recombination. *Nature*, **440**, 1157-1162.

18. Bouvier, M., Ducos-Galand, M., Loot, C., Bikard, D. and Mazel, D. (2009) Structural features of single-stranded integron cassette attC sites and their role in strand selection. *PLoS Genet*, **5**, e1000632.
19. Messier, N. and Roy, P.H. (2001) Integron integrases possess a unique additional domain necessary for activity. *J Bacteriol*, **183**, 6699-6706.
20. Frumerie, C., Ducos-Galand, M., Gopaul, D.N. and Mazel, D. (2010) The relaxed requirements of the integron cleavage site allow predictable changes in integron target specificity. *Nucleic Acids Res*, **38**, 559-569.
21. Hall, R.M. and Collis, C.M. (1995) Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol*, **15**, 593-600.
22. Mazel, D., Dychinco, B., Webb, V.A. and Davies, J. (1998) A distinctive class of integron in the Vibrio cholerae genome. *Science*, **280**, 605-608.
23. Partridge, S.R., Tsafnat, G., Coiera, E. and Iredell, J.R. (2009) Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol Rev*, **33**, 757-784.
24. Rowe-Magnus, D.A., Guerout, A.M., Biskri, L., Bouige, P. and Mazel, D. (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res*, **13**, 428-442.
25. Boucher, Y., Labbate, M., Koenig, J.E. and Stokes, H.W. (2007) Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol*, **15**, 301-309.
26. Diaz-Mejia, J.J., Amabile-Cuevas, C.F., Rosas, I. and Souza, V. (2008) An analysis of the evolutionary relationships of integron integrases, with emphasis on the prevalence of class 1 integrons in Escherichia coli isolates from clinical and environmental origins. *Microbiology*, **154**, 94-102.
27. Nemergut, D.R., Robeson, M.S., Kysela, R.F., Martin, A.P., Schmidt, S.K. and Knight, R. (2008) Insights and inferences about integron evolution from genomic data. *BMC Genomics*, **9**, 261.
28. Hall, R.M. (2012) Integrons and gene cassettes: hotspots of diversity in bacterial genomes. *Ann N Y Acad Sci*, **1267**, 71-78.
29. Gillings, M., Boucher, Y., Labbate, M., Holmes, A., Krishnan, S., Holley, M. and Stokes, H.W. (2008) The evolution of class 1 integrons and the rise of antibiotic resistance. *J Bacteriol*, **190**, 5095-5100.
30. Dobrindt, U., Hochhut, B., Hentschel, U. and Hacker, J. (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*, **2**, 414-424.
31. Guglielmini, J., Quintais, L., Garcillan-Barcia, M.P., de la Cruz, F. and Rocha, E.P. (2011) The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet*, **7**, e1002222.
32. Hochhut, B., Lotfi, Y., Mazel, D., Faruque, S.M., Woodgate, R. and Waldor, M.K. (2001) Molecular analysis of antibiotic resistance gene clusters in vibrio cholerae O139 and O1 SXT constins. *Antimicrob Agents Chemother*, **45**, 2991-3000.
33. Iwanaga, M., Toma, C., Miyazato, T., Insisiengmay, S., Nakasone, N. and Ehara, M. (2004) Antibiotic resistance conferred by a class I integron and SXT constin in Vibrio cholerae O1 strains isolated in Laos. *Antimicrob Agents Chemother*, **48**, 2364-2369.

34. Gillings, M.R., Holley, M.P., Stokes, H.W. and Holmes, A.J. (2005) Integrons in Xanthomonas: a source of species genome diversity. *Proc Natl Acad Sci U S A*, **102**, 4419-4424.
35. Rowe-Magnus, D.A., Guerout, A.M., Ploncard, P., Dychinco, B., Davies, J. and Mazel, D. (2001) The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc Natl Acad Sci U S A*, **98**, 652-657.
36. Holmes, A.J., Gillings, M.R., Nield, B.S., Mabbutt, B.C., Nevalainen, K.M. and Stokes, H.W. (2003) The gene cassette metagenome is a basic resource for bacterial genome evolution. *Environ Microbiol*, **5**, 383-394.
37. Moura, A., Henriques, I., Ribeiro, R. and Correia, A. (2007) Prevalence and characterization of integrons from bacteria isolated from a slaughterhouse wastewater treatment plant. *The Journal of antimicrobial chemotherapy*, **60**, 1243-1250.
38. Stalder, T., Barraud, O., Casellas, M., Dagot, C. and Ploy, M.C. (2012) Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol*, **3**, 119.
39. Gillings, M.R., Gaze, W.H., Pruden, A., Smalla, K., Tiedje, J.M. and Zhu, Y.G. (2015) Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *ISME J*, **9**, 1269-1279.
40. Koenig, J.E., Boucher, Y., Charlebois, R.L., Nesbo, C., Zhaxybayeva, O., Bapteste, E., Spencer, M., Joss, M.J., Stokes, H.W. and Doolittle, W.F. (2008) Integron-associated gene cassettes in Halifax Harbour: assessment of a mobile gene pool in marine sediments. *Environ Microbiol*, **10**, 1024-1038.
41. Elsaied, H., Stokes, H.W., Kitamura, K., Kurusu, Y., Kamagata, Y. and Maruyama, A. (2011) Marine integrons containing novel integrase genes, attachment sites, attI, and associated gene cassettes in polluted sediments from Suez and Tokyo Bays. *ISME J*, **5**, 1162-1177.
42. Tsafnat, G., Coiera, E., Partridge, S.R., Schaeffer, J. and Iredell, J.R. (2009) Context-driven discovery of gene cassettes in mobile integrons using a computational grammar. *BMC Bioinformatics*, **10**, 281.
43. Eddy, S.R. and Durbin, R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res*, **22**, 2079-2088.
44. Nawrocki, E.P. and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933-2935.
45. Neron, B., Menager, H., Maufrais, C., Joly, N., Maupetit, J., Letort, S., Carrere, S., Tuffery, P. and Letondal, C. (2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005-3011.
46. Moura, A., Soares, M., Pereira, C., Leitao, N., Henriques, I. and Correia, A. (2009) INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics*, **25**, 1096-1098.
47. Cambray, G., Sanchez-Alberola, N., Campoy, S., Guerin, E., Da Re, S., Gonzalez-Zorn, B., Ploy, M.C., Barbe, J., Mazel, D. and Erill, I. (2011) Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mobile DNA*, **2**, 6.
48. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460-2461.
49. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30**, 772-780.

