Host range and genetic plasticity explain the co-existence of 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. Plasmids encode more replicases, partition systems, and antibiotic resistance genes, whereas ICEs encode more integrases and metabolism-associated genes. 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, and this drives the conversion of plasmids into ICEs after transfer to distantly related hosts. Hence, differential plasticity and transmissibility explain the occurrence of both integrative and extra-chromosomal elements in microbial populations.

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

32 ## Introduction

33 The genomes of Prokaryotes have mobile genetic elements (MGEs) integrated in the
34 chromosome or replicating as extrachromosomal elements. These MGEs usually encode
35 non-essential but ecologically important traits (1, 2). Extra-chromosomal elements, such as
36 conjugative plasmids (CPs) and lytic phages, replicate autonomously in the cell using
37 specialized replicases to recruit the bacterial DNA replication machinery (or to use their
38 own). Plasmids and extra-chromosomal prophages can also increase their stability in cellular
39 lineages using partition systems, for proper segregation during bacterial replication (3),
40 resolution systems, to prevent accumulation of multimers (4), and restriction-modification
41 or toxin-antitoxin systems, for post-segregation killing of their hosts (5). Alternatively, many
42 MGEs integrate into the chromosome. This is the case of the vast majority of known
43 prophages, of most conjugative elements (ICEs), and of many elements with poorly
44 characterized mechanisms of genetic mobility (*e.g.*, many pathogenicity islands)(6–8). The
45 integrated elements are replicated along with the host chromosome and require an
46 additional step of excision before being transferred between cells. The existence of both
47 integrative and extra-chromosomal elements was a fruitful source of controversy in the
48 dawn of molecular biology, eventually leading to the discovery of the molecular
49 mechanisms allowing both states (9, 10). Yet, a complementary question does not seem to
50 have been addressed in the literature: Why are there both types of elements? What are the
51 relative benefits and disadvantages of the integrated and extrachromosomal MGE?

52 To address these questions, we analyzed the differences and similarities between ICEs and
53 CPs. We focused on these elements because both forms are frequently found in bacteria,
54 they can be easily detected in genomes, and the mechanism of conjugation is well known.
55 Conjugative elements have a crucial role in spreading antibiotic resistance and virulence
56 genes among bacterial pathogens (11–14). Recently, several works suggested that the line
57 separating integrative ICEs and CPs could be thinner than anticipated (15), because some
58 ICEs encode plasmid-associated functions like replication (16) or partition (17), some
59 plasmids encode integrases (18), and ICEs and CPs are intermingled in the phylogenetic tree
60 of conjugative systems (19). Finally, both forms – ICEs and CPs – are found throughout the
61 bacterial kingdom, but their relative frequency depends on the taxa and on the mechanisms

62 of conjugation (7). These differences suggest that conjugative elements endure diverse
63 selective pressures for being integrative or extrachromosomal depending on unknown
64 environmental, genetic, or physiological variables.

65 We thought that key differences in the biology of integrative and extrachromosomal
66 elements might provide them with different types of advantages. ICEs require an additional
67 step of integration/excision during transfer, which may take time and requires genetic
68 regulation. Their integration in the chromosome may affect the latter's organization and
69 structure, and these collateral effects might depend on the size of the element. On the
70 other hand, ICEs replicate as part of chromosomes and could thus be lost from the cell at
71 lower rates than plasmids. Furthermore, plasmids must recruit the host replication
72 machinery, which may render elements incompatible and is known to constrain their host
73 range: many plasmids are able to conjugate into distantly related hosts, but are unable to
74 replicate there (20–22). We thus hypothesize that ICEs might be favored when transfers
75 occur between distant hosts, whereas plasmids might provide more genetic plasticity
76 because their size is not constrained by chromosomal organization.

77 Here, we study conjugative elements of the type MPF_T. This is the most frequent and best-
78 studied type of conjugative systems (19), and the only one for which we can identify
79 hundreds of elements of each of the forms (ICEs and CPs). We restricted our analysis to
80 genera containing both CPs and ICEs, to avoid, as much as possible, taxonomical biases. We
81 first describe the content of both types of elements and highlight their differences and
82 similarities. Next, we quantify their genetic similarity and the extent of their gene
83 exchanges. Finally, we show that chromosomal integration facilitates the colonization of
84 novel taxa by a conjugative element.

