Host range expansion and genetic plasticity drive the trade-off between integrative and extrachromosomal mobile genetic elements

Jean Cury[1,2], Pedro H. Oliveira[1,2,*], Fernando de la Cruz[3], Eduardo P.C. Rocha[1,2]

Affiliations.

[1]Microbial Evolutionary Genomics, Institut Pasteur, 28, rue Dr Roux, Paris, 75015, France,
[2]CNRS, UMR3525, 28, rue Dr Roux, Paris, 75015, France,
[3] Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC),

Universidad de Cantabria, Calle Albert Einstein 22, 39011 Santander, Spain

*Currently at: Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA

Classification: Biological Sciences/Evolution

Corresponding author: Jean Cury (jean.cury@normalesup.org)

Abstract

Self-transmissible mobile genetic elements drive horizontal gene transfer between prokaryotes. Some of these elements integrate in the chromosome, whereas others replicate autonomously as plasmids. Recent works showed the existence of few differences, and occasional interconversion, between the two types of elements. Here, we enquired on why evolutionary processes have maintained the two types of mobile genetic elements by comparing integrative and conjugative elements (ICE) with extrachromosomal ones (conjugative plasmids) of the highly abundant MPF_T conjugative type. Expectedly, we found that plasmids encode more replicases and partition systems, whereas ICEs encode more integrases. Yet, the two types of elements often clustered together in terms of gene repertoires because they encode both types of functions. Plasmids and ICEs encode similar accessory functions, but in different quantity; ICEs encode more metabolic genes and conjugative plasmids over-represent antibiotic resistance genes. We hypothesized that plasmids might be more plastic, whereas ICEs could have broader host ranges. Indeed, plasmids are more variable in size, have more DNA repeats, and exchange genes more frequently. On the other hand, ICEs are more frequently transferred between distant taxa. We found few very similar ICEs and plasmids, but when they are present in distant bacterial hosts, their composition suggests that long-range transfer of a plasmid can be stabilized if it becomes an ICE. This work suggests that the differential plasticity and transmissibility of the elements drives their abundance and functional content, thus explaining the co-existence of plasmids and integrative elements in Prokaryotes.

Keywords: Mobile genetic elements, horizontal gene transfer, molecular evolution, microbial genomics, conjugation

39 ## Significance statement

40 Horizontal transfer of genetic information between bacteria drives their rapid adaptation
41 and is often mediated by mobile genetic elements. We identify differences between
42 conjugative elements that may explain why some of them integrate in the chromosome
43 (ICE) whereas others remain as independent replicons in the cell (plasmids). We show that
44 plasmids are more plastic: they have wider variations in gene repertoires, more repeats, and
45 transfer more genes between elements. On the other hand, ICEs have broader host ranges.
46 Plasmids transferring to very distinct hosts seem to require integration – as ICEs - to be
47 maintained in the genome. This contributes to elucidate the longstanding question of why
48 there is stable co-existence of integrative and extrachromosomal mobile elements in
49 Prokaryotes.

50 ## Introduction

51 The genomes of Prokaryotes have mobile genetic elements (MGEs) integrated in the
52 chromosome or replicating as extrachromosomal elements. These MGEs usually encode
53 non-essential but ecologically important traits (1, 2). Extra-chromosomal elements, such as
54 conjugative plasmids (CPs) and lytic phages, replicate autonomously in the cell using
55 specialized replicases to recruit the bacterial DNA replication machinery (or to use their
56 own). Plasmids and extra-chromosomal prophages can also increase their stability in cellular
57 lineages using partition systems, for proper segregation during bacterial replication (3),
58 resolution systems, to prevent accumulation of multimers (4), and restriction-modification
59 or toxin-antitoxin systems, for post-segregation killing of their hosts (5). Alternatively, many
60 MGEs integrate into the chromosome. This is the case of the vast majority of known
61 prophages, of most conjugative elements (ICEs), and of many elements with poorly
62 characterized mechanisms of genetic mobility (*e.g.*, many pathogenicity islands)(6–8). The
63 integrated elements are replicated along with the host chromosome and require an
64 additional step of excision before being transferred between cells. This raises a question
65 that has not been addressed in the literature: what are the relative benefits and
66 disadvantages of the integrated and extrachromosomal forms of a MGE?

67 To address this question, we analyzed the differences and similarities between ICEs and CPs.
68 We focused on these elements because both forms are frequently found in bacteria, they
69 can be easily detected in genomes, and the mechanism of conjugation is well known.
70 Conjugative elements have a crucial role in spreading antibiotic resistance and virulence
71 genes among bacterial pathogens (9–12). Recently, several works suggested that the line
72 separating integrative ICEs and CPs could be thinner than anticipated (13), because some
73 ICEs encode plasmid-associated functions like replication (14) or partition (15), some
74 plasmids encode integrases (16), and ICEs and CPs are intermingled in the phylogenetic tree
75 of conjugative systems (17). Finally, both forms – ICEs and CPs – are found throughout the
76 bacterial kingdom, but their relative frequency depends on the taxa and on the mechanisms
77 of conjugation (7). These differences suggest that conjugative elements endure diverse
78 selective pressures for being integrative or extrachromosomal depending on unknown
79 environmental, genetic, or physiological variables.

80 We thought that key differences in the biology of integrative and extrachromosomal
81 elements might provide them with different types of advantages. ICEs require an additional
82 step of integration/excision during transfer, which may take time and requires genetic
83 regulation. Their integration in the chromosome may affect the latter's organization and
84 structure, and these collateral effects might increase with the size of the element. On the
85 other hand, ICEs replicate as part of chromosomes and could thus be lost from the cell at
86 lower rates than plasmids. Furthermore, plasmids must recruit the host replication
87 machinery, which may render elements incompatible and is known to constrain their host
88 range: many plasmids are able to conjugate into distantly related hosts, but are unable to
89 replicate there (18–20). We thus hypothesize that ICEs might be favored when transfers
90 occur between distant hosts, whereas plasmids might provide more genetic plasticity
91 because their size is not constrained by chromosomal organization.

92 Here, we study conjugative elements of the type MPF_T. This is the most frequent and best-
93 studied type of conjugative systems (17), and the only one for which we can identify
94 hundreds of elements of each of the forms (ICEs and CPs). We restricted our analysis to
95 genera containing both CPs and ICEs, to avoid, as much as possible, taxonomical biases. We
96 first describe the content of both types of elements and highlight their differences and
97 similarities. Next, we quantify their genetic similarity and the extent of their gene
98 exchanges. Finally, we show that chromosomal integration facilitates the colonization of
99 novel taxa by a conjugative element.

100

101

102 ## Results

103 ### Functional and genetic differences between ICEs and CPs

104 We analyzed a set of 151 ICEs and 136 CPs of the same genera and of type MPF_T, most of
105 which were from Proteobacteria (96.9%). Both ICEs and CPs were found to be AT-richer than
106 their host chromosomes, which is a common feature in MGEs and horizontally transferred
107 genes (21). However, the difference was three times smaller for ICEs (Fig. 1A), presumably
108 because they replicate with the chromosome or remain a longer time in the same host. The
109 average size of CPs is slightly larger (75kb vs 59kb), and the median slightly smaller (46kb vs
110 52kb) than that of ICEs. In contrast, CPs have more diverse sizes than ICEs (Fig. 1B), showing
111 a coefficient of variation twice as large (1.05 vs 0.49). The size of the conjugative elements
112 depends on the size of the bacterial genome (after discounting the size of their conjugative
113 elements), this effect being much stronger for CPs (Fig. 1C). CPs also have four times higher
114 density of large DNA repeats than ICEs (Fig. 1D). These results suggest that CPs diversify
115 faster than ICEs.

116 HGT concentrates in a few hotspots in bacterial chromosomes, presumably to minimize
117 disruption in their organization (22). We used HTg50, a measure of the concentration of
118 HGT in chromosomes that corresponds to the minimal number of spots required to account
119 for 50% of horizontally transferred genes (22), to test if chromosomes with fewer
120 integration hotspots had more plasmids. Indeed, there is a negative association between
121 the number of plasmids, weighted by their size, and the chromosomes' HTg50 (Spearman
122 ρ=-0.35, p-value=0.0016, Fig. S1).

123 We then quantified the differences between ICEs and CPs in terms of functions associated
124 with their biology, with a focus on stabilization functions. Relaxases are part of the rolling
125 circle replication initiator proteins and some have been shown to act as replicases(14, 23) or
126 site-specific recombinases (24, 25). Since all conjugative elements encode a relaxase, by
127 definition, they may also have these functions. In the following, we focused on typical
128 plasmid replication initiator proteins (more than 95% of them are involved in theta-
129 replication, and none is matched by the protein profiles of relaxases), and serine or tyrosine
130 recombinases as integrases. Expectedly, ICEs showed higher frequency of integrases, while
131 CPs had more frequently identifiable partition and replication systems. Some ICEs encode

132 partition systems (11%) and many encode a replicase (40%), while 37% of CPs encode at
133 least one tyrosine or serine recombinase (Fig. 2A). These results further illustrate a
134 continuum between the two types of elements: about half of the elements (40% and 48%,
135 ICEs and CPs, respectively) have functions usually associated with the other type and may
136 (rarely) lack functions typically associated with its own type (Fig. 2B). We identified plasmid
137 incompatibility systems of diverse types, whereas no ICE could not be typed in the current
138 scheme (Fig. S2).

