Pseudogenes in plasmid genomes reveal past transitions in plasmid mobility

Dustin M. Hanke[1*], Yiqing Wang[1], Tal Dagan[1]

[1]Institute of General Microbiology, Kiel University, Kiel, Germany

[*]Corresponding author e-mail: dhanke@ifam.uni-kiel.de

Abstract

Evidence for gene non-functionalization due to mutational processes is found in genomes in the form of pseudogenes. Pseudogenes are known to be rare in prokaryote chromosomes, with the exception of lineages that underwent an extreme genome reduction (e.g., obligatory pathogens). Much less is known about the frequency of pseudogenes in prokaryotic plasmids; those are genetic elements that can transfer between cells and may encode beneficial traits for their host. Non-functionalization of plasmid-encoded genes may alter the plasmid characteristics, e.g., mobility, or their effect on the host. Analyzing 10,832 prokaryotic genomes, we find that plasmid genomes are characterized by threefold-higher pseudogene density compared to chromosomes. The majority of plasmid pseudogenes correspond to deteriorated transposable elements. A detailed analysis of enterobacterial plasmids furthermore reveals frequent gene non-functionalization events associated with the loss of plasmid self-transmissibility. Reconstructing the evolution of closely related plasmids revealed that non-functionalization of the conjugation machinery led to the emergence of non-mobilizable plasmid types. Examples are virulence plasmids in *Escherichia* and *Salmonella*. Our study highlights non-functionalization of core plasmid mobility functions as one route for the evolution of domesticated plasmids. Pseudogenes in plasmids supply insights into past transitions in plasmid mobility that are akin to transitions in bacterial lifestyle.

Keywords: Phylogenomics, Plasmid evolution, Structural variants, pO157, pSENV, pSLT

Introduction

Pseudogenes are gene copies that have been rendered non-functional due to mutations that disrupt their expression and/or function. Non-functionalization events that generate unitary pseudogenes, which lack a functional homolog in the genome are rather rare (e.g., (1)). In eukaryotes, pseudogenes frequently originate from non-functionalization of gene copies following gene duplication (2). In prokaryotes, protein family diversification is more often mediated by horizontal gene transfer (HGT) that introduces xenologs in the genome, compared to gene duplication (i.e., paralogs) (3, 4). Nonetheless, acquired genes may become non-functional upon the integration into the genome (5), e.g., due to incompatibility with the transcriptional or regulatory network of the recipient organism. Gene families of transposable elements (TEs) (e.g., transposases and site-specific recombinases) are characterized by a high frequency of pseudogenes, indicating that transposition and translocation play a major role in the evolution of pseudogenes (6). Pseudogenes correspond to molecular fossil record of past genes; the sequence similarity between pseudogenes and their functional homolog is expected to decay due to neutral mutational processes over time (7).

The frequency of pseudogenes in prokaryotic genomes is lower compared to eukaryotes, with several exceptions, including the genome *Mycobacterium leprea*, *Rickettsia prowazekii* (8), and *Salmonella enterica* serovars Typhi and Paratyphi A (9). It has been previously suggested that non-functional DNA in prokaryotes is rapidly eliminated by deletions due to the energetic cost of gene expression (10). The prevalence of pseudogenes in prokaryotes may be furthermore associated with bacterial lifestyles. Genomes of facultative pathogens are characterized by a high pseudogene frequency compared to free-living bacteria (11, 12), which may correspond to an initial step in the evolution of genome reduction that is common in the adaptation to an obligate host-associated lifestyle (13).

Plasmids are extrachromosomal elements that play an important role in prokaryote evolution (14). The evolution of plasmids differs from that of bacterial chromosomes due to their relatively small genome size, variable copy number in the cell, and their ability to horizontally transfer between cells. Plasmids can be either self-transmissible (i.e., conjugative), mobilizable, or non-mobilizable, and this property can be inferred from their gene content. Plasmids having the complete set of genetic features required for conjugation, e.g., genes encoding the mating pair formation (MPF) machinery, are classified as self-transmissible; a plasmid is classified as mobilizable, if it encodes an *oriT* binding site or relaxase genes, but lacks the MPF machinery gene markers (e.g., (15)). Mobilizable plasmids are able to hijack the conjugation machinery of

other plasmids for horizontal transfer (reviewed in (16)). Plasmids lacking an *oriT* and a relaxase gene are typically classified as non-mobilizable (e.g., (15)). Nonetheless, it has been suggested that plasmids classified as non-mobilizable with unidentified *oriT* might utilize relaxases of conjugative or mobilizable plasmids for transfer (17). Plasmid mobility is an evolvable property, for example, an evolution experiment of plasmid persistence in *Escherichia coli* under strong selection exemplified a loss of plasmid self-transmissibility due to a large deletion of genes required for conjugation (18). Indeed, a recent large-scale study of plasmid genomes suggested that conjugative plasmids frequently lose their self-transmissibility, with consequent changes in the plasmid genetic repertoire due to gene non-functionalization and gene loss (19). Hence, plasmids that harbor conjugative origins of transfer but lacking a relaxase and MPF machinery may evolved from ancestral conjugative plasmids (17).

Here we study the extent of gene non-functionalization in plasmid genomes. For that purpose, we compare the prevalence of pseudogenes between plasmids and chromosomes using a large-scale analysis of 10,832 publicly available genomes of 738 prokaryote genera. We further study the evolutionary history of plasmid-encoded pseudogenes in 2,442 *Enterobacteriaceae* isolates including the genera *Klebsiella*, *Escherichia*, and *Salmonella*. Comparing homologous plasmids, we further reconstruct the origin of pseudogenes in the context of transitions in plasmid mobility.

Results

Higher density of pseudogenes in plasmids compared to chromosomes

To examine the presence of pseudogenes in plasmids and chromosomes we surveyed 10,832 genome annotation from the Prokaryotic Genome Annotation Pipeline of NCBI's RefSeq dataset (20). Pseudogenes were annotated in 21,564 (74%) plasmids and 10,832 (100%) chromosomes. Our results reveal that plasmids are characterized by a higher pseudogene density (pseudogene per CDS) compared to chromosomes (Fig. 1A), with a median of 0.070 pseudogenes per coding sequence (CDS) for plasmids and 0.024 for chromosomes ($P < 0.001$, using Wilcoxon test). This equates to approximately 14 CDS per pseudogene for plasmids and 42 CDS per pseudogene for chromosomes. Thus, the genomic composition of plasmids is enriched by a 3-fold higher pseudogene density compared to chromosomes.

The size difference between plasmids and chromosomes raises the question of whether replicon size may explain the different pseudogene densities. To test the impact of replicon size differences, we constructed a dataset comprising chromosomes and plasmids with similar size distributions (Supplementary Fig. S1). Comparing the pseudogene density between plasmids and

97 chromosomes in that set showed that, also for plasmids and chromosomes of similar size,
98 plasmids have a significantly higher pseudogene density compared to chromosomes ($P < 0.001$,
99 using permutation test, $median_{plasmid}=0.06$, $median_{chromosome}=0.03$). Hence, the smaller plasmid
100 genome size cannot alone explain the high pseudogene density in plasmids.

101 To test if the high pseudogene density in plasmids is a general phenomenon, we
102 compared the pseudogene density between plasmids and chromosomes within different genera.
103 Our results show that the pseudogene density was consistently higher in plasmids compared to
104 chromosomes for all taxa, except *Shigella* ($P < 0.001$, using Wilcoxon test with FDR correction,
105 Supplementary Fig. S2). The high frequency of gene non-functionalization in *Shigella* has been
106 previously associated with adaptation of a host-associated lifestyle in this genus (11, 21). Hence,
107 the elevated occurrence of pseudogenes in plasmids is common to most taxonomic groups.

108 To study plasmid gene non-functionalization in detail, we focused on plasmids residing in
109 2,442 *Klebsiella*, *Escherichia*, and *Salmonella* (KES) isolates. These three enterobacterial taxa
110 have been extensively sampled for sequencing and plasmids in these taxa include diverse
111 plasmid types (22, 23). The density of pseudogenes in KES plasmids is significantly higher
112 compared to chromosomes ($P < 0.001$, using Wilcoxon test, Fig. 1A). This reflects an observed
113 4-fold increase in pseudogene density in plasmids compared to chromosomes and equates to
114 approximately 9 CDS per pseudogene in plasmids and 40 CDS per pseudogene in chromosomes.
115 Hence, enterobacterial genomes within the KES dataset capture the trend of increased
116 pseudogene density in plasmids compared to chromosomes that we observed within the
117 taxonomically diverse RefSeq dataset.