50. Eddy, S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol*, **7**, e1002195.
51. Gibson, M.K., Forsberg, K.J. and Dantas, G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*, **9**, 207-216.
52. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res*, **36**, D281-288.
53. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422-1423.
54. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W. and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
55. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647-1649.
56. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792-1797.
57. Criscuolo, A. and Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol Biol*, **10**, 210.
58. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, **32**, 268-274.
59. Oliveira, P.H., Touchon, M. and Rocha, E.P.C. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research*, **42**, 10618-U10803.
60. Lee, B.M., Park, Y.J., Park, D.S., Kang, H.W., Kim, J.G., Song, E.S., Park, I.C., Yoon, U.H., Hahn, J.H., Koo, B.S. *et al.* (2005) The genome sequence of Xanthomonas oryzae pathovar oryzae KACC10331, the bacterial blight pathogen of rice. *Nucleic Acids Res*, **33**, 577-586.
61. Miele, V., Penel, S. and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
62. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. and Chandler, M. (2006) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res*, **34**, D32-36.
63. Touchon, M. and Rocha, E.P. (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol*, **24**, 969-981.
64. Brown, H.J., Stokes, H.W. and Hall, R.M. (1996) The integrons In0, In2, and In5 are defective transposon derivatives. *J Bacteriol*, **178**, 4429-4437.
65. Cambray, G., Guerout, A.M. and Mazel, D. (2010) Integrons. *Annu Rev Genet*, **44**, 141-166.
66. Silva, F.J., Latorre, A. and Moya, A. (2003) Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet*, **19**, 176-180.
67. Canback, B., Tamas, I. and Andersson, S.G. (2004) A phylogenomic study of endosymbiotic bacteria. *Mol Biol Evol*, **21**, 1110-1122.

68. McCutcheon, J.P. and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol*, **10**, 13-26.
69. Ochman, H., Lawrence, J.G. and Groisman, E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299-304.
70. Cordero, O.X. and Hogeweg, P. (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci U S A*, **106**, 21748-21753.
71. Bikard, D., Julie-Galau, S., Cambray, G. and Mazel, D. (2010) The synthetic integron: an in vivo genetic shuffling device. *Nucleic Acid Res*, **38**, e153.
72. Baltrus, D.A. (2013) Exploring the costs of horizontal gene transfer. *Trends Ecol Evol*, **28**, 489-495.
73. Mazodier, P. and Davies, J. (1991) Gene transfer between distantly related bacteria. *Annu Rev Genet*, **25**, 147-171.
74. Kloesges, T., Popa, O., Martin, W. and Dagan, T. (2011) Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. *Mol Biol Evol*, **28**, 1057-1074.
75. Guglielmini, J., de la Cruz, F. and Rocha, E.P. (2013) Evolution of conjugation and type IV secretion systems. *Mol Biol Evol*, **30**, 315-331.
76. Nandi, S., Maurer, J.J., Hofacre, C. and Summers, A.O. (2004) Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc Natl Acad Sci U S A*, **101**, 7118-7122.
77. Salah, P., Bisaglia, M., Aliprandi, P., Uzan, M., Sizun, C. and Bontems, F. (2009) Probing the relationship between Gram-negative and Gram-positive S1 proteins by sequence analysis. *Nucleic Acids Res*, **37**, 5578-5588.
78. Rowe-Magnus, D.A. and Mazel, D. (1999) Resistance gene capture. *Curr Opin Microbiol*, **2**, 483-488.
79. Le Roux, F., Zouine, M., Chakroun, N., Binesse, J., Saulnier, D., Bouchier, C., Zidane, N., Ma, L., Rusniok, C., Lajus, A. *et al.* (2009) Genome sequence of Vibrio splendidus: an abundant planctonic marine species with a large genotypic diversity. *Environ Microbiol*, **11**, 1959-1970.
80. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*, **14**, 1188-1190.

Figure Legends

Figure 1. Schema of an integron and the three types of elements detected by IntegronFinder. (A) The integron is composed of a specific integron integrase gene (*intI*, orange), an *attI* recombination site (red), and an array of gene cassettes (blue, yellow and green). A cassette is composed of an ORF flanked by two *attC* recombination sites. The integron integrase has its own promoter (PintI). There is one constitutive promoter (Pc) for the array of cassettes. Cassettes rarely contain promoters. The integrase can excise a cassette (1) and/or integrate it at the *attI* site (2). (B) Complete integrons include an integrase and at least one *attC* site. (C) The In0 elements are composed of an integron integrase and no *attC* sites. (D) The attC0 elements are composed of at least two *attC* sites (but no integrase).

Figure 2. Diagram describing the different steps used by IntegronFinder to identify and annotate integrons. Solid lines represent the default mode, dotted lines optional modes. Blue boxes indicate the main dependency used for a given step. Green boxes indicate the format of the file needed for a given step.

Figure 3. Characteristics of the *attC* sites. (A) Scheme of the secondary structure of a folded *attC* site. EHB stands for Extra Helical Bases. (B) Analysis of the *attC* sites used to build the model, including the WebLogo (80) of the R and L box and unpaired central spacers (UCS) and the histogram (and kernel density estimation) of the size of the variable terminal structure (VTS). The Weblogo represents the information contained in a column of a multiple sequence alignment (using the log2 transformation). The taller the letter is, the more conserved is the character at that position. The width of each column of symbols takes into account the presence of gaps. Thin columns are mostly composed of gaps. (C) Same as (B) but with the set of *attC* sites identified in complete integrons found in complete bacterial genomes. (D) Secondary structure used in the model in WUSS format, colors match those of (A).

Figure 4. Quality assessment of the *attC* sites covariance model on pseudo-genomes with varying G+C content and depending on the run mode (default and local_max). (Top) Table resuming the results. The mean time is the average running time per pseudo-genome on a Mac Pro, 2 x 2.4 GHz 6-Core Intel Xeon, 16 Gb RAM, with options --cpu 20 and --no-proteins. (Middle) Rate of false positives per mega-base (Mb) as function of the G+C content. The same result is obtained for both modes (default and --local_max). (Bottom) Sensitivity (or true positive rate) as function of the G+C content. The red line depicts results obtained with the default parameters, and the blue line represents results obtained with the accurate parameters (--local_max option).

Figure 5. Taxonomic distribution of integrons in clades with more than 50 complete genomes sequenced. The grey bar represents the number of genomes sequenced for a given clade. The blue bar represents the number of complete integrons, the red bar the number of In0, and the yellow bar the number of attC0.

Figure 6. Frequency of integrons and related elements as a function of the genome size.

Figure 7. Histogram of the distribution of the number of *attC* sites per integron. Mobile integrons are depicted in dark orange; non-mobile integrons (persistent in more than 60% of the genomes of a species) are depicted in green; undetermined elements are depicted in grey. The largest mobile integron has 15 *attC* sites.

