A replication-centered phylogeny illuminates the evolutionary landscape of bacterial plasmids

Yosuke Nishimura[1*+], Kensei Kaneko[2+], Tatsuya Kamijo[2+], Nanako Isogai[2], Maho Tokuda[2,3], Hui Xie[4], Yusuke Tsuda[5], Aki Hirabayashi[6], Ryota Moriuchi[7], Hideo Dohra[7,8,9], Kazuhide Kimbara[2], Chiho Suzuki-Minakuchi[4], Hideaki Nojiri[4], Haruo Suzuki[10]*, Masato Suzuki[6]* and Masaki Shintani[2,8,11]*

[1]Research Center for Bioscience and Nanoscience (CeBN), Japan Agency for Marine Earth Science and Technology (JAMSTEC), 2-15 Natsushima-cho, Yokosuka-city, Kanagawa, 237-0061, Japan
[2]Department of Engineering, Graduate School of Integrated Science and Technology, Shizuoka University, 3-5-1 Johoku, Chuo-ku, Hamamatsu, Shizuoka, 432-8561, Japan
[3]Faculty of Science, Tokyo Metropolitan University, 1-1 Minamiosawa, Hachioji, Tokyo, 192-0397, Japan
[4]Agro-Biotechnology Research Center (AgTECH), Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo, 113-8657, Japan
[5]Department of Clinical Laboratory Medicine, Kyoto University Graduate School of Medicine, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8507, Japan
[6]Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Japan Institute for Health Security, 4-2-1 Aobacho, Higashi-murayama, Tokyo 189-0002, Japan
[7]Shizuoka Instrumental Analysis Center, Shizuoka University, 836 Ohya, Suruga-ku, Shizuoka, 422-8529, Japan
[8]Research Institute of Green Science and Technology, Shizuoka University, 3-5-1, Johoku, Chuo-ku, Hamamatsu, Shizuoka, 432-8561, Japan
[9]Department of Science, Graduate School of Integrated Science and Technology, Shizuoka University, 3-5-1, Johoku, Chuo-ku, Hamamatsu, Shizuoka, 432-8561, Japan
[10]Faculty of Environment and Information Studies, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
[11]Japan Collection of Microorganisms, RIKEN BioResource Research Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

†Joint authors

*To whom correspondence should be addressed.

Tel:+81-46-867-9084; Email: nishimuray@jamstec.go.jp (Yosuke Nishimura);
Tel:+81-466-47-5099; Email: haruo@sfc.keio.ac.jp (Haruo Suzuki);

Tel:+81-3-5285-1111; Email: suzuki.mas@jihs.go.jp (Masato Suzuki);
Tel:+81-53-478-1181; Email: shintani.masaki@shizuoka.ac.jp (Masaki Shintani)

SUMMARY

Plasmids are the most influential engines of bacterial evolution and horizontal gene transfer, fueling the global spread of traits such as antimicrobial resistance, yet their deep evolutionary relationships remain difficult to resolve because current classification schemes are constrained by host range and nucleotide similarity. Replication initiation proteins (RIPs), which govern plasmid persistence and diversification, also remain poorly annotated across public databases. Here we establish PInc, a curated and experimentally grounded replicon classification framework anchored in historically defined incompatibility groups of *Pseudomonas* plasmids. Homology searches beyond PInc revealed that most replication initiators analyzed here share a conserved winged-helix (WH) domain that interacts with DNA, defining a broad WH RIP superfamily. Using the conserved WH region, we reconstructed a large-scale phylogeny that linked WH RIPs to nearly one hundred thousand plasmids, representing approximately half of all plasmids, across public databases and resolved deep evolutionary structure, clade-specific host and environmental distributions, and substantial RIP diversity not captured by current typing tools or annotation schemes. This framework overcomes the historical host bias of plasmid typing and provides a replication-centered view of plasmid diversification across bacterial lineages and environments.

INTRODUCTION

Plasmids are among the most influential mobile genetic elements in bacteria, driving rapid evolutionary innovation through horizontal gene transfer[1,2]. Plasmids are recognized as major vehicles for the global spread of antimicrobial resistance genes (ARGs), representing a critical threat to public health[3–5]. As recent large-scale sequencing efforts have revealed an unexpectedly vast and heterogeneous collection of plasmids[6,7], a systematic framework is essential for organizing this complexity and for interpreting plasmid evolution in a coherent manner. Therefore, a standardized plasmid classification framework is urgently required.

A historical scheme known as incompatibility (Inc) groups provided the conceptual foundation for plasmid classification by exploiting phenotypic traits, defined by the inability of plasmids with similar replication systems to stably coexist in the same host[8]. More recently, sequence-based approaches have been developed that incorporate Inc group names into their classification schemes. Although sequence-based methods have improved classification coverage, they still fall short of delivering a macroevolutionary perspective. For example, replicon typing[9] assigns plasmids to groups based on nucleotide sequence homology of characteristic *rep* genes encoding replication initiation proteins (RIPs). Plasmid taxonomic units (PTUs)[10,11] cluster plasmids into taxon-like groups using whole-plasmid nucleotide similarity. Relaxase-based

76 approaches[10,12,13] have also expanded classification by grouping plasmids according to
77 the phylogeny of their relaxase proteins, which serve as key markers of conjugative and
78 mobilizable plasmids. Limitations arise because replicon- and whole-sequence-based
79 approaches rely solely on nucleotide similarity, which obscures remote homology
80 among deeply divergent plasmids. In contrast, the relaxase-based schemes exclude
81 more than half of plasmids, which lack a relaxase gene[8], and thus cannot serve as a
82 universal backbone for plasmid classification. Consequently, current classification
83 schemes do not fully resolve how plasmids are related across deep evolutionary time.

84 *Pseudomonas* plasmids provide an ideal starting point because this genus spans an
85 exceptionally wide ecological and clinical range from soils and aquatic environments to
86 hospital settings, and its plasmids serve as major vectors connecting environmental and
87 clinical antimicrobial-resistance (AMR) reservoirs[14]. Phenotype-based 14 Inc groups
88 (IncP-1 through IncP-14) have been designated for *Pseudomonas* plasmids and include
89 some of the most clinically significant resistance plasmids identified in *P. aeruginosa* [15,16].
90 However, despite their importance, some are not only under-characterized but also lack
91 nucleotide sequence information, relying solely on their original incompatibility
92 phenotypes.

93 We here constructed a curated and experimentally validated replicon classification
94 system, termed PInc (*Pseudomonas* Incompatibility-informed replicon classification),
95 which provides a sequence-based framework anchored in historical Inc groups. This
96 framework is grounded in the experimental identification of RIPs from representative
97 plasmids hosted by *Pseudomonas*. Keeping focus on RIPs, we identified conserved but
98 distantly homologous winged-helix (WH) domain within RIPs in most PInc groups and
99 other many plasmids, suggesting a common evolutionary origin of the RIPs. Building on
100 this observation, we reconstructed extended phylogeny of RIPs with the conserved WH
101 domain. The phylogeny classifies many yet unclassified plasmids and uncovers the
102 diversification and macroevolutionary trajectory of plasmids based on the replication
103 system.

104 **Results and Discussion**

105 **Sequencing and molecular characterization of *Pseudomonas* plasmids**

106 We determined the complete nucleotide sequences of six *Pseudomonas* plasmids:
107 Rms139 (IncP-2)[17], Rms163 (IncP-5)[18,19], Rsu2 (IncP-9)[20,21], RP1-1 (IncP-11)[22], R716 (IncP-
108 12)[23], and pMG26 (IncP-13)[24]. Although these plasmids represent the archetype
109 members of their respective Inc groups in *Pseudomonas*, no full-length reference
110 sequences had been available, leaving the replication systems of historically important
111 plasmids unresolved (Table 1). For IncP-10 plasmids, because only a partial plasmid

112 sequence of the representative plasmid R91-5[26,27] is available, we selected pPAB546 as
113 a representative (Supplemental Text S1). Sequencing of pMG26, previously assigned as
114 IncP-13 plasmid[28], yielded no circular contigs after removal of reads mapping to the *P.*
115 *aeruginosa* PAO1 chromosome. Detailed molecular characterization indicated that
116 pMG26 is not a plasmid but an integrative and conjugative element (ICE), and therefore
117 IncP-13 plasmids should be regarded as ICEs (Supplemental Text S1). The other five
118 (Rms139, Rms163, Rsu2, RP1-1, and R716) are shown to be self-transmissible plasmids
119 (Supplemental Text S1). The RIPs and origin of vegetative replication (*oriV*) of these
120 plasmids (Table 1) were identified by sequence- and structure-based searches, and their
121 functions were experimentally validated using mini-replicon constructs (see Methods,
122 Figure S1, S2, S3, and Supplemental Text S1).

123 **PInc classify unclassified plasmids in PLSDB**

124 To provide a biologically grounded and sequence-based framework for *Pseudomonas*
125 plasmids, we constructed a novel replicon-typing scheme named PInc, using
126 experimentally validated RIPs as reference sequences (Table 1). This scheme follows the
127 traditional Inc group numbering and its subgroup lettering (in Greek character) —for
128 instance, PInc-1α corresponds to IncP-1α, and PInc-2 to IncP-2. Because the original Inc
129 were defined on the basis of phenotypic incompatibility in different host lineages such
130 as *Pseudomonas* and Enterobacterales, their names often overlap each other. For
131 example, *Pseudomonas* IncP-1 is equivalent to Enterobacterales IncP (Table 1).
132 Introducing the PInc scheme allows us to retain meaningful correspondence with the
133 original Inc names while eliminating confusion caused by the overlapped or inconsistent
134 Inc labels[29].

135 Some Inc groups were excluded from PInc, because IncP-8[30] and IncP-13 (this study)
136 were identified as ICEs, or due to material unavailability for IncP-14[31]. Additionally,
137 certain *Pseudomonas* plasmid groups that are not included in the traditional Inc
138 grouping were incorporated into the PInc system, as they are widely distributed and
139 known to disseminate antimicrobial resistance. These include PInc-15 (PromA
140 group[32,33]), PInc-16 (pSN1216-29-like group[34]), PInc-17 (pQBR103-like group[35]), and
141 PInc-18 (pSTY-like group[36]) (Table 1, Supplemental Text S1). PInc subgroups were
142 defined according to subgroups of the Inc groups or newly defined if a group contains a
143 large nucleotide diversity of the RIP genes.

144 Plasmid sequences were collected and the RIP genes in the plasmids were identified
145 based on sequence and structural homology to known RIPs (Figures S1, S2). Reference
146 plasmids (n=87), including those newly sequenced in this study, were selected to cover
147 the 15 original Inc groups and 22 subgroups (Table S1). Each PInc group and its

148 subgroups were defined as follows: a plasmid was assigned to a group if its RIP gene
149 showed nucleotide similarity to a reference RIP (BLASTn; ≥90% identity and ≥60% length
150 coverage), considering only the best hit. Plasmids harboring multiple distinct RIPs were
151 recognized as multi-replicons (e.g., PInc-10 and PInc-11).

152 Accordingly, we assigned RIPs of 2,351 plasmids in the PLSDB (59,895 plasmids) to PInc
153 (Tables S2), including 530 plasmids unclassified by the existing tools, PlasmidFinder and
154 1,514 unclassified by MOB-suite. Thus, PInc substantially expands the coverage of
155 replicon-based classification for *Pseudomonas*-associated plasmids, increasing the
156 number of classifiable plasmids by 29.1 % for PlasmidFinder and 181% for MOB-suite.
157 Notably, PInc-classified plasmids span not only *Pseudomonas* but also diverse
158 *Enterobacterales* hosts, reflecting the broad host range of several PInc plasmids
159 (https://doi.org/10.6084/m9.figshare.26768173). Overall, PInc provides a systematic
160 sequence-based framework that improves classification depth and coverage across
161 plasmids in PLSDB.

162 **Comparative genomics analysis highlights PInc group characteristics**

163 To characterize the genomic and evolutionary diversity of plasmids classified by PInc, we
164 conducted a comprehensive pangenomic analysis across the PInc groups (Tables S2, S3-
165 1~S3-14). Three key patterns emerged. First, several PInc groups (notably PInc-2, PInc-
166 17, and PInc-18) exhibited exceptionally large genome sizes (Figure 1A). Second, specific
167 groups including PInc-4, PInc-10 and PInc-11, displayed distinctive multi-replicon
168 coupling patterns, suggesting long-term modular evolution of their replication systems.
169 Third, PInc-1 stood out for its broad genomic diversity consistent with its presence
170 across diverse environments and host lineages.

171 Plasmid size distribution across PInc groups (Figure 1A) revealed that PInc-2, PInc-17,
172 and PInc-18 harbor markedly larger plasmids than other groups, indicating expanded
173 accessory gene content and potential ecological specialization. %GC content (Figure 1B)
174 varied widely in PInc-1 and PInc-4, suggesting flexible amelioration in multiple hosts [37–
175 39], whereas other groups showed narrow %GC distributions consistent with more
176 specialized ecological niches. Core-genome analysis (Figure 1C) showed that PInc-2,
177 PInc-12, and PInc-17 maintain unusually rich core sets compared with other groups,
178 implying conserved functional backbones within these groups. In contrast, the low core-
179 gene proportion in PInc-1 reflects its high genomic and ecological heterogeneity.

180 Multi-replicon analysis (Figure 1D, Table S2) revealed substantial variation among PInc
181 groups. More than half of PInc-4 (70.9%), PInc-11 (65.2%), PInc-6 (60.5%), and PInc-10
182 (51.0%) plasmids are multi-replicons, whereas more than 90% of PInc-1, PInc-5, PInc-7,

183 PInc-9, and PInc-15 plasmids are single-replicon. These rates indicate a distinct
184 evolutionary history of the replicon acquisition and coevolution. Specifically, PInc-4
185 plasmids are coupled to many types of replicons (Figure 1D, Table S2). In contrast, most
186 multi-replicons of PInc-10 and PInc-11 were coupled to each other or to only a few types
187 of replicons (IncU or rep_cluster_2025, Table S2). Importantly, more than 60% of the
188 PInc-2, PInc-3, PInc-4, PInc-6, PInc-10, and PInc-17 plasmids harbored ARGs, with several
189 of these groups, PInc-2, PInc-6, PInc-10, and PInc-17 plasmids, particularly containing
190 clinically important ARGs, such as carbapenem, tigecycline, and colistin resistance genes
191 (Figure 1D). More detailed pangenomic characterization for each PInc was provided in
192 Supplemental Text S1.