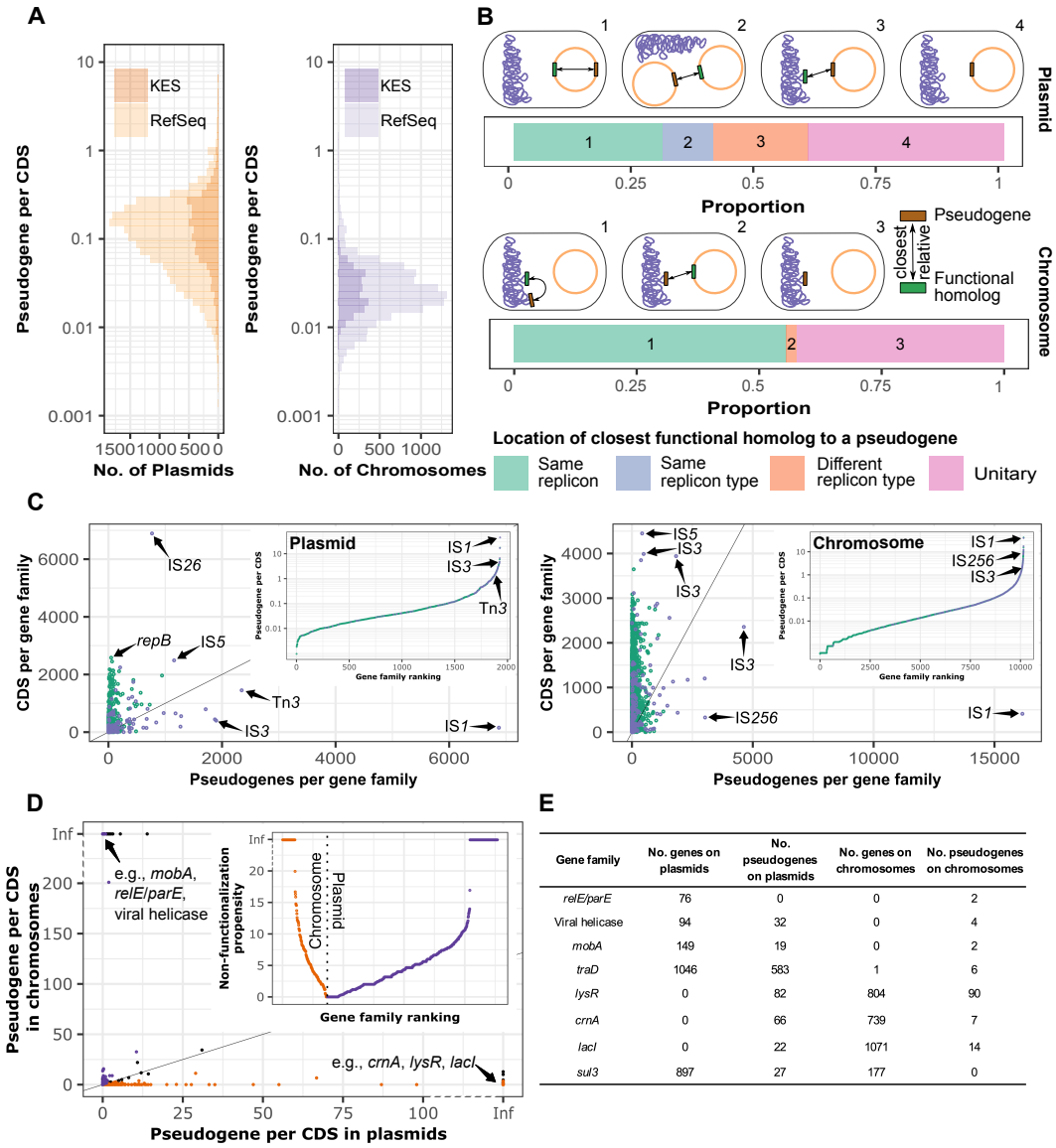
118 Pseudogenes having a functional homolog in the isolate genome, including the
119 chromosome and extra-chromosomal elements, may have evolved from gene non-
120 functionalization following a gene duplication (either within the same replicon, or via translocation
121 of duplicates between replicons). Pseudogenes that do not have a functional homolog within the
122 same isolate may have evolved either from non-functionalization of a single-copy gene, or
123 alternatively, non-functionalization of a horizontally acquired gene upon arrival. The comparison
124 of pseudogenes with their functional homologs showed that most pseudogenes are truncated
125 (480,349, 71%). Recent studies suggest that early stop codons in many truncated pseudogenes
126 correspond to nonsense mutations (24), or transient alleles that may be reversed under purifying
127 selection (25). Comparing the evolutionary rates of truncated and non-truncated pseudogenes we
128 find that truncated pseudogenes are characterized by higher dN and dS, as well as higher dN/dS
129 compared to non-truncated pseudogenes (Supplementary Figure S3), as is expected for DNA
130 evolution following a relaxation of selection pressure (i.e., gene non-functionalization).

131 Furthermore, the divergence of pseudogenes in plasmids and chromosomes compared to their
132 closest homologous gene had similar characteristics (Supplementary Fig. S3), hence the high
133 density of pseudogenes in plasmids cannot be explained by a bias in the genome annotation
134 pipeline.

135 Pseudogenes were further classified into two groups: unitary and non-unitary
136 pseudogenes, based on whether a functional homolog to the pseudogene was present within the
137 same isolate. Our results show that the proportion of pseudogenes that are accompanied by a
138 homologous gene on the same replicon is lower for plasmid pseudogenes (Fig. 1B). The
139 proportion of unitary pseudogenes at the isolate level is comparable between plasmids and
140 chromosomes. Plasmid-encoded pseudogenes may furthermore have a homologous gene
141 encoded either in the chromosome (19%) or in another plasmid (10%) within the same isolate.
142 Chromosome-encoded pseudogenes having a plasmid-encoded homologous gene are rather
143 rare (2%). Plasmid-encoded pseudogenes are thus more likely to have a homologous gene
144 encoded on a different replicon within the isolate genome compared to chromosome-encoded
145 pseudogenes. Accordingly, we conclude that gene non-functionalization in plasmids frequently
146 entails the loss of redundant functions at the level of the whole isolate genome.

147 **The majority of pseudogenes in plasmids correspond to deteriorated IS-elements**

148 Which genes are frequently non-functionalized? In plasmids, gene families comprising
149 transposition-related gene functions such as IS*1*, Tn*3*, and IS*3* are characterized by the highest
150 frequency of pseudogenes (Fig. 1C). Similarly, transposases such as those encoded by IS*1*,
151 IS*256*, and IS*3*, are frequently non-functionalized in chromosomes (Fig. 1C). Non-functional IS
152 elements accounted for a substantial portion of pseudogenes in our dataset with 88,107 (13%)
153 pseudogene instances. For example, the IS*1* family transposase alone comprised approximately
154 every 20th pseudogene (5.85%), with an average of 33 pseudogenes per gene. Furthermore,
155 pseudogenes in the IS*1* family transposases are the most abundant in the data and occur in 72%
156 of the chromosomes (Supplementary Fig. S4). Our results are thus in agreement with previous
157 reports on frequent non-functionalization of TEs in prokaryote genomes (6) and furthermore show
158 a similar trend in plasmid genomes.



Gene family	No. genes on plasmids	No. pseudogenes on plasmids	No. genes on chromosomes	No. pseudogenes on chromosomes
relE/parE	76	0	0	2
Viral helicase	94	32	0	4
mobA	149	19	0	2
traD	1046	583	1	6
lysR	0	82	804	90
crnA	0	66	739	7
lacI	0	22	1071	14
sul3	897	27	177	0

Figure 1. Evolution of gene non-functionalization in plasmids compared to chromosomes. **(A)** Distribution of pseudogene density (pseudogene per CDS) per plasmid and chromosome for the RefSeq and KES datasets. The graph includes plasmids and chromosomes with pseudogene content (RefSeq dataset: 21,564 (74%) plasmids and 10,832 (100%) chromosomes; KES dataset: 5,573 (82%) plasmids and 2,442 (100%) chromosomes. **(B)** Proportions of pseudogenes according to the location of the closest related homologous CDSs in plasmids and chromosomes. **(C)** Non-functional to functional ratio per gene family in plasmids and chromosomes. Each dot corresponds to a gene family (green). Transposase gene families are distinguished (purple). **(D)** Ratio of pseudogene per CDS in plasmids (x-axis) to pseudogene per CDS in chromosomes (y-axis) for 6,396 gene families that occur on both plasmids and chromosomes. Gene functions with a significant enrichment for pseudogenes are highlighted for plasmids (orange, n=840) and chromosomes (purple, n=221) ($P < 0.001$, using Fisher's exact test). Infinite (Inf) values correspond to gene families lacking pseudogenes (i.e., divisions by zero). **(E)** Example gene families that are enriched for pseudogenes. Names of listed gene families (C, D, E) have been validated with Pfam (26) and TnCentral (27) databases using sequence search with the most frequent sequence variant of the gene families.

176 The comparison between plasmids and chromosomes further reveals that the propensity for non-
177 functionalization may be different depending on the replicon type. We identified 221 (0.64%)
178 plasmid-encoded gene families that are significantly enriched for pseudogenes in the
179 chromosome (Fig. 1D, Supplementary Table 1). These families often correspond to plasmid-
180 encoded genes, such as mobilization genes (*mobA*), toxin-antitoxin systems (*relE*/*parE*), and viral
181 helicases (Fig. 1E). Additionally, we identified 840 (2.4%) chromosome-encoded gene families
182 that are significantly enriched for pseudogenes in plasmids (Fig. 1D, Supplementary Table 1).
183 These families include genes such as LysR-type transcriptional regulators (*lysR*), LacI-related
184 transcriptional regulators (*lacI*), and creatininase (*crnA*), that are specifically encoded on the
185 chromosome (Fig. 1E). We hypothesize that these gene families correspond to replicon-specific
186 genes that underwent non-functionalization upon translocation to the other replicon, rendering
187 them essentially dead-on-arrival. Barriers for gene transfer from chromosomes to plasmids have
188 been previously shown, for example due to dose effect (28). Our results suggest the presence of
189 similar barriers in gene transfer from plasmids to chromosomes.

