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),

Instituto de Biomedicina y Biotecnología de Cantabria, Universidad de Cantabria-CSIC, 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. We observed that plasmids encode more replicases, partition systems, and antibiotic resistance genes, whereas ICEs encode more integrases and metabolism-associated genes. ICEs and plasmids have similar average sizes, but plasmids are much more variable, have more DNA repeats, and exchange genes more frequently. On the other hand, we found that ICEs are more frequently transferred between distant taxa. We propose a model where differential plasticity and transmissibility range explain the co-occurrence of integrative and extra-chromosomal elements in microbial populations. In particular, the conversion from ICE to plasmid allows ICE to be more plastic, while the conversion from plasmid to ICE allows the expansion of the element's host range.

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

35 ## Introduction

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

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

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

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

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

88

89

Results

Functional and genetic differences between ICEs and CPs

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

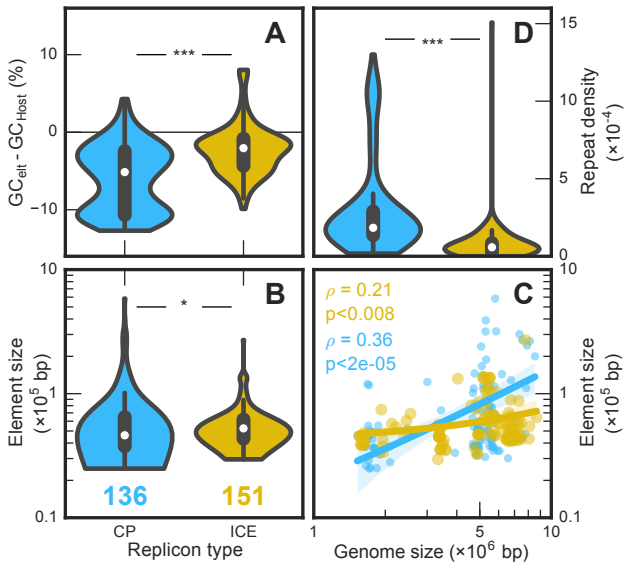


Fig. 1: Comparison between 136 CPs (blue) and 151 ICEs (yellow) in composition and sizes. **A**. CPs are AT richer than ICEs relative to their hosts (Wilcoxon rank sum test, p-value<10^{-3}). **B**. ICEs and CPs have different distributions of size (same test, p-value<0.05). Median sizes: ICEs (52.5 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

111 in panel B. **D**. Density of repeats is higher in CPs than in ICEs (0.30 vs 0.078 repeats per kb,
112 same test, p-value < 10^{-10}).
113

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

121 We then quantified the differences between ICEs and CPs in terms of functions associated
122 with their biology, with a focus on stabilization functions. Relaxases are part of the rolling
123 circle replication initiator proteins and some have been shown to act as replicases(16, 25) or
124 site-specific recombinases (26, 27). Since all conjugative elements encode a relaxase, by
125 definition, they may also have these functions. In the following, we focused on typical
126 plasmid replication initiator proteins (more than 95% of them are involved in theta-
127 replication, and none is matched by the protein profiles of relaxases), and serine or tyrosine
128 recombinases as integrases. Expectedly, ICEs showed higher frequency of integrases, while
129 CPs had more frequently identifiable partition and replication systems. Some ICEs encode
130 partition systems (11%) and many encode a replicase (40%), while 37% of CPs encode at
131 least one tyrosine or serine recombinase (Fig. 2A). These results further illustrate a
132 continuum between the two types of elements: about half of the elements (40% and 48%,
133 ICEs and CPs, respectively) have functions usually associated with the other type and may
134 (rarely) lack functions typically associated with its own type (Fig. 2B). Interestingly, we
135 observe that ICEs containing replication or partition systems contain more repeats per
136 nucleotide than the others (Wilcoxon rank-sum test, p-value = 0.02). We identified plasmid
137 incompatibility systems of diverse types, whereas ICE could not be typed in the current
138 scheme (Fig. S2).

