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]Science and Technology Avant-garde Research of Life (X-star), 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. Their deep evolutionary relationships, however, 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, defining a broad WH RIP superfamily. Using the conserved WH region, we reconstructed a large-scale phylogeny that linked WH RIPs to over 100,000 plasmids, representing approximately half of those analyzed across public databases. This phylogeny resolved eight major clades and the deep split between the single- and double-winged-helix superclades, while revealing clade-specific host and environmental distributions and substantial RIP diversity not captured by current typing tools or annotation schemes. Together, these results overcome the historical host bias of plasmid typing and provide 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]. They are also major vehicles for the global spread of antimicrobial resistance genes (ARGs), representing a critical threat to public health[3–5]. Recent large-scale sequencing efforts have revealed an unexpectedly vast and heterogeneous collection of plasmids[6,7],underscoring the need for a systematic framework to organize this diversity and interpret plasmid evolution coherently.

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 incorporated Inc group names into their classification schemes, improving classification coverage, but still falling short of delivering a macroevolutionary perspective. Replicon typing[9] and plasmid taxonomic units (PTUs)[10,11] rely on nucleotide similarity of rep genes or whole plasmid sequences, respectively, which obscures remote homology among deeply divergent plasmids. Relaxase-based approaches[10,12–14] have expanded classification based on detectable relaxase homology, but exclude more than half of plasmids that lack relaxase genes[8], and therefore cannot provide a universal backbone for plasmid classification. Consequently, current schemes do not fully resolve how plasmids are related across deep evolutionary time. By contrast, in virology, higher-order classification has increasingly been anchored by conserved hallmark genes, enabling molecular evolutionary analysis to be incorporated directly into large-scale taxonomic frameworks[15,16]. This contrast highlights the absence of an equivalent organizing principle for plasmid classification.

Pseudomonas plasmids provide an ideal starting point because this genus spans an exceptionally wide ecological and clinical range from soils and aquatic environments to hospital settings, and its plasmids serve as major vectors connecting environmental and clinical antimicrobial-resistance (AMR) reservoirs[17]. Phenotype-based 14 Inc groups (IncP-1 through IncP-14) have been designated for *Pseudomonas* plasmids and include some of the most clinically significant resistance plasmids identified in *P. aeruginosa* [18,19]. However, despite their importance, some are not only under-characterized but also lack nucleotide sequence information, relying solely on their original incompatibility phenotypes.

Here, we establish a curated and experimentally validated replicon classification system, termed PInc (*Pseudomonas* Incompatibility-informed replicon

classification), which provides a sequence-based framework anchored in historical Inc groups. This framework is grounded in the experimental identification of RIPs from representative *Pseudomonas* plasmids. Focusing on RIPs, we identified a conserved but distantly homologous winged-helix (WH) domain in most PInc groups and many additional plasmids, pointing to a shared evolutionary origin. Building on this observation, we reconstructed an extended phylogeny of WH-containing RIPs and used it to place many previously unclassified plasmids into a unified evolutionary framework. In this way, our study provides a replication-centered basis for higher-order plasmid classification and reveals the diversification and macroevolutionary trajectories of plasmids through their replication systems.

Results and Discussion

Sequencing and molecular characterization of *Pseudomonas* plasmids

We determined the complete nucleotide sequences of six *Pseudomonas* plasmids: Rms139 (IncP-2)[20], Rms163 (IncP-5)[21,22], Rsu2 (IncP-9)[23,24], RP1-1 (IncP-11)[25], R716 (IncP-12)[26], and pMG26 (IncP-13)[27,28]. Although these plasmids represent the archetype members of their respective Inc groups in *Pseudomonas*, no full-length reference sequences had been available, leaving the replication systems of historically important plasmids unresolved (Table 1). For IncP-10 plasmids, because only a partial plasmid sequence of the representative plasmid R91-5[29,30] is available, we selected pPAB546 as a representative (Supplemental Text S1). Sequencing of pMG26, previously assigned as IncP-13 plasmid[31], yielded no circular contigs after removal of reads mapping to the *P. aeruginosa* PAO1 chromosome. Detailed molecular characterization indicated that pMG26 is not a plasmid but an integrative and conjugative element (ICE), and therefore IncP-13 plasmids should be regarded as ICEs (Supplemental Text S1). The other five (Rms139, Rms163, Rsu2, RP1-1, and R716) are shown to be self-transmissible plasmids (Supplemental Text S1). The RIPs and origin of vegetative replication (*oriV*) of these plasmids (Table 1) were identified by sequence- and structure-based searches, and their functions were experimentally validated using mini-replicon constructs (see Methods, Figure S1, S2, S3, and Supplemental Text S1).

PInc classify unclassified plasmids in PLSDB

To provide a biologically grounded and sequence-based framework for *Pseudomonas* plasmids, we constructed a novel replicon-typing scheme named PInc, using experimentally validated RIPs as reference sequences (Table 1). This scheme follows the traditional Inc group numbering and its subgroup lettering (in Greek character) —for instance, PInc-1α corresponds to IncP-1α,

137 and PInc-2 to IncP-2. Because the original Inc were defined on the basis of
138 phenotypic incompatibility in different host lineages such as *Pseudomonas* and
139 Enterobacterales, their names often overlap each other. For example,
140 *Pseudomonas* IncP-1 is equivalent to Enterobacterales IncP (Table 1).
141 Introducing the PInc scheme allows us to retain meaningful correspondence
142 with the original Inc names while eliminating confusion caused by the
143 overlapped or inconsistent Inc labels[32].

144 Some Inc groups were excluded from PInc, because IncP-8[33] and IncP-13 (this
145 study) were identified as ICEs, or due to material unavailability for IncP-14[34].
146 Additionally, certain *Pseudomonas* plasmid groups that are not included in the
147 traditional Inc grouping were incorporated into the PInc system, as they are
148 widely distributed and known to disseminate antimicrobial resistance. These
149 include PInc-15 (PromA group[35,36]), PInc-16 (pSN1216-29-like group[37]), PInc-17
150 (pQBR103-like group[38]), and PInc-18 (pSTY-like group[39]) (Table 1,
151 Supplemental Text S1). PInc subgroups were defined according to subgroups
152 of the Inc groups or newly defined if a group contains a large nucleotide
153 diversity of the RIP genes.

154 Plasmid sequences were collected and the RIP genes in the plasmids were
155 identified based on sequence and structural homology to known RIPs (Figures
156 S1, S2). Reference plasmids (n=87), including those newly sequenced in this
157 study, were selected to cover the 15 original Inc groups and 22 subgroups
158 (Table S1) and their RIP genes were extracted and compiled into the repP
159 database (n = 87, https://doi.org/10.6084/m9.figshare.31883911). Each PInc
160 group and its subgroups were defined as follows: a plasmid was assigned to a
161 group if its RIP gene showed nucleotide similarity to a reference RIP (BLASTn;
162 ≥90% identity and ≥60% length coverage), considering only the best hit.
163 Plasmids harboring multiple distinct RIPs were recognized as multi-replicons
164 (e.g., PInc-10 and PInc-11).

165 Accordingly, we assigned 2,351 plasmids in PLSDB (59,895 plasmids) to PInc
166 based on RIP genes detected using the repP database (Tables S2) and curated
167 into the PseudomonasRepDB database after redundancy removal (Table 2, n =
168 404, https://doi.org/10.6084/m9.figshare.26778175), including 530 plasmids
169 unclassified by the existing tools, PlasmidFinder and 1,514 unclassified by
170 MOB-suite. Thus, PInc substantially expands the coverage of replicon-based
171 classification for *Pseudomonas*-associated plasmids, increasing the number of
172 classifiable plasmids by 29.1 % for PlasmidFinder and 181% for MOB-suite.
173 Notably, PInc-classified plasmids span not only *Pseudomonas* but also diverse
174 *Enterobacterales* hosts, reflecting the broad host range of several PInc
175 plasmids (https://doi.org/10.6084/m9.figshare.26768173). Overall, PInc provides

176 a systematic sequence-based framework that improves classification depth and
177 coverage across plasmids in PLSDB.

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

179 To characterize the genomic and evolutionary diversity of plasmids classified by
180 PInc, we conducted a comprehensive pangenomic analysis across the PInc
181 groups (Tables S2, S3-1~S3-14). Three key patterns emerged. First, several
182 PInc groups (notably PInc-2, PInc-17, and PInc-18) exhibited exceptionally
183 large genome sizes (Figure 1A). Second, specific groups including PInc-4, PInc-
184 10 and PInc-11, displayed distinctive multi-replicon coupling patterns,
185 suggesting long-term modular evolution of their replication systems. Third, PInc-
186 1 stood out for its broad genomic diversity consistent with its presence across
187 diverse environments and host lineages.

188 Plasmid size distribution across PInc groups (Figure 1A) revealed that PInc-2,
189 PInc-17, and PInc-18 harbor markedly larger plasmids than other groups,
190 indicating expanded accessory gene content and potential ecological
191 specialization. %GC content (Figure 1B) varied widely in PInc-1 and PInc-4,
192 suggesting flexible amelioration in multiple hosts[40–42], whereas other groups
193 showed narrow %GC distributions consistent with more specialized ecological
194 niches. Core-genome analysis (Figure 1C) showed that PInc-2, PInc-12, and
195 PInc-17 maintain unusually rich core sets compared with other groups, implying
196 conserved functional backbones within these groups. In contrast, the low core-
197 gene proportion in PInc-1 reflects its high genomic and ecological
198 heterogeneity.

199 Multi-replicon analysis (Figure 1D, Table S2) revealed substantial variation
200 among PInc groups. More than half of PInc-4 (70.9%), PInc-11 (65.2%), PInc-6
201 (60.5%), and PInc-10 (51.0%) plasmids are multi-replicons, whereas more than
202 90% of PInc-1, PInc-5, PInc-7, PInc-9, and PInc-15 plasmids are single-
203 replicon. These rates indicate a distinct evolutionary history of the replicon
204 acquisition and coevolution. Specifically, PInc-4 plasmids are coupled to many
205 types of replicons (Figure 1D, Table S2). In contrast, most multi-replicons of
206 PInc-10 and PInc-11 were coupled to each other or to only a few types of
207 replicons (IncU or rep_cluster_2025, Table S2). Importantly, more than 60% of
208 the PInc-2, PInc-3, PInc-4, PInc-6, PInc-10, and PInc-17 plasmids harbored
209 ARGs, with several of these groups, PInc-2, PInc-6, PInc-10, and PInc-17
210 plasmids, particularly containing clinically important ARGs, such as
211 carbapenem, tigecycline, and colistin resistance genes (Figure 1D). More
212 detailed pangenomic characterization for each PInc was provided in
213 Supplemental Text S1.

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

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

To extend the analysis beyond the sensitivity of BLASTp, we next performed iterative sequence searches using RIPs from the PInc groups as queries to collect more diverse homologs from PLSDB and the UniProt Archive (UniParc). To improve coverage of RIP sequence diversity, we also included nine additional RIPs outside the PInc groups as additional queries (Table S5). The detailed procedures are described in Supplemental Text S1 and Figure S4. In total, this approach retrieved 195,236 sequences, which were classified into 14 homologous groups. This expanded sequence set provided the basis for resolving deeper relationships among plasmid RIPs.

Remote homology analysis reveals evolutionary divergence among plasmid RIPs

Remote homology among RIPs from the 14 homologous groups was assessed using HHblits[43], which performs sensitive sequence-based homology detection through HMM-HMM comparison. With the exception of G-PInc-6, which comprises sequences collected using PInc-6 RIPs as queries, RIPs from the remaining 13 groups shared a homologous region (Figure S5). Based on the significance of homology among groups, these 13 groups could be further divided into two supergroups. One supergroup, comprising seven groups, showed homology with >95% probability in 20 of 21 possible pairwise group comparisons, whereas the other, comprising six groups, did so in all 15 possible comparisons (Figure S5).

Considering available experimental structures and AlphaFold3[44] predictions of RIPs (Figure S6), we classified these two supergroups as single winged-helix (sWH) and double winged-helix (dWH) according to their domain architecture. The sWH supergroup comprises seven groups with a single WH domain, whereas the dWH supergroup comprises six groups with two WH domains. In

252 dWH RIPs, available structures of RepE (plasmid F; PDB: 2Z9O), Pi (R6K;
253 PDB: 2NRA), RepA (pPS10; PDB: 1HKQ), and TrfA (RK2/RP4)[45] all contain two
254 conserved winged-helix domains, here termed WH1 and WH2. By contrast, only
255 two crystal structures are available for sWH RIPs[46] (pSK41, PDB: 5KBJ;
256 pTZ2162, PDB: 4PT7), although AlphaFold3 predicts that the remaining
257 members also adopt a single-WH-domain architecture. HHblits detected
258 significant homology (>80% probability) between the sWH and dWH
259 supergroups in 18 of 42 possible pairwise group comparisons (6 × 7). In all
260 cases, the homologous region mapped to the WH domain of sWH RIPs and to
261 WH2, the C-terminal WH domain, of dWH RIPs.

262 By contrast, RIPs of G-PInc-6 showed no detectable homology to any other
263 groups by HHblits, with probabilities of less than 1% in all comparisons. We
264 therefore examined their evolutionary origin separately by searching against the
265 Pfam database using HHblits. G-PInc-6 RIPs matched PF03090, corresponding
266 to the replicase family, with 98.4% probability. This domain is present in the
267 ColE2 RIP[47], a member of the archaeo-eukaryotic primases (AEPs)[48]
268 superfamily. Consistent with this assignment, G-PInc-6 RIPs[49,50] also contain a
269 PriCT domain implicated in DNA unwinding[47]. Together, these observations
270 identify G-PInc-6 RIPs as AEP-superfamily initiators and therefore as
271 evolutionarily distinct from the WH-containing RIPs.

272 On the basis of this shared homologous domain across the sWH and dWH
273 supergroups, we define these proteins as a WH RIPs superfamily. This
274 framework also provides a basis for interpreting their deeper evolutionary
275 relationships. A previous structure-based study reported deep evolutionary
276 relationships among helix-turn-helix (HTH) domains, including WH domains[51],
277 although plasmid RIPs were not examined in this context. Moreover, the N-
278 terminal WH1 domain of dWH RIPs has been proposed to be evolutionarily
279 related to the WH domain of archaeal Orc1/Cdc6 and eukaryotic Orc proteins[52],
280 both of which function in origin recognition for the chromosomal replication[53]. By
281 contrast, the evolutionary origin of WH2 has remained unresolved. Our analysis
282 identifies homology between WH2 and the WH domain of sWH RIPs,
283 uncovering an unexpected evolutionary link within plasmid replication initiators.
284 Notably, the WH2 domain of pPS10 RepA binds the 3′ end of the iteron and is
285 the major determinant of sequence specificity[45], suggesting that this conserved
286 WH module represents a core origin-recognition element across the WH RIP
287 superfamily. More broadly, the coexistence of such divergent RIPs lineages,
288 including both WH RIPs and AEP-superfamily initiators, within the traditional Inc
289 framework exposes the limitations of historical phenotypic classification and
290 supports the need for a sequence-based PInc system. The conserved WH

291 module across the 13 homologous groups then provided a common anchor for
292 large-scale phylogenetic reconstruction.