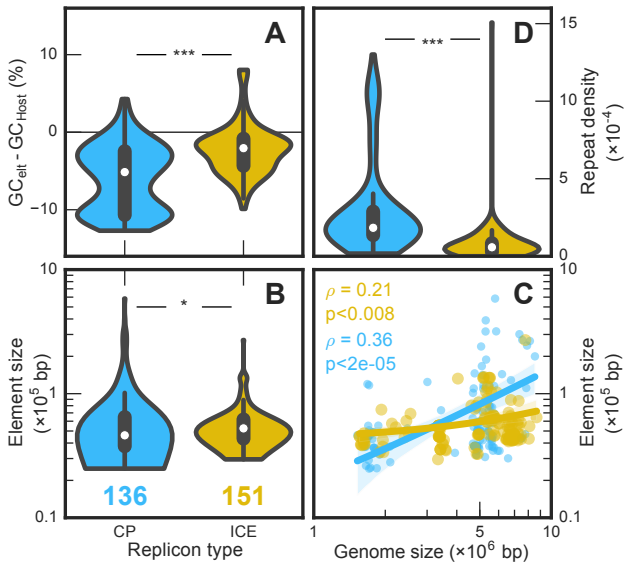
85

86

87 ## Results

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

89 We analyzed a set of 151 ICEs and 136 CPs of the same genera and of type MPF_T, most of
90 which were from Proteobacteria (96.9%). Both ICEs and CPs were found to be AT-richer than
91 their host chromosomes, which is a common feature in MGEs and horizontally transferred
92 genes (23). However, the difference was three times smaller for ICEs (Fig. 1A), presumably
93 because they replicate with the chromosome or remain a longer time in the same host. The
94 average size of CPs is slightly larger (75kb vs 59kb), and the median slightly smaller (46kb vs
95 52kb) than that of ICEs. In contrast, CPs have more diverse sizes than ICEs (Fig. 1B), showing
96 a coefficient of variation twice as large (1.05 vs 0.49). The size of the conjugative elements
97 depends on the size of the bacterial genome (after discounting the size of their conjugative
98 elements), this effect being much stronger for CPs (Fig. 1C). CPs also have four times higher
99 density of large DNA repeats than ICEs (Fig. 1D). These results suggest that CPs diversify
100 faster than ICEs.



101

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

108 in panel B. **D**. Density of repeats is higher in CPs than in ICEs (0.30 vs 0.078 repeats per kb,
109 same test, p-value < 10^{-10}).
110

111 HGT concentrates in a few hotspots in bacterial chromosomes, presumably to minimize
112 disruption in their organization (24). We used HTg50, a measure of the concentration of
113 HGT in chromosomes that corresponds to the minimal number of spots required to account
114 for 50% of horizontally transferred genes (24), to test if chromosomes with fewer
115 integration hotspots had more plasmids. Indeed, there is a negative association between
116 the number of plasmids, weighted by their size, and the chromosomes' HTg50 (Spearman
117 ρ=-0.35, p-value=0.0016, Fig. S1).

118 We then quantified the differences between ICEs and CPs in terms of functions associated
119 with their biology, with a focus on stabilization functions. Relaxases are part of the rolling
120 circle replication initiator proteins and some have been shown to act as replicases(16, 25) or
121 site-specific recombinases (26, 27). Since all conjugative elements encode a relaxase, by
122 definition, they may also have these functions. In the following, we focused on typical
123 plasmid replication initiator proteins (more than 95% of them are involved in theta-
124 replication, and none is matched by the protein profiles of relaxases), and serine or tyrosine
125 recombinases as integrases. Expectedly, ICEs showed higher frequency of integrases, while
126 CPs had more frequently identifiable partition and replication systems. Some ICEs encode
127 partition systems (11%) and many encode a replicase (40%), while 37% of CPs encode at
128 least one tyrosine or serine recombinase (Fig. 2A). These results further illustrate a
129 continuum between the two types of elements: about half of the elements (40% and 48%,
130 ICEs and CPs, respectively) have functions usually associated with the other type and may
131 (rarely) lack functions typically associated with its own type (Fig. 2B). We identified plasmid
132 incompatibility systems of diverse types, whereas no ICE could not be typed in the current
133 scheme (Fig. S2).

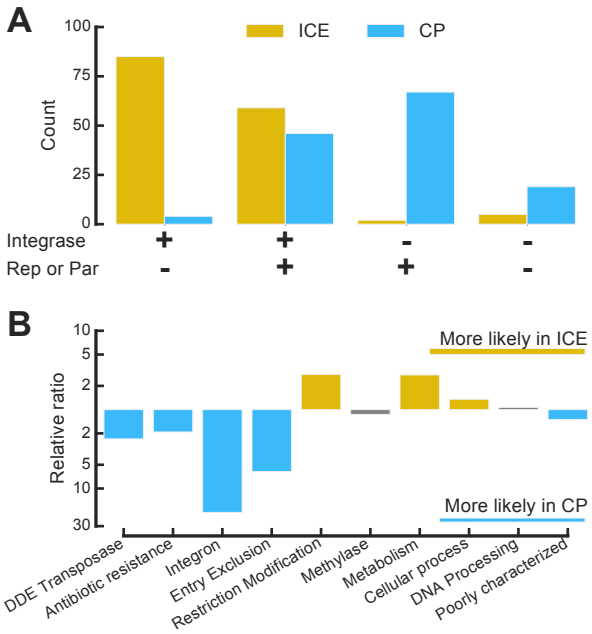


Fig. 2: Comparisons of the functions carried by ICEs and CPs. **A.** Elements encoding (+) or lacking (-) replication or partition systems ("Rep or Par") or an integrase. **B.** Accessory functions over-represented (yellow) or under-represented (blue) in ICEs relative to CPs. Colored bars: significantly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

We then made similar analyses for functions usually regarded as accessory or unrelated to the biology of MGEs (Fig. 2B). ICEs were more likely to carry restriction-modification systems (x2.8) than CPs (but not orphan methylases), suggesting that ICEs endure stronger selective pressure for stabilization in the genome. In contrast, they were significantly less likely to carry antibiotic resistance genes or integrons. They also had fewer identifiable entry-exclusion systems, which may reflect the ability of ICEs to tolerate the presence of multiple similar elements in the cell (28). The classification of genes in the four major functional categories of the EggNOG database, showed that ICEs had relatively more genes encoding metabolic and cellular processes. We have previously shown that genes of unknown or poorly characterized function were over-represented in ICEs relative to their host chromosome (29). The frequency of these genes is even higher in plasmids (61% vs 46%). Hence, both types of elements have many functions in common, but their relative frequency often differs significantly.