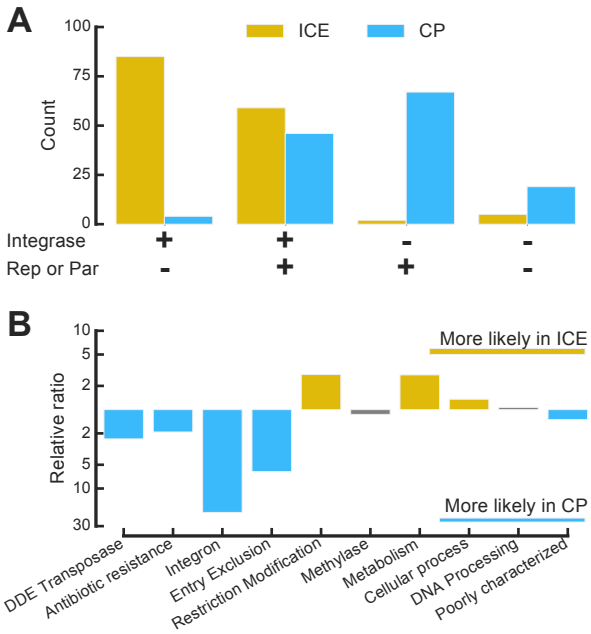


Fig. 2: Comparisons of the functions carried by ICEs and CPs. **A.** Elements encoding (+) or lacking (-) replication or partition systems ("Rep or Par") or an integrase. **B.** Accessory functions over-represented (yellow) or under-represented (blue) in ICEs relative to CPs. Colored bars: significantly 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.

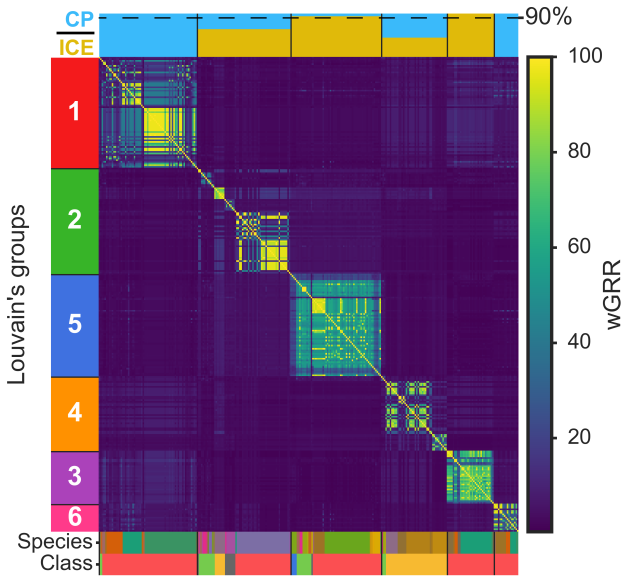
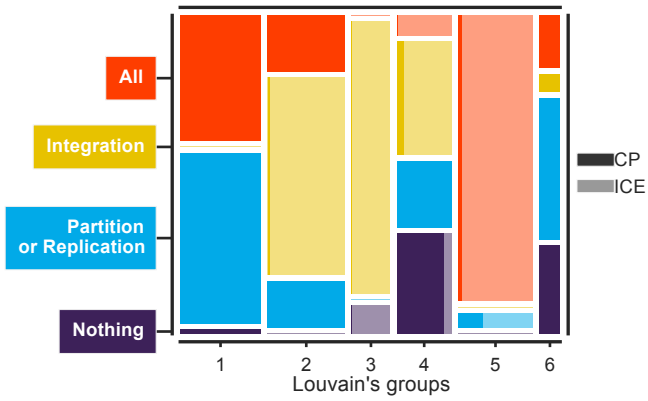


Fig. 3: Heatmap of the wGRR scores, ordered after the size of Louvain's group (depicted on the left bar). The top bar represents the proportion of ICEs (yellow) and CPs (blue) for each