193 **Homology-guided collection of RIP sequences expands the diversity of plasmid**
194 **initiators**

195 To investigate the evolutionary relationships of RIPs, we first assessed protein sequence
196 homology among 251 non-redundant RIP sequences from PInc plasmids using BLASTp
197 (Figure S3). Several PInc groups contained substantial sequence diversity, including PInc-
198 1, PInc-4, and PInc-15, in which the most divergent RIP pairs shared only 34.05%, 28.81%,
199 56.14% identity, respectively. By contrast, RIPs in the remaining groups were less diverse,
200 with >73% identity within each PInc group (Table S4). Sequence homology between RIPs
201 from different groups was also detected for specific combinations, including PInc-
202 2/PInc-3/PInc-17, PInc-1/PInc-12, and PInc-10/PInc-11, suggesting evolutionary
203 relatedness among these plasmid groups (Figures S3 and S4).

204 To extend the analysis beyond the sensitivity of BLASTp, we next performed iterative
205 sequence searches using RIPs from the PInc groups as queries to collect more diverse
206 homologs from PLSDB and the UniProt Archive (UniParc). To improve coverage of RIP
207 sequence diversity, we also included nine additional RIPs outside the PInc groups as
208 additional queries (Table S5). The detailed procedures are described in Supplemental
209 Text S1 and Figure S4. In total, this approach retrieved 195,236 sequences, which were
210 classified into 14 homologous groups.

211 **Remote homology analysis reveals evolutionary divergence among plasmid RIPs**

212 Remote homology among RIPs from the 14 homologous groups was assessed using
213 HHblits[40], which performs sensitive sequence-based homology detection through HMM-
214 HMM comparison. With the exception of G-PInc-6, which comprises sequences
215 collected using PInc-6 RIPs as queries, RIPs from the remaining 13 groups shared a
216 homologous region (Figure S5). Based on the significance of homology among groups,
217 these 13 groups could be further divided into two supergroups. One supergroup,

218 comprising seven groups, showed homology with >95% probability in 20 of 21 possible
219 pairwise group comparisons, whereas the other, comprising six groups, did so in all 15
220 possible comparisons (Figure S5).

221 Considering available experimental structures and AlphaFold3[41] predictions of RIPs
222 (Figure S6), we classified these two supergroups as single winged-helix (sWH) and
223 double winged-helix (dWH) according to their domain architecture. The sWH
224 supergroup comprises seven groups with a single WH domain, whereas the dWH
225 supergroup comprises six groups with two WH domains. In dWH RIPs, available
226 structures of RepE (plasmid F; PDB: 2Z9O), Pi (R6K; PDB: 2NRA), RepA (pPS10; PDB:
227 1HKQ), and TrfA (RK2/RP4)[42] all contain two conserved winged-helix domains, here
228 termed WH1 and WH2. By contrast, only two crystal structures are available for sWH
229 RIPs[43] (pSK41, PDB: 5KBJ; pTZ2162, PDB: 4PT7), although AlphaFold3 predicts that the
230 remaining members also adopt a single-WH-domain architecture. HHblits detected
231 significant homology (>80% probability) between the sWH and dWH supergroups in 18
232 of 42 possible pairwise group comparisons (6 × 7). In all cases, the homologous region
233 mapped to the WH domain of sWH RIPs and to WH2, the C-terminal WH domain, of
234 dWH RIPs.

235 By contrast, RIPs of G-PInc-6 showed no detectable homology to any other groups by
236 HHblits, with probabilities of less than 1% in all comparisons. To determine their
237 evolutionary origin, we next searched against Pfam database using HHblits. G-PInc-6
238 RIPs matched PF03090, corresponding to the replicase family, with 98.4% probability.
239 This domain is present in the ColE2 RIP[44], a member of the archaeo-eukaryotic primases
240 (AEPs)[45] superfamily. Consistent with this assignment, G-PInc-6 RIPs[46,47] also contain a
241 PriCT domain implicated in DNA unwinding[44]. Together, these observations identify G-
242 PInc-6 RIPs as AEP-superfamily initiators and therefore as evolutionarily distinct from
243 the WH-containing RIPs.

244 Based on this shared homologous domain across the sWH and dWH supergroups, we
245 define these proteins as a WH RIPs superfamily. A previous structure-based study
246 reported deep evolutionary relationships among helix-turn-helix (HTH) domains,
247 including WH domains[48], although plasmid RIPs were not examined in this context.
248 Moreover, the N-terminal WH1 domain of dWH RIPs has been proposed to be
249 evolutionarily related to the WH domain of archaeal Orc1/Cdc6 and eukaryotic Orc
250 proteins[49], both of which function in origin recognition for the chromosomal
251 replication[50]. By contrast, the evolutionary origin of WH2 has remained unresolved. Our
252 analysis identifies homology between WH2 and the WH domain of sWH RIPs, uncovering
253 an unexpected evolutionary link within plasmid replication initiators. Notably, the WH2
254 domain of pPS10 RepA binds the 3' end of the iteron and is the major determinant of

255 sequence specificity[45], suggesting that this conserved WH module represents a core
256 origin-recognition element across the WH RIP superfamily. More broadly, the
257 coexistence of such divergent RIPs lineages, including both WH RIPs and AEP-
258 superfamily initiators, within the traditional Inc framework exposes the limitations of
259 historical phenotypic classification and supports the need for a sequence-based PInc
260 system.

261 **WH RIP phylogeny resolves RIP diversity across from nearly 100K plasmids**

262 A phylogeny of WH RIPs from the 13 homologous groups was reconstructed using the
263 shared homologous region identified by HHblits (Figure 2A). The maximum likelihood
264 tree was built from 4,769 representative sequences clustered at 50% identity, selected
265 from RIPs identified in 34,724 PLSDB plasmids and 95,057 UniParc proteins (Figure S4B).
266 On the basis of branch support, we defined eight clades (A to H) and 42 subclades using
267 high-confidence thresholds (UFBoot2: ≥95% or SH-aLRT: ≥80%) (Figure 2B). At the
268 deepest level, sWH and dWH RIPs formed two superclades separated by a major split in
269 the tree, supported by both UFBoot2 (89.4%) and SH-aLRT (88%). This split was also
270 consistently identified as a phylogenetic root by minimal ancestor deviation analysis[51].
271 Together, these results suggest an early evolutionary transition in WH RIPs, involving
272 either acquisition of an additional WH domain or loss of one ancestral WH domain (i.e.
273 WH1 of dWH RIPs).

274 To place this phylogeny in the context of plasmid diversity, we surveyed associated
275 plasmids in the public plasmid databases PLSDB and IMG/PR. PLSDB is enriched in
276 plasmids from cultivated isolates, whereas IMG/PR contains many metagenome-derived
277 plasmids. Protein sequences predicted from 214,575 presumably complete plasmids
278 (59,895 from PLSDB and 154,680 from IMG/PR) were mapped to the closest leaves in
279 the WH RIP tree (Figure 2C). This analysis identified 118,302 phylogeny-linked RIPs
280 associated with 99,395 plasmids. These included 14,745 plasmids (14.8% of phylogeny-
281 linked plasmids) identified as multi-replicons carrying multiple WH RIPs (Figure 2A).
282 Most of these were derived from Gammaproteobacteria, although other taxa, including
283 Bacillota, Actinomycetota, Bacteroidota, and Campylobacterota, were also represented
284 (Figure 2D). Of the 118,302 phylogeny-linked RIPs, 62,046 (52.4%) from 58,878 plasmids
285 were newly identified and were not detected by existing replicon typing tools, including
286 PlasmidFinder, MOB-suite, and AcinetobacterPlasmidTyping[52], or by the PInc framework
287 defined here. Furthermore, 28,951 RIPs (24.5%) from 25,675 plasmids lacked any Pfam
288 domain annotation (Figure 2E), and in clades A, B, and F, more than half of the RIPs were
289 unannotated. Together, these results indicate that a substantial fraction of RIP diversity
290 remains uncaptured by current annotation frameworks.

291 The distribution of plasmids across the tree was highly uneven. Only 2,081 leaves
292 (43.6%) were linked to plasmids, and just 26 leaves were associated with more than
293 1,000 plasmids each. Together, these accounted for 61,767 plasmids (62.1% of
294 phylogeny-linked plasmids). Mapping of PInc and other replicon types using
295 PlasmidFinder and AcinetobacterPlasmidTyping further showed that annotated replicon
296 types were broadly distributed across clades (Figure 2B). RIPs from 2,205 PInc plasmids
297 mapped to 17 leaves spanning seven clades except for clade D. In addition, 38, 32, and
298 32 leaves were associated with gram-positive, Enterobacterales, and *Acinetobacter*
299 replicon types, respectively. In total, only 118 leaves (2.5% of all 4,769 leaves) were
300 linked to current typing schemes, indicating that much of the RIP diversity remains
301 uncaptured by existing plasmid typing frameworks.

302 **Structural and sequence diversification across WH RIP clades.**

303 We examined the structural diversity of representative WH RIPs from PInc plasmids,
304 focusing on domain organization and predicted three-dimensional structures (Figure 3
305 and Figure S7). Pfam-based domain annotation, together with WH1/WH2 assignments
306 based on previous studies[42,53] (see Methods), identified WH domains linked to
307 conserved replication-associated domains, including RepA_N/C, Rep3_N/C, IncFII_repA.
308 The Bac_RepA_C domain, which does not overlap the WH region, has been implicated
309 in interactions with DnaG primase[43]. Based on available crystal structures, we inferred
310 that sWH RIPs bind DNA as monomers, whereas dWH RIPs bind DNA as dimers. These
311 observations indicate that WH RIPs combine a conserved WH core with distinct
312 replication-associated domain architecture across clades.

313 Sequence features derived from the multiple alignment used for phylogenetic
314 reconstruction further distinguished clades A–H (Figure S8). In particular, both gap
315 patterns in the alignment and the lengths of the N- and C-terminal regions flanking the
316 conserved WH domain showed clade-specific features, while still exhibiting substantial
317 variation within individual clades, such as clades C and E. Overall, these results highlight
318 the structural diversification of Rep proteins underlying plasmid evolution.

319 **WH RIP phylogeny highlights evolutionary constraints for plasmid dispersion.**

320 We analyzed the taxonomic distribution of WH RIP clades using metadata from PLSDB
321 and IMG/PR. PLSDB follows NCBI taxonomy, whereas IMG/PR follows GTDB taxonomy
322 (Figure 4A). Despite differences in the taxonomic systems and data sources, several
323 clade–taxonomy associations were consistent across both datasets. Clades B and D were
324 strongly associated with Bacillota, clade C with Gammaproteobacteria, and clade E with
325 Alphaproteobacteria. Clade H showed the broadest taxonomic distribution and, in
326 IMG/PR, accounted for most Bacteroidota and large fractions of Campylobacterota and

327 Fusobacteriota. This was particularly evident in IMG/PR, which contained 22,359
328 Bacteroidota plasmids, most of which were assigned to clades A and H. Clade G also
329 showed a broad host range, with strong representation in Actinomycetota. These
330 patterns indicate that WH RIP evolution is structured by host lineage, while some clades
331 have undergone broader host transitions than others.

332 For environmental distribution, we analysed IMG/PR-linked plasmids, for which
333 environmental metadata were available from metagenomic and metatranscriptomic
334 datasets. Environmental distributions were non-random across WH RIP clades and were
335 not fully explained by host taxonomy alone (Figure 4B). Clade H was broadly distributed
336 and accounted for a major fraction of RIPs across most environments, whereas clade
337 composition varied substantially among habitat categories. Host-associated
338 environments, including human, mammals and built environment, were dominated
339 primarily by clades H and A. By contrast, plants, soil, freshwater and bioreactor samples
340 showed more diverse clade mixtures, with increased contributions from clades C, E, F
341 and G. These observations indicate that plasmid ecological distribution is shaped not
342 only by host lineage, but also by clade-specific constraints on dispersal.

343 **Further evolutionary insights into the eight clades.**

344 The eight WH RIP clades also differed markedly in plasmid size, RIP length and GC
345 composition. Clade A was associated with the smallest plasmids and the shortest RIPs,
346 with median plasmid length and RIP length of 3.2 kb and 185 aa, respectively (Figure 4C).
347 Clade H also tended to occur in relatively small plasmids (median 6.8 kb), although its
348 RIPs were substantially longer than those of clade A (median 325 aa). By contrast, clades
349 C, E and F were enriched in larger plasmids, with median lengths of 98.5, 112.6 and 143.9
350 kb, respectively, whereas clade B was also associated with relatively large plasmids
351 despite its limited taxonomic range. Clade G showed particularly broad size variation,
352 spanning both small and very large plasmids, indicating substantial architectural
353 diversity within this clade.

354 Clear clade-specific differences were also observed in GC composition. Clades B and D
355 were associated with the lowest-GC plasmids, which may partly reflect their taxonomic
356 bias towards Bacillota, whereas clades E and G showed the highest plasmid GC values,
357 with medians of 52.6% and 60.6% (Figure 4D). RIP GC content broadly tracked plasmid-
358 wide GC content, but clade C was distinctive in showing substantially higher GC in RIPs
359 than in their plasmid backgrounds (median 57.0% versus 51.5%). A similar but weaker
360 tendency was also evident in clades E and G. In addition, N-ARSC, defined here as the
361 mean number of nitrogen atoms in amino acid side chains, was markedly elevated in
362 clade C RIPs relative to the plasmid proteome as a whole (median 0.565 versus 0.396)

363 (Figure 4E). This difference is likely explained by enrichment of positively charged
364 residues, particularly lysine and arginine, and may reflect compositional adaptation
365 related to DNA binding (Figure S8). Together, these patterns indicate that the eight
366 clades occupy distinct regions of plasmid architectural and compositional space,
367 reflecting long-term divergence in both replication systems and their genomic contexts.