Genetic similarities between ICEs and CPs

The results of the previous section, together with previously published studies (see Introduction), suggest that ICEs and CPs either share a common history or often exchange genes (or both). We detailed the relationships of homology between ICEs and CPs using the weighted Gene Repertoire Relatedness (wGRR), which measures the frequency of bi-directional best hits between two elements weighted by their sequence similarity (see Methods). We clustered the matrix of wGRR using the Louvain algorithm (30), and found six well-distinguished groups (Fig. 3). Two groups (1 and 6) are only constituted of CPs, two are composed of more than 90% of ICEs (3 and 5) and two have a mix of both types of elements (2 and 4) (Fig.3, top bar). Bacterial species are scattered between groups, showing that they are not the key determinant of the clustering. Some groups are only from γ-proteobacteria, but others include bacteria from different classes. Groups where elements are from the same taxonomic classes tend to have either CPs or ICEs, whereas the others have mixtures of both elements. Group 4, includes many ICEs and CPs, where all ICEs have integrases while more than half of the CPs lack both replication and partition systems (Fig. 4). This group includes almost only elements from ε-proteobacteria that may have specificities that we were not able to take into account. In contrast, almost all ICEs of groups 2 and 3 encode an integrase and all CPs have partition or replication systems.

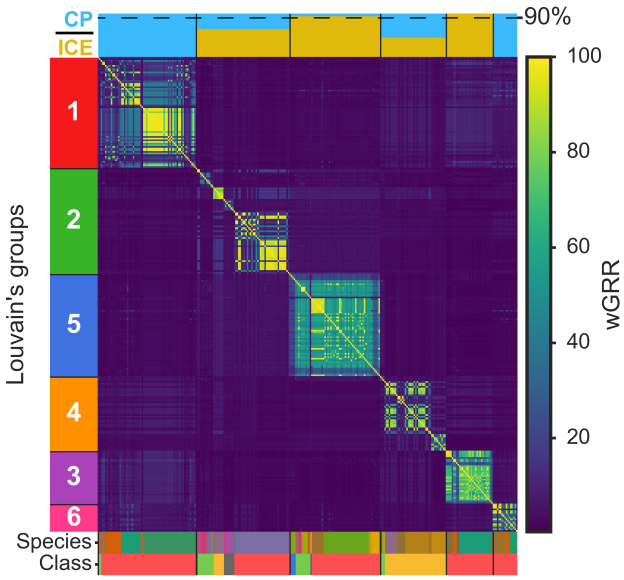
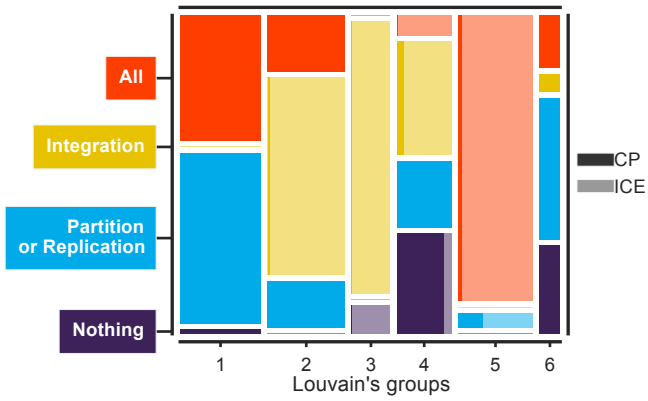


Fig. 3: Heatmap of the wGRR scores, ordered after the size of Louvain's group (depicted on the left bar). The top bar represents the proportion of ICEs (yellow) and CPs (blue) for each

175 group. The bottom bar assigns a color corresponding to the host's species or class (γ-
176 proteobacteria in red, β-proteobacteria in green, α-proteobacteria in blue, ε-
177 proteobacteria in orange, Fusobacteria and Acidobacteria in grey).