190 **Frequent pseudogenes in large mobilizable and non-mobilizable plasmids**

191 Which plasmid types are most associated with a high frequency of pseudogenes? Initially we
192 focus on two prominent plasmid characteristics: plasmid size and mobility class. The small
193 plasmid types in the KES dataset mostly correspond to ColE1-like plasmids; accordingly, here we
194 consider plasmids of <19Kb as small (23). Our analysis shows that most of the large plasmid
195 types (4,659; 97%) and about half of the small plasmid types (1,045; 54%) contain pseudogenes.
196 The pseudogene density is highest in small non-mobilizable plasmids (small plasmid types with
197 pseudogenes: $\text{median}_{\text{small mobilizable}}=0.2$, $\text{median}_{\text{small non-mobilizable}}=0.33$) compared to large plasmid
198 types and chromosomes, but that property is tightly linked to their small size encoding only few
199 genes. Among the large plasmids, pseudogene density is highest in mobilizable plasmids followed
200 by non-mobilizable plasmids ($P < 0.05$, using Wilcoxon test with FDR correction, $\text{median}_{\text{large}}$
201 $_{\text{mobilizable}}=0.27$, $\text{median}_{\text{large non-mobilizable}}=0.15$, $\text{median}_{\text{large conjugative}}=0.1$, Fig. 2A). Thus, approximately
202 every fourth gene is non-functionalized in some mobilizable plasmids. Taken together, our results
203 point towards an association between increased pseudogene densities in large plasmids and a
204 lack of plasmid mobility.

205 Plasmids in the different plasmid size and mobility categories are typically different in their
206 composition. We therefor compared the propensity of gene non-functionalization among the
207 plasmid size and mobility groups. Only seven gene families were enriched for pseudogenes in
208 the small plasmid types including antibiotic resistance genes (Supplementary Fig. S5). A total of
209 488 gene families were enriched for pseudogenes in at least one of the large plasmid mobility

classes (Fig. 2B). The comparison between large plasmids in different mobility classes revealed a significant enrichment for pseudogenes in gene families encoding for plasmid transfer mechanisms including *traI*, *nikB*, and *traV* (Fig. 2B; Supplementary Table S3). Consequently, we hypothesized that the origin of these pseudogenes is gene non-functionalization following loss of self-transmissibility of conjugative plasmids.

To gain further understanding of gene non-functionalization events in the context of large plasmid types, we compared the pseudogene density among plasmids in different plasmid taxonomic units (PTUs) (29). PTUs that comprise a high pseudogene density typically include a high proportion of mobilizable and non-mobilizable plasmid types (Fig. 2C). Gene deletions and changes in genetic repertoire of plasmids may lead to a transition of plasmid mobility, which may lead to PTU divergence and the evolution of novel PTUs in the long-term (17, 19). Indeed, here we observe several PTUs where transfer-related gene families frequently occur as pseudogenes; these PTUs are highlighted as putative plasmid types where transitions in plasmid mobility may have occurred (red labeled PTUs in Fig. 2C). Transfer-related pseudogenes thus correspond to molecular fossils of lost mobility mechanisms in plasmid genomes.

Vertical inheritance of pseudogenes in evolutionary-related plasmid clusters

Pseudogenes observed within the same gene family may have originated either from multiple independent gene non-functionalization events, or alternatively, from few non-functionalization events in ancestral plasmids followed by vertical inheritance during plasmid diversification. If vertical inheritance of pseudogenes is common in plasmid evolution, then the number of observed pseudogenes supplies an overestimation for the frequency of gene non-functionalization events. To evaluate the role of vertical inheritance in pseudogene evolution, we compared shared gene and pseudogene content among closely related plasmids. For that purpose, we clustered all plasmids into clusters of homologous plasmids based on shared gene content. We obtained 353 clusters comprising 85% large plasmids and 206 clusters comprising 89% of the small plasmids; these clusters largely correspond to the PTUs (see Supplementary Fig. S6). A visual inspection of the shared genes matrices suggests that pseudogenes tend to be shared among plasmids that also share genes (i.e., clusters of closely related plasmid) (Fig. 3AB). Indeed, the proportion of shared genes and shared pseudogenes between plasmid pairs are positively correlated ($r_s = 0.47$, $P < 0.001$, using Spearman's correlation), suggesting frequent vertical inheritance (e.g., closely related plasmids and loss if plasmids are restructured (e.g., distantly related-plasmids) in the evolution of plasmid-encoded pseudogenes.

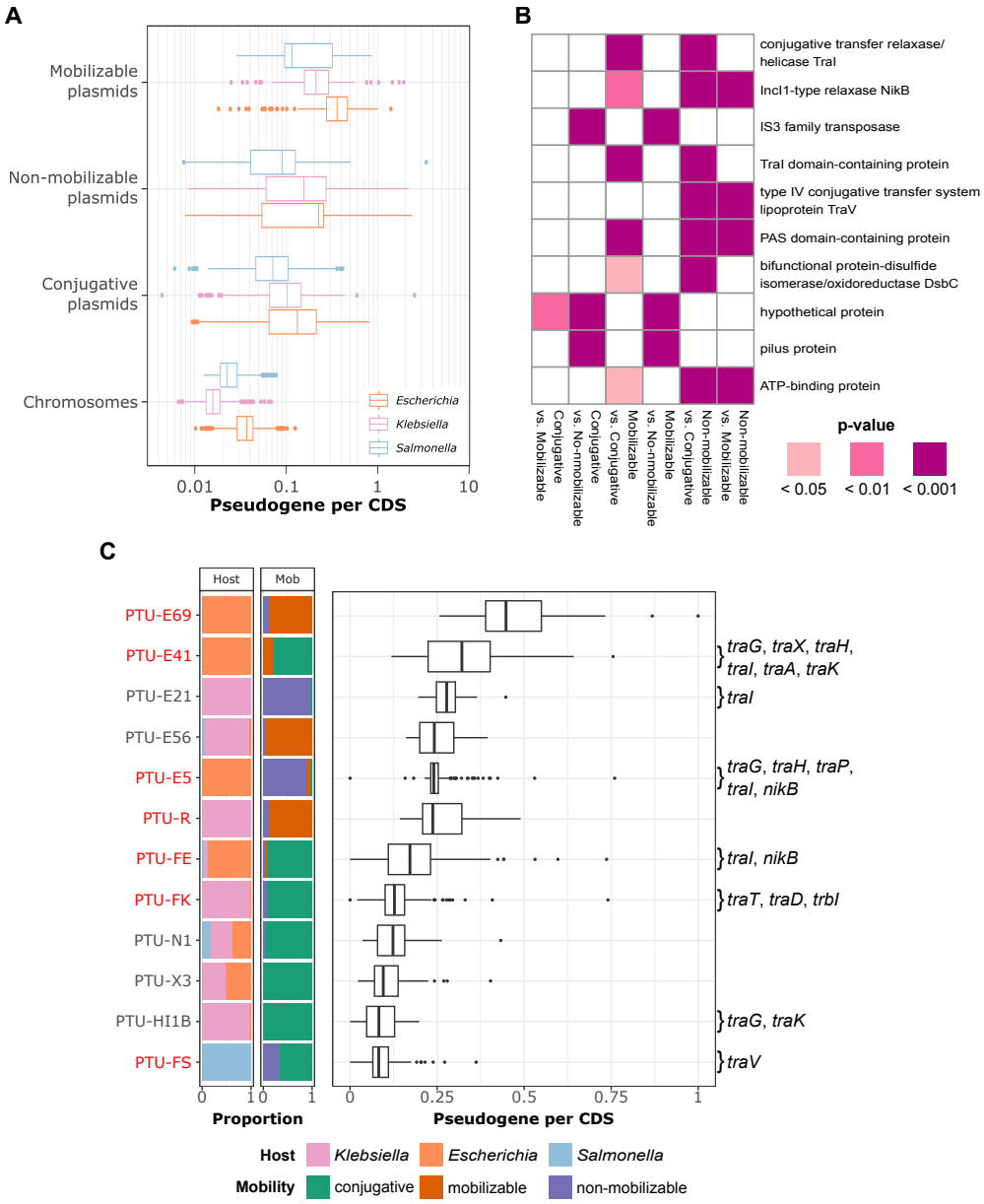


Figure 2. Pseudogene density and gene function in large (≥19Kb) plasmid types. (A) Ratio of pseudogene per CDS shown for large plasmid types (≥19Kb) and chromosomes (pseudogene-less replicons were excluded). **(B)** Enrichment for pseudogenes in gene families significantly depends on the plasmid mobility types in large plasmids ($P < 0.05$, using Fisher's exact test with FDR correction). First mentioned plasmid mobility type indicates in which type the enrichment of pseudogenes for gene functions has been statistically observed. The second mentioned plasmid mobility type indicates the type from which the frequencies of pseudogenes and CDSs were used for the statistical comparison. **(C)** Distribution of pseudogene per CDS ratio for plasmid taxonomic units (PTUs) of large plasmid type (only frequent ($n_{PTU} \geq 30$) PTUs are shown). Stacked bar plots show proportions of plasmid mobility type and the host taxonomy per PTU. Labels besides the boxplot report PTU-specific non-functionalized gene families corresponding to transfer-related (*tra*) functions ($P < 0.05$, using Fisher's exact test with FDR correction). PTUs highlighted in red are characterized by multiple plasmid mobility types (where the main plasmid mobility type < 95%).

