Integrative and conjugative elements and their hosts: composition, distribution and organization

Jean Cury, Marie Touchon, Eduardo P. C. Rocha

Published online 13 July 2017

Nucleic Acids Research, 2017, Vol. 45, No. 15 ***8943–8956***
doi: 10.1093/nar/gkx607

Integrative and conjugative elements and their hosts: composition, distribution and organization

Jean Cury[1,2,*], Marie Touchon[1,2] and Eduardo P. C. Rocha[1,2]

[1]Microbial Evolutionary Genomics, Institut Pasteur, 28, rue du Dr Roux, Paris 75015, France and [2]CNRS, UMR3525, 28, rue Dr Roux, Paris 75015, France

Received April 20, 2017; Revised June 30, 2017; Editorial Decision July 04, 2017; Accepted July 04, 2017

ABSTRACT

Conjugation of single-stranded DNA drives horizontal gene transfer between bacteria and was widely studied in conjugative plasmids. The organization and function of integrative and conjugative elements (ICE), even if they are more abundant, was only studied in a few model systems. Comparative genomics of ICE has been precluded by the difficulty in finding and delimiting these elements. Here, we present the results of a method that circumvents these problems by requiring only the identification of the conjugation genes and the species' pan-genome. We delimited 200 ICEs and this allowed the first large-scale characterization of these elements. We quantified the presence in ICEs of a wide set of functions associated with the biology of mobile genetic elements, including some that are typically associated with plasmids, such as partition and replication. Protein sequence similarity networks and phylogenetic analyses revealed that ICEs are structured in functional modules. Integrases and conjugation systems have different evolutionary histories, even if the gene repertoires of ICEs can be grouped in function of conjugation types. Our characterization of the composition and organization of ICEs paves the way for future functional and evolutionary analyses of their cargo genes, composed of a majority of unknown function genes.

INTRODUCTION

Bacterial diversification occurs rapidly by the constant influx of exogenous DNA by horizontal gene transfer (HGT) (1–3). As a consequence, the diversity of genes found in the strains of a species, its pangenome, is usually much higher than the number of genes found in a single bacterial genome at a given time (4). The pangenome represents a huge reservoir of potentially adaptive genes, whose potential has become evident in the rapid spread of antibiotic resistance in the last decades (5), and in the emergence of novel pathogens (6). Mobile genetic elements (MGE) drive the spread of genes in populations using a variety of mechanisms, often encoded by the elements themselves.

Conjugative MGEs carry between a few dozens to many hundreds of genes (7). They can be extra-chromosomal (plasmids) or integrative (ICEs). Conjugation requires an initial step of cell-to-cell contact during mating pair formation (MPF). The mechanism is the same for plasmids and ICEs, apart from the initial and final steps of chromosomal excision and integration (8,9). The mechanism of transfer proceeds in three steps. Initially, the relaxase (MOB) nicks the DNA at the origin of transfer (*oriT*), and binds covalently to one of the DNA strands. The nucleoprotein filament is then coupled to the type 4 secretion system (T4SS) and transferred to the recipient cell. Finally, the element is replicated in the original and novel hosts leading to double stranded DNA molecules in each cell. One should note that some integrative elements are transferred using another mechanism, also called conjugation, relying on double-stranded DNA. They are restricted to certain Actinobacteria, have been recently described in detail (10,11), and will not be mentioned in this work. Integration of ICEs is usually mediated by a Tyrosine recombinase (integrase), but some ICEs use Serine or DDE recombinases instead (9,12–14). There are eight types of MPF (15,16), each based on a model system as described in Figure 1. Among those, six MPF types (B, C, F, G, I, T) are specific to diderms, i.e. bacteria with an outer membrane (typically gram negative), while two others (FA, FATA) are specific to monoderms, i.e. bacteria lacking an outer membrane (typically gram positive). Phylogenetic analyses suggest that ssDNA conjugation evolved initially in diderms and was then transferred to monoderms (15). The identification of a valid conjugative system requires the presence of a relaxase, of the VirB4 ATPase, a coupling ATPase (T4CP), and several other proteins that may differ between types (although they are sometimes distant homologs or structural analogs) (16). Both ICEs and plasmids can be of any MPF type, but some preferential associations have been observed: there are few plasmids of MPF_G and few ICEs of MPF_I and MPF_F (17).

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

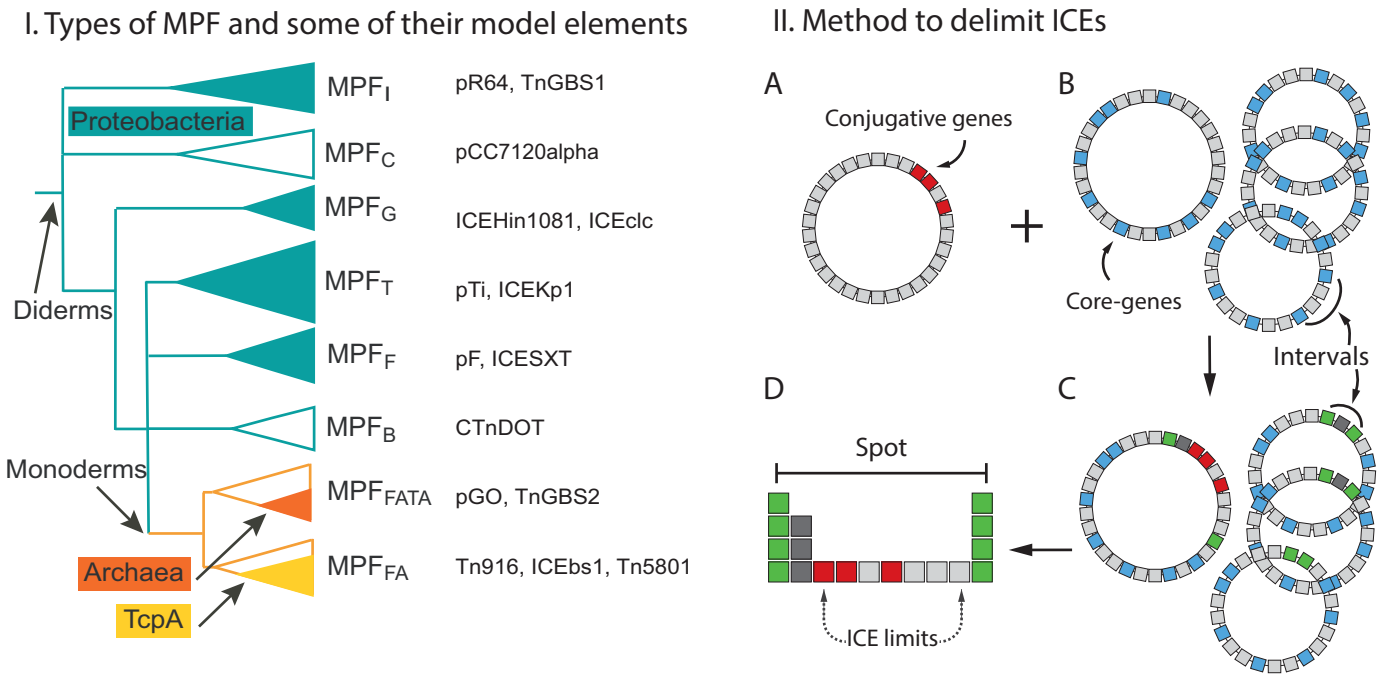

Figure 1. Mating Pair Formation (MPF) types and procedure for ICE delimitation. (**I**) The phylogenetic tree displays the evolutionary relationships between MPF types as given by the VirB4 phylogeny. Most lineages are from diderms (green branches), the systems from monoderms (yellow branches, including Firmicutes, Actinobacteria, Archaea, and Tenericutes) being derived from these. MPF$_B$ (Bacteroides) and MPF$_C$ (Cyanobacteria) were absent from our data because not enough genomes were sequenced in those clades. The full green clades indicate systems that are typically found in Proteobacteria. The label TcpA indicates a clade that uses this protein as T4CP (an homologous ATPase from the typical T4CP – VirD4). In front of each tip of the tree, we indicate a non-exhaustive list of well-known conjugative elements (ICE (starting with 'ICE', or '(C)Tn') or plasmid (starting with 'p')) for each MPF type. The phylogenetic tree was adapted from (15). (**II**) Scheme of the method. Boxes represent genes, circles represent chromosomes. (**A**) Genes encoding conjugative systems (Red) were detected in bacterial genomes using MacSyFinder. At this stage, this indicates the presence of an ICE that remains to be delimited. (**B**) We restricted the dataset of ICEs to those present in the 37 species for which we had at least four genomes (and a chromosomal conjugative system). We built the core genome (core-genes are represented in blue) of each species. The regions between two consecutive core-genes are defined as an interval. (**C**) The information on the conjugative system and the core-genes is used to delimit the chromosomal interval harboring the ICE. Hence, two core genes flank the ICE (in green). They define an upper bound for its limits. (**D**) Representation of the spot. The two families of core genes (green) define intervals in several genomes of the species (typically in all of them). The set of such intervals is called a spot and is here represented from the point of view of the interval that contains an ICE. We built the spot pan-genome, i.e. we identified the gene families present in the spot, and mapped this information on the interval with the conjugative system. Hence, the bottom layer of genes represents the genes of the interval with the ICE. The upper layers represent other genomes (each layer represents one genome), and the boxes correspond to genes that are orthologs of the genes in the interval with the ICE (genes lacking orthologs are omitted to simplify the representation). Finally, the manual delimitation is based on a visual representation of the spot including this information and the G+C content (see Supplementary Figure S1 and Materials and Methods).

ICEs are more numerous than conjugative plasmids among sequenced genomes (17), but their study is still in the infancy. Beyond the fact that they were discovered more recently, the extremities of ICEs are difficult to delimit precisely in genomic data. Hence, most data available on the biology of ICEs comes from a small number of experimental models, such as the ICE SXT of the MPF$_F$ type, ICEclc (MPF$_G$), MlSym^{R7A} (MPF$_T$), Tn916 and ICEBs1 (MPF$_{FA}$), CTnDOT (MPF$_B$) (18–21). Several of these elements encode traits associated with pathogenicity, mutualism, or the spread of antibiotic resistance, which spurred the initial interest on ICEs. It has been suggested that this has biased the study of ICE biology and that analyses with fewer *a priori* are needed to appreciate the evolutionary relevance of these elements (9). Some ICEs encode mechanisms typically found in plasmids, such as replication and partition (22–25), and phylogenetic studies showed that interconversions between ICE and conjugative plasmids were frequent in the evolutionary history of conjugation (17). Together, these results suggest that ICE and CP are more similar than previously thought (26). Interestingly, ICEs also encode functions typically associated with other MGEs, such as phage-related recombinases (27), and transposable elements (28).

The unbiased study of the gene repertoires and structural traits of ICEs is important to improve the current knowledge on these elements. Here, we developed a method based on the comparison of multiple genomes in a species to identify, class, and study ICEs in bacteria. The method allowed us to delimit ICEs, analyze their gene content, study the resemblance between elements and their internal organization, and to study their distribution in chromosomes. Our methodology does not use any *a priori* knowledge on the organization of previously known ICEs, apart from requiring the presence of a conjugative system. Hence, it should not be affected by ascertainment biases caused by the use of model systems to identify novel ICEs. We show that it provides a broad view of the diversity of ICEs among bacterial genomes.

MATERIALS AND METHODS

Data

The main dataset used in this study concerns 2484 complete genomes of Bacteria that were downloaded from NCBI RefSeq (http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/), in November 2013. A post-hoc validation dataset was obtained from the same database in November 2016 and included a total of genomes. We used the classification of replicons in plasmids and chromosomes as provided in the GenBank files. We searched for conjugation systems in all replicons of all genomes of the two datasets. Yet, the delimitation of ICEs was restricted to species having at least one chromosomally encoded conjugative system and at least four genomes completely sequenced (37 species, 506 genomes) in the main dataset. The validation dataset was used to delimit novel ICEs of the MPF_T type. These were used for post-hoc validation only. Sequences from experimentally validated ICEs were retrieved form the ICEberg database version 1.0 (http://db-mml.sjtu.edu.cn/ICEberg/).

Detection of conjugative systems

Conjugative systems were found with the CONJscan module of MacSyFinder (29), using protein profiles and definitions following a previous work (16) (File S1). Protein profiles are probabilistic models built from the information contained in proteins alignments. They allow more sensitive identification of distant homologs than classical pairwise sequence-search approaches (30). MacSyFinder uses the protein profiles and a set of rules (defined in *models*) about their presence in a given MPF type and their genetic organization. For the latter, we used definitions from previous works from our laboratory: two components of the conjugation genes must be separated by less than 31 genes, an exception being granted for relaxases that can be distant by as much as 60 genes. An element was considered as conjugative when it contained the following components of the conjugative system: VirB4/TraU, a relaxase, a T4CP, and a minimum number of MPF type-specific genes: two for types MPF_{FA} and MPF_{FATA}, or three for the others. MacSyFinder was ran independently for each given MPF type with default parameters (hmmer *e*-value < 0.001, protein profile coverage in the alignment higher than 50%). Conjugative elements of some taxa lack known relaxases, this is the case of some Tenericutes and some Archaea. Since T4SS can be mistaken by protein secretion systems in the absence of relaxases, such systems were excluded from the analysis. The models in CONJscan can be modified by the user. The CONJscan module for MacSyFinder (downloadable at https://github.com/gem-pasteur/Macsyfinder_models) can be used with command lines in a unix-like terminal, or in a webserver (https://galaxy.pasteur.fr, see availability section).

Identification of gene families, core and pan-genomes, spots and intervals

The identification of an ICE at the locus of the conjugative system uses information from comparative genomics. It requires the definition within each species of a core genome, a set of intervals, and a set of spots.

The *core genome* is the set of families of orthologous proteins present in all genomes of the species. We computed the core genome of each species as in (31). Briefly, orthologous genes were identified as the bi-directional best hits (BBH, using global end-gap-free alignments with more than 80% of protein similarity, <20% of difference in length, and having at least four other pairs of BBH hits within a neighborhood of ten genes), and the core genome was defined as the intersection of the pairwise lists of orthologs between genomes using the reference strain as a pivot.