178

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

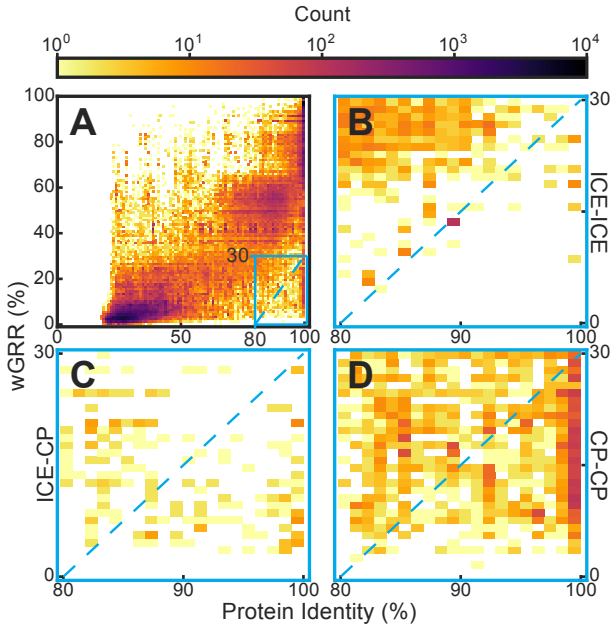
185

186 We controlled for the effect of the MPF genes in the previous clustering analysis by re-doing
187 it without these genes (Fig. S3). This produced the same number of groups - N1 to N6 - and
188 90% of the elements of the former groups were classed in the same novel groups (Fig. S4).
189 The only qualitatively significant difference between the two analyses concerned the group
190 2 for which 36% of the elements were now classed in groups N4 or N6. Overall, these
191 controls confirm that ICEs and CPs can be grouped together, and apart from other elements
192 of the same type. The grouping is not caused by sequence similarity between conjugative
193 systems. Instead, it probably reflects either within group genetic exchanges between ICEs
194 and CPs, or interconversions of the two types of elements.

195 Genetic exchanges: CPs become ICEs for broader host range

196 The clustering of ICEs and CPs could be explained by genetic transfer between them. To
197 address this question, we searched for pairs of conjugative elements with low wGRR (<30%)

198 but some highly similar homologs (>80% sequence identity). This identified several cases of
199 recent transfer of a single or a few genes between elements (Fig. 5A). In agreement with our
200 observations of higher genetic plasticity in CPs, most transfers took place between these
201 types of elements (Fig. 5.B-D).



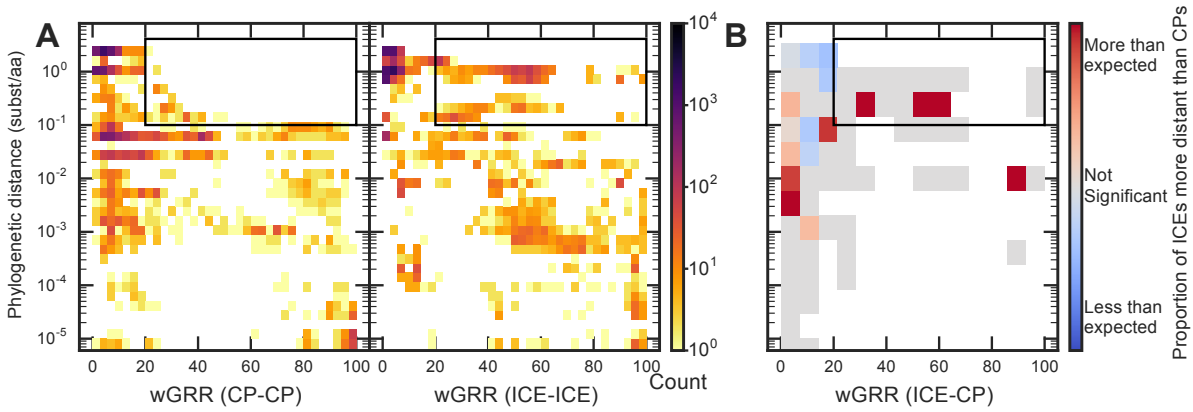
202

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

210

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

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



222
223 **Fig. 6**: wGRR as function of the host phylogenetic distance **A**. 2D histogram of the
224 distribution of the wGRR score as a function of the phylogenetic distance for pairs of CPs
225 (left) and pairs of ICEs (right). The bottom row corresponds to all pairs with distance lower
226 than 10^{-5} (including those in the same host). Elements in the black rectangle are depicted in
227 a phylogenetic tree in Fig. S5. **B**. Proportion of ICEs more distinct in terms of tri-nucleotides
228 from their host than CPs. Bins are larger than in A. to increase the power of the statistical
229 analysis. Color code: ICEs (Red) or CPs (Blue) are more distinct from the host in terms of tri-
230 nucleotide composition than the other element. Grey: not significant (binomial test, p-value
231 $> 10^{-2}$). White: no elements in the bin.