The observation of shared pseudogenes in plasmids may be either due to retention of non-functional DNA or because non-functionalization occurs in a narrow range of gene family functions. In the letter case, even if these non-functionalization events emerged independently, it would generate a signal of shared pseudogenes that are not vertically inherited. The strongest patterns of shared pseudogenes in our dataset are mostly restricted to closely related plasmid clusters. A total of 874 (10.1%) gene families comprising pseudogenes are plasmid cluster-specific ($P < 0.05$, using Fisher's exact test with FDR correction). Hence, pseudogene sharing between plasmids clusters, is expected to correspond to diverse non-functional categories. A total of 356 (4.4%) gene families correspond to frequent, non-functional transposase-related functions such as IS*1*, Tn*3*, and IS*26*. Additionally, gene family size and the frequency of non cluster-specific pseudogenes are significantly correlated ($r_s = 0.58$, $P < 0.001$, using Spearman's correlation). Shared pseudogene content between plasmid clusters thus corresponds either to horizontal gene transfer or commonly non-functionalized gene families.

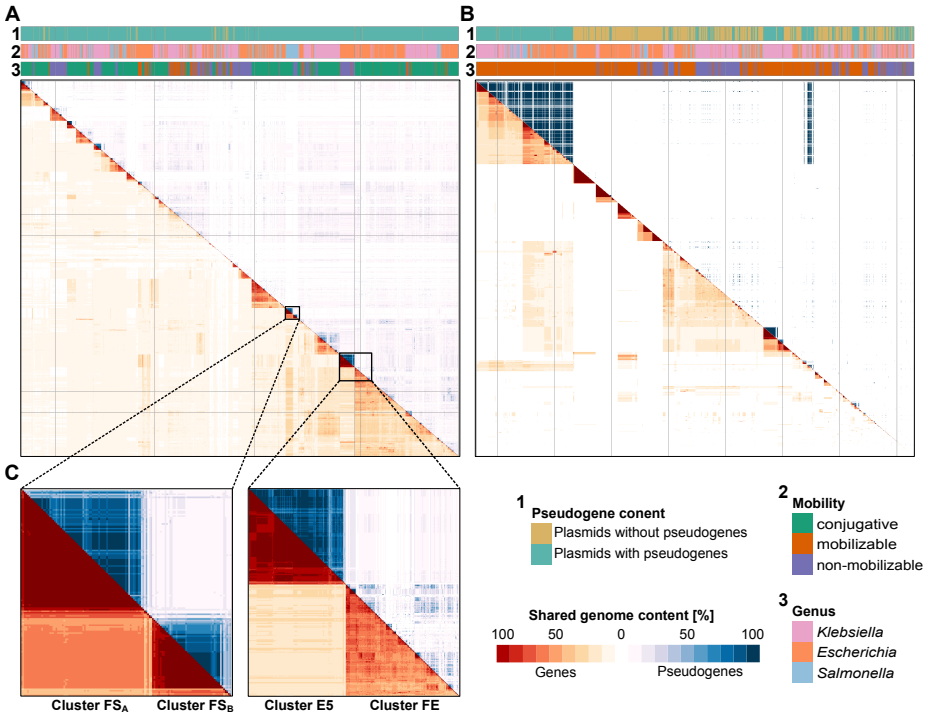


Figure 3. Shared gene and pseudogene content are similar in enterobacterial plasmids. (A, B) Pairwise shared gene (red) and pseudogene (blue) matrices of large plasmid types (A) and small plasmid types (B) among the KES dataset. The matrices were sorted using hierarchical clustering of shared gene content. **(C)** Shared genes and pseudogenes of adjacent plasmid clusters (FS_A, FS_B, FE, and E5) in the hierarchically clustered shared gene content matrix. Labeled plasmid clusters are based on the Markov Cluster Algorithm (MCL) using pairs of plasmids sharing ≥ 60% gene content (see methods).

279 **Non-functionalization following segmental deletion is a signature for the loss of self-**
280 **transmissibility**
281 To gain a better understanding on the evolutionary events at the basis of pseudogene vertical
282 inheritance, we reconstructed the evolution of closely related plasmid clusters. For that purpose,
283 we selected two clusters of *Salmonella* plasmids that share a high proportion of their genes and
284 correspond mostly to PTU-FS (Supplementary Table S4, Supplementary Fig. S6). The shared
285 gene content within the selected clusters FS_A and FS_B is 92% and 86%, respectively, with an
286 average of 50% shared genes among plasmids in both clusters (Fig. 3C). Plasmids of cluster FS_A
287 are homologous to plasmid pSLT (NC_003277.2) of *S. enterica* serovar Typhimurium and
288 plasmids of cluster FS_B are homologous to plasmid pSENV (NZ_CP063701.1) of *S. enterica*
289 subsp. *enterica* serovar Enteritidis. Cluster FS_A comprises only 83 conjugative plasmids and
290 cluster FS_B includes mostly (46; 92%) non-mobilizable plasmids, in addition to four conjugative
291 plasmids. The average shared pseudogene content among plasmids within the clusters is 67%
292 or 65% for clusters FS_A and FS_B, respectively. The shared pseudogene content among plasmids
293 from different clusters was only 9%, on average, thus plasmids in the two clusters are diverged in
294 their pseudogene content. The high proportion of shared gene (and pseudogene) content among
295 plasmids in the two PTU-FS clusters supports the notion that these plasmids are of common
296 ancestry, i.e., they are homologs, albeit with alterations in their mobility class.

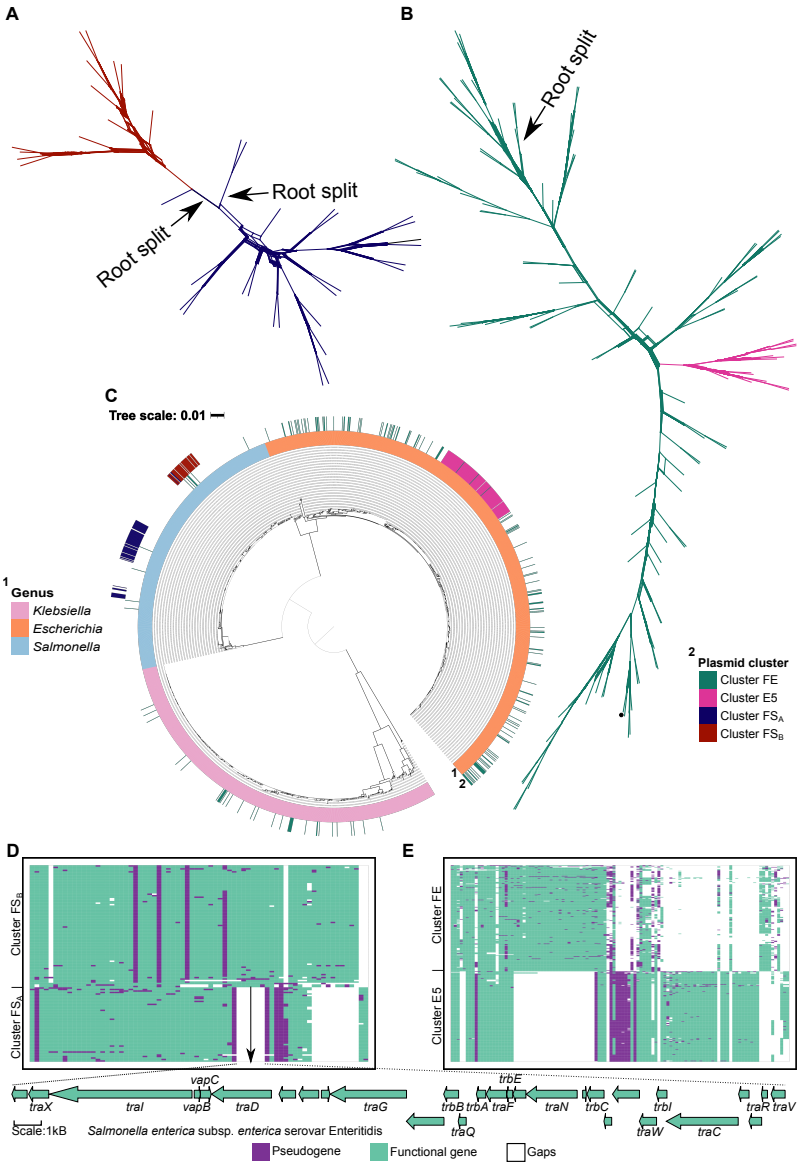


Figure 4. Phylogenetic relationship of self-transmissible and non-mobilizable plasmid clusters and their hosts. (A, B) Phylogenetic networks including 11 single-copy genes of plasmid clusters FS_A and FS_B (PTU-FS) and 10 single-copy genes of plasmid clusters FE (PTU-FE) and E5 (PTU-E5) (see Supplementary Table S7 & S8). Three plasmids (ca. 2%) and 40 plasmids (ca. 11%) were excluded from the phylogenetic reconstruction of (A) and (B), respectively, to increase the number of complete single-copy gene families and hence the robustness of the phylogenetic inference. Arrows indicate inferred root positions resulting from root neighborhood inference. Branches are colored according to plasmid assignment into plasmid clusters. The black dot in (B) labels a mobilizable plasmid that deviates from the main mobility class of cluster FE (PTU-FE, conjugative). **(C)** Host phylogenetic tree based on 133 chromosomal single-copy genes. The tree was rooted using midpoint rooting. **(D, E)** Presence Absence Pattern (PAP) of pseudogenes (purple) and functional genes (green) for plasmid clusters FS_A/FS_B (D) and plasmid clusters FE/E5 (E). The PAPs are sorted using hierarchical clustering on shared gene content. The arrow depicts the closest plasmid to cluster FS_B (non-mobilizable) of cluster FS_A (conjugative), which bears transfer-related genes that are absent in cluster FS_B (genomic arrangement shown below).