We defined *intervals* as the loci between two consecutive core genes in a genome (Figure 1). If these core genes have no intervening gene, then the interval is empty; otherwise it contains a number of *accessory* genes (i.e. genes not present in the core genome). We defined a spot as the set of intervals flanked by members of the same pair of core gene families (Figure 1). Consider two families of core genes X and Y that are consecutive in all N genomes of a species (i.e. no core gene is between them). Each genome has thus an interval (I_i) at this location that is flanked by the members of X (X_i) and Y (Y_i) in the genome (G_i). The spot i is the set of the intervals I_i in the species. Note that by definition there cannot be more than one interval per spot in a genome. If the region has not endured chromosomal rearrangements (the most typical situation), then the spot will contain as many intervals as the number of genomes. We identified the gene families of the spots that encode at least one ICE (spot pan-genome). For this, we searched for sequence similarity between all proteins in the spot using blastp (version 2.2.15, default parameters). The output was then clustered to identify protein families using Silix (version 1.2.8) (32). Proteins whose alignments had more than 80% identity and at least 80% of coverage were grouped in the same family. The members of the spot pan-genome that were not part of the core genome constituted the accessory genome.

Delimitation of ICEs

We analyzed conjugative systems encoded in chromosomes, and delimited the corresponding ICEs using comparative genomics. There is usually a high turnover of ICEs at the species level (i.e. most elements are present in only a few strains), implicating that a few genomes are usually sufficient to delimit the element by analyzing the patterns of gene presence and absence. We restricted our analysis to species with at least four genomes completely sequenced and assembled (without gaps). ICEs were delimited in two steps. First, we identified the spots encoding conjugative systems (see definition above). The core genes flanking these spots provide upper bounds for the limits of the ICE. Second, we analyzed the interval with the ICE and identified the limits of the element by overlaying the information on the presence of genes of the conjugation system, on G+C content, and on the frequency of accessory genes in the spot pan-genome. The genes of the ICE are expected to be present in the spot pan-genome at similar frequencies (some differences may be caused by mutations, deletions, transposable elements, and annotation errors) and this information is usually sufficient to delimit the ICE. We produced a

visual representation of this data in the context of the spot, and used it to precisely delimit the ICE at the gene-level (see Figure 1 and Supplementary Figure S1).

Specific functional analyses

Antibiotic resistance genes were annotated with the Resfams profiles (core version, v1.1) (33) using HMMER 3.1b1 (34), with the option –cut_ga. The cellular localization was determined with PsortB (version 3.0) (35), using the default parameters for diderms and monoderms separately. Genes encoding stable RNAs were annotated using Infernal (36) and Rfam covariance models (37) (hits were regarded as significant when *e*-value $\leq 10^{-5}$). Integrons were detected using IntegronFinder v1.5 with the –local_max option (38). DDE transposes were annotated with MacSyFinder (29) following the procedure described in (31). Integrases were detected with the PFAM profile PF00589 for tyrosine recombinases and the pair PF00239 and PF07508 for Serine recombinases (http://pfam.xfam.org/) (39). HMMER hits were regarded as significant when their e-value was smaller than 10^{-3} and their alignment covered at least 50% of the protein profile.

Specific HMM protein profiles were built with HMMER v3.1b1 for partition systems (40,41), replication proteins (42), and entry exclusion systems (43). In the general case, we started from a few proteins with experimental evidence of the given function, curated by experts, or reported in published databases. Since these sets were usually small and present in a small number of species, we used a two-step procedure (described below): we started by building preliminary profiles, used them to scan the complete genome database, and then used these results to make the final profiles. First, the proteins of experimental model systems were aligned with mafft v7.154b (with –auto parameter) (44), and manually trimmed at the N- and C-terminal ends with SeaView v.4.4.1 (45). The alignments were used to make preliminary HMM profiles using hmmbuild from HMMER v.3.1b1. A first round of searches with these profiles using hmmsearch (*e*-value $< 10^{-3}$ and coverage >50%) returned hits that were clustered with usearch (–cluster_fast at 90% identity) (46). We took only the longest protein of each cluster, to remove redundant sequences, and searched for sequence identity between all pairs of these representative hits using blastp v.2.2.15 (with the –F F parameter to not filter query sequences). The output was then clustered to identify protein families using Silix (version 1.2.8, 40% identity and 80% of coverage). We made multiple alignments of the resulting families and used them to build a novel set of HMM protein profiles (alignment and trimming as above). The detailed procedure used for building the protein profiles of each function is given in the supplementary material.

Functional annotation

We used HMMER v.3.1b1 (*e*-value <0.001 and coverage of 50%) to search ICEs for hits against the EggNOG Database of hmm profiles (Version 4.5, bactNOG). These results were used to class genes in broad functional categories. We added a class 'Unknown' for genes lacking hits when queried with EggNOG. We tested the over-representation of given functional categories in ICE using binomial tests where the successes were given by the number of times that the relative frequency of a given functional category was higher in the ICE than in the rest of the host chromosome. Under the null hypothesis, any given category is as frequent in the ICE as in the rest of the chromosome (relative to the number of genes). Formally:

$$H0 : \forall x \in Cat_{EggNOG} : N(f(x_{ICE}) > f(x_{HOST})) \sim \mathcal{B}(N_{ICE}, 0.5).$$

With $N(f(x_{ICE}) > f(x_{HOST}))$ being the number of times the EggNOG category $\boldsymbol{x}$ had a higher relative frequency in the ICE than in the rest of the host chromosome, and with $\mathcal{B}(N_{ICE}, 0.5)$ being a binomial distribution with N_{ICE} trials (199 typed ICEs) and an *a priori* probability of 50%.

Networks of homology

We searched for sequence similarity between all proteins of all ICE elements using blastp v.2.2.15 (default parameters) and kept all bi-directional best hits with an e-value lower than 10^{-5}. We used the results to compute a score of gene repertoire relatedness for each pair of ICEs weighted by sequence identity:

$$wGRR_{A,B} = \sum_i \frac{id(A_i, B_i)}{\min(A, B)} \underset{iff}{\longleftrightarrow} evalue(A_i, B_i) < 10^{-5}$$

where (A_i, B_i) is the pair i of homologous proteins in ICEs A and B, id(A_i,B_i) is the sequence identity of their alignment, min(A,B) is the number of proteins of the element with fewest proteins (A or B). The wGRR varies between zero and one, it represents the sum of the identity for all pairs of orthologs of the two ICEs, divided by the number of proteins of the smallest ICE. It is zero if there are no orthologs between the elements and one if all genes of the smaller element have an ortholog 100% identical in the other element. A similar score was previously used to compare prophages (47). We kept pairs of ICEs with a wGRR higher than a certain threshold (5% for the general analysis and 30% for the supplementary analysis). A wGRR of 5% represents, for instance, the occurrence of one homologous gene with 100% identity between two ICEs, where the smallest encodes 20 proteins. We computed two versions of this score for each pair of ICEs, one with all proteins and another after excluding the proteins from the conjugative system.

Phylogenetic tree

The phylogenetic analysis used the 134 integrases from ICEs containing exactly one tyrosine recombinase. We excluded the other elements because additional recombinases might be involved in other functions (dimmer resolution, DNA inversions) and could confound the results. We added to this dataset: 60 phage integrases randomly selected from a database of 296 non-redundant (<90% identity) phage integrases from RefSeq; 11 integrases from pathogenicity islands (Supplementary Table S2 from (48)); 25 experimentally studied XerC and XerD from (49), XerS from (50), XerH from (51,52); seven integrases representing the diversity of integron integrases (38). The final dataset was composed of 237 integrases. The initial alignment was made

with mafft (with parameters –maxiterate 1000 –genafpair), and 100 alternative guide-trees were built. We then removed non-informative positions in the multiple alignment (columns) when they had less than 50% of confidence score, as calculated by GUIDANCE 2 (53). We used Phylobayes MPI (version 1.7, model CAT+GTR+Gamma, two chains for 30 000 iterations) to build the phylogeny from that alignment (54). Following the guidelines of Phylobayes, we considered that chains had converged enough to give a good picture of the posterior consensus when the maximum difference across all bipartitions was lower or equal than 0.3 (we obtained 0.24). The tree was represented with Figtree v1.4.2.

Identification of origin and terminus of replication

The predicted origin (ori) and terminus (ter) of replication were taken from DoriC (55). When one predicted replichore was more than 20% larger than the other, the genome was removed from the corresponding analysis. The leading strand was defined as the one showing a positive GC skew (56).

RESULTS AND DISCUSSION

Identification and delimitation of ICEs by comparative genomics

We developed a procedure to identify and delimit ICEs in two stages (see Materials and Methods). First, we identified conjugative systems in bacterial chromosomes using the CONJscan module of MacSyFinder, a methodology that we have previously shown to be highly accurate (17,57). We then concentrated our attention on the conjugative systems of species for which at least four complete genomes were available. The comparative genomics data provides information on the maximal size of the ICE, given by the flanking core genes, and on the frequency of the genes in the locus. Hence, we defined for each species: the core-genome, the intervals (locations between consecutive core genes), the spots (sets of intervals flanked by the same families of core genes, see Methods), and the pan-genomes of each spot. ICEs were always in a single interval, ICEs were never part of the core genome, and could be preliminarily delimited at their flaking core genes. We then analyzed the frequency of gene families in the spot pan-genome, the position of conjugative systems, and the G+C content (Figure 1 and Supplementary Figure S1). The integration of this information allows to identify the set of genes that are part of the ICE. Of note, we were not able to obtain a general method to identify accurately the attachment sites (*attL* and *attR)* delimiting ICEs. Hence, we placed the elements' borders at the edges of their flanking genes. The customizable standalone and online tools to identify ICEs and a tutorial for their use are here made freely available (see Availability section).

We identified five pairs of ICEs in tandem. They were relatively easy to identify with our procedure because the corresponding intervals had two copies of the conjugation apparatus (and typically two integrases). The tandem ICEs corresponded to different MPF types in four out of five pairs. This fits previous observations that tandems of identical ICEs are very rare in *recA*$^+$ backgrounds (all the ICEs analyzed in this work are in these circumstances), presumably because they are rapidly deleted by homologous recombination (58). Some elements encoded two or more integrases or relaxases and only one conjugation system. They may be composite elements resulting from the independent integration of different elements, or they may correspond to ICE encoding additional tyrosine recombinases with other functions (than that of being an integrase). When faced with such elements, and when the frequency of the genes was homogeneous, we included them in one single ICE. This choice was based on the published works showing that ICEs can mobilize composite elements, including genomic islands or IMEs (59–61), and on the observation that conjugative plasmids sometimes also include other integrative elements and multiple relaxases (7).

The curation process started with 601 conjugative systems detected among 2484 complete genomes. Among these, we selected 207 (∼35%) elements that were present in species with more than four complete sequenced genomes, and we ultimately were able to delimit 200 ICEs within 37 species. A total of 41 had several ICEs, typically in different genomes, accounting for a total of 118 elements. Hence, integration of different ICEs in the same locus is common. To assess if these ICEs are different elements or the result of single ancestral integrations, we computed the weighted gene repertoire relatedness (wGRR), which represents the proportion of homologous genes between two ICEs weighted by their sequence identity (see Methods, Supplementary Figure S2). Most ICEs had very low wGRR when compared with any other element. Yet, there were 61 ICEs very similar, i.e. with wGRR>90% (70 ICEs with wGRR>80%), to at least one of the ICEs in the same spot. If one had kept only one ICE per family with this threshold of wGRR>90%, this would have reduced the number of ICEs from 200 to 160 (Supplementary Figure S2). Yet, these elements are not necessarily derived from the same ancestral event of integration since we found 29 ICEs with similarly high wGRR values in different species and 24 in different genera. These latter elements are most likely the result of multiple independent integrations in the same locus, not of a single ancestral integration, since ICEs are never part of the species core genome. Hence, instead of arbitrarily selecting one ICE per spot, we opted to maintain all elements in the subsequent analyses and control for this effect, when necessary.

We used information from the ICEberg database to validate the delimitation procedure. We could not use the whole database because there was no complete conjugative system in 51% of the elements of ICEberg (184 out of 358), and 40% of them even lacked the essential protein VirB4/TraU. Hence, we restricted our comparisons to the 19 ICEs in ICEberg that were derived from experimental data or predicted from literature and that were in a genome available in our dataset (Supplementary Table S1). Among these elements, 16 ICEs had their start and end positions within less than 2.5 kb of ours (typically less than 500bp, Supplementary Figure S3). We analyzed in detail the three cases showing discordance between the two datasets. Two of them corresponded to a tandem of ICEs that we split in our procedure because they had two different complete conjugative systems. They were identified as a single ICE in ICEberg.

The third discordance arose because the ICEberg annotation included an MPF_{FATA} ICE interrupted by an MPF_{FA} ICE. Our procedure spotted the complete MPF_{FA} ICE. These results suggest that our delimitation is more accurate. It is important to note that our procedure does not use any information on the experimental models of ICE; it is only based on the identification of conjugation proteins, and the frequency of accessory genes. Hence, the accuracy of our method is expected to remain high even when studying poorly known ICE families. It is also expected to improve upon inclusion of more genomes within a species.

By the end of this study there was an influx of novel complete genomes in the public databases. We took advantage of this data to make a *post hoc* validation of our results. We concentrated our attention on MPF_T elements because they are the most abundant and the best studied. We identified 1181 conjugative systems in the chromosomes of the novel genomes. We delimited 124 novel MPF_T ICEs in the novel genomes (out of 498 detected) and compared them with the ICEs of the same type in the main dataset, according to five measures. The results showed no significant difference between the two sets (Supplementary Tables S7 and S8), suggesting that our results are robust to sampling effects, *i.e.*, adding novel data will not significantly affect the main conclusions of our work.

General features of ICEs

We grouped the 200 ICEs (delimited in the main dataset) on the basis of the MPF and relaxase (MOB) types. We identified five of the eight previously defined MPF types among these elements: type F (9), FA (72), FATA (29), T (49) and G (40). Three types were absent from our data because they are strictly associated with clades for which not enough complete genomes were available (MPF_B for Bacteroides and MPF_C for cyanobacteria) or corresponded to systems that were previously shown to be extremely rare among ICE (MPF_I) (17). A preliminary analysis of the *post hoc* validation dataset mentioned above revealed no ICEs for MPF_I and MPF_C, and very few MPF_B (because they are in poorly sampled species). Study of these elements will have to wait before more data becomes available. As a result, the distribution of MPF types differed from that of the 601 conjugative systems identified in all chromosomes of the main dataset (χ^2, P-value < 0.001, Supplementary Figure S4), and included mostly Firmicutes (for MPF_{FA} and MPF_{FATA}) and Proteobacteria (for the others). One ICE had an undetermined MPF type (albeit it included a VirB4, a MOB and a coupling protein), and was excluded from the remaining analyses (except from the wGRR network, see below). Expectedly, the types of relaxases identified in the ICEs corresponded to those frequently found in Proteobacteria and Firmicutes (17) (Supplementary Figure S5).