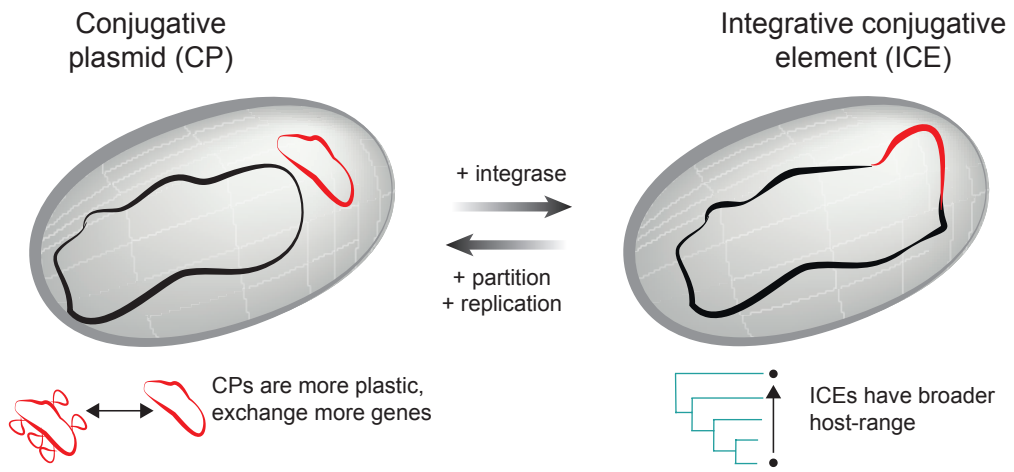
232

233 We then analyzed the pairs ICE-CP. We found few pairs of highly similar ICEs and CPs in
234 closely related hosts (bottom right corner of Fig. 6, n=8 for wGRR>50% and $d < 10^{-2}$),
235 suggesting that interconversion between these elements remains rare within a clade. A
236 larger number of ICE-CP pairs were very similar but present in distant hosts (n=38, Fig. 6).
237 The most parsimonious explanation for these observations, is the recent transfer of one of
238 the elements (ICE or CP) to a distant bacterial host. We identified the latter element based
239 on the differences in terms of tri-nucleotide composition between the elements and the
240 host chromosomes (defined as pvalue in (31)). We then computed for each ICE-CP pair the
241 difference between the pvalue of the pair ICE-host and that of the pair CP-host (see
242 Methods). In agreement to our observation that ICEs have broader host ranges, these

243 differences indicate that ICEs are relatively more distant from the host chromosome for
244 pairs with high wGRR in distant hosts than for the rest of the pairs (Wilcoxon rank sum test,
245 p-value<10^{-20}, Fig. 6B, and Fig. S7). The rarity of ICE-CP pairs in closely related hosts, their
246 abundance in distant hosts, and the identification that ICEs are the most compositionally
247 atypical relative to the host in the latter, suggest that successful transfer of CPs to distant
248 hosts is favored when they integrate the chromosome and become ICEs.
249

250 # Discussion

251 In this study, we quantified the differences between ICEs and CPs to evaluate the claims that
252 they are essentially equivalent MGEs (15, 16, 29). We found that numerous CPs have
253 integrases (although these may serve for dimer resolution and not integration (32)),
254 numerous ICEs encode replication and partition functions, the elements often cluster them
255 together, and they exchange genetic information. Furthermore, relaxases – present in both
256 ICEs and CPs – have been shown to act as integrases or replication initiators (33). Hence,
257 ICEs and CPs constitute closely related elements sharing many functions beyond those
258 related to conjugation. Yet, there are also some clear differences between them (Fig. 7).
259 First, genes encoding plasmid replicases and partition systems are more frequent in
260 plasmids, and tyrosine and serine recombinases are more frequent in ICEs. Interestingly, we
261 could not attribute incompatibility groups to ICEs, suggesting that the replication module is
262 rarely exchanged between ICEs and CPs. Second, the frequency of certain accessory traits is
263 different: plasmids are more likely to encode antibiotic resistance genes whereas ICEs
264 encode more metabolism-related genes, even if this could result from biases in the
265 database towards nosocomial pathogens. Finally, the %GC relative to the host, the number
266 of repeats, the patterns of gene variation and exchange, and the host range are
267 quantitatively different in the two types of elements. After integrating all this information,
268 we propose that in spite of their similarities each type has specific advantages.



269 **Fig. 7:** To integrate or not to integrate? ICEs and CPs have similar conjugative systems but
270 persist in different ways (resp. integrating the chromosome or replicating independently).

271 ICEs over-represent integrases, whereas plasmids over-represent replication and partition
272 functions (although both can have all of these functions). ICEs seem at an advantage when
273 conjugating to distant hosts, presumably because they integrate the chromosome. On the
274 other hand, plasmids have wider distribution of size and exchange more genetic
275 information.

276

277 Even if there are some known families of large (>200kb) ICEs (34, 35), most of these
278 elements in our dataset have a narrower variation in size than CPs. This suggests that CPs
279 are more flexible than ICEs in terms of the amount of genetic information they can carry and
280 in their ability to accommodate novel information. CPs also exchange genes more
281 frequently. Mechanistically, the rate of recombination between plasmids may be higher
282 because they encode more repeats, integrons, and transposable elements. Plasmid copy
283 number, when high, may also contribute to increase recombination rates. Interestingly,
284 recombination mediated by transposable elements has been shown to be drive the
285 evolution of certain plasmids(36, 37), and to accelerate the reduction of plasmid cost thus
286 stabilizing the element after horizontal transfer (38). The restrictions in size variation of ICEs
287 are probably not due to the mechanism of integration or excision because such reactions
288 occur between very distant recombination sites (39). Instead, very large ICEs may disrupt
289 chromosome organization by affecting the distribution of motifs, changing chromosome
290 folding domains, or unbalancing the sizes of replichores (40). Repeat-mediated
291 recombination leads to replicon rearrangements, and may lead to stronger counter-
292 selection of DNA repeats in ICEs than in CPs, further restricting their size variation and their
293 flux of gene exchange. Interestingly, the size of plasmids varies much more steeply with
294 genome size than the size of ICEs, suggesting that CPs may play a particularly important role
295 in the evolution of large bacterial genomes of Proteobacteria, which have higher rates of
296 genetic exchanges (41), and often contain mega-plasmids (42).