314 To examine the evolution of PTU-FS plasmids, we reconstructed the plasmid phylogeny based
315 on eleven single-copy gene families present in 122 (92%) plasmids. The phylogeny reveals a
316 clear split between plasmid clusters FS_A and FS_B (Fig. 4A). To reconstruct the ancestral mobility
317 state of PTU-FS plasmids, we inferred the root of PTU-FS using a phylogenomic rooting approach
318 (30). The phylogenetic inference revealed a root neighborhood in the deepest split between
319 plasmid clusters FS_A and FS_B that includes branches leading to several conjugating plasmids of
320 cluster FS_A (Fig. 4A). This root position implies that the ancestor of PTU-FS plasmids was
321 conjugative and the non-mobilizable plasmids in cluster FS_B are derived plasmids (Fig. 4A).
322 Notably, the plasmid genome size differs significantly between the non-mobilizable cluster FS_B
323 plasmids and the conjugative FS_A plasmids, with the non-mobilizable plasmids having a smaller
324 genome size ($P < 0.001$, using Wilcoxon test, $median_{non\text{-}mobilizable}$=59,372 bp,
325 $median_{conjugative}$=93,865 bp). Further comparison of the gene content among the PTU-FS reveals
326 a high conservation of plasmid gene order, except for the absence of transfer (*tra*) genes in the
327 non-mobilizable FS_B plasmids (Fig. 4D, Supplementary Fig. S10A).
328 The loss of the conjugation machinery genes may have followed two alternative scenarios:
329 gradual non-functionalization and gene loss, or a large-scale deletion followed by non-
330 functionalization of functionally-related genes. Since the absent genes are consecutively
331 arranged in the ancestral plasmid (Fig. 4D), we infer that a large segmental deletion of transfer-
332 related genes occurred in the ancestor of cluster FS_B non-mobilizable plasmids (Supplementary
333 Fig. S7). The transfer-related genes *traV* and *FinO* (a repressor of conjugative transfer in F-like
334 plasmids) correspond to the boundaries of the deleted transfer-related gene arrangement in the
335 ancestor and are found as pseudogenes in derived plasmids. Notably, the *Salmonella* host
336 phylogeny shows that plasmid clusters FS_A and FS_B reside in distinct groups of *Salmonella*
337 isolates where hosts of the conjugative FS_A plasmids appear as more distantly related in
338 comparison to hosts of the non-mobilizable FS_B plasmids (Fig. 4C, Supplementary Fig. S9). Taken
339 together, our results reveal a transition in plasmid mobility from self-transmissible to non-
340 mobilizable plasmids within *Salmonella* PTU-FS plasmids due to large segmental deletion
341 followed by gene non-functionalization events.
342 To examine further scenarios for gene non-functionalization in plasmid genomes, we
343 investigated two additional adjacent plasmid clusters in the shared gene matrix where the pattern
344 of gene and pseudogene sharing indicates that they are closely related (Fig. 3C). The plasmids
345 in these clusters correspond to the evolutionary related PTUs FE and E5 (19). Most (167; 87%)
346 of the plasmids in cluster FE were assigned to the conjugative PTU-FE, with the remaining
347 plasmids being unassigned (Supplementary Table S4, Supplementary Fig. S6). All 164 plasmids

348 in cluster E5 were classified into the non-mobilizable PTU-E5 (Supplementary Table S4,
349 Supplementary Fig. S6), which correspond to homologs of pO157 in *E. coli* strain O157:H7 and
350 are known to be non-transmissible (17, 31). Plasmids in cluster FE share, on average, 53% of
351 their genes and 12% of their pseudogenes. In comparison, gene sharing among plasmids in
352 cluster E5 shows a higher shared content with 90% of the genes and 80% of the pseudogenes
353 shared, on average. Plasmids in FE and E5 clusters share, on average, 18% gene content and
354 5% pseudogene content. The shared gene content among plasmids in the two clusters reflects
355 evolutionary relationships where the different plasmid mobility types may be explained by the
356 differential gene content.

357 To examine the evolutionary history of plasmids in clusters FE and E5, we reconstructed the
358 plasmid backbone phylogeny using ten single-copy gene families present in 316 (89%) plasmids.
359 The phylogeny revealed a clear split between the plasmid clusters (Fig. 4B). The topology of
360 cluster E5 plasmids reveal two main plasmid lineages of closely related plasmids. In contrast, the
361 phyletic pattern of cluster FE conjugative plasmids revealed a high divergence among the
362 plasmids (Fig. 4B). The inferred root position is located in a branch leading to two conjugative
363 cluster FE plasmids. Notably, two plasmids in cluster E5 encode a partial repertoire of the
364 conjugation machinery and they were accordingly predicted as conjugative; these plasmids
365 branch closely to the split with cluster FE plasmids (Supplementary Fig. S8) and their syntenic
366 gene order reveals differential loss of transfer-related genes (Supplementary Fig. S8). Comparing
367 the sequences of the conjugative plasmids of cluster E5 to their closest non-mobilizable neighbor
368 plasmid revealed evidence for loss of plasmid mobility genes and shared genomic arrangements
369 in the non-mobilizable cluster E5 plasmids (Fig. 4B, Supplementary Fig. S10B). The phylogenetic
370 reconstruction suggests that the ancestral plasmid of cluster E5 plasmids (or PTU-E5) was
371 conjugative and the plasmid mobility was lost due to a segmental deletion of the *tra* genes (Fig.
372 4B, Supplementary Fig. S8). Plasmids in cluster E5 share 80% of their pseudogenes, including
373 transposases such as IS*3* (508), IS*1* (336), and IS*91* (320), as well as transfer-related gene
374 families *traI* (164) and *nikB* (164) that are non-functionalized in 164 (98%) of the cluster E5
375 plasmids. Plasmids in cluster FE reside in a diverse set of hosts including *Klebsiella*, *Escherichia*,
376 and *Salmonella*. In contrary, plasmids in cluster E5 are only hosted by closely related *Escherichia*
377 isolates including strain O157:H7 (Fig. 4C); these plasmids are homologs of plasmid pO157 that
378 is non-mobilizable (e.g., NZ_CP017435.1). Note that the hosts of cluster E5 plasmids are
379 diverged, i.e., they do not correspond to isogenic strains (Supplementary Fig. S9). Hence the
380 topology of cluster E5 phylogeny likely corresponds to within-host evolution via vertical
381 inheritance. Notably, cluster E5 plasmids are characterized by significantly larger plasmid genome

382 size, but a smaller number of CDSs compared to the conjugative cluster FE plasmids, which is
383 well explained by the frequent pseudogenes in cluster E5 plasmids ($P < 0.001$, using Wilcoxon
384 test, $median_{Cluster\ E5}$=78 CDS, 19 pseudogenes, 93,179 bp; $median_{Cluster\ FE}$=86 CDS, 9
385 pseudogenes, 74,922.5 bp). Hence pseudogenes may be retained as non-coding DNA in the
386 evolution of the non-mobilizable plasmids after the loss of self-transmissibility. Our results thus
387 reveal that plasmid divergence events may give rise to new integral replicons within their host
388 lineage.

389 **Discussion**

390 Gene non-functionalization is considered a common event following gene duplication and gene
391 acquisition via lateral gene transfer. Previous studies suggested that bacterial genomes are often
392 devoid of pseudogenes due to purifying selection (8). This implies that pseudogenes in bacterial
393 genomes should be considered as 'garbage DNA', that is, non-functional DNA that may have an
394 effect on the organism fitness, rather than 'Junk DNA' (see (32) for definitions). Our results show
395 that pseudogenes are found in bacterial genomes and their density is highest in plasmids. We
396 identify two main processes in plasmid evolution that are accompanied by frequent gene non-
397 functionalization: proliferation of mobile genetic elements (MGEs) and reductive evolution of self-
398 transmissible plasmids.

399 Mobile genetic elements, including transposons and integrons, are known to translocate
400 or proliferate within and between genomes (33), thus, facilitating gene transfer, as well as
401 genomic rearrangements (reviewed in (34)). At the same time, the integration of MGEs in bacterial
402 genomes may lead to gene non-functionalization at the insertion locus, e.g., due to disruption of
403 an open reading frame (e.g., (35)). The high frequency of MGE pseudogenes in plasmids reported
404 here conforms previous studies highlighting the role of failed transposition events in the
405 generation of pseudogenes in bacterial genomes (e.g., (6)). We observed that most pseudogenes
406 are due to truncations and that pseudogenes in plasmids are more deteriorated. Indeed, site-
407 specific recombination of transposable elements generates genomic hotspots for their insertion,
408 where repeating insertion at the same locus may lead to gene truncations (reviewed in (36)). That
409 being said, recent studies suggest that many pseudogenes are transcribed, and sometime even
410 translated (37). Pseudogenes of MGE origin may thus correspond to proto-genes and a source
411 for genetic innovation during MGE diversification (25). This suggestion is in line with recent
412 studies that report pseudogene resurrection under strong selective conditions for the lost gene

413 function. Examples include the resurrection of iron uptake in *E. coli* (38) and the resurrection of
414 CO_2 assimilation in *Brucella* spp. (39).