293 **WH RIP phylogeny reveals replication initiator diversity in over 100,000**
294 **plasmids**

295 We reconstructed a phylogeny of WH RIPs from the 13 homologous groups
296 using HHblits-defined homologous regions spanning the conserved WH domain
297 and its flanking sequences. The maximum-likelihood tree was built from 4,769
298 representative sequences clustered at 50% identity (Figure 2A, Figure S4B). On
299 the basis of branch support, we defined eight clades (A to H) and 42 subclades
300 using high-confidence thresholds (UFBoot2: ≥95% or SH-aLRT: ≥80%) (Figure
301 2B). At the deepest level, sWH and dWH RIPs formed two superclades
302 separated by a major split in the tree, supported by both UFBoot2 (89.4%) and
303 SH-aLRT (88%). This split was also consistently identified as a phylogenetic
304 root by minimal ancestor deviation analysis[54]. These results define a deeply
305 structured WH RIP phylogeny and suggest an early evolutionary transition
306 involving either gain of an additional WH domain or loss of one ancestral WH
307 domain.

308 We next placed this phylogeny in the context of plasmid diversity by surveying
309 associated plasmids in the public databases PLSDB and IMG/PR. PLSDB is
310 enriched in plasmids from cultivated isolates, whereas IMG/PR is largely
311 composed of metagenome-derived plasmids. Protein sequences predicted from
312 214,575 presumably complete plasmids (59,895 from PLSDB and 154,680 from
313 IMG/PR) were mapped to the closest leaves in the WH RIP tree (Figure 2C).
314 This analysis identified 122,369 phylogeny-linked RIPs associated with 102,879
315 plasmids (47.9% of all analyzed plasmids), with linked plasmids represented in
316 all subclades. Linked plasmids were distributed broadly across the tree, but
317 their abundance varied markedly among leaves. A total of 2,137 leaves (44.8%
318 of all 4,769 leaves) were linked to plasmids, and 162 of these leaves (3.4%)
319 were each associated with more than 100 plasmids (Figure 2D). Together,
320 these leaves accounted for 82,735 unique plasmids (80.4% of phylogeny-linked
321 plasmids). Of the linked plasmids, 15,131 (14.7% of phylogeny-linked plasmids)
322 were multi-replicons carrying multiple WH RIPs (Figure 2E). Most of these multi-
323 replicons were derived from Gammaproteobacteria, although other taxa,
324 including Bacillota, Actinomycetota, Bacteroidota, and Campylobacterota, were
325 also represented. Overall, these observations show that the WH RIP phylogeny
326 captures broad plasmid diversity across all subclades.

327 We then asked how much of this diversity is captured by existing annotation
328 frameworks. Of the 122,369 phylogeny-linked RIPs, 60,202 (49.2%) from

329 57,087 plasmids were newly identified and were not detected by existing
330 replicon typing tools, including PlasmidFinder, MOB-suite, and
331 AcinetobacterPlasmidTyping[55], or by the PInc framework defined here. When
332 mapped onto the WH RIP phylogeny, replicon types were broadly distributed
333 across clades (Figure 2B). In particular, RIPs from 2,768 PInc-assigned
334 plasmids were distributed across 16 leaves spanning seven of the eight clades,
335 except for clade D. Likewise, 44, 33, and 33 leaves were associated with Gram-
336 positive, Enterobacterales, and *Acinetobacter* replicon types, respectively. In
337 total, only 125 leaves (2.5% of all 4,769 leaves) were linked to current typing
338 schemes, indicating that much of the RIP diversity remains uncaptured by
339 existing plasmid typing frameworks. Furthermore, 29,600 RIPs (24.2%) from
340 26,133 plasmids lacked any Pfam domain annotation (Figure 2F), and in clades
341 A, B, and F, more than half of the RIPs were unannotated. Together, these
342 results indicate that a substantial fraction of RIP diversity remains uncaptured
343 by both nucleotide-level replicon typing and protein-level domain annotation.

344 **Structural and sequence diversification across WH RIP clades**

345 We examined the structural diversity of representative WH RIPs from PInc
346 plasmids, focusing on domain organization and predicted three-dimensional
347 structures (Figure 3 and Figure S7). Pfam-based domain annotation, together
348 with WH1/WH2 assignments based on previous studies[45,56] (see Methods),
349 identified WH domains linked to conserved replication-associated domains,
350 including RepA_N/C, Rep3_N/C, and IncFII_repA. The Bac_RepA_C domain,
351 which does not overlap the WH region, has been implicated in interactions with
352 DnaG primase[46]. On the basis of available crystal structures, we inferred that
353 sWH RIPs bind DNA as monomers, whereas dWH RIPs bind DNA as dimers.
354 These observations indicate that WH RIPs combine a conserved WH core with
355 distinct replication-associated domain architecture across clades.

356 Sequence features derived from the multiple alignment used for phylogenetic
357 reconstruction further distinguished clades A–H (Figure S8). In particular, both
358 gap patterns in the alignment and the lengths of the N- and C-terminal regions
359 flanking the conserved WH domain showed clade-specific features, while still
360 exhibiting substantial variation within individual clades, such as clades C and E.
361 Overall, these results highlight structural diversification of WH RIPs across
362 clades, reflecting the diversification of replication systems underlying plasmid
363 evolution.

364 **WH RIP phylogeny reveals host and environmental constraints on plasmid**
365 **dispersal**

366 We examined the taxonomic distribution of WH RIP clades using metadata from
367 PLSDB and IMG/PR. PLSDB follows NCBI taxonomy, whereas IMG/PR follows
368 GTDB taxonomy (Figure 4A). Despite differences in the taxonomic systems and
369 data sources, several clade–taxonomy associations were consistent across
370 both datasets. Clades B and D were strongly associated with Bacillota, clades C
371 and F with Gammaproteobacteria, and clade E with Alphaproteobacteria. Clade
372 H showed the broadest taxonomic distribution and, in IMG/PR, accounted for
373 most Bacteroidota and large fractions of Campylobacterota and Fusobacteriota.
374 This was particularly evident in IMG/PR, which contained 22,390 Bacteroidota
375 plasmids, most of which were assigned to clades A and H. At the subclade
376 level, however, clade H was partitioned into groups with distinct taxonomic
377 associations, indicating that its broad host range is composed of multiple more
378 restricted evolutionary units (Figure S8). Clade G also showed a broad host
379 range, with strong representation in Actinomycetota. These patterns indicate
380 that WH RIP evolution is structured by host lineage, while some clades have
381 undergone broader host transitions than others.

382 For environmental distribution, we analyzed IMG/PR-linked plasmids, for which
383 environmental metadata were available from metagenomic and
384 metatranscriptomic datasets. Environmental distributions also showed clear
385 clade specificity (Figure 4B). Clade H was broadly distributed and accounted for
386 a major fraction of RIPs across most environments, whereas clade composition
387 varied substantially among habitat categories. This broad distribution also
388 masked substantial heterogeneity within clade H, as individual subclades
389 showed more specific ecological associations (Figures 2A, S8). Host-associated
390 environments, including human, other mammals, and insects, as well as the
391 built environment, were dominated primarily by clades H and A. By contrast,
392 marine, freshwater, soil, saline and alkaline environments, as well as plant-
393 associated and bioreactor samples, showed more diverse clade mixtures, with
394 increased contributions from clades C, E, F and G. Among these, clade E was
395 more prominent in marine-associated plasmids, whereas clades C and F
396 contributed more strongly to freshwater- and plant-associated environments,
397 and clade G was more strongly associated with soil. These observations
398 suggest that plasmid ecological distributions reflect clade-specific constraints on
399 dispersal and environmental association. We therefore asked whether the same
400 clades also differed in broader architectural, compositional and functional
401 properties.

402 **Distinct architectural and functional profiles of the eight WH RIP clades**

403 The eight WH RIP clades differed markedly in plasmid size, RIP length, GC
404 composition and isoelectric point (pI) (Figure 4C–E). Clades A and H occupied

405 the lower end of the plasmid size range but differed substantially in RIP length.
406 Clade A was associated with the smallest plasmids and the shortest RIPs
407 (median plasmid length 3.2 kb; median RIP length 185 aa), whereas clade H
408 also tended to occur in relatively small plasmids (median 6.8 kb) but carried
409 substantially longer RIPs (median 325 aa). By contrast, the remaining clades
410 were associated with progressively larger plasmids, with median lengths
411 ranging from 19.4 kb to 144 kb, indicating that plasmid size is itself a
412 characteristic feature of WH RIP clades. GC composition was likewise strongly
413 structured by clade identity. Clades B and D were associated with the lowest-
414 GC plasmids (median 33.4% and 33.6%, respectively), which may reflect their
415 taxonomic bias towards Bacillota, whereas clades E and G showed the highest
416 plasmid GC values. RIP GC content broadly tracked plasmid-wide GC content,
417 but clade C was distinctive in showing higher GC in RIPs than in their plasmid
418 backgrounds, with a similar but weaker tendency in clades E, F and G.
419 Differences in pI were even more pronounced. Clade C showed a marked
420 positive shift relative to the plasmid proteome, clade D showed the opposite
421 trend, and clades E, F and G were more moderately shifted towards higher pI.
422 The elevated pI of clade C RIPs is likely explained by enrichment of positively
423 charged residues, particularly lysine and arginine, and may reflect
424 compositional adaptation related to DNA binding (Figure S9). Together, these
425 patterns indicate that each WH RIP clade is associated with a distinct
426 combination of plasmid architecture, sequence composition and RIP
427 biochemical properties. We then asked whether this clade specificity extended
428 to transfer-related and accessory functions.

429 Mobility and accessory functions also varied sharply across WH RIP clades.
430 MOBP and MOBQ were detected across most clades, whereas other MOB
431 classes showed much stronger clade specificity (Figure 4F). Notably, clade A
432 was enriched in MOBV despite its association with small plasmids, suggesting
433 retained mobilization capacity. By contrast, clades C, F and G were enriched in
434 MOBF, whereas clades E and H showed relatively stronger contributions from
435 MOBQ. MPF repertoires also differed among clades. MPF modules were nearly
436 absent from clade A, whereas clades C and F were dominated by MPF_F and
437 clade D by MPF_FATA. AMR and defense systems showed similarly uneven
438 distributions across clades (Figure 4G). Clades C, D and F were relatively
439 enriched in ARG-rich plasmids, whereas clades A, B, G and H were dominated
440 by plasmids with few ARGs (Figure 4H). Defense system content showed a
441 partly overlapping but distinct pattern, with clades B, C, D and F more frequently
442 carrying multiple defense systems, whereas clades A and H showed low
443 defense systems content (Figure 4I). Metal resistance genes also showed
444 strong clade specificity (Figure 4J). Clades C and F were enriched in plasmids

445 carrying multiple metal resistance determinants, with tellurite resistance being
446 especially prominent in clade F. Clade E was distinctive in its association with
447 cadmium resistance, whereas clade D showed a more selective pattern
448 dominated by cadmium and arsenic resistance. These patterns indicate that
449 WH RIP clades differ in the functional modules that shape plasmid transmission
450 and adaptation.

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

452 To summarize the major features associated with each lineage, we integrated
453 the structural, compositional, taxonomic and ecological characteristics of the
454 eight WH RIP clades (Figure 5). This comparison shows that the major split
455 between the sWH and dWH superclades is accompanied by broad differences
456 in domain organization and plasmid architecture, whereas individual clades are
457 further distinguished by characteristic combinations of plasmid size, RIP length,
458 GC composition, biochemical properties, host distribution, environmental
459 association and accessory functional traits. Notably, clades G and H, as well as
460 clades B, C and D, form well-supported higher-level assemblages, with UFBoot
461 support values of 99 and 92, respectively, suggesting that these sets of clades
462 represent coherent evolutionary subdivisions within the WH RIP phylogeny.
463 Together, these patterns support the view that the eight WH RIP clades
464 represent distinct evolutionary trajectories of plasmid replication systems.
465 Extending this replication-centred framework to additional initiator classes,
466 including AEP-type and rolling-circle replication proteins, should further broaden
467 the evolutionary map of plasmids.

468 **Resources for WH RIP identification and plasmid classification**

469 We provide a software package for identification and classification of WH RIPs
470 based on the protein homology relationships defined in this study
471 (https://github.com/yosuken/WHTreeMapper). In addition, although the analyses
472 described above relied on amino acid-level detection of WH domains, practical
473 application to sequence-based identification requires extending detection to the
474 nucleotide level. To this end, total of 48,743 nucleotide sequences
475 corresponding to the conserved WH region of RIP genes
476 (https://doi.org/10.6084/m9.figshare.31969347), including clades A to H and
477 their subclades, were extracted from PLSDB plasmids and curated into the
478 WHRepDB database after redundancy removal (Table 2, n = 12,923,
479 https://doi.org/10.6084/m9.figshare.31883887). This database enables high-
480 resolution detection and typing of plasmids, even from complex metagenomic
481 data, thereby facilitating comprehensive characterization of plasmid diversity,
482 host range, and transmission dynamics across diverse environments. As such,
483 it represents a valuable resource for advancing plasmid surveillance and

484 improving our understanding of the ecological and epidemiological roles of
485 plasmids in the dissemination of antimicrobial resistance.

486 Conclusion

487 In this study, we establish PInc, a curated and experimentally grounded replicon
488 classification framework anchored in historically defined incompatibility groups
489 of *Pseudomonas* plasmids. By coupling experimental identification of replication
490 initiation proteins (RIPs) with sequence-based classification, PInc expands the
491 repertoire of identifiable plasmid replicons and provides a framework for
492 classifying plasmids through their replication systems.

493 Homology searches beyond PInc plasmids revealed that most replication
494 initiators analysed here share a conserved winged-helix (WH) domain, defining
495 a WH RIP superfamily that unifies a broad range of plasmid replication systems.
496 Remote homology analysis further resolved this superfamily into single- and
497 double-winged-helix superclades and uncovered homology between the WH
498 domain of single-WH RIPs and the WH2 domain of double-WH RIPs,
499 suggesting a shared origin-recognition module across diverse plasmids. By
500 contrast, PInc-6 RIPs were assigned to the AEP superfamily, showing that
501 plasmids grouped historically by incompatibility can encode fundamentally
502 distinct classes of replication initiators.