139 We then made similar analyses for functions usually regarded as accessory or unrelated to
140 the biology of MGEs (Fig. 2B). ICEs were more likely to carry restriction-modification systems
141 (x2.8) than CPs (but not orphan methylases), suggesting that ICEs endure stronger selective
142 pressure for stabilization in the genome. In contrast, they were significantly less likely to
143 carry antibiotic resistance genes or integrons. They also had fewer identifiable entry-
144 exclusion systems, which may reflect the ability of ICEs to tolerate the presence of multiple
145 similar elements in the cell (26). The classification of genes in the four major functional
146 categories of the EggNOG database, showed that ICEs had relatively more genes encoding
147 metabolic and cellular processes. We have previously shown that genes of unknown or
148 poorly characterized function were over-represented in ICEs relative to their host
149 chromosome (27). The frequency of these genes is even higher in plasmids (61% vs 46%).
150 Hence, both types of elements have many functions in common, but their relative frequency
151 often differs significantly.

152 ## Genetic similarities between ICEs and CPs

153 The results of the previous section, together with previously published studies (see
154 Introduction), suggest that ICEs and CPs either share a common history or often exchange
155 genes (or both). We detailed the relationships of homology between ICEs and CPs using the
156 weighted Gene Repertoire Relatedness (wGRR), which measures the frequency of bi-
157 directional best hits between two elements weighted by their sequence similarity (see
158 Methods). We clustered the matrix of wGRR using the Louvain algorithm (28), and found six
159 well-distinguished groups (Fig. 3). Two groups (1 and 6) are only constituted of CPs, two are
160 composed of more than 90% of ICEs (3 and 5) and two have a mix of both types of elements
161 (2 and 4) (Fig.3, top bar). Bacterial species are scattered between groups, showing that they

162 are not the key determinant of the clustering. Some groups are only from γ-proteobacteria,
163 but others include bacteria from different classes. Groups where elements are from the
164 same taxonomic classes tend to have either CPs or ICEs, whereas the others have mixtures
165 of both elements. Group 4, includes many ICEs and CPs, where all ICEs have integrases while
166 more than half of the CPs lack both replication and partition systems (Fig. 4). This group
167 includes almost only elements from ε-proteobacteria that may have specificities that we
168 were not able to take into account. In contrast, almost all ICEs of groups 2 and 3 encode an
169 integrase and all CPs have partition or replication systems.

170 We controlled for the effect of the MPF genes in the previous clustering analysis by re-doing
171 it without these genes (Fig. S3). This produced the same number of groups - N1 to N6 - and
172 90% of the elements of the former groups were classed in the same novel groups (Fig. S4).
173 The only qualitatively significant difference between the two analyses concerned the group
174 2 for which 36% of the elements were now classed in groups N4 or N6. Overall, these
175 controls confirm that ICEs and CPs can be grouped together, and apart from other elements
176 of the same type. The grouping is not caused by sequence similarity between conjugative
177 systems. Instead, it probably reflects either within group genetic exchanges between ICEs
178 and CPs, or interconversions of the two types of elements.

179 ## Genetic exchanges: CPs become ICEs for broader host range

180 The clustering of ICEs and CPs could be explained by genetic transfer between them. To
181 address this question, we searched for pairs of conjugative elements with low wGRR (<30%)
182 but some highly similar homologs (>80% sequence identity). This identified several cases of
183 recent transfer of a single or a few genes between elements (Fig. 5A). In agreement with our
184 observations of higher genetic plasticity in CPs, most transfers took place between these
185 types of elements (Fig. 5.B-D).

186 We hypothesized that ICEs could hold an advantage over CPs to colonize novel hosts,
187 because replication restricts plasmid host range. To test this hypothesis, we analyzed the
188 wGRR between pairs of ICEs and pairs of CPs in function of the phylogenetic distance
189 between their bacterial hosts. This showed similar patterns for the two types of elements,
190 with the notable exception that there are no pairs of highly similar plasmids (wGRR>50%) in
191 distant hosts (more than 0.1 substitutions/position, *e.g.,* the average distance in the tree

192 between *E. coli* and *P. aeruginosa*). In contrast, a third of all ICEs (n=50) are in these
193 conditions (Fig. 6A, Fig. S5). The same analysis after removing the MPF genes shows wGRR
194 values shifted to lower values for all elements, but qualitatively similar trends (Fig. S6). This
195 suggests a major difference in the ability of ICEs and CPs to be stably maintained after their
196 transfer into a distant host.

197 We then analyzed the pairs ICE-CP. We found few pairs of highly similar ICEs and CPs in
198 closely related hosts (bottom right corner of Fig. 6, n=8 for wGRR>50% and $d < 10^{-2}$),
199 suggesting that interconversion between these elements remains rare within a clade. A
200 larger number of ICE-CP pairs were very similar but present in distant hosts (n=38, Fig. 6).
201 The most parsimonious explanation for these observations, is the recent transfer of one of
202 the elements (ICE or CP) to a distant bacterial host. We identified the latter element based
203 on the differences in terms of tri-nucleotide composition between the elements and the
204 host chromosomes (defined as pvalue in (29)). We then computed for each ICE-CP pair the
205 difference between the pvalue of the pair ICE-host and that of the pair CP-host (see
206 Methods). In agreement to our observation that ICEs have broader host ranges, these
207 differences indicate that ICEs are relatively more distant from the host chromosome for
208 pairs with high wGRR in distant hosts than for the rest of the pairs (Wilcoxon rank sum test,
209 p-value$<10^{-20}$, Fig. 6B, and Fig. S7). The rarity of ICE-CP pairs in closely related hosts, their
210 abundance in distant hosts, and the identification that ICEs are the most compositionally
211 atypical relative to the host in the latter, suggest that successful transfer of CPs to distant
212 hosts is favored when they integrate the chromosome and become ICEs.
213

214 # Discussion

215 In this study, we quantified the differences between ICEs and CPs to evaluate the claims that
216 they are essentially equivalent MGEs (13, 14, 27). We found that numerous CPs have
217 integrases (although these may serve for dimer resolution and not integration (30)),
218 numerous ICEs encode replication and partition functions, the elements often cluster them
219 together, and they exchange genetic information. Furthermore, relaxases – present in both
220 ICEs and CPs – have been shown to act as integrases or replication initiators (31). Hence,
221 ICEs and CPs constitute closely related elements sharing many functions beyond those
222 related to conjugation. Yet, there are also some clear differences between them (Fig. 7).
223 First, genes encoding plasmid replicases and partition systems are more frequent in
224 plasmids, and tyrosine and serine recombinases are more frequent in ICEs. Interestingly, we
225 could not attribute incompatibility groups to ICEs, suggesting that the replication module is
226 rarely exchanged between ICEs and CPs. Second, the frequency of certain accessory traits is
227 different: plasmids are more likely to encode antibiotic resistance genes whereas ICEs
228 encode more metabolism-related genes, even if this could result from biases in the
229 database towards nosocomial pathogens. Finally, the %GC relative to the host, the number
230 of repeats, the patterns of gene variation and exchange, and the host range are
231 quantitatively different in the two types of elements. After integrating all this information,
232 we propose that in spite of their similarities each type has specific advantages.

233 Even if there are some known families of large (>200kb) ICEs (32, 33), most of these
234 elements in our dataset have a narrower variation in size than CPs. This suggests that CPs
235 are more flexible than ICEs in terms of the amount of genetic information they can carry and
236 in their ability to accommodate novel information. CPs also exchange genes more
237 frequently. Mechanistically, the rate of recombination between plasmids may be higher
238 because they encode more repeats, integrons, and transposable elements. Plasmid copy
239 number, when high, may also contribute to increase recombination rates. Interestingly,
240 recombination mediated by transposable elements has been shown to be drive the
241 evolution of certain plasmids(34, 35), and to accelerate the reduction of plasmid cost thus
242 stabilizing the element after horizontal transfer (36). The restrictions in size variation of ICEs
243 are probably not due to the mechanism of integration or excision because such reactions

244 occur between very distant recombination sites (37). Instead, very large ICEs may disrupt
245 chromosome organization by affecting the distribution of motifs, changing chromosome
246 folding domains, or unbalancing the sizes of replichores (38). Repeat-mediated
247 recombination leads to replicon rearrangements, and may lead to stronger counter-
248 selection of DNA repeats in ICEs than in CPs, further restricting their size variation and their
249 flux of gene exchange. Interestingly, the size of plasmids varies much more steeply with
250 genome size than the size of ICEs, suggesting that CPs may play a particularly important role
251 in the evolution of large bacterial genomes of Proteobacteria, which have higher rates of
252 genetic exchanges (39), and often contain mega-plasmids (40).

253 Some plasmids are known to be broad-host range and adapt to novel hosts, especially if
254 they carry adaptive traits that compensate for the initially poor intrinsic persistence of the
255 element (41, 42). However, within the large phylogenetic span considered in this work,
256 MPF_T ICEs have broader host ranges. Actually, the first known ICE, Tn916 (not MPF_T, thus
257 not included in this study), became notorious due to its ability to spread antibiotic
258 resistance between distant phyla (43). We thus propose that ICEs are the ones
259 compositionally atypical in the pairs of very similar elements ICE-CPs present in distant
260 hosts, because CPs conjugating to distantly related bacteria have a higher probability of
261 being maintained in the novel lineage if they become ICEs.