415 Our analysis furthermore reveals that the propensity for specific gene non-
416 functionalization depends on the replicon type, i.e., plasmid or chromosome. This observation
417 may be explained by selection against the presence of essential (chromosomal) genes in
418 plasmids, e.g., due to unreliable plasmid inheritance (40) or deleterious effects of gene duplication
419 (28). Thus, the higher pseudogene density in plasmids compared to chromosomes may be
420 explained by the propensity of plasmid gene content to include accessory and dispensable genes.
421 We note however, that the classification of genes into core and accessory may vary depending
422 on the bacterial taxon, and that has an effect on the evolution of plasmid gene content in different
423 bacterial taxa (e.g., in the context of antibiotic resistance plasmids; (41)). We hypothesize that the
424 horizontal transfer (or translocation) of chromosomal genes into plasmids (and *vice versa*) is
425 bound to yield effectively 'dead-on-arrival' pseudogenes.

426 Inferring rooted plasmid phylogenies that comprise related non-mobilizable and
427 conjugative plasmid enabled us to infer the ancestral and derived plasmid variants in plasmid
428 evolution. According to our inference, non-functionalized copies of transfer-related genes in large
429 non-mobilizable and mobilizable plasmids are molecular fossils underlining a self-transmissible
430 plasmid origin. The loss of plasmid self-transmissibility may lead to plasmid 'domestication' in the
431 host lineage (or plasmid 'fixation' in the host pangenome). Domesticated plasmids are
432 characterized by a stable vertical inheritance within the lineage that can be considered as
433 ancestral states of chromids (42). Examples for the evolution of chromids from essential plasmids
434 are found in plant-associated bacteria such as *Sinorhizobium meliloti* harboring a chromid,
435 pSYMB, that encodes essential genes (43). Nonetheless, evolution of domesticated plasmids
436 following the acquisition of essential genes was so far not considered in *Enterobacteriaceae* (28,
437 40). Previous studies reported a stable inheritance of plasmids pO157 in *Escherichia* and pSLT
438 in *Salmonella* (44–46). Although, pSLT is self-transmissible, it was shown that most *Salmonella*
439 virulence plasmids are rather vertically inherited except for pSENV of *S. enterica* subsp. *enterica*
440 serovar Enteritidis (47). Plasmids of *S. enteritidis* bear incomplete *tra* regions (48) and virulence
441 plasmids of serovars Enteritidis and Typhimurium are not mobilizable by other F-like plasmids
442 even in the presence of an *oriT* region (49). The results of our analysis support the view that
443 pSENV plasmid homologs were acquired by horizontal transfer subsequent to a segmental
444 deletion of transfer-related gene (the evolution of serovar specific virulence plasmids reviewed in
445 (50)). Furthermore, the phylogeny of isolates hosting pO157- and pSENV-like plasmids indicates

446 that these plasmids were fixed in those lineages following the loss of self-transmissibility (Fig.
447 4C), hence they are better considered as domesticated plasmids (see also (51)).
448 The pseudogene repertoire and order in large non-mobilizable plasmids indicate that
449 segmental deletions within the conjugation machinery play a major role in the initial steps of
450 transitions in plasmid mobility. Such events may be caused, for example, by a transposon-
451 mediated segmental deletion, as observed in an experimental evolution study of plasmid pKP33
452 adaptation to a naïve *E. coli* host under selection for the plasmid-encoded antibiotic resistance
453 (18). Such a change in the plasmid genomic composition that leads to the loss of self-
454 transmissibility is likely to trigger a 'domino effect' of sequential non-functionalization events of
455 other transfer-related genes. Similar dynamics of gene non-functionalization have been
456 previously suggested for the evolution of pseudogene content in obligatory symbionts (52). The
457 loss of self-transmissibility may be deleterious for plasmids due to the inevitable the restriction in
458 the host range. Nonetheless, plasmids encoding a trait that is essential for the host will be
459 maintained in the population, also if their inheritance is unstable due to selection acting at the
460 level of host (e.g., (53)). Overtime, the co-adaptation of the plasmid and host can lead to the
461 evolution of stable plasmid inheritance (18, 54, 55). Notably, the genomes of facultative
462 pathogens are also characterized by a high frequency of MGEs and pseudogenes that likely play
463 a role in the evolution of genome miniaturization (13). The effect of selection at multiple levels
464 during plasmid evolution is thus akin to the evolution of obligatory symbionts and the mechanisms
465 involved in these processes – transposable element proliferation and gradual gene non-
466 functionalization – are likely similar.

467 **Materials & Methods**

468 **Data**

469 Genomes of isolates comprising plasmids were downloaded from the NCBI RefSeq database
470 (version 8/2022). Genome assemblies containing replicons labeled other than chromosomes or
471 plasmids were excluded from the analysis. Only genomes with a single chromosome were
472 retained. Contigs suspected to be erroneously annotated chromosomes were discarded
473 according to their size using a threshold of <550 kb. Here we used *Mycoplasma genitalium*
474 (chromosome size ca. 580 kb) as reference for the minimal chromosome size (56). The curated
475 data consisted 10,832 genome assemblies from 738 prokaryotic genera (Supplementary Table
476 S5). For detailed genome analysis, enterobacterial genomes from the NCBI RefSeq database
477 (KES dataset; version 01/2021) were used (as previously reported (23)). The KES dataset
478 includes 1,115 chromosomes and 3,098 plasmids from *Escherichia*; 755 chromosomes and 2,693

479 plasmids from *Klebsiella*; 572 chromosomes and 993 plasmids from *Salmonella* (Supplementary
480 Table S6).

481 **Construction of gene and pseudogene families**

482 Protein coding genes in the KES genomes set were clustered into gene families previously
483 described in (23). Briefly, reciprocal best hits (RBHs) of protein sequences between all replicon
484 pairs were identified using MMseqs2 ((57), v.13.45111, with module easy-rbh applying a threshold
485 of E-value ≤ 1×10^{-10}). RBHs were further compared by global alignment using parasail-python
486 ((58), v. 1.2.4, with the Needleman-Wunsch algorithm). Sequence pairs with ≥30% identical amino
487 acids were clustered into protein families using a high-performance parallel implementation of the
488 Markov clustering algorithm ((59), HipMCL with parameter --abc -I 2.0). Pseudogene nucleotide
489 sequences were extracted from genome assemblies using their genomic coordinates annotated
490 by the Prokaryotic Genome Annotation Pipeline (20). Proteins and pseudogenes of the KES
491 dataset have been aligned using MMseqs2 ((57), v.13.45111, with module search applying a
492 threshold of E-value ≤ 1×10^{-10}). Afterwards, pseudogenes have been assigned to a gene family,
493 if the gene family revealed the highest number of significantly aligned proteins to the pseudogene.
494 The final KES dataset encompassed 34,416 gene families consisting of 17,182,509 protein coding
495 genes and 658,133 (97%) pseudogenes in 13,375 (38%) gene families. Specific gene families
496 discussed in detail in the text were validated using sequence searches of TnCentral (27) and
497 Pfam (26) with the most frequent sequence variant of the gene families. Pseudogenes in one
498 relaxase/nuclease domain-containing gene family were suspected as erroneously classified as
499 such, and were excluded from the analysis of ColE1-like plasmids. The pseudogene had a highly
500 conserved synteny within mobilization related functions such as *mobC*, *mbeB,* and *mbeD*. In
501 addition, ORFs homologous to *mbeA* proteins were identified at the same genomic location
502 supporting the classification of that locus as a gene rather than pseudogene.

503 **Inference of dN and dS distances**

504 As a first step, the closest gene sequence to the pseudogene was identified. For that, pseudogene
505 nucleotide sequences were translated into amino acid sequences using transeq from EMBOSS
506 ((60, 61), v.6.6.0.0, with parameters -frame 6 and -table 11). Translated pseudogenes were
507 searched against the KES protein sequences using blastp ((62), v. 2.14.0+). The best hit was
508 used for further comparison. A pairwise alignment between the translated pseudogene and the
509 best matching amino acid sequence was inferred using clustalo ((63), v.1.2.4, --outfmt=clu).
510 Codon alignment was calculated using pal2nal ((64), v.14.1, with parameters -output paml and -
511 codontable 11). The dN and dS distances were calculated using codeml from PAML ((65), v.4.9,

512 with runmode = -2). Additionally, the maximum truncation size of pseudogenes compared to their
513 homologous genes was determined by calculating the leading and trailing truncations between
514 the pseudogene and the best matching homologous gene.