368 Accessory functions also differed sharply among WH RIP clades. MOBP and MOBQ were
369 detected across most clades, but other MOB classes were much more clade-specific
370 (Figure 4F). Notably, clade A showed a remarkably high prevalence of MOBV despite
371 being associated with relatively small plasmids, suggesting that compact plasmid
372 architecture does not preclude efficient mobilization potential. MPF repertoires also
373 varied among clades: some MPF types were broadly distributed, whereas others showed
374 clear clade bias, and MPF modules were nearly absent from clade A. AMR and defense
375 systems were similarly unevenly distributed, showing marked clade-to-clade variation
376 both in the proportion of plasmids carrying these functions and in the number of genes
377 encoded per plasmid (Figure 4G).

378 **Integrated evolutionary landscape of the eight WH RIP clades**

379 To summarize the major features associated with each lineage, we integrated the
380 structural, compositional, taxonomic and ecological characteristics of the eight WH RIP
381 clades (Figure 5). This comparison shows that the major split between the sWH and dWH
382 superclades is accompanied by broad differences in domain organization and plasmid
383 architecture, whereas individual clades are further distinguished by characteristic
384 combinations of RIP length, GC composition, host distribution and environmental
385 association. Notably, clades G and H, as well as clades B, C and D, form well-supported
386 higher-level assemblages, suggesting that these sets of clades represent coherent
387 evolutionary subdivisions within the WH RIP phylogeny. Together, these patterns
388 support the view that the eight WH RIP clades represent distinct evolutionary
389 trajectories of plasmid replication systems.

390 **Conclusion**

391 In this study, we establish PInc, a curated and experimentally grounded replicon
392 classification framework anchored in historically defined incompatibility groups of
393 *Pseudomonas* plasmids. By coupling experimental identification of RIPs with sequence-
394 based classification, PInc expands the repertoire of identifiable plasmid replicons and
395 provides a framework for plasmid classification through their replication systems.

396 Homology searches beyond PInc plasmids revealed that most replication initiators
397 analyzed here share a conserved WH domain, defining a WH RIP superfamily. Further

398 remote homology analysis resolved this superfamily into sWH and dWH superclades and
399 uncovered homology between the WH domain of sWH RIPs and the WH2 domain of
400 dWH RIPs, suggesting a shared origin-recognition module across a broad range of
401 plasmids. By contrast, G-PInc-6 RIPs were assigned to the AEP superfamily, showing that
402 plasmids grouped historically by incompatibility can encode fundamentally distinct
403 classes of replication initiators.

404 Using this conserved WH region, we reconstructed a large-scale phylogeny that linked
405 WH RIPs to nearly 100,000 plasmids across public databases. This phylogeny resolved
406 deep evolutionary structure, revealed clade-specific host and environmental
407 distributions, and captured a substantial fraction of RIP diversity not detected by existing
408 typing tools or current annotation schemes. Notably, several lineages not covered by
409 current plasmid typing systems, including Bacteroidota, Campylobacterota,
410 Thermodesulfobacterota and Fusobacterota, could be placed within the WH RIP
411 phylogeny. In this way, our approach overcomes the historical host bias of plasmid
412 classification and connects plasmids from both well-studied and environmental bacteria
413 within a single phylogenetic framework. More broadly, this phylogeny illuminates
414 relationships among WH RIPs beyond lineage-restricted typing systems and provides a
415 basis for understanding the evolutionary trajectories of host transitions and replication-
416 system diversification in plasmids.

417 At present, this evolutionary landscape is limited to plasmids carrying WH RIPs.
418 Extending the same framework to additional initiator classes, including AEP-type and
419 rolling-circle replication proteins, should further broaden the phylogenetic landscape of
420 plasmids and provide a route towards a more comprehensive evolutionary map of
421 bacterial mobile DNA.

422 **Methods**

423 **Bacterial strains, plasmids, and culture conditions**

424 The bacterial strains and plasmids used are listed in Table S6. *E. coli* and *Pseudomonas*
425 strains were cultured in Luria broth (LB) or M9 broth[54] at 30°C (*P. putida* KT2440) or 37°C
426 (*P. aeruginosa* PAO1 and its derivatives). LB plates containing 1.5% agar were used for
427 filter mating. Ampicillin (Ap, 50 μg/mL), piperacillin (Pip, 16 μg/mL), rifampicin (Rif, 50
428 μg/mL), streptomycin (Sm 150 μg/mL), and tetracycline (Tc, 50 μg/mL for the others)
429 were added to the medium.

430 **DNA manipulation**

431 Total DNA from the bacterial strains was extracted using a NucleoSpin Tissue Kit
432 (TAKARA BIO, Japan). Polymerase chain reaction (PCR) was performed on a T100 thermal

433 cycler (Bio-Rad Laboratories, USA) using the KOD One PCR Master Mix (TOYOBO, Japan).
434 The primers used in this study are listed in Table S7. The following amplification
435 conditions were used: 30 cycles at 98°C for 10 s, 67°C for 5 s, and 68°C for 1 s. The pUC-
436 GW plasmid with DNA fragments including putative RIP genes and *oriV*s of each plasmid,
437 except those for RP1-1 and pMG26, was synthesized by the custom service of Azenta
438 Life Science, USA. Restriction enzymes (New England Biolabs, USA or TAKARA BIO), the
439 HiYieldTM Gel/PCR DNA fragments Extraction kit (RBC Bioscience, Taiwan), NEBuilder
440 Hifi DNA Assembly system (New England Biolabs), and competent *E. coli* JM109 and
441 DH5α cells (RBC Bioscience) were used for cloning the DNA fragments. Plasmid DNA
442 from the bacterial strains was extracted using the NucleoSpin Plasmid EasyPure Kit
443 (TAKARA BIO). Electroporation to introduce the plasmids into different bacterial strains
444 was performed using a MicroPulser electroporator (Bio-Rad Laboratories). All other
445 procedures were performed according to standard methods[54].

446 **Complete sequencing of the plasmids by next-generation sequencing**

447 The nucleotide sequences of the plasmid DNAs were determined as follows: whole
448 genomic DNA extracted from their host (*P. aeruginosa* PAO1808, a derivative strain of
449 PAO1[55]) was sequenced using the HiSeq2500 platform (Illumina, USA). Trimmed high-
450 quality short reads (read length >140 bp and quality score >15) were assembled using
451 SPAdes v3.14.0 using the plasmid option[56]. If circular contig(s) could not be found, the
452 host chromosomal DNA was removed by mapping the resultant contigs on the host
453 genome sequences [their deposited sequences, that is, NCBI RefSeq accession number
454 NC_002516 (*P. aeruginosa* PAO1)] using Geneious Prime software v2020.01[57].
455 Subsequently, the plasmid reads were extracted from the remaining contigs using
456 SAMtools v1.7[58] and SeqKit v0.8.0[59] and reassembled using SPAdes[60]. For RP1-1 and
457 pMG26, gaps in plasmids were closed by PCR with the designed primer sets based on
458 the ends of contigs and Sanger sequencing of the PCR products.

459 **Gene annotation of plasmids**

460 Gene annotations of the plasmids were performed using DFAST-core v1.2.5[61] with an in-
461 house database of plasmid sequences collected from NCBI RefSeq and then corrected
462 manually. Synteny analyses of plasmids were performed using Easyfig v2.2.2[62]. Other *in*
463 *silico* analyses were performed using Geneious Prime v2020.01. The annotated plasmid
464 sequences were deposited under the following GenBank accession numbers: LC685025,
465 LC685026, LC685027, LC685593, and LC700336.

466 **Detection and phenotype identification of ARGs**

467 ARG sequences were detected using Staramr v0.10.0 and ResFinder database v2.2.1
468 updated on 2023-10-27 (https://bitbucket.org/genomicepidemiology/resfinder_db)[59,60]
469 with the following criteria: ≥ 90% identity and ≥ 60% coverage. Clinically important ARGs,
470 including genes conferring resistance to carbapenem, colistin, or tigecycline, in Gram-
471 negative bacteria were further detected based on the phenotype information of ARGs
472 in AMRFinderPlus v3.11.26 updated on 2023-11-16
473 (https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/)[61].

474 **Prediction of RIP genes and *oriV*s of each plasmid**

475 To identify the genes encoding the RIP and origin of vegetative replication (*oriV*) regions
476 of Rms139, Rms163, RP1-1, and R716, two prediction approaches were used; after
477 coding sequence (CDS) annotations using DFAST-core, protein.faa files generated by
478 DFAST-core were used for the following: (i) nucleotide sequence-based searches were
479 performed using BLASTn; (ii) amino acid sequence-based protein structure prediction
480 and the structure-based searches were performed as described later. The GC-skew value,
481 which has been commonly used to measure the asymmetry in the composition of two
482 nucleotides, guanine (G) and cytosine (C), along a DNA strand and is defined as (G-
483 C)/(G+C), of the plasmids was used by Webskew
484 (https://genskew.csb.univie.ac.at/webskew) to detect the changes in the GC skew,
485 which indicates the DNA replication origins or termini[63,64]. Subsequently, the nucleotide
486 or amino acid sequences of putative CDSs around the above GC-skew shift point(s) were
487 subjected to methods (i) or (ii). The putative promoter region of the RIP gene was
488 predicted using BPROM
489 (http://www.softberry.com/berry.phtml?topic=bprom&group=help&subgroup=gfindb)
490 [65]. Subsequently, the predicted RIP and *oriV* sequences of each PInc groups were cloned
491 into the pUC-GW-Amp vector (Azenta Life Science). Transformation of *P. aeruginosa*
492 PAO1808 was performed with each of the resultant plasmids (at most 472 ng/µL), and
493 the cultures were spread onto LB+Pip plates. After incubation at 37ºC overnight,
494 colonies on the plates were isolated. Next, the presence of the mini-plasmid was
495 confirmed by PCR with specific primers after extraction of total DNA of transformants.

496 **Protein structure prediction of RIPs**

497 AlphaFold3-mediated protein structure prediction was performed using the amino acid
498 sequences of the RIP gene products in representative *Pseudomonas* plasmids
499 (https://doi.org/10.6084/m9.figshare.26778898) on the AlphaFold server[41]. A structure-
500 based protein homology search using the predicted structural models was performed
501 for the PDB100 database on the FoldSeek server in 3Di/AA mode[66]. The predicted and
502 experimental structures were visualised and compared using PyMOL v3.1.3.1. The

503 predicted structures were coloured based on the root mean square deviation (RMSD)
504 from the reference experimental structures, i.e., the Pi or Rep19A[43,67] proteins, which
505 are RIPs of the R6K and pCA-347 plasmids, respectively, using the ColorByRMSD script
506 (https://pymolwiki.org/index.php/ColorByRMSD). The distances between aligned C-
507 alpha atom pairs were stored as B-factors of these residues, which are coloured using a
508 colour spectrum, with blue specifying the minimum pairwise RMSD and red indicating
509 the maximum. Unaligned residues were coloured grey.

510 **Identification, retrieval, and re-annotation of plasmids from public databases**

511 Complete plasmid sequences were downloaded in FASTA format using the Entrez
512 Programming Utilities (E-utilities; https://www.ncbi.nlm.nih.gov/books/NBK25501/)
513 from a public database (PLSDB 2023_11_23_v2)
514 (https://doi.org/10.6084/m9.figshare.26768173)[7,68]. To classify the plasmids in the
515 PLSDB based on their rep sequence similarity, the RIP gene sequences were detected
516 using Staramr v0.10.0 (https://github.com/phac-nml/staramr) and rep databases of
517 MOB-suite v3.1.8 updated on 2023-12-12 (https://github.com/phac-nml/mob-suite)
518 (https://doi.org/10.6084/m9.figshare.26768170)[69], PlasmidFinder v2.1 updated on
519 2023-12-04 (https://bitbucket.org/genomicepidemiology/plasmidfinder_db)
520 (https://doi.org/10.6084/m9.figshare.26778199)[70], our custom-made Pseudomonas
521 plasmid RIP gene database (termed repP and PseudomonasRepDB database,
522 https://doi.org/10.6084/m9.figshare.31883911 and
523 https://doi.org/10.6084/m9.figshare.26778175), and our custom-made plasmid RIP WH
524 database (termed WHRepDB database, https://doi.org/10.6084/m9.figshare.31883887)
525 with the following criteria: ≥90% identity and ≥60% coverage. The Tus protein, known
526 for binding to DNA replication terminus (ter) sites of replicons, encoded by
527 rep_cluster_1115 in the MOB-suite database, is not included in our custom-made RIP
528 reference sequences. The repP database was constructed using experimentally
529 identified RIP genes from the 87 reference *Pseudomonas* plasmids shown in Table S1.
530 Furthermore, duplicated RIP sequences between the MOB-suite and repP databases
531 (Table S6) and incorrect RIP sequences, including uncertain IncP and rep_cluster_1115,
532 were manually removed from the MOB-suite database. The modified MOB-suite
533 database (https://doi.org/10.6084/m9.figshare.26768167) was used to detect multi-
534 replicon *Pseudomonas* plasmids, which are single plasmids that can belong to more than
535 one group other than the PInc groups. Among the candidate PInc-1 (IncP-1) plasmid
536 sequences based on the similarity to the RIP sequences, artificially constructed
537 sequences not explicitly labelled as such ("other entries;other sequences;artificial
538 sequences") in the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy)
539 were also present. We identified six PInc-1 plasmids (PD584_L3, PD689_delTT1, pJM105,

540 pJM50, pRK2013, and pUZ8002) as artificially constructed sequences based on a
541 literature survey and removed them from the putative PInc-1 plasmid set[71].