The size of ICEs described in the scientific literature varies between ~13kb (ICE*Sa1* in *Staphylococcus aureus* (62)) and ~500kb (ICEMlSymR71 in *Mesorhizobium loti* R7A (63)). In our dataset, the smallest ICE was also identified in *S. aureus* (strain USA300-FPR3757) and was only 11.5 kb long. The largest ICEs were found in *Rhodopseudomonas palustris BisB18* (MPF_T) *and Pseudomonas putida S16* (MPF_G) and were around 155 kb long. The distribution

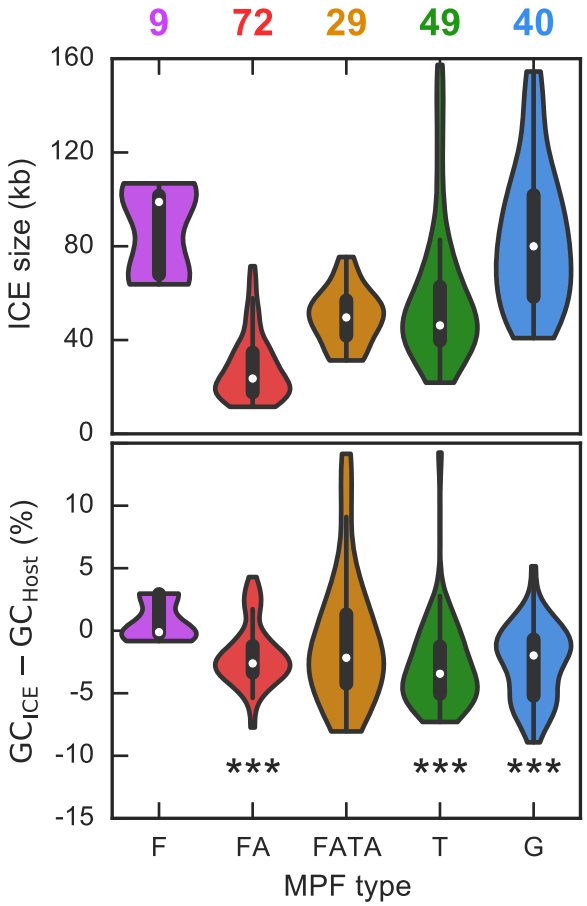

Figure 2. ICE statistics as a function of the MPF type. **Top**. Distribution of the size of ICEs (in kb). The numbers above each violin plot represent the number of elements in each category. **Bottom**. Distribution of pairwise differences between the GC content of the ICE and that of its host. The violin plots represent the kernel density estimation of the underlying values. Here the violin plots are limited by the minimum and maximum values. ***P-value <0.001, Wilcoxon signed-rank test (rejecting the null hypothesis that the difference is equal to zero).

of ICE size per MPF type showed that type F (median size of 99 kb) and G (median 80 kb) were the largest, whereas those of type FA were the smallest (median 23.5 kb) (Figure 2). The distribution of the size of ICE in our dataset is close to that of ICEs reported in the literature, even if we could not identify any ICE of a size comparable to ICEMlSymR71. Such large ICEs may be rare or specific of taxa not sampled in our study. The majority of ICEs were found to be AT-rich compared to their host's chromosome (Figure 2). This is, with some exceptions, a general trend for mobile genetic elements (64). The distribution of sizes of these 160 families of ICEs is similar to the one for the entire dataset (Supplementary Figure S6), validating our decision to keep all ICEs in our analysis.

It was known that the largest plasmids are typically in the largest genomes (7). We observed a positive correlation between the size of ICEs and that of their host chromosomes ($\rho = 0.47$, P-value < 0.0001, after discounting the ICE size from chromosome size), even if this is partly because the smaller types of elements are associated with the

clades with the smallest genomes (Supplementary Figure S7). Larger ICEs may be disfavored in smaller genomes because they are harder to accommodate in terms of genome organization (65), thus leading to higher fitness cost, or because smaller genomes endure lower rates of HGT (66). Interestingly, the average size of ICEs (52.4 kb) could be a general feature of integrative elements, since it comes near to that of temperate phages (∼50 kb) of enterobacteria (67), and of known pathogenicity islands (ranging between 10kb and 100 kb (68)).

The families of ICEs

We analyzed the network of protein sequence similarity between ICEs in relation to some of the best-known experimental models (SXT, ICEclc, Tn916, ICEBs1, TnGBS2, ICEKp1). To this end, we used the abovementioned weighted gene repertoire relatedness (wGRR) scores between every pair of ICEs. The network of wGRR-based relationships between ICEs was represented as an undirected graph, where nodes represent ICEs and edges were weighted according to the wGRR (if wGRR > 5%) (Figure 3). This graph showed that all ICEs were connected in a single component (all nodes can be accessed from any others) with the exception of a group of ICEs from *H. pylori*. ICEs grouped predominantly according to their MPF types, which were sometimes split in several clusters. Interestingly, ICEs in Firmicutes (FA and FATA) and Proteobacteria (T, G and F) formed two main groups, and their sub-groups typically included different species. This fits the observation that FA and FATA are sister-clades in the phylogeny of VirB4 (15). Interestingly, a similar graph where MPF proteins were excluded from the calculation of wGRR produced very similar results, suggesting that the effect of the host taxa may be preponderant in the split into two groups according to the major phyla (Supplementary Figure S8). To detail the similarities between ICEs, we also clustered them using a more restrictive threshold (wGRR > 30%, Supplementary Figure S9). This graph is composed of many connected components of single MPF types, highlighting the huge diversity of ICEs.

Independent integrase acquisitions by ICE

Integrases allow the integration of ICEs in the chromosome and are one of their most distinctive features (relative to conjugative plasmids). Around 70% of the ICEs encoded a single tyrosine recombinase, nine encoded a Serine recombinase, 21 had at least two integrases, among which six had both Serine and Tyrosine recombinases. A total of 37 ICEs lacked integrases, among which six encoded one or more DDE recombinases (in two cases the genes are at the edge of the ICE) and six encoded pseudogenized integrases. DDE recombinases-mediated ICE integration was previously described for ICE TnGBs2 in *Streptococcus agalactiae* (13) and for ICEA in *Mycoplasma agalactiae* (14). Experimental work will be necessary to test if some of the six ICEs with only DDE recombinases use them to replace integrases. Of note, we actually found more transposases in ICEs with Serine or Tyrosine recombinases than in those lacking them (χ^2 on a contingency table, *P*-value < 0.01). Recently, it has been shown that some ICEs may use relaxases instead of integrases to integrate the chromosome when there is an *oriT* in the genome that can be recognized by the relaxase (69).

A previous analysis using integrases from the tyrosine recombinase family of genomic islands, phages, and six ICEs (of which four of the SXT family), showed that ICE integrases clustered separately (70). However, doubts have been casted on this analysis because of the small number of ICEs that had been used (71). We have thus made a phylogenetic tree of tyrosine recombinases from the ICEs encoding one single integrase (Figure 4). We analyzed a total of 237 tyrosine recombinases from ICEs and other elements including integrons, four different types of recombinases involved in chromosome dimer resolution (XerCD, XerS, XerH), pathogenicity islands, and phages (see Methods, Supplementary Table S2). This tree showed that the Xer recombinases and the integron integrases were all monophyletic. In contrast, genomic islands and ICEs were scattered in the tree. Even ICEs of similar MPF types are systematically paraphyletic. A particularly striking example is provided by a clade in the tree (arc in Figure 4) that contains integrases from several phages, and two types of ICE (Type G and type T) from different species (*K. pneumonia*e and *P. fluorescens*), as well as three different groups of pathogenicity islands. The clear paraphyly of the integrases at the level of MPF types suggests that conjugative elements often exchange the key genes allowing chromosomal integration with otherwise unrelated mobile genetic elements.

The functional repertoires of ICEs

We investigated the functional classification of the genes in ICEs in relation to those in the rest of the host chromosome using the EggNOG database (see Methods). Unknown or unannotated functions accounted for 61% of all genes in ICEs, showing how much remains to be known about the functions carried by these elements. We observed three functional categories that were systematically more frequent in ICEs (*P*-value<0.01 with Bonferroni correction for multiple test, Figure 5), including typical ICE functions: secretion (genes related to conjugation), replication/recombination/repair (integrases, relaxases, and transposable elements), and to a lesser extent cell cycle control/cell division/chromosome partitioning. The category associated with transcription (gene expression regulation) showed similar frequencies in the ICE and in the host chromosome. Most functions were systematically less frequent in ICEs. Removing from the analysis the proteins implicated in conjugation did not reveal novel families overrepresented in ICEs (Supplementary Figure S10).

Since the functional analysis of ICEs pinpointed an overrepresentation of functions typical of plasmids, we developed more specific approaches to characterize them (see Materials and Methods). We identified 23 partition systems, 13 from type Ia, five of type Ib, and five of type II (no type III). The presence of partition systems suggests the existence of replication systems (even if relaxases can themselves be implicated in ICE replication (25,72)). We found 16 ICEs with proteins predicted to be associated with theta replication (15 in MPF_T and one in MPF_G) and two associated with rolling circle replication (all in MPF_{FATA}). In-

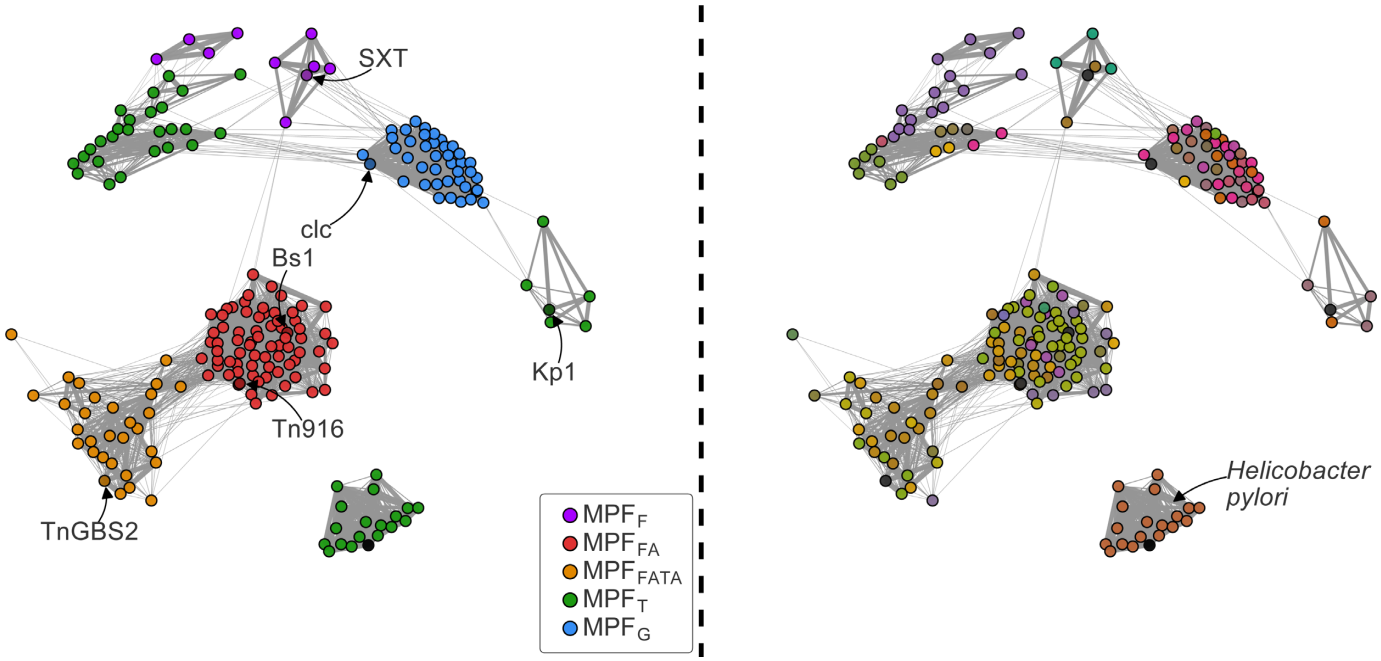

Figure 3. Representation of the wGRR-based network of ICEs. The nodes represent the ICEs and the edges link pairs of ICEs with wGRR score >5% (the thickness of the edge is proportional to the score). **Left**. Nodes are colored according to the MPF type. Darker nodes represent ICEs commonly used as experimental models, and are indicated by an arrow. **Right**. Nodes are colored according to the species of the host to highlight the distribution of the 37 species. The information of the species and type are in Supplementary Table S4. The position of the point has been determined by the Fruchterman-Reingold force-directed algorithm, as implemented in the NetworkX python library (spring layout).

terestingly, all six MPF$_T$ ICEs with a partition system also encoded a replication protein. Apart from these traits (MPF type, partition and replication systems), these six ICEs are very different (average wGRR score of 34%). To the best of our knowledge ICE replication has not been reported in this family (the most numerous one among complete genomes). Naturally, if some of the relaxases act as replicases, then the actual number of replicases in ICE could be much larger. This might explain why few or no replication systems were found in type F and G although they encode partition systems.

Several ICEs encode accessory functions typical of mobile genetic elements, such as integrons (73), restriction-modification (R-M, (74)), toxin–antitoxin (TA, (75)) and exclusion (43) systems. We identified two integrons in ICEs of the SXT family (MPF$_F$) (as first shown in (76)) and one array of *attC* sites lacking the integron–integrase (CALIN elements (38)) in another ICE. We identified 23 ICEs with at least one R-M system (all four types of R–M systems could be identified). The frequency of R-M systems in ICE (12% of all elements encoded at least one complete system) is similar to that observed in plasmids (10.5%) and higher than in phages (1%). Interestingly, as it was the case in these MGEs (74), the frequency of solitary methylases (25%) was higher than that of the complete systems. These solitary methylases might provide a broad protection from the host R–M systems. Most RNA genes identified in ICE corresponded to intron group II associated RNAs (36% of all detected RNA, excluding those from type G), but MPF$_G$ ICEs encoded many *radC* and STAXI RNA (representing 80% of RNA in MPF$_G$). Both genes are associated with anti-restriction functions (77). They might defend the element from R-M systems, thus explaining the relative rarity of the latter in this family of ICEs (7.5% versus 11%). Few ICEs encoded recognizable entry exclusion systems (6%, mostly in MPF$_T$). One ICE contained a type II CRISPR-Cas system in *Legionella pneumophila* str. Paris. Overall, genes encoding many molecular systems associated with plasmid biology could be identified in ICEs, even if some were relatively rare.

The organization of ICE

We grouped ICEs by their MPF types and analyzed their genetic organization (Figure 6 and Supplementary Figure S11). We restricted our attention to ICEs with one single integrase of the Serine or Tyrosine recombinase families (141/199), to avoid the inclusion of recombinases with functions unrelated to the integration of the element and to facilitate the representation of the ICE organization. We represented ICEs in such a way that the integrase was located in the first half of the element. Actually, almost 90% of the Tyrosine and Serine recombinases were within the five first percent of the ICE, as expected given their role in the integration of the element. DDE transposases were randomly distributed within the ICE (Kolmogorov–Smirnov test, *P*-value = 0.27), which suggests that most of them are not involved in the integration of the ICE. The transposases in the inner parts of the element may be involved in accretion and deletion of parts of the element, and can lead to its occasional integration in spots that would not be targeted otherwise.