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



183

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

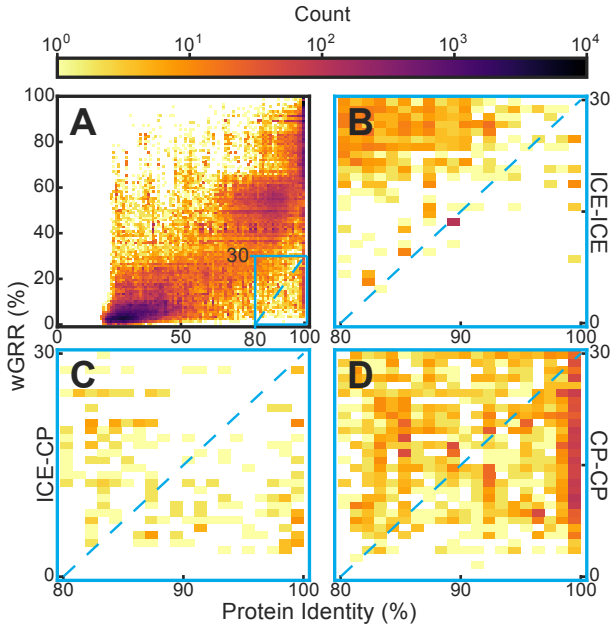
190

191 We controlled for the effect of the MPF genes in the previous clustering analysis by re-doing
192 it without these genes (Fig. S3). This produced the same number of groups - N1 to N6 - and
193 90% of the elements of the former groups were classed in the same novel groups (Fig. S4).
194 The only qualitatively significant difference between the two analyses concerned the group
195 2 for which 36% of the elements were now classed in groups N4 or N6. Overall, these
196 controls confirm that ICEs and CPs can be grouped together, and apart from other elements
197 of the same type. The grouping is not caused by sequence similarity between conjugative
198 systems. Instead, it probably reflects either within group genetic exchanges between ICEs
199 and CPs, or interconversions of the two types of elements.

200 Genetic exchanges: CPs become ICEs for broader host range

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

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



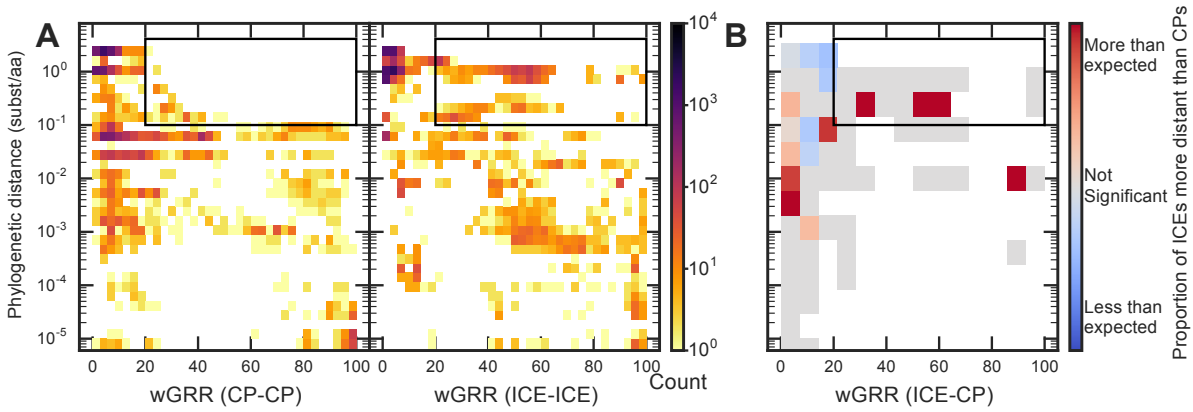
207

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

215

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

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



227
228 **Fig. 6**: wGRR as function of the host phylogenetic distance **A**. 2D histogram of the
229 distribution of the wGRR score as a function of the phylogenetic distance for pairs of CPs
230 (left) and pairs of ICEs (right). The bottom row corresponds to all pairs with distance lower
231 than 10^{-5} (including those in the same host). Elements in the black rectangle are depicted in
232 a phylogenetic tree in Fig. S5. **B**. Proportion of ICEs more distinct in terms of tri-nucleotides
233 from their host than CPs. Bins are larger than in A. to increase the power of the statistical
234 analysis. Color code: ICEs (Red) or CPs (Blue) are more distinct from the host in terms of tri-
235 nucleotide composition than the other element. Grey: not significant (binomial test, p-value
236 $> 10^{-2}$). White: no elements in the bin.