515 **Inference of unitary and non-unitary pseudogenes**

516 Homologous genes of pseudogenes in the KES dataset have been determined using MMseqs2
517 ((57), v.13.45111, with module search applying thresholds of E-value≤ 1×10^{-10} and sequence
518 identity ≥ 30%). Thereby, four scenarios (categories) of the homologous gene locus were
519 examined in the following hierarchical order: (I) on the same replicon, (II) on the same replicon
520 type, (III) on a different replicon type, and (IV) unitary pseudogene (see illustration of Fig. 1B). In
521 detail, in the example of plasmid pseudogenes, this analysis determined homologous protein-
522 coding sequences from the same plasmid, other plasmids of the same isolate, or the chromosome
523 of the same isolate. Following this procedure, the locations of closest related homologous coding-
524 sequences of pseudogenes were identified, which classified pseudogenes. If no homolog was
525 found in the same bacterial isolate, the pseudogene was classified as unitary. The same
526 procedure was applied for chromosomal pseudogenes identifying homologous protein-coding
527 sequences on the same chromosome or plasmids within the same bacterial isolate.

528 **Plasmid typing and clustering of homologous plasmids**

529 Plasmid mobility type was identified using MOB-suite ((15), v.3.0.3). Plasmids were assigned to
530 plasmid taxonomic units (PTUs) using COPLA supplying taxonomic information of residing hosts
531 ((29), see Supplementary Table S4). Additionally, plasmids were clustered into clusters of
532 homologous plasmids based on shared gene content. Pairwise shared gene and pseudogene
533 content of plasmids was calculated with the Jaccard similarity index from the gene and
534 pseudogene families. The shared gene and pseudogene content was visualized with hierarchical
535 clustering using pheatmap (v.1.0.12). Plasmid pairs sharing ≥ 60% of their genes were filtered to
536 construct plasmid clusters using the mcl algorithm ((66), v.12-135, parameter --abc -I 2.0).

537 **Phylogenetic analysis**

538 Plasmid and host phylogenies were constructed from complete single copy gene families for the
539 analyzed set of plasmids or chromosomes. Sequences were aligned using MAFFT ((67), v.7.520,
540 with parameter --maxiterate 1000 --localpair). Host phylogeny using 133 universal chromosomal
541 genes was reconstructed using IQ-TREE 2 ((68), v.2.2.2.7) with Le and Cascuel (LG) model
542 based on amino acid sequences (with parameter -m LG -B 1000, (69)). The host phylogeny was
543 rooted at midpoint using iTOL ((70), v.6.8). The reconstruction of the phylogenetic trees for
544 plasmids was based on nucleotide sequences of complete single-copy genes (see

Supplementary Table S7 & S8) applying IQ-TREE 2 ((68), v.2.2.2.7) with a transition model (with parameter -m TIM -B 1000, (69)). A rooting position for the plasmid trees was inferred by applying phylogenomic rooting using nucleotide sequences from plasmid cluster gene families ((30), Supplementary Table S7 & S8). Obtained phylogenetic trees were visualized and annotated using iTOL ((70), v.6.8) or SplitsTree4 ((71), v.4.17.0).

Statistics

All statistics were performed using R (v.4.0.3).

Data accessibility

The data are provided in electronic supplementary material.

Author contributions

D.M.H. and T.D. conceived the study. D.M.H. designed and performed the data analysis and visualizations. Y.W. prepared and provided complementary data. D.M.H. and T.D. interpreted the results and wrote the manuscript with comments and additions from Y.W.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the German Science Foundation (RTG2501 (TransEvo), grant number: 456882089), the Leibniz Science Campus EvoLUNG, the European Research Council (pMolEvol, grant gumber: 101043835), and the China Scholarship Council (CSC scholarship to Y.W.).

Acknowledgements

We thank Nils Hülter, Devani Picazo, Fabian Nies, Lisa Hartmann, Johannes Effe, Ishan Bhatt, and Mario Santer for critical comments on the manuscript. This research was supported in part through high-performance computing resources available at the Kiel University Computing Centre.

References

1. Zhao,H., Yang,J.-R., Xu,H. and Zhang,J. (2010) Pseudogenization of the umami taste receptor gene *Tas1r1* in the giant panda coincided with its dietary switch to bamboo. *Mol. Biol. Evol.*, **27**, 2669–2673.

2. Ohno,S. (1972) An argument for the genetic simplicity of man and other mammals. *J. Hum. Evol.*, **1**, 651–662.

576 3. Treangen,T.J. and Rocha,E.P.C. (2011) Horizontal transfer, not duplication, drives the
577 expansion of protein families in prokaryotes. *PLoS Genet.*, **7**, e1001284.

578 4. Tria,F.D.K. and Martin,W.F. (2021) Gene duplications are at least 50 times less frequent than
579 gene transfers in prokaryotic genomes. *Genome Biol. Evol.*, **13**, evab224.

580 5. van Passel,M.W.J., Marri,P.R. and Ochman,H. (2008) The emergence and fate of horizontally
581 acquired genes in *Escherichia coli*. *PLoS Comput. Biol.*, **4**, e1000059.

582 6. Liu,Y., Harrison,P.M., Kunin,V. and Gerstein,M. (2004) Comprehensive analysis of
583 pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally
584 transferred genes. *Genome Biol.*, **5**, R64.

585 7. Li,W.-H., Gojobori,T. and Nei,M. (1981) Pseudogenes as a paradigm of neutral evolution.
586 *Nature*, **292**, 237–239.

587 8. Lawrence,J.G., Hendrix,R.W. and Casjens,S. (2001) Where are the pseudogenes in bacterial
588 genomes? *Trends in Microbiol.*, **9**, 535–540.

589 9. Holt,K.E., Thomson,N.R., Wain,J., Langridge,G.C., Hasan,R., Bhutta,Z.A., Quail,M.A.,
590 Norbertczak,H., Walker,D., Simmonds,M., *et al.* (2009) Pseudogene accumulation in the
591 evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi. *BMC*
592 *Genom.*, **10**, 36.

593 10. Kuo,C.-H. and Ochman,H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS*
594 *Genet.*, **6**, e1001050.

595 11. Jin,Q. (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through
596 comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**,
597 4432–4441.

598 12. Moran,N.A. and Plague,G.R. (2004) Genomic changes following host restriction in bacteria.
599 *Curr. Opin. Genet. Dev.*, **14**, 627–633.

600 13. Ochman,H. and Davalos,L.M. (2006) The nature and dynamics of bacterial genomes.
601 *Science*, **311**, 1730–1733.

602 14. Rodríguez-Beltrán,J., DelaFuente,J., León-Sampedro,R., MacLean,R.C. and San Millán,Á.
603 (2021) Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat.*
604 *Rev. Microbiol.*, **19**, 347–359.

605 15. Robertson,J. and Nash,J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction
606 and typing of plasmids from draft assemblies. *Microb. Genom.*, **4**.

607 16. Ramsay,J.P. and Firth,N. (2017) Diverse mobilization strategies facilitate transfer of non-
608 conjugative mobile genetic elements. *Curr. Opin. Microbiol.*, **38**, 1–9.

609 17. Ares-Arroyo,M., Coluzzi,C. and Rocha,E.P.C. (2023) Origins of transfer establish networks of
610 functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Res.*, **51**,
611 3001–3016.

612 18. Porse,A., Schønning,K., Munck,C. and Sommer,M.O.A. (2016) Survival and evolution of a
613 large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Biol. Evol.*, **33**,
614 2860–2873.

615 19. Coluzzi,C., Garcillán-Barcia,M.P., de la Cruz,F. and Rocha,E.P.C. (2022) Evolution of plasmid
616 mobility: origin and fate of conjugative and nonconjugative plasmids. *Mol. Biol. Evol.*, **39**,
617 msac115.

618 20. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G.,
619 Chitsaz,F., Derbyshire,M.K., Durkin,A.S., *et al.* (2021) RefSeq: expanding the Prokaryotic
620 Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*,
621 **49**, D1020–D1028.

622 21. Feng,Y., Chen,Z. and Liu,S.-L. (2011) Gene decay in *Shigella* as an incipient stage of host-
623 adaptation. *PLoS One*, **6**, e27754.

624 22. Redondo-Salvo,S., Fernández-López,R., Ruiz,R., Vielva,L., De Toro,M., Rocha,E.P.C.,
625 Garcillán-Barcia,M.P. and De La Cruz,F. (2020) Pathways for horizontal gene transfer in
626 bacteria revealed by a global map of their plasmids. *Nat. Commun.*, **11**, 3602.

627 23. Wang,Y. and Dagan,T. (2023) Interaction between mobile genetic elements and plasmids
628 drives the evolution of multidrug resistant bacteria. *In Revision*.

629 24. Belinky,F., Ganguly,I., Poliakov,E., Yurchenko,V. and Rogozin,I.B. (2021) Analysis of stop
630 codons within prokaryotic protein-coding genes suggests frequent readthrough events.
631 *Int. J. Mol. Sci.*, **22**, 1876.

632 25. Feng,Y., Wang,Z., Chien,K.-Y., Chen,H.-L., Liang,Y.-H., Hua,X. and Chiu,C.-H. (2022)
633 "Pseudo-pseudogenes" in bacterial genomes: proteogenomics reveals a wide but low
634 protein expression of pseudogenes in *Salmonella enterica*. *Nucleic Acids Res.*, **50**, 5158–
635 5170.

636 26. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L.,
637 Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: The protein
638 families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

639 27. Ross,K., Varani,A.M., Snesrud,E., Huang,H., Alvarenga,D.O., Zhang,J., Wu,C., McGann,P.
640 and Chandler,M. (2021) TnCentral: a prokaryotic transposable element database and web
641 portal for transposon analysis. *Mbio*, **12**, e02060-21.