297 Some plasmids are known to be broad-host range and adapt to novel hosts, especially if
298 they carry adaptive traits that compensate for the initially poor intrinsic persistence of the
299 element (43, 44). However, within the large phylogenetic span considered in this work,
300 MPF_T ICEs have broader host ranges than CPs. Actually, the first known ICE, Tn916 (not

301 MPF_T, thus not included in this study), became notorious due to its ability to spread
302 antibiotic resistance between distant phyla (45).

303 Finally, we show novel evidence for interconversion between the two types of elements,
304 which had been proposed based on the phylogeny of the conjugative system (7). This is
305 consistent with the clustering of ICEs and CPs in terms of gene repertoires, even when
306 removing the conjugation system from the analysis, in certain groups and not in others.
307 Also, the transition of CPs to ICEs is the best explanation for the observations of the
308 presence of pairs of similar elements ICE-CP in distant hosts.

309 Other traits may provide advantages to ICEs or CPs. The ability of plasmids to modify their
310 copy number may accelerate adaptive evolutionary processes, such as the acquisition of
311 antibiotic resistance (46). On the other hand, ICEs might be more stably maintained in
312 lineages because they replicate within the chromosome. Finally, the carriage of ICEs and CPs
313 may have different costs. The cost of plasmids has been extensively studied and is strongly
314 dependent on the traits they encode (47). Much less is known about the cost of ICEs; several
315 reports suggest that they lead to low fitness costs when conjugation is not expressed, but
316 their fitness cost varies much more between elements during transfer (14). Direct
317 comparisons of the cost of carriage of ICE and CPs carrying similar traits are unavailable.
318 Further work will be needed to test these hypotheses.

319 Occasional transfers between CPs and ICEs allow them to access the other elements' gene
320 pool. These events may create elements with traits of ICEs and of CPs, as observed in more
321 than a third of all conjugative elements, and lead to their clustering in the wGRR group 5.
322 Additionally, they facilitate the interconversion of one type of element into the other.

323 Many elements are mobilizable but not able to conjugate independently (42, 48). These
324 elements often encode a relaxase that recognizes the element's origin of transfer and is able
325 to interact with a T4SS from an autonomously conjugative element to transfer to other cells.
326 Many of the disadvantages of conjugative plasmids and ICEs are similar to those of
327 mobilizable plasmids and integrative mobilizable elements, whether they encode a relaxase
328 or not. Notably, the former must be replicated in the extrachromosomal state, and the

329 latter integrate the genome where they must not disrupt genome organization. Patterns
330 observed in conjugative elements are thus likely to be applicable to mobilizable ones.

331 These results may also be relevant to understand lysogeny by temperate phages. The vast
332 majority of known prophages are integrated in the chromosome, but some replicate like
333 plasmids (49, 50). Considering that prophages share some of the constraints of conjugative
334 elements, they are likely to be under similar trade-offs. However, phages are under
335 additional constraints. Notably, their genome size is much less variable than that of
336 conjugative elements, because it must be packaged into the virion (51), and this may render
337 the extrachromosomal prophages less advantageous in terms of accumulating novel genes.
338 This could explain why most prophages are integrative whereas conjugative systems are
339 more evenly split between integrative and extrachromosomal elements.
340

341 ## Material and Methods

342 ### Data

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

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

352 Conjugative systems were detected using the CONJscan module of MacSyFinder (52), with
353 protein profiles and definitions of the MPF type T, published previously (53). ICEs were
354 delimited with the same methodology, as developed in a previous work (29). Briefly, we
355 identified the core genomes of the species. The region between two consecutive genes of
356 the core genome defined an interval in each chromosome. We then defined spots as the
357 sets of intervals in the chromosome flanked by genes of the same two families of the core
358 genome (24). We then identified the intervals and the spots with conjugative systems. The
359 information on the sets of gene families of the spots with ICEs (i.e., the spot pan-genome)
360 was used to delimit the element boundaries (script available at
361 https://gitlab.pasteur.fr/gem/spot_ICE). This methodology was shown to be accurate at the
362 gene level (precise nucleotide boundaries are not identifiable by this method, see (29)).