237

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

248 differences indicate that ICEs are relatively more distant from the host chromosome for
249 pairs with high wGRR in distant hosts than for the rest of the pairs (Wilcoxon rank sum test,
250 p-value<10^{-20}, Fig. 6B, and Fig. S7). The rarity of ICE-CP pairs in closely related hosts, their
251 abundance in distant hosts, and the identification that ICEs are the most compositionally
252 atypical relative to the host in the latter, suggest that successful transfer of CPs to distant
253 hosts is favored when they integrate the chromosome and become ICEs.
254

255 # Discussion

256 In this study, we compared the genetic organization of ICEs and CPs to evaluate the
257 hypothesis that they are essentially equivalent MGEs (15, 16, 29). We found that numerous
258 CPs have integrases (although these may serve for dimer resolution and not integration
259 (32)), and numerous ICEs encode replication and partition functions. Relaxases – present in
260 both ICEs and CPs – have been shown to act as integrases or replication initiators in certain
261 elements (33), which provides further functional overlapping between the elements.
262 Hence, ICEs and CPs share many functions beyond those related to conjugation. These
263 similarities explain why they can cluster together in terms of gene repertoires, even when
264 excluding conjugation functions.

265 There are also some clear differences between CPs and ICEs explaining why they can also
266 group separately. First, genes encoding replicases and partition systems are more frequent
267 in plasmids, while tyrosine and serine recombinases are more frequent in ICEs. Interestingly,
268 we could not attribute incompatibility groups to ICEs, suggesting that, either the replication
269 module is rarely exchanged between ICEs and CPs, or it evolves too rapidly. Second, the
270 frequency of certain accessory traits is different: plasmids are more likely to encode
271 antibiotic resistance genes whereas ICEs encode more metabolism-related genes. We
272 controlled for taxonomical bias by analyzing ICEs and CPs from the same genera, yet the
273 precise functions we observed are dependent on the dataset which, here, is biased toward
274 nosocomial pathogens. A dataset of bacteria from other environments might present other
275 functional differences in terms of the traits they carry. Finally, the difference in %GC content
276 relative to the host, the number of repeats, the patterns of gene variation and exchange,
277 and the host range are quantitatively different in the two types of elements. After
278 integrating all this information, we propose that, in spite of their similarities, each type has
279 advantages favored in specific situations. In particular, our results suggest that
280 interconversion between ICEs and CPs gives them access to the higher genetic plasticity of
281 CPs and the broader host range of ICEs (Fig. 7).

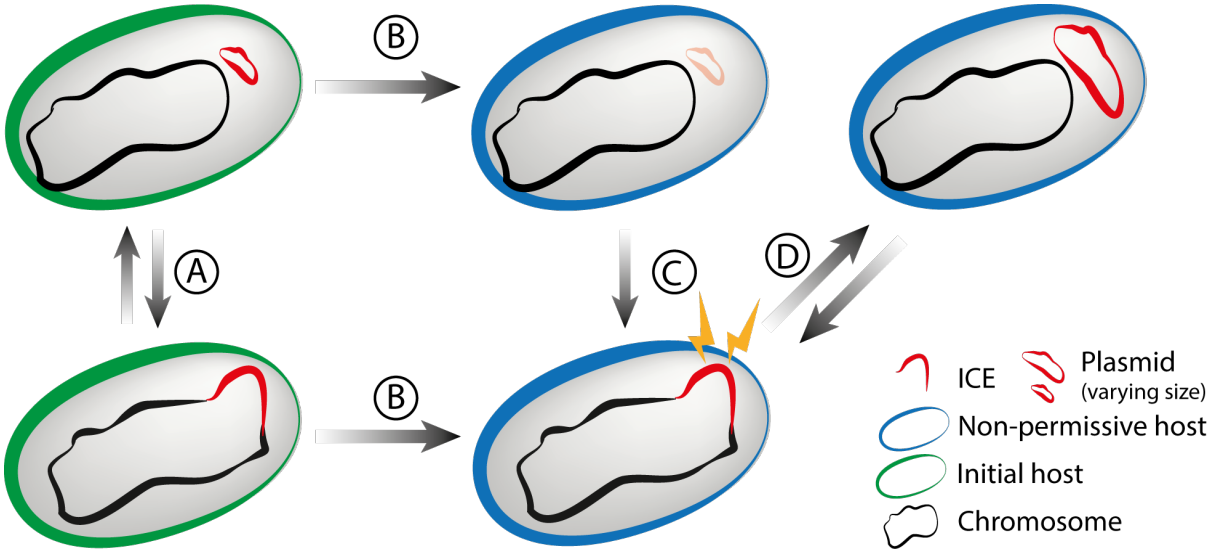


Fig. 7: Interconversion between ICEs and CPs allows access to the higher genetic plasticity of CPs and the broader host range of ICEs. **A**. Many ICEs and CPs encode the functions of the other element, facilitating CP<->ICE interconversion. **B**. Conjugative transfer to a taxonomically distant host often precludes stabilization of CPs because they are non-permissive for plasmid replication. ICEs seem at an advantage when conjugating to distant hosts, presumably because they integrate the chromosome. **C**. CPs can be salvaged after transfer in a non-permissive host if they integrate the chromosome (as ICE). We call this effect host-range expansion. **D**. Mutations and/or gene acquisitions can result in an ICE with the ability to replicate as a plasmid. CPs have wider distribution of size and exchange more genetic information, which allows them to acquire novel genetic information at higher rates, which may eventually be adaptive for their hosts.

Even if there are some known families of large (>200kb) ICEs (34, 35), these elements have a remarkably narrower variation in size than CPs in our dataset. This suggests that CPs are more flexible than ICEs in terms of the amount of genetic information they can carry and in their ability to accommodate novel information. We also show that CPs exchange genes more frequently. Mechanistically, this higher rate of gene exchange between CPs may result from recombination between DNA repeats, gene acquisition by integrons, and genome rearrangements by transposable elements. We showed that CPs encode all of these in higher number. Plasmid copy number, when high, may also contribute to increase recombination rates in plasmids relative to ICEs. Interestingly, recombination mediated by transposable elements has been shown to drive the evolution of certain plasmids (36, 37) and to accelerate the reduction of plasmid cost, thus stabilizing the element after horizontal

306 transfer (38). The restrictions in size variation of ICEs are probably not due to the
307 mechanism of integration or excision, because such reactions occur between very distant
308 recombination sites (39). Instead, very large ICEs may disrupt chromosome organization by
309 affecting the distribution of motifs, changing chromosome folding domains, or unbalancing
310 the sizes of replichores (40). Repeat-mediated recombination leads to replicon
311 rearrangements, and may lead to stronger counter-selection of DNA repeats in ICEs than in
312 CPs, further restricting their size variation and their flux of gene exchange. Interestingly, the
313 size of plasmids varies much more steeply with genome size than the size of ICEs, suggesting
314 that CPs may play a particularly important role in the evolution of large bacterial genomes
315 of Proteobacteria, which have higher rates of genetic exchanges (41), and often contain
316 mega-plasmids (42).

317 Some plasmids are known to be broad-host range and adapt to novel hosts, especially if
318 they carry adaptive traits that compensate for the initially poor intrinsic persistence of the
319 element (43, 44). However, within the large phylogenetic span considered in this work,
320 MPF_T ICEs have broader host ranges than CPs. Actually, the first known ICE, Tn916 (not
321 MPF_T, thus not included in this study), became notorious due to its ability to spread
322 antibiotic resistance between distant phyla (45). We propose that the plasmid to ICE
323 conversion can elicit conjugative transmission to otherwise non-permissive hosts. This
324 results in an effective expansion of the host range of the element (now an ICE). Once
325 installed in the recipient chromosome, the ICE can incorporate a new replication system, or
326 freely mutate its own. If it obtains a functional replication system in the new host the
327 reverse interconversion may occur (Fig. 7). This results in a plasmid with a new host range
328 and the higher evolvability of an autonomous plasmid.

329 The similarities between certain groups of ICEs and CPs in terms of gene repertoires, the
330 integration of CPs as ICEs upon long-range transfer, and the exchange of genetic information
331 between them are novel evidence for interconversion between the two types of elements.
332 This had previously been proposed based on the phylogeny of the conjugative system (7).
333 Inter-conversion of elements and occasional transfers between CPs and ICEs allow them to
334 access the other elements' gene pool. These events may result in conjugative elements
335 sharing many traits of the other type of element, as we observed in more than a third of all

336 conjugative elements, and should produce many genetic similarities between ICEs and CPs.
337 The latter probably facilitate further interconversions between the elements.

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

348 Many mobile elements are mobilizable but not able to conjugate independently (42, 48).
349 These elements often encode a relaxase that recognizes the element's origin of transfer and
350 is able to interact with a T4SS from an autonomously conjugative element to transfer to
351 other cells. Some other elements only contain an origin of transfer that is recognized by a
352 relaxase of another element. Many of the disadvantages of CPs and ICEs are similar to those
353 of mobilizable plasmids and integrative mobilizable elements, whether they encode a
354 relaxase or not. Notably, the former must be replicated in the extrachromosomal state, and
355 the latter integrate the genome where they must not disrupt genome organization. Patterns
356 observed in conjugative elements are thus likely to be applicable to mobilizable ones.

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

366 In summary, our results show that there are specific fitness benefits associated with the
367 divergent lifestyles identified for pairs of highly similar ICEs and CPs. We should emphasize
368 that our model proposes that plasmid to ICE transition results in broadening the host range
369 of the element, with a concomitant fitness benefit associated with higher rates of its
370 horizontal transfer. The ICE to plasmid transition, on the other hand, results in added
371 versatility of the cargo content (increased size range, etc.), which enhances the evolvability
372 of the mobile element. These factors may promote or even drive the interconversion
373 between plasmids and ICEs in an ever-changing environment. The concepts presented in
374 this work will provide a better understanding of the evolution of bacterial genomes and
375 their mobile genetic elements.
376

377 ## Material and Methods

378 ### Data

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

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

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

399 ### Functional analyses

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

406 using the --cut_ga option. We determined the functional categories of genes using their
407 annotation as provided by their best hits to the protein profiles of the EggNOG database for
408 bacteria (version 4.5, bactNOG) (57). Genes not annotated by the EggNOG profiles were
409 classed as "Unknown" and included in the "Poorly characterized" group. The HMM profiles
410 were used to search the genomes with HMMER 3.1b2 (58), and we retrieved the hits with
411 an e-value lower than 10^{-3} and with alignments covering at least 50% of the profile.
412 Integrons were detected using IntegronFinder version 1.5.1 with the --local_max option for
413 higher accuracy (59). Repeats (direct and inverted) were detected with Repseek (version
414 6.6) (60) using the option -p 0.001 which set the p-value for determining the minimum seed
415 length.

416 Statistics

417 We tested the over-representation of a given function or group of functions using Fisher's
418 exact tests on contingency tables. For partition, replication and integration, the contingency
419 table was made by splitting replicons in those encoding or not encoding the function and
420 between ICEs and CPs. The use of presence/absence data instead of the absolute counts
421 was made because the presence of at least one occurrence of a system is sufficient to have
422 the function and because the counts were always low. For the other functions, the
423 contingency table was made by splitting the proteins of the element in those annotated for
424 a given function and the remaining ones. This allowed to take into account the differences in
425 the number of genes between elements. The Fisher-exact tests were considered as
426 significant after sequential Holm-Bonferroni correction, with a family-wise error rate of 5%
427 (the probability of making at least one false rejection in the multiple tests, the type I error).
428 From the contingency table, we computed the relative ratio (or relative risk) of having a
429 given function more often in ICEs than in CPs. The relative ratio is computed as follow:
430 $RR = \frac{ICEwF/N_{ICE}}{CPwF/N_{CP}}$ where *ICEwF* is the number of ICE (or proteins in ICEs) with the given
431 function, and N_{ICE}, the total number of ICE (or proteins in ICEs), and likewise for CP. The
432 term $ICEwF/N_{ICE}$ is an estimation of the probability of an ICE (or a protein in an ICE) to
433 carry a given a function.

434 ## Phylogenetic distances

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

448 ## Distance to the host

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

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

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

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

467 We compared ICEs and CPs in relation to their compositional distance to the host. For this,
468 we made the null hypothesis that the proportion of ICEs having a p-value lower than CPs
469 follows a binomial distribution whose expected proportion is that of the entire dataset (the
470 proportion of ICEs having a p-value lower than CP), precisely:
471 $H_0 = N(pvalue_{ICE} < pvalue_{CP})/N_{Comparisons}$, where $N_{Comparisons}$ is the total number of
472 ICE-CP pairs, *i.e.* $151 \times 136 = 20536$.

473 Network analysis of gene repertoire relatedness

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

$$wGRR_{A,B} = \sum_{i=1}^{N} \frac{id(A_i, B_i)}{min(A, B)} \overset{\leftrightarrow}{iff} evalue(A_i, B_i) < 10^{-5}$$

480 Where (A_i, B_i) is the ith pair among N pairs of homologous proteins between element A and
481 element B, id(A_i,B_i) is the sequence identity of their alignment, min(A,B) is the number of
482 proteins of the element with fewest proteins (A or B). The sequence identity was computed
483 with blastp v.2.2.15 (default parameters)(66) and kept all bi-directional best hits with an e-
484 value lower than 10^{-5}.

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

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

491 ### Incompatibility typing

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

497

498 ## Acknowledgements

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

506

507 ## References

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

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

512 3. Ebersbach G, Gerdes K (2005) Plasmid segregation mechanisms. *Annu Rev Genet*
513 39:453–79.

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

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

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

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

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

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

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

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

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

531 38:720–760.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

571 30. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of
572 communities in large networks. *J Stat Mech*. doi:10.1088/1742-
573 5468/2008/10/P10008.

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

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

578 33. Wawrzyniak P, Płucienniczak G, Bartosik D (2017) The Different Faces of Rolling-Circle
579 Replication and Its Multifunctional Initiator Proteins. *Front Microbiol* 8(November):1–
580 13.

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

583 95(April):5145–5149.

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

586 36. Hall JPJ, Williams D, Paterson S, Harrison E, Brockhurst MA (2017) Positive selection
587 inhibits gene mobilization and transfer in soil bacterial communities. *Nat Ecol Evol*
588 1(9):1348–1353.

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

591 38. Porse A, Schønning K, Munck C, Sommer MOA (2016) Survival and Evolution of a
592 Large Multidrug Resistance Plasmid in New Clinical Bacterial Hosts. *Mol Biol Evol*
593 33(11):2860–2873.

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

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

599 41. Oliveira PH, Touchon M, Rocha EPC (2016) Regulation of genetic flux between
600 bacteria by restriction-modification systems. *Proc Natl Acad Sci U S A* 113(20):5658–
601 63.

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

604 43. De Gelder L, Williams JJ, Ponciano JM, Sota M, Top EM (2008) Adaptive plasmid
605 evolution results in host-range expansion of a broad-host-range plasmid. *Genetics*
606 178(4):2179–2190.

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

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

611 46. San Millan A, Escudero JA, Gifford DR, Mazel D, Maclean RC (2016) Multicopy
612 plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat Ecol Evol*
613 1(10):1–8.

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

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

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

620 50. Łobocka MB, et al. (2004) Genome of Bacteriophage P1. *J Bacteriol* 186(21):7032–
621 7068.

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

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

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

631 54. Oliveira PH, Touchon M, Rocha EPC (2014) The interplay of restriction-modification
632 systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res*
633 42(21):1–14.

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

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

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

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

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

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

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

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

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

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

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

656 65. Suzuki H, Sota M, Brown CJ, Top EM (2008) Using Mahalanobis distance to compare
657 genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res*
658 36(22):e147.

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

661 67. Carattoli A, et al. (2014) In Silico detection and typing of plasmids using plasmidfinder
662 and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58(7):3895–
663 3903.

664

665