542 For the final dataset, the sequence length (bp) and GC content (%, %GC) were calculated
543 using SeqKit v2.3.0 (https://bioinf.shenwei.me/seqkit/)[59]. Plasmid sequences belonging
544 to each PInc groups were re-annotated using DFAST-core, with a reference genome from
545 each group to ensure consistency in the functional annotations.

546 **Pangenome analysis of plasmids**

547 A pangenome, i.e., the entire set of protein genes from all members of a defined group
548 (all the plasmids of the PInc groups), was analyzed using PIRATE[72]
549 (https://github.com/SionBayliss/PIRATE). Here, we used DIAMOND v0.9.14[73] rather
550 than BLAST[74] to identify homologous genes (protein families) in PIRATE. PIRATE was run
551 on default parameters with a high-scoring segment pair (HSP) query length threshold of
552 0.5, a sequence similarity step range of 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 98%,
553 and an inflation value of 2 in MCL (v14.137), a Markov cluster algorithm
554 (https://micans.org/mcl/). The core genes in each PInc group were those that were
555 present in more than 95% of the plasmids included in the pangenome analysis. Statistical
556 computing and result visualization were performed using R v4.1.3 (https://www.r-
557 project.org/). The output of SeqKit and PIRATE are available in Figshare
558 (https://doi.org/10.6084/m9.figshare.26789416).

559 **Homolog collection and phylogenetic analysis of RIPs**

560 Detailed methods are provided in Supplemental Text S1 and illustrated in Figure S6. In
561 brief, to analyze the evolutionary relationships of RIPs, we identified 2,374 RIPs that
562 originated from 2,332 PInc plasmids. Subsequently, 251 non-redundant proteins at
563 100% identity were extracted using MMseqs cluster v15-6f452[75]. Sequence homology
564 between the non-redundant proteins was assessed using BLASTp[74], and the resultant
565 hits were visualized via CPAP (https://github.com/yosuken/CPAP). To detect remote
566 homology between PInc groups and collect comprehensive RIP sequences, we extended
567 the sequence diversity through a multistep homology search process. First, we searched
568 ORFs of PLSDB plasmids predicted using orfM (n = 20,331,985)[76] for homologous RIPs
569 using hmmsearch of HMMER v3.3.2[77]. In addition, 8 additional RIPs sequences that are
570 distant from RIPs were manually picked up and homology searches using diamond blastp
571 against the ORFs of PLSDB plasmids were performed. This process generated 19
572 homologous groups of RIPs. Second, UniParc proteins (downloaded on Feb 2024; n =
573 607,912,929) were searched with hmmsearch using 19 HMMs generated from the
574 homologous groups. We merged the UniParc sequences into 19 homologous groups,
575 and filtered out potential contamination of other functional proteins, including various

576 transcriptional factors, using hmmsearch against the Pfam database. After merging the
577 19 homologous groups into 14 based on the UniParc sequence overlapping, we selected
578 5,488 representative sequences for the 14 homologous groups. To investigate whether
579 these groups exhibit homology with each other, a sensitive hidden Markov model
580 HMM–HMM comparison was performed using HHblits v3.3.0[78]. We extracted the
581 homologous regions of the alignments of the 13 groups (containing 6,793 sequences in
582 total). These regions were aligned using MAFFT v7.455[79]. After removing the gappy
583 sequences in the alignment, a maximum-likelihood phylogeny of WH RIPs was
584 reconstructed using the alignment and IQ-TREE v2.2.2.6[80]. We used the parameters "--
585 perturb 0.2 --nstop 500", following the recommendation of the reference manual of the
586 software, in case that many short sequences are included. The best fit model
587 (Q.pfam+F+R10) was selected using ModelFinder[81]. We evaluated the supporting values
588 of the best tree with UFBoot2[82] and SH-aLRT[83] via IQ-TREE. We discarded the abnormally
589 long branches using treeshrink[84]. Overall, the tree structure of sWH RIPs and dWH RIPS
590 exhibits a single, highly supported split (UFBoot2: 89.4%, SH-aLRT: 88%). Since our
591 assumption is that this split represents the root of the tree and corresponds to an
592 ancestral event that led to either the acquisition or the loss of a WH domain, the
593 resultant tree was provisionally rooted by the sWH-dWH split using gotree v0.4.0[85]. The
594 tree-associated PLSDB plasmids, UniParc proteins, and IMG/PR[6] plasmids were
595 identified based on sequence homology search results. Taxonomic affiliation was
596 assigned based on PLSDB plasmids and UniParc proteins, and environmental affiliation
597 was based on IMG/PR plasmids. A detailed information related to the phylogenetic tree
598 is available in Figshare (https://doi.org/10.6084/m9.figshare.28837559).

599 **Data Availability**

600 The annotated plasmid sequences have been deposited in the GenBank/EMBL/DDBJ
601 under accession numbers LC685025, LC685026, LC685027, LC685593, and LC700336.

602 **Supplementary Data Statement**

603 Supplementary Data is available online.

604 **Contributions**

605 YN, HS, MS, MS conceived, designed, and supervised the study. KK, TK, NI, MT, HX, YT,
606 AH, RM, HD, MS, HS, MS performed the experiments and data analysis. YN, RM, HD,
607 KK, HN, MS, HS, MS wrote, reviewed, and edited the manuscript. All authors
608 contributed to manuscript revision, read and approved the submitted version.
609

610 **Acknowledgements**

611 We are grateful to Prof. Dr. S. Iyobe of the Gumma University School of Medicine and
612 Prof. Dr. M. Tsuda of Tohoku University for providing Rms139, Rms163, RP1-1, R716,
613 and pMG26. Computation time was provided by the Super Computer System, Institute
614 for Chemical Research, Kyoto University.

615 **Funding**

616 This work was supported by grants (JP24wm0225029 to Y. Tsuda; JP24gm1610007 to A.
617 Hirabayashi; JP24fk0108665, JP24fk0108683, JP24fk0108712, JP24fk0108642,
618 JP24gm1610003, JP24wm0225029, and JP24wm0225022 to M. Suzuki;
619 JP22wm0325022 and JP22wm0225008 to M. Shintani) from the Japan Agency for
620 Medical Research and Development (AMED), grants (JPMJAX21BK to Y. Nishimura;
621 19H02869 and 19H05679 to H. Nojiri; 20K10436 and JPMJCR20H1 to H. Suzuki;
622 22KK0058 and 23K06556 to M. Suzuki; 16H06279, 19H02869, 19H05686, and
623 JP23H02124 to M. Shintani) from the Ministry of Education, Culture, Sports, Science and
624 Technology (MEXT), Japan, grants to Y. Nishimura, M. Tokuda, X. Hui, C. Suzuki-
625 Minakuchi, H. Nojiri, H. Suzuki, M. Suzuki, and M. Shintani from Consortium for the
626 Exploration of Microbial Functions of Ohsumi Frontier Science Foundation, Japan, a
627 grant from Asahi Glass Foundation to M. Shintani, a grant to M. Shintani from Institute
628 of Fermentation Osaka (L-2023-1-002), National University Corporation Shizuoka
629 University, Japan, and a grant to M. Shintani from the Research Institute of Green
630 Science and Technology Fund for Research Project Support (2023-RIGST-23104),
631 National University Corporation Shizuoka University, Japan.

632

633 **Conflict of interests**

634 The authors declare no conflict of interest.

635

636 **References**

637 1. Wiedenbeck, J., and Cohan, F.M. (2011). Origins of bacterial diversity through
638 horizontal genetic transfer and adaptation to new ecological niches. FEMS
639 Microbiol. Rev. *35*, 957–976.

640 2. Haudiquet, M., de Sousa, J.M., Touchon, M., and Rocha, E.P.C. (2022). Selfish,
641 promiscuous and sometimes useful: how mobile genetic elements drive horizontal
642 gene transfer in microbial populations. Philos. Trans. R. Soc. Lond. B Biol. Sci. *377*,

643 20210234.

644 3. Helinski, D.R. (2022). A Brief History of Plasmids. EcoSal Plus *10*, eESP00282021.

645 4. Starling, S. (2025). Plasmid evolution in the time of antibiotics. Nat. Rev. Microbiol.
646 https://doi.org/10.1038/s41579-025-01273-9.

647 5. Cazares, A., Figueroa, W., Cazares, D., Lima, L., Turnbull, J.D., McGregor, H., Dicks,
648 J., Alexander, S., Iqbal, Z., and Thomson, N.R. (2025). Pre- and postantibiotic
649 epoch: The historical spread of antimicrobial resistance. Science *390*, eadr1522.

650 6. Camargo, A.P., Call, L., Roux, S., Nayfach, S., Huntemann, M., Palaniappan, K.,
651 Ratner, A., Chu, K., Mukherjeep, S., Reddy, T.B.K., et al. (2024). IMG/PR: a
652 database of plasmids from genomes and metagenomes with rich annotations and
653 metadata. Nucleic Acids Res. *52*, D164–D173.

654 7. Schmartz, G.P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Müller, R., and
655 Keller, A. (2022). PLSDB: advancing a comprehensive database of bacterial
656 plasmids. Nucleic Acids Res. *50*, D273–D278.

657 8. Garcillán-Barcia, M.P., Redondo-Salvo, S., and de la Cruz, F. (2023). Plasmid
658 classifications. Plasmid *126*, 102684.

659 9. Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O., Villa,
660 L., Moller Aarestrup, F., and Hasman, H. (2014). *In silico* detection and typing of
661 plasmids using PlasmidFinder and plasmid multilocus sequence typing.
662 Antimicrob. Agents Chemother. *58*, 3895–3903.

663 10. Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M.P., and de la Cruz, F. (2017).
664 Towards a taxonomy of conjugative plasmids. Curr. Opin. Microbiol. *38*, 106–113.

665 11. Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M., Rocha,
666 E.P.C., Garcillán-Barcia, M.P., and de la Cruz, F. (2020). Pathways for horizontal
667 gene transfer in bacteria revealed by a global map of their plasmids. Nat.
668 Commun. *11*, 3602.

669 12. Garcillán-Barcia, M.P., Redondo-Salvo, S., Vielva, L., and de la Cruz, F. (2020).
670 MOBscan: Automated annotation of MOB relaxases. Methods Mol. Biol. *2075*,
671 295–308.

672 13. Coluzzi, C., Garcillán-Barcia, M.P., de la Cruz, F., and Rocha, E.P.C. (2022).
673 Evolution of plasmid mobility: origin and fate of conjugative and nonconjugative
674 plasmids. Mol. Biol. Evol. *39*. https://doi.org/10.1093/molbev/msac115.

675 14. Moradali, M.F., Ghods, S., and Rehm, B.H.A. (2017). *Pseudomonas aeruginosa*
676 lifestyle: A paradigm for adaptation, survival, and persistence. Front. Cell. Infect.
677 Microbiol. *7*, 39.

678 15. Thomas, C.M., and Haines, A.S. (2004). Plasmids of the genus *Pseudomonas*. In:
679 Ramos, JL. (eds) Pseudomonas. Springer, Boston, MA.
680 https://doi.org/10.1007/978-1-4419-9086-0_7

681 16. Shintani, M., and Suzuki, H. (2019). Plasmids and their hosts. In: Nishida, H.,
682 Oshima, T. (eds) DNA Traffic in the Environment. Springer, Singapore.
683 https://doi.org/10.1007/978-981-13-3411-5_6

684 17. Sawada, Y., Yaginuma, S., Tai, M., Iyobe, S., and Mitsuhashi, S. (1976). Plasmid-
685 mediated penicillin beta-lactamases in *Pseudomonas aeruginosa*. Antimicrob.
686 Agents Chemother. *9*, 55–60.

687 18. Sagai, H., Hasuda, K., Iyobe, S., Bryan, L.E., Holloway, B.W., and Mitsuhashi, S.
688 (1976). Classification of R Plasmids by incompatibility in *Pseudomonas aeruginosa*.
689 Antimicrob. Agents Chemother. *10*, 573–578.

690 19. Summers, A.O., and Jacoby, G.A. (1978). Plasmid-determined resistance to boron
691 and chromium compounds in *Pseudomonas aeruginosa*. Antimicrob. Agents
692 Chemother. *13*, 637–640.

693 20. Kawakami, Y., Mikoshiba, F., Nagasaki, S., Matsumoto, H., and Tazaki, T. (1972).
694 Prevalence of *Pseudomonas aeruginosa* strains possessing R factor in a hospital. J.
695 Antibiot. *25*, 607–609.

696 21. Hedges, R.W., and Jacob, A.E. (1975). Mobilization of plasmid-borne drug
697 resistance determinants for transfer from *Pseudomonas aeruginosa* to *Escherichia*
698 *coli*. Mol. Gen. Genet. *140*, 69–79.

699 22. Ingram, L., Sykes, R.B., Grinsted, J., Saunders, J.R., and Richmond, M.H. (1972). A
700 transmissible resistance element from a strain of *Pseudomonas aeruginosa*
701 containing no detectable extrachromosomal DNA. J. Gen. Microbiol. *72*, 269–279.

702 23. Bryan, L.E., Semaka, S.D., Van den Elzen, H.M., Kinnear, J.E., and Whitehouse, R.L.
703 (1973). Characteristics of R931 and other *Pseudomonas aeruginosa* R factors.
704 Antimicrob. Agents Chemother. *3*, 625–637.

705 24. Bradley, D.E. (1983). Specification of the conjugative pili and surface mating
706 systems of *Pseudomonas* plasmids. J. Gen. Microbiol. *129*, 2545–2556.

707 26. Cain, D., and Holloway, B.W. (1984). Prime plasmids derived from the IncP-10
708 plasmid R91-5 in *Pseudomonas putida*. FEMS Microbiol. Lett. *24*, 97–101.

709 27. Davies, S., and Krishnapillai, V. (1990). DNA sequence analysis of the replication
710 region of the *Pseudomonas aeruginosa* plasmid R91-5. J. Genet. *69*, 101–112.