262 Other traits may provide advantages to ICEs or CPs. The ability of plasmids to modify their
263 copy number may accelerate adaptive evolutionary processes, such as the acquisition of
264 antibiotic resistance (44). On the other hand, ICEs might be more stably maintained in
265 lineages because they replicate within the chromosome. Finally, the carriage of ICEs and CPs
266 may have different costs. The cost of plasmids has been extensively studied and is strongly
267 dependent on the traits they encode (45). Much less is known about the cost of ICEs; several
268 reports suggest that they lead to low fitness costs when conjugation is not expressed, but
269 their fitness cost varies much more between elements during transfer(12). Direct
270 comparisons of the cost of carriage of ICE and CPs carrying similar traits are unavailable.
271 Further work will be needed to test these hypotheses.

272 Occasional transfers between CPs and ICEs allow them to access the other elements' gene
273 pool. These events may create elements with traits of ICEs and of CPs, as observed in more

274 than a third of all conjugative elements, and lead to their clustering in the wGRR group 5.
275 Additionally, they facilitate the interconversion of one type of element into the other.

276 Many elements are mobilizable but not able to conjugate independently (40, 46). These
277 elements often encode a relaxase that recognizes the element's origin of transfer and is able
278 to interact with a T4SS from an autonomously conjugative element to transfer to other cells.
279 Many of the disadvantages of conjugative plasmids and ICEs are similar to those of
280 mobilizable plasmids and integrative mobilizable elements, whether they encode a relaxase
281 or not. Notably, the former must be replicated in the extrachromosomal state, and the
282 latter integrate the genome where they must not disrupt genome organization. Patterns
283 observed in conjugative elements are thus likely to be applicable to mobilizable ones.

284 These results may also be relevant to understand lysogeny by temperate phages. The vast
285 majority of known prophages are integrated in the chromosome, but some replicate like
286 plasmids (47, 48). Considering that prophages share some of the constraints of conjugative
287 elements, they are likely to be under similar trade-offs. However, phages are under
288 additional constraints. Notably, their genome size is much less variable than that of
289 conjugative elements, because it must be packaged into the virion (49), and this may render
290 the extrachromosomal prophages less advantageous in terms of accumulating novel genes.
291 This could explain why most prophages are integrative whereas conjugative systems are
292 more evenly split between integrative and extrachromosomal elements.
293

294 ## Material and Methods

295 **Data**. Conjugative systems of type T (MPF_T) were searched in the set of complete bacterial
296 genomes from NCBI RefSeq (http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/, last
297 accessed in November 2016). We analyzed 5562 complete genomes from 2268 species,
298 including 4345 plasmids and 6001 chromosomes. The classification of the replicon in
299 plasmid or chromosome was taken from the information available in the GenBank file. Our
300 method to delimit ICEs is based on comparative genomics of closely related strains. Hence,
301 we restricted our search for conjugative systems to the species for which we had at least
302 five genomes completely sequenced (164 species, 2990 genomes).

303 **Detection of conjugative systems and delimitation of ICEs**. Conjugative systems were
304 detected using the CONJscan module of MacSyFinder (50), with protein profiles and
305 definitions of the MPF type T, published previously (51). ICEs were delimited with a
306 comparative genomics method, as developed in a previous work (27).

307 **Weighted gene repertoire relatedness**. This score represents the number of homologous
308 proteins between two elements, weighted by their sequence identity, as described in (27).
309 The formula is:

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

310 Where (A_i, B_i) is the ith pair of homologous protein between element A and element B,
311 $id(A_i,B_i)$ is the sequence identity of their alignment, min(A,B) is the number of proteins of
312 the element with fewest proteins (A or B). The sequence identity was computed with blastp
313 v.2.2.15 (default parameters)(52) and kept all bi-directional best hits with an e-value lower
314 than 10^{-5}. The groups were made using the Louvain algorithm (28). We controlled for the
315 consistency of the heuristic used to assess that the group found are not form a local
316 optimal. We performed 100 clustering, which led to the same classification in 95% of the
317 time.

318

Acknowledgements

This work was supported by an European Research Council grant [EVOMOBILOME n°281605], and a grant from the Agence National de la Recherche [MAGISBAC, ANR-14-CE10-0007]. Work in FdlC lab was supported by grants BFU2014-55534-C2-1-P and BFU2014-62190-EXP from the Spanish Ministry of Economy and Competitiveness. We thank Alan Grossman and Marie Touchon for comments and suggestions, and Aude Bernheim for providing the phylogenetic tree of proteobacteria. J.C. is a member of the 'Ecole Doctorale Frontière du Vivant (FdV) – Programme Bettencourt'.

327

328 # References

329 1. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: the
330 agents of open source evolution. *Nat Rev Microbiol* 3(9):722–32.

331 2. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of
332 bacterial innovation. *Nature* 405(6784):299–304.

333 3. Ebersbach G, Gerdes K (2005) Plasmid segregation mechanisms. *Annu Rev Genet*
334 39:453–79.

335 4. Summers DK (1991) The kinetics of plasmid loss. *Trends Biotechnol* 9(8):273–278.

336 5. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile
337 elements and their impact on genome evolution. *Nucleic Acids Res* 29(18):3742–3756.

338 6. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic
339 and environmental microorganisms. *Nat Rev Microbiol* 2(5):414–24.

340 7. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC (2011) The
341 repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of
342 conjugation. *PLoS Genet* 7(8):e1002222.

343 8. Canchaya C, Proux C, Fournous G, Bruttin A, Brussow H (2003) Prophage Genomics.
344 *Microbiol Mol Biol Rev* 67(2):238–276.

345 9. Johnson CM, Grossman AD (2015) Integrative and Conjugative Elements (ICEs): What
346 They Do and How They Work. *Annu Rev Genet* 49(1):annurev-genet-112414-055018.

347 10. Bellanger X, Payot S, Leblond-bourget N, Guédon G (2014) Conjugative and
348 mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*
349 38:720–760.

350 11. Carraro N, Burrus V (2014) Biology of Three ICE Families: SXT/R391, ICEBs1, and
351 ICESt1/ICESt3. *Microbiol Spectr* 2(6):1–20.

352 12. Delavat F, Miyazaki R, Carraro N, Pradervand N, van der Meer JR (2017) The hidden
353 life of integrative and conjugative elements. *FEMS Microbiol Rev* 41(4):512–537.

354 13. Carraro N, Burrus V (2015) The dualistic nature of integrative and conjugative
355 elements. *Mob Genet Elements* 5(6):98–102.

356 14. Lee C a, Babic A, Grossman AD (2010) Autonomous plasmid-like replication of a
357 conjugative transposon. *Mol Microbiol* 75(2):268–79.

358 15. Carraro N, Poulin D, Burrus V (2015) Replication and Active Partition of Integrative
359 and Conjugative Elements (ICEs) of the SXT/R391 Family: The Line between ICEs and
360 Conjugative Plasmids Is Getting Thinner. *PLOS Genet* 11(6):e1005298.

361 16. Nunes-Düby SE, Kwon HJ, Tirumalai RS, Ellenberger T, Landy a (1998) Similarities and
362 differences among 105 members of the Int family of site-specific recombinases.
363 *Nucleic Acids Res* 26(2):391–406.

364 17. Guglielmini J, de la Cruz F, Rocha EPC (2013) Evolution of conjugation and type IV
365 secretion systems. *Mol Biol Evol* 30(2):315–31.

366 18. Guiney DG (1982) Host range of conjugation and replication functions of the
367 Escherichia coli sex plasmid Flac. Comparison with the broad host-range plasmid RK2.
368 *J Mol Biol* 162(3):699–703.

369 19. Zhong Z, Helinski D, Toukdarian A (2005) Plasmid host-range: Restrictions to F
370 replication in Pseudomonas. *Plasmid* 54(1):48–56.

371 20. Klümper U, et al. (2015) Broad host range plasmids can invade an unexpectedly
372 diverse fraction of a soil bacterial community. *ISME J* 9:934–945.

373 21. Rocha EPC, Danchin A (2002) Base composition bias might result from competition for
374 metabolic resources. *Trends Genet* 18(6):291–4.

375 22. Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of
376 horizontal gene transfer in Bacteria. *Nat Commun* 8(841):1–10.

377 23. Carraro N, et al. (2016) Plasmid-like replication of a minimal streptococcal integrative

378 and conjugative element. *Microbiol (United Kingdom)* 162(4):622–632.

379 24. Francia MV, Clewell DB (2002) Transfer origins in the conjugative Enterococcus
380 faecalis plasmids pAD1 and pAM373: Identification of the pAD1 nic site, a specific
381 relaxase and a possible TraG-like protein. *Mol Microbiol* 45(2):375–395.

382 25. César CE, Machón C, De La Cruz F, Llosa M (2006) A new domain of conjugative
383 relaxase TrwC responsible for efficient oriT-specific recombination on minimal target
384 sequences. *Mol Microbiol* 62(4):984–996.

385 26. Garcillán-Barcia MP, de la Cruz F (2008) Why is entry exclusion an essential feature of
386 conjugative plasmids? *Plasmid* 60(1):1–18.

387 27. Cury J, Touchon M, Rocha EPC (2017) Integrative and conjugative elements and their
388 hosts: composition, distribution and organization. *Nucleic Acids Res* (17):1–14.

389 28. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of
390 communities in large networks. *J Stat Mech*. doi:10.1088/1742-
391 5468/2008/10/P10008.

392 29. Suzuki H, Yano H, Brown CJ, Top EM (2010) Predicting plasmid promiscuity based on
393 genomic signature. *J Bacteriol* 192(22):6045–55.

394 30. Carnoy C, Roten C-A (2009) The dif/Xer recombination systems in proteobacteria.
395 *PLoS One* 4(9):e6531.