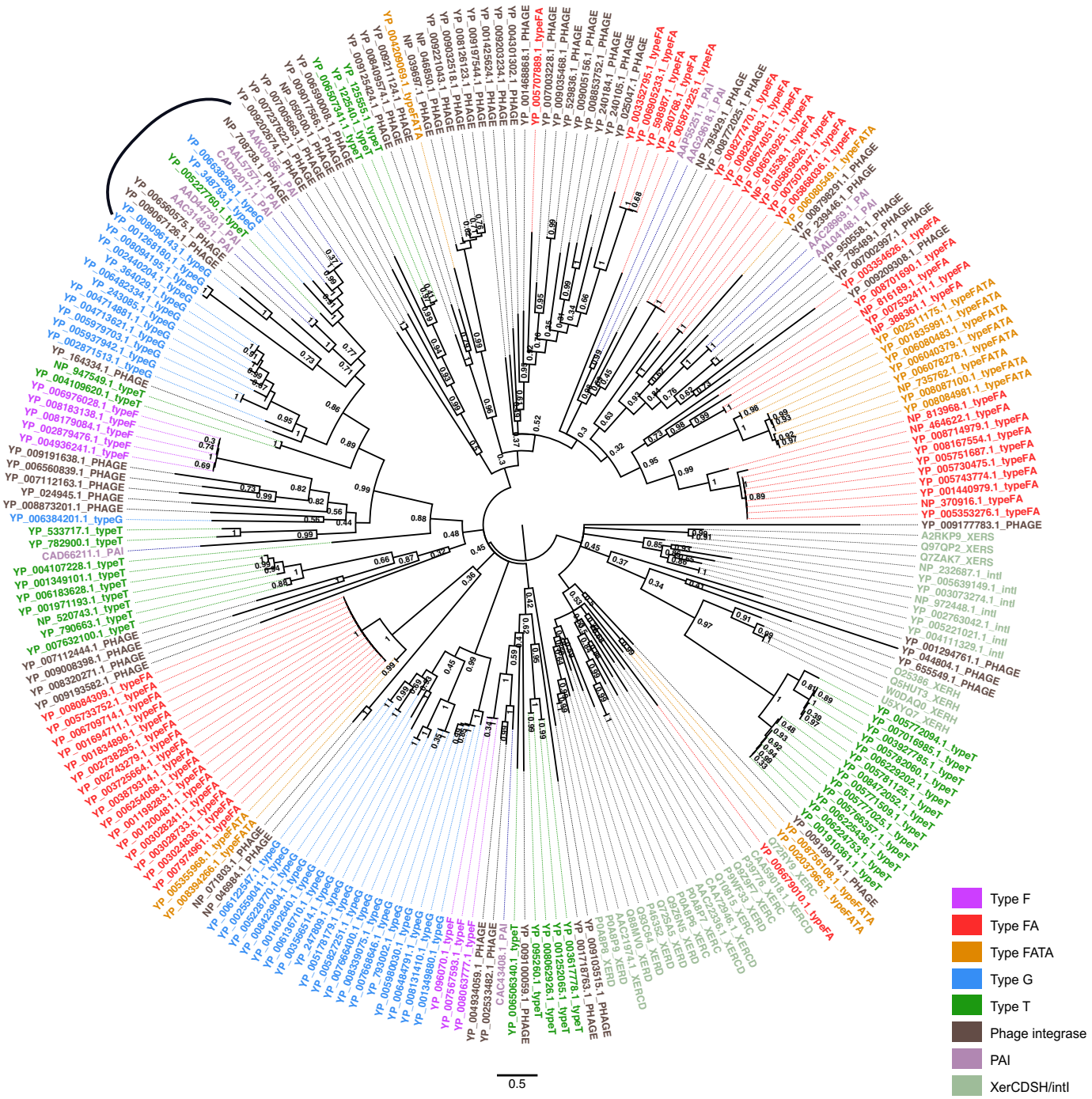

Figure 4. Phylogenetic tree of tyrosine recombinases. The phylogenetic tree was built using 60 prophage integrases (labelled as '…_PHAGE', brown), 11 integrases from pathogenicity islands ('…_PAI', mauve), 25 XerC,D,S or H ('..._XER..', greenish grey), 7 integron-integrases ('..intI', greenish grey), and 134 integrases from ICEs (colored after the MPF type). The tree was built using Phylobayes and the values represent posterior probabilities support of the partition, with a cut off equal to 0.3, below which the nodes were collapsed because there is insufficient resolution (see Materials and Methods). The black arc denotes a clade with good support, which contains integrases from prophages, PAI and different MPF types, that is explicitly cited in the text.

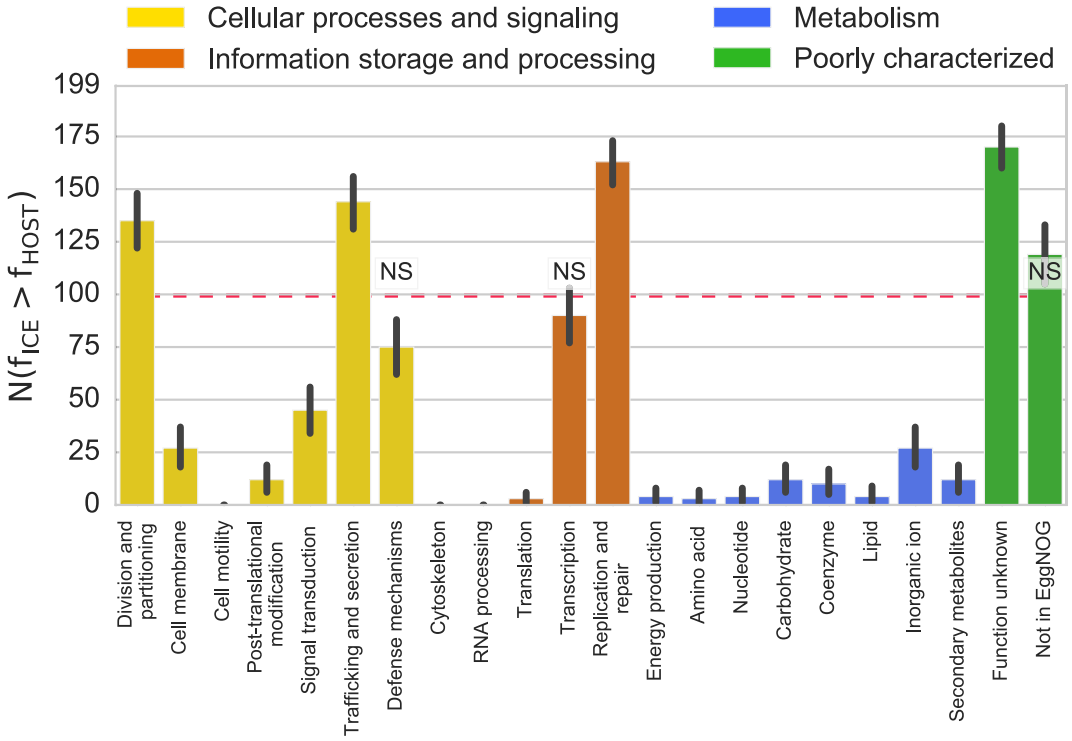

Figure 5. Representation of EggNOG functional categories in ICEs relative to the host chromosome. The bars represent the number of times a given category is found more frequently in an ICE than in its host chromosome ($N(f_{ICE}>f_{HOST})$). The red dotted line represents the expected value under the null hypothesis, where a category is in similar proportion in ICE and its host's chromosome. Bars marked as NS represent a lack of significant difference ($P > 0.05$, Binomial test with 199 trials and expected value of 0.5), whereas the others are all significantly different ($P < 0.05$, same test). Note that there are 199 trials because one of the 200 ICEs could not be types and was thus excluded, see text. Error bars represent 95% confidence interval computed with 1000 bootstraps. 'Not in EggNOG' represents the class of genes that didn't match any EggNOG profile.

To facilitate the representation of the organization of ICE, we oriented the genes relative to *virB4*, which was placed on the top strand. Expectedly, given that they are often part of the same operon, the remaining components of the T4SS were usually found in a single locus and almost always (>96%) on the same strand as *virB4* (Supplementary Figure S11A). The T4SS locus spanned, on average, 26% of the ICEs (Figure 6 and Supplementary Figure S12). As observed in plasmids (7), the relaxase (MOB) gene was sometimes encoded close to the T4SS genes (MPF_F, MPF_T, MPF_{FA}) and sometimes apart (MPF_G, MPF_{FATA}). Interestingly, the relaxase and *virB4* were encoded in the same strand in most cases (86%).

Most accessory functions were encoded apart from the T4SS genes, with the known exception of the entry exclusion systems (43) (Supplementary Figure S11C). We showed above that partition and replication functions co-occurred in ICEs. Here, we show that they co-localize within the element. In MPF_G, they are often found in the edge opposite to the integrase, whereas they are close to the integrase in MPF_F. The genes classed as 'Metabolism' were also encoded away from the region of the T4SS, and were typically found in discrete modules. Intriguingly, some regions were particularly rich in genes of unknown function, e.g. integrase-proximal regions, and also at the opposite end of the elements in MPF_G and MPF_F ICEs (Figure 6). The MPF_T ICEs were an exception to most of these trends, since their genes were almost uniformly distributed along the elements. This group may be genetically more diverse than the others, which would explain these results and the scattering of these ICEs in the homology network.

Interestingly, almost all genes in the ICEs were encoded in the strand of *virB4* (>80%), including the RNA genes (Supplementary Figure S11F). Furthermore, genes were predominantly encoded in the leading strand for all types of ICEs but MPF_G (Supplementary Figure S13). Overall, these results show a certain level of modularity in the organization of ICEs, as previously described in phages and plasmids (78,79), and frequent co-orientation of genes, as identified in different ICE families (SXT (80), Tn916 (81)) and in lambdoid phages (82).

The chromosomal context of ICE

We analyzed the chromosomal context of ICEs to characterize their integration patterns. The analysis of the two chromosomal genes bordering the elements showed that ICEs are often integrated near hypothetical proteins (52%) or tRNAs (30%). These tRNAs decoded 12 different amino acids, in most cases Leucine, Lysine, and Glycine. The tropism towards integration near a tRNA varied with type of ICE, it was high for MPF_G (75%) and null for MPF_{FATA} (0%) (Figure 7).

We then analyzed the distribution of ICEs in the larger context of the bacterial chromosome. Based on the analysis of 15 ICEs, it had recently been suggested that ICEs

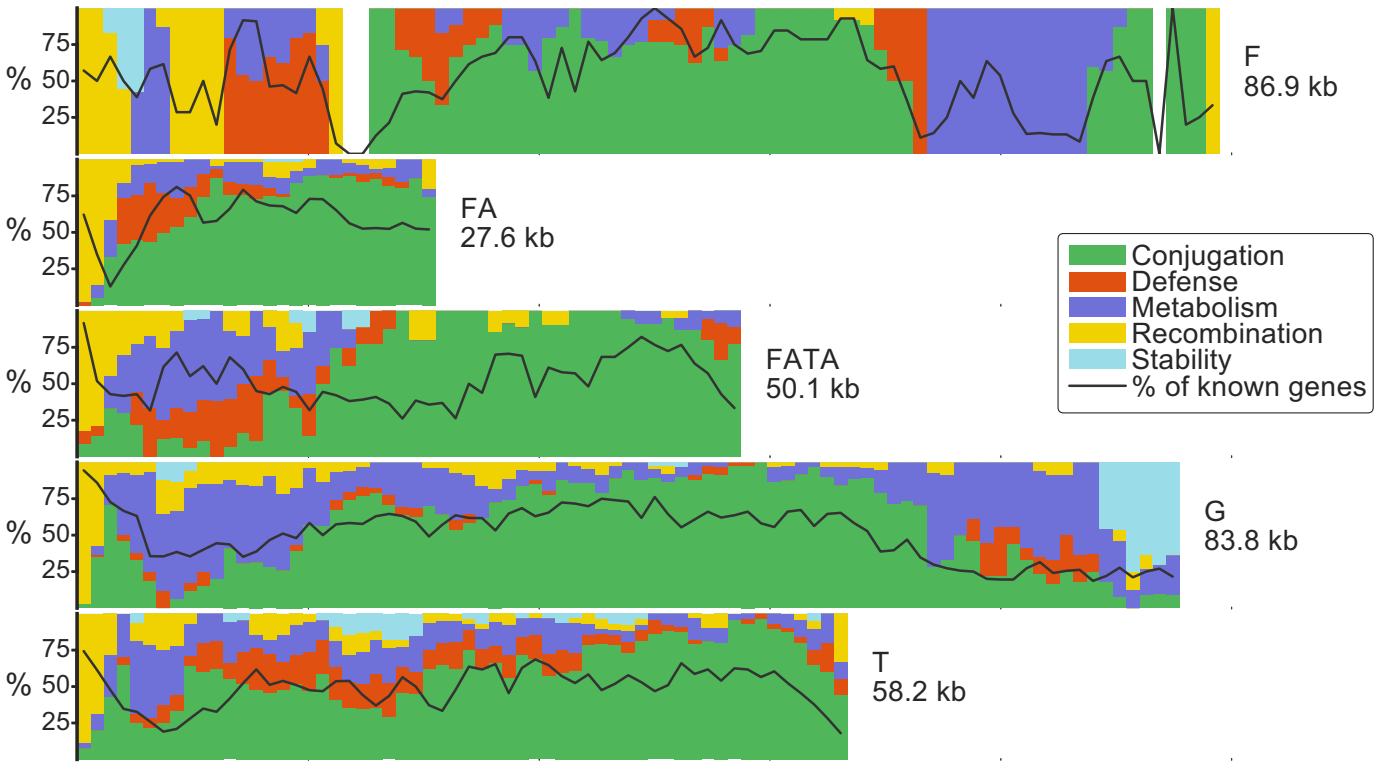

Figure 6. Average organization of ICEs. Each row represents an MPF type and has a length proportional to the mean size of the ICEs of the corresponding type. Colors represent different classes of functions. The black line represents the proportion of genes with a predicted function per bin. The classes of functions correspond to those described in detail in Supplementary Figure S9A-B-C-F. More precisely, *Conjugation* includes MPF-associated genes, and the relaxase. *Defense* includes antibiotic resistance genes, restriction modification, solitary methylases. *Metabolism* includes genes annotated by EggNOG as such. *Recombination* includes tyrosine and serine recombinases, and DDE transposases. *Stability* includes replication, partition and entry exclusion systems. Bar heights are proportional to the proportion of genes of a given function at that position among the genes with a predicted function. The width of each bin is 1kb.

