The Rare Plasmid Biosphere: A Hidden Reservoir of Genetic Diversity

João S. Rebelo[1,4], Célia P.F. Domingues[1,2,4], Luís Borda-de-Água[3], Teresa Nogueira[1,2], and Francisco Dionisio[1,*]

[1] cE3c—Centre for Ecology, Evolution and Environmental Changes & CHANGE, Global Change and Sustainability Institute, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal
[2] INIAV—Instituto Nacional de Investigação Agrária e Veterinária, 4485-655 Vairão, Portugal
[3] Centro de Ecologia Aplicada Professor Baeta Neves (CEABN), Instituto Superior de Agronomia, Tapada da Ajuda, 1349-017 Lisboa, Portugal
[4] These authors contributed equally: João S. Rebelo, Célia P.F. Domingues
*Correspondence: fadionisio@fc.ul.pt

Abstract: Bacterial communities typically display highly uneven abundance patterns, with a few dominant taxa and many low-abundance ones contributing to extensive genetic diversity [1–4]. Notably, this 'rare biosphere' [5] includes species performing critical ecological functions, such as biogeochemical cycling and resisting invasions [6–9]. While bacterial abundance patterns have been extensively studied, the distribution of plasmids - extrachromosomal, self-replicating genetic elements ubiquitous in bacteria—remains poorly understood. Using a dataset of 52,909 plasmids from isolates or bacteria and archaea [10], we found that their 16,547 Plasmid Taxonomic Units (PTUs) [11,12] exhibit a distribution with a long tail, whether in rank abundance or relative abundance distributions: a few highly prevalent PTUs and many rare. The relative abundance distributions are well described by a Poisson log-normal distribution, coherent with recent findings for species abundance distributions across several taxonomic groups [13]. The host distribution also presents a long tail; however, we show that rare PTUs are not necessarily associated with rare bacterial species nor are common PTUs exclusively found in common hosts. This indicates that PTUs' distribution is not a direct consequence of hosts' distribution. Per plasmid, the host-range of rare PTUs is higher than that of common PTUs, at all levels, from species to phyla. Yet, plasmids from common PTUs are more mobile, likely explaining their success. The large group of rare PTUs constitutes an exceptionally diverse reservoir of genetic material, much more diverse than the group of common PTUs. Under appropriate selective pressures, some of these rare plasmids could spread not only by hitchhiking with their hosts but also through horizontal transfer.

Keywords: Rare biosphere; Plasmid, antimicrobial resistance, virulence factors

Plasmids are pervasive in several bacterial species. For example, *Staphylococcus aureus* strains carry an average of 0.8 distinct plasmids, *Salmonella* 1.7, *Escherichia coli* 2.4 and *Klebsiella* 3.5 [14,15]. Most plasmids can transfer between different bacterial species, and because they contribute to bacterial virulence, drug resistance, heavy metals resistance, and biofilm formation, they have been extensively studied [16–18]. In addition, plasmids play a crucial role in bioremediation by carrying genes that encode enzymes and metabolic pathways capable of degrading environmental pollutants and facilitate the metabolism of xenobiotics—synthetic compounds not naturally present in the biosphere. Through horizontal transfer, plasmids can disseminate these traits across bacterial communities, thereby enhancing the bioremediation potential of microbial populations [19].

Here, we ask how plasmid diversity is distributed according to the abundance of their taxonomic units (see next paragraph). Understanding the distribution of plasmids may be relevant, particularly considering the peculiar distribution of their hosts (mostly bacteria or archaea) within communities. In most ecosystems, bacterial communities are characterized by a few dominant taxa and many low-abundance taxa – the 'rare biosphere', representing most species. An interesting characteristic of these rare biospheres [5] is their potential for dynamic adaptation, where a vast repertoire of individual cells from low-abundance bacterial taxa can proliferate rapidly through clonal replication when environmental conditions shift to favour their growth [4,5].

Answering our question requires a robust plasmid classification, which has been challenging due to the dynamic nature of plasmids, marked by extensive shuffling and recombination [11]. However, like bacterial chromosomes, plasmids can be clustered, enabling their classification into discrete groups known as Plasmid Taxonomic Units (PTUs) [11,12]. These clusters meet the criteria of genetic similarity, where members are more similar to each other than to those outside the cluster, and phylogenetic coherence [12]. Therefore, these PTUs are coherent genomic clusters [20], similar to the concept of operational taxonomic units (OTUs) in metagenomics, or of bacterial species.

Given the close relationship between plasmids and their prokaryotic hosts, should one expect that a few PTUs dominate, while numerous low-abundance taxonomic units account for most taxa, as is generally observed in the case of bacterial communities [1,4,21]? In the bacterial world, rare biospheres represent a vast source of genomic innovