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 symbionts). 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 reveals 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 recipient 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 thus 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 that include the genomes of *Mycobacterium leprea*, *Rickettsia prowazekii* (8), as well as *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 yet 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 for the plasmid-encoded trait 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 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 2,441 *Enterobacteriaceae* isolates including the genera *Klebsiella*, *Escherichia*, and *Salmonella*. Comparing closely related plasmids (i.e., homologous plasmids), we further reconstruct the origin of pseudogenes in the context of transitions in plasmid mobility.

Materials & Methods

Genomes data and gene families

Genomes of isolates comprising plasmids were downloaded from the NCBI RefSeq database (version 8/2022). Genome assemblies containing replicons labeled other than chromosomes or plasmids were excluded from the analysis. Only genomes with a single chromosome were retained. To avoid potential inclusion of experimentally miniaturized or genomes of low quality, we excluded genomes with chromosome size below the norm of free-living bacterial organisms (<550 kb). This threshold was determined according to the genome size of *Mycoplasma genitalium* (chromosome size ca. 580 kb) that we considered as reference for the minimal chromosome size of a free-living prokaryote species (20). The curated data consisted of 10,832 genome assemblies from 738 prokaryotic genera (Supplementary Table S1). For a detailed analysis of pseudogenes in plasmids, we used the previously established KES dataset of enterobacterial genomes (21). The KES dataset includes 1,114 chromosomes and 3,098 plasmids from *Escherichia*; 755 chromosomes and 2,693 plasmids from *Klebsiella*; 572 chromosomes and 993 plasmids from *Salmonella* (all from RefSeq version 01/2021;

97 Supplementary Table S2). Protein coding genes in the KES genomes set are clustered into
98 32,623 gene families as previously described in (21). Briefly, reciprocal best hits (RBHs) of protein
99 sequences between all replicon pairs were identified using MMseqs2 (22) (v.13.45111, with
100 module easy-rbh applying a threshold of E-value ≤ 1×10^{-10}). RBHs were further compared by
101 global alignment using parasail-python (23) (v. 1.2.4, with the Needleman-Wunsch algorithm).
102 Sequence pairs with ≥30% identical amino acids were clustered into gene families using a high-
103 performance parallel implementation of the Markov clustering algorithm (24) (HipMCL with
104 parameter --abc -I 2.0).

105 **Retrieval of pseudogenes**

106 Pseudogenes were extracted from the RefSeq genomes according to their annotation and
107 coordinates in the RefSeq genome feature tables. Pseudogenes in RefSeq genomes are
108 annotated by the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (25). Briefly, PGAP
109 utilizes sequence similarity-based approaches and *ab initio* gene prediction to annotate genomes
110 (26). Pseudogene annotation is part of the PGAP pipeline for the annotation of protein-coding
111 genes. Genomic loci that have sequence similarity to known genes, yet are lacking evidence of
112 other gene features (e.g., complete open reading frame), are annotated as pseudogenes.

113 The quality of PGAP pseudogene annotation in KES dataset was further validated using
114 Pseudofinder (27) (v.1.1.0, with module annotate and parameters -it 0.95 -s 0.8 -e 1e-9). The
115 output of Pseudofinder was filtered to include only pseudogenes identified in intergenic regions.
116 That is, pseudogenes annotated by Pseudofinder in the same locus as functional genes in the
117 PGAP annotation were excluded. All of the PGAP annotated pseudogenes could be validated
118 using Pseudofinder (see Supplementary text, Supplementary Fig. S1 & S2). Pseudogenes
119 identified solely by Pseudofinder were excluded.

120 All KES pseudogenes were furthermore tested for programmed frameshifts (as described,
121 e.g., for bacterial insertion sequences (ISs) IS*1*, IS*150*, and IS*911* (28)). Therefore, we searched
122 for each pseudogene for a similar, large complete protein sequence in the KES dataset using
123 MMseqs2 (22) (v.13.45111, with module search and parameters: --min-seq-id 0.95 -e 1.000E-09
124 -c 0.95 --cov-mode 2). The search result was filtered for the best alignment for each pseudogene
125 using MMseqs2 (22) (v.13.45111, with module filterdb and parameter --extract-lines 1). The
126 protein sequences were used as a query for sequence search against the contig where the
127 corresponding pseudogene was found with BLAST (29) (v.2.12.0+, with module tblastn applying
128 a threshold of E-value ≤ 1×10^{-9}). This procedure yielded 135 (0.03%) pseudogenes that were
129 identical to a complete protein sequence. The annotated function of those protein-coding genes

130 corresponded mostly to hypothetical proteins (115 out of 135). Hence, we find no significant
131 evidence for PGAP misclassification of pseudogenes due to programmed frameshifts of ISs.

132 **Assignment of pseudogenes into gene families**

133 Pseudogenes were assigned into the KES gene families using their nucleotide sequences that
134 were extracted from genome assemblies according to the genomic coordinates annotated by
135 PGAP. Pseudogenes were aligned with the amino-acid sequences of the KES gene families using
136 MMseqs2 (22) (v.13.45111, with module search applying a threshold of E-value ≤ 1×10^{-10}).
137 These MMseqs2 parameters trigger an alignment of all six translated reading frames of the
138 pseudogene nucleotide sequence to the target amino acids sequence. Pseudogenes were
139 assigned into the gene family that had the highest number of significantly aligned gene family
140 members to the pseudogene. The final KES dataset encompassed 32,623 gene families
141 consisting of 11,995,860 protein coding genes and 393,350 (96% or the total) pseudogenes in
142 13,375 gene families. The annotation of gene function was extracted from the RefSeq database
143 and gene families were assigned the most frequent annotation among their member genes. The
144 annotation of specific gene families discussed in detail in the results was further validated using
145 sequence search against TnCentral (30) and Pfam (31) using the most frequent sequence variant
146 of the gene families as query. During this analysis, pseudogenes in a single relaxase/nuclease
147 domain-containing gene family were suspected as erroneously classified as such, and were
148 excluded from the analysis of ColE1-like plasmids. The excluded pseudogenes were annotated
149 in the sequences of 487 plasmids. The pseudogene had a highly conserved synteny within
150 mobilization related functions such as *mobC*, *mbeB,* and *mbeD*. In addition, homologous ORFs
151 to *mbeA* (e.g., WP_077943844.1) were identified at the same genomic location supporting the
152 classification of that locus as a gene rather than pseudogene.

153 **Inference of dN and dS distances**

154 As a first step, the most similar gene sequence to the pseudogene was identified. For that,
155 pseudogene nucleotide sequences were translated into amino acid sequences using transeq from
156 EMBOSS (32, 33) (v.6.6.0.0, with parameters -frame 6 and -table 11). Translated pseudogenes
157 were searched against the KES protein sequences using BLAST (29) (v.2.14.0+, with module
158 blastp). The best hit was used for further comparison. A pairwise alignment between the
159 translated pseudogene and the best matching amino acid sequence was inferred using Clustal
160 Omega (34) (v.1.2.4, --outfmt=clu). Codon alignment was calculated using pal2nal (35) (v.14.1,
161 with parameters -output paml and -codontable 11). The dN and dS distances were calculated
162 using codeml from PAML (36) (v.4.9, with runmode = -2). Additionally, the maximum truncation

163 size of pseudogenes compared to their homologous genes was determined by calculating the
164 leading and trailing truncations between the pseudogene and the best matching homologous
165 gene.

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

167 The closest homologous gene of pseudogenes in the KES dataset were inferred using MMseqs2
168 (22) (v.13.45111, with module search applying thresholds of E-value ≤ 1×10^{-10} and amino acid
169 identity ≥ 30%). Thereby, four scenarios (categories) of the homologous gene locus were
170 examined in the following hierarchical order: (I) on the same replicon, (II) on the same replicon
171 type, (III) on a different replicon type, and (IV) unitary pseudogene (see illustration of Fig. 1B). In
172 the example of plasmid pseudogenes, this analysis determined homologous protein-coding genes
173 from the same plasmid, other plasmids of the same isolate, or the chromosome of the same
174 isolate. If no homolog was found in the same bacterial isolate, the pseudogene was classified as
175 unitary. The same procedure was applied for chromosomal pseudogenes identifying homologous
176 genes on the same chromosome or plasmids within the same bacterial isolate.