396 31. Wawrzyniak P, Płucienniczak G, Bartosik D (2017) The Different Faces of Rolling-Circle
397 Replication and Its Multifunctional Initiator Proteins. *Front Microbiol* 8(November):1–
398 13.

399 32. Sullivan JT, Ronson CW (1998) Evolution of rhizobia by acquisition of a 500-kb
400 symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci*
401 95(April):5145–5149.

402 33. Kers JA, et al. (2005) A large, mobile pathogenicity island confers plant pathogenicity
403 on Streptomyces species. *Mol Microbiol* 55(4):1025–1033.

404 34. Hall JPJ, Williams D, Paterson S, Harrison E, Brockhurst MA (2017) Positive selection
405 inhibits gene mobilization and transfer in soil bacterial communities. *Nat Ecol Evol*
406 1(9):1348–1353.

407 35. He S, et al. (2016) Mechanisms of Evolution in High-Consequence Drug Resistance.
408 *MBio* 7(6):e01987-16.

409 36. Porse A, Schønning K, Munck C, Sommer MOA (2016) Survival and Evolution of a
410 Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol Biol Evol*
411 33(11):2860–2873.

412 37. Wu LJ, Errington J, Rossignol M, Cornet F, Boccard F (2002) A large dispersed
413 chromosomal region required for chromosome segregation in sporulating cells of
414 Bacillus subtilis. *EMBO J* 21(15):4001–11.

415 38. Touchon M, Rocha EPC (2016) Coevolution of the organization and structure of
416 prokaryotic genomes. *Cold Spring Harb Perspect Biol* 8(1):1–18.

417 39. Oliveira PH, Touchon M, Rocha EPC (2016) Regulation of genetic flux between
418 bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 113(20):5658–
419 63.

420 40. Smillie C, Garcillán-Barcia MP, Francia MV, Rocha EPC, de la Cruz F (2010) Mobility of
421 plasmids. *Microbiol Mol Biol Rev* 74(3):434–52.

422 41. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM (2008) Adaptive plasmid
423 evolution results in host-range expansion of a broad-host-range plasmid. *Genetics*
424 178(4):2179–2190.

425 42. Loftie-Eaton W, et al. (2017) Compensatory mutations improve general
426 permissiveness to antibiotic resistance plasmids. *Nat Ecol Evol* 1:1354–1363.

427 43. Clewell DB, Flannagan SE, Jaworski DD (1995) Unconstrained bacterial promiscuity:
428 the Tn916–Tn1545 family of conjugative transposons. *Trends Microbiol* 3(6):229–236.

429 44. San Millan A, Escudero JA, Gifford DR, Mazel D, Maclean RC (2016) Multicopy

430 plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol*
431 1(10):1–8.

432 45. San Millan A, MacLean RC (2017) Fitness Costs of Plasmids: a Limit to Plasmid
433 Transmission. *Microbiol Spectr* 5(5):1–12.

434 46. Guédon G, Libante V, Coluzzi C, Payot S, Leblond-Bourget N (2017) The obscure world
435 of integrative and mobilizable elements, highly widespread elements that pirate
436 bacterial conjugative systems. *Genes (Basel)* 8(11). doi:10.3390/genes8110337.

437 47. Ravin N V. (2011) N15: The linear phage-plasmid. *Plasmid* 65(2):102–109.

438 48. Łobocka MB, et al. (2004) Genome of Bacteriophage P1. *J Bacteriol* 186(21):7032–
439 7068.

440 49. Touchon M, Moura de Sousa JA, Rocha EP (2017) Embracing the enemy: The
441 diversification of microbial gene repertoires by phage-mediated horizontal gene
442 transfer. *Curr Opin Microbiol* 38:66–73.

443 50. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC (2014) MacSyFinder: A
444 Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas
445 Systems. *PLoS One* 9(10):e110726.

446 51. Guglielmini J, et al. (2014) Key components of the eight classes of type IV secretion
447 systems involved in bacterial conjugation or protein secretion. *Nucleic Acids*
448 *Res*:gku194-.

449 52. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein
450 database search programs. *Nucleic Acids Res* 25(17):3389–3402.

451 53. Oliveira PH, Touchon M, Rocha EPC (2014) The interplay of restriction-modification
452 systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res*
453 42(21):1–14.

454 54. Touchon M, et al. (2014) The Genomic Diversification of the Whole Acinetobacter
455 Genus: Origins, Mechanisms, and Consequences. *Genome Biol Evol* 6(10):2866–2882.

456 55. Gibson MK, Forsberg KJ, Dantas G (2014) Improved annotation of antibiotic resistance
457 determinants reveals microbial resistomes cluster by ecology. *ISME J* 9(1):207–216.

458 56. Huerta-Cepas J, et al. (2016) eggNOG 4.5: a hierarchical orthology framework with
459 improved functional annotations for eukaryotic, prokaryotic and viral sequences.
460 *Nucleic Acids Res* 44(D1):D286-93.

461 57. Eddy SR (2011) Accelerated Profile HMM Searches. *PLoS Comput Biol* 7(10):e1002195.

462 58. Cury J, Jové T, Touchon M, Néron B, Rocha EP (2016) Identification and analysis of
463 integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res* 44(10):4539–50.

464 59. Achaz G, Boyer F, Rocha EPC, Viari A, Coissac E (2007) Repseek, a tool to retrieve
465 approximate repeats from large DNA sequences. *Bioinformatics* 23(1):119–121.

466 60. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:
467 Improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.

468 61. Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a
469 new software for selection of phylogenetic informative regions from multiple
470 sequence alignments. *BMC Evol Biol* 10:210.

471 62. Filipski A, Murillo O, Freydenzon A, Tamura K, Kumar S (2014) Prospects for building
472 large timetrees using molecular data with incomplete gene coverage among species.
473 *Mol Biol Evol* 31(9):2542–2550.

474 63. Price MN, Dehal PS, Arkin AP (2009) Fasttree: Computing large minimum evolution
475 trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641–1650.

476 64. Suzuki H, Sota M, Brown CJ, Top EM (2008) Using Mahalanobis distance to compare
477 genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res*
478 36(22):e147.

479 65. Carattoli A, et al. (2014) In Silico detection and typing of plasmids using plasmidfinder
480 and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58(7):3895–
481 3903.

Figure Legends

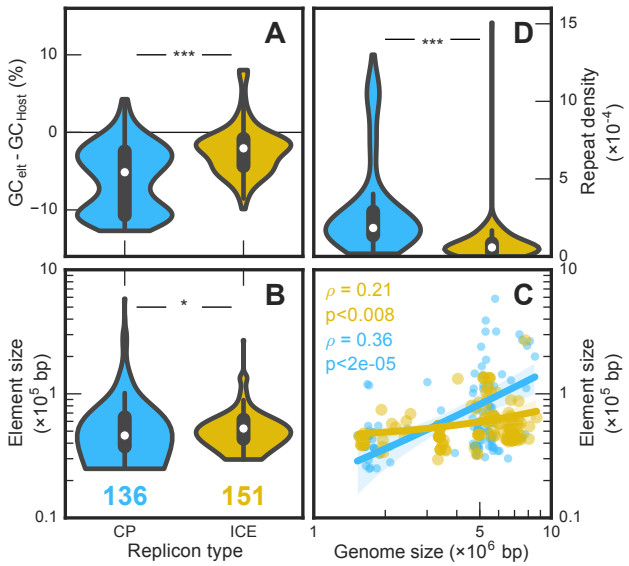
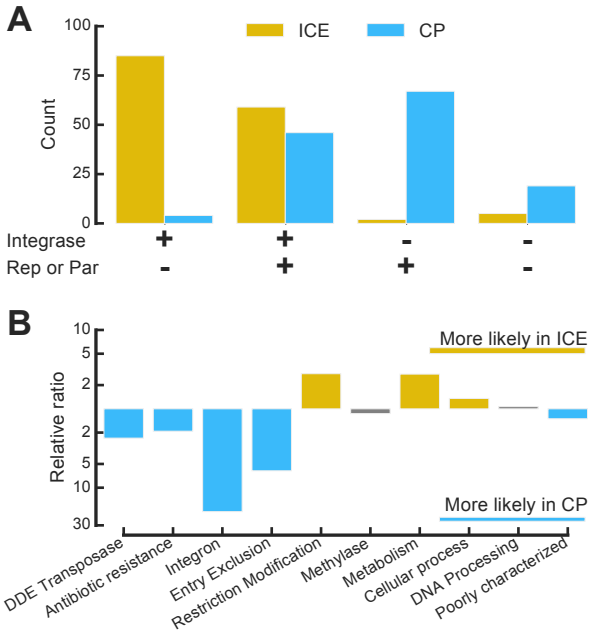


Fig. 1: Comparison between 136 CPs (blue) and 151 ICEs (yellow) in composition and sizes. **A**. CPs are AT richer than ICEs relative to their hosts (Wilcoxon rank sum test, p-value<10^{-3}). **B**. ICEs and CPs have different distributions of size (same test, p-value<0.05). Median sizes: ICEs (52.5 kb) > CPs (46.1 kb). Averages: ICEs (59 kb) < CPs (74.6 kb) **C.** Size of the element as a function of the genome size of its host (decreased by the size of the mobile element itself). Shaded regions indicate the 95% confidence interval. The Y-axis is identical to the one in panel B. **D**. Density of repeats is higher in CPs than in ICEs (0.30 vs 0.078 repeats per kb, same test, p-value < 10^{-10}).