Figure 8. Comparison of *attC* sites. A column represents comparisons of *attC* sites between (inter) or within (intra) element(s) (integron complete or attC0) depending on the type of complete integron (mobile or persistent) or attC0 element. (A) Distribution of the sequence distance between two R-UCS-L boxes of two *attC* sites. (B) Distribution of the difference in VTS size. Other-inter integron comparisons are shown in Figure S7. Mann-Whitney rank tests: ***: $p < 10^{-6}$; **: $p < 10^{-3}$; * : $p < 0.05$; ns: $p > 0.05$. ND: Not determined.

Figure 9. Relationship between the number of *attC* sites in an integron and the mean sequence distance between *attC* sites within an integron. The x-axis is in log10 scale. The association is significant: spearman rho = -0.58, $P < 0.001$.

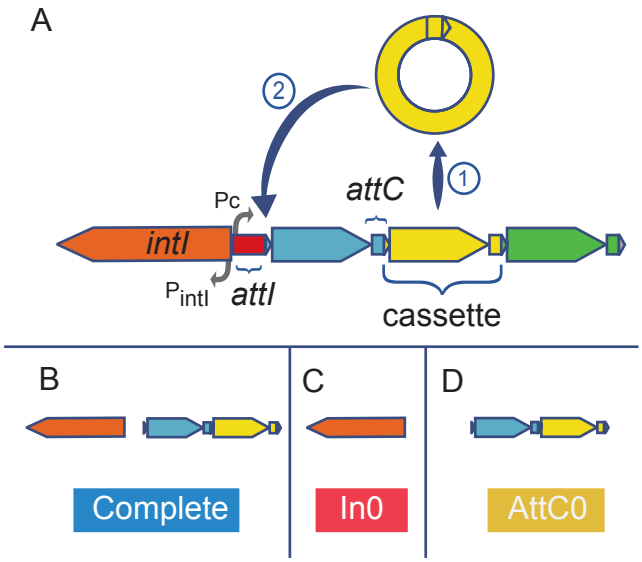

Figure 1

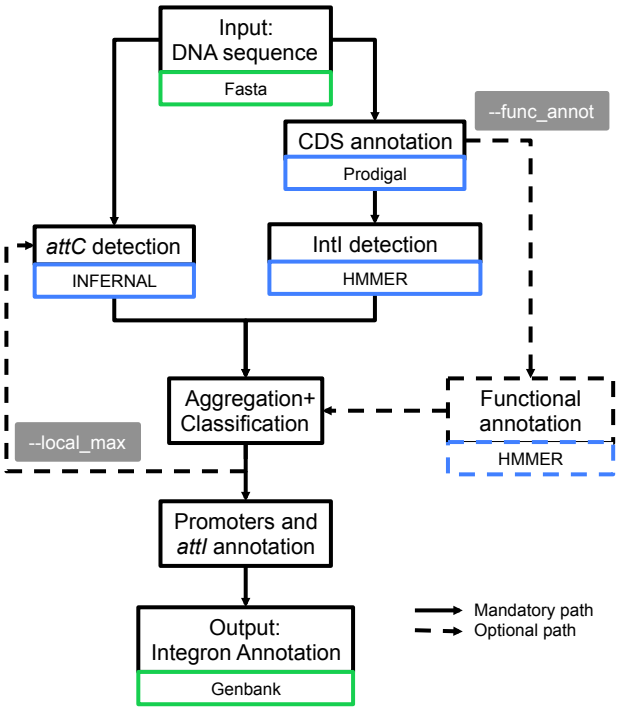
Input:
DNA sequence
Fasta
--func_annot
CDS annotation
Prodigal
attC detection
INFERNAL
IntI detection
HMMER
Aggregation+
Classification
Functional
annotation
HMMER
--local_max
Promoters and
attI annotation
Output:
Integron Annotation
Genbank
Mandatory path
Optional path

Figure 2

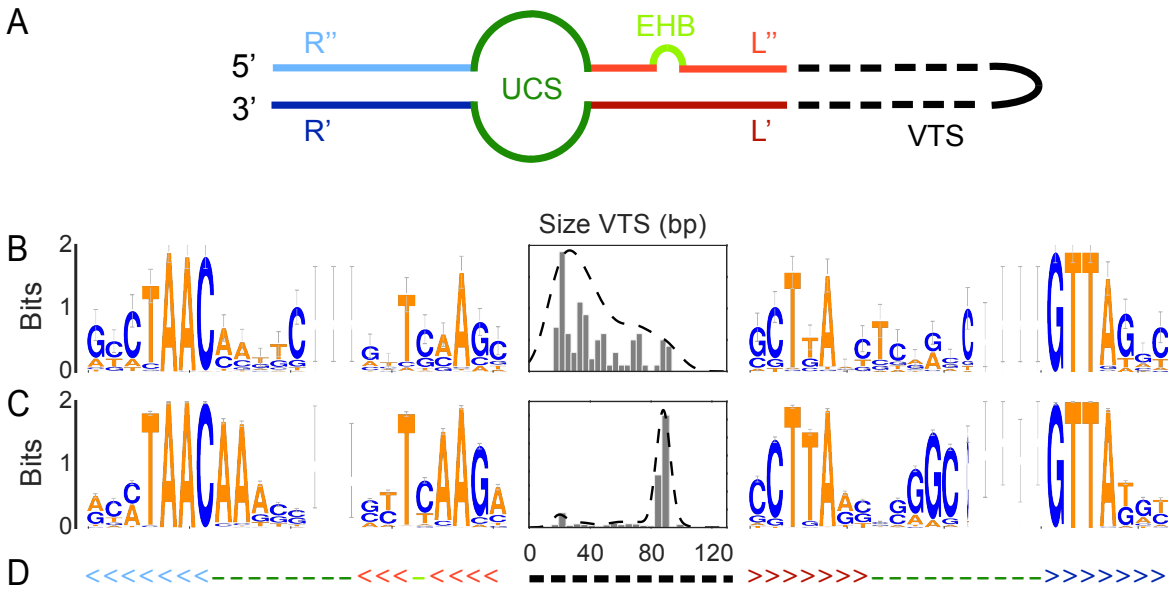

Figure 3

	Default	local_max
Sensitivity	60.43%	82.91%
FP rate	0.02 FP/Mb	0.02 FP/Mb
Mean time	2.09 s	56.24 s