363 ### Functional analyses

364 Partition systems, replication systems, entry-exclusion systems and restriction modification
365 systems were annotated with HMM profiles, as described in our previous work (29, 54).
366 Integrases were annotated with the PFAM profile PF00589 for the Tyrosine recombinases
367 and the combination of PFAM profiles PF00239 and PF07508 for Serine recombinases. DDE
368 Transposases were detected with Macsyfinder (52) with models used previously (55).
369 Antibiotic resistance genes were detected with ResFams profiles (core version v1.1) (56)

370 using the --cut_ga option. We determined the functional categories of genes using their
371 annotation as provided by their best hits to the protein profiles of the EggNOG database for
372 bacteria (version 4.5, bactNOG) (57). Genes not annotated by the EggNOG profiles were
373 classed as "Unknown" and included in the "Poorly characterized" group. The HMM profiles
374 were used to search the genomes with HMMER 3.1b2 (58), and we retrieved the hits with
375 an e-value lower than 10^{-3} and with alignments covering at least 50% of the profile.
376 Integrons were detected using IntegronFinder version 1.5.1 with the --local_max option for
377 higher accuracy (59). Repeats (direct and inverted) were detected with Repseek (version
378 6.6) (60) using the option -p 0.001 which set the p-value for determining the minimum seed
379 length.

380 Statistics

381 We tested the over-representation of a given function or group of functions using Fisher's
382 exact tests on contingency tables. For partition, replication and integration, the contingency
383 table was made by splitting replicons in those encoding or not encoding the function and
384 between ICEs and CPs. The use of presence/absence data instead of the absolute counts
385 was made because the presence of at least one occurrence of a system is sufficient to have
386 the function and because the counts were always low. For the other functions, the
387 contingency table was made by splitting the proteins of the element in those annotated for
388 a given function and the remaining ones. This allowed to take into account the differences in
389 the number of genes between elements. The Fisher-exact tests were considered as
390 significant after sequential Holm-Bonferroni correction, with a family-wise error rate of 5%
391 (the probability of making at least one false rejection in the multiple tests, the type I error).
392 From the contingency table, we computed the relative ratio (or relative risk) of having a
393 given function more often in ICEs than in CPs. The relative ratio is computed as follow:
394 $RR = \frac{ICEwF/N_{ICE}}{CPwF/N_{CP}}$ where *ICEwF* is the number of ICE (or proteins in ICEs) with the given
395 function, and N_{ICE}, the total number of ICE (or proteins in ICEs), and likewise for CP. The
396 term $ICEwF/N_{ICE}$ is an estimation of the probability of an ICE (or a protein in an ICE) to
397 carry a given a function.

398 ## Phylogenetic distances

399 Phylogenetic distances were extract from the Proteobacterial tree of the Core-genome. To
400 build the tree, we identified the genes present in at least 90% of the 2897 genomes of
401 Proteobacteria larger than 1 Mb that were available in GenBank RefSeq in November 2016.
402 A list of orthologs was identified as reciprocal best hits using end-gap free global alignment.
403 Hits with less than 37% similarity in amino acid sequence and more than 20% difference in
404 protein length were discarded. We then identified the protein families with relations of
405 orthology in at least 90% of the genomes. They represent 341 protein families. We made
406 multiple alignments of each protein family with MAFFT v.7.205 (with default options) (61)
407 and removed poorly aligned regions with BMGE (with default options) (62). Genes missing in
408 a genome were replaced by stretches of "–" in each multiple alignment, which has been
409 shown to have little impact in phylogeny reconstruction (63). The tree of the concatenate
410 alignment was computed with FastTree version 2.1 under the LG model (64). We chose the
411 LG model because it was the one that minimized the AIC.

412 ## Distance to the host

413 We used the differences in tri-nucleotide composition to compute the genetic distance
414 between the mobile element and its host chromosome, as previously proposed (65). The
415 analysis of ICEs was done by comparing the element with the chromosome after the
416 removal of its sequence from the latter. Briefly, we computed the trinucleotide relative
417 abundance ($x_{ijk} \forall i, j, k \in \{A, T, C, G\}$) for the chromosomes (in windows of 5 kb) and for the
418 conjugative elements (entire replicon), which is given by: $x_{ijk} = f_{ijk}/f_i f_j f_k$, with f the
419 frequency of a given k-mer in the sequence (31). We first computed the Mahalanobis
420 distance between each window and the host chromosome as follow:

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

421 with *w*, the vector of tri-nucleotide abundances (x_{ijk}) in a given window, and *h*, the mean of
422 the vector of x_{ijk} (*i.e.*, the average tri-nucleotide abundance in the chromosome). *H* is the
423 covariance matrix of the tri-nucleotide relative abundances. The inverse of the covariance
424 matrix (H^{-1}) downweights frequent trinucleotides, like the tri-nucleotides corresponding to
425 start codons, which are common to conjugative elements and chromosome and could bias

426 the distance. We computed the Mahalanobis distance between conjugative elements and
427 their hosts' chromosomes (same formula as above, but *w* is now for a conjugative element
428 instead of a chromosome window). We then computed the probability (p-value) that the
429 measured distance between a conjugative element and the host's chromosome is the same
430 as any fragment of the host's chromosome.

431 We compared ICEs and CPs in relation to their compositional distance to the host. For this,
432 we made the null hypothesis that the proportion of ICEs having a p-value lower than CPs
433 follows a binomial distribution whose expected proportion is that of the entire dataset (the
434 proportion of ICEs having a p-value lower than CP), precisely:
435 $H_0 = N(pvalue_{ICE} < pvalue_{CP})/N_{Comparisons}$, where $N_{Comparisons}$ is the total number of
436 ICE-CP pairs, *i.e.* $151 \times 136 = 20536$.