503 Using the conserved WH region, we reconstructed a large-scale phylogeny that
504 linked WH RIPs to over 100,000 plasmids across public databases. This
505 phylogeny resolved eight major clades, revealed clade-specific host and
506 environmental distributions, and captured substantial RIP diversity not
507 represented in current typing tools or annotation schemes. It also placed
508 several lineages that are poorly covered by existing plasmid typing systems,
509 including Bacteroidota, Campylobacterota, Thermodesulfobacterota and
510 Fusobacteriota, within a unified WH RIP phylogeny, while bringing together
511 plasmids from cultivated isolates and metagenomic datasets within a single
512 evolutionary framework. More broadly, our results suggest that RIP-based
513 phylogeny can provide the long-missing evolutionary backbone for higher-order
514 plasmid classification and begin to do for plasmids what hallmark gene-based
515 phylogenies have enabled in virology. At present, this framework is limited to
516 plasmids carrying WH RIPs. Expanding comparable phylogenetic frameworks
517 across additional initiator classes will be an important next step towards a more
518 comprehensive evolutionary map of bacterial mobile DNA.

519 Methods

520 Bacterial strains, plasmids, and culture conditions

521 The bacterial strains and plasmids used are listed in Table S6. *E. coli* and
522 *Pseudomonas* strains were cultured in Luria broth (LB) or M9 broth[57] at 30°C
523 (*P. putida* KT2440) or 37°C (*P. aeruginosa* PAO1 and its derivatives). LB plates
524 containing 1.5% agar were used for filter mating. Ampicillin (Ap, 50 μg/mL),
525 piperacillin (Pip, 16 μg/mL), rifampicin (Rif, 50 μg/mL), streptomycin (Sm 150
526 μg/mL), and tetracycline (Tc, 50 μg/mL for the others) were added to the
527 medium.

528 **DNA manipulation**

529 Total DNA from the bacterial strains was extracted using a NucleoSpin Tissue
530 Kit (TAKARA BIO, Japan). Polymerase chain reaction (PCR) was performed on
531 a T100 thermal cycler (Bio-Rad Laboratories, USA) using the KOD One PCR
532 Master Mix (TOYOBO, Japan). The primers used in this study are listed in
533 Table S7. The following amplification conditions were used: 30 cycles at 98°C
534 for 10 s, 67°C for 5 s, and 68°C for 1 s. The pUC-GW plasmid with DNA
535 fragments including putative RIP genes and *oriV*s of each plasmid, except those
536 for RP1-1 and pMG26, was synthesised by the custom service of Azenta Life
537 Science, USA. Restriction enzymes (New England Biolabs, USA or TAKARA
538 BIO), the HiYieldTM Gel/PCR DNA fragments Extraction kit (RBC Bioscience,
539 Taiwan), NEBuilder Hifi DNA Assembly system (New England Biolabs), and
540 competent *E. coli* JM109 and DH5α cells (RBC Bioscience) were used for
541 cloning the DNA fragments. Plasmid DNA from the bacterial strains was
542 extracted using the NucleoSpin Plasmid EasyPure Kit (TAKARA BIO).
543 Electroporation to introduce the plasmids into different bacterial strains was
544 performed using a MicroPulser electroporator (Bio-Rad Laboratories). All other
545 procedures were performed according to standard methods[57].

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

547 The nucleotide sequences of the plasmid DNAs were determined as follows:
548 whole genomic DNA extracted from their host (*P. aeruginosa* PAO1808, a
549 derivative strain of PAO1[58]) was sequenced using the HiSeq2500 platform
550 (Illumina, USA). Trimmed high-quality short reads (read length >140 bp and
551 quality score >15) were assembled using SPAdes v3.14.0 using the plasmid
552 option[59]. If circular contig(s) could not be found, the host chromosomal DNA
553 was removed by mapping the resultant contigs on the host genome sequences
554 [their deposited sequences, that is, NCBI RefSeq accession number
555 NC_002516 (*P. aeruginosa* PAO1)] using Geneious Prime software v2020.01[60].
556 Subsequently, the plasmid reads were extracted from the remaining contigs
557 using SAMtools v1.7[61] and SeqKit v0.8.0[62] and reassembled using SPAdes[63].
558 For RP1-1 and pMG26, gaps in plasmids were closed by PCR with the

559 designed primer sets based on the ends of contigs and Sanger sequencing of
560 the PCR products.

561 **Gene annotation of plasmids**

562 Gene annotations of the plasmids were performed using DFAST-core v1.2.5[64]
563 with an in-house database of plasmid sequences collected from NCBI RefSeq
564 and then corrected manually. Synteny analyses of plasmids were performed
565 using Easyfig v2.2.2[65]. Other *in silico* analyses were performed using Geneious
566 Prime v2020.01. The annotated plasmid sequences were deposited under the
567 following GenBank accession numbers: LC685025, LC685026, LC685027,
568 LC685593, and LC700336.

569 **Detection and phenotype identification of ARGs**

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

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

580 To identify the genes encoding the RIP and origin of vegetative replication (*oriV*)
581 regions of Rms139, Rms163, RP1-1, and R716, two prediction approaches
582 were used; after coding sequence (CDS) annotations using DFAST-core,
583 protein.faa files generated by DFAST-core were used for the following: (i)
584 nucleotide sequence-based searches were performed using BLASTn; (ii) amino
585 acid sequence-based protein structure prediction and the structure-based
586 searches were performed as described later. The GC-skew value, which has
587 been commonly used to measure the asymmetry in the composition of two
588 nucleotides, guanine (G) and cytosine (C), along a DNA strand and is defined
589 as (G-C)/(G+C), of the plasmids was used by Webskew
590 (https://genskew.csb.univie.ac.at/webskew) to detect the changes in the GC
591 skew, which indicates the DNA replication origins or termini[66,67]. Subsequently,
592 the nucleotide or amino acid sequences of putative CDSs around the above
593 GC-skew shift point(s) were subjected to methods (i) or (ii). The putative
594 promoter region of the RIP gene was predicted using BPROM
595 (http://www.softberry.com/berry.phtml?topic=bprom&group=help&subgroup=gfi

596 ndb) [68]. Subsequently, the predicted RIP and *oriV* sequences of each PInc
597 groups were cloned into the pUC-GW-Amp vector (Azenta Life Science).
598 Transformation of *P. aeruginosa* PAO1808 was performed with each of the
599 resultant plasmids (at most 472 ng/μL), and the cultures were spread onto
600 LB+Pip plates. After incubation at 37ºC overnight, colonies on the plates were
601 isolated. Next, the presence of the mini-plasmid was confirmed by PCR with
602 specific primers after extraction of total DNA of transformants.

603 **Protein structure prediction of RIPs**

604 AlphaFold3-mediated protein structure prediction was performed using the
605 amino acid sequences of the RIP gene products in representative
606 *Pseudomonas* plasmids (https://doi.org/10.6084/m9.figshare.26778898) on the
607 AlphaFold server[44]. A structure-based protein homology search using the
608 predicted structural models was performed for the PDB100 database on the
609 FoldSeek server in 3Di/AA mode[69]. The predicted and experimental structures
610 were visualised and compared using PyMOL v3.1.3.1. The predicted structures
611 were coloured based on the root mean square deviation (RMSD) from the
612 reference experimental structures, i.e., the Pi or Rep19A[46,70] proteins, which are
613 RIPs of the R6K and pCA-347 plasmids, respectively, using the ColorByRMSD
614 script (https://pymolwiki.org/index.php/ColorByRMSD). The distances between
615 aligned C-alpha atom pairs were stored as B-factors of these residues, which
616 are coloured using a colour spectrum, with blue specifying the minimum
617 pairwise RMSD and red indicating the maximum. Unaligned residues were
618 coloured gray.

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

621 Complete plasmid sequences were downloaded in FASTA format using the
622 Entrez Programming Utilities (E-utilities;
623 https://www.ncbi.nlm.nih.gov/books/NBK25501/) from a public database
624 (PLSDB 2023_11_23_v2) (https://doi.org/10.6084/m9.figshare.26768173)[7,71].
625 To classify the plasmids in the PLSDB based on their rep sequence similarity,
626 the RIP gene sequences were detected using Staramr v0.10.0
627 (https://github.com/phac-nml/staramr) and rep databases of MOB-suite v3.1.8
628 updated on 2023-12-12 (https://github.com/phac-nml/mob-suite)
629 (https://doi.org/10.6084/m9.figshare.26768170)[72], PlasmidFinder v2.1 updated
630 on 2023-12-04 (https://bitbucket.org/genomicepidemiology/plasmidfinder_db)
631 (https://doi.org/10.6084/m9.figshare.26778199)[73], our custom-made
632 *Pseudomonas* plasmid RIP gene database (termed repP and
633 PseudomonasRepDB database, https://doi.org/10.6084/m9.figshare.31883911
634 and https://doi.org/10.6084/m9.figshare.26778175), and our custom-made

635 plasmid RIP WH database (termed WHRepDB database,
636 https://doi.org/10.6084/m9.figshare.31883887) with the following criteria: ≥90%
637 identity and ≥60% coverage (Table 2). The Tus protein, known for binding to
638 DNA replication terminus (*ter*) sites of replicons, encoded by rep_cluster_1115
639 in the MOB-suite database, is not included in our custom-made RIP reference
640 sequences. The repP database was constructed using experimentally identified
641 RIP genes from the 87 reference *Pseudomonas* plasmids shown in Table S1.
642 Furthermore, duplicated RIP sequences between the MOB-suite and repP
643 databases (Table S6) and incorrect RIP sequences, including uncertain IncP
644 and rep_cluster_1115, were manually removed from the MOB-suite database.
645 The modified MOB-suite database
646 (https://doi.org/10.6084/m9.figshare.26768167) was used to detect multi-
647 replicon *Pseudomonas* plasmids, which are single plasmids that can belong to
648 more than one group other than the PInc groups. Among the candidate PInc-1
649 (IncP-1) plasmid sequences based on the similarity to the RIP sequences,
650 artificially constructed sequences not explicitly labelled as such ("other
651 entries;other sequences;artificial sequences") in the NCBI taxonomy database
652 (https://www.ncbi.nlm.nih.gov/taxonomy) were also present. We identified six
653 PInc-1 plasmids (PD584_L3, PD689_delTT1, pJM105, pJM50, pRK2013, and
654 pUZ8002) as artificially constructed sequences based on a literature survey and
655 removed them from the putative PInc-1 plasmid set.

656 For the final dataset, the sequence length (bp) and GC content (%, %GC) were
657 calculated using SeqKit v2.3.0 (https://bioinf.shenwei.me/seqkit/)[62]. Plasmid
658 sequences belonging to each PInc groups were re-annotated using DFAST-
659 core, with a reference genome from each group to ensure consistency in the
660 functional annotations.

661 **Comparative genomics of plasmids**

662 Comparative genomics of all the plasmids of the PInc groups were performed
663 using PIRATE[74] (https://github.com/SionBayliss/PIRATE). Here, we used
664 DIAMOND[75] v0.9.14 rather than BLAST[76] to identify homologous genes (protein
665 families) in PIRATE. PIRATE was run on default parameters with a high-scoring
666 segment pair (HSP) query length threshold of 0.5, a sequence similarity step
667 range of 10, 20, 30, 40, 50, 60, 70, 80, 90, 95, and 98%, and an inflation value
668 of 2 in MCL (v14.137), a Markov cluster algorithm (https://micans.org/mcl/). The
669 core genes in each PInc group were those that were present in more than 95%
670 of the plasmids included in the pangenome analysis. Statistical computing and
671 result visualisation were performed using R v4.1.3 (https://www.r-project.org/).
672 The output of SeqKit and PIRATE are available in Figshare
673 (https://doi.org/10.6084/m9.figshare.26789416).

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

675 Detailed methods are provided in Supplemental Text S1 and illustrated in
676 Figure S6. In brief, to analyse the evolutionary relationships of RIPs, we
677 identified 2,374 RIPs that originated from 2,332 PInc plasmids. Subsequently,
678 251 non-redundant proteins at 100% identity were extracted using MMseqs[77]
679 cluster v15-6f452. Sequence homology between the non-redundant proteins
680 was assessed using BLASTp[76], and the resultant hits were visualised via CPAP
681 (https://github.com/yosuken/CPAP). To detect remote homology between PInc
682 groups and collect comprehensive RIP sequences, we extended the sequence
683 diversity through a multistep homology search process. First, we searched
684 ORFs of PLSDB plasmids predicted using orfM[78] (n = 20,331,985) for
685 homologous RIPs using hmmsearch in HMMER[79] v3.3.2. In addition, 8
686 additional RIPs sequences that are distant from RIPs were manually picked up
687 and homology searches using diamond blastp against the ORFs of PLSDB
688 plasmids were performed. This process generated 19 homologous groups of
689 RIPs. Second, UniParc proteins (downloaded in Feb 2024; n = 607,912,929)
690 were searched with hmmsearch using 19 HMMs generated from the
691 homologous groups. We merged the UniParc sequences into 19 homologous
692 groups, and filtered out potential contamination of other functional proteins,
693 including various transcriptional factors, using hmmsearch against the Pfam
694 database. After merging the 19 homologous groups into 14 based on the
695 UniParc sequence overlapping, we selected 5,488 representative sequences for
696 the 14 homologous groups. To investigate whether these groups exhibit
697 homology with each other, a sensitive hidden Markov model HMM–HMM
698 comparison was performed using HHblits[80] v3.3.0. We extracted the
699 homologous regions of the alignments of the 13 groups (containing 6,793
700 sequences in total). These regions were aligned using MAFFT[81] v7.455.

701 After removing gappy sequences in the alignment, a maximum-likelihood
702 phylogeny of WH RIPs was reconstructed from the resulting 150-position
703 alignment using IQ-TREE[82] v2.2.2.6. We used the parameters "--perturb 0.2 --
704 nstop 500", following the software manual recommendation for datasets
705 containing many short sequences. The best fit model (Q.pfam+F+R10) was
706 selected using ModelFinder[83]. Branch support was assessed with UFBoot2[84]
707 and SH-aLRT[85] in IQ-TREE. We discarded the abnormally long branches using
708 TreeShrink[86]. To evaluate the root of the WH RIP phylogeny independently of
709 domain-based interpretation, we applied minimal ancestor deviation (MAD)
710 rooting to the unrooted maximum-likelihood tree using MAD[54] v2.2 with default
711 settings. The MAD-inferred root was then compared with the major split
712 separating the sWH and dWH superclades. The final tree was rooted at the
713 sWH-dWH split using gotree[87] v0.4.0.