177 **Plasmid typing and inference of closely related (homologous) plasmids**

178 Plasmid mobility type was predicted using MOB-suite (15) (v.3.0.3). The inference of closely
179 related plasmids was based on a new scheme of plasmid taxonomic units (PTUs) (37) and shared
180 gene content. Plasmids were assigned to plasmid taxonomic units (PTUs) using COPLA (38)
181 supplying taxonomic information of their hosts (Supplementary Table S3). Plasmid taxonomic
182 units correspond to a classification scheme of closely related plasmids as inferred from average
183 nucleotide identity (ANI) networks of whole plasmid sequences (38). Additionally, plasmids were
184 clustered into clusters of homologous plasmids based on shared gene content. Pairwise shared
185 gene and pseudogene content of plasmids were calculated with the Jaccard similarity index from
186 the gene and pseudogene families. The shared gene and pseudogene content were visualized
187 with hierarchical clustering using pheatmap (v.1.0.12). Plasmid pairs sharing ≥ 60% of their genes
188 were filtered to construct plasmid clusters using the mcl algorithm (39) (v.12-135, parameter --
189 abc -I 2.0).

190 **Phylogenetic analysis**

191 Plasmid and host phylogenies were constructed from complete single copy gene families for the
192 analyzed set of plasmids or chromosomes. Sequences were aligned using MAFFT (40) (v.7.520,
193 with parameter --maxiterate 1000 --localpair). Host phylogeny in 133 universal chromosomal
194 genes was reconstructed using IQ-TREE 2 (41) (v.2.2.2.7) with Le and Gascuel (LG) model based
195 on amino acid sequences (with parameter -m LG -B 1000, (42)). The host phylogeny was rooted

196 at midpoint using iTOL (43) (v.6.8). The reconstruction of plasmid phylogenetic trees was based
197 on nucleotide sequences of complete single-copy genes (see Supplementary Table S4 & S5)
198 applying IQ-TREE 2 (41) (v.2.2.2.7) with a transition model (with parameter -m TIM -B 1000,
199 (42)). The root position in the plasmid phylogenies was inferred by applying phylogenomic rooting
200 using nucleotide sequences from plasmid gene families (44) (Supplementary Table S4 & S5). The
201 obtained phylogenetic trees were visualized and annotated using iTOL (43) (v.6.8).

202 **Statistics**

203 All statistical tests were performed using R (v.4.0.3).

204 **Results**

205 **Higher density of pseudogenes in plasmids compared to chromosomes**

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

215 The size difference between plasmids and chromosomes raises the question of whether
216 replicon size (i.e., plasmid or chromosome size) may explain the different pseudogene densities.
217 To test the impact of replicon size differences, we constructed a dataset comprising chromosomes
218 and plasmids with similar size distributions (Supplementary Fig. S3). Comparing the pseudogene
219 density between plasmids and chromosomes in that set showed that, also for plasmids and
220 chromosomes of similar size, plasmids have a significantly higher pseudogene density compared
221 to chromosomes ($P < 0.001$, using permutation test, $median_{plasmid}=0.06$, $median_{chromosome}=0.03$).
222 Hence, the smaller plasmid genome size cannot alone explain the high pseudogene density in
223 plasmids.

224 To test if the high pseudogene density in plasmids is a general phenomenon, we
225 compared the pseudogene density between plasmids and chromosomes within different genera.
226 Our results show that the pseudogene density was consistently higher in plasmids compared to
227 chromosomes for all taxa, except *Shigella* ($P < 0.001$, using Wilcoxon test with FDR correction,

228 Supplementary Fig. S4). The high frequency of gene non-functionalization in *Shigella* has been
229 previously associated with adaptation of a host-associated lifestyle in this genus (11, 45). Hence,
230 the elevated occurrence of pseudogenes in plasmids is common to most taxonomic groups.

231 To study plasmid gene non-functionalization in detail, we focused on plasmids residing in
232 2,441 *Klebsiella*, *Escherichia*, and *Salmonella* (KES) isolates. These three enterobacterial taxa
233 have been extensively sampled for sequencing and plasmids in these taxa include diverse
234 plasmid types (21, 37). The density of pseudogenes in KES plasmids is significantly higher
235 compared to chromosomes ($P < 0.001$, using Wilcoxon test, Fig. 1A). This reflects an observed
236 4-fold increase in pseudogene density in plasmids compared to chromosomes and equates to
237 approximately 9 CDS per pseudogene in plasmids and 40 CDS per pseudogene in chromosomes.
238 Hence, enterobacterial genomes within the KES dataset capture the trend of increased
239 pseudogene density in plasmids compared to chromosomes that we observed within the
240 taxonomically diverse RefSeq dataset.

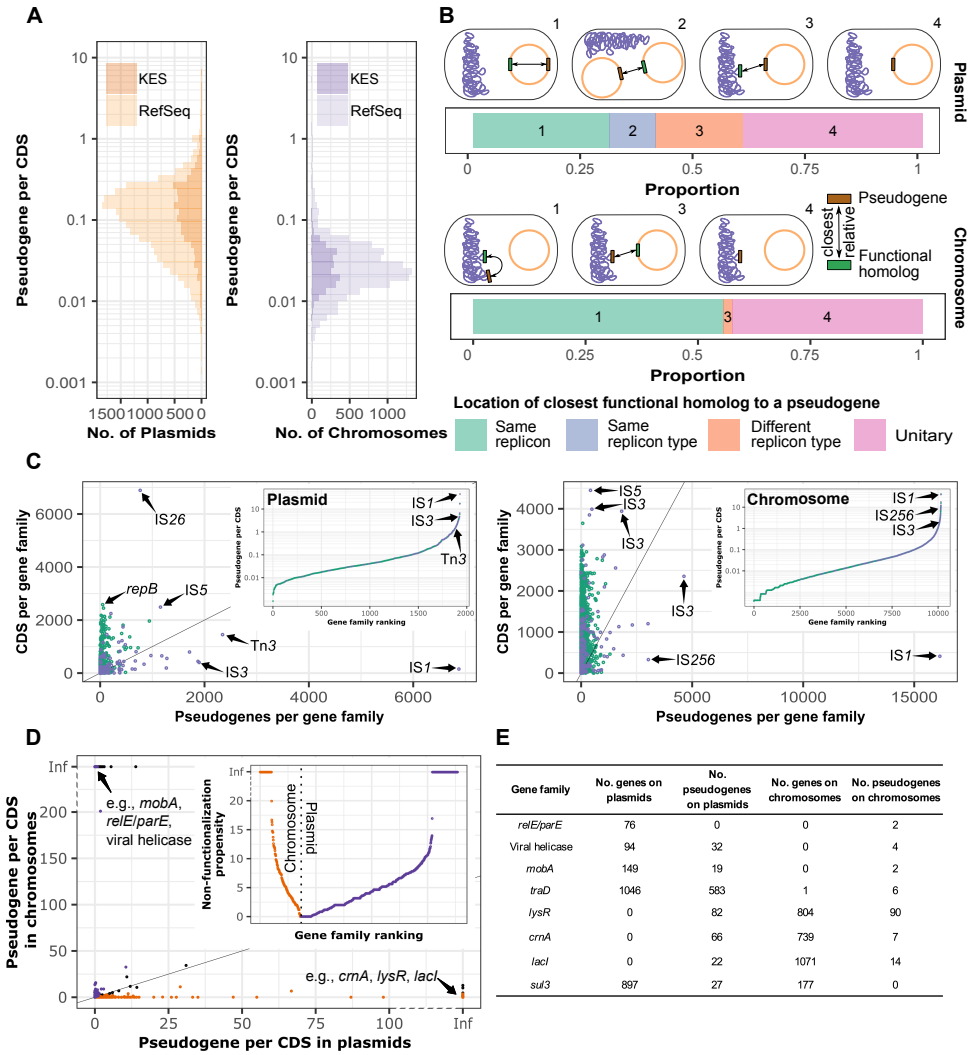
241 Pseudogenes having a functional homolog in the isolate genome, including the
242 chromosome and extra-chromosomal elements, may have evolved from gene non-
243 functionalization following a gene duplication (either within the same replicon, or via translocation
244 of duplicates between replicons). Pseudogenes that do not have a functional homolog within the
245 same isolate may have evolved either from non-functionalization of a single-copy gene, or
246 alternatively, non-functionalization of a horizontally acquired gene upon arrival. The comparison
247 of pseudogenes with their functional homologs showed that most pseudogenes are truncated
248 (289,746, 71%). Recent studies suggest that early stop codons in many truncated pseudogenes
249 correspond to nonsense mutations (46), or transient alleles that may be reversed under purifying
250 selection (47). Comparing the evolutionary rates of truncated and non-truncated pseudogenes we
251 find that truncated pseudogenes are characterized by higher dN and dS, as well as higher dN/dS
252 compared to non-truncated pseudogenes (Supplementary Fig. S5), as is expected for DNA
253 evolution following a relaxation of selection pressure (i.e., gene non-functionalization).
254 Furthermore, the divergence of pseudogenes in plasmids and chromosomes compared to their
255 closest homologous gene had similar characteristics (Supplementary Fig. S5), hence the high
256 density of pseudogenes in plasmids cannot be explained by a bias in the genome annotation
257 pipeline.

258 Pseudogenes were further classified into two groups: unitary and non-unitary
259 pseudogenes, based on whether a functional homolog to the pseudogene was present within the
260 same isolate. Our results show that the proportion of pseudogenes that are accompanied by a
261 homologous gene on the same replicon is lower for plasmid pseudogenes (Fig. 1B, see also

262 Supplementary Fig. S6). The proportion of unitary pseudogenes at the isolate level is comparable
263 between plasmids (40%) and chromosomes (42%). Plasmid pseudogenes may furthermore have
264 a homologous gene either in the chromosome (19%) or in another plasmid (10%) within the same
265 isolate. Chromosomal pseudogenes having a plasmid homologous gene are rather rare (2%).
266 Plasmid pseudogenes are thus more likely to have a homologous gene on a different replicon
267 within the isolate genome compared to chromosomal pseudogenes. Accordingly, we conclude
268 that gene non-functionalization in plasmids frequently entails the loss of redundant functions at
269 the level of the whole isolate genome.

270 **The majority of pseudogenes in plasmids correspond to deteriorated insertion sequences**
271 Which genes are frequently non-functionalized? In plasmids, gene families comprising
272 transposition-related gene functions such as IS*1*, Tn*3*, and IS*3* are characterized by the highest
273 frequency of pseudogenes (Fig. 1C). Similarly, transposases such as those encoded by IS*1*,
274 IS*256*, and IS*3*, are frequently non-functionalized in chromosomes (Fig. 1C). Non-functional
275 insertion sequences (ISs) accounted for a substantial portion of pseudogenes in our dataset with
276 88,107 (22%) pseudogene instances. For example, the IS*1* family transposase alone comprised
277 approximately every 20th pseudogene (5.85%), with an average of 33 pseudogenes per gene.
278 Furthermore, pseudogenes in the IS*1* family transposases are the most abundant in the data and
279 occur in 72% of the chromosomes (Supplementary Fig. S7). Our results are thus in agreement
280 with previous reports on frequent non-functionalization of TEs in prokaryote genomes (6) and
281 furthermore show a similar trend in plasmid genomes.

282 The comparison between plasmids and chromosomes further reveals that the propensity
283 for non-functionalization may be different depending on the replicon type. We identified 221
284 (0.64%) plasmid-encoded gene families that are significantly enriched for pseudogenes in the
285 chromosome (Fig. 1D, Supplementary Table 6); of these families, 84 (38.01%) include a
286 homologous gene in the plasmid within the same isolate (i.e., as category 3 in Fig. 1B). These
287 families often correspond to plasmid genes, such as mobilization genes (*mobA*), toxin-antitoxin
288 systems (*relE*/*parE*), and viral helicases (Fig. 1E). Additionally, we identified 840 (2.4%)
289 chromosomal gene families that are significantly enriched for pseudogenes in plasmids (Fig. 1D,
290 Supplementary Table S6); of these families, 498 (59.28%) have a chromosomal homologous
291 gene in the same isolate. These families include genes such as LysR-type transcriptional
292 regulators (*lysR*), LacI-related transcriptional regulators (*lacI*), and creatininase (*crnA*), that are
293 specifically encoded on the chromosome (Fig. 1E). We hypothesize that these gene families
294 correspond to replicon-specific genes that underwent non-functionalization upon translocation to
295 the other replicon, rendering them essentially dead-on-arrival. Barriers for gene transfer from



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 2,441 (100%) chromosomes). An increased pseudogene density in plasmids compared to chromosomes was observed also when analyzing pseudogenes detected with Pseudofinder (see Supplementary text, Supplementary Fig. S1) **(B)** Proportions of pseudogenes according to the location of the closest related homologous CDSs in plasmids and chromosomes. The median sequence similarity between pseudogenes and their homologous genes ranged between 59-87% in the five location categories (Supplementary Fig. S6). **(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 marked (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 (31) and TnCentral (30) databases using sequence search with the most frequent sequence variant of the gene families.

317 chromosomes to plasmids have been previously shown, for example due to dose effect (termed
318 also dose repetition) (48). The acquisition of an extra copy of a chromosomal protein-coding gene
319 may lead to an increased dose of the product protein, which may have a negative effect on the
320 host fitness. For example, the presence of an additional copy of the core chromosomal chaperonin
321 genes *groEL*/*groES* on a plasmid in *E. coli* was shown to have a negative fitness effect on the
322 plasmid host (48). Our results suggest the presence of similar barriers for gene transfer from
323 plasmids to chromosomes.

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

325 Which plasmid types are most associated with a high frequency of pseudogenes? Initially we
326 focus on two prominent plasmid characteristics: plasmid size and mobility class. The small
327 plasmid types in the KES dataset mostly correspond to ColE1-like plasmids; accordingly, here we
328 consider plasmids of <19Kbp as small (21). Our analysis shows that most of the large plasmid
329 types (4,659, 97%) and about half of the small plasmid types (1,045, 54%) contain pseudogenes.
330 The pseudogene density is highest in small non-mobilizable plasmids (small plasmid types with
331 pseudogenes: $median_{small\ mobilizable}$=0.2, $median_{small\ non\text{-}mobilizable}$=0.33) compared to large plasmid
332 types and chromosomes, but that property is tightly linked to their small size encoding only few
333 genes. Among the large plasmids, pseudogene density is highest in mobilizable plasmids followed
334 by non-mobilizable plasmids ($P < 0.05$, using pairwise Wilcoxon test with FDR correction,
335 $median_{large\ mobilizable}$=0.27, $median_{large\ non\text{-}mobilizable}$=0.15, $median_{large\ conjugative}$=0.1, Fig. 2A). Thus,
336 approximately every fourth gene is non-functionalized in some mobilizable plasmids. Taken
337 together, our results point towards an association between increased pseudogene densities in
338 large plasmids and a lack of plasmid mobility.

339 Plasmids in the different plasmid size and mobility categories are typically different in their
340 composition. We therefor compared the propensity of gene non-functionalization among the
341 plasmid size and mobility groups. Only seven gene families were enriched for pseudogenes in
342 the small plasmid types including antibiotic resistance genes (Supplementary Fig. S8,
343 Supplementary Table S7). A total of 488 gene families were enriched for pseudogenes in at least
344 one of the large plasmid mobility classes (Fig. 2B). The comparison between large plasmids in
345 different mobility classes revealed a significant enrichment for pseudogenes in gene families
346 encoding for plasmid transfer mechanisms including *traI*, *nikB*, and *traV* (Fig. 2B, Supplementary
347 Table S8). Consequently, we hypothesized that the origin of these pseudogenes is gene non-
348 functionalization following loss of self-transmissibility of conjugative plasmids.

349

350

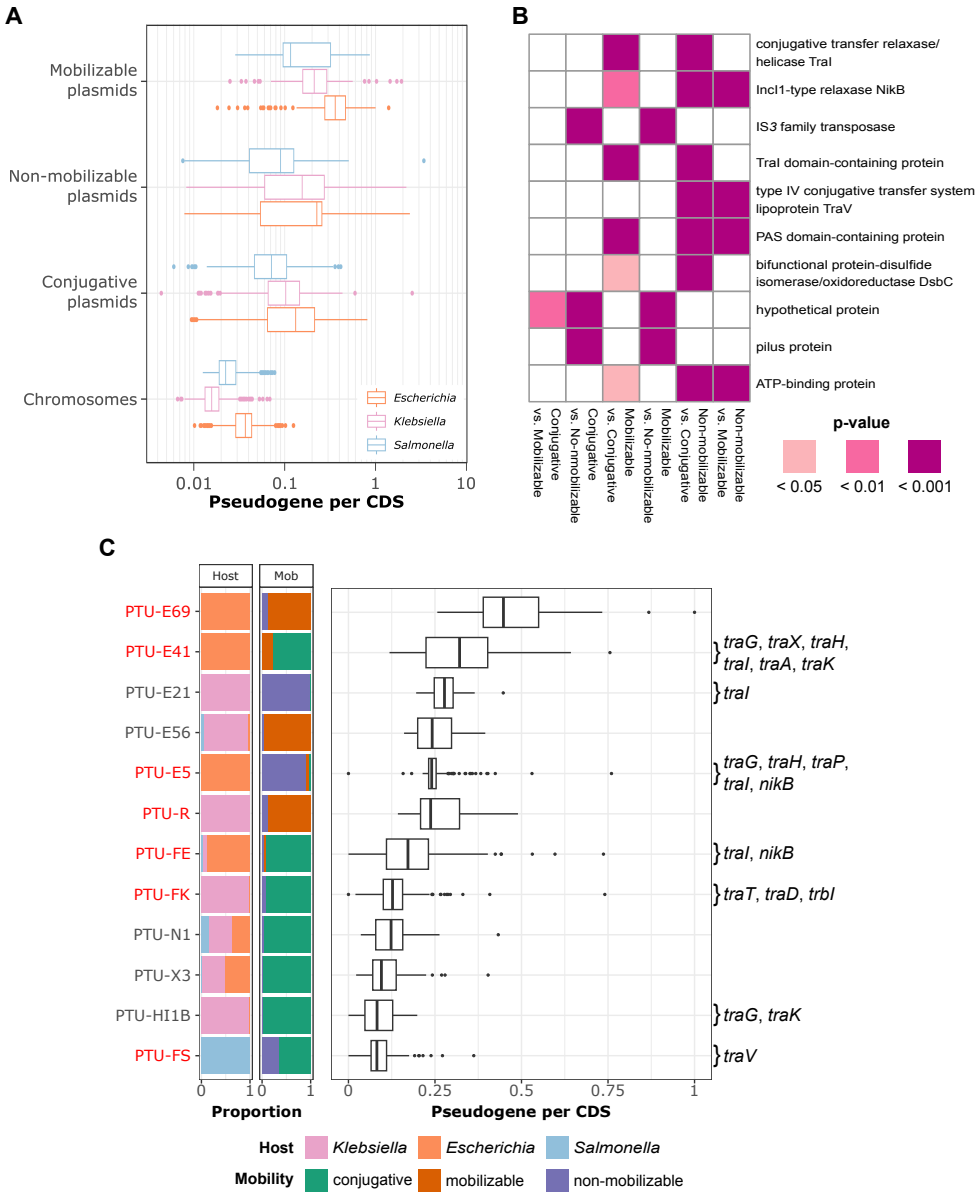


Figure 2. Pseudogene density and gene function in large (≥19Kb) plasmid types. (A) Ratio of pseudogene per CDS shown for large plasmid types (≥19Kb) and chromosomes (Replicons lacking pseudogenes were excluded. This includes 94 (3%) conjugative, 9 (1.7%) mobilizable, and 61 (5.1%) non-mobilizable plasmids). **(B)** Enrichment for pseudogenes in gene families significantly depends on the plasmid mobility type in large plasmids ($P < 0.05$, using Fisher's exact test with FDR correction). The 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%).

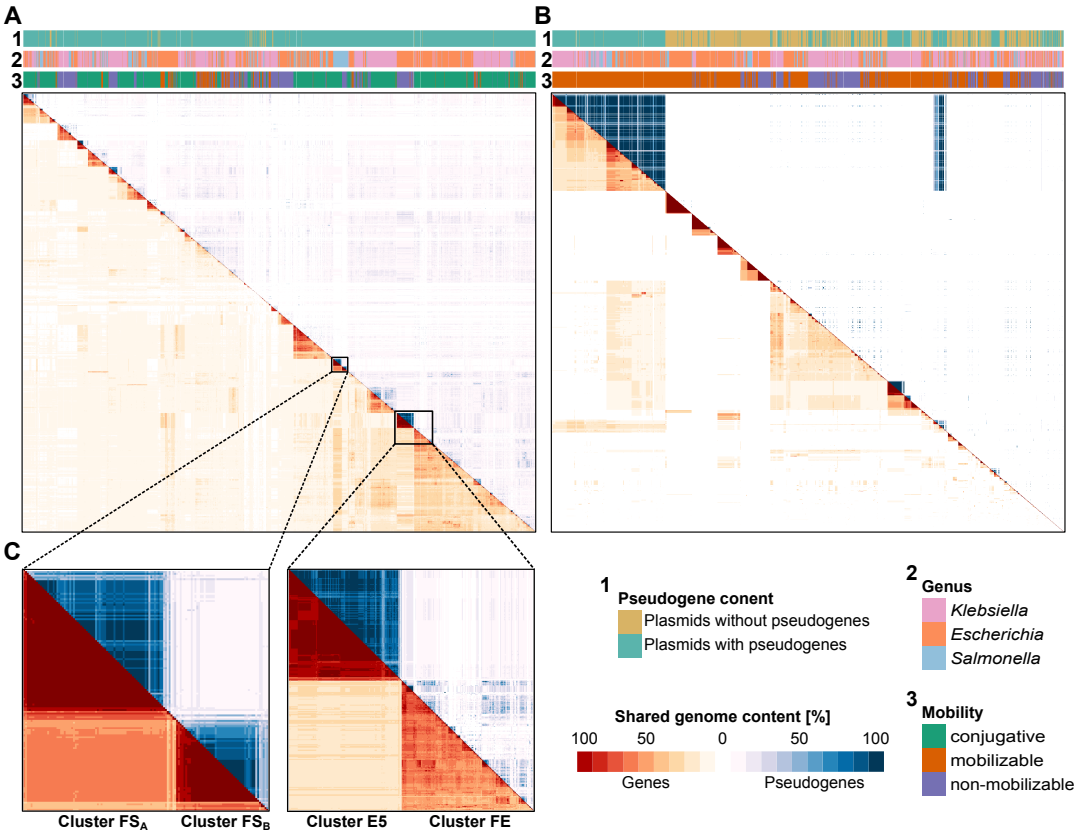
367 To gain further understanding of gene non-functionalization events in the context of large plasmid
368 types, we compared the pseudogene density among closely related plasmids as inferred from
369 plasmid taxonomic units (PTUs) (37). PTUs that comprise a high pseudogene density typically
370 include a high proportion of mobilizable and non-mobilizable plasmid types (Fig. 2C). Gene
371 deletions and changes in genetic repertoire of plasmids may lead to a transition in plasmid
372 mobility, which may lead to PTU divergence and the evolution of novel PTUs in the long-term (17,
373 19). Indeed, here we observe several PTUs where transfer-related gene families frequently occur
374 as pseudogenes; these PTUs are highlighted as putative plasmid types where transitions in
375 plasmid mobility may have occurred (red labeled PTUs in Fig. 2C). Transfer-related pseudogenes
376 thus correspond to molecular fossils of lost mobility mechanisms in plasmid genomes.

377 **Vertical inheritance of pseudogenes in evolutionary-related plasmid clusters**

378 Pseudogenes observed within the same gene family may have originated either from multiple
379 independent gene non-functionalization events, or alternatively, from few non-functionalization
380 events in ancestral plasmids followed by vertical inheritance during plasmid diversification. If
381 vertical inheritance of pseudogenes is common in plasmid evolution, then the number of observed
382 pseudogenes supplies an overestimation for the frequency of gene non-functionalization events.
383 To evaluate the role of vertical inheritance in pseudogene evolution, we compared shared gene
384 and pseudogene content among closely related plasmids. For that purpose, we clustered all
385 plasmids into clusters of homologous plasmids based on shared gene content. We obtained 353
386 clusters comprising 85% large plasmids and 206 clusters comprising 89% of the small plasmids;
387 these clusters largely correspond to the PTUs (see Supplementary Fig. S9). A visual inspection
388 of the shared genes matrices suggests that pseudogenes tend to be shared among plasmids that
389 also share genes (i.e., clusters of closely related plasmid) (Fig. 3AB). Indeed, the proportion of
390 shared genes and shared pseudogenes between plasmid pairs are positively correlated ($r_s = 0.47$,
391 $P < 0.001$, using Spearman's correlation).

392 The observation of shared pseudogenes in plasmids may be due to ancestral gene non-
393 functionalization and vertical inheritance (i.e., retention) of non-functional DNA. An alternative
394 explanation for shared pseudogene content is independent gene non-functionalization events in
395 a narrow range of gene families. The strongest patterns of shared pseudogenes in our dataset
396 are mostly restricted to closely related plasmid clusters. A total of 874 (10.1%) gene families
397 comprising pseudogenes are plasmid cluster-specific ($P < 0.05$, using Fisher's exact test with
398 FDR correction). Hence, shared pseudogene content between plasmids clusters, is expected to
399 correspond to non-functionalization events in diverse gene families. A total of 356 (4.4%) gene
400 families correspond to frequent, non-functional transposase-related functions such as IS*1*, Tn*3*,

401 and IS*26*. Additionally, gene family size and the frequency of non cluster-specific pseudogenes
402 are significantly correlated (r_s = 0.58, P < 0.001, using Spearman's correlation). Shared
403 pseudogene content between plasmid clusters thus corresponds either to horizontal gene transfer
404 or commonly non-functionalized gene families.



405

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

413 **Non-functionalization following segmental deletion is a signature for the loss of self-**
414 **transmissibility**

415 To gain a better understanding on the evolutionary events at the basis of pseudogene vertical
416 inheritance, we reconstructed the evolution of closely related plasmid clusters. For that purpose,
417 we selected two clusters of *Salmonella* plasmids that share a high proportion of their genes and
418 correspond mostly to PTU-FS (Supplementary Table S3, Supplementary Fig. S9). The shared
419 gene content within the selected clusters FS_A and FS_B is 92% and 86%, respectively, with an
420 average of 50% shared genes among plasmids in both clusters (Fig. 3C). Plasmids of cluster FS_A

421 are homologous to plasmid pSLT (NC_003277.2) of *S. enterica* serovar Typhimurium and
422 plasmids of cluster FS_B are homologous to plasmid pSENV (NZ_CP063701.1) of *S. enterica*
423 subsp. *enterica* serovar Enteritidis. Cluster FS_A comprises only 83 conjugative plasmids and
424 cluster FS_B includes mostly (46, 92%) non-mobilizable plasmids, in addition to four conjugative
425 plasmids. The average shared pseudogene content among plasmids within the clusters is 67%
426 or 65% for clusters FS_A and FS_B, respectively. The shared pseudogene content among plasmids
427 from different clusters was only 9%, on average, thus plasmids in the two clusters are diverged in
428 their pseudogene content. The high proportion of shared gene (and pseudogene) content among
429 plasmids in the two PTU-FS clusters supports the notion that these plasmids are of common
430 ancestry, i.e., they are homologs, albeit with alterations in their mobility class.
431 To examine the evolution of PTU-FS plasmids, we reconstructed the plasmid phylogeny
432 based on eleven single-copy gene families present in 122 (92%) plasmids. The phylogeny reveals
433 a clear split between plasmid clusters FS_A and FS_B (Fig. 4A). To reconstruct the ancestral mobility
434 state of PTU-FS plasmids, we inferred the root of PTU-FS using a phylogenomic rooting approach
435 (44). The phylogenetic inference revealed a root neighborhood in the deepest split between
436 plasmid clusters FS_A and FS_B that includes branches leading to several conjugating plasmids of
437 cluster FS_A (Fig. 4A, see also Supplementary Fig. S10). This root position implies that the
438 ancestor of PTU-FS plasmids was conjugative and the non-mobilizable plasmids in cluster FS_B
439 are derived plasmids (Fig. 4A). Notably, the plasmid genome size differs significantly between the
440 non-mobilizable cluster FS_B plasmids and the conjugative FS_A plasmids, with the non-mobilizable
441 plasmids having a smaller genome size ($P < 0.001$, using Wilcoxon test, $median_{non-}$
442 $_{mobilizable}$=59,372 bp, $median_{conjugative}$=93,865 bp). Further comparison of the gene content among
443 the PTU-FS reveals a high conservation of plasmid gene order, except for the absence of transfer
444 (*tra*) genes in the non-mobilizable FS_B plasmids (Fig. 5A, Supplementary Fig. S11A).
445