642 28. Wein,T., Wang,Y., Barz,M., Stücker,F.T., Hammerschmidt,K. and Dagan,T. (2021) Essential
643 gene acquisition destabilizes plasmid inheritance. *PLoS Genet.*, **17**, e1009656.

644 29. Redondo-Salvo,S., Bartomeus-Peñalver,R., Vielva,L., Tagg,K.A., Webb,H.E., Fernández-
645 López,R. and De La Cruz,F. (2021) COPLA, a taxonomic classifier of plasmids. *BMC*
646 *Bioinform.*, **22**, 390.

647 30. Tria,F.D.K., Landan,G., Picazo,D.R. and Dagan,T. (2023) Phylogenomic testing of root
648 hypotheses. *Genome Biol. Evol.*, **15**, evad096.

649 31. Lim,J.Y., Sheng,H., Seo,K.S., Park,Y.H. and Hovde,C.J. (2007) Characterization of an
650 *Escherichia coli* O157:H7 plasmid O157 deletion mutant and its survival and persistence
651 in cattle. *Appl. Environ. Microbiol.*, **73**, 2037–2047.

652 32. Graur,D., Zheng,Y. and Azevedo,R.B.R. (2015) An evolutionary classification of genomic
653 function. *Genome Biol. Evol.*, **7**, 642–645.

654 33. Plague,G.R., Dunbar,H.E., Tran,P.L. and Moran,N.A. (2008) Extensive proliferation of
655 transposable elements in heritable bacterial symbionts. *J. Bacteriol.*, **190**, 777–779.

656 34. Gray,Y.H.M. (2000) It takes two transposons to tango:transposable-element-mediated
657 chromosomal rearrangements. *Trends Genet.*, **16**, 461–468.

658 35. Belda,E., Moya,A., Bentley,S. and Silva,F.J. (2010) Mobile genetic element proliferation and
659 gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis*
660 *glossinidius*, the secondary endosymbiont of tsetse flies. *BMC Genom.*, **11**, 449.

661 36. Partridge,S.R., Kwong,S.M., Firth,N. and Jensen,S.O. (2018) Mobile genetic elements
662 associated with antimicrobial resistance. *Clin. Microbiol. Rev.*, **31**, e00088-17.

663 37. Soler-Camargo,N.C., Silva-Pereira,T.T., Zimpel,C.K., Camacho,M.F., Zelanis,A., Aono,A.H.,
664 Patané,J.S., dos Santos,A.P. and Guimarães,A.M.S. (2022) The rate and role of
665 pseudogenes of the *Mycobacterium tuberculosis* complex. *Microb. Genom.*, **8**.

666 38. Anand,A., Olson,C.A., Yang,L., Sastry,A.V., Catoiu,E., Choudhary,K.S., Phaneuf,P.V.,
667 Sandberg,T.E., Xu,S., Hefner,Y., *et al.* (2019) Pseudogene repair driven by selection
668 pressure applied in experimental evolution. *Nat. Microbiol.*, **4**, 386–389.

669 39. Varesio,L.M., Willett,J.W., Fiebig,A. and Crosson,S. (2019) A carbonic anhydrase
670 pseudogene sensitizes select *Brucella* lineages to low CO_2 tension. *J. Bacteriol.*, **201**.

671 40. Tazzyman,S.J. and Bonhoeffer,S. (2014) Why there are no essential genes on plasmids. *Mol.*
672 *Biol. Evol.*, **32**, msu293.

673 41. Wang,Y., Batra,A., Schulenburg,H. and Dagan,T. (2022) Gene sharing among plasmids and
674 chromosomes reveals barriers for antibiotic resistance gene transfer. *Phil. Trans. R. Soc.*
675 *B*, **377**, 20200467.

676 42. Vial,L. and Hommais,F. (2020) Plasmid-chromosome cross-talks. *Environ. Microbiol.*, **22**,
677 540–556.

678 43. Galibert,F., Finan,T.M., Long,S.R., Pühler,A., Abola,P., Ampe,F., Barloy-Hubler,F.,
679 Barnett,M.J., Becker,A., Boistard,P., *et al.* (2001) The composite genome of the legume
680 symbiont *Sinorhizobium meliloti*. *Science*, **293**, 668–672.

681 44. Makino,K. (1998) Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an
682 enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res.*, **5**,
683 1–9.

684 45. Lim,J.Y., Yoon,J. and Hovde,C.J. (2010) A brief overview of *Escherichia coli* O157:H7 and its
685 plasmid O157. *J. Microbiol. Biotechnol.*, **20**, 5–14.

686 46. Lobato-Márquez,D., Molina-García,L., Moreno-Córdoba,I., García-del Portillo,F. and Díaz-
687 Orejas,R. (2016) Stabilization of the virulence plasmid pSLT of *Salmonella* Typhimurium
688 by three maintenance systems and its evaluation by using a new stability test. *Front. Mol.*
689 *Biosci.*, **3**.

690 47. Feng,Y., Liu,J., Li,Y.-G., Cao,F.-L., Johnston,R.N., Zhou,J., Liu,G.-R. and Liu,S.-L. (2012)
691 Inheritance of the *Salmonella* virulence plasmids: mostly vertical and rarely horizontal.
692 *Infect. Genet. Evol.*, **12**, 1058–1063.

693 48. Rotger,R. and Casadesús,J. (1999) The virulence plasmids of *Salmonella*. *Int. Microbiol.*, **2**,
694 177–184.

695 49. Ou,J.T., Lin,M.-Y. and Chao,H.-L. (1994) Presence of F-like OriT base-pair sequence on the
696 virulence plasmids of *Salmonella* serovars Gallinarum, Enteritidis, and Typhimurium, but
697 absent in those of Choleraesuis and Dublin. *Microb. Pathog.*, **17**, 13–21.

698 50. Rychlik,I., Gregorova,D. and Hradecka,H. (2006) Distribution and function of plasmids in
699 *Salmonella enterica*. *Vet. Microbiol.*, **112**, 1–10.

700 51. Mebrhatu,M.T., Cenens,W. and Aertsen,A. (2014) An overview of the domestication and
701 impact of the *Salmonella* mobilome. *Crit. Rev. Microbiol.*, **40**, 63–75.

702 52. Dagan,T., Blekhman,R. and Graur,D. (2006) The "Domino Theory" of gene death: gradual
703 and mass gene extinction events in three lineages of obligate symbiotic bacterial
704 pathogens. *Mol. Biol. Evol.*, **23**, 310–316.

705 53. Wein,T., Wang,Y., Hülter,N.F., Hammerschmidt,K. and Dagan,T. (2020) Antibiotics interfere
706 with the evolution of plasmid stability. *Curr. Biol.*, **30**, 3841-3847.e4.

707 54. Harrison,E., Guymer,D., Spiers,A.J., Paterson,S. and Brockhurst,M.A. (2015) Parallel
708 compensatory evolution stabilizes plasmids across the parasitism-mutualism continuum.
709 *Curr. Biol.*, **25**, 2034–2039.

710 55. San Millan,A., Escudero,J.A., Gifford,D.R., Mazel,D. and MacLean,R.C. (2016) Multicopy
711 plasmids potentiate the evolution of antibiotic resistance in bacteria. *Nat. Ecol. Evol.*, **1**,
712 0010.

713 56. Glass,J.I., Assad-Garcia,N., Alperovich,N., Yooseph,S., Lewis,M.R., Maruf,M.,
714 Hutchison,C.A., Smith,H.O. and Venter,J.C. (2006) Essential genes of a minimal
715 bacterium. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 425–430.

716 57. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching
717 for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

718 58. Daily,J. (2016) Parasail: SIMD C library for global, semi-global, and local pairwise sequence
719 alignments. *BMC Bioinform.*, **17**, 81.

720 59. Azad,A., Pavlopoulos,G.A., Ouzounis,C.A., Kyrpides,N.C. and Buluç,A. (2018) HipMCL: a
721 high-performance parallel implementation of the Markov clustering algorithm for large-
722 scale networks. *Nucleic Acids Res.*, **46**, e33–e33.

723 60. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open
724 Software Suite. *Trends Genet.*, **16**, 276–277.

725 61. Goujon,M., McWilliam,H., Li,W., Valentin,F., Squizzato,S., Paern,J. and Lopez,R. (2010) A
726 new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, **38**, W695–
727 W699.

728 62. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L.
729 (2009) BLAST+: architecture and applications. *BMC Bioinform.*, **10**, 421.

730 63. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H.,
731 Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein
732 multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

733 64. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: robust conversion of protein sequence
734 alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–
735 W612.

736 65. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**,
737 1586–1591.

738 66. Van Dongen,S. (2008) Graph clustering via a discrete uncoupling process. *SIAM J. Matrix*
739 *Anal. Appl.*, **30**, 121–141.

740 67. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7:
741 improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

742 68. Minh,B.Q., Schmidt,H.A., Chernomor,O., Schrempf,D., Woodhams,M.D., Von Haeseler,A.
743 and Lanfear,R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic
744 inference in the genomic era. *Mol. Biol. Evol.*, **37**, 1530–1534.

745 69. Hoang,D.T., Chernomor,O., Von Haeseler,A., Minh,B.Q. and Vinh,L.S. (2018) UFBoot2:
746 improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.*, **35**, 518–522.

747 70. Letunic,I. and Bork,P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic
748 tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.

749 71. Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies.
750 *Mol. Biol. Evol.*, **23**, 254–267.

751