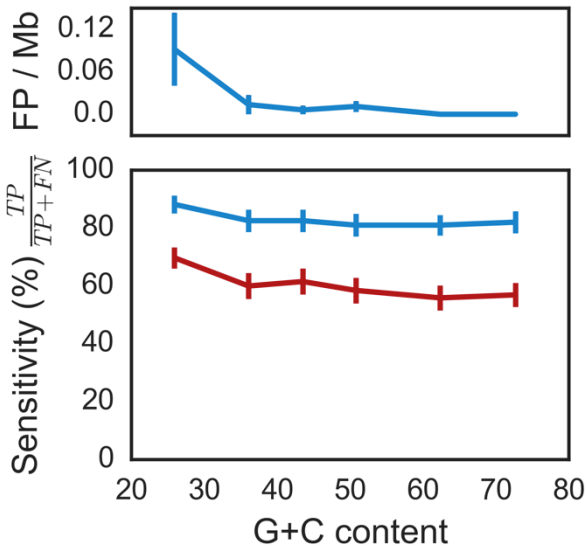

Figure 4

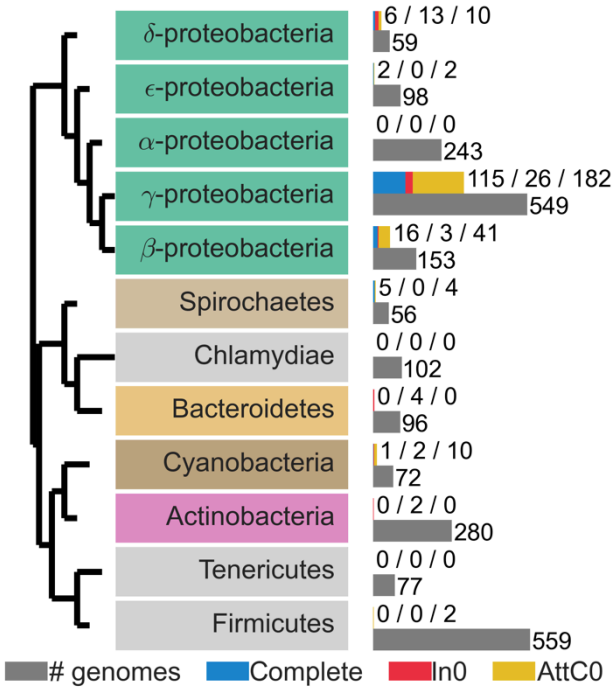
δ-proteobacteria
6 / 13 / 10
59
ϵ-proteobacteria
2 / 0 / 2
98
α-proteobacteria
0 / 0 / 0
243
γ-proteobacteria
115 / 26 / 182
549
β-proteobacteria
16 / 3 / 41
153
Spirochaetes
5 / 0 / 4
56
Chlamydiae
0 / 0 / 0
102
Bacteroidetes
0 / 4 / 0
96
Cyanobacteria
1 / 2 / 10
72
Actinobacteria
0 / 2 / 0
280
Tenericutes
0 / 0 / 0
77
Firmicutes
0 / 0 / 2
559
genomes
Complete
In0
AttC0

Figure 5

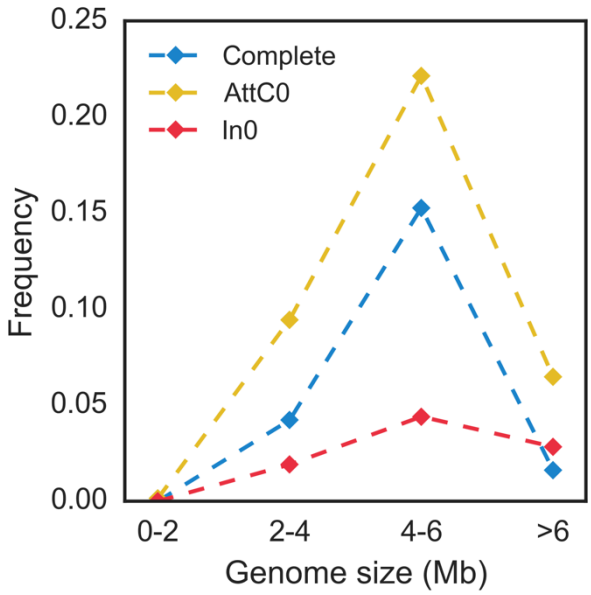
0.25
0.20
0.15
0.10
0.05
0.00
Frequency
Complete
AttC0
In0
0-2
2-4
4-6
>6
Genome size (Mb)

Figure 6

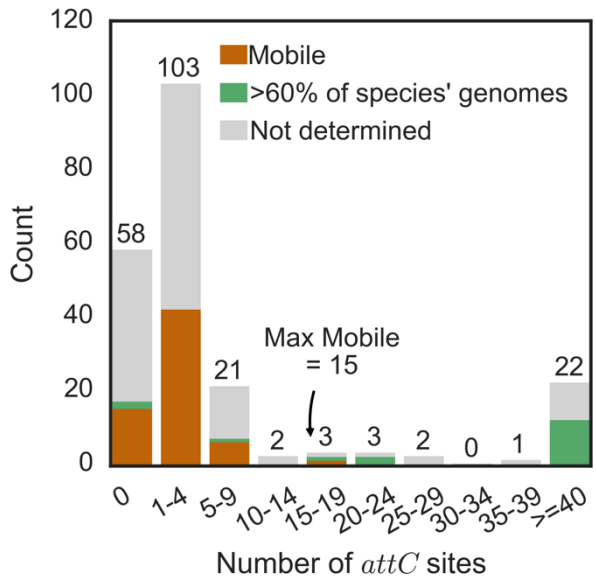
120
100
80
60
40
20
0
Count
Mobile
>60% of species' genomes
Not determined
58
103
21
2
3
3
2
0
1
22
Max Mobile
= 15
0
1-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
>=40
Number of attC sites