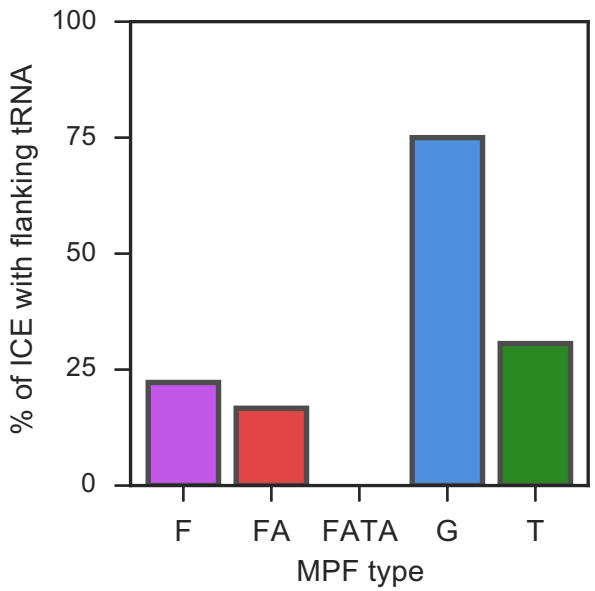

Figure 7. Proportion of ICEs with flanking tRNAs on either side of the ICE.

would be more frequent close to the origin of replication because they target essential highly conserved genes (26), which are more frequent near the origin of replication in fast growing bacteria (83). However, we could not find a significant correlation between the frequency of ICEs and their position in the origin-to-terminus axis of replication (Supplementary Figure S13). When we analyzed the chromosomal distribution of ICEs across MPF types, MPF_{FATA} were over-abundant in the terminus region (χ^2, P-value < 0.003), whereas the others didn't show significant trends (P-value > 0.1, same test). Neither the strand location (χ^2, P-value $= 0.39$) nor the size of the ICE (Spearman-ρ, P-value $= 0.52$) were associated with its distance to the origin of replication.

We identified the origins and terminus of replication of genomes and inferred the leading and lagging strand of each gene in ICEs (see Materials and Methods). Most genes were oriented in the same direction as the replication fork (leading strand, χ^2, P-value $= 10^{-5}$), with the exception of MPF_G ICEs that showed an opposite trend (Supplementary Figure S14). The high frequency of leading strand MPF_{FATA} ICEs may be associated with the hosts' genome organization, in this case they are all Firmicutes, because genomes from this phyla show high frequency of genes in the leading strand (84).

CONCLUSION

Our work shows that one can identify and delimit ICEs from genome data using comparative genomics. The pre-

cise identification of integration sites and the validation of the functions of these ICEs will require further experimental work by experts on a large number of different species and conjugation systems. In this respect, a current limitation of our approach is the reliance on a set of genomes for a given species. This means that ICEs from poorly covered taxa could not be studied. We have restricted our study to complete genomes to avoid using poor quality data. This is not a restriction of the method: draft genomes can also be analyzed with our method, as long as the ICE is in the same contig as the flanking core genes. Unfortunately, our experience is that the presence of repeats in ICEs, like transposases, often splits these elements in several contigs when genomes are sequenced using short-read technologies. The identification ICEs split in several contigs requires the availability of a very similar ICE for reference, which may bias the study of these elements. The increasing use of long-read technologies to sequence bacterial genomes will soon solve this limitation.

Even if our dataset is representative of the diversity of experimentally studied ICEs, we lacked ICEs from types B (Bacteroidetes), of which there are experimental systems (85), and C (Cyanobacteria), for which there is no experimental system. Further data will be necessary to study these elements. Another limitation of this work is the assumption that the presence of a certain number of components of the T4SS system and a relaxase are necessary and sufficient to define an ICE. While our previous studies have shown that we are able to identify known conjugative systems accurately, we cannot exclude the possibility that some of the identified ICE are defective for transfer. This may explain the existence of some elements that lack identifiable integrases. In spite of these limitations, the availability for the first time of large dataset of ICEs having undergone a systematic expert curation allowed to quantify many traits associated with ICEs and confirm, and sometimes infirm, observations made from the small number of well-known ICE models. It also allowed to characterize their genetic organization, and identify common traits.

Integration and conjugation are the only functions that we found to be present in most ICEs. Interestingly, even these functions had very different phylogenetic histories, as revealed by the scattered distribution of ICEs per MPF type in the phylogenetic tree of the tyrosine recombinases. A number of other functions were often identified in some types of ICE, notably defense systems, partition, and replication. Collectively, they reinforce the suggestions of a thin line separating ICEs from conjugative plasmids (26). The analysis of the genetic organization of ICEs suggests that they are organized in functional modules. Together, these results suggest that ICEs are highly modular, which may contribute to the evolution of their gene repertoires by genetic exchange between elements, as previously observed in temperate phages (78). If so, our data suggests that either ICEs tend to recombine more with elements of the same MPF type, or that the fitness of the products of recombination tends to be higher when recombination takes place within ICE of the same type (e.g. for functional reasons).

AVAILABILITY

The program to identify conjugative systems is available on https://github.com/gem-pasteur/Macsyfinder_models

The webserver is hosted on: https://galaxy.pasteur.fr/ > Search > CONJScan

The program and data to make representations like those of Supplementary Figure S1 is available at https://gitlab.pasteur.fr/gem/spot_ICE.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

J.C. is a member of the ≪ École Doctorale Frontière du Vivant (FdV) – Programme Bettencourt ≫. We thank Christine Citti for insightful comments on a previous version of this manuscript, Fernando de la Cruz for insights and providing the list of rep proteins of PLACNET, Bertrand Néron and Olivia Doppelt-Azeroual for setting up the webserver version of CONJscan.

Author contributions: J.C. and E.P.C.R. designed the study. J.C. and M.T. produced the data. J.C. made the analysis. J.C. and E.P.C.R. drafted the manuscript. All authors contributed to the final text of the manuscript.

FUNDING

European Research Council [EVOMOBILOME, 281605 to E.P.C.R.]. Funding for open access charge: ERC EVOMOBILOME.

Conflict of interest statement. None declared.

REFERENCES

1. Ochman,H., Lawrence,J.G. and Groisman,E.A. (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
2. Popa,O. and Dagan,T. (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr. Opin. Microbiol.*, **14**, 615–623.
3. Polz,M.F., Alm,E.J. and Hanage,W.P. (2013) Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.*, **29**, 170–175.
4. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
5. Davies,J. and Davies,D. (2010) Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.: MMBR*, **74**, 417–433.
6. Boyd,E.F. and Brussow,H. (2002) Common themes among bacteriophage-encoded virulence factors and diversity among the bacteriophages involved. *Trends Microbiol.*, **10**, 521–529.
7. Smillie,C., Pilar Garcillan-Barcia,M., Victoria Francia,M., Rocha,E.P.C. and de la Cruz,F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
8. Burrus,V., Pavlovic,G., Decaris,B. and Guedon,G. (2002) Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.*, **46**, 601–610.
9. Johnson,C.M. and Grossman,A.D. (2015) Integrative and conjugative elements (ICEs): what they do and how they work. *Annu. Rev. Genet.*, **49**, 577–601.
10. Ghinet,M.G., Bordeleau,E., Beaudin,J., Brzezinski,R., Roy,S. and Burrus,V. (2011) Uncovering the prevalence and diversity of integrating conjugative elements in Actinobacteria. *PLoS ONE*, **6**, e27846.
11. Goessweiner-Mohr,N., Arends,K., Keller,W. and Grohmann,E. (2013) Conjugative type IV secretion systems in Gram-positive bacteria. *Plasmid*, **70**, 289–302.

12. Wang,H. and Mullany,P. (2000) The large resolvase TndX is required and sufficient for integration and excision of derivatives of the novel conjugative transposon Tn5397. *J. Bacteriol.*, **182**, 6577–6583.
13. Brochet,M., Da Cunha,V., Couve,E., Rusniok,C., Trieu-Cuot,P. and Glaser,P. (2009) Atypical association of DDE transposition with conjugation specifies a new family of mobile elements. *Mol. Microbiol.*, **71**, 948–959.
14. Dordet Frisoni,E., Marenda,M.S., Sagne,E., Nouvel,L.X., Guerillot,R., Glaser,P., Blanchard,A., Tardy,F., Sirand-Pugnet,P., Baranowski,E. *et al.* (2013) ICEA of Mycoplasma agalactiae: a new family of self-transmissible integrative elements that confers conjugative properties to the recipient strain. *Mol. Microbiol.*, **89**, 1226–1239.
15. Guglielmini,J., de la Cruz,F. and Rocha,E.P.C. (2013) Evolution of conjugation and type IV secretion systems. *Mol. Biol. Evol.*, **30**, 315–331.
16. Guglielmini,J., Neron,B., Abby,S.S., Garcillan-Barcia,M.P., la Cruz,F.D. and Rocha,E.P. (2014) Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. *Nucleic Acids Res.*, **42**, 5715–5727.
17. Guglielmini,J., Quintais,L., Pilar Garcillan-Barcia,M., de la Cruz,F. and Rocha,E.P.C. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
18. Juhas,M., van der Meer,J.R., Gaillard,M., Harding,R.M., Hood,D.W. and Crook,D.W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.*, **33**, 376–393.
19. Roberts,A.P. and Mullany,P. (2011) Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. *FEMS Microbiol. Rev.*, **35**, 856–871.
20. Carraro,N. and Burrus,V. (2014) Biology of Three ICE Families: SXT/R391, ICEBs1, and ICESt1/ICESt3. *Microbiol. Spectr.*, **2**, doi:10.1128/microbiolspec.MDNA3-0008-2014.
21. Auchtung,J.M., Aleksanyan,N., Bulku,A. and Berkmen,M.B. (2016) Biology of ICEBs1, an integrative and conjugative element in Bacillus subtilis. *Plasmid*, **86**, 14–25.
22. Wang,J., Wang,G.R., Shoemaker,N.B. and Salyers,A.A. (2001) Production of two proteins encoded by the Bacteroides mobilizable transposon NBU1 correlates with time-dependent accumulation of the excised NBu1 circular form. *J. Bacteriol.*, **183**, 6335–6343.
23. Lee,C.A., Babic,A. and Grossman,A.D. (2010) Autonomous plasmid-like replication of a conjugative transposon. *Mol. Microbiol.*, **75**, 268–279.
24. Grohmann,E. (2010) Autonomous plasmid-like replication of Bacillus ICEBs1: a general feature of integrative conjugative elements? *Mol. Microbiol.*, **75**, 261–263.
25. Carraro,N., Poulin,D. and Burrus,V. (2015) Replication and active partition of integrative and conjugative elements (ICEs) of the SXT/R391 family: the line between ICEs and conjugative plasmids is getting thinner. *PLoS Genet.*, **11**, e1005298.
26. Carraro,N. and Burrus,V. (2015) The dualistic nature of integrative and conjugative elements. *Mobile Genet. Elem.*, **5**, 98–102.
27. Garriss,G., Waldor,M.K. and Burrus,V. (2009) Mobile antibiotic resistance encoding elements promote their own diversity. *PLoS Genet.*, **5**, e1000775.
28. Guerillot,R., Siguier,P., Gourbeyre,E., Chandler,M. and Glaser,P. (2014) The diversity of prokaryotic DDE transposases of the mutator superfamily, insertion specificity, and association with conjugation machineries. *Genome Biol. Evol.*, **6**, 260–272.
29. Abby,S.S., Neron,B., Menager,H., Touchon,M. and Rocha,E.P. (2014) MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, **9**, e110726.
30. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
31. Touchon,M., Cury,J., Yoon,E.-J., Krizova,L., Cerqueira,G.C., Murphy,C., Feldgarden,M., Wortman,J., Clermont,D., Lambert,T. *et al.* (2014) The genomic diversification of the whole Acinetobacter genus: origins, mechanisms, and consequences. *Genome Biol. Evol.*, **6**, 2866–2882.
32. Miele,V., Penel,S. and Duret,L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
33. 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.
34. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
35. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Sahinalp,S.C., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
36. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
37. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
38. Cury,J., Jove,T., Touchon,M., Neron,B. and Rocha,E.P. (2016) Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.*, **44**, 4539–4550.
39. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
40. Gerdes,K., Moller-Jensen,J. and Bugge Jensen,R. (2000) Plasmid and chromosome partitioning: surprises from phylogeny. *Mol. Microbiol.*, **37**, 455–466.
41. Larsen,R.A., Cusumano,C., Fujioka,A., Lim-Fong,G., Patterson,P. and Pogliano,J. (2007) Treadmilling of a prokaryotic tubulin-like protein, TubZ, required for plasmid stability in Bacillus thuringiensis. *Genes Dev.*, **21**, 1340–1352.
42. Lanza,V.F., de Toro,M., Garcillan-Barcia,M.P., Mora,A., Blanco,J., Coque,T.M. and de la Cruz,F. (2014) Plasmid flux in Escherichia coli ST131 sublineages, analyzed by plasmid constellation network (PLACNET), a new method for plasmid reconstruction from whole genome sequences. *PLoS Genet.*, **10**, e1004766.
43. Garcillan-Barcia,M.P. and de la Cruz,F. (2008) Why is entry exclusion an essential feature of conjugative plasmids? *Plasmid*, **60**, 1–18.
44. 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.
45. Gouy,M., Guindon,S. and Gascuel,O. (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
46. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
47. Bobay,L.-M., Rocha,E.P.C. and Touchon,M. (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.*, **30**, 737–751.
48. Yoon,S.H., Hur,C.G., Kang,H.Y., Kim,Y.H., Oh,T.K. and Kim,J.F. (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. *BMC Bioinformatics*, **6**, 184.
49. 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.
50. Le Bourgeois,P., Bugarel,M., Campo,N., Daveran-Mingot,M.L., Labonte,J., Lanfranchi,D., Lautier,T., Pages,C. and Ritzenthaler,P. (2007) The unconventional Xer recombination machinery of Streptococci/Lactococci. *PLoS Genet.*, **3**, e117.
51. Debowski,A.W., Carnoy,C., Verbrugghe,P., Nilsson,H.O., Gauntlett,J.C., Fulurija,A., Camilleri,T., Berg,D.E., Marshall,B.J. and Benghezal,M. (2012) Xer recombinase and genome integrity in Helicobacter pylori, a pathogen without topoisomerase IV. *PLoS One*, **7**, e33310.
52. Leroux,M., Rezoug,Z. and Szatmari,G. (2013) The Xer/dif site-specific recombination system of Campylobacter jejuni. *Mol. Genet. Genomics*, **288**, 495–502.
53. Sela,I., Ashkenazy,H., Katoh,K. and Pupko,T. (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res.*, **43**, W7–W14.
54. Lartillot,N., Rodrigue,N., Stubbs,D. and Richer,J. (2013) PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.*, **62**, 611–615.