495

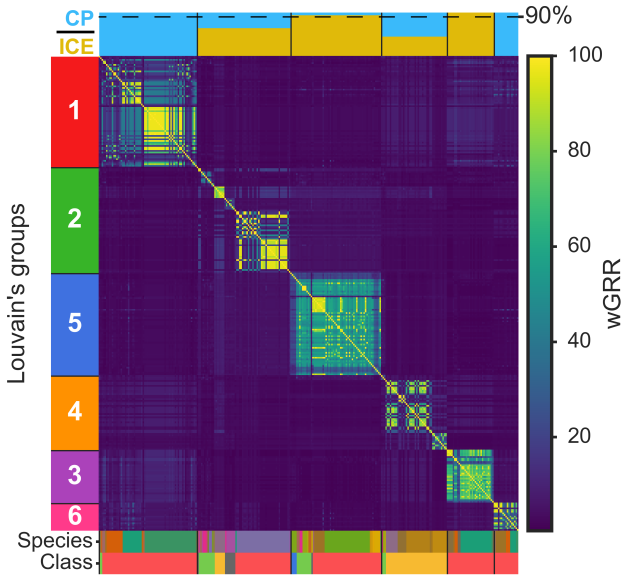


496

497 **Fig. 2**: Comparisons of the functions carried by ICEs and CPs. **A.** Elements encoding (+) or
498 lacking (-) replication or partition systems ("Rep or Par") or an integrase. **B.** Accessory
499 functions over-represented (yellow) or under-represented (blue) in ICEs relative to CPs.
500 Colored bars: significantly different from zero, p-value < 0.05 Fisher exact test with
501 Bonferroni-Holm correction for multiple tests. Grey bars otherwise.

502

503



504

505 **Fig. 3**: Heatmap of the wGRR scores, ordered after the size of Louvain's group (depicted on
506 the left bar). The top bar represents the proportion of ICEs (yellow) and CPs (blue) for each
507 group. The bottom bar assigns a color corresponding to the host's species or class (γ-
508 proteobacteria in red, β-proteobacteria in green, α-proteobacteria in blue, ε-
509 proteobacteria in orange, Fusobacteria and Acidobacteria in grey).
510

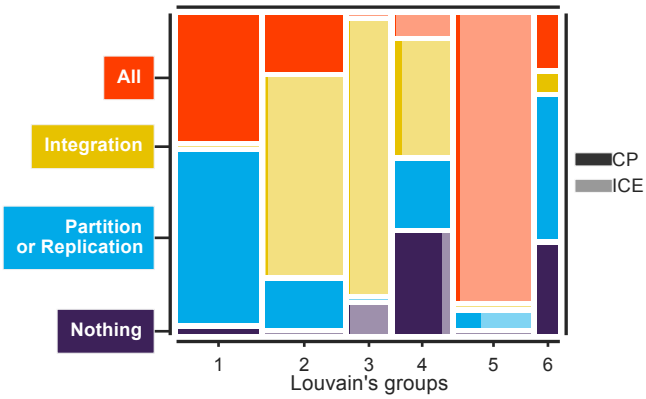


Fig. 4: Mosaic plot representing the frequency of key functions of conjugative elements in terms of the Louvain's groups (Fig. 3). The width of the bar is proportional to the number of elements in a given Louvain's group (see the number of elements of each group on the top of the bars) and the areas reflect the proportion of the elements with the function. The colors represent the type of function, and their tint represent the part of ICEs (lighter) and CPs (darker).

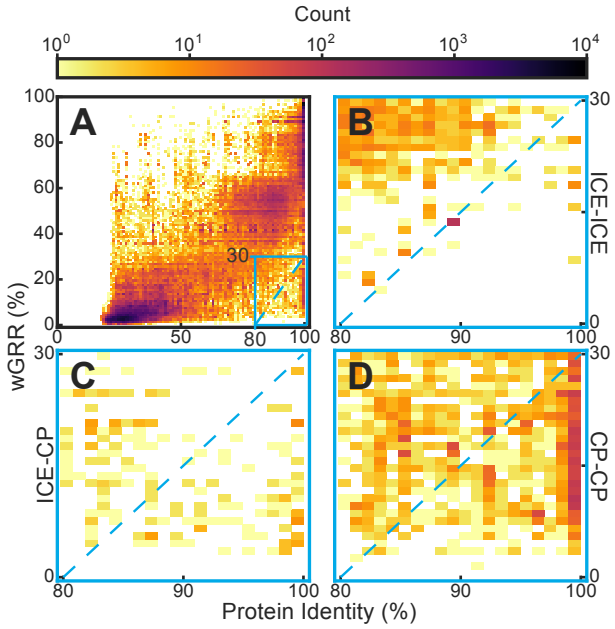


Fig. 5: 2D histogram of the wGRR score of pairs of conjugative elements as a function of the protein identity of their homologues. **A**. Distribution for the entire dataset. wGRR values are correlated with protein identity of the elements' homologs ($\rho_{ICE\text{-}ICE}$ = 0.90, $\rho_{CP\text{-}CP}$ = 0.83). The black rectangle zooms on a region where the pairs of elements are very different (GRR<30%), yet they encode at least one very similar protein (identity > 80%). The dashed line separates the elements where protein identity is higher than wGRR x 2/3. **B**. Zoom for ICE-ICE comparisons. **C**. ICE-CP comparisons. **D**. CP-CP comparisons.

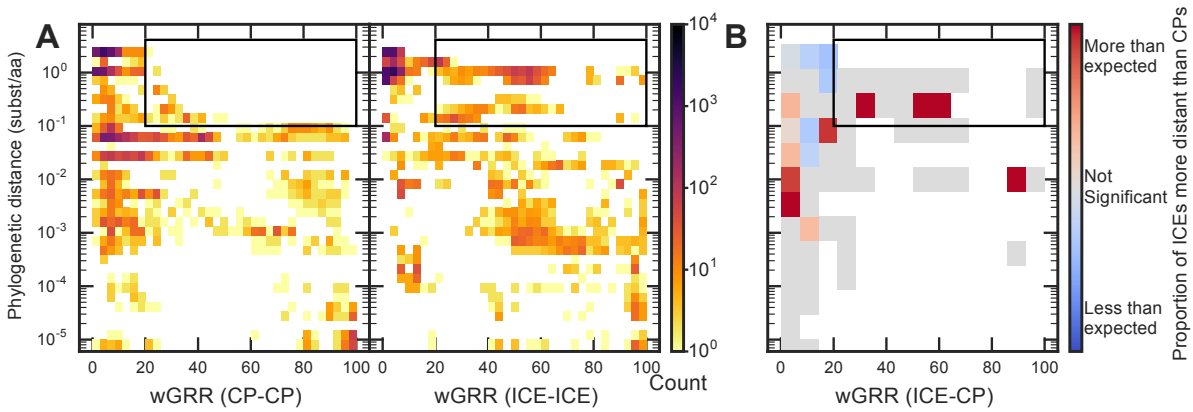


Fig. 6: wGRR as function of the host phylogenetic distance **A**. 2D histogram of the distribution of the wGRR score as a function of the phylogenetic distance for pairs of CPs (left) and pairs of ICEs (right). The bottom row corresponds to all pairs with distance lower than 10^{-5} (including those in the same host). Elements in the black rectangle are depicted in a phylogenetic tree in Fig. S5. **B**. Proportion of ICEs more distinct in terms of tri-nucleotides from their host than CPs. Bins are larger than in A. to increase the power of the statistical analysis. Color code: ICEs (Red) or CPs (Blue) are more distinct from the host in terms of tri-nucleotide composition than the other element. Grey: not significant (binomial test, p-value $> 10^{-2}$). White: no elements in the bin.

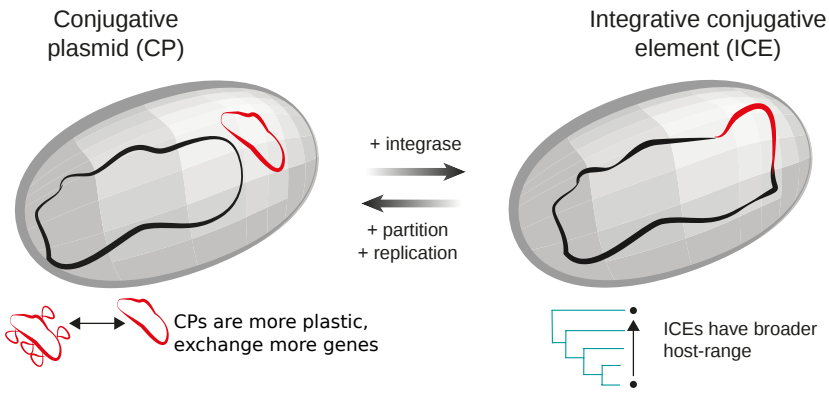


Fig. 7: To integrate or not to integrate? ICEs and CPs have similar conjugative systems but persist in different ways (resp. integrating the chromosome or replicating independently). ICEs over-represent integrases, whereas plasmids over-represent replication and partition functions (although both can have all of these functions). ICEs seem at an advantage when conjugating to distant hosts, presumably because they integrate the chromosome. On the other hand, plasmids have broader distribution of size and exchange more genetic information.