437 Network analysis of gene repertoire relatedness

438 We built a network describing the relations of homology between the elements. The nodes
439 in the network are conjugative elements and they are linked if they share a given relation of
440 homology. More precisely, the relationship between two elements was quantified with the
441 weighted Gene Repertoire Relatedness score (wGRR). This score represents the number of
442 homologous proteins between two elements, weighted by their sequence identity, as
443 described in (29). The formula is:

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

444 Where (A_i, B_i) is the ith pair of homologous protein between element A and element B,
445 $id(A_i,B_i)$ is the sequence identity of their alignment, min(A,B) is the number of proteins of
446 the element with fewest proteins (A or B). The sequence identity was computed with blastp
447 v.2.2.15 (default parameters)(66) and kept all bi-directional best hits with an e-value lower
448 than 10^{-5}.

449 The network was built based on the wGRR matrix. Its representation was made using the
450 Fruchterman-Reingold force-directed algorithm as implemented in the NetworkX v1.11
451 python library. The groups were made using the Louvain algorithm (30). We controlled for

452 the consistency of the heuristic used to assess that the group found are not form a local
453 optimal. We performed 100 clustering, which led to the same classification in 95% of the
454 time.

455 ### Incompatibility typing

456 We determined the incompatibility group of replicons using the method of PlasmidFinder
457 (67). We used BLASTN (66) to search the replicons for sequences matching the set of 116
458 probes used by PlasmidFinder. We kept the hits with a coverage above 60% and sequence
459 identity above 80%, as recommended by the authors. Around 3% of the elements had
460 multiple incompatibility types attributed.

461

462 ## Acknowledgements

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

470

471 # References

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

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

476 3. Ebersbach G, Gerdes K (2005) Plasmid segregation mechanisms. *Annu Rev Genet*
477 39:453–79.

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

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

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

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

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

488 9. Lederberg J (1998) Plasmid (1952 – 1997). *Plasmid* 9:1–9.

489 10. Jacob F, Schaeffer P, Wollman EL (1960) Episomic Element in Bacteria. *Symp. Soc.*
490 *Gen. Microbiol*, pp 67–91.

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

493 12. Bellanger X, Payot S, Leblond-bourget N, Guédon G (2014) Conjugative and
494 mobilizable genomic islands in bacteria: evolution and diversity. *FEMS Microbiol Rev*

495 38:720–760.

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

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

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

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

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

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

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

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

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

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

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

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

523 25. Carraro N, et al. (2016) Plasmid-like replication of a minimal streptococcal integrative
524 and conjugative element. *Microbiol (United Kingdom)* 162(4):622–632.

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

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

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

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

535 30. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of
536 communities in large networks. *J Stat Mech*. doi:10.1088/1742-
537 5468/2008/10/P10008.

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

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

542 33. Wawrzyniak P, Płucienniczak G, Bartosik D (2017) The Different Faces of Rolling-Circle
543 Replication and Its Multifunctional Initiator Proteins. *Front Microbiol* 8(November):1–
544 13.

545 34. Sullivan JT, Ronson CW (1998) Evolution of rhizobia by acquisition of a 500-kb
546 symbiosis island that integrates into a phe-tRNA gene. *Proc Natl Acad Sci*

547 95(April):5145–5149.

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

550 36. Hall JPJ, Williams D, Paterson S, Harrison E, Brockhurst MA (2017) Positive selection
551 inhibits gene mobilization and transfer in soil bacterial communities. *Nat Ecol Evol*
552 1(9):1348–1353.

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

555 38. Porse A, Schønning K, Munck C, Sommer MOA (2016) Survival and Evolution of a
556 Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol Biol Evol*
557 33(11):2860–2873.

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

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

563 41. Oliveira PH, Touchon M, Rocha EPC (2016) Regulation of genetic flux between
564 bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 113(20):5658–
565 63.

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

568 43. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM (2008) Adaptive plasmid
569 evolution results in host-range expansion of a broad-host-range plasmid. *Genetics*
570 178(4):2179–2190.

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

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

575 46. San Millan A, Escudero JA, Gifford DR, Mazel D, Maclean RC (2016) Multicopy
576 plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol*
577 1(10):1–8.

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

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

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

584 50. Łobocka MB, et al. (2004) Genome of Bacteriophage P1. *J Bacteriol* 186(21):7032–
585 7068.

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

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

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

595 54. Oliveira PH, Touchon M, Rocha EPC (2014) The interplay of restriction-modification
596 systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res*
597 42(21):1–14.

598 55. Touchon M, et al. (2014) The Genomic Diversification of the Whole Acinetobacter

599 Genus: Origins, Mechanisms, and Consequences. *Genome Biol Evol* 6(10):2866–2882.

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

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

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

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

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

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

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

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

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

620 65. Suzuki H, Sota M, Brown CJ, Top EM (2008) Using Mahalanobis distance to compare
621 genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res*
622 36(22):e147.

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

625 67. Carattoli A, et al. (2014) In Silico detection and typing of plasmids using plasmidfinder
626 and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58(7):3895–
627 3903.

628

629