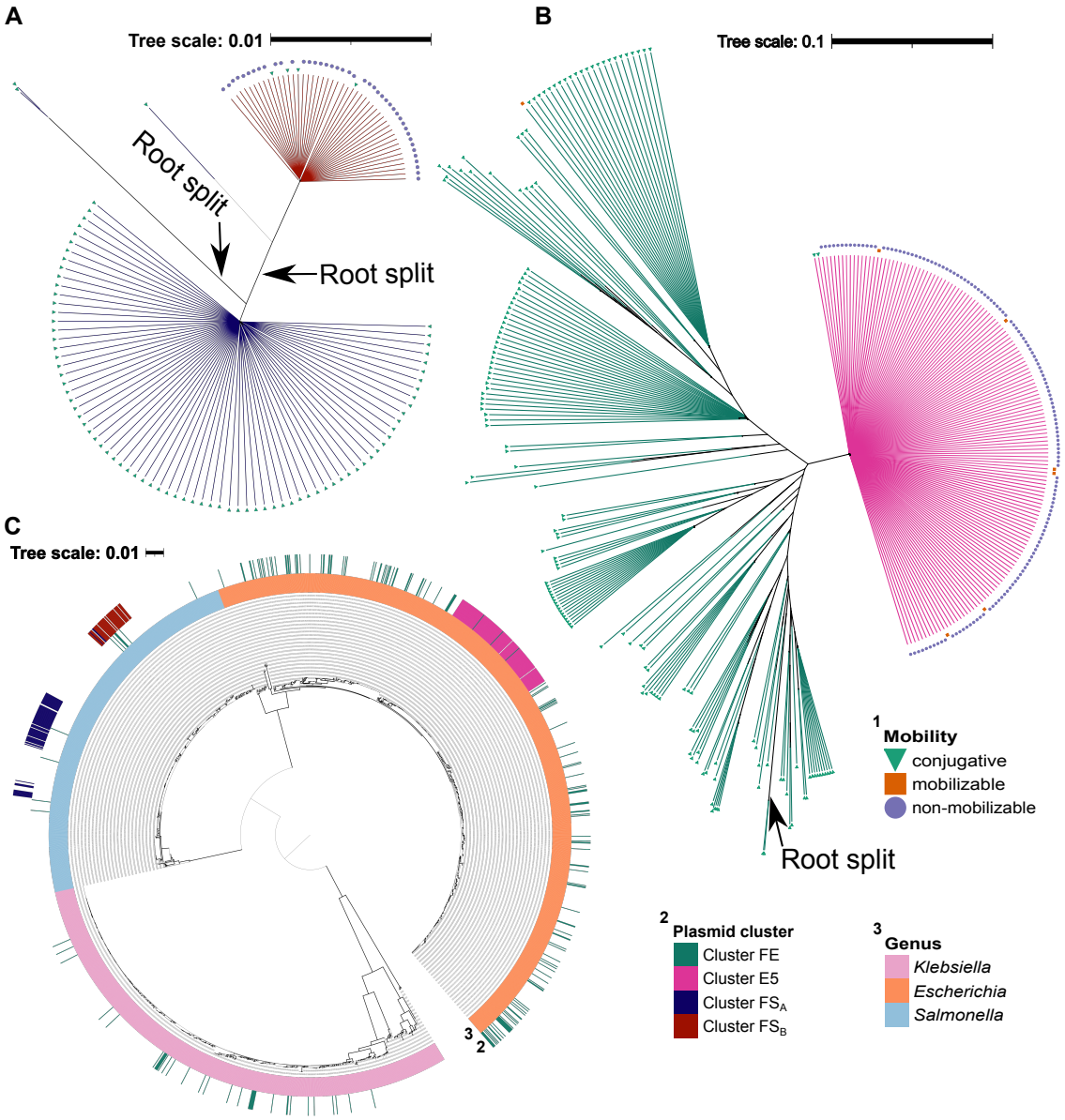


Figure 4. Phylogenetic relationship of self-transmissible and non-mobilizable plasmid clusters and their hosts. (A, B) Phylogenetic maximum-likelihood trees including 11 single-copy genes of plasmid clusters FS_A and FS_B (PTU-FS) and 10 single-copy genes of plasmid clusters FE (PTU-FE) and E5 (PTU-E5) (see Supplementary Table S4 & S5). 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 the inferred root positions resulting from root neighborhood inference (see supporting gene trees in Supplementary Fig. S10). The operational taxonomic unit symbols correspond to predicted plasmid mobility type (legend 1). The reconstructed tree branches are colored black (note the short branch length in nearly polytomic ancestral node). The plasmid clusters (PTUs) are shown by the colored lines extending from the tree branches (legend 2). **(C)** Host phylogenetic tree based on 133 chromosomal single-copy genes. The tree was rooted using midpoint rooting. The outer color strips of the phylogenetic host tree show the cluster affiliation of plasmids residing in the hosts (legend 2) and the host genera (legend 3).

462 The loss of the conjugation machinery genes may have followed two alternative scenarios:
463 gradual non-functionalization and gene loss, or a large-scale deletion followed by non-
464 functionalization of functionally-related genes. Since the absent genes are consecutively
465 arranged in the ancestral plasmid (Fig. 5C), we infer that a large segmental deletion of transfer-
466 related genes occurred in the ancestor of cluster FS_B non-mobilizable plasmids (Supplementary
467 Fig. S12). The transfer-related genes *traV* and *FinO* (a repressor of conjugative transfer in F-like
468 plasmids) correspond to the boundaries of the deleted transfer-related gene arrangement in the
469 ancestor and are found as pseudogenes in derived plasmids. Notably, the *Salmonella* host
470 phylogeny shows that plasmid clusters FS_A and FS_B reside in distinct groups of *Salmonella*
471 isolates where hosts of the conjugative FS_A plasmids appear as more distantly related in
472 comparison to hosts of the non-mobilizable FS_B plasmids (Fig. 4C, Supplementary Fig. S13).
473 Taken together, our results reveal a transition in plasmid mobility from self-transmissible to non-
474 mobilizable plasmids within *Salmonella* PTU-FS plasmids due to large segmental deletion
475 followed by gene non-functionalization events.
476 To examine further scenarios for gene non-functionalization in plasmid genomes, we
477 investigated two additional adjacent plasmid clusters in the shared gene matrix where the pattern
478 of gene and pseudogene sharing indicates that they are closely related (Fig. 3C). The plasmids
479 in these clusters correspond to the evolutionary related PTUs FE and E5 (19). Most (167, 87%)
480 of the plasmids in cluster FE were assigned to the conjugative PTU-FE, with the remaining
481 plasmids being unassigned (Supplementary Table S3, Supplementary Fig. S9). All 164 plasmids
482 in cluster E5 were classified into the non-mobilizable PTU-E5 (Supplementary Table S3,
483 Supplementary Fig. S9), which correspond to homologs of pO157 in *E. coli* strain O157:H7 and
484 are known to be non-transmissible (17, 49). Plasmids in cluster FE share, on average, 53% of
485 their genes and 12% of their pseudogenes. In comparison, gene sharing among plasmids in
486 cluster E5 shows a higher shared content with 90% of the genes and 80% of the pseudogenes
487 shared, on average. Plasmids in FE and E5 clusters share, on average, 18% gene content and
488 5% pseudogene content. The shared gene content among plasmids in the two clusters reflects
489 evolutionary relationships where the different plasmid mobility types may be explained by the
490 differential gene content.
491

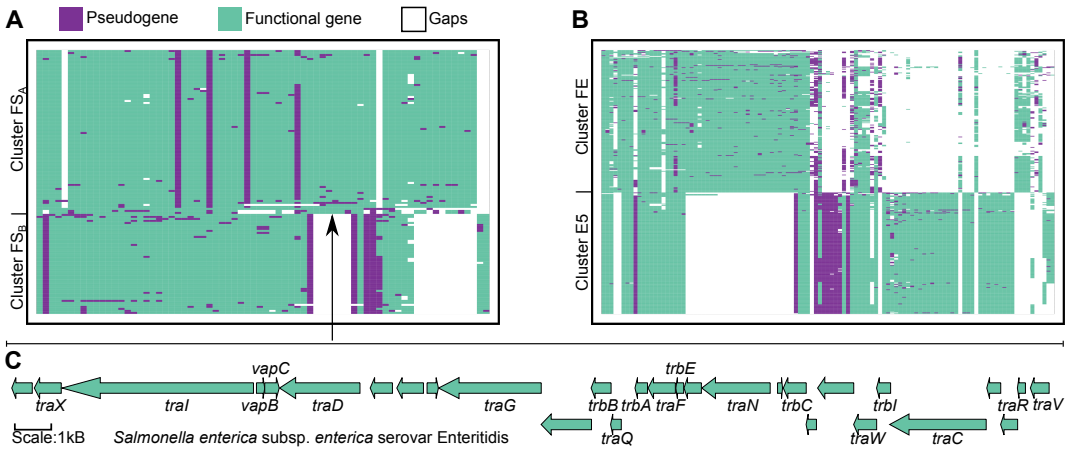


Figure 5. Pseudogene and gene family Presence Absence Pattern (PAP) of self-transmissible and non-mobilizable plasmid clusters. PAP of plasmid clusters FS_A/FS_B **(A)** and FE/E5 **(B)**. The PAPs are sorted using hierarchical clustering of the shared gene family content. **(C)** Gene content and order in the segmental deletion inferred for the ancestor of FS_B (non-mobilizable) plasmids (pointed by an arrow). The inference was based on the transfer-related gene neighborhood in cluster FS_A (conjugative) plasmids (see also Supplementary Fig. S12).

To examine the evolutionary history of plasmids in clusters FE and E5, we reconstructed the plasmid backbone phylogeny using ten single-copy gene families present in 316 (89%) plasmids. The phylogeny revealed a clear split between the plasmid clusters (Fig. 4B). The topology of cluster E5 plasmids reveal two main plasmid lineages of closely related plasmids. In contrast, the phyletic pattern of cluster FE conjugative plasmids revealed a high divergence among the plasmids (Fig. 4B). The inferred root position is located in a branch leading to two conjugative cluster FE plasmids (Supplementary Fig. S10). Notably, two plasmids in cluster E5 encode a partial repertoire of the conjugation machinery and they were accordingly predicted as conjugative; these plasmids branch closely to the split with cluster FE plasmids (Supplementary Fig. S14) and their syntenic gene order reveals differential loss of transfer-related genes (Supplementary Fig. S14). Comparing the sequences of the conjugative plasmids of cluster E5 to their closest non-mobilizable neighbor plasmid revealed evidence for loss of plasmid mobility genes and shared genomic arrangements in the non-mobilizable cluster E5 plasmids (Fig. 5B, Supplementary Fig. S12B). The phylogenetic reconstruction suggests that the ancestral plasmid of cluster E5 plasmids (or PTU-E5) was conjugative and the plasmid mobility was lost due to a segmental deletion of the *tra* genes (Fig. 4B, Supplementary Fig. S14). Plasmids in cluster E5 share 80% of their pseudogenes, including transposases such as IS*3* (n=508), IS*1* (n=336), and IS*91* (n=320), as well as transfer-related gene families *traI* (n=164) and *nikB* (n=164) that are non-functionalized in 164 (98%) of the cluster E5 plasmids. Plasmids in cluster FE reside in a diverse