714 **Recruitment of associated plasmids onto the WH RIP phylogeny**

715 The tree-associated plasmids from PLSDB[7] (version 2023_11_23_v2) and
716 IMG/PR[6] (version 2023-08-08_1) were identified as follows. Open reading
717 frames (ORFs) were predicted using Prodigal[88]. Following the developer's
718 recommendations, plasmids longer than 100 kb were processed in the default
719 mode, whereas plasmids of 100 kb or shorter were processed in metagenomic
720 mode "--meta". The predicted ORFs were then searched against HMM profiles
721 representing the 42 WH RIP subclades using PiPP v0.3.0
722 (https://github.com/yosuken/PiPP), which runs hmmsearch in HMMER v3.3.2[79],
723 with the options "--evalue 1e-5 --evaluedom 1e-2 --minhmmcovdom 0.2 --
724 minhmmcov 0.8 --only-detect". For each best hit, the region of the ORF covered
725 by the HMM match was extracted. These extracted regions were then assigned
726 to the aligned representative WH RIP sequences used for phylogenetic
727 reconstruction (n = 4,769) using diamond[75] blastp (v2.1.9) with the options "--
728 ultra-sensitive --id 40 --subject-cover 80 --max-target-seqs 1 --dbsize 1e+9 --
729 evalue 1e-5". The extracted regions mapped to the tree were then organized at
730 both the ORF and plasmid levels. For plasmid-level analyses, mapped WH RIP-
731 containing ORFs were grouped according to their source plasmids, and
732 plasmids carrying two or more such ORFs were classified as multi-replicons.
733 For summaries at the clade, subclade, or individual RIP level, plasmids were
734 deduplicated within each category, so that multiple ORFs from the same
735 plasmid assigned to the same category were counted once. The number of WH
736 replicons per plasmid, however, was defined by the total number of mapped
737 WH RIP-containing ORFs detected in that plasmid.

738 **Plasmid metadata and functional annotation associated with the WH RIP**
739 **phylogeny**

740 Taxonomic metadata for tree-associated plasmids were obtained from the
741 source databases: NCBI taxonomy for PLSDB and GTDB taxonomy for
742 IMG/PR. Taxonomic distributions were summarized using the manually curated
743 higher-rank host categories shown in Figure 4A, with class-level categories
744 used for the phylum Pseudomonadota and phylum-level categories used for
745 several other bacterial lineages. Because nomenclature differs between the two
746 databases, original names were retained for each source but interpreted as
747 corresponding higher-level lineages when appropriate (e.g., Bacillota in PLSDB
748 versus Firmicutes in IMG/PR, Actinomycetota versus Actinobacteriota, and
749 Thermodesulfobacteriota versus Desulfobacterota). Low-frequency taxa were
750 grouped into an "others" category for visualization.

751 Environmental analyses were performed using IMG/PR-linked plasmids only,
752 because environmental metadata was available for these plasmids. Habitat
753 categories shown in Figure 4B were defined based on the original metadata
754 labels, including host-associated environments, aquatic environments (marine,
755 freshwater, and non-marine saline and alkaline), terrestrial soil, engineered
756 environments (bioreactor, built environment, and wastewater/WWTP), and
757 "others". Records annotated as "Engineered; Modeled; Simulated communities"
758 were treated as lacking environmental information and excluded from this
759 analysis, because they do not specify a source habitat.

760 Replicon typing was performed using PlasmidFinder[9] v2.1 updated on 2023-12-
761 04, AcinetobacterPlasmidTyping with database version feb2025_v3, and
762 PseudomonasRepDB for PInc assignment. For all three databases, hits were
763 identified using PlasmidFinder with a minimum nucleotide identity threshold of
764 90% "--threshold 90" and a minimum coverage of 60% "--mincov 60". Additional
765 replicon typing was performed using MOB-typer in MOB-suite v3.1.8 with the
766 equivalent options "--min_rep_ident 90 --min_rep_cov 60". Pfam annotations of
767 tree-associated RIPs were identified using hmmsearch in HMMER v3.3.2
768 against the Pfam v38.1 database and parsed using parse_hmmsearch.rb
769 (https://github.com/yosuken/parse_hmmsearch) with the thresholds "--gene-
770 evalue 1e-5 --evalue 1e-2 --min-ali-cov-dom 0.2 --min-ali-cov 0.5". For
771 visualization of the annotation rings in Figure 2A, taxonomic, environmental,
772 and replicon-typing information associated with mapped plasmids was
773 summarized as stacked-bar compositions for each tree-linked leaf. The WH RIP
774 phylogeny and its annotation rings in Figure 2A were visualized using iTOL v7[89].

775 MOB and MPF classes were identified using MacSyFinder[90] v2.1.4 with the
776 CONJScan/Plasmids models[91] under default settings. For MOB classification,
777 hits with gene_name annotated as T4SS_MOB(X) and model_fqn annotated as
778 CONJScan/Plasmids/MOB were used. For MPF classification, hits with
779 model_fqn annotated as CONJScan/Plasmids/T4SS_type(X) were used. For
780 Figures 4F and 4G, MOB and MPF annotations were summarized at the
781 plasmid level for clade-associated plasmids. Plasmids carrying two or more
782 MOB classes or MPF classes were assigned to the category "multiple".
783 Plasmids without detectable MOB or MPF annotations were retained in the
784 calculations of clade-wise proportions but were not displayed as separate
785 categories. Antimicrobial resistance (AMR) genes were identified using
786 AMRFinderPlus v3.12.8 with database version 2024-07-22.1 under default
787 settings. Hits annotated with Element subtype = AMR were retained. Metal
788 resistance genes were identified in the same manner as AMR genes, except
789 that hits annotated with Element subtype = METAL were retained. For Figure
790 4J, metal resistance was summarized at the plasmid level for clade-associated

plasmids. The number of resistant metals per plasmid was defined as the number of distinct metal resistance classes detected in each plasmid, irrespective of the number of genes assigned to each class. Defense systems were identified using DefenseFinder[92] v1.3.0 with default settings. For Figure 4I, defense systems were summarized at the plasmid level for clade-associated plasmids, and the number of defense systems per plasmid was defined on a system basis. Anti-defense systems were excluded from these counts. Plasmids were grouped into the bins shown in Figure 4I.

For Figures 4C–E, analyses were performed at the level of clade-associated RIPs. Plasmid length was defined as the total nucleotide length of the source plasmid, and RIP length as the amino-acid length of the corresponding RIP. RIP GC content was calculated from the nucleotide sequence of the RIP-encoding CDS, whereas plasmid GC content was calculated from the full plasmid nucleotide sequence. The RIP–plasmid GC difference was defined as RIP %GC minus plasmid %GC. The isoelectric point (pI) of RIPs and plasmid-encoded proteins was calculated using pepstats in EMBOSS v6.6.0 with default settings. Plasmid proteome pI was defined as the median pI of all predicted proteins encoded by each plasmid, and the RIP–plasmid proteome pI difference was defined as RIP pI minus plasmid proteome pI. Unless otherwise noted, analyses were performed either at the plasmid level or at the RIP level, depending on the panel. Figures 2C–E and 4F–J summarize plasmid-level properties, whereas Figure 2F and Figures 4C–E summarize RIP-level properties. Figures 2A and 2B are based on the representative WH RIP sequences used for phylogenetic reconstruction.

Data visualization for Figures 2 and 4 was performed in R v4.1.2 using the ggplot2 package, except that phylogenetic tree visualization was carried out using iTOL. Figures 2, 4, and 5 were subsequently assembled and finalized in Adobe Illustrator based on the resulting plots. The phylogenetic tree, multiple sequence alignment, associated plasmid metadata, and iTOL visualization files are available in Figshare (https://doi.org/10.6084/m9.figshare.28837559).

Data Availability

All datasets generated and analyzed during this study have been deposited in publicly accessible repositories and are freely available to reviewers as follows:

Replicon reference database (repP database): Experimentally validated replication initiation protein (RIP) sequences from representative *Pseudomonas* plasmids used for PInc classification are available at Figshare: https://doi.org/10.6084/m9.figshare.31883911 (nucleotides), https://doi.org/10.6084/m9.figshare.26778898 (amino acids).

Pseudomonas plasmid dataset with PInc classification (PseudomonasRepDB): Curated plasmid dataset derived from PLSDB, including PInc classification results and associated annotations: https://doi.org/10.6084/m9.figshare.26778175.
WH RIP database (WHRepDB): Database of winged-helix (WH) domain-containing replication initiation proteins used for large-scale homology search and classification: https://doi.org/10.6084/m9.figshare.31883887.
WH-domain sequence dataset: A total of 48,743 WH-domain sequences identified from RIP genes and used for downstream phylogenetic and structural analyses: https://doi.org/10.6084/m9.figshare.31969347.
Phylogenetic datasets and tree files: Multiple sequence alignment, maximum-likelihood phylogenetic tree, associated plasmid metadata, and iTOL visualization files: https://doi.org/10.6084/m9.figshare.28837559.
PLSDB-derived plasmid dataset (version 2023_11_23_v2): https://doi.org/10.6084/m9.figshare.26768173.
Original and modified MOB-suite database used for replicon detection: https://doi.org/10.6084/m9.figshare.26768170 (original), https://doi.org/10.6084/m9.figshare.26768167 (modified)
PlasmidFinder database used in this study: https://doi.org/10.6084/m9.figshare.26778199.
Representative rep
Additional analysis outputs (pangenome analysis, sequence statistics, etc.): https://doi.org/10.6084/m9.figshare.26789416.
The annotated plasmid sequences have been deposited in the GenBank/EMBL/DDBJ under accession numbers LC685025, LC685026, LC685027, LC685593, and LC700336.

Supplementary Data Statement

Supplementary Data (Supplemental Text S1, Supplemental Figures: Figs. S1-S13, and Supplemental Tables S1-S9) are available online.

Contributions

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

Acknowledgements

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

Funding

This work was supported by grants (JP25wm0225029 to Y. Nishimura; JP25wm0225029 to Y. Tsuda; JP25gm1610007 to A. Hirabayashi; JP25fk0108665, JP25fk0108683, JP25fk0108712, JP26fk0108755, JP26fk0108756, JP25gm1610003, JP25wm0225029, JP25wm0225054, and JP25wm0325076 to M. Suzuki; JP25wm0225029 and JP26fk0108755 to M. Shintani) from the Japan Agency for Medical Research and Development (AMED), grants (JPMJAX21BK to Y. Nishimura; 19H02869 and 19H05679 to H. Nojiri; 20K10436 and JPMJCR20H1 to H. Suzuki; 22KK0058 and 23K06556 to M. Suzuki; 16H06279, 19H02869, 19H05686, and JP23H02124 to M. Shintani) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, grants to Y. Nishimura, M. Tokuda, X. Hui, C. Suzuki-Minakuchi, H. Nojiri, H. Suzuki, M. Suzuki, and M. Shintani from Consortium for the Exploration of Microbial Functions of Ohsumi Frontier Science Foundation, Japan, a grant from Asahi Glass Foundation to M. Shintani, a grant to M. Shintani from Institute of Fermentation Osaka (L-2023-1-002), National University Corporation Shizuoka University, Japan, and a grant to M. Shintani from the Research Institute of Green Science and Technology Fund for Research Project Support (2023-RIGST-23104), National University Corporation Shizuoka University, Japan.

Conflict of interests

The authors declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

During the preparation of this work, the authors used ChatGPT (OpenAI) for language editing and improving the clarity of English expression. The authors also used GitHub Copilot (GitHub/Microsoft) and Claude Code (Anthropic) to assist in the development of parts of the code. All generated code and final outputs were reviewed and verified by the authors, who take full responsibility for them. These tools were not used to generate scientific content or interpret

data. All content was critically reviewed and edited by the authors, who take full responsibility for the final manuscript.

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	[93]
PInc-2 (IncP-2)	Rms139* [LC653116]	MOB_P MPF_I	87	2	E	[20,94]
PInc-3 (IncP-3, IncA/C)	pRA1 (IncA) [FJ705807] pR55 (IncC) [JQ010984]	MOB_H MPF_F	921	2	E	[95,96]
PInc-4 (IncP-4, IncQ)	RSF1010 [M28829]	MOB_Q -	695	2	F	[97]

PInc-5 **(IncP-5)**	Rms163* [LC685027]	NA NA	2	1	B	[21]
PInc-6 **(IncP-6, IncG)**	Rms149 [AJ877225]	MOB_P -	119	NA	-	[49]
PInc-7 **(IncP-7)**	pCAR1 [AB088420]	MOB_H MPF_F	23	2	H	[98–100]
PInc-9 **(IncP-9)**	pWW0 [AJ344068] NAH7 [AB237655] Rsu2* [LC685593]	MOB_F MPF_T	34	2	A	[24,101,102]
PInc-10 **(IncP-10)**	R91-5 [X54695[d]] pPAB546 [MN433456]	NA NA	49	1	C	[29,30]
PInc-11	RP1-1* [LC700336]	MOB_c	23	1	C	[25]

(IncP-11)		NA				
PInc-12 **(IncP-12)**	R716* [LC685026]	NA NA	9	2	F	[26]
- **(IncP-13)**	pMG26*[e] [LC685025]	MOB_F MPF_G	NA	NA	-	[27,28]
PInc-15 **(PromA)**	pSN1104-11 [AP018707]	MOB_P MPF_T	42	2	G	[103]
PInc-16	pSN1216-29 [AP018710]	MOB_P MPF_T	24	2	E	[37]
PInc-17	pQBR103 [AM235768]	NA NA MOB_F MPF_F	11	2	E	[38]
PInc-18 **(Inc_{pSTY})**	pSTY [CP003962]	MOB_F MPF_F	15	2	H	[39]

	pPT23-C1_1 [AP040308]						

[a]Names of plasmid Inc groups originating from Pseudomonas and/or Enterobacterales are shown in parentheses. Some Inc groups have been historically given different names depending on their host, *Pseudomonas* or *Enterobacterales*, but they represent the same plasmid group (e.g., IncP-1 in *Pseudomonas* corresponds to IncP in *Enterobacterales*). For PInc-18, the name "Inc_{pSTY}" was adopted from a previous study by another research group[39]. [b]Representative plasmid name [GenBank accession number]. Asterisks (*) indicate plasmids whose complete nucleotide sequences were determined in this study (Rms139, Rms163, Rsu2, RP1-1, R716, and pMG26). [c]Classification of MOB classes and MPF types was identified by the MOB-suite with criteria of >=90% identity and >=60% coverage. "-" indicates that the genes involved in conjugation have not been detected, whereas "NA" indicates that the classes and types of MOB or MPF were not identified although the plasmid was self-transmissible. [d]Partial DNA sequences are available. [e]pMG26 was shown to be not a plasmid but an integrative and conjugative element in this study.

Table 2. Representative plasmid replicon databases and their sequence features.

Host bacteria	Database	Latest version	Sequence	Coding region	Reference
Enterococcus* spp.** ***Staphylococcus* spp.** ***Enterobacterales	PlasmidFinder	v2.1 (Apr 19, 2025)	488	Full-length, partial, or not included	[9]
Enterococcus* spp.** ***Staphylococcus* spp.** ***Enterobacterales	MOB-typer*	v3.1.9 (Jun 4, 2024)	2,685	Full-length, partial, or not included	[72,104]
***Acinetobacter* spp.**	AcinetobacterPlasmidTyping	v3.0 (Feb 12, 2025)	555	Full-length	[55]
***Pseudomonas* spp.**	PseudomonasRepDB	v1.0 (This study)	404	Full-length, or partial	This study
All bacteria	WHRepDB	v1.0 (This study)	12,923	Partial (WH domain)	This study

*The "IncP" labels by the MOB-typer database should be interpreted with caution. In some cases, sequences are reported as "IncP", although they do not correspond to the correct incompatibility group classification[32].