549 # Supplementary Material for the Manuscript :

550 # Host range expansion and genetic plasticity drive the trade-off
551 # between integrative and extrachromosomal mobile genetic elements

552 Jean Cury, Pedro H. Oliveira, Fernando de la Cruz, Eduardo P.C. Rocha

553

554 ## Material and Methods: Extended version

555 ### Data

556 Conjugative systems of type T (MPF_T) were searched in the set of complete bacterial
557 genomes from NCBI RefSeq (http://ftp.ncbi.nih.gov/genomes/refseq/bacteria/, last
558 accessed in November 2016). We analyzed 5562 complete genomes from 2268 species,
559 including 4345 plasmids and 6001 chromosomes. The classification of the replicon in
560 plasmid or chromosome was taken from the information available in the GenBank file. Our
561 method to delimit ICEs is based on comparative genomics of closely related strains. Hence,
562 we restricted our search for conjugative systems to the species for which we had at least
563 five genomes completely sequenced (164 species, 2990 genomes).

564 ### Detection of conjugative systems and delimitation of ICEs

565 Conjugative systems were detected using the CONJscan module of MacSyFinder (50), with
566 protein profiles and definitions of the MPF type T, published previously (51). ICEs were
567 delimited with the same methodology, as developed in a previous work (27). Briefly, we
568 identified the core genomes of the species. The region between two consecutive genes of
569 the core genome defined an interval in each chromosome. We then defined spots as the
570 sets of intervals in the chromosome flanked by genes of the same two families of the core
571 genome (22). We then identified the intervals and the spots with conjugative systems. The
572 information on the sets of gene families of the spots with ICEs (i.e., the spot pan-genome)
573 was used to delimit the element boundaries (script available at
574 https://gitlab.pasteur.fr/gem/spot_ICE). This methodology was shown to be accurate at the
575 gene level (precise nucleotide boundaries are not identifiable by this method, see (27)).

576 ## Functional analyses

577 Partition systems, replication systems, entry-exclusion systems and restriction modification
578 systems were annotated with HMM profiles, as described in our previous work (27, 53).
579 Integrases were annotated with the PFAM profile PF00589 for the Tyrosine recombinases
580 and the combination of PFAM profiles PF00239 and PF07508 for Serine recombinases. DDE
581 Transposases were detected with Macsyfinder (50) with models used previously (54).
582 Antibiotic resistance genes were detected with ResFams profiles (core version v1.1) (55)
583 using the --cut_ga option. We determined the functional categories of genes using their
584 annotation as provided by their best hits to the protein profiles of the EggNOG database for
585 bacteria (version 4.5, bactNOG) (56). Genes not annotated by the EggNOG profiles were
586 classed as "Unknown" and included in the "Poorly characterized" group. The HMM profiles
587 were used to search the genomes with HMMER 3.1b2 (57), and we retrieved the hits with
588 an e-value lower than 10^{-3} and with alignments covering at least 50% of the profile.
589 Integrons were detected using IntegronFinder version 1.5.1 with the --local_max option for
590 higher accuracy (58). Repeats (direct and inverted) were detected with Repseek (version
591 6.6) (59) using the option -p 0.001 which set the p-value for determining the minimum seed
592 length.

593 ## Statistics

594 We tested the over-representation of a given function or group of functions using Fisher's
595 exact tests on contingency tables. For partition, replication and integration, the contingency
596 table was made by splitting replicons in those encoding or not encoding the function and
597 between ICEs and CPs. The use of presence/absence data instead of the absolute counts
598 was made because the presence of at least one occurrence of a system is sufficient to have
599 the function and because the counts were always low. For the other functions, the
600 contingency table was made by splitting the proteins of the element in those annotated for
601 a given function and the remaining ones. This allowed to take into account the differences in
602 the number of genes between elements. The Fisher-exact tests were considered as
603 significant after sequential Holm-Bonferroni correction, with a family-wise error rate of 5%
604 (the probability of making at least one false rejection in the multiple tests, the type I error).
605 From the contingency table, we computed the relative ratio (or relative risk) of having a

606 given function more often in ICEs than in CPs. The relative ratio is computed as follow:
607 $RR = \frac{ICEwF/N_{ICE}}{CPwF/N_{CP}}$ where *ICEwF* is the number of ICE (or proteins in ICEs) with the given
608 function, and N_{ICE}, the total number of ICE (or proteins in ICEs), and likewise for CP. The
609 term $ICEwF/N_{ICE}$ is an estimation of the probability of an ICE (or a protein in an ICE) to
610 carry a given a function.

611 ## Phylogenetic distances

612 Phylogenetic distances were extract from the Proteobacterial tree of the Core-genome. To
613 build the tree, we identified the genes present in at least 90% of the 2897 genomes of
614 Proteobacteria larger than 1 Mb that were available in GenBank RefSeq in November 2016.
615 A list of orthologs was identified as reciprocal best hits using end-gap free global alignment.
616 Hits with less than 37% similarity in amino acid sequence and more than 20% difference in
617 protein length were discarded. We then identified the protein families with relations of
618 orthology in at least 90% of the genomes. They represent 341 protein families. We made
619 multiple alignments of each protein family with MAFFT v.7.205 (with default options) (60)
620 and removed poorly aligned regions with BMGE (with default options) (61). Genes missing in
621 a genome were replaced by stretches of "–" in each multiple alignment, which has been
622 shown to have little impact in phylogeny reconstruction (62). The tree of the concatenate
623 alignment was computed with FastTree version 2.1 under the LG model (63). We chose the
624 LG model because it was the one that minimized the AIC.

625 ## Distance to the host

626 We used the differences in tri-nucleotide composition to compute the genetic distance
627 between the mobile element and its host chromosome, as previously proposed (64). The
628 analysis of ICEs was done by comparing the element with the chromosome after the
629 removal of its sequence from the latter. Briefly, we computed the trinucleotide relative
630 abundance ($x_{ijk} \forall i, j, k \in \{A, T, C, G\}$) for the chromosomes (in windows of 5 kb) and for the
631 conjugative elements (entire replicon), which is given by: $x_{ijk} = f_{ijk}/f_i f_j f_k$, with f the
632 frequency of a given k-mer in the sequence (29). We first computed the Mahalanobis
633 distance between each window and the host chromosome as follow:

$$D = \sqrt{(w-h)^T H^{-1} (w-h)}$$

634 with *w*, the vector of tri-nucleotide abundances (x_{ijk}) in a given window, and *h*, the mean of
635 the vector of x_{ijk} (*i.e.*, the average tri-nucleotide abundance in the chromosome). *H* is the
636 covariance matrix of the tri-nucleotide relative abundances. The inverse of the covariance
637 matrix (H^{-1}) downweights frequent trinucleotides, like the tri-nucleotides corresponding to
638 start codons, which are common to conjugative elements and chromosome and could bias
639 the distance. We computed the Mahalanobis distance between conjugative elements and
640 their hosts' chromosomes (same formula as above, but *w* is now for a conjugative element
641 instead of a chromosome window). We then computed the probability (p-value) that the
642 measured distance between a conjugative element and the host's chromosome is the same
643 as any fragment of the host's chromosome.

644 We compared ICEs and CPs in relation to their compositional distance to the host. For this,
645 we made the null hypothesis that the proportion of ICEs having a p-value lower than CPs
646 follows a binomial distribution whose expected proportion is that of the entire dataset (the
647 proportion of ICEs having a p-value lower than CP), precisely:
648 $H_0 = N(pvalue_{ICE} < pvalue_{CP})/N_{Comparisons}$, where $N_{Comparisons}$ is the total number of
649 ICE-CP pairs, *i.e.* $151 \times 136 = 20536$.

650 Network analysis of gene repertoire relatedness

651 We built a network describing the relations of homology between the elements. The nodes
652 in the network are conjugative elements and they are linked if they share a given relation of
653 homology. More precisely, the relationship between two elements was quantified with the
654 weighted Gene Repertoire Relatedness score (wGRR). This score represents the number of
655 homologous proteins between two elements, weighted by their sequence identity, as
656 described in (27). The formula is:

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

657 Where (A_i, B_i) is the ith pair of homologous protein between element A and element B,
658 $id(A_i,B_i)$ is the sequence identity of their alignment, min(A,B) is the number of proteins of
659 the element with fewest proteins (A or B). The sequence identity was computed with blastp

660 v.2.2.15 (default parameters)(52) and kept all bi-directional best hits with an e-value lower
661 than 10^{-5}.

662 The network was built based on the wGRR matrix. Its representation was made using the
663 Fruchterman-Reingold force-directed algorithm as implemented in the NetworkX v1.11
664 python library. The groups were made using the Louvain algorithm (28). We controlled for
665 the consistency of the heuristic used to assess that the group found are not form a local
666 optimal. We performed 100 clustering, which led to the same classification in 95% of the
667 time.