Figure 7

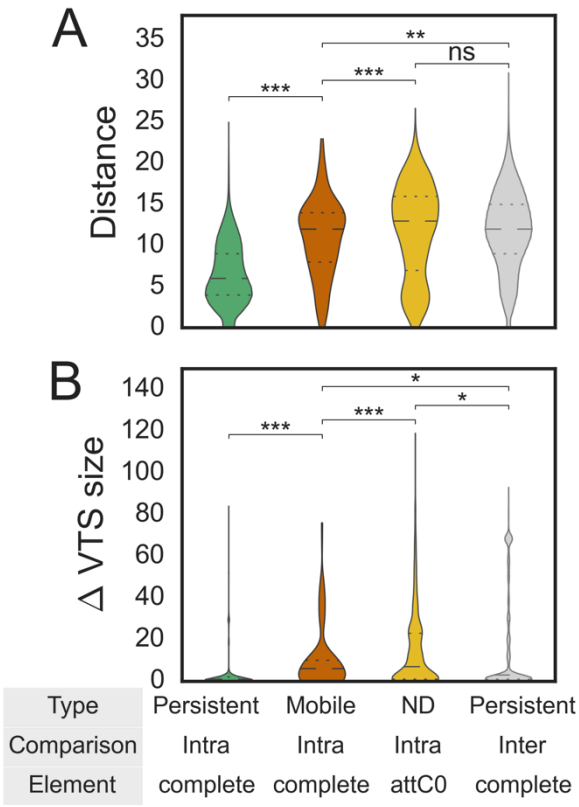
A
Distance
35
30
25
20
15
10
5
0
***
***
**
ns
B
Δ VTS size
140
120
100
80
60
40
20
0
***
***
*
*
Type
Persistent
Mobile
ND
Persistent
Comparison
Intra
Intra
Intra
Inter
Element
complete
complete
attC0
complete

Figure 8

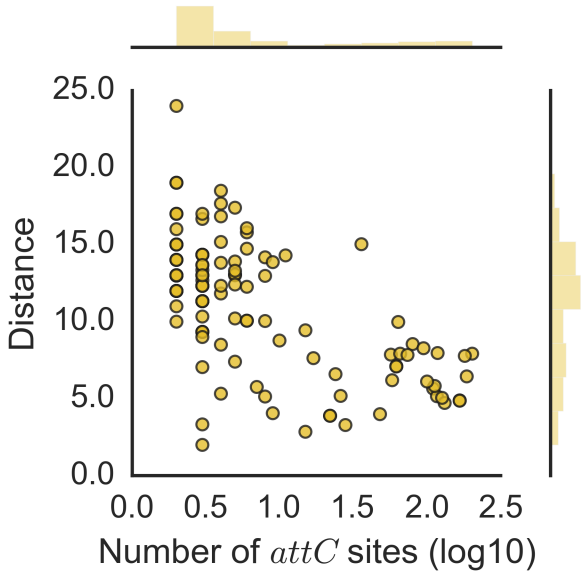
25.0
20.0
15.0
10.0
5.0
0.0
Distance
0.0
0.5
1.0
1.5
2.0
2.5
Number of $attC$ sites (log10)

Figure 9

Supplementary material for:

Automatic and accurate identification of integrons and cassette arrays in bacterial genomes reveals unexpected patterns.

Jean Cury[1,2], Thomas Jové[3], Marie Touchon[1,2], Bertrand Néron[4], Eduardo PC Rocha[1,2]

[1] Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France

[2] CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France

[3] Univ. Limoges, INSERM, CHU Limoges, UMR_S 1092, F-87000 Limoges, France.

[4] Centre d'Informatique pour la Biologie, C3BI, Institut Pasteur, Paris, France

Table of contents

Figure S1 – Phylogenetic tree of tyrosine recombinases.
Figure S2 – Phylogenetic tree of IntI.
Figure S3 – Distribution of hits of PF00589 and intI_Cterm in function of each other and the existence of neighboring attC sites.
Figure S4 – Taxonomic distribution of integrons in clades with less than 50 genomes fully sequenced.
Figure S5 – Frequency of integrons and related elements as a function of the genome size when the analysis is restricted to Gammaproteobacteria.
Figure S6 – Histogram of the distribution of the number of attC sites in attC0 elements.
Figure S7 – All comparisons between attC sites.

Table S1 – Sequence of the promoter of the integrase (Pint), the promoter of the cassette (Pc), and of the attI site in 3 classes of integron when available.
Table S2 – List of genome used to create the 6 pseudo-genomes with different GC% background composition.
Table S3 – List of IntI protein from tree of figure S2 with the data (attached as a text file).
Table S4 – List of all integrons identified in bacterial genomes per element (attached as text file)
Table S5 – List of all integrons identified in bacterial genomes per integron (attached as text file)
Table S6 – List of all integrons identified in bacterial genomes per genome (attached as text file)

Table S7a/b – List of expected (a) and observed (b) position of the 596 *attC* sites from INTEGRALL (attached as text file)

File S1 – Covariance Model of the *attC* site.

Tree S1 – Phylogenetic tree of tyrosine recombinases, corresponding to Figure S1 (attached as a nexus file).
Tree S2 – Phylogenetic tree of integron integrases, corresponding to Figure S2 (attached as a newick file)