55. Gao,F., Luo,H. and Zhang,C.T. (2013) DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.*, **41**, D90–D93.
56. Lobry,J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, **13**, 660–665.
57. Abby,S.S., Cury,J., Guglielmini,J., Neron,B., Touchon,M. and Rocha,E.P. (2016) Identification of protein secretion systems in bacterial genomes. *Scientific Rep.*, **6**, 23080.
58. Burrus,V. and Waldor,M.K. (2004) Formation of SXT tandem arrays and SXT-R391 hybrids. *J. Bacteriol.*, **186**, 2636–2645.
59. Paauw,A., Leverstein-van Hall,M.A., Verhoef,J. and Fluit,A.C. (2010) Evolution in quantum leaps: multiple combinatorial transfers of HPI and other genetic modules in Enterobacteriaceae. *PLoS One*, **5**, e8662.
60. Lechner,M., Schmitt,K., Bauer,S., Hot,D., Hubans,C., Levillain,E., Locht,C., Lemoine,Y. and Gross,R. (2009) Genomic island excisions in Bordetella petrii. *BMC Microbiol.*, **9**, 141.
61. Bellanger,X., Morel,C., Gonot,F., Puymege,A., Decaris,B. and Guedon,G. (2011) Site-specific accretion of an integrative conjugative element together with a related genomic island leads to cis mobilization and gene capture. *Mol. Microbiol.*, **81**, 912–925.
62. Schijffelen,M.J., Boel,C.H., van Strijp,J.A. and Fluit,A.C. (2010) Whole genome analysis of a livestock-associated methicillin-resistant Staphylococcus aureus ST398 isolate from a case of human endocarditis. *BMC Genomics*, **11**, 376.
63. Ramsay,J.P., Sullivan,J.T., Stuart,G.S., Lamont,I.L. and Ronson,C.W. (2006) Excision and transfer of the Mesorhizobium loti R7A symbiosis island requires an integrase IntS, a novel recombination directionality factor RdfS, and a putative relaxase RlxS. *Mol. Microbiol.*, **62**, 723–734.
64. Rocha,E. and Danchin,A. (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet.*, **18**, 291–294.
65. Touchon,M., Bobay,L.M. and Rocha,E.P. (2014) The chromosomal accommodation and domestication of mobile genetic elements. *Curr. Opin. Microbiol.*, **22**, 22–29.
66. Oliveira,P.H., Touchon,M. and Rocha,E.P. (2016) Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, 5658–5663.
67. Bobay,L.M., Touchon,M. and Rocha,E.P.C. (2014) Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 12127–12132.
68. Gal-Mor,O. and Finlay,B.B. (2006) Pathogenicity islands: a molecular toolbox for bacterial virulence. *Cell Microbiol.*, **8**, 1707–1719.
69. Agundez,L., Gonzalez-Prieto,C., Machon,C. and Llosa,M. (2012) Site-specific integration of foreign DNA into minimal bacterial and human target sequences mediated by a conjugative relaxase. *PLoS One*, **7**, e31047.
70. Boyd,E.F., Almagro-Moreno,S. and Parent,M.A. (2009) Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.*, **17**, 47–53.
71. Bellanger,X., Payot,S., Leblond-Bourget,N. and Guedon,G. (2014) Conjugative and mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol. Rev.*, **38**, 720–760.
72. Thomas,J., Lee,C.A. and Grossman,A.D. (2013) A conserved helicase processivity factor is needed for conjugation and replication of an integrative and conjugative element. *PLoS Genet.*, **9**, e1003198.
73. 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.
74. Oliveira,P.H., Touchon,M. and Rocha,E.P. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 10618–10631.
75. Wozniak,R.A. and Waldor,M.K. (2009) A toxin-antitoxin system promotes the maintenance of an integrative conjugative element. *PLoS Genet.*, **5**, e1000439.
76. 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.
77. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol.*, **11**, R31.
78. Botstein,D. (1980) A theory of modular evolution for bacteriophages. *Ann. N. Y. Acad. Sci.*, **354**, 484–490.
79. Toussaint,A. and Merlin,C. (2002) Mobile elements as a combination of functional modules. *Plasmid*, **47**, 26–35.
80. Beaber,J.W., Hochhut,B. and Waldor,M.K. (2002) Genomic and functional analyses of SXT, an integrating antibiotic resistance gene transfer element derived from Vibrio cholerae. *J. Bacteriol.*, **184**, 4259–4269.
81. Clewell,D.B., Flannagan,S.E. and Jaworski,D.D. (1995) Unconstrained bacterial promiscuity: the Tn916-Tn1545 family of conjugative transposons. *Trends Microbiol.*, **3**, 229–236.
82. Campbell,A.M. (2002) Preferential orientation of natural lambdoid prophages and bacterial chromosome organization. *Theor. Popul. Biol.*, **61**, 503–507.
83. Couturier,E. and Rocha,E. (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.*, **59**, 1506–1518.
84. Rocha,E. (2002) Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.*, **10**, 393–395.
85. Bonheyo,G.T., Hund,B.D., Shoemaker,N.B. and Salyers,A.A. (2001) Transfer region of a Bacteroides conjugative transposon contains regulatory as well as structural genes. *Plasmid*, **46**, 202–209.

Supplementary material for:

Integrative and conjugative elements and their hosts: composition, distribution, and organization

Jean Cury[1,2,*], Marie Touchon[1,2], Eduardo P. C. 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

* Corresponding author jean.cury@normalesup.org

Table of contents

Table S3 - Annotation of all elements (proteins, RNA, recombination site) found in ICEs (attached as text file).

Table S4 - Count of the different functions found in ICEs (attached as text file).

Table S5 – Genomes with ICEs (attached as text file).

Table S6 – List of wGRR scores for all ICEs combinations (attached as text file).

Table S7 – Post-hoc validation

Table S8 – List of 124 new ICEs type T with counts of functions and values used for the post-hoc validation (attached as text file).

Supplementary files (annex)

File S1 - Archive containing Model and profile for the detection of conjugative systems (attached as zip file).

File S2 – Multiple alignment at the basis of the phylogenetic analysis of Tyrosine recombinases.

Tree S1 - Phylogenetic tree of the tyrosine recombinase used for the Figure 4 (attached as in newick format).

Supplementary information on the construction of HMM profiles.

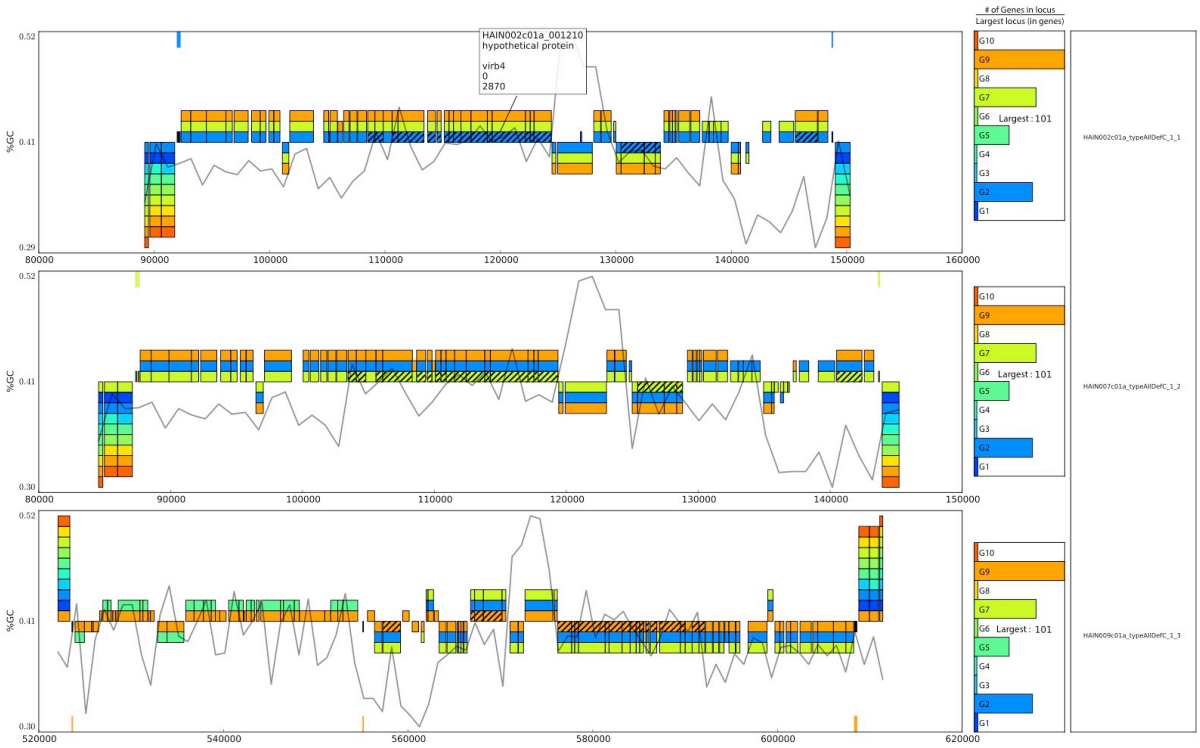

Figure S1: Example of a visualization plot.

We define an interval as the region between consecutive core genes in the genome. Other genomes of the species typically have intervals flanked by genes of the same two families of the core genome. The set of these intervals is called a spot. The plot shows GC content computed in non-overlapping windows of 1kb and gene repertoires in a spot. This spot has one ICE in three genomes (three intervals), and none in seven. The figure is divided in three sub-plots, one per interval containing an ICE. The line represents the GC% along the ICE. Each box is a gene whose width is proportional to its size. Each color represents a genome in the species. A box has hatches if the gene is annotated by a conjugation profile. Genes in the same gene family are stacked on top of the gene of the focal ICE (the one in the center). Genes in a non-focal genome lacking a homolog in the focal genome are not represented. The number of genes in each interval is displayed on the right panel (normalized by the number of genes in the largest interval). This figure is clickable (a text box triggered by a click is depicted in the first plot) and we can report the first and last gene we believe belong to the ICE.

The figure clearly shows that there are three ICEs (from genomes G2, G5 and G9) in the same spot. The flash green bar in genome G5 on the right-hand graphs shows

that there is an additional mobile element in the spot. This element is homologous to an element also present in the genome G9 represented in orange (see left part of bottom graph, indicated by the dotted ellipse) contiguous to the ICE.

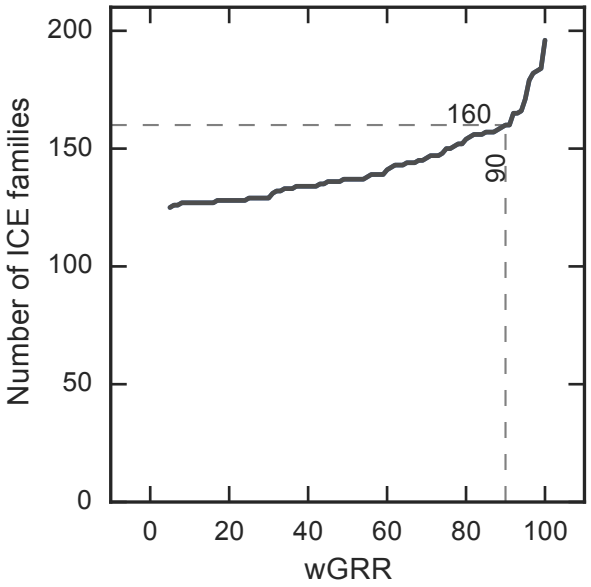

Figure S2 – Number of ICE families with different wGRR threshold values. ICEs are grouped in a family if they have a wGRR above the given threshold and if they are in the same spot. At wGRR=90, there are 160 ICE families.

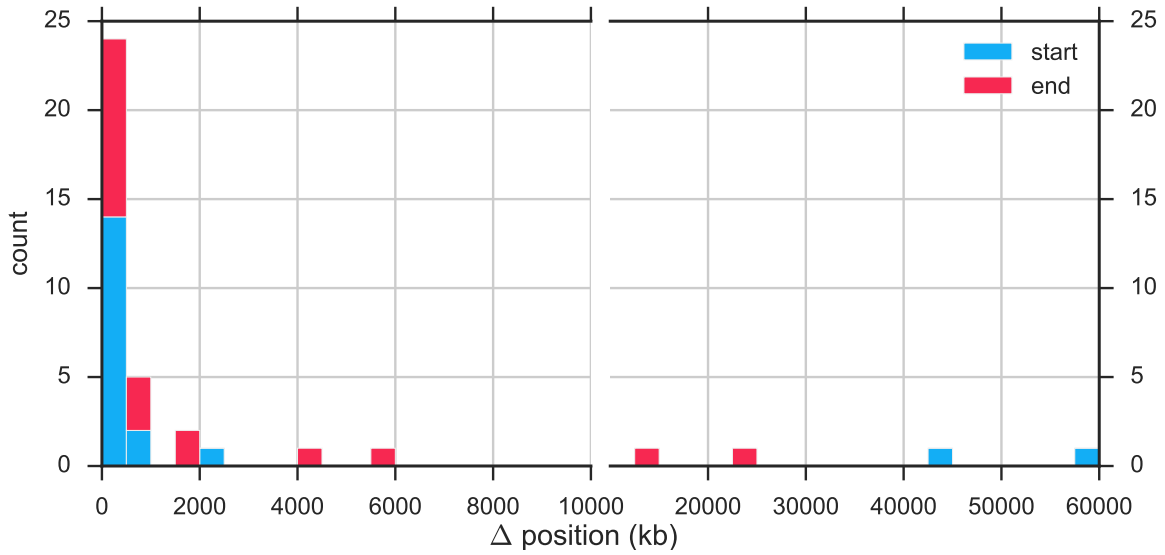

Figure S3: Verification of the 19 delimited reference ICEs.

The plot shows the distribution of the difference (Δ) between the start (blue) and end (red) positions of our delimited ICEs and those annotated by ICEberg.