519 set of hosts including *Klebsiella*, *Escherichia*, and *Salmonella*. In contrary, plasmids in cluster E5
520 are only hosted by closely related *Escherichia* isolates including strain O157:H7 (Fig. 4C); these
521 plasmids are homologs of plasmid pO157 that is non-mobilizable (e.g., NZ_CP017435.1). Note
522 that the hosts of cluster E5 plasmids are diverged, i.e., they do not correspond to isogenic strains
523 (Supplementary Fig. S13). Hence the topology of cluster E5 phylogeny likely corresponds to
524 within-host evolution via vertical inheritance. Notably, cluster E5 plasmids are characterized by
525 significantly larger plasmid genome size, but a smaller number of CDSs compared to the
526 conjugative cluster FE plasmids, which is well explained by the frequent pseudogenes in cluster
527 E5 plasmids ($P < 0.001$, using Wilcoxon test, $\text{median}_{\text{Cluster E5}}$=78 CDS, 19 pseudogenes, 93,179
528 bp; $\text{median}_{\text{Cluster FE}}$=86 CDS, 9 pseudogenes, 74,922.5 bp). Hence pseudogenes may be retained
529 as non-coding DNA in the evolution of the non-mobilizable plasmids after the loss of self-
530 transmissibility. Our results thus reveal that plasmid divergence events may give rise to new
531 integral replicons within their host lineage.

532 **Discussion**

533 Gene non-functionalization is considered a common event following gene duplication and gene
534 acquisition via lateral gene transfer. Previous studies suggested that bacterial genomes are often
535 devoid of pseudogenes due to purifying selection (8). This implies that pseudogenes in bacterial
536 genomes should be considered as 'garbage DNA', that is, non-functional DNA that may have an
537 effect on the organism fitness, rather than 'Junk DNA' (see (50) for definitions). Our results show
538 that pseudogenes are found in bacterial genomes and their density is highest in plasmids. We
539 identify two main processes in plasmid evolution that are accompanied by frequent gene non-
540 functionalization: proliferation of mobile genetic elements (MGEs) and loss of plasmid mobility
541 (i.e., reductive plasmid evolution).

542 Mobile genetic elements, including transposons and integrons, are known to translocate
543 or proliferate within and between genomes (51), thus, facilitating gene translocation and transfer,
544 as well as genomic rearrangements (reviewed in (52)). Specifically in conjugative plasmids, ISs
545 are known to mediate and facilitate the transfer of antibiotic resistance genes (53). At the same
546 time, the integration of MGEs in bacterial genomes may lead to gene non-functionalization due to
547 disruption of open reading frames at the insertion locus. Examples are previous reports on IS-
548 mediated inactivation of the arginine biosynthesis pathway in the tsetse fly symbiont *Sodalis*
549 *glossinidius* (54) and an inactivation of the plasmid-encoded nitrogen fixation genes (*nif*) in
550 *Bradyrhizobium* isolates (55). The high frequency of MGE pseudogenes in plasmids reported here

551 conforms previous studies highlighting the role of failed transposition events in the generation of
552 pseudogenes in bacterial genomes (e.g., (6)). We observed that most pseudogenes are due to
553 gene truncations and that pseudogenes in plasmids are more deteriorated. Indeed, site-specific
554 recombination of transposable elements generates genomic hotspots for their insertion, where
555 repeated insertions at the same locus may lead to gene truncations (reviewed in (56)). That being
556 said, recent studies suggest that many pseudogenes are transcribed, and sometime even
557 translated (57). Pseudogenes of MGE origin may thus correspond to proto-genes and a source
558 for genetic innovation during MGE diversification (47). This suggestion is in line with recent
559 studies that report pseudogene resurrection under strong selective conditions for the lost gene
560 function. Examples include the resurrection of iron uptake in *E. coli* (58) and the resurrection of
561 CO_2 assimilation in *Brucella* spp. (59).

562 Our analysis furthermore reveals that the propensity for specific gene non-
563 functionalization depends on the replicon type, i.e., plasmid or chromosome. This observation
564 may be explained by selection against the presence of essential (chromosomal) genes in
565 plasmids, e.g., due to unreliable plasmid inheritance (60) or deleterious effects of gene duplication
566 (48). Thus, the higher pseudogene density in plasmids compared to chromosomes may be
567 explained by the propensity of plasmid gene content to include accessory and dispensable genes.
568 We note however, that the classification of genes into core and accessory may vary depending
569 on the bacterial taxon, and that has an effect on the evolution of plasmid gene content in different
570 bacterial taxa (e.g., in the context of antibiotic resistance plasmids; (61)). We hypothesize that the
571 horizontal transfer (or translocation) of chromosomal genes into plasmids (and *vice versa*) is
572 bound to yield effectively 'dead-on-arrival' pseudogenes.

573 Inferring rooted plasmid phylogenies that comprise related non-mobilizable and
574 conjugative plasmids enabled us to infer the ancestral and derived plasmid variants in plasmid
575 evolution. According to our inference, non-functionalized copies of transfer-related genes in large
576 non-mobilizable and mobilizable plasmids are molecular fossils underlining a self-transmissible
577 plasmid origin. The loss of plasmid self-transmissibility may lead to plasmid 'domestication' in the
578 host lineage (or plasmid 'fixation' in the host pangenome). Domesticated plasmids are
579 characterized by a stable vertical inheritance within the lineage that can be considered as
580 ancestral states of chromids (62). Examples for the evolution of chromids from essential plasmids
581 are found in plant-associated bacteria such as *Sinorhizobium meliloti* harboring a chromid,
582 pSYMB, that encodes essential genes (63). Nonetheless, evolution of domesticated plasmids
583 following the acquisition of essential genes was so far not considered in *Enterobacteriaceae* (48,
584 60). Previous studies reported a stable inheritance of plasmids pO157 in *Escherichia* and pSLT

585 in *Salmonella* (64–66). Although, pSLT is self-transmissible, it was shown that most *Salmonella*
586 virulence plasmids are rather vertically inherited except for pSENV of *S. enterica* subsp. *enterica*
587 serovar Enteritidis (67). Plasmids of *S. enteritidis* bear incomplete *tra* regions (68) and virulence
588 plasmids of serovars Enteritidis and Typhimurium are not mobilizable by other F-like plasmids
589 even in the presence of an *oriT* region (69). The results of our analysis support the view that
590 pSENV plasmid homologs were acquired by horizontal transfer subsequent to a segmental
591 deletion of transfer-related gene (the evolution of serovar specific virulence plasmids reviewed in
592 (70)). Furthermore, the phylogeny of isolates hosting pO157- and pSENV-like plasmids indicates
593 that these plasmids were fixed in those lineages following the loss of self-transmissibility (Fig.
594 4C), hence they are better considered as domesticated plasmids (see also (71)).
595 The pseudogene repertoire and order in large non-mobilizable plasmids indicate that
596 segmental deletions within the conjugation machinery play a major role in the initial steps of
597 transitions in plasmid mobility. Such events may be caused, for example, by a transposon-
598 mediated segmental deletion. A similar event was observed in an experimental evolution study of
599 plasmid pKP33 adaptation to a naïve *E. coli* host under selection for the plasmid-encoded
600 antibiotic resistance (18). Such a change in the plasmid genomic composition that leads to the
601 loss of self-transmissibility is likely to trigger a 'domino effect' of sequential non-functionalization
602 events of other transfer-related genes. Similar dynamics of gene non-functionalization have been
603 previously suggested for the evolution of pseudogene content in obligatory symbionts (72). The
604 loss of self-transmissibility may be deleterious for plasmids due to the inevitable restriction in the
605 host range. Nonetheless, plasmids encoding a trait that is essential for the host will be maintained
606 in the population, also if their inheritance is unstable due to selection acting at the level of host
607 (e.g., see (73)). Another commonality in the evolution of plasmids and facultative pathogens is
608 the interplay of MGE proliferation and pseudogene accumulation that is associated with genome
609 miniaturization. Examples for pathogens having highly reduced (functional) genome size following
610 massive gene decay are *Mycobacterium leprae*, *Yersinia pestis*, *Shigella flexneri*, *Rickettsia*
611 *prowazekii*, *Salmonella enterica* serovars Paratyphi A, and Typhi (11, 74–78). Genome
612 miniaturization following gene non-functionalization and loss has been furthermore described in
613 the evolution of obligatory symbionts in heritable symbioses where deleterious mutations can be
614 fixed by genetic drift (reviewed in (79)). Similarly to plasmids, frequent gene non-functionalization
615 in the evolution of symbionts can lead to host range restriction that is accompanied by a shift from
616 facultative to obligatory symbiosis. Co-adaptation plasmids that lost their mobility and their host
617 can lead to the evolution of stable plasmid inheritance (e.g., as observed in experimental plasmid
618 evolution studies (18, 76, 77)) and eventually plasmid domestication. Reductive genome evolution

of plasmids is thus akin to the evolution of obligatory symbionts and the mechanisms involved in these processes – transposable element proliferation and gradual gene non-functionalization leading to shifts in the plasmid lifestyle – are likely similar.