913

914 # Figure legend

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

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

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

924 (A) Maximum-likelihood phylogeny of WH RIPs reconstructed from HHblits-defined
925 homologous regions spanning the conserved WH domain and its flanking
926 sequences. The tree was built from 4,769 representative sequences clustered at
927 50% identity. Annotation rings show clade and subclade assignments together with
928 stacked-bar summaries of Pfam annotation, taxonomy, environment, and replicon-
929 typing information associated with linked plasmids. Curved lines drawn in the centre
930 connect WH RIPs co-occurring in the same multi-replicon plasmids; line colours
931 indicate the taxonomy of the corresponding plasmid hosts. Links for plasmids without
932 taxonomic assignment were omitted from this visualization. (B) Summary of the eight
933 clades (A–H) and 42 subclades, with SH-aLRT and UFBoot2 support values shown
934 for major nodes. Representative PInc groups and other known replicon types
935 associated with each clade are indicated. (C) Numbers and proportions of plasmids
936 in PLSDB and IMG/PR linked to leaves in the WH RIP phylogeny. (D) Rank-
937 abundance distribution of linked plasmids per tree leaf. (E) Taxonomic composition
938 of multi-replicon plasmids carrying multiple WH RIPs in PLSDB and IMG/PR,
939 stratified by the number of WH replicons per plasmid. Gray indicates plasmids
940 without taxonomic assignment; in IMG/PR, all gray fractions correspond to
941 taxonomy-undefined plasmids. (F) Pfam annotation profiles of phylogeny-linked RIPs
942 across clades (left) and clade composition within each Pfam category (right). Sample
943 sizes are indicated next to the bars.

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

946 (A) Domain organization of representative Rep proteins from PInc plasmids,
947 including Rep pSN2, Rep p14035, Rep pOXA-48, Rep pSK41, RepP-16A, RepP-4A,
948 RepP-15A, and RepP-7A. N- and C-termini are indicated. Winged-helix (WH)

949 domains (WH1, WH2) and conserved replication-associated domains (e.g., RepL
950 [PF05732], Bac_RepA_C [PF18008], IncFII_repA [PF02387], RepA_N [PF06970],
951 RPA [PF10134], RepC [PF06504], RepA_C [PF04796], Rep3_N [PF01051], Rep3_C
952 [PF21205]) are shown. Predicted oligomeric states (dimer or monomer) are
953 indicated. (B) Pairwise protein sequence identity among representative Rep proteins.
954 (C) Predicted three-dimensional structures of representative Rep proteins and their
955 DNA-binding modes. Structural models highlight the arrangement of WH domains
956 and predicted interactions with DNA, illustrating diversity in oligomeric state and DNA
957 recognition mechanisms among PInc replicons.

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

960 (A) Taxonomic distribution of WH RIP clades among linked plasmids in PLSDB and
961 IMG/PR. Upper panels show the percentage distribution of taxa across clades; lower
962 panels show clade composition within each taxon. PLSDB follows NCBI taxonomy
963 and IMG/PR follows GTDB taxonomy. (B) Environmental distribution of WH RIP
964 clades among IMG/PR-linked plasmids derived from metagenomic and
965 metatranscriptomic datasets. Upper panel shows the percentage distribution of
966 environments across clades; lower panel shows clade composition within each
967 environment. (C) RIP amino-acid length versus plasmid length across clades.
968 Scatter plots show RIP length versus plasmid length, and boxplots summarize clade-
969 wise distributions. (D) RIP GC content versus RIP–plasmid GC difference across
970 clades. Scatter plots show RIP GC content versus the difference between RIP and
971 plasmid GC content, and boxplots summarize clade-wise distributions. (E) RIP
972 isoelectric point (pI) versus RIP–plasmid proteome pI difference across clades.
973 Scatter plots show RIP pI versus the difference between RIP and plasmid proteome
974 pI, and boxplots summarize clade-wise distributions. (F) Distribution of MOB classes
975 across clades. Upper panel shows the proportions within each clade; lower panel
976 shows clade composition within each MOB class. (G) Distribution of MPF classes
977 across clades. Upper panel shows the proportions within each clade; lower panel
978 shows clade composition within each MPF class. (H) Distribution of antimicrobial
979 resistance (AMR) genes across clades, shown as the proportions of plasmids
980 carrying different numbers of ARGs per plasmid. (I) Distribution of defense systems
981 across clades, shown as the proportions of plasmids carrying different numbers of
982 defense systems per plasmid. (J) Distribution of metal resistance genes across
983 clades. Upper panel shows the proportions of plasmids carrying different numbers of
984 metal resistance determinants per plasmid; lower panels show the proportions of
985 individual metal resistance classes within each clade and clade composition within
986 each metal resistance class. Sample sizes are indicated next to the bars.

987 **Figure 5. Integrated overview of the eight WH RIP clades.**

988 The tree is partitioned into the sWH and dWH superclades, corresponding to single-
989 and double-winged-helix replication initiation proteins, with branch support values

shown at the major internal nodes. For each clade, the figure summarizes domain arrangement, representative Pfam domains, characteristic plasmid and RIP features, compositional properties, predominant host taxonomic distribution, environmental association, and major accessory functional traits. Feature labels were assigned as an integrated summary of the dominant structural, compositional, taxonomic, ecological, and functional patterns observed across Figures 2–4; terms such as “enriched,” “poor,” and “associated” summarize the predominant tendencies within each clade rather than fixed categorical thresholds. Branch support values were derived from the WH RIP phylogeny shown in Figure 2.

Contributions

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

1007 **References**

1008 1. Wiedenbeck, J., and Cohan, F.M. (2011). Origins of bacterial diversity through
1009 horizontal genetic transfer and adaptation to new ecological niches. FEMS
1010 Microbiol. Rev. *35*, 957–976.

1011 2. Haudiquet, M., de Sousa, J.M., Touchon, M., and Rocha, E.P.C. (2022). Selfish,
1012 promiscuous and sometimes useful: how mobile genetic elements drive
1013 horizontal gene transfer in microbial populations. Philos. Trans. R. Soc. Lond. B
1014 Biol. Sci. *377*, 20210234.

1015 3. Helinski, D.R. (2022). A Brief History of Plasmids. EcoSal Plus *10*,
1016 eESP00282021.

1017 4. Starling, S. (2025). Plasmid evolution in the time of antibiotics. Nat. Rev.
1018 Microbiol. https://doi.org/10.1038/s41579-025-01273-9.

1019 5. Cazares, A., Figueroa, W., Cazares, D., Lima, L., Turnbull, J.D., McGregor, H.,
1020 Dicks, J., Alexander, S., Iqbal, Z., and Thomson, N.R. (2025). Pre- and
1021 postantibiotic epoch: The historical spread of antimicrobial resistance. Science
1022 *390*, eadr1522.

1023 6. Camargo, A.P., Call, L., Roux, S., Nayfach, S., Huntemann, M., Palaniappan,
1024 K., Ratner, A., Chu, K., Mukherjeep, S., Reddy, T.B.K., et al. (2024). IMG/PR: a
1025 database of plasmids from genomes and metagenomes with rich annotations
1026 and metadata. Nucleic Acids Res. *52*, D164–D173.

1027 7. Schmartz, G.P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Müller, R., and
1028 Keller, A. (2022). PLSDB: advancing a comprehensive database of bacterial
1029 plasmids. Nucleic Acids Res. *50*, D273–D278.

1030 8. Garcillán-Barcia, M.P., Redondo-Salvo, S., and de la Cruz, F. (2023). Plasmid
1031 classifications. Plasmid *126*, 102684.

1032 9. Carattoli, A., Zankari, E., Garcia-Fernandez, A., Voldby Larsen, M., Lund, O.,
1033 Villa, L., Moller Aarestrup, F., and Hasman, H. (2014). In silico detection and
1034 typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing.
1035 Antimicrob. Agents Chemother. *58*, 3895–3903.

1036 10. Fernandez-Lopez, R., Redondo, S., Garcillan-Barcia, M.P., and de la Cruz, F.
1037 (2017). Towards a taxonomy of conjugative plasmids. Curr. Opin. Microbiol. *38*,
1038 106–113.

1039 11. Redondo-Salvo, S., Fernández-López, R., Ruiz, R., Vielva, L., de Toro, M.,
1040 Rocha, E.P.C., Garcillán-Barcia, M.P., and de la Cruz, F. (2020). Pathways for
1041 horizontal gene transfer in bacteria revealed by a global map of their plasmids.
1042 Nat. Commun. *11*, 3602.

1043 12. Garcillán-Barcia, M.P., Redondo-Salvo, S., Vielva, L., and de la Cruz, F. (2020).
1044 MOBscan: Automated annotation of MOB relaxases. Methods Mol. Biol. *2075*,
1045 295–308.

1046 13. Coluzzi, C., Garcillán-Barcia, M.P., de la Cruz, F., and Rocha, E.P.C. (2022).
1047 Evolution of Plasmid mobility: Origin and fate of conjugative and nonconjugative
1048 plasmids. Mol. Biol. Evol. *39*. https://doi.org/10.1093/molbev/msac115.

1049 14. Garcillán-Barcia, M.P., Francia, M.V., and de la Cruz, F. (2009). The diversity of
1050 conjugative relaxases and its application in plasmid classification. FEMS
1051 Microbiol Rev *33*, 657–687.

1052 15. Koonin, E.V., Kuhn, J.H., Dolja, V.V., and Krupovic, M. (2024). Megataxonomy
1053 and global ecology of the virosphere. ISME J. *18*.
1054 https://doi.org/10.1093/ismejo/wrad042.

1055 16. International Committee on Taxonomy of Viruses Executive Committee (2020).
1056 The new scope of virus taxonomy: partitioning the virosphere into 15
1057 hierarchical ranks. Nat Microbiol *5*, 668–674.

1058 17. Moradali, M.F., Ghods, S., and Rehm, B.H.A. (2017). *Pseudomonas aeruginosa*
1059 lifestyle: A paradigm for adaptation, survival, and persistence. Front. Cell. Infect.
1060 Microbiol. *7*, 39.

1061 18. Thomas, C.M., and Haines, A.S. (2004). Plasmids of the genus Pseudomonas.
1062 In *Pseudomonas* (Springer US), pp. 197–231.

1063 19. Shintani, M., and Suzuki, H. (2019). Plasmids and their hosts. In: Nishida, H.,
1064 Oshima, T. (eds) DNA Traffic in the Environment. Springer, Singapore.
1065 https://doi.org/10.1007/978-981-13-3411-5_6

1066 20. Sawada, Y., Yaginuma, S., Tai, M., Iyobe, S., and Mitsuhashi, S. (1976).
1067 Plasmid-mediated penicillin beta-lactamases in *Pseudomonas aeruginosa*.
1068 Antimicrob. Agents Chemother. *9*, 55–60.

1069 21. Sagai, H., Hasuda, K., Iyobe, S., Bryan, L.E., Holloway, B.W., and Mitsuhashi,
1070 S. (1976). Classification of R Plasmids by Incompatibility in *Pseudomonas*
1071 *aeruginosa*. Antimicrob. Agents Chemother. *10*, 573–578.

1072 22. Summers, A.O., and Jacoby, G.A. (1978). Plasmid-determined resistance to
1073 boron and chromium compounds in *Pseudomonas aeruginosa*. Antimicrob.
1074 Agents Chemother. *13*, 637–640.

1075 23. Kawakami, Y., Mikoshiba, F., Nagasaki, S., Matsumoto, H., and Tazaki, T.
1076 (1972). Prevalence of *Pseudomonas aeruginosa* strains possessing R factor in
1077 a hospital. J. Antibiot. *25*, 607–609.

1078 24. Hedges, R.W., and Jacob, A.E. (1975). Mobilization of plasmid-borne drug
1079 resistance determinants for transfer from *Pseudomonas aeruginosa* to
1080 Escherichia coli. Mol. Gen. Genet. *140*, 69–79.

1081 25. Ingram, L., Sykes, R.B., Grinsted, J., Saunders, J.R., and Richmond, M.H.
1082 (1972). A transmissible resistance element from a strain of *Pseudomonas*
1083 *aeruginosa* containing no detectable extrachromosomal DNA. J. Gen. Microbiol.
1084 *72*, 269–279.

1085 26. Bryan, L.E., Semaka, S.D., Van den Elzen, H.M., Kinnear, J.E., and
1086 Whitehouse, R.L. (1973). Characteristics of R931 and other *Pseudomonas*
1087 *aeruginosa* R factors. Antimicrob. Agents Chemother. *3*, 625–637.

1088 27. Jacoby, G.A. (1980). Plasmid determined resistance to carbenicillin and
1089 gentamicin in *Pseudomonas aeruginosa*. In: Stuttard, C., Rozee, K, R, (eds.)
1090 Plasmids and Transposons, Academic Press, https://doi.org/10.1016/b978-0-12-
1091 675550-3.50011-3

1092 28. Bradley, D.E. (1983). Specification of the conjugative pili and surface mating
1093 systems of *Pseudomonas* plasmids. J. Gen. Microbiol. *129*, 2545–2556.

1094 29. Cain, D., and Holloway, B.W. (1984). Prime plasmids derived from the IncP-10
1095 plasmid R91-5 in *Pseudomonas putida*. FEMS Microbiol. Lett. *24*, 97–101.

1096 30. Davies, S., and Krishnapillai, V. (1990). DNA sequence analysis of the
1097 replication region of the *Pseudomonas aeruginosa* plasmid R91-5. J. Genet. *69*,
1098 101–112.

1099 31. Jacoby, G.A. (1986). Resistance Plasmids of *Pseudomonas*. In: Sokatch, J.R. (Ed.)
1100 The Biology of *Pseudomonas*, Academic Press, https://doi.org/10.1016/b978-0-12-
1101 307210-8.50013-0.

1102 32. Suzuki, M., Suzuki, H., Nishimura, Y., Nojiri, H., and Shintani, M. (2025).
1103 Enhancing plasmid typing with MOB-typer: resolving IncP and other
1104 incompatibility group misclassifications in *Pseudomonas*. Microb. Genom. *11*.
1105 https://doi.org/10.1099/mgen.0.001491.

1106 33. Kawalek, A., Kotecka, K., Modrzejewska, M., Gawor, J., Jagura-Burdzy, G., and
1107 Bartosik, A.A. (2020). Genome sequence of *Pseudomonas aeruginosa*
1108 PAO1161, a PAO1 derivative with the ICEPae1161 integrative and conjugative
1109 element. BMC Genomics *21*, 14.

1110 34. Boronin, A.M. (1992). Diversity of *Pseudomonas* plasmids: To what extent?
1111 FEMS Microbiol. Lett. *100*, 461–467.

1112 35. Van der Auwera, G.A., Krol, J.E., Suzuki, H., Foster, B., Van Houdt, R., Brown,
1113 C.J., Mergeay, M., and Top, E.M. (2009). Plasmids captured in *C. metallidurans*
1114 CH34: defining the PromA family of broad-host-range plasmids. Antonie Van
1115 Leeuwenhoek *96*, 193–204.