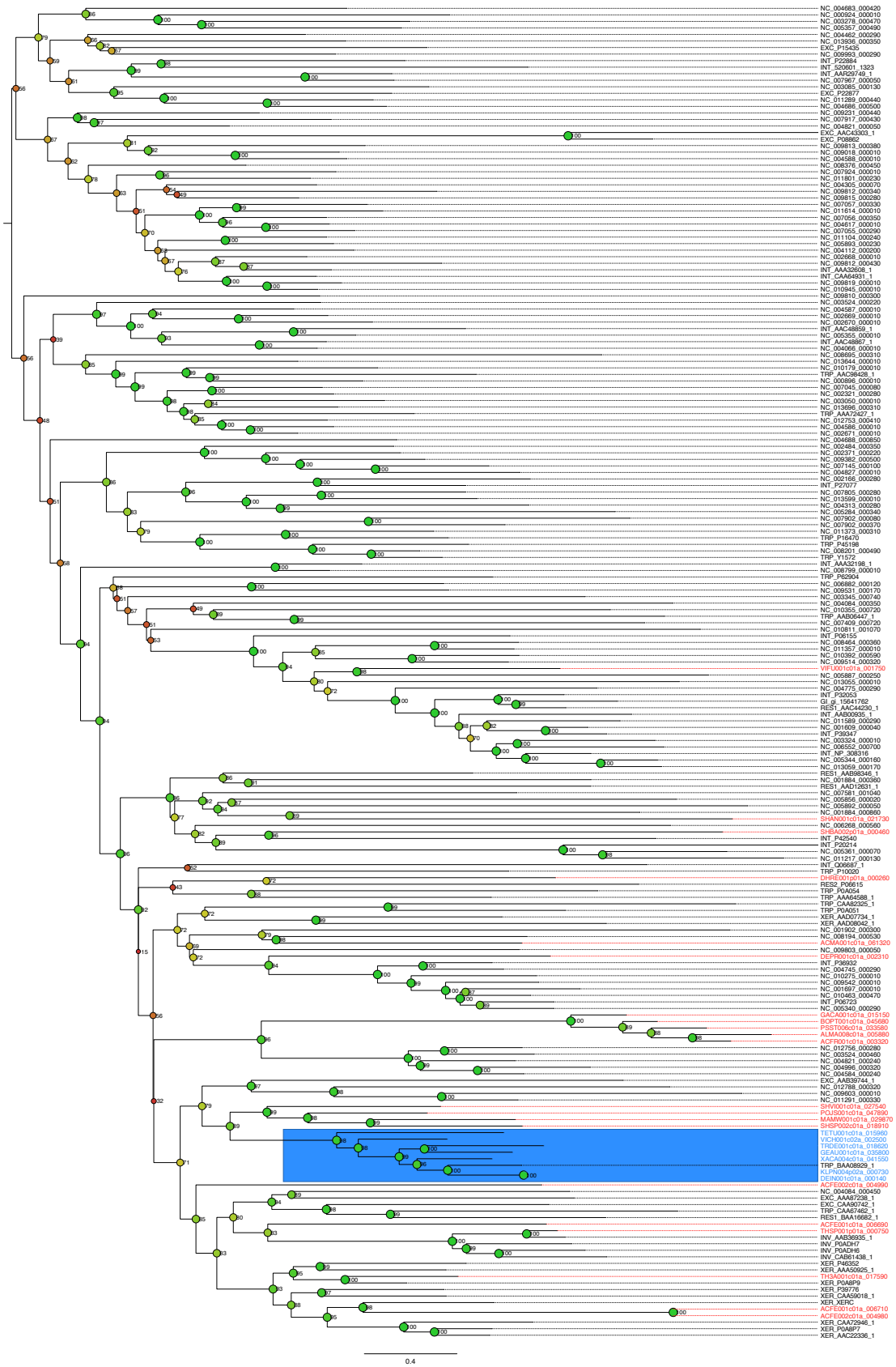

Figure S1 – Phylogenetic tree of tyrosine recombinases including the 21 proteins, which match the profile PF00589 but not intI_Cterm (red) and IntI from complete integron (blue).

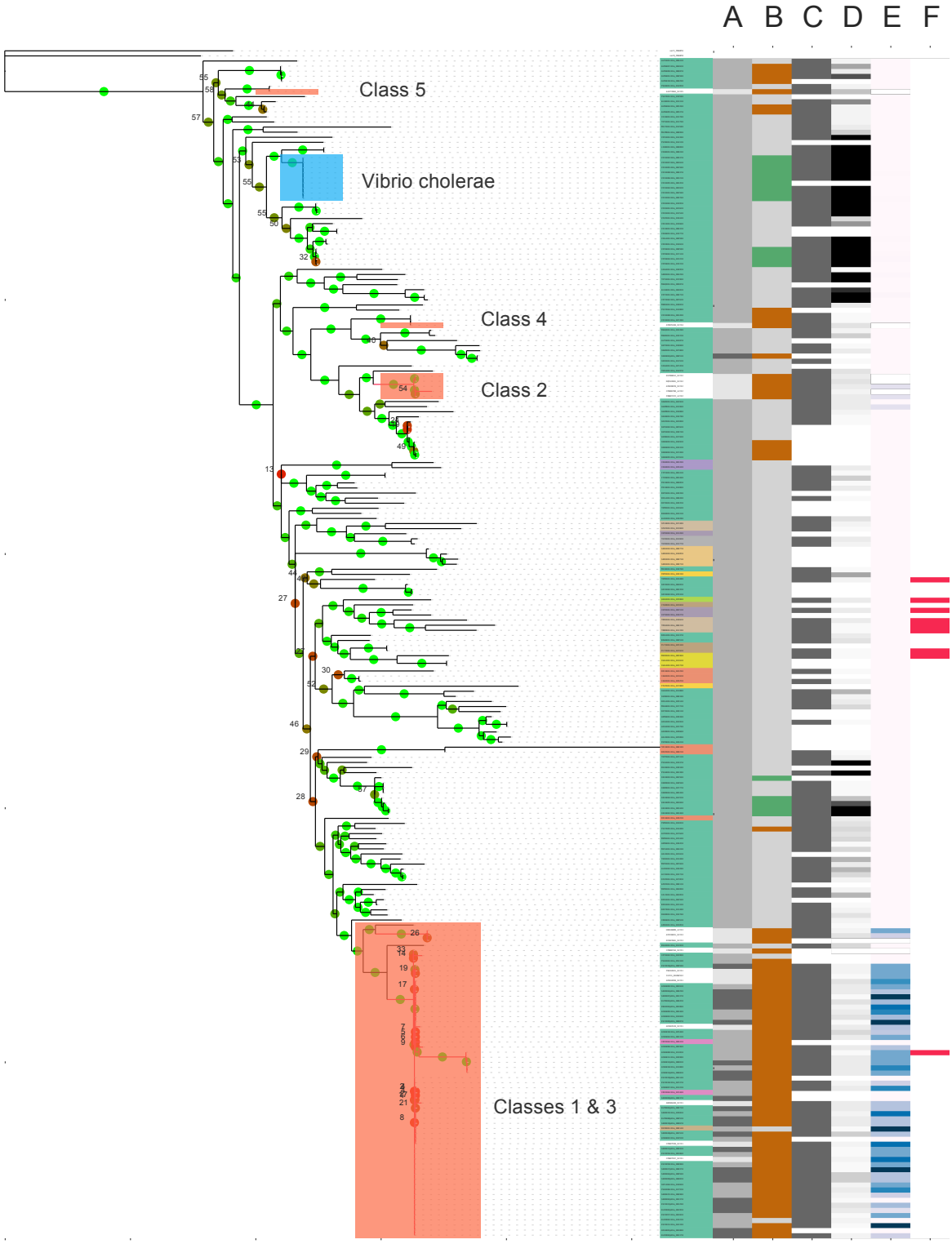

Figure S2 - Phylogenetic tree of IntI. The tree was made using IQ-Tree on the multiple alignments of the integrases carrying the two profiles (intI_Cterm and PF00589) using XerC and XerD as outgroups (see Methods for details). Bootstrap values are indicated if lower than 60%. The matrix on the right represents traits associated with the integron. Column (A) represents whether the corresponding

integron is on a plasmid (dark grey) or on a chromosome (grey). Lighter grey corresponds to integrons of classes 1 to 5, but whose replicon types are unknown (unavailable genomes). The column (B) represents whether the corresponding integron is persistent (green) or mobile (dark orange), or not determinable (grey). Column (C) represents the integrases associated with at least one attC site (grey). The column (D) represents the number of *attC* sites with a gradient of grey (the darker the more *attC* sites, with saturation from 20 attC sites). The column (E) represents the frequency of resistance genes among cassettes (darker blue indicates higher values). The last column (F) represents the integron with inverted integrase (red). The leaves of the tree are colored according to their clade. For the names corresponding to the colors, see Figure 5 and Figure S4. See Table S3 for full data.