711 28. Jacoby, G.A. (1986). Resistance Plasmids of *Pseudomonas*. In: Sokatch, J.R. (Ed.)
712 The Biology of *Pseudomonas*, Academic Press, https://doi.org/10.1016/b978-0-12-

713 307210-8.50013-0.

714 29. Suzuki, M., Suzuki, H., Nishimura, Y., Nojiri, H., and Shintani, M. (2025). Enhancing
715 plasmid typing with MOB-typer: resolving IncP and other incompatibility group
716 misclassifications in *Pseudomonas*. Microb. Genom. *11*.
717 https://doi.org/10.1099/mgen.0.001491.

718 30. Kawalek, A., Kotecka, K., Modrzejewska, M., Gawor, J., Jagura-Burdzy, G., and
719 Bartosik, A.A. (2020). Genome sequence of *Pseudomonas aeruginosa* PAO1161, a
720 PAO1 derivative with the ICEPae1161 integrative and conjugative element. BMC
721 Genomics *21*, 14.

722 31. Boronin, A.M. (1992). Diversity of *Pseudomonas* plasmids: To what extent? FEMS
723 Microbiol. Lett. *100*, 461–467.

724 32. Van der Auwera, G.A., Krol, J.E., Suzuki, H., Foster, B., Van Houdt, R., Brown, C.J.,
725 Mergeay, M., and Top, E.M. (2009). Plasmids captured in *C. metallidurans* CH34:
726 defining the PromA family of broad-host-range plasmids. Antonie Van
727 Leeuwenhoek *96*, 193–204.

728 33. Law, A., Solano, O., Brown, C.J., Hunter, S.S., Fagnan, M., Top, E.M., and Stalder, T.
729 (2021). Biosolids as a source of antibiotic resistance plasmids for commensal and
730 pathogenic bacteria. Front. Microbiol. *12*, 606409.

731 34. Yanagiya, K., Maejima, Y., Nakata, H., Tokuda, M., Moriuchi, R., Dohra, H., Inoue,
732 K., Ohkuma, M., Kimbara, K., and Shintani, M. (2018). Novel self-transmissible and
733 broad-host-range plasmids exogenously captured from anaerobic granules or cow
734 manure. Front. Microbiol. *9*, 2602.

735 35. Tett, A., Spiers, A.J., Crossman, L.C., Ager, D., Ciric, L., Dow, J.M., Fry, J.C., Harris,
736 D., Lilley, A., Oliver, A., et al. (2007). Sequence-based analysis of pQBR103; a
737 representative of a unique, transfer-proficient mega plasmid resident in the
738 microbial community of sugar beet. ISME J. *1*, 331–340.

739 36. Köhler, K.A.K., Rückert, C., Schatschneider, S., Vorhölter, F.-J., Szczepanowski, R.,
740 Blank, L.M., Niehaus, K., Goesmann, A., Pühler, A., Kalinowski, J., et al. (2013).
741 Complete genome sequence of *Pseudomonas* sp. strain VLB120 a solvent tolerant,
742 styrene degrading bacterium, isolated from forest soil. J. Biotechnol. *168*, 729–
743 730.

744 37. Lawrence, J.G., and Ochman, H. (1997). Amelioration of bacterial genomes: rates
745 of change and exchange. J. Mol. Evol. *44*, 383–397.

746 38. Rocha, E.P., and Danchin, A. (2002). Base composition bias might result from
747 competition for metabolic resources. Trends Genet. *18*, 291–294.

748 39. Dietel, A.-K., Merker, H., Kaltenpoth, M., and Kost, C. (2019). Selective advantages

749 favour high genomic AT-contents in intracellular elements. PLoS Genet. *15*,
750 e1007778.

751 40. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast
752 iterative protein sequence searching by HMM-HMM alignment. Nat. Methods *9*,
753 173–175.

754 41. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O.,
755 Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction
756 of biomolecular interactions with AlphaFold 3. Nature *630*, 493–500.

757 42. Wegrzyn, K., Zabrocka, E., Bury, K., Tomiczek, B., Wieczor, M., Czub, J.,
758 Uciechowska, U., Moreno-Del Alamo, M., Walkow, U., Grochowina, I., et al.
759 (2021). Defining a novel domain that provides an essential contribution to site-
760 specific interaction of Rep protein with DNA. Nucleic Acids Res.
761 https://doi.org/10.1093/nar/gkab113.

762 43. Schumacher, M.A., Tonthat, N.K., Kwong, S.M., Chinnam, N.B., Liu, M.A., Skurray,
763 R.A., and Firth, N. (2014). Mechanism of staphylococcal multiresistance plasmid
764 replication origin assembly by the RepA protein. Proc. Natl. Acad. Sci. U. S. A. *111*,
765 9121–9126.

766 44. Itou, H., Yagura, M., Shirakihara, Y., and Itoh, T. (2015). Structural basis for
767 replication origin unwinding by an initiator primase of plasmid ColE2-P9: duplex
768 DNA unwinding by a single protein. J. Biol. Chem. *290*, 3601–3611.

769 45. Iyer, L.M., Koonin, E.V., Leipe, D.D., and Aravind, L. (2005). Origin and evolution of
770 the archaeo-eukaryotic primase superfamily and related palm-domain proteins:
771 structural insights and new members. Nucleic Acids Res. *33*, 3875–3896.

772 46. Haines, A.S., Jones, K., Cheung, M., and Thomas, C.M. (2005). The IncP-6 plasmid
773 Rms149 consists of a small mobilizable backbone with multiple large insertions. J.
774 Bacteriol. *187*, 4728–4738.

775 47. Haines, A.S., Akhtar, P., Stephens, E.R., Jones, K., Thomas, C.M., Perkins, C.D.,
776 Williams, J.R., Day, M.J., and Fry, J.C. (2006). Plasmids from freshwater
777 environments capable of IncQ retrotransfer are diverse and include pQKH54, a
778 new IncP-1 subgroup archetype. Microbiology *152*, 2689–2701.

779 48. Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The
780 many faces of the helix-turn-helix domain: transcription regulation and beyond.
781 FEMS Microbiol Rev *29*, 231–262.

782 49. Giraldo, R., and Fernandez-Tresguerres, M.E. (2004). Twenty years of the pPS10
783 replicon: insights on the molecular mechanism for the activation of DNA
784 replication in iteron-containing bacterial plasmids. Plasmid *52*, 69–83.

785 50. Giraldo, R. (2003). Common domains in the initiators of DNA replication in
786 *Bacteria, Archaea* and *Eukarya*: combined structural, functional and phylogenetic
787 perspectives. FEMS Microbiol. Rev. *26*, 533–554.

788 51. Tria, F.D.K., Landan, G., and Dagan, T. (2017). Phylogenetic rooting using minimal
789 ancestor deviation. Nat. Ecol. Evol. *1*, 193.

790 52. Lam, M.M.C., and Hamidian, M. (2024). Examining the role of Acinetobacter
791 baumannii plasmid types in disseminating antimicrobial resistance. NPJ
792 Antimicrob Resist *2*, 1.

793 53. Komori, H., Matsunaga, F., Higuchi, Y., Ishiai, M., Wada, C., and Miki, K. (1999).
794 Crystal structure of a prokaryotic replication initiator protein bound to DNA at 2.6
795 A resolution. EMBO J. *18*, 4597–4607.

796 54. Sambrook, J., and Russell, D. (2001). Molecular Cloning: A Laboratory Manual 3rd
797 edn. (Cold Spring Harbor Laboratory).

798 55. Royle, P.L., Matsumoto, H., and Holloway, B.W. (1981). Genetic circularity of the
799 *Pseudomonas aeruginosa* PAO chromosome. J. Bacteriol. *145*, 145–155.

800 56. plasmidSPAdes – Center for Algorithmic Biotechnology
801 https://cab.spbu.ru/software/plasmid-spades/.

802 57. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton,
803 S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an
804 integrated and extendable desktop software platform for the organization and
805 analysis of sequence data. Bioinformatics *28*, 1647–1649.

806 58. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
807 Abecasis, G., Durbin, R., and 1000 Genome project data processing subgroup
808 (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*,
809 2078–2079.

810 59. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-platform and ultrafast
811 toolkit for FASTA/Q file manipulation. PLoS One *11*, e0163962.

812 60. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S.,
813 Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new
814 genome assembly algorithm and its applications to single-cell sequencing. J.
815 Comput. Biol. *19*, 455–477.

816 61. Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018). DFAST: a flexible prokaryotic
817 genome annotation pipeline for faster genome publication. Bioinformatics *34*,
818 1037–1039.

819 62. Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011). Easyfig: a genome comparison

820 visualizer. Bioinformatics *27*, 1009–1010.

821 63. Lobry, J.R. (1996). Origin of Replication of *Mycoplasma genitalium*. Science. *272*,
822 745–746.

823 64. Necşulea, A., and Lobry, J.R. (2007). A new method for assessing the effect of
824 replication on DNA base composition asymmetry. Mol. Biol. Evol. *24*, 2169–2179.

825 65. V. Solovyev, A Salamov (2011) Automatic annotation of microbial genomes and
826 metagenomic sequences. In: Li, R.W. (Ed.) Metagenomics and its applications in
827 agriculture, biomedicine and environmental studies, Nova Science Publishers,
828 p.61-78.

829 66. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M.,
830 Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search
831 with Foldseek. Nat. Biotechnol. *42*, 243–246.

832 67. Swan, M.K., Bastia, D., and Davies, C. (2006). Crystal structure of π initiator
833 protein–iteron complex of plasmid R6K: Implications for initiation of plasmid DNA
834 replication. Proc. Nat. Acad. Sci. USA. *103*, 18481–18486.

835 68. Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2018). PLSDB: a resource of
836 complete bacterial plasmids. Nucleic Acids Res. *47*, D195-D202.

837 69. Robertson, J., and Nash, J.H.E. (2018). MOB-suite: software tools for clustering,
838 reconstruction and typing of plasmids from draft assemblies. Microb. Genom. *27*,
839 e000206.

840 70. Carattoli, A., and Hasman, H. (2020). PlasmidFinder and in silico pMLST:
841 Identification and typing of plasmid replicons in whole-genome sequencing (WGS).
842 Methods Mol. Biol. *2075*, 285–294.

843 71. Suzuki, H., Moriguchi, K., Shintani, M., Suzuki, M., and Nojiri, H. (2025). Insights
844 from public database sequences related to the replication initiation protein TrfA of
845 the IncP-1 plasmid RK2. Plasmid *134*, 102756.

846 72. Bayliss, S.C., Thorpe, H.A., Coyle, N.M., Sheppard, S.K., and Feil, E.J. (2019).
847 PIRATE: A fast and scalable pangenomics toolbox for clustering diverged
848 orthologues in bacteria. Gigascience *8*.
849 https://doi.org/10.1093/gigascience/giz119.

850 73. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at
851 tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368.

852 74. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and
853 Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein
854 database search programs. Nucleic Acids Res. *25*, 3389–3402.

855 75. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein
856 sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*,
857 1026–1028.

858 76. Woodcroft, B.J., Boyd, J.A., and Tyson, G.W. (2016). OrfM: a fast open reading
859 frame predictor for metagenomic data. Bioinformatics *32*, 2702–2703.

860 77. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*,
861 e1002195.

862 78. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and
863 Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein
864 annotation. BMC Bioinformatics *20*, 473.

865 79. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment
866 software version 7: improvements in performance and usability. Mol. Biol. Evol.
867 *30*, 772–780.

868 80. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von
869 Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods
870 for phylogenetic inference in the fenomic era. Mol. Biol. Evol. *37*, 1530–1534.

871 81. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S.
872 (2017). ModelFinder: fast model selection for accurate phylogenetic estimates.
873 Nat. Methods *14*, 587–589.

874 82. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018).
875 UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*,
876 518–522.

877 83. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O.
878 (2010). New algorithms and methods to estimate maximum-likelihood
879 phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

880 84. Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier
881 long branches in collections of phylogenetic trees. BMC Genomics *19*, 272.

882 85. Lemoine, F., and Gascuel, O. (2021). Gotree/Goalign: toolkit and Go API to
883 facilitate the development of phylogenetic workflows. NAR Genom Bioinform *3*,
884 lqab075.

885 86. Pansegrau, W., Lanka, E., Barth, P.T., Figurski, D.H., Guiney, D.G., Haas, D.,
886 Helinski, D.R., Schwab, H., Stanisich, V.A., and Thomas, C.M. (1994). Complete
887 nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and
888 comparative analysis. J. Mol. Biol. *239*, 623–663.

889 87. Shintani, M., Suzuki, H., Nojiri, H., and Suzuki, M. (2022). Precise classification of

890 antimicrobial resistance-associated IncP-2 megaplasmids for molecular
891 epidemiological studies on Pseudomonas species. J. Antimicrob. Chemother. *77*,
892 1203–1205.

893 88. Llanes, C., Gabant, P., Couturier, M., Bayer, L., and Plesiat, P. (1996). Molecular
894 analysis of the replication elements of the broad-host-range RepA/C replicon.
895 Plasmid *36*, 26–35.

896 89. Fricke, W.F., Welch, T.J., McDermott, P.F., Mammel, M.K., LeClerc, J.E., White,
897 D.G., Cebula, T.A., and Ravel, J. (2009). Comparative genomics of the IncA/C
898 multidrug resistance plasmid family. J. Bacteriol. *191*, 4750–4757.

899 90. Scholz, P., Haring, V., Wittmann-Liebold, B., Ashman, K., Bagdasarian, M., and
900 Scherzinger, E. (1989). Complete nucleotide sequence and gene organization of
901 the broad-host-range plasmid RSF1010. Gene *75*, 271–288.

902 91. Maeda, K., Nojiri, H., Shintani, M., Yoshida, T., Habe, H., and Omori, T. (2003).
903 Complete nucleotide sequence of carbazole/dioxin-degrading plasmid pCAR1 in
904 *Pseudomonas resinovorans* strain CA10 indicates its mosaicity and the presence of
905 large catabolic transposon Tn*4676*. J. Mol. Biol. *326*, 21–33.

906 92. Shintani, M., Yano, H., Habe, H., Omori, T., Yamane, H., Tsuda, M., and Nojiri, H.
907 (2006). Characterization of the replication, maintenance, and transfer features of
908 the IncP-7 plasmid pCAR1, which carries genes involved in carbazole and dioxin
909 degradation. Appl. Environ. Microbiol. *72*, 3206–3216.

910 93. Takahashi, Y., Shintani, M., Yamane, H., and Nojiri, H. (2009). The complete
911 nucleotide sequence of pCAR2: pCAR2 and pCAR1 were structurally identical IncP-
912 7 carbazole degradative plasmids. Biosci. Biotechnol. Biochem. *73*, 744–746.

913 94. Greated, A., Lambertsen, L., Williams, P.A., and Thomas, C.M. (2002). Complete
914 sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. Environ.
915 Microbiol. *4*, 856-871.

916 95. Sota, M., Yano, H., Ono, A., Miyazaki, R., Ishii, H., Genka, H., Top, E.M., and Tsuda,
917 M. (2006). Genomic and functional analysis of the IncP-9 naphthalene-catabolic
918 plasmid NAH7 and its transposon Tn*4655* suggests catabolic gene spread by a
919 tyrosine recombinase. J. Bacteriol. *188*, 4057–4067.

920 96. Hayakawa, M., Tokuda, M., Kaneko, K., Nakamichi, K., Yamamoto, Y., Kamijo, T.,
921 Umeki, H., Chiba, R., Yamada, R., Mori, M., et al. (2022). Hitherto-unnoticed self-
922 transmissible plasmids widely distributed among different environments in Japan.
923 Appl. Environ. Microbiol., *88*, e0111422.

924 **Table 1. Lists of representative plasmids in the PInc groups.**

PInc(s) (classical Inc name[a])	Representative plasmid[b] [accession no.]	MOB MPF[c]	No. of plasmids in PLSDB	Numbers of WH	Clade	References
PInc-1 **(IncP-1, IncP)**	RK2/RP4 [BN000925]	MOB_P MPF_T	321	2	F	[86]
PInc-2 **(IncP-2)**	Rms139* [LC653116]	MOB_P MPF_I	87	2	E	[17,87]
PInc-3 **(IncP-3, IncA/C)**	pRA1 (IncA) [FJ705807] pR55 (IncC) [JQ010984]	MOB_H MPF_F	921	2	E	[88,89]
PInc-4 **(IncP-4, IncQ)**	RSF1010 [M28829]	MOB_Q -	695	2	F	[90]
PInc-5 **(IncP-5)**	Rms163* [LC685027]	NA NA	2	1	B	[18]

PInc-6 **(IncP-6, IncG)**	Rms149 [AJ877225]	MOB_P -	119	NA	-	[46]
PInc-7 **(IncP-7)**	pCAR1 [AB088420]	MOB_H MPF_F	23	2	H	[91–93]
PInc-9 **(IncP-9)**	pWW0 [AJ344068] NAH7 [AB237655] Rsu2* [LC685593]	MOB_F MPF_T	34	2	A	[21,94,95]
PInc-10 **(IncP-10)**	R91-5 [X54695[d]] pPAB546 [MN433456]	NA NA	49	1	C	[26,27]
PInc-11 **(IncP-11)**	RP1-1* [LC700336]	MOB_c NA	23	1	C	[22]

PInc-12 **(IncP-12)**	R716* [LC685026]	NA NA	9	2	F	[23]
- **(IncP-13)**	pMG26*[e] [LC685025]	MOB_F MPF_G	NA	NA	-	[24,25]
PInc-15 **(PromA)**	pSN1104-11 [AP018707]	MOB_P MPF_T	42	2	G	[96]
PInc-16	pSN1216-29 [AP018710]	MOB_P MPF_T	24	2	E	[34]
PInc-17	pQBR103 [AM235768]	NA NA MOB_F MPF_F	11	2	E	[35]
PInc-18 **(Inc_{pSTY})**	pSTY [CP003962] pPT23-C1_1 [AP040308]	MOB_F MPF_F	15	2	H	[36]

925 [a]Names of plasmid Inc groups originating from Pseudomonas and/or Enterobacterales are shown in parentheses. Some Inc groups
926 have been historically given different names depending on their host, *Pseudomonas* or *Enterobacterales*, but they represent the same
927 plasmid group (e.g., IncP-1 in *Pseudomonas* corresponds to IncP in *Enterobacterales*). For PInc-18, the name "Inc_{pSTY}" was adopted
928 from a previous study by another research group[36]. [b]Representative plasmid name [GenBank accession number]. Asterisks (*) indicate
929 plasmids whose complete nucleotide sequences were determined in this study (Rms139, Rms163, Rsu2, RP1-1, R716, and pMG26).
930 [c]Classification of relaxase-based MOB classes and MPF types was identified by the MOB-suite with criteria of >=90% identity and
931 >=60% coverage. "-" indicates that the genes involved in conjugation have not been detected, whereas "NA" indicates that the classes
932 and types of MOB or MPF were not identified although the plasmid was self-transmissible. [d]Partial DNA sequences are available.
933 [e]pMG26 was shown to be not a plasmid but an integrative and conjugative element in this study.

934

935

936 ## Figure legend

937 **Figure 1. Features of plasmids of PInc groups in the PLSDB.**

938 Distribution of (A) plasmid sizes (kb) and (B) GC contents (%), and (C) ratio of core and
939 accessory genes of each PInc plasmids in the PLSDB. (D) Relationship of each PInc plasmids
940 to other plasmid replicons (left) and ARGs (right) in the PLSDB. “rep_cluster” contains
941 different types of replicons identified by MOB-typer. ’N.D.' indicates not detected. A sankey
942 diagram was generated using the SankeyWidget library in Python through in-house
943 programming.

944 **Figure 2. Large-scale phylogeny of WH RIPs and its links to plasmid diversity across public**
945 **databases.**

946 (A) Maximum-likelihood phylogeny of WH RIPs reconstructed from the conserved
947 homologous region shared across 13 RIP groups. The tree was built from 4,769
948 representative sequences clustered at 50% identity. Outer annotations indicate clade and
949 subclade assignment, Pfam annotation, taxonomy of linked plasmids in PLSDB and IMG/PR,
950 environmental category of linked IMG/PR plasmids, and replicon typing information. The
951 major split separates the sWH and dWH superclades, corresponding to RIPs with one or two
952 winged-helix domains, respectively. Curved lines in the center indicate co-occurrence of WH
953 RIPs in the same multi-replicon plasmids, colored by host taxonomy. (B) Summary of the
954 eight clades (A–H) and 42 subclades, with SH-aLRT and UFBoot2 support values shown for
955 major nodes. Representative PInc groups and other known replicon types associated with
956 each clade are indicated. (C) Numbers and proportions of plasmids in PLSDB and IMG/PR
957 linked to leaves in the WH RIP phylogeny. (D) Taxonomic composition of multi-replicon
958 plasmids carrying multiple WH RIPs in PLSDB and IMG/PR, stratified by the number of WH
959 replicons per plasmid. (E) Distribution of Pfam annotations across clades (left) and clade
960 composition within each Pfam category (right). Sample sizes are indicated next to the bars.

961 **Figure 3. Domain architecture and predicted structural features of representative Rep**
962 **proteins from PInc plasmids.**

963 (A) Domain organization of representative Rep proteins from PInc plasmids, including Rep
964 pSN2, Rep p14035, Rep pOXA-48, Rep pSK41, RepP-16A, RepP-4A, RepP-15A, and RepP-7A.
965 N- and C-termini are indicated. Winged-helix (WH) domains (WH1, WH2) and conserved
966 replication-associated domains (e.g., RepL [PF05732], Bac_RepA_C [PF18008], IncFII_repA
967 [PF02387], RepA_N [PF06970], RPA [PF10134], RepC [PF06504], RepA_C [PF04796], Rep3_N
968 [PF01051], Rep3_C [PF21205]) are shown. Predicted oligomeric states (dimer or monomer)

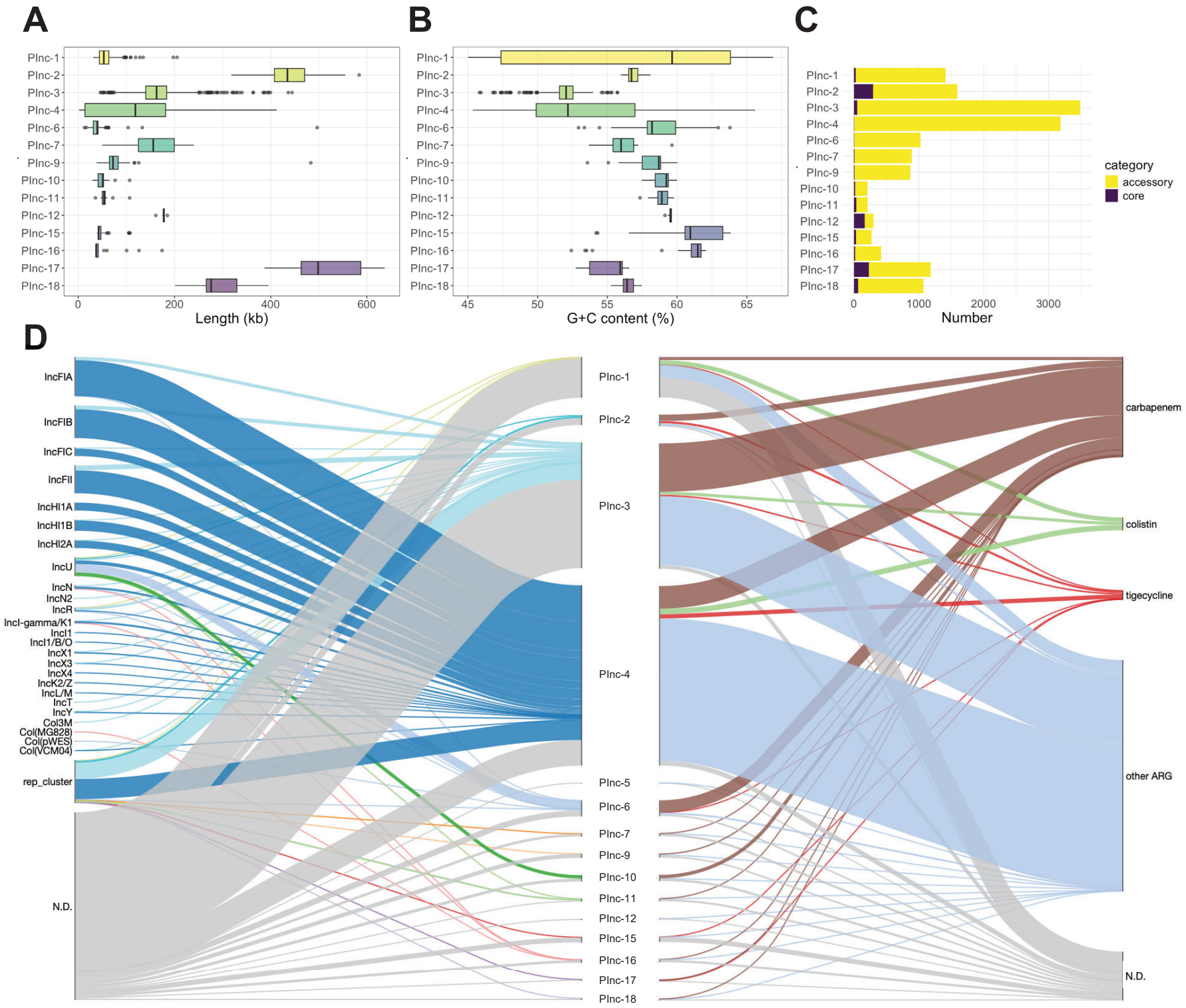
969 are indicated. (B) Pairwise protein sequence identity among representative Rep proteins. (C)
970 Predicted three-dimensional structures of representative Rep proteins and their DNA-
971 binding modes. Structural models highlight the arrangement of WH domains and predicted
972 interactions with DNA, illustrating diversity in oligomeric state and DNA recognition
973 mechanisms among PInc replicons.

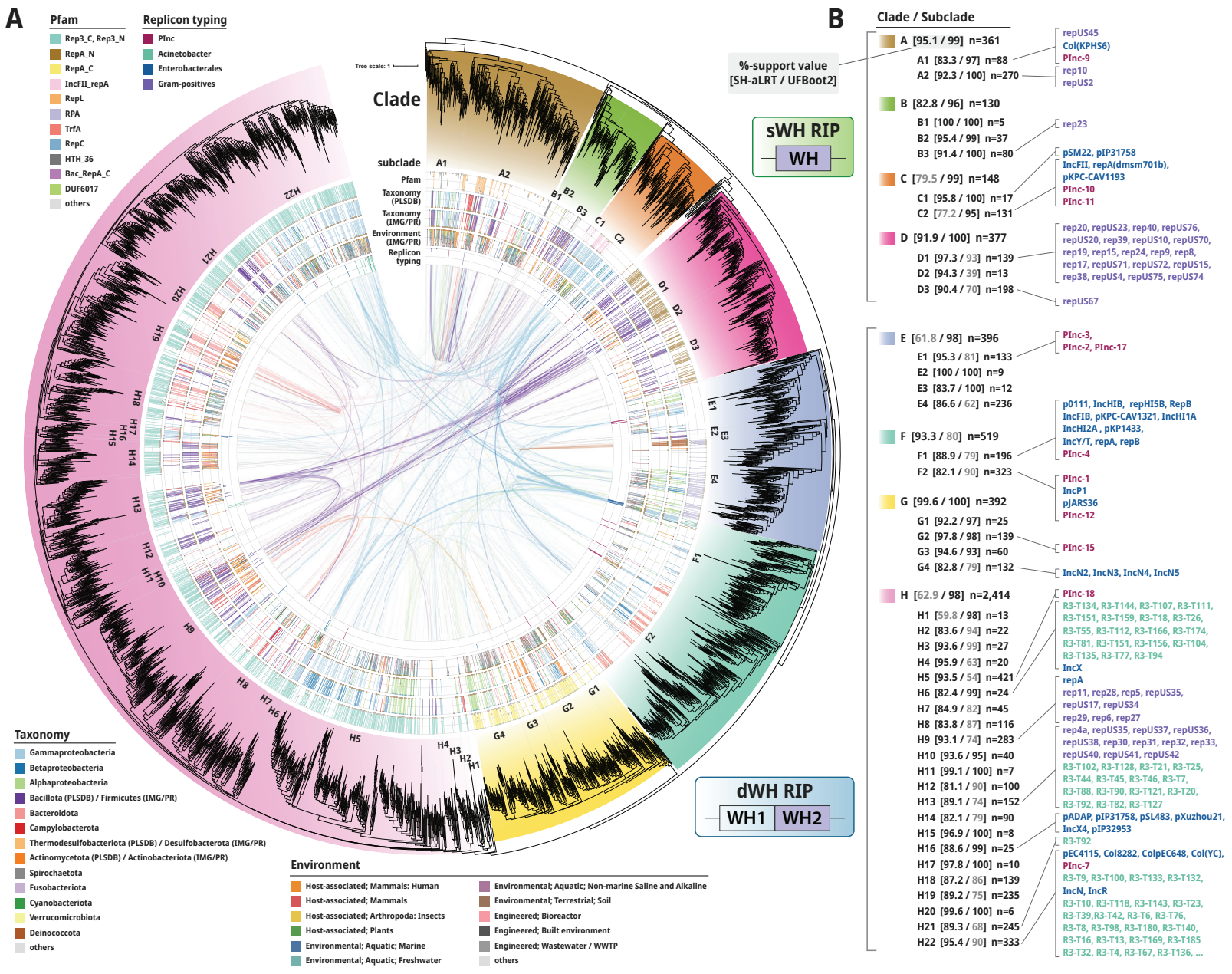
974 **Figure 4. Taxonomic, ecological, compositional and functional diversification across the**
975 **eight WH RIP clades.**

976 (A) Taxonomic distribution of WH RIP clades among linked plasmids in PLSDB and IMG/PR.
977 Upper panels show the percentage distribution of taxa across clades; lower panels show
978 clade composition within each taxon. PLSDB follows NCBI taxonomy and IMG/PR follows
979 GTDB taxonomy. (B) Environmental distribution of WH RIP clades among IMG/PR-linked
980 plasmids derived from metagenomic and metatranscriptomic datasets. Upper panel,
981 percentage distribution of environments across clades; lower panel, clade composition
982 within each environment. (C) RIP amino-acid length versus plasmid length across clades,
983 with violin plots summarizing clade-wise distributions. (D) RIP GC content versus the
984 difference between RIP and plasmid GC content, with violin plots summarizing clade-wise
985 distributions. (E) RIP N-ARSC (nitrogen atoms per residue side chain) versus their differences
986 from the corresponding plasmid proteomes, with violin plots summarizing clade-wise
987 distributions. (F) Distribution of MOB classes, MPF classes and Pfam domain annotations
988 across clades. Upper panels show the proportions within each clade; lower panels show
989 clade composition within each category. (G) Distribution of antimicrobial resistance (AMR)
990 genes and defense systems across clades, shown as the proportions of plasmids carrying
991 different numbers of AMR or defense genes. Sample sizes are indicated next to the bars.

992 **Figure 5. Integrated evolutionary landscape of the eight WH RIP clades.**

993 Summary of the major features distinguishing the eight WH RIP clades. The tree is
994 partitioned into the sWH and dWH superclades, corresponding to single- and double-
995 winged-helix replication initiation proteins, with branch support values shown at the major
996 internal nodes. For each clade, the figure summarizes domain arrangement, representative
997 Pfam domains, characteristic plasmid and RIP size, GC composition, predominant host
998 taxonomic distribution and environmental association. This integrated view illustrates that
999 the major split between sWH and dWH RIPs is accompanied by broad differences in domain
1000 organization and plasmid architecture, whereas the individual clades are further
1001 distinguished by characteristic combinations of compositional, taxonomic and ecological
1002 features.


A
PInc-1
PInc-2
PInc-3
PInc-4
PInc-6
PInc-7
PInc-9
PInc-10
PInc-11
PInc-12
PInc-15
PInc-16
PInc-17
PInc-18
0
200
400
600
Length (kb)
B
45
50
55
60
65
G+C content (%)
C
0
1000
2000
3000
Number
category
accessory
core
D
IncFIA
IncFIB
IncFIC
IncFII
IncHI1A
IncHI1B
IncHI2A
IncU
IncN
IncN2
IncR
IncI-gamma/K1
IncI1
IncI1/B/O
IncX1
IncX3
IncX4
IncK2/Z
IncL/M
IncT
IncY
Col3M
Col(MG828)
Col(pWES)
Col(VCM04)
rep_cluster
N.D.
carbapenem
colistin
tigecycline
other ARG
N.D.



B

Clade / Subclade

%-support value
[SH-aLRT / UFBoot2]

A [95.1 / 99] n=361
- A1 [83.3 / 97] n=88 — repUS45, Col(KPHS6), PInc-9
- A2 [92.3 / 100] n=270 — rep10, repUS2

B [82.8 / 96] n=130
- B1 [100 / 100] n=5 — rep23
- B2 [95.4 / 99] n=37
- B3 [91.4 / 100] n=80

C [79.5 / 99] n=148
- C1 [95.8 / 100] n=17 — pSM22, pIP31758, IncFII, repA(dmsm701b), pKPC-CAV1193, PInc-10, PInc-11
- C2 [77.2 / 95] n=131

D [91.9 / 100] n=377
- D1 [97.3 / 93] n=139 — rep20, repUS23, rep40, repUS76, repUS20, rep39, repUS10, repUS70, rep19, rep15, rep24, rep9, rep8, rep17, repUS71, repUS72, repUS15, rep38, repUS4, repUS75, repUS74
- D2 [94.3 / 39] n=13
- D3 [90.4 / 70] n=198 — repUS67

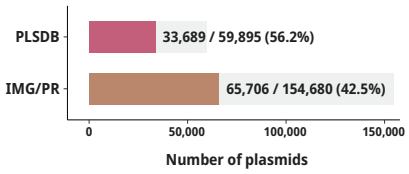
E [61.8 / 98] n=396
- E1 [95.3 / 81] n=133 — PInc-3, PInc-2, PInc-17
- E2 [100 / 100] n=9
- E3 [83.7 / 100] n=12
- E4 [86.6 / 62] n=236

F [93.3 / 80] n=519
- F1 [88.9 / 79] n=196 — p0111, IncHIB, repHI5B, RepB, IncFIB, pKPC-CAV1321, IncHI1A, IncHI2A , pKP1433, IncY/T, repA, repB, PInc-4
- F2 [82.1 / 90] n=323 — PInc-1, IncP1, pJARS36, PInc-12

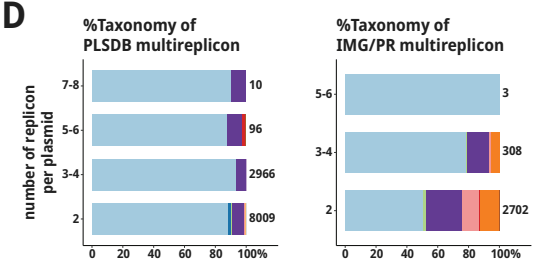
G [99.6 / 100] n=392
- G1 [92.2 / 97] n=25
- G2 [97.8 / 98] n=139 — PInc-15
- G3 [94.6 / 93] n=60
- G4 [82.8 / 79] n=132 — IncN2, IncN3, IncN4, IncN5

H [62.9 / 98] n=2,414
- H1 [59.8 / 98] n=13 — PInc-18, R3-T134, R3-T144, R3-T107, R3-T111, R3-T151, R3-T159, R3-T18, R3-T26, R3-T55, R3-T112, R3-T166, R3-T174, R3-T81, R3-T151, R3-T156, R3-T104, R3-T135, R3-T77, R3-T94, IncX
- H2 [83.6 / 94] n=22
- H3 [93.6 / 99] n=27
- H4 [95.9 / 63] n=20
- H5 [93.5 / 54] n=421
- H6 [82.4 / 99] n=24
- H7 [84.9 / 82] n=45 — repA, rep11, rep28, rep5, repUS35, repUS17, repUS34, rep29, rep6, rep27
- H8 [83.8 / 87] n=116
- H9 [93.1 / 74] n=283 — rep4a, repUS35, repUS37, repUS36, repUS38, rep30, rep31, rep32, rep33, repUS40, repUS41, repUS42
- H10 [93.6 / 95] n=40
- H11 [99.1 / 100] n=7
- H12 [81.1 / 90] n=100 — R3-T102, R3-T128, R3-T21, R3-T25, R3-T44, R3-T45, R3-T46, R3-T7, R3-T88, R3-T90, R3-T121, R3-T20, R3-T92, R3-T82, R3-T127
- H13 [89.1 / 74] n=152
- H14 [82.1 / 79] n=90 — pADAP, pIP31758, pSL483, pXuzhou21, IncX4, pIP32953
- H15 [96.9 / 100] n=8
- H16 [88.6 / 99] n=25 — R3-T92
- H17 [97.8 / 100] n=10 — pEC4115, Col8282, ColpEC648, Col(YC), PInc-7
- H18 [87.2 / 86] n=139 — R3-T9, R3-T100, R3-T133, R3-T132, IncN, IncR
- H19 [89.2 / 75] n=235
- H20 [99.6 / 100] n=6 — R3-T10, R3-T118, R3-T143, R3-T23, R3-T39,R3-T42, R3-T6, R3-T76, R3-T8, R3-T98, R3-T180, R3-T140, R3-T16, R3-T13, R3-T169, R3-T185 R3-T32, R3-T4, R3-T67, R3-T136, ...
- H21 [89.3 / 68] n=245
- H22 [95.4 / 90] n=333