1116 36. Law, A., Solano, O., Brown, C.J., Hunter, S.S., Fagnan, M., Top, E.M., and
1117 Stalder, T. (2021). Biosolids as a source of antibiotic resistance plasmids for
1118 commensal and pathogenic bacteria. Front. Microbiol. *12*, 606409.

1119 37. Yanagiya, K., Maejima, Y., Nakata, H., Tokuda, M., Moriuchi, R., Dohra, H.,
1120 Inoue, K., Ohkuma, M., Kimbara, K., and Shintani, M. (2018). Novel self-
1121 transmissible and broad-host-range plasmids exogenously captured from
1122 anaerobic granules or cow manure. Front. Microbiol. *9*, 2602.

1123 38. Tett, A., Spiers, A.J., Crossman, L.C., Ager, D., Ciric, L., Dow, J.M., Fry, J.C.,
1124 Harris, D., Lilley, A., Oliver, A., et al. (2007). Sequence-based analysis of

1125 pQBR103; a representative of a unique, transfer-proficient mega plasmid
1126 resident in the microbial community of sugar beet. ISME J. *1*, 331–340.

1127 39. Köhler, K.A.K., Rückert, C., Schatschneider, S., Vorhölter, F.-J.,
1128 Szczepanowski, R., Blank, L.M., Niehaus, K., Goesmann, A., Pühler, A.,
1129 Kalinowski, J., et al. (2013). Complete genome sequence of *Pseudomonas* sp.
1130 strain VLB120 a solvent tolerant, styrene degrading bacterium, isolated from
1131 forest soil. J. Biotechnol. *168*, 729–730.

1132 40. Lawrence, J.G., and Ochman, H. (1997). Amelioration of bacterial genomes:
1133 rates of change and exchange. J. Mol. Evol. *44*, 383–397.

1134 41. Rocha, E.P., and Danchin, A. (2002). Base composition bias might result from
1135 competition for metabolic resources. Trends Genet. *18*, 291–294.

1136 42. Dietel, A.-K., Merker, H., Kaltenpoth, M., and Kost, C. (2019). Selective
1137 advantages favour high genomic AT-contents in intracellular elements. PLoS
1138 Genet. *15*, e1007778.

1139 43. Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-
1140 fast iterative protein sequence searching by HMM-HMM alignment. Nat.
1141 Methods *9*, 173–175.

1142 44. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A.,
1143 Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024).
1144 Accurate structure prediction of biomolecular interactions with AlphaFold 3.
1145 Nature *630*, 493–500.

1146 45. Wegrzyn, K., Zabrocka, E., Bury, K., Tomiczek, B., Wieczor, M., Czub, J.,
1147 Uciechowska, U., Moreno-Del Alamo, M., Walkow, U., Grochowina, I., et al.
1148 (2021). Defining a novel domain that provides an essential contribution to site-
1149 specific interaction of Rep protein with DNA. Nucleic Acids Res. *49*, 3394–3408.

1150 46. Schumacher, M.A., Tonthat, N.K., Kwong, S.M., Chinnam, N.B., Liu, M.A.,
1151 Skurray, R.A., and Firth, N. (2014). Mechanism of staphylococcal
1152 multiresistance plasmid replication origin assembly by the RepA protein. Proc.
1153 Natl. Acad. Sci. U. S. A. *111*, 9121–9126.

1154 47. Itou, H., Yagura, M., Shirakihara, Y., and Itoh, T. (2015). Structural basis for
1155 replication origin unwinding by an initiator primase of plasmid ColE2-P9: duplex
1156 DNA unwinding by a single protein. J. Biol. Chem. *290*, 3601–3611.

1157 48. Iyer, L.M., Koonin, E.V., Leipe, D.D., and Aravind, L. (2005). Origin and
1158 evolution of the archaeo-eukaryotic primase superfamily and related palm-
1159 domain proteins: structural insights and new members. Nucleic Acids Res. *33*,
1160 3875–3896.

1161 49. Haines, A.S., Jones, K., Cheung, M., and Thomas, C.M. (2005). The IncP-6
1162 plasmid Rms149 consists of a small mobilizable backbone with multiple large
1163 insertions. J. Bacteriol. *187*, 4728–4738.

1164 50. Haines, A.S., Akhtar, P., Stephens, E.R., Jones, K., Thomas, C.M., Perkins,

1165 C.D., Williams, J.R., Day, M.J., and Fry, J.C. (2006). Plasmids from freshwater
1166 environments capable of IncQ retrotransfer are diverse and include pQKH54, a
1167 new IncP-1 subgroup archetype. Microbiology *152*, 2689–2701.

1168 51. Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005).
1169 The many faces of the helix-turn-helix domain: transcription regulation and
1170 beyond. FEMS Microbiol Rev *29*, 231–262.

1171 52. Giraldo, R., and Fernandez-Tresguerres, M.E. (2004). Twenty years of the
1172 pPS10 replicon: insights on the molecular mechanism for the activation of DNA
1173 replication in iteron-containing bacterial plasmids. Plasmid *52*, 69–83.

1174 53. Giraldo, R. (2003). Common domains in the initiators of DNA replication
1175 in*Bacteria, Archaea*and*Eukarya*: combined structural, functional and
1176 phylogenetic perspectives. FEMS Microbiol. Rev. *26*, 533–554.

1177 54. Tria, F.D.K., Landan, G., and Dagan, T. (2017). Phylogenetic rooting using
1178 minimal ancestor deviation. Nat. Ecol. Evol. *1*, 193.

1179 55. Lam, M.M.C., and Hamidian, M. (2024). Examining the role of *Acinetobacter*
1180 *baumannii* plasmid types in disseminating antimicrobial resistance. NPJ
1181 Antimicrob Resist *2*, 1.

1182 56. Komori, H., Matsunaga, F., Higuchi, Y., Ishiai, M., Wada, C., and Miki, K. (1999).
1183 Crystal structure of a prokaryotic replication initiator protein bound to DNA at 2.6
1184 A resolution. EMBO J. *18*, 4597–4607.

1185 57. Sambrook, J., and Russell, D. (2001). Molecular Cloning: A Laboratory Manual
1186 3rd edn. (Cold Spring Harbor Laboratory).

1187 58. Royle, P.L., Matsumoto, H., and Holloway, B.W. (1981). Genetic circularity of
1188 the *Pseudomonas aeruginosa* PAO chromosome. J. Bacteriol. *145*, 145–155.

1189 59. plasmidSPAdes – Center for Algorithmic Biotechnology
1190 https://cab.spbu.ru/software/plasmid-spades/.

1191 60. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S.,
1192 Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic:
1193 an integrated and extendable desktop software platform for the organization and
1194 analysis of sequence data. Bioinformatics *28*, 1647–1649.

1195 61. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
1196 Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup
1197 (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*,
1198 2078–2079.

1199 62. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and
1200 Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962.

1201 63. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S.,
1202 Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a
1203 new genome assembly algorithm and its applications to single-cell sequencing.

1204 J. Comput. Biol. *19*, 455–477.

1205 64. Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018). DFAST: a flexible
1206 prokaryotic genome annotation pipeline for faster genome publication.
1207 Bioinformatics *34*, 1037–1039.

1208 65. Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011). Easyfig: a genome
1209 comparison visualizer. Bioinformatics *27*, 1009–1010.

1210 66. Lobry, J.R. (1996). Origin of Replication of *Mycoplasma genitalium*. Science.
1211 *272*, 745–746.

1212 67. Necşulea, A., and Lobry, J.R. (2007). A new method for assessing the effect of
1213 replication on DNA base composition asymmetry. Mol. Biol. Evol. *24*, 2169–
1214 2179.

1215 68. Solovyev, V., and Salamov, A. (2011) Automatic annotation of microbial
1216 genomes and metagenomic sequences. In: Li, R.W. (Ed.) Metagenomics and its
1217 applications in agriculture, biomedicine and environmental studies, Nova
1218 Science Publishers, p.61-78.

1219 69. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist,
1220 C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein
1221 structure search with Foldseek. Nat. Biotechnol. *42*, 243–246.

1222 70. Swan, M.K., Bastia, D., and Davies, C. (2006). Crystal structure of π initiator
1223 protein–iteron complex of plasmid R6K: Implications for initiation of plasmid
1224 DNA replication. Proc. Nat. Acad. Sci. U. S. A. *103*, 18481–18486.

1225 71. Galata, V., Fehlmann, T., Backes, C., and Keller, A. (2018). PLSDB: a resource
1226 of complete bacterial plasmids. Nucleic Acids Res. *47*, D195-D202.

1227 72. Robertson, J., and Nash, J.H.E. (2018). MOB-suite: software tools for clustering,
1228 reconstruction and typing of plasmids from draft assemblies. Microb Genom. *27*,
1229 e000206.

1230 73. Carattoli, A., and Hasman, H. (2020). PlasmidFinder and in silico pMLST:
1231 Identification and typing of Plasmid replicons in whole-genome sequencing
1232 (WGS). Methods Mol. Biol. *2075*, 285–294.

1233 74. Bayliss, S.C., Thorpe, H.A., Coyle, N.M., Sheppard, S.K., and Feil, E.J. (2019).
1234 PIRATE: A fast and scalable pangenomics toolbox for clustering diverged
1235 orthologues in bacteria. Gigascience *8*, giz119.

1236 75. Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at
1237 tree-of-life scale using DIAMOND. Nat. Methods *18*, 366–368.

1238 76. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and
1239 Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of
1240 protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

1241 77. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein

1242 sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*,
1243 1026–1028.

1244 78. Woodcroft, B.J., Boyd, J.A., and Tyson, G.W. (2016). OrfM: a fast open reading
1245 frame predictor for metagenomic data. Bioinformatics *32*, 2702–2703.

1246 79. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*,
1247 e1002195.

1248 80. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and
1249 Söding, J. (2019). HH-suite3 for fast remote homology detection and deep
1250 protein annotation. BMC Bioinformatics *20*, 473.

1251 81. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment
1252 software version 7: improvements in performance and usability. Mol. Biol. Evol.
1253 *30*, 772–780.

1254 82. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von
1255 Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient
1256 methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. *37*,
1257 1530–1534.

1258 83. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin,
1259 L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic
1260 estimates. Nat. Methods *14*, 587–589.

1261 84. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S.
1262 (2018). UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol.
1263 Evol. *35*, 518–522.

1264 85. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and
1265 Gascuel, O. (2010). New algorithms and methods to estimate maximum-
1266 likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*,
1267 307–321.

1268 86. Mai, U., and Mirarab, S. (2018). TreeShrink: fast and accurate detection of
1269 outlier long branches in collections of phylogenetic trees. BMC Genomics *19*,
1270 272.

1271 87. Lemoine, F., and Gascuel, O. (2021). Gotree/Goalign: toolkit and Go API to
1272 facilitate the development of phylogenetic workflows. NAR Genom Bioinform *3*,
1273 lqab075.

1274 88. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser,
1275 L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site
1276 identification. BMC Bioinformatics *11*, 119.

1277 89. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent
1278 updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res
1279 *52*, W78–W82.

1280 90. Néron, B., Denise, R., Coluzzi, C., Touchon, M., Rocha, E.P.C., and Abby, S.S.

1281 (2023). MacSyFinder v2: Improved modelling and search engine to identify
1282 molecular systems in genomes. Peer Community J. *3* e28.

1283 91. Cury, J., Abby, S.S., Doppelt-Azeroual, O., Néron, B., and Rocha, E.P.C.
1284 (2020). Identifying conjugative plasmids and integrative conjugative elements
1285 with CONJscan. Methods Mol Biol *2075*, 265–283.

1286 92. Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., and
1287 Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of
1288 prokaryotes. Nat Commun *13*, 2561.

1289 93. Pansegrau, W., Lanka, E., Barth, P.T., Figurski, D.H., Guiney, D.G., Haas, D.,
1290 Helinski, D.R., Schwab, H., Stanisich, V.A., and Thomas, C.M. (1994). Complete
1291 nucleotide sequence of Birmingham IncP alpha plasmids. Compilation and
1292 comparative analysis. J. Mol. Biol. *239*, 623–663.

1293 94. Shintani, M., Suzuki, H., Nojiri, H., and Suzuki, M. (2022). Precise classification
1294 of antimicrobial resistance-associated IncP-2 megaplasmids for molecular
1295 epidemiological studies on *Pseudomonas species*. J. Antimicrob. Chemother.
1296 *77*, 1203–1205.

1297 95. Llanes, C., Gabant, P., Couturier, M., Bayer, L., and Plesiat, P. (1996).
1298 Molecular analysis of the replication elements of the broad-host-range RepA/C
1299 replicon. Plasmid *36*, 26–35.

1300 96. Fricke, W.F., Welch, T.J., McDermott, P.F., Mammel, M.K., LeClerc, J.E., White,
1301 D.G., Cebula, T.A., and Ravel, J. (2009). Comparative genomics of the IncA/C
1302 multidrug resistance plasmid family. J. Bacteriol. *191*, 4750–4757.

1303 97. Scholz, P., Haring, V., Wittmann-Liebold, B., Ashman, K., Bagdasarian, M., and
1304 Scherzinger, E. (1989). Complete nucleotide sequence and gene organization of
1305 the broad-host-range plasmid RSF1010. Gene *75*, 271–288.

1306 98. Maeda, K., Nojiri, H., Shintani, M., Yoshida, T., Habe, H., and Omori, T. (2003).
1307 Complete nucleotide sequence of carbazole/dioxin-degrading plasmid pCAR1 in
1308 *Pseudomonas resinovorans* strain CA10 indicates its mosaicity and the
1309 presence of large catabolic transposon Tn*4676*. J. Mol. Biol. *326*, 21–33.

1310 99. Shintani, M., Yano, H., Habe, H., Omori, T., Yamane, H., Tsuda, M., and Nojiri,
1311 H. (2006). Characterization of the replication, maintenance, and transfer
1312 features of the IncP-7 plasmid pCAR1, which carries genes involved in
1313 carbazole and dioxin degradation. Appl. Environ. Microbiol. *72*, 3206–3216.