668 ## Incompatibility typing

669 We determined the incompatibility group of replicons using the method of PlasmidFinder
670 (65). We used BLASTN (52) to search the replicons for sequences matching the set of 116
671 probes used by PlasmidFinder. We kept the hits with a coverage above 60% and sequence
672 identity above 80%, as recommended by the authors. Around 3% of the elements had
673 multiple incompatibility types attributed.
674

Supplementary Figures

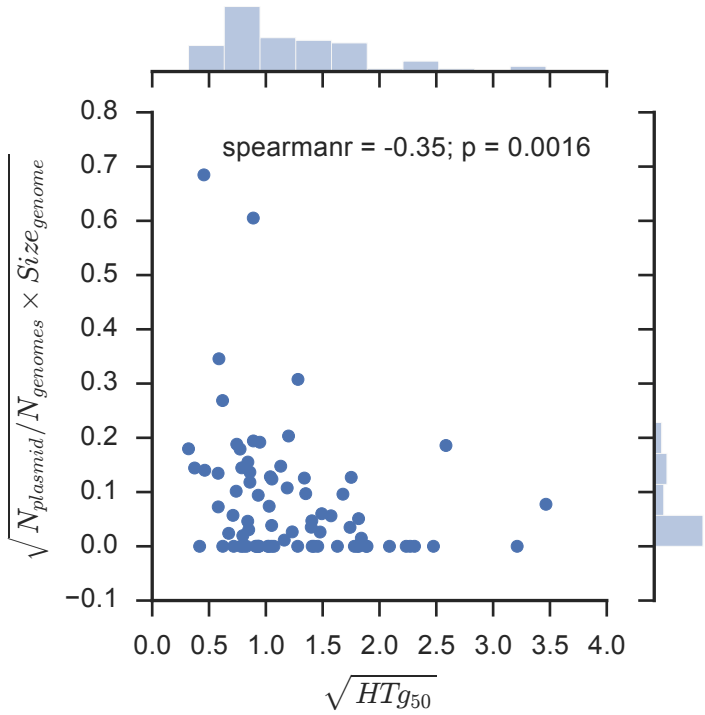


Fig. S1: Distribution of the frequency of plasmids in genomes of a species as function of the HTg50 in the corresponding species (HTg50 values taken from (22)). There is a negative correlation as measured with the Spearman coefficient rho=-0.35; p-value= 0.0016.

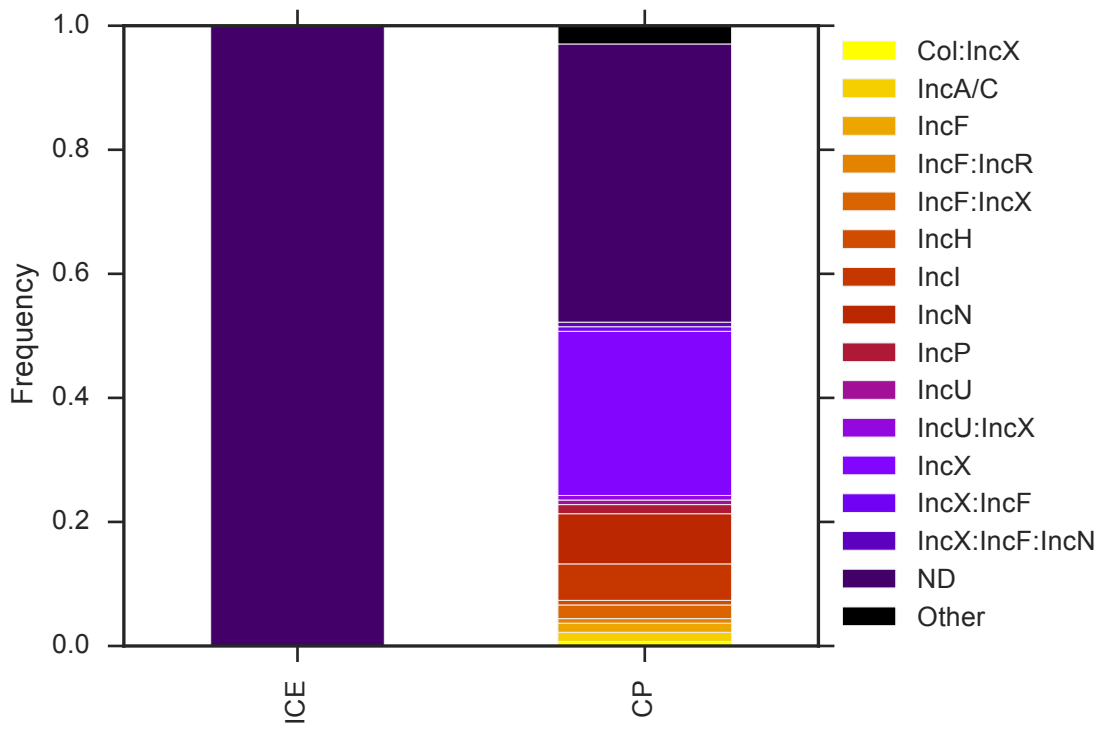


Fig. S2: Distribution of inc type among ICE and CP. No ICEs could be typed.

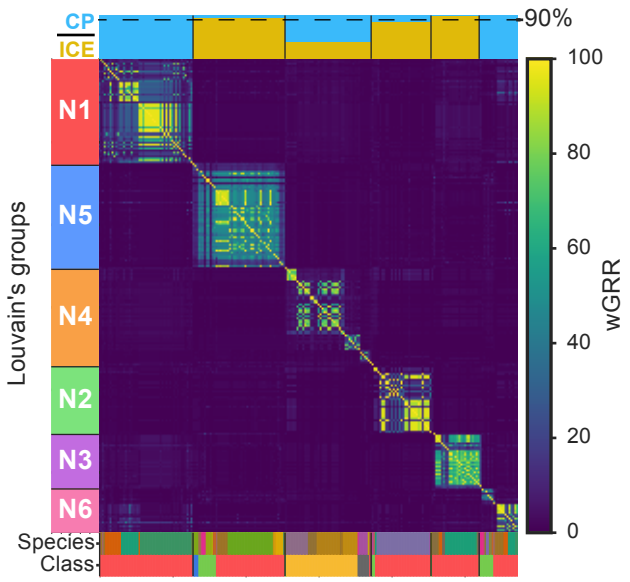
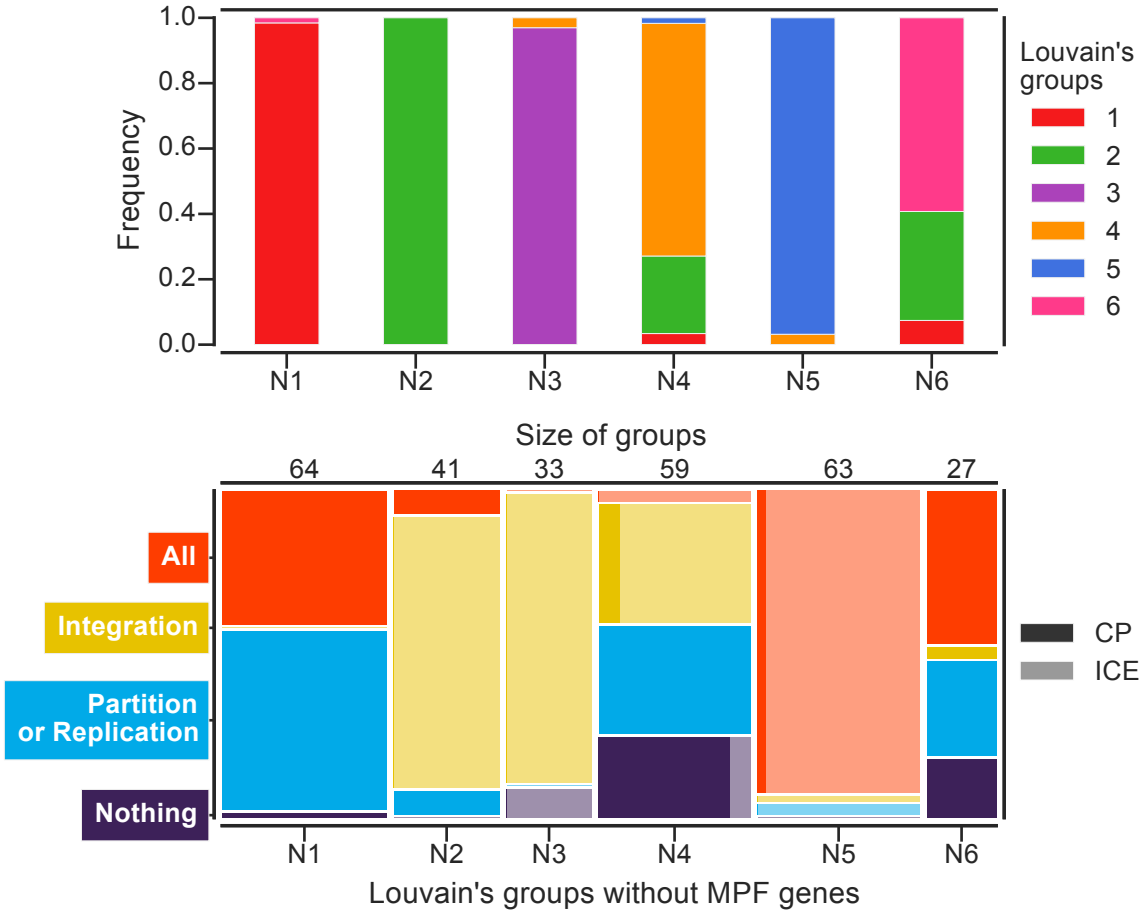


Fig. S3 : Heatmap of the wGRR scores without MPF genes, ordered after the size of Louvain's group (depicted on the left bar). The top bar represents the proportion of ICEs (yellow) and CPs (blue) for each group. The bottom bar assigns a color corresponding to the host's species or class (γ-proteobacteria in red, β-proteobacteria in green, α-proteobacteria in blue, ε-proteobacteria in orange, Fusobacteria and Acidobacteria in grey).

690 clustering.



691

692 **Fig. S4**: **Top.** Distribution of the Louvain's group in the new groups formed after removing
693 the MPF genes. Each new group correspond to the first groups, at the exception of group 2.
694 **Bottom.** Mosaic plot representing the frequency of key functions of conjugative elements in
695 terms of the Louvain's groups (Fig. 3). The width of the bar is proportional to the number of
696 elements in a given Louvain's group (see the number of elements of each group on the top
697 of the bars) and the areas reflect the proportion of the elements with the function. The
698 colors represent the type of function, and their tint represent the part of ICEs (lighter) and
699 CPs (darker).