Data availability

The data and accession numbers (RefSeq database) are provided in electronic supplementary material.

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 number: 101043835), and the China Scholarship Council (CSC scholarship to Y.W.).

Competing interests

The authors declare no competing interests.

Acknowledgements

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

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.

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.

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

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

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

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

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

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

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

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

668 11. Jin,Q. (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through
669 comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**,
670 4432–4441.

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

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

675 14. Rodríguez-Beltrán,J., DelaFuente,J., León-Sampedro,R., MacLean,R.C. and San Millán,Á.
676 (2021) Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. *Nat.*
677 *Rev. Microbiol.*, **19**, 347–359.

678 15. Robertson,J. and Nash,J.H.E. (2018) MOB-suite: software tools for clustering, reconstruction
679 and typing of plasmids from draft assemblies. *Microb. Genom.*, **4**, e000206.

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

682 17. Ares-Arroyo,M., Coluzzi,C. and Rocha,E.P.C. (2023) Origins of transfer establish networks of
683 functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Res.*, **51**,
684 3001–3016.

685 18. Porse,A., Schønning,K., Munck,C. and Sommer,M.O.A. (2016) Survival and evolution of a
686 large multidrug resistance plasmid in new clinical bacterial hosts. *Mol. Biol. Evol.*, **33**,
687 2860–2873.

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

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

694 21. Wang,Y. and Dagan,T. (2024) The evolution of antibiotic resistance islands occurs within the
695 framework of plasmid lineages. *bioRxiv*, doi: 10.1101/2024.02.20.581145.

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

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

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

703 25. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G.,
704 Chitsaz,F., Derbyshire,M.K., Durkin,A.S., *et al.* (2021) RefSeq: expanding the Prokaryotic
705 Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.*,
706 **49**, D1020–D1028.

707 26. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for
708 prediction of gene starts in microbial genomes. Implications for finding sequence motifs in
709 regulatory regions. *Nucleic Acids Research*, **29**, 2607–2618.

710 27. Syberg-Olsen,M.J., Garber,A.I., Keeling,P.J., McCutcheon,J.P. and Husnik,F. (2022)
711 Pseudofinder: detection of pseudogenes in prokaryotic genomes. *Mol. Biol. Evol.*, **39**,
712 msac153.

713 28. Chandler,M. and Fayet,O. (1993) Translational frameshifting in the control of transposition in
714 bacteria. *Mol. Biol.*, **7**, 497–503.

715 29. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L.
716 (2009) BLAST+: architecture and applications. *BMC Bioinform.*, **10**, 421.

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

720 31. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L.,
721 Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families
722 database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

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

725 33. 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 34. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H.,
729 Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein
730 multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

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

734 36. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**,
735 1586–1591.

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

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

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

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

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

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

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

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

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

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

760 47. Feng,Y., Wang,Z., Chien,K.-Y., Chen,H.-L., Liang,Y.-H., Hua,X. and Chiu,C.-H. (2022)
761 "Pseudo-pseudogenes" in bacterial genomes: proteogenomics reveals a wide but low

762 protein expression of pseudogenes in *Salmonella enterica*. *Nucleic Acids Res.*, **50**, 5158–
763 5170.

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

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

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

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

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

775 53. Che,Y., Yang,Y., Xu,X., Břinda,K., Polz,M.F., Hanage,W.P. and Zhang,T. (2021) Conjugative
776 plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial
777 resistance genes. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2008731118.

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

781 55. Arashida,H., Odake,H., Sugawara,M., Noda,R., Kakizaki,K., Ohkubo,S., Mitsui,H., Sato,S.
782 and Minamisawa,K. (2022) Evolution of rhizobial symbiosis islands through insertion
783 sequence-mediated deletion and duplication. *ISME J*, **16**, 112–121.

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

786 57. Soler-Camargo,N.C., Silva-Pereira,T.T., Zimpel,C.K., Camacho,M.F., Zelanis,A., Aono,A.H.,
787 Patané,J.S., dos Santos,A.P. and Guimarães,A.M.S. (2022) The rate and role of
788 pseudogenes of the *Mycobacterium tuberculosis* complex. *Microb. Genom.*, **8**, e000876.

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

792 59. Varesio,L.M., Willett,J.W., Fiebig,A. and Crosson,S. (2019) A carbonic anhydrase
793 pseudogene sensitizes select *Brucella* lineages to low CO_2 tension. *J. Bacteriol.*, **201**,
794 e00509-19.

795 60. Tazzyman,S.J. and Bonhoeffer,S. (2014) Why there are no essential genes on plasmids. *Mol.*
796 *Biol. Evol.*, **32**, 3079–3088.

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

800 62. Vial,L. and Hommais,F. (2020) Plasmid-chromosome cross-talks. *Environ. Microbiol.*, **22**,
801 540–556.

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

805 64. Makino,K. (1998) Complete nucleotide sequences of 93-kb and 3.3-kb plasmids of an
806 enterohemorrhagic *Escherichia coli* O157:H7 derived from Sakai outbreak. *DNA Res.*, **5**,
807 1–9.

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

810 66. Lobato-Márquez,D., Molina-García,L., Moreno-Córdoba,I., García-del Portillo,F. and Díaz-
811 Orejas,R. (2016) Stabilization of the virulence plasmid pSLT of *Salmonella* Typhimurium
812 by three maintenance systems and its evaluation by using a new stability test. *Front. Mol.*
813 *Biosci.*, **3**, 00066.

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

817 68. Rotger,R. and Casadesús,J. (1999) The virulence plasmids of *Salmonella*. *Int. Microbiol.*, **2**,
818 177–184.

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

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

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

826 72. Dagan,T., Blekhman,R. and Graur,D. (2006) The “Domino Theory” of gene death: gradual
827 and mass gene extinction events in three lineages of obligate symbiotic bacterial
828 pathogens. *Mol. Biol. Evol.*, **23**, 310–316.

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

831 74. Cole,S.T., Eiglmeier,K., Parkhill,J., James,K.D., Thomson,N.R., Wheeler,P.R., Honoré,N.,
832 Garnier,T., Churcher,C., Harris,D., *et al.* (2001) Massive gene decay in the leprosy
833 bacillus. *Nature*, **409**, 1007–1011.

75. McClelland,M., Sanderson,K.E., Clifton,S.W., Latreille,P., Porwollik,S., Sabo,A., Meyer,R., Bieri,T., Ozersky,P., McLellan,M., *et al.* (2004) Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat. Genet.*, **36**, 1268–1274.

76. Parkhill,J., Wren,B.W., Thomson,N.R., Titball,R.W., Holden,M.T.G., Prentice,M.B., Sebaihia,M., James,K.D., Churcher,C., Mungall,K.L., *et al.* (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.

77. Andersson,J.O. and Andersson,S.G.E. (2001) Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol. Biol. Evol.*, **18**, 829–839.

78. Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G., Rose,D.J., Darling,A., *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun*, **71**, 2775–2786.

79. Bennett,G.M. and Moran,N.A. (2015) Heritable symbiosis: the advantages and perils of an evolutionary rabbit hole. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 10169–10176.

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

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