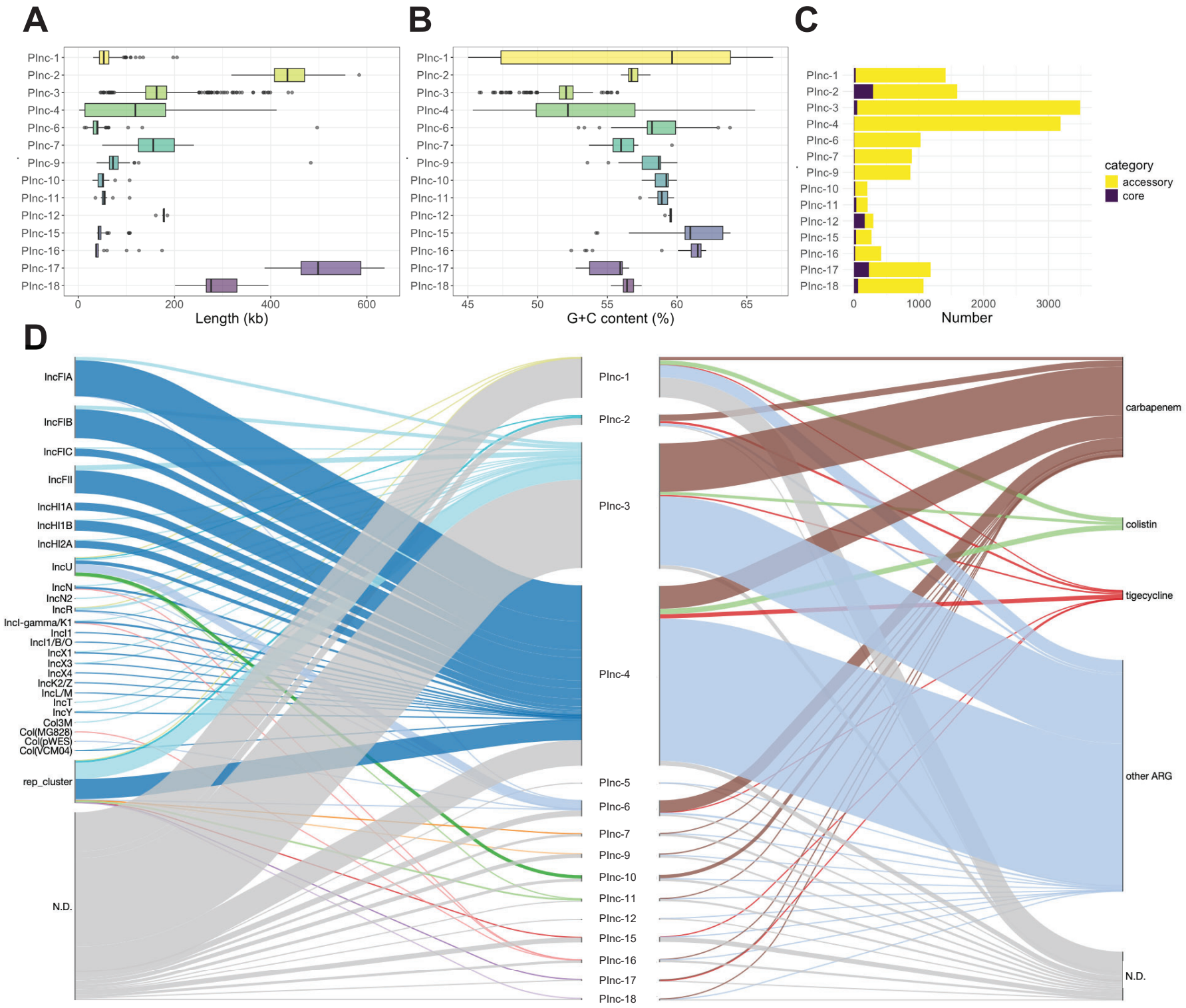
1314 100. Takahashi, Y., Shintani, M., Yamane, H., and Nojiri, H. (2009). The complete
1315 nucleotide sequence of pCAR2: pCAR2 and pCAR1 were structurally identical
1316 IncP-7 carbazole degradative plasmids. Biosci. Biotechnol. Biochem. *73*, 744–
1317 746.

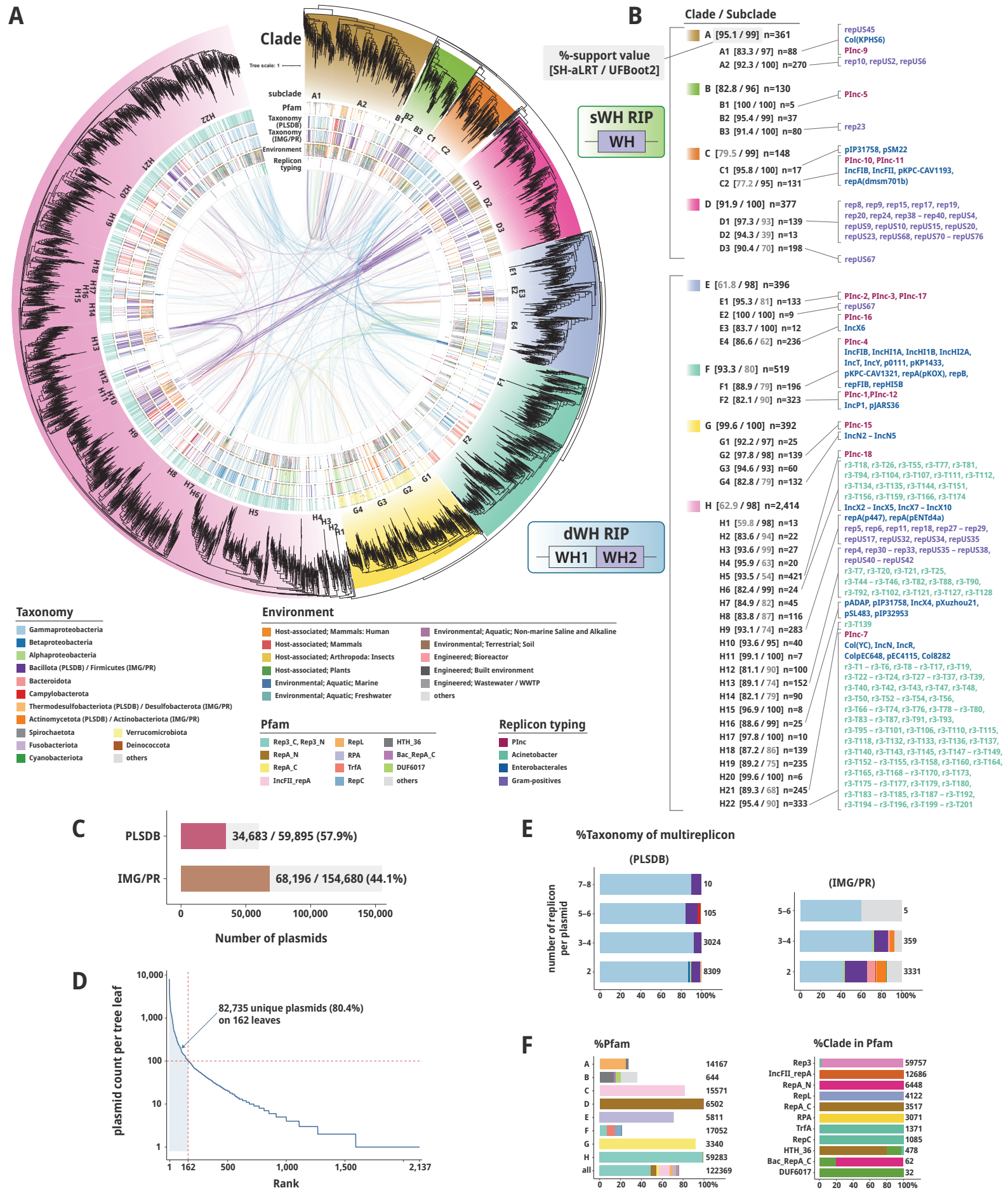
1318 101. Greated, A., Lambertsen, L., Williams, P.A., and Thomas, C.M. (2002).
1319 Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas*
1320 *putida*. Environ. Microbiol. *4*, 856-871.

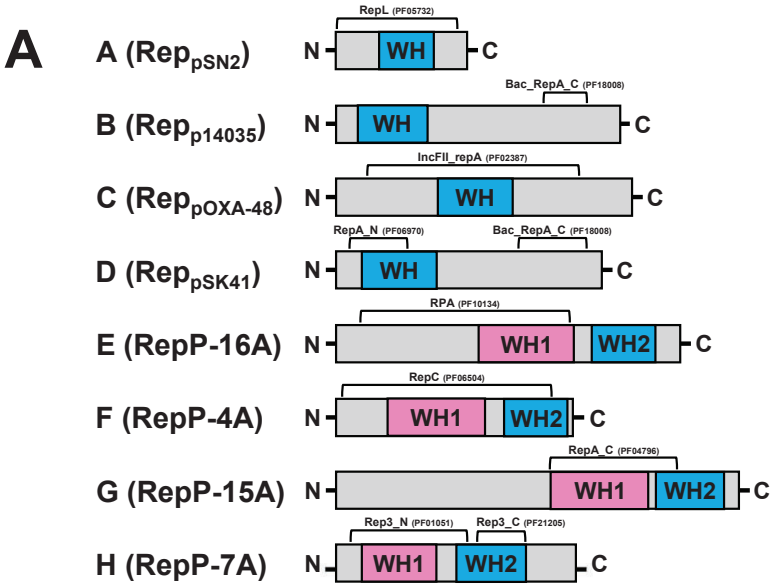
1321 102. Sota, M., Yano, H., Ono, A., Miyazaki, R., Ishii, H., Genka, H., Top, E.M., and
1322 Tsuda, M. (2006). Genomic and functional analysis of the IncP-9 naphthalene-
1323 catabolic plasmid NAH7 and its transposon Tn*4655* suggests catabolic gene
1324 spread by a tyrosine recombinase. J. Bacteriol. *188*, 4057–4067.

1325 103. Hayakawa, M., Tokuda, M., Kaneko, K., Nakamichi, K., Yamamoto, Y., Kamijo,
1326 T., Umeki, H., Chiba, R., Yamada, R., Mori, M., et al. (2022). Hitherto-unnoticed
1327 self-transmissible plasmids widely distributed among different environments in
1328 Japan. Appl. Environ. Microbiol., *88*, e0111422.

1329 104. Douarre, P.-E., Mallet, L., Radomski, N., Felten, A., and Mistou, M.-Y. (2020).
1330 Analysis of COMPASS, a new comprehensive plasmid database revealed
1331 prevalence of multireplicon and extensive diversity of IncF plasmids. Front.
1332 Microbiol. *11*, 483.

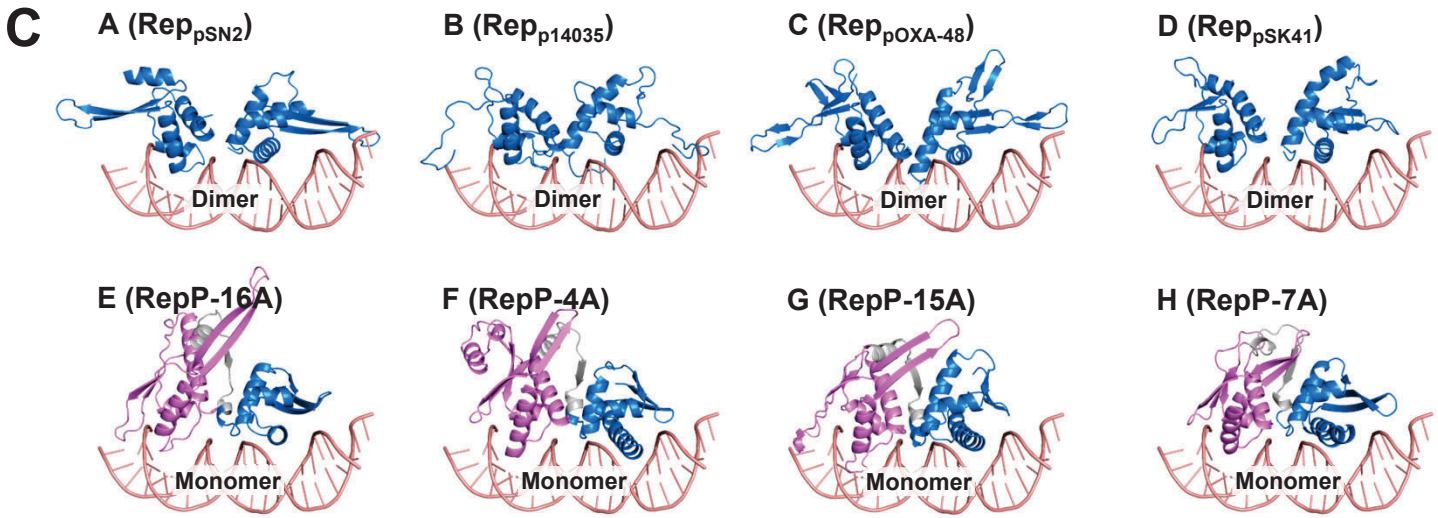

A
PInc-1
PInc-2
PInc-3
PInc-4
PInc-6
PInc-7
PInc-9
PInc-10
PInc-11
PInc-12
PInc-15
PInc-16
PInc-17
PInc-18
0
200
400
600
Length (kb)
B
45
50
55
60
65
G+C content (%)
C
0
1000
2000
3000
Number
category
accessory
core
D
IncFIA
IncFIB
IncFIC
IncFII
IncHI1A
IncHI1B
IncHI2A
IncU
IncN
IncN2
IncR
IncI-gamma/K1
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.