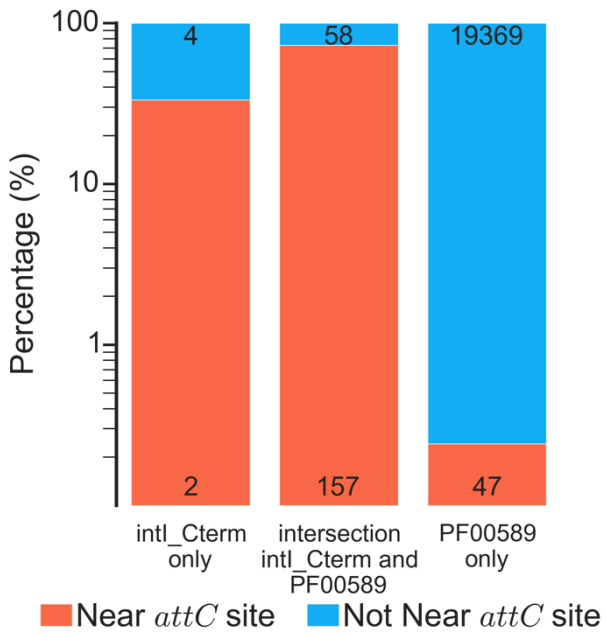

Figure S3 - Distribution of hits of PF00589 and intI_Cterm in function of each other and the existence of neighboring *attC* sites. The y-axis represents percentage in a log10 scale. Numbers on the bar represent the actual quantity for the underlying bar.

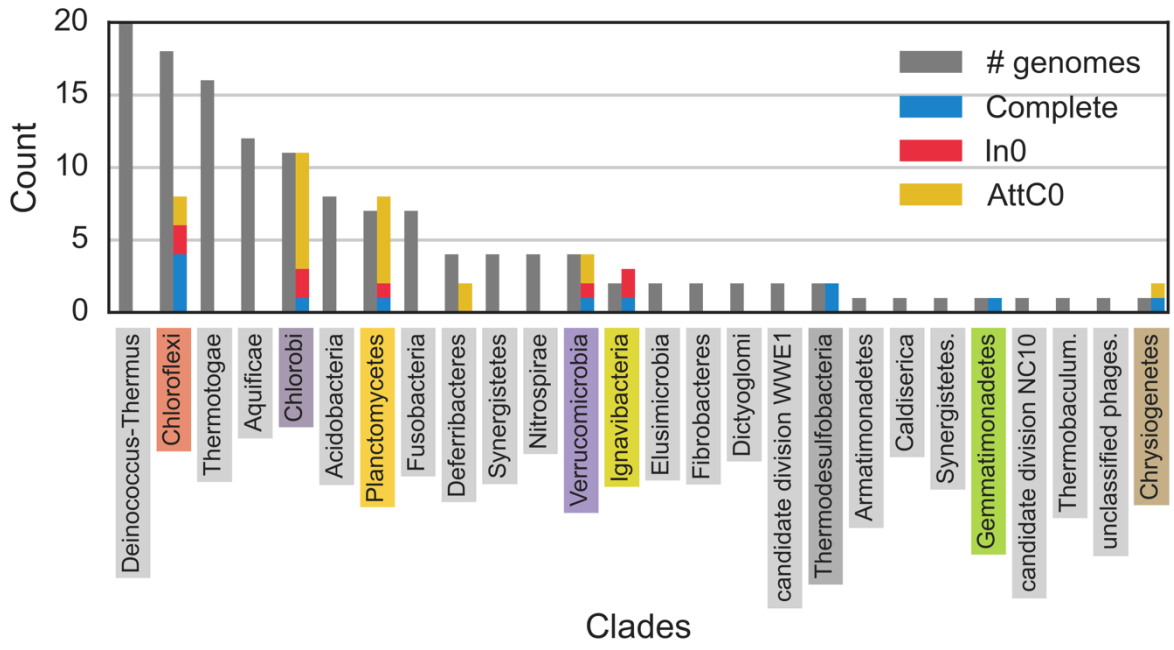

Figure S4 – Taxonomic distribution of integrons in clades with less than 50 genomes available in our dataset. The grey bar represents the number of genome sequenced for a given clade. The blue bar represents the number of complete integron, the red bar number of In0 and the yellow bar the number of attC0.

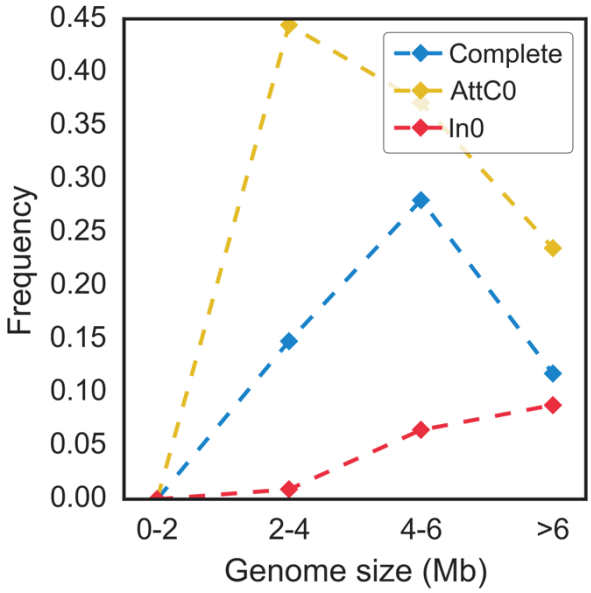

Figure S5 - Frequency of integrons and related elements as a function of the genome size when the analysis is restricted to Gammaproteobacteria.

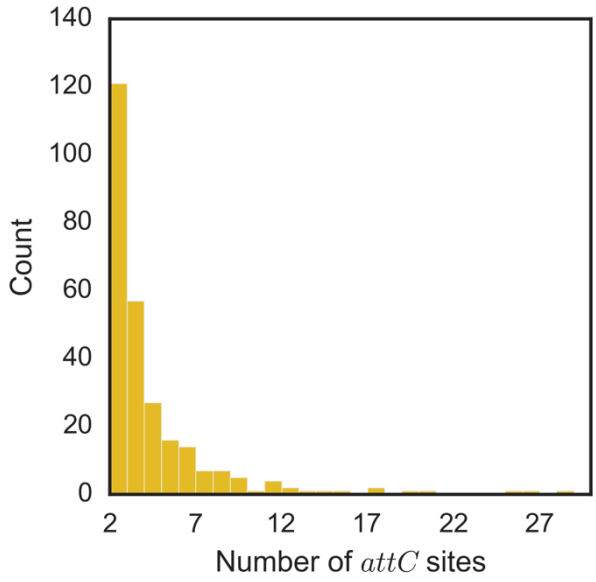

Figure S6 - Histogram of the distribution of the number of *attC* sites in attC0 elements.