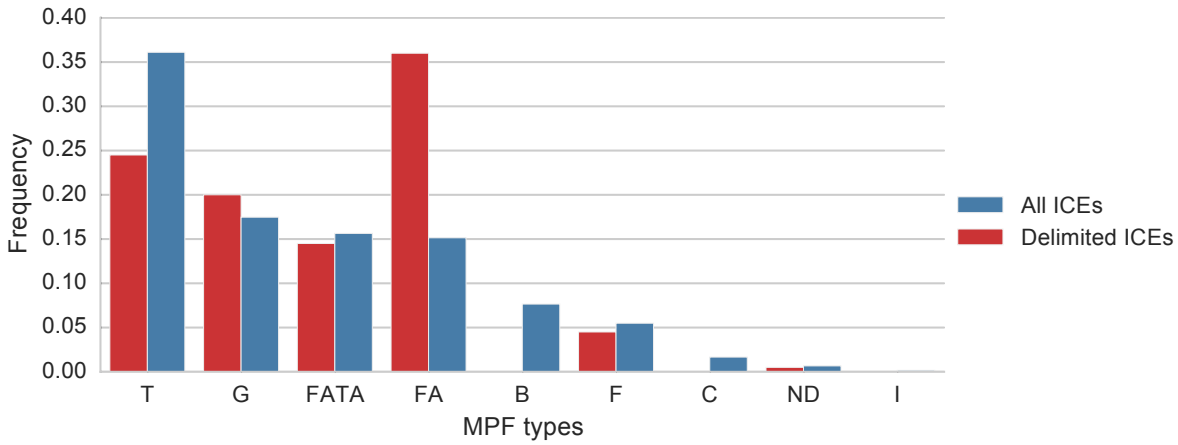

Figure S4: Comparison of the distribution of the different MPF types in all genomes ("All ICEs", 2484 genomes, 601 elements, blue) and on the set of 200 delimited ICEs ("Delimited ICEs", 506 genomes, red). ND represents one ICE that could not be classed in any MPF type.

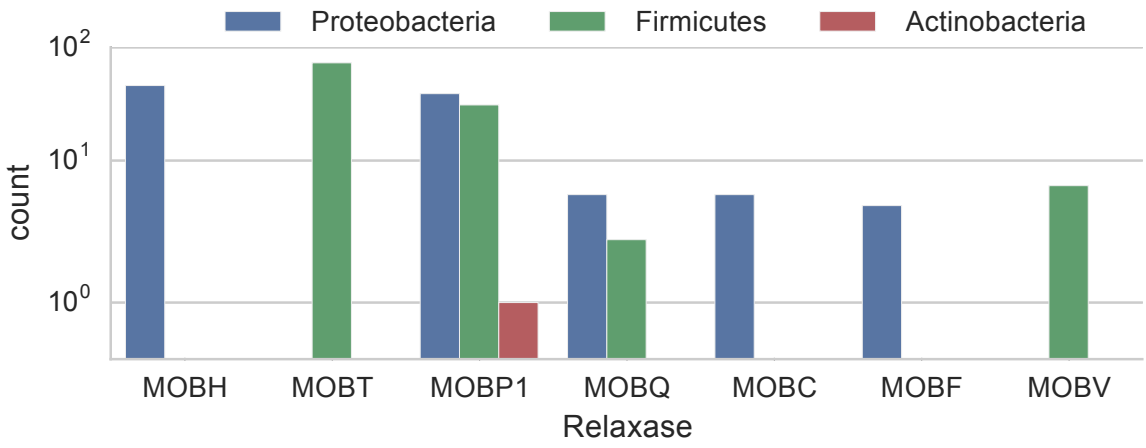

Figure S5: Repartition of the different type of MOB found in the different phyla.

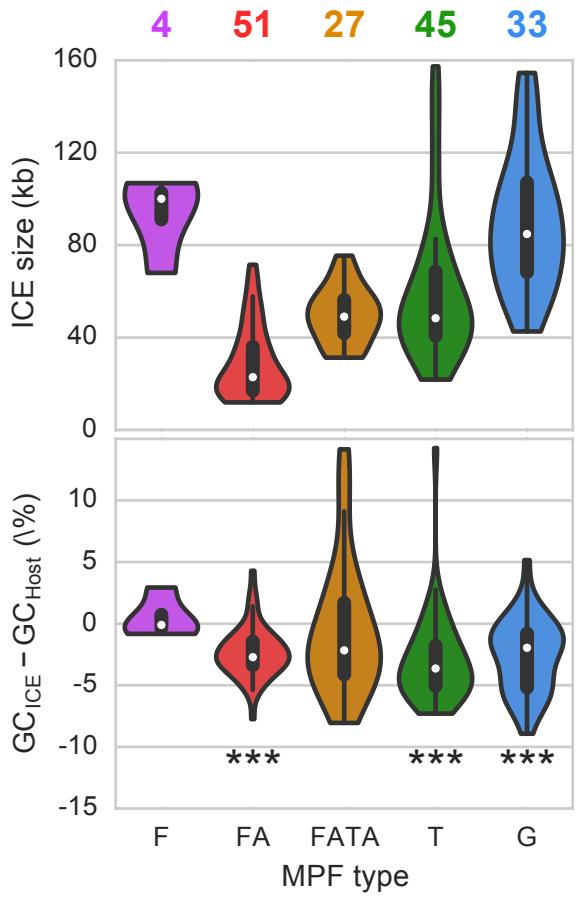

Figure S6: ICE family statistics as a function of the MPF type. This figure is the same as Figure 2, but without possibly redundant ICE from the same family with a threshold of wGRR>90 to define the families.
Top. Distribution of the size of ICEs families (mean size per family in kb). The numbers above each violin plot represent the number of ICE families in each category.
Bottom. Distribution of pairwise differences between the mean GC content of the ICE family and that of its host.
The violin plots represent the kernel density estimation of the underlying values. Here the violin plots is limited by the minimum and maximum values.
***: p-value<0.001, Wilcoxon signed-rank test (rejecting the null hypothesis that the difference is equal to zero).

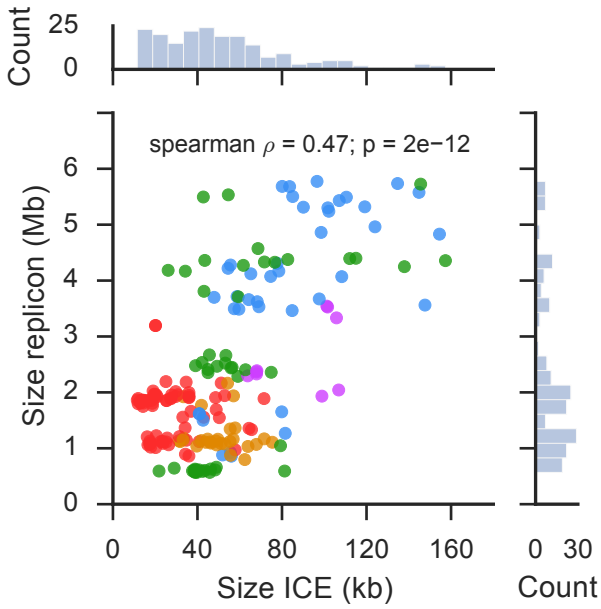

Figure S7: Scatter plot of the size of the host's replicon as a function of the size of the ICE. In case of multi chromosome, the size is that of the chromosome in which is inserted the ICE. Colors correspond to different MPF types, and match the color code from the other figures (F: purple; FA: red, FATA: orange; T: green; G: blue). The association between replicon and ICE sizes is significant in general, and individually for the types T and G (green and blue, $\rho>0.54$, p-values$<10^{-4}$), but not for the others (all $\rho<0.23$ and p-value>0.4). The histograms at the edges of the scatterplot show the distribution of the points for each variable.

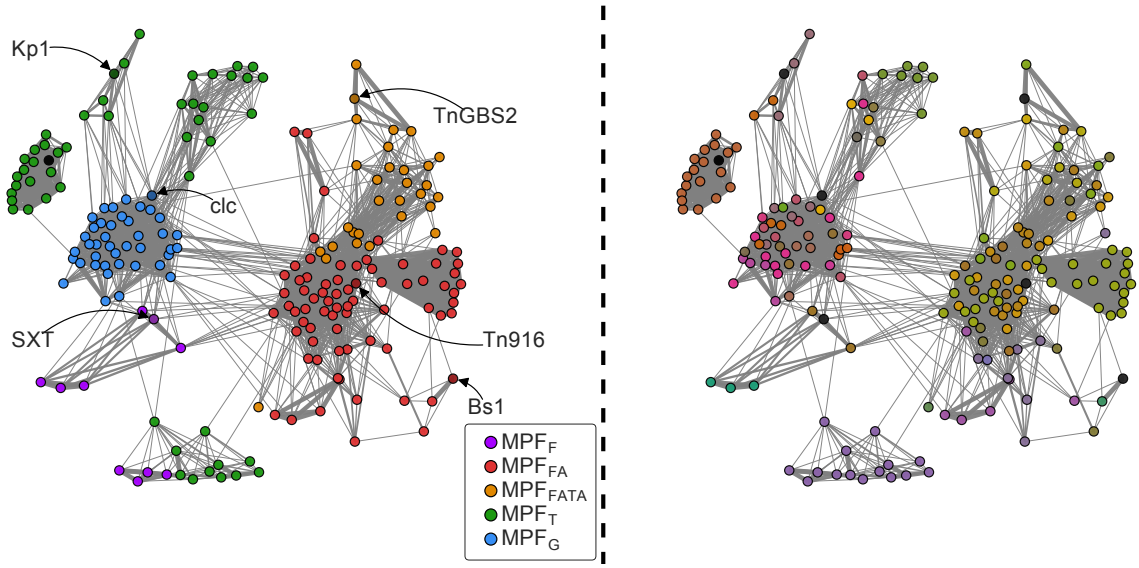

Figure S8 ICE network (same as Figure 3 but ignoring MPF genes to calculate the wGRR). Representation of the wGRR-based network of ICEs. The nodes represent the ICEs and the edges link pairs of ICEs with wGRR score above 5% (the thickness of the edge is proportional to the score). Nodes are colored after the MPF type. Darker nodes represent model ICEs.

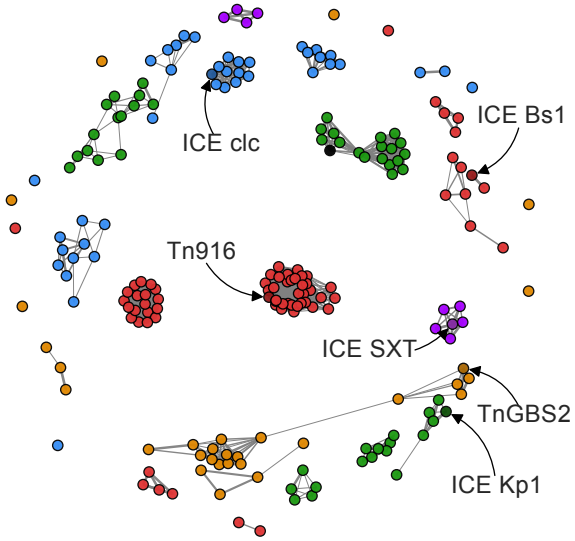

Figure S9: ICE network (same as Figure 3 but with a threshold wGRR>30%). Representation of the wGRR-based network of ICEs. The nodes represent the ICEs and the edges link pairs of ICEs with wGRR score above 30% (the thickness of the edge is proportional to the score). Nodes are colored after the MPF type. Darker nodes represent model ICEs.

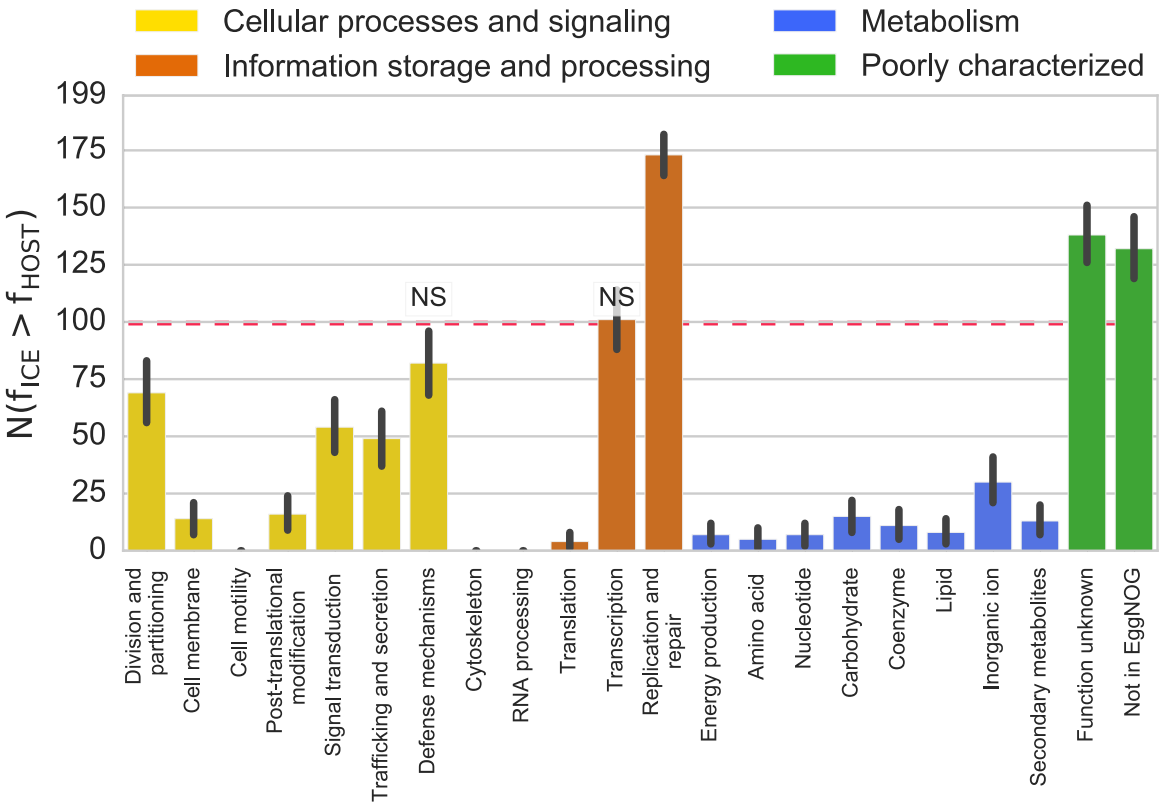

Figure S10: Representation of EggNOG functional categories in ICEs relative to the host chromosome (ignoring MPF genes in the analysis).

The bars represent the number of times a given category is found more frequently in an ICE than in its host chromosome ($N(f_{ICE}>f_{HOST})$). The red dotted line represents the expected value under the null hypothesis, where a category is in similar proportion in ICE and its host's chromosome. Bars marked as NS represent a lack of significant difference ($p>0.05$, Binomial test with 199 trials and expected value of 0.5), whereas the others are all significantly different ($p<0.05$, same test). Note that there are 199 trials because one of the 200 ICEs could not be types and was thus excluded, see text. Error bars represent 95% confidence interval computed with 1000 bootstraps. "Not in EggNOG" represents the class of genes that didn't match any EggNOG profile.