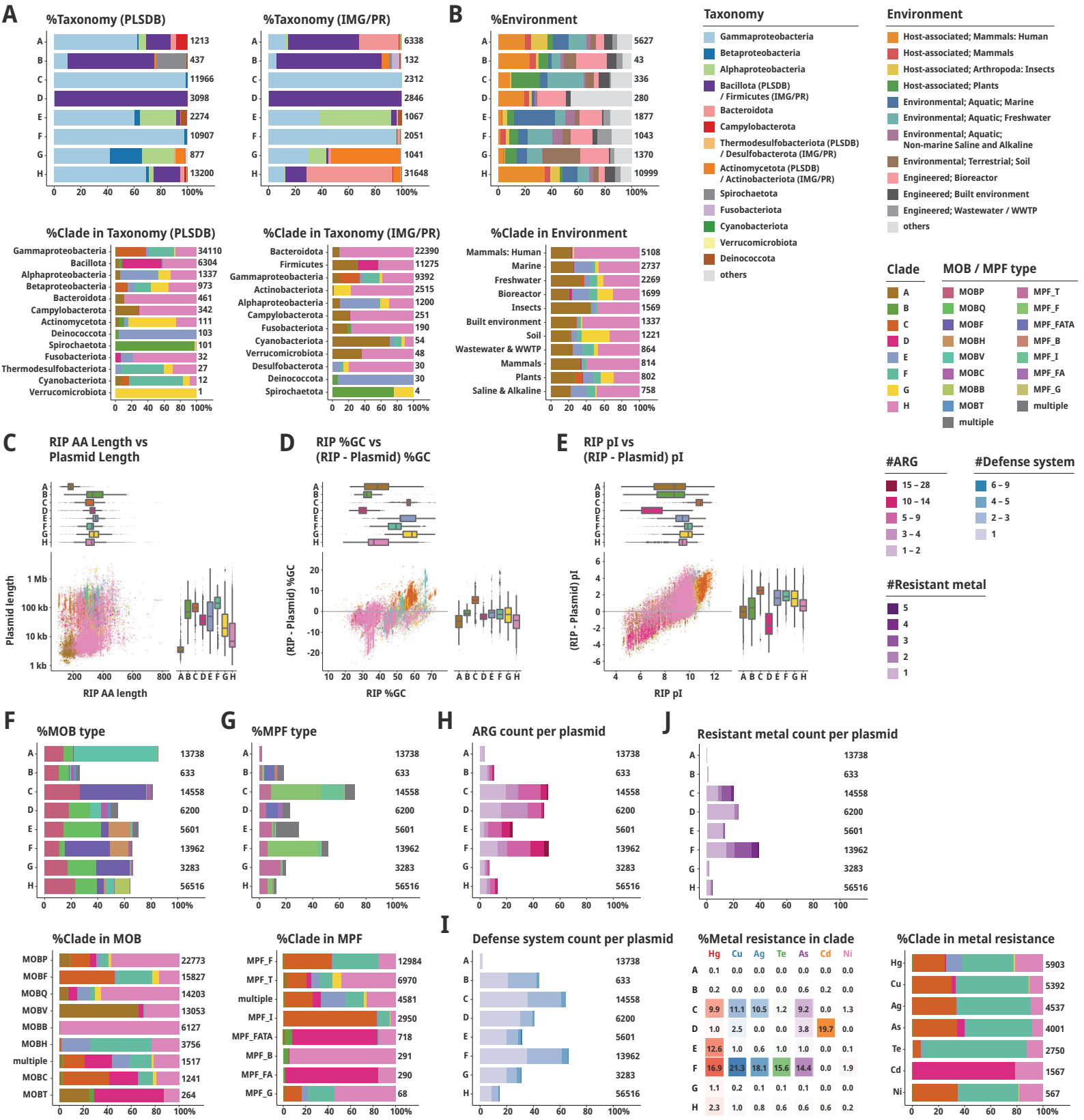

A
Clade
Tree scale: 1
subclade
Pfam
Taxonomy (PLSDB)
Taxonomy (IMG/PR)
Environment
Replicon typing
A1
A2
B1
B2
B3
C1
C2
D1
D2
D3
E1
E2
E3
E4
F1
F2
G1
G2
G3
G4
H1
H2
H3
H4
H5
H6
H7
H8
H9
H10
H11
H12
H13
H14
H15
H16
H17
H18
H19
H20
H21
H22
%-support value
[SH-aLRT / UFBoot2]
sWH RIP
WH
dWH RIP
WH1
WH2
B
Clade / Subclade
A [95.1 / 99] n=361
A1 [83.3 / 97] n=88
A2 [92.3 / 100] n=270
repUS45
Col(KPHS6)
PInc-9
rep10, repUS2, repUS6
B [82.8 / 96] n=130
B1 [100 / 100] n=5
B2 [95.4 / 99] n=37
B3 [91.4 / 100] n=80
PInc-5
rep23
C [79.5 / 99] n=148
C1 [95.8 / 100] n=17
C2 [77.2 / 95] n=131
pIP31758, pSM22
PInc-10, PInc-11
IncFIB, IncFII, pKPC-CAV1193,
repA(dmsm701b)
D [91.9 / 100] n=377
D1 [97.3 / 93] n=139
D2 [94.3 / 39] n=13
D3 [90.4 / 70] n=198
rep8, rep9, rep15, rep17, rep19,
rep20, rep24, rep38 – rep40, repUS4,
repUS9, repUS10, repUS15, repUS20,
repUS23, repUS68, repUS70 – repUS76
repUS67
E [61.8 / 98] n=396
E1 [95.3 / 81] n=133
E2 [100 / 100] n=9
E3 [83.7 / 100] n=12
E4 [86.6 / 62] n=236
PInc-2, PInc-3, PInc-17
repUS67
PInc-16
IncX6
F [93.3 / 80] n=519
F1 [88.9 / 79] n=196
F2 [82.1 / 90] n=323
PInc-4
IncFIB, IncHI1A, IncHI1B, IncHI2A,
IncT, IncY, p0111, pKP1433,
pKPC-CAV1321, repA(pKOX), repB,
repFIB, repHI5B
PInc-1,PInc-12
IncP1, pJARS36
G [99.6 / 100] n=392
G1 [92.2 / 97] n=25
G2 [97.8 / 98] n=139
G3 [94.6 / 93] n=60
G4 [82.8 / 79] n=132
PInc-15
IncN2 – IncN5
H [62.9 / 98] n=2,414
H1 [59.8 / 98] n=13
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
H8 [83.8 / 87] n=116
H9 [93.1 / 74] n=283
H10 [93.6 / 95] n=40
H11 [99.1 / 100] n=7
H12 [81.1 / 90] n=100
H13 [89.1 / 74] n=152
H14 [82.1 / 79] n=90
H15 [96.9 / 100] n=8
H16 [88.6 / 99] n=25
H17 [97.8 / 100] n=10
H18 [87.2 / 86] n=139
H19 [89.2 / 75] n=235
H20 [99.6 / 100] n=6
H21 [89.3 / 68] n=245
H22 [95.4 / 90] n=333
PInc-18
r3-T18, r3-T26, r3-T55, r3-T77, r3-T81,
r3-T94, r3-T104, r3-T107, r3-T111, r3-T112,
r3-T134, r3-T135, r3-T144, r3-T151,
r3-T156, r3-T159, r3-T166, r3-T174
IncX2 – IncX5, IncX7 – IncX10
repA(p447), repA(pENTd4a)
rep5, rep6, rep11, rep18, rep27 – rep29,
repUS17, repUS32, repUS34, repUS35
rep4, rep30 – rep33, repUS35 – repUS38,
repUS40 – repUS42
r3-T7, r3-T20, r3-T21, r3-T25,
r3-T44 – r3-T46, r3-T82, r3-T88, r3-T90,
r3-T92, r3-T102, r3-T121, r3-T127, r3-T128
pADAP, pIP31758, IncX4, pXuzhou21,
pSL483, pIP32953
r3-T139
PInc-7
Col(YC), IncN, IncR,
ColpEC648, pEC4115, Col8282
r3-T1 – r3-T6, r3-T8 – r3-T17, r3-T19,
r3-T22 – r3-T24, r3-T27 – r3-T37, r3-T39,
r3-T40, r3-T42, r3-T43, r3-T47, r3-T48,
r3-T50, r3-T52 – r3-T54, r3-T56,
r3-T66 – r3-T74, r3-T76, r3-T78 – r3-T80,
r3-T83 – r3-T87, r3-T91, r3-T93,
r3-T95 – r3-T101, r3-T106, r3-T110, r3-T115,
r3-T118, r3-T132, r3-T133, r3-T136, r3-T137,
r3-T140, r3-T143, r3-T145, r3-T147 – r3-T149,
r3-T152 – r3-T155, r3-T158, r3-T160, r3-T164,
r3-T165, r3-T168 – r3-T170, r3-T173,
r3-T175 – r3-T177, r3-T179, r3-T180,
r3-T183 – r3-T185, r3-T187 – r3-T192,
r3-T194 – r3-T196, r3-T199 – r3-T201
Taxonomy
Gammaproteobacteria
Betaproteobacteria
Alphaproteobacteria
Bacillota (PLSDB) / Firmicutes (IMG/PR)
Bacteroidota
Campylobacterota
Thermodesulfobacteriota (PLSDB) / Desulfobacterota (IMG/PR)
Actinomycetota (PLSDB) / Actinobacteriota (IMG/PR)
Spirochaetota
Verrucomicrobiota
Fusobacteriota
Deinococcota
Cyanobacteriota
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
Pfam
Rep3_C, Rep3_N
RepA_N
RepA_C
IncFII_repA
RepL
RPA
TrfA
RepC
HTH_36
Bac_RepA_C
DUF6017
others
Replicon typing
PInc
Acinetobacter
Enterobacterales
Gram-positives
C
PLSDB
34,683 / 59,895 (57.9%)
IMG/PR
68,196 / 154,680 (44.1%)
0
50,000
100,000
150,000
Number of plasmids
D
plasmid count per tree leaf
10,000
1,000
100
10
1
82,735 unique plasmids (80.4%)
on 162 leaves
1
162
500
1,000
1,500
2,137
Rank
E
%Taxonomy of multireplicon
(PLSDB)
(IMG/PR)
number of replicon per plasmid
7–8
5–6
3–4
2
10
105
3024
8309
5
359
3331
0
20
40
60
80
100%
F
%Pfam
A
B
C
D
E
F
G
H
all
14167
644
15571
6502
5811
17052
3340
59283
122369
%Clade in Pfam
Rep3
IncFII_repA
RepA_N
RepL
RepA_C
RPA
TrfA
RepC
HTH_36
Bac_RepA_C
DUF6017
59757
12686
6448
4122
3517
3071
1371
1085
478
62
32

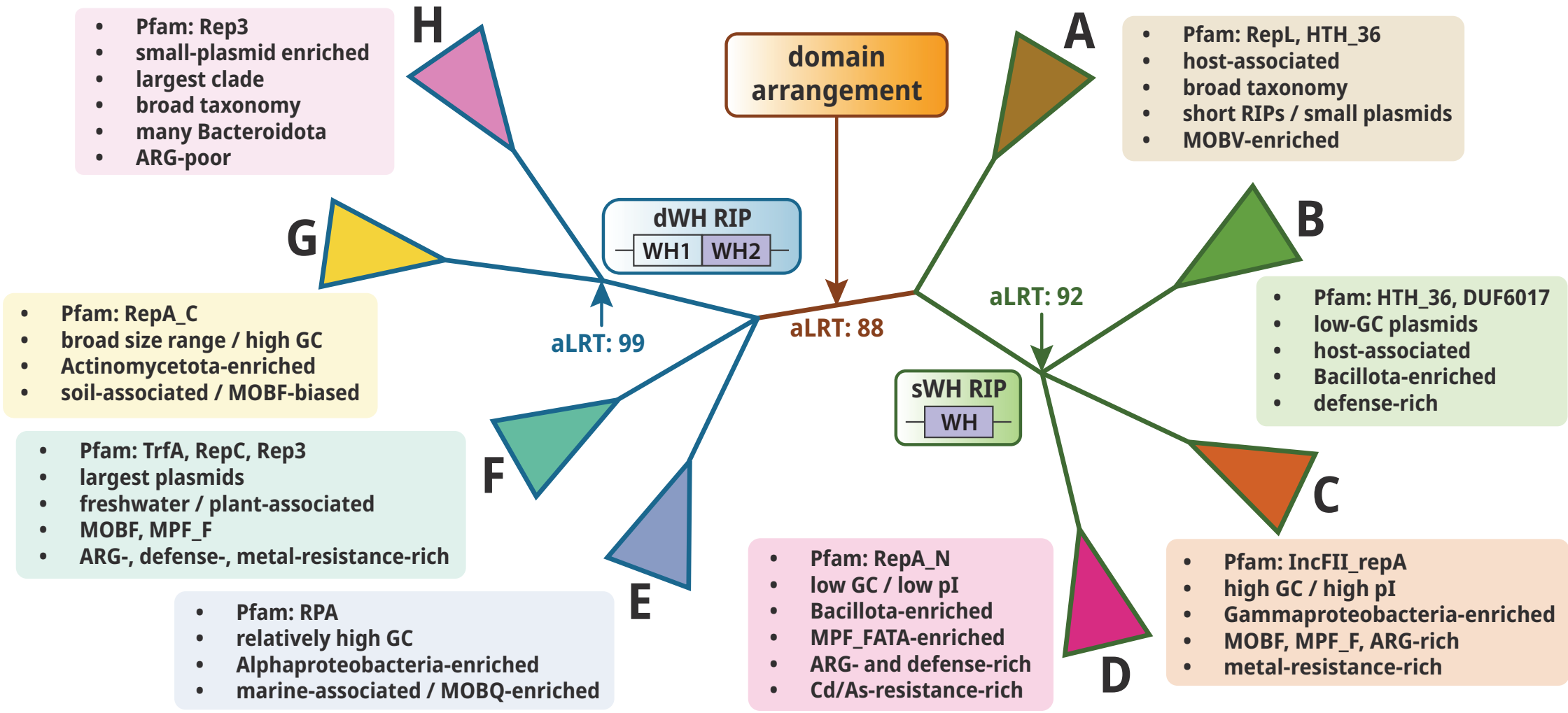


B

	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
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
C
RIP AA Length vs Plasmid Length
Plasmid length
RIP AA length
D
RIP %GC vs (RIP - Plasmid) %GC
(RIP - Plasmid) %GC
RIP %GC
E
RIP pI vs (RIP - Plasmid) pI
(RIP - Plasmid) pI
RIP pI
#ARG
15 – 28
10 – 14
5 – 9
3 – 4
1 – 2
#Defense system
6 – 9
4 – 5
2 – 3
1
#Resistant metal
5
4
3
2
1
F
%MOB type
G
%MPF type
H
ARG count per plasmid
J
Resistant metal count per plasmid
%Clade in MOB
%Clade in MPF
I
Defense system count per plasmid
%Metal resistance in clade
Hg Cu Ag Te As Cd Ni
A 0.1 0.0 0.0 0.0 0.0 0.0 0.0
B 0.2 0.0 0.0 0.0 0.6 0.2 0.0
C 9.9 11.1 10.5 1.2 9.2 0.0 1.3
D 1.0 2.5 0.0 0.0 3.8 19.7 0.0
E 12.6 1.0 0.6 1.0 1.0 0.0 0.1
F 16.9 21.3 18.1 15.6 14.4 0.0 1.9
G 1.1 0.2 0.1 0.1 0.1 0.0 0.0
H 2.3 1.0 0.8 0.6 0.6 0.6 0.2
%Clade in metal resistance


H
Pfam: Rep3
small-plasmid enriched
largest clade
broad taxonomy
many Bacteroidota
ARG-poor
domain arrangement
A
Pfam: RepL, HTH_36
host-associated
broad taxonomy
short RIPs / small plasmids
MOBV-enriched
dWH RIP
WH1
WH2
G
B
Pfam: RepA_C
broad size range / high GC
Actinomycetota-enriched
soil-associated / MOBF-biased
aLRT: 99
aLRT: 88
aLRT: 92
Pfam: HTH_36, DUF6017
low-GC plasmids
host-associated
Bacillota-enriched
defense-rich
sWH RIP
WH
F
C
Pfam: TrfA, RepC, Rep3
largest plasmids
freshwater / plant-associated
MOBF, MPF_F
ARG-, defense-, metal-resistance-rich
E
Pfam: RPA
relatively high GC
Alphaproteobacteria-enriched
marine-associated / MOBQ-enriched
Pfam: RepA_N
low GC / low pI
Bacillota-enriched
MPF_FATA-enriched
ARG- and defense-rich
Cd/As-resistance-rich
D
Pfam: IncFII_repA
high GC / high pI
Gammaproteobacteria-enriched
MOBF, MPF_F, ARG-rich
metal-resistance-rich