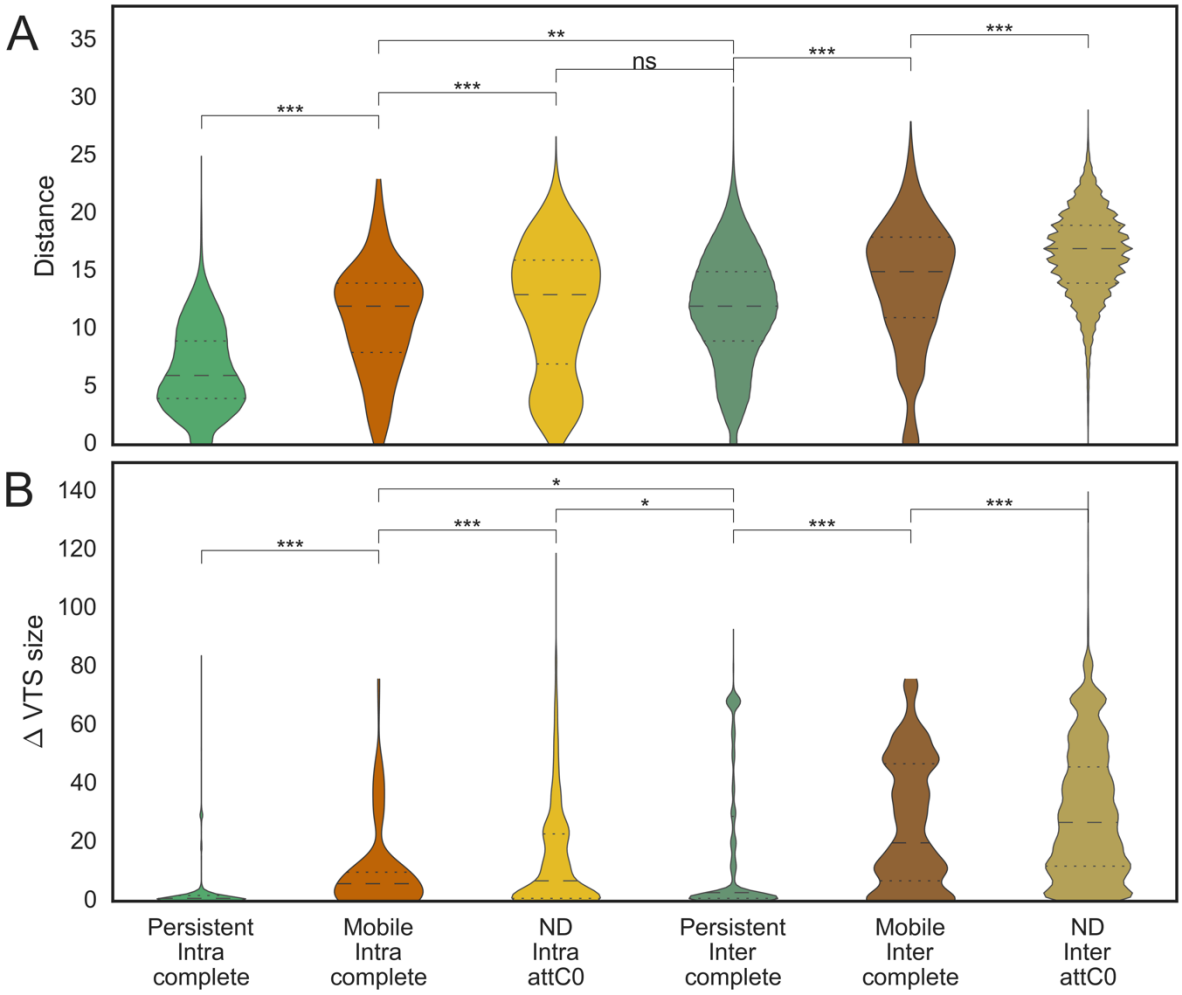

Figure S7 Comparison of *attC* sites. A column represents comparisons of *attC* sites between (inter) or within (intra) element(s) (integron complete or attC0) depending on the type of element (mobile or persistent). (A) Distribution of the sequence distance between two R-UCS-L boxes of two *attC* sites. (B) Distribution of the difference in VTS size. ***: $p < 10^{-6}$; **: $p < 10^{-3}$; * : $p < 0.05$; ns: $p > 0.05$, Mann-Whitney rank test.

Table S1 – Sequence of the promoter of the integrase (Pint), the promoter of the cassette (Pc), and of the attI site in 3 classes of integron when available. Sequences were provided by INTEGRALL. Square brackets indicate that it can be one of the letters at that position. Curly brackets indicate that the square brackets before is repeated a number of times comprised between a minimum and a maximum.

Class	P_{int}	P_c	*attI*
1	TTGCTGCTTGGATGCCCGAGGCATAGACTGTACA	T[GT]G[AG][CT]ATAAGCCTGTTCGGTT[CG]GT[AG]A[AG]CTGTAATCGCA	TGATGTTATGGAGCAGCAACGATGTTACGCAGCAGGGCAGTCGCCCTAAAACAAAGTT
		TTGTTATGACTGTTTTTTT[G-]{1,4}[GT]ACA[GCA][AT]	
2	ND	ND	TTAATTAACGGTAAGCATCAGCGGGTGACAAAACGAGCATGCTTACTAATAAAATGTT
3	ND	TAGACATAAGCTTTCTCGGTCTGTAGG[CA]TGTAATG	CTTTGTTTAACGACCACGGTTGTGGGTATCCGGTGTTTGGTCAGATAAACCACAAGTT

Table S2 – List of genome used to create the 6 pseudo-genomes with different GC% background composition.

Genome	Size (pb)	%GC
Mycoplasma hyorhinis GDL-1	837480	25.91
Anaerococcus prevotii DSM 20548	1883067	36.07
Bacilllus subtilis subsp. subtilis str. 168	4215606	43.51
Escherichia coli str. K-12 substr. MG1655	4639675	50.79
Arthrobacter aurescens TC1	4597686	62.34
Clavibacter michiganensis subsp. michiganensis NCPPB 382	3297891	72.66