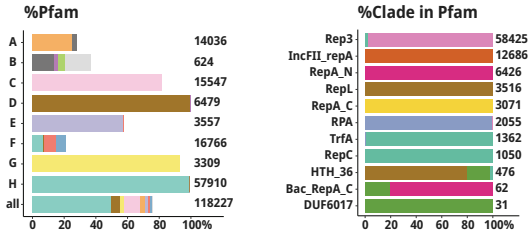
C

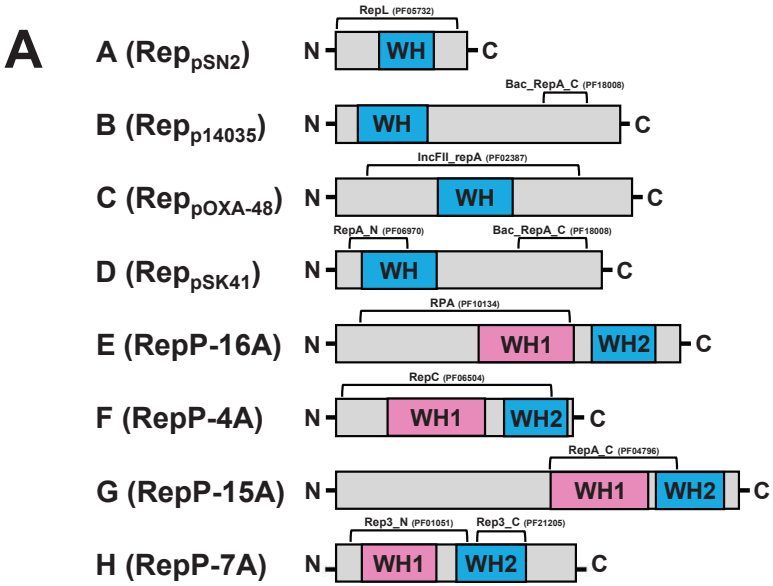


D



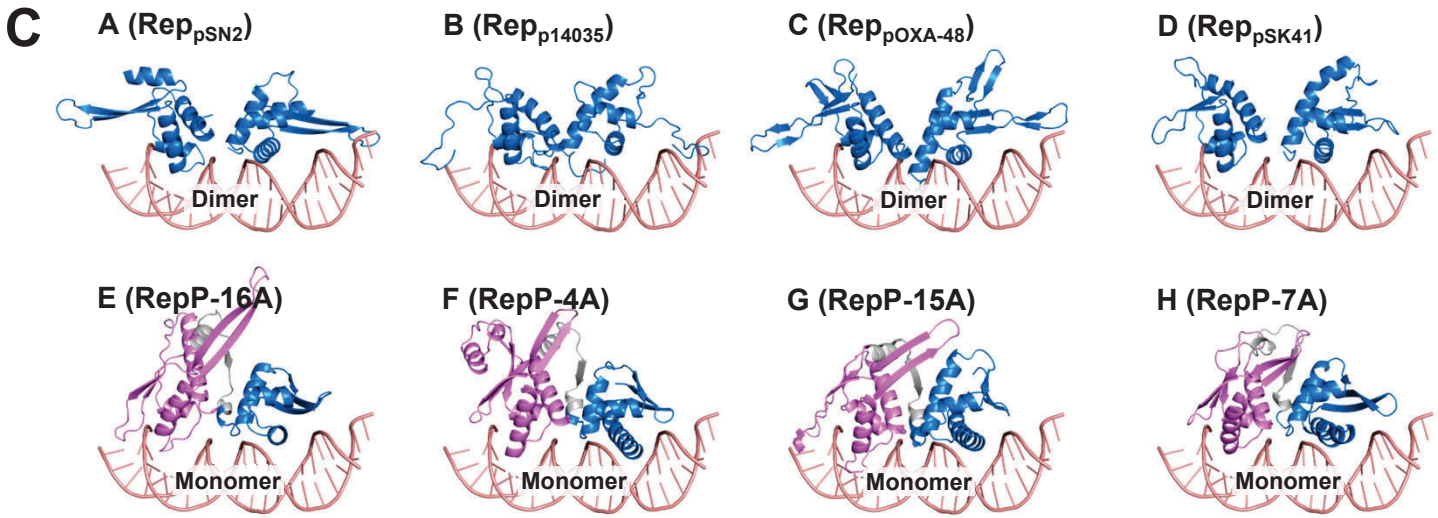
E

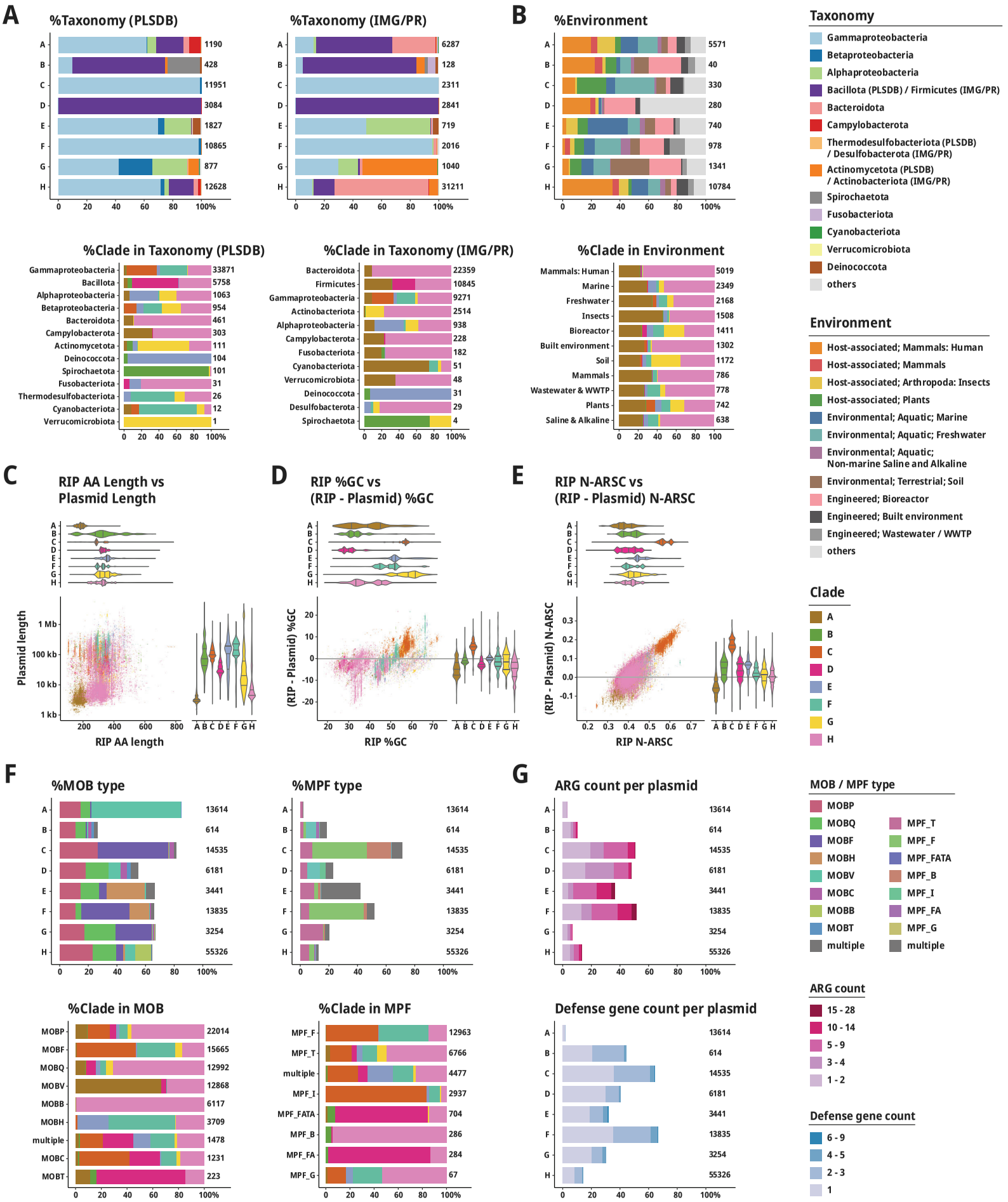


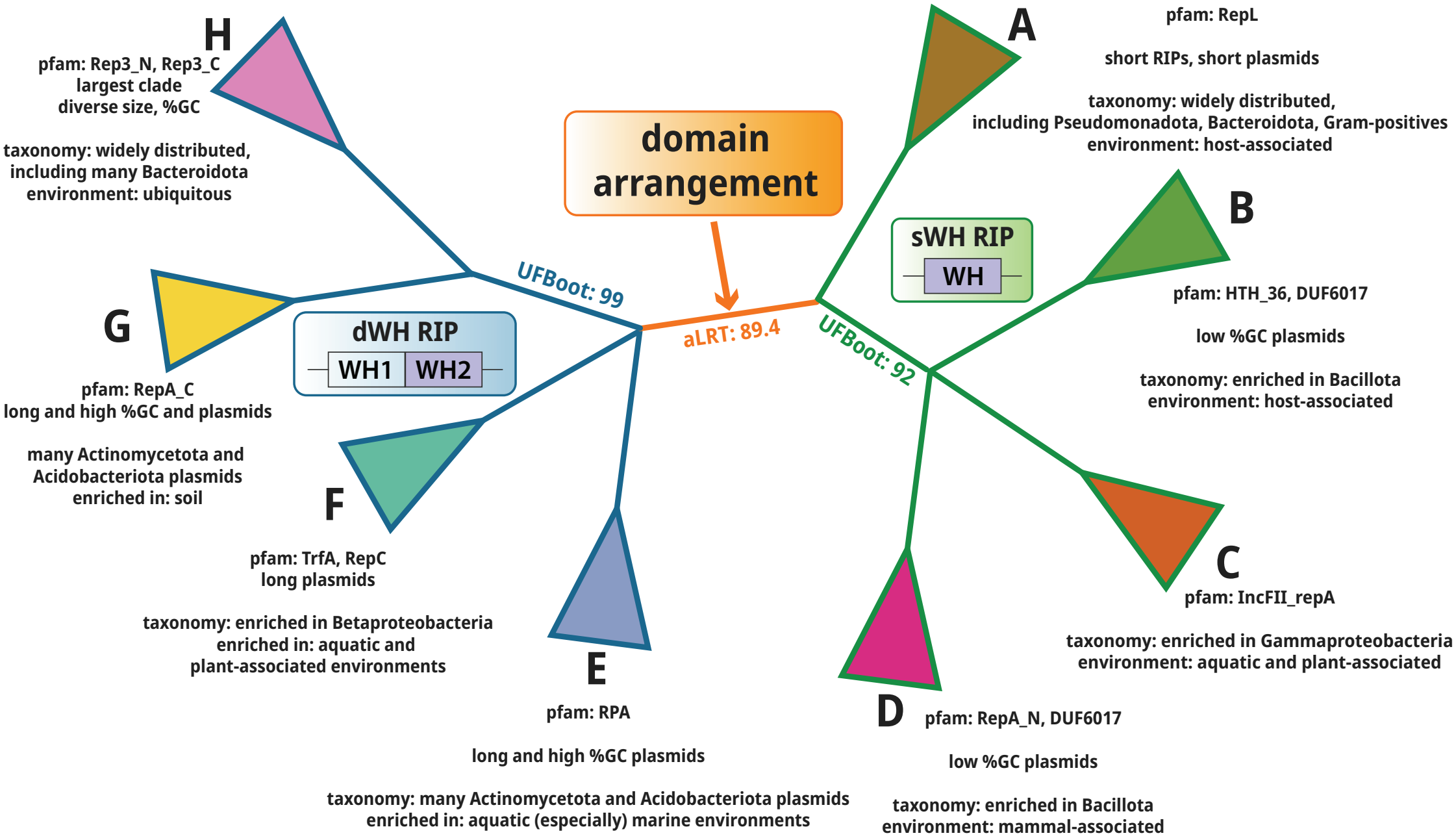


B

	A (Rep$_{pSN2}$)	A (RepP-9A)	B (Rep$_{p14035}$)	B (RepP-5A)	C (Rep$_{pOXA-48}$)	C (RepP-10A)	C (RepP-11A)	D (Rep$_{pSK41}$)	E (RepP-2A)	E (RepP-3A)	E (RepP-16A)	E (RepP-17A)	F (RepP-1A)	F (RepP-4A)	F (RepP-12A)	G (RepP-15A)	H (RepP-7A)	H (RepP-18A)
A (Rep$_{pSN2}$)	100	18	18	26	20	17	15	27	21	22	21	22	11	15	13	21	14	19
A (RepP-9A)	18	100	18	16	20	23	23	19	16	22	10	16	12	22	21	17	20	19
B (Rep$_{p14035}$)	18	18	100	10	14	20	20	17	15	13	15	23	23	15	18	14	17	15
B (RepP-5A)	26	16	10	100	18	18	18	16	16	13	15	13	20	18	18	19	10	18
C (Rep$_{pOXA-48}$)	20	20	14	18	100	14	18	18	17	10	18	14	20	25	18	14	20	14
C (RepP-10A)	17	23	20	18	14	100	66	16	15	17	19	14	21	22	27	21	22	13
C (RepP-11A)	15	23	20	18	18	66	100	15	14	22	21	12	17	20	22	20	22	12
D (Rep$_{pSK41}$)	27	19	17	16	18	16	15	100	20	14	18	18	13	14	16	21	18	15
E (RepP-2A)	21	16	15	16	17	15	14	20	100	42	20	62	24	21	22	20	19	19
E (RepP-3A)	22	22	13	13	10	17	22	14	42	100	21	38	23	18	24	23	17	14
E (RepP-16A)	21	10	15	15	18	19	21	18	20	21	100	20	21	20	20	19	16	19
E (RepP-17A)	22	16	23	13	14	14	12	18	62	38	20	100	24	21	26	19	20	20
F (RepP-1A)	11	12	23	20	20	21	17	13	24	23	21	24	100	25	25	18	24	21
F (RepP-4A)	15	22	15	18	25	22	20	14	21	18	20	21	25	100	23	20	27	20
F (RepP-12A)	13	21	18	18	18	27	22	16	22	24	20	26	25	23	100	27	19	21
G (RepP-15A)	21	17	14	19	14	21	20	21	20	23	19	19	18	20	27	100	14	19
H (RepP-7A)	14	20	17	10	20	22	22	18	19	17	16	20	24	27	19	14	100	22
H (RepP-18A)	19	19	15	18	14	13	12	15	19	14	19	20	21	20	21	19	22	100




A
%Taxonomy (PLSDB)
%Taxonomy (IMG/PR)
B
%Environment
%Clade in Taxonomy (PLSDB)
%Clade in Taxonomy (IMG/PR)
%Clade in Environment
C
RIP AA Length vs Plasmid Length
D
RIP %GC vs (RIP - Plasmid) %GC
E
RIP N-ARSC vs (RIP - Plasmid) N-ARSC
F
%MOB type
%MPF type
%Clade in MOB
%Clade in MPF
G
ARG count per plasmid
Defense gene count per plasmid
Taxonomy
Gammaproteobacteria
Betaproteobacteria
Alphaproteobacteria
Bacillota (PLSDB) / Firmicutes (IMG/PR)
Bacteroidota
Campylobacterota
Thermodesulfobacteriota (PLSDB) / Desulfobacterota (IMG/PR)
Actinomycetota (PLSDB) / Actinobacteriota (IMG/PR)
Spirochaetota
Fusobacteriota
Cyanobacteriota
Verrucomicrobiota
Deinococcota
others
Environment
Host-associated; Mammals: Human
Host-associated; Mammals
Host-associated; Arthropoda: Insects
Host-associated; Plants
Environmental; Aquatic; Marine
Environmental; Aquatic; Freshwater
Environmental; Aquatic; Non-marine Saline and Alkaline
Environmental; Terrestrial; Soil
Engineered; Bioreactor
Engineered; Built environment
Engineered; Wastewater / WWTP
others
Clade
A
B
C
D
E
F
G
H
MOB / MPF type
MOBP
MOBQ
MOBF
MOBH
MOBV
MOBC
MOBB
MOBT
multiple
MPF_T
MPF_F
MPF_FATA
MPF_B
MPF_I
MPF_FA
MPF_G
multiple
ARG count
15 - 28
10 - 14
5 - 9
3 - 4
1 - 2
Defense gene count
6 - 9
4 - 5
2 - 3
1


domain arrangement
sWH RIP
WH
dWH RIP
WH1
WH2
UFBoot: 99
aLRT: 89.4
UFBoot: 92
A
pfam: RepL
short RIPs, short plasmids
taxonomy: widely distributed, including Pseudomonadota, Bacteroidota, Gram-positives
environment: host-associated
B
pfam: HTH_36, DUF6017
low %GC plasmids
taxonomy: enriched in Bacillota
environment: host-associated
C
pfam: IncFII_repA
taxonomy: enriched in Gammaproteobacteria
environment: aquatic and plant-associated
D
pfam: RepA_N, DUF6017
low %GC plasmids
taxonomy: enriched in Bacillota
environment: mammal-associated
E
pfam: RPA
long and high %GC plasmids
taxonomy: many Actinomycetota and Acidobacteriota plasmids
enriched in: aquatic (especially) marine environments
F
pfam: TrfA, RepC
long plasmids
taxonomy: enriched in Betaproteobacteria
enriched in: aquatic and plant-associated environments
G
pfam: RepA_C
long and high %GC and plasmids
many Actinomycetota and Acidobacteriota plasmids
enriched in: soil
H
pfam: Rep3_N, Rep3_C
largest clade
diverse size, %GC
taxonomy: widely distributed, including many Bacteroidota
environment: ubiquitous