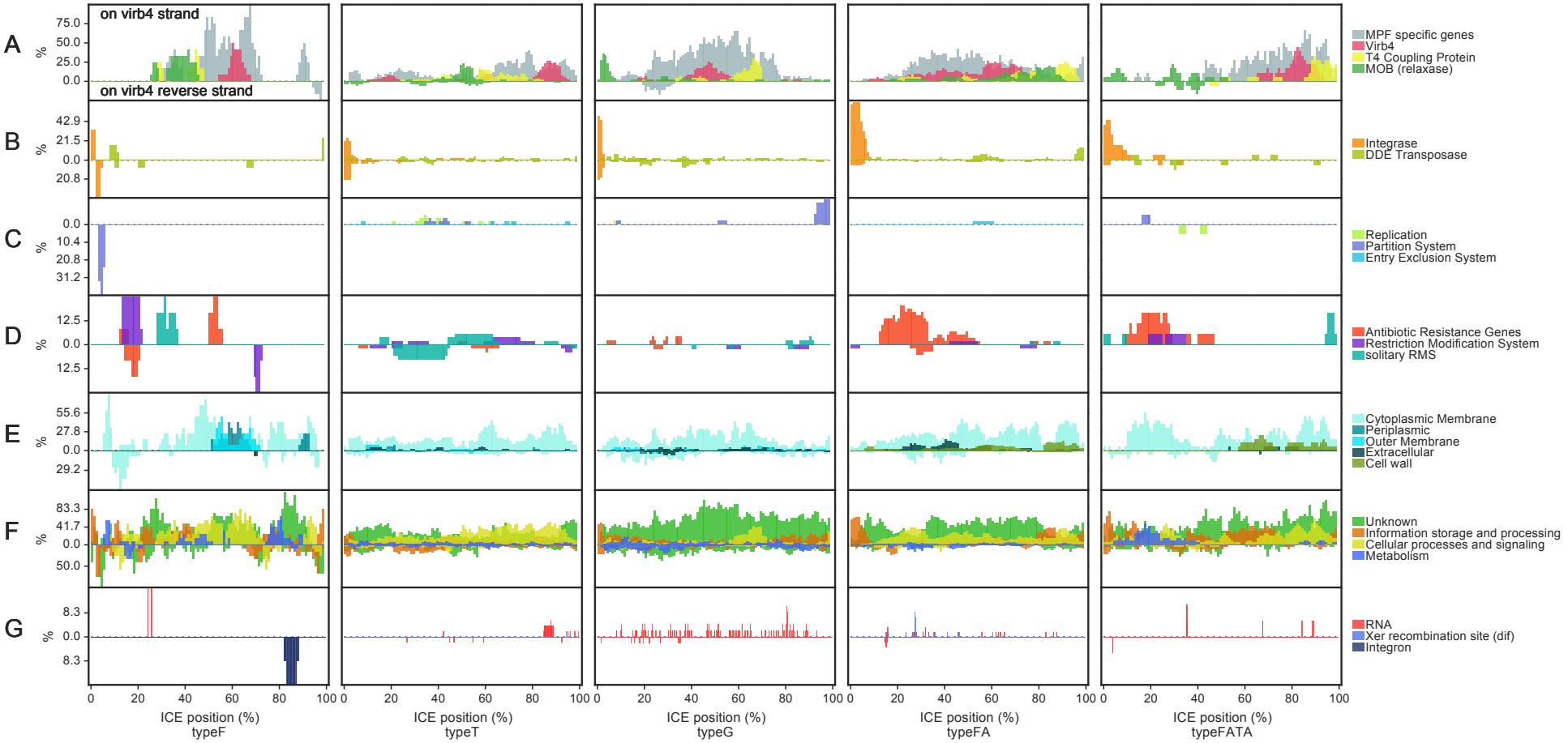
A
B
C
D
E
F
G
%
on virb4 strand
on virb4 reverse strand
75.0
50.0
25.0
0.0
42.9
21.5
0.0
20.8
0.0
10.4
20.8
31.2
12.5
0.0
12.5
55.6
27.8
0.0
29.2
83.3
41.7
0.0
50.0
8.3
0.0
8.3
0
20
40
60
80
100
ICE position (%)
typeF
typeT
typeG
typeFA
typeFATA
MPF specific genes
Virb4
T4 Coupling Protein
MOB (relaxase)
Integrase
DDE Transposase
Replication
Partition System
Entry Exclusion System
Antibiotic Resistance Genes
Restriction Modification System
solitary RMS
Cytoplasmic Membrane
Periplasmic
Outer Membrane
Extracellular
Cell wall
Unknown
Information storage and processing
Cellular processes and signaling
Metabolism
RNA
Xer recombination site (dif)
Integron

Figure S11: Organization of ICEs

Each column corresponds to a given MPF type. Each row corresponds to a given set of functions. In each panel, the functions are mapped at their relative position in a given type of ICE. When the same function is found many times in the same type of ICEs, the height of the bars increases. The height is then normalized to the number of ICE in a given MPF type. The values of the bar are positive if the corresponding function is found on the same strand than virb4, and negative otherwise.

The peak of RNA in type T corresponds to the presence of a CRISPR array.

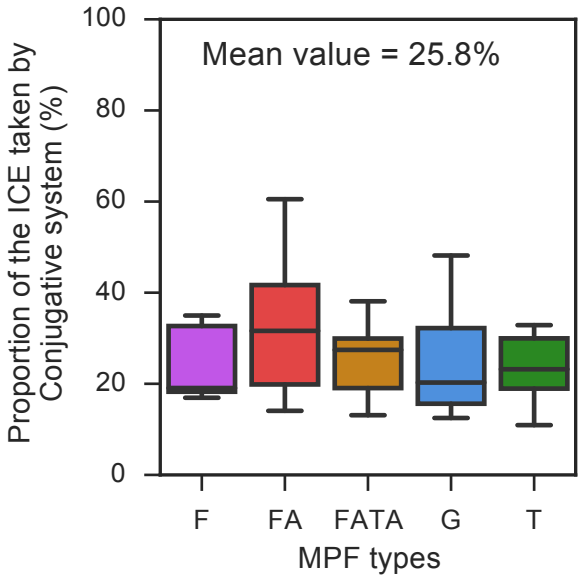

Figure S12 - Distribution of the proportion of the ICE taken per the conjugative system for the different MPF types.

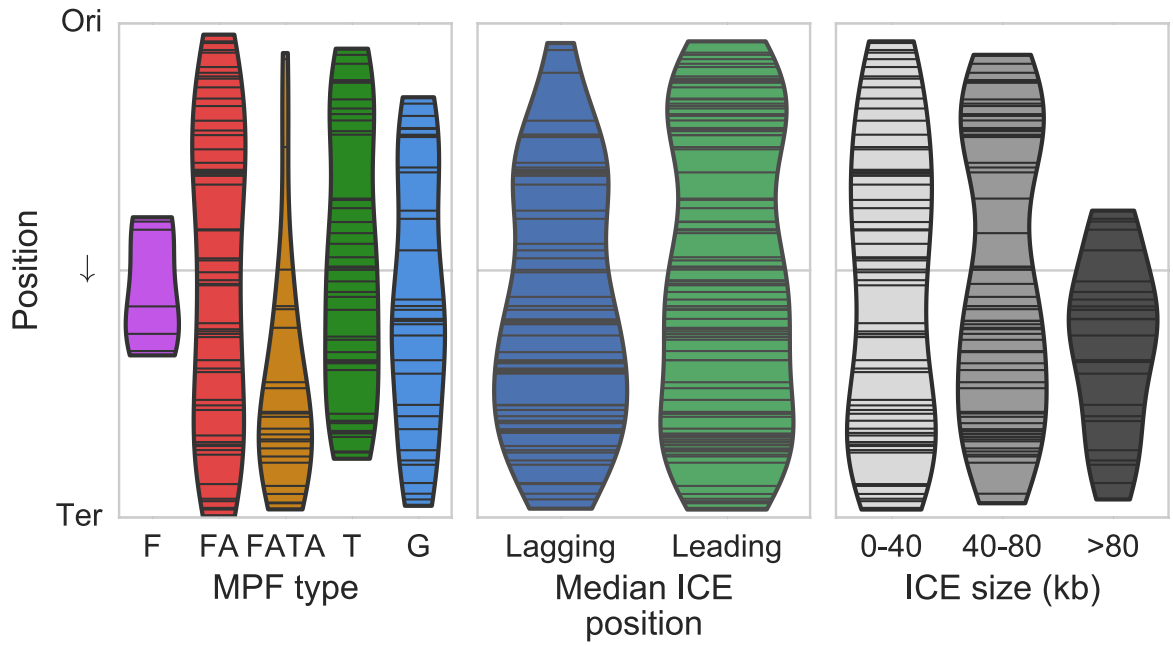

Figure S13: Position of ICEs along the Ori-Ter axis according to their MPF type.

Left: distribution of the ICE locations as a function of the MPF type. Only type FATA is more frequent towards the half closer to the Ter region (χ^2 test, p-value<0.001). Center: distribution of ICE locations as a function of its median strand position. Right: distribution of ICE locations as a function of the size category of the ICE.

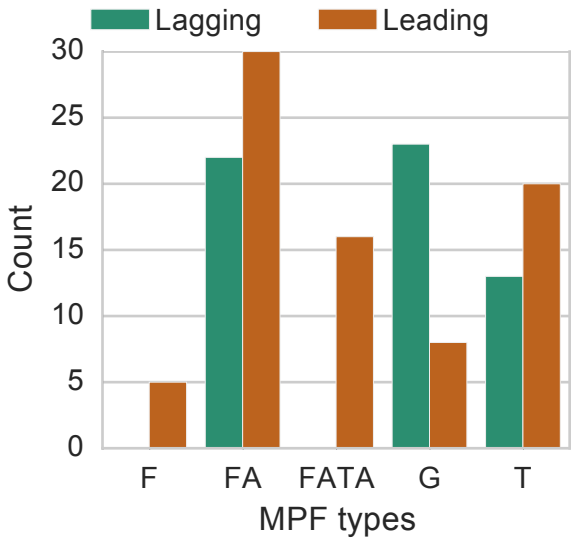

Figure S14: Number of ICEs with the majority of its genes on the leading (brown) or lagging (green) strand (relative to chromosome replication) for each MPF type.

Table S7: Comparisons of key traits of MPF T ICEs from the main dataset and from the novel genomes for post hoc validation

Database / Measure	November 2013	November 2016	p-value (Bonferroni corrected)	Statistical test
Number of delimited ICE	49	124	NA	NA
Mean size (kb)	58.5	58.6	> 0.05	Wilcoxon rank-sum test
Proportion of ICE with integrase (%)	85.7	96.0	> 0.05	Fisher exact test
GC% difference with host	-2.74	-2.05	> 0.05	Wilcoxon rank-sum test
Mean *virb4* – integrase distance (kb)	29.7	31.3	> 0.05	Wilcoxon rank-sum test

Supplementary information on HMM construction

Partition systems

ParAB

1. We started with proteins with experimental evidences from (Gerdes 2000)
2. We aligned them (mafft --auto)
3. We made HMM profiles with them (hmmbuild default option)
4. We searched on all proteins (RefSeq 2013 with hmmsearch default option)
5. We got hits with e-value<1e-2 for parA and parB type Ib and parB type Ia (because more degenerated) and evalue<1e-3 for parA typeIa, and a coverage of 50%.
6. We got hits co-localizing with the other protein of the same type with 1 gene in between maximum.
7. We removed redundancy above 90% of identity (uclust)
8. We made protein families at 40% identity (Blastall + Silix)
9. We re-did steps 2-6 with each protein families (except threshold for step 5: all e-values < 1e-2)
10. We searched against all protein from a database of plasmids (Ref Seq 2015, hmmsearch default option)
11. We gathered all the hits at ± 2 genes around parA genes that are not near a parB genes
12. We made protein families at 40% identity (blastall + silix)
13. We manually checked whether the annotation could be partition protein. (For families with >40 proteins for parA type Ia and >15 for par B type Ib)
14. We selected 4 families for parAB type Ia and 3 for parAB type Ib which have annotations related to partition system
15. We made hmm profiles with them
16. We added them to the hmm profiles for parB type Ia and par B type Ib. We also added parG profile from Pfam. (PF09274)

ParMR

We used the same pipeline as for parAB, but:

- Steps 2-6 were done 3 times (instead of 2).
- 2 profiles PFAM were added (PF06406 (parM) and PF10784 (parR))
- Steps 11 to 15 led to nothing.

TubZR

We used the same pipeline as for parAB, but:

- We started with proteins with experimental evidence. Only 4 tubZ was available from Bacillus genus. (Larsen 2007)
- We took proteins near those tubZ having expected tubR size (about 100 aa).
- We used wHTH profile ((PF10771) - Winged helix-turn-helix domain), supposed to match tubR, but it did not really help.

- Steps 2-6, were done 4 times. One repeat was done with blast instead of hmmsearch.
- Steps 11 to 15 led to nothing.

Replication systems

Replication initiation proteins

We started with proteins from Lanza et al. (Lanza 2014), and added some PFAM profiles associated with plasmid replication:

- Rep_trans PF02486.14
- RepA_C PF04796.9
- Rep_1 PF01446.14
- Rep_2 PF01719.14
- Rep_3 PF01051.18
- Rop PF01815.11
- Replicase PF03090.12
- RepC PF06504.8
- rep_Rpt-F
- rep_Rpt-A3 -
- rep_Rpt-A2
- rep_Rpt-A1
- rep_Rpt-L3
- rep_Rpt-Z
- rep_Rpt-A5
- rep_Rpt-C
- rep_Rpt-14
- rep_Rpt-L2
- rep_Rpt-A4
- rep_Rpt-L1
- rep_null

We ran hmmsearch default values on the entire database of proteins, and selected hits whose coverage of the hit was above 50% of the length of the profile, and if the e-value was less than 10^{-6}

We removed duplicated proteins (hit by different profiles), and kept the profiles with the best e-value.

Rolling Circle Replication proteins

We used the DPR database (Database of Plasmid Replicons) from Dr. Mark Osborn, however the site does not exist anymore, but we retrieved the data from the Internet cache (archived in 2009). We built HMM profiles with 15 out of 17 groups, after making the alignment with mafft (--auto). Then, we used hmmsearch with default parameters, and filter out hits with coverage of the profile being less than 50% and evalue above 10^{-6}.

Restriction and modification systems

We use data from Oliveira et al., NAR, 2014 to build HMM profiles for RMS, which we used to search for complete RMS. We were able to find more RMS in ICEs than with the original dataset, while all the RMS from the original dataset were found. The additional hits are likely to be true positives given the e-value threshold at 10^{-6}, the coverage of 80% of the profile and the association of two hits at a distance of 4 genes.

We use the set of solitary methylases from the above-mentioned article.