700

701

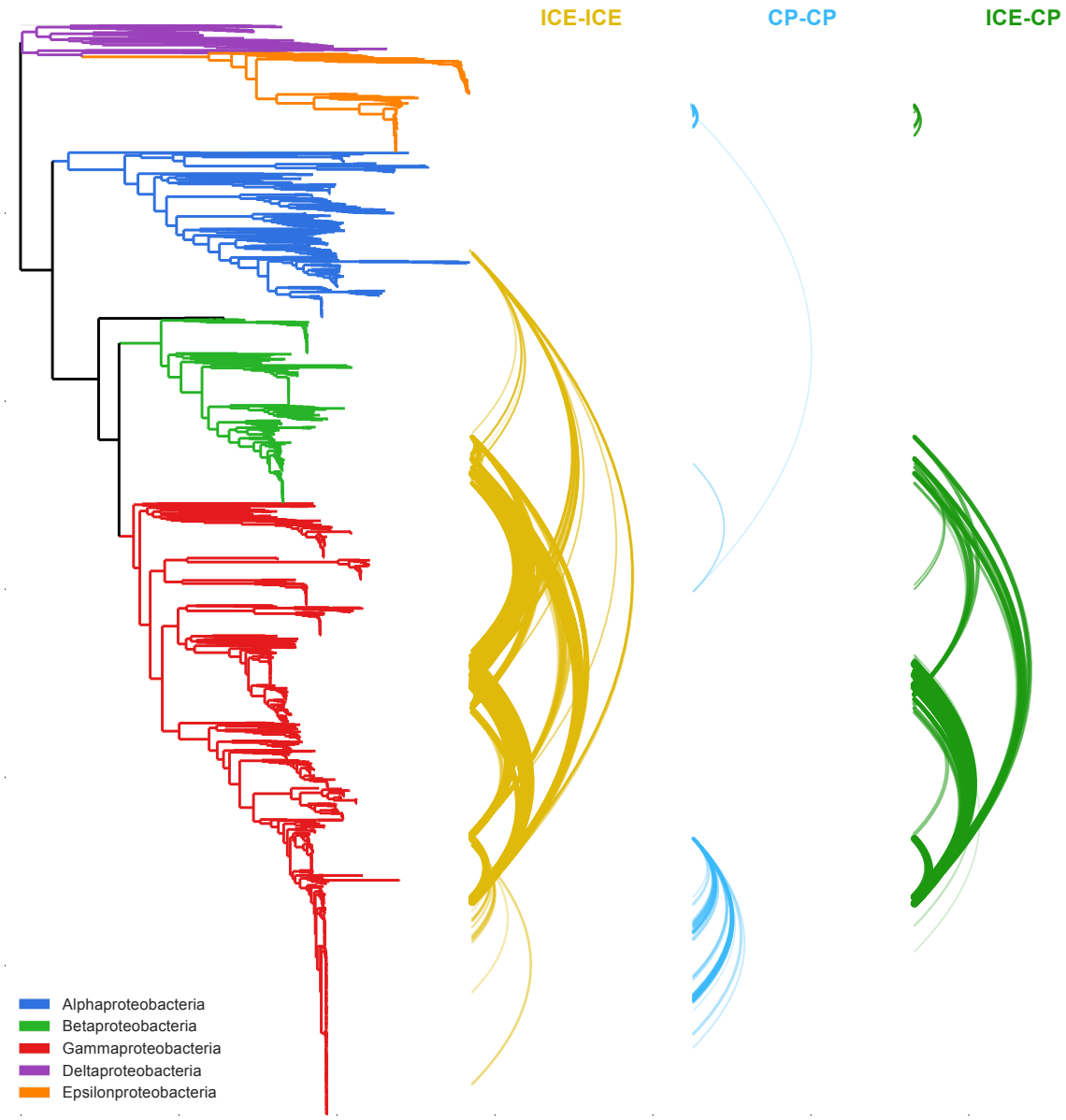


Fig. S5: Taxonomic distribution of each element in the pairs of elements found in the black rectangle of Fig. 6. The phylogenetic tree of Proteobacteria is depicted on the left, and the major clades are colored according to taxonomic classes. The single genome in black is an Acidithiobacillia. On the right, a line links two hosts if they contain two elements with a wGRR higher than 20% and a phylogenetic distance higher than 0.1. The thickness of the line is proportional to the wGRR. ICE-ICE relationships are in yellow, CP-CP in blue and ICE-CP in green.

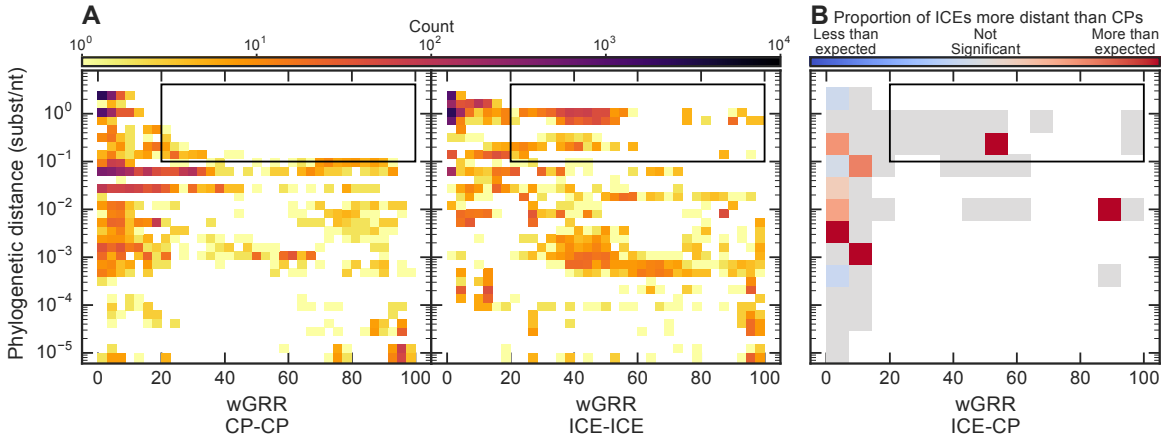
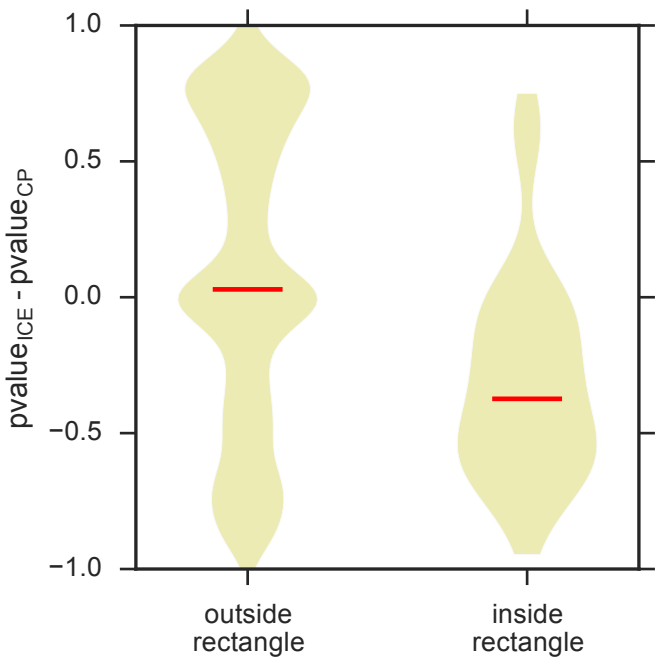


Fig. S6: wGRR computed without MPF genes as function of the host phylogenetic distance **A**. 2D histogram of the distribution of the wGRR score computed without the MPF genes as a function of the phylogenetic distance for pairs of CPs (left) and pairs of ICEs (right). The bottom row corresponds to all pairs with distance lower than 10^{-5} (including those in the same host). **B**. Proportion of ICEs more distinct in terms of tri-nucleotides from their host than CPs. Bins are larger than in A. to increase the power of the statistical analysis. Color code: ICEs (Red) or CPs (Blue) are more distinct from the host in terms of tri-nucleotide composition than the other element. Grey: not significant (binomial test, p-value > 10^{-2}). White: no elements in the bin.

722



723

724 **Fig. S7**: Distribution of the difference in p-values between ICEs and CPs, relative to the
725 regions highlighted in Fig. 6B. The lower the p-value the more distant is the element from its
726 host's chromosome. The rectangle corresponds to the black rectangle in Fig. 6B. There is a
727 significant difference between both distributions (Wilcoxon rank-sum test, p-value<10^{-20}).
728 Outside of the rectangle ICEs and CPs show similar average p-values when compared with
729 their hosts, whereas inside the difference is negative, indicating lower p-values in ICEs.

730

731 ## Supplementary Tables

732 Supplementary tables can be found at the following address:
733 https://gitlab.pasteur.fr/gem/ICE_CP

734 Table 1: List of elements (protein, recombination site) found in conjugative element of type
735 T, with corresponding functional annotation when available.

736 Table 2: Summary information per element (ICE or CP).

737 Table 3: list of pairs of elements with corresponding information (e.g. wGRR, phylogenetic
738 distance of the host)

739 Table 4: Same as Table 3 but wGRR is computed without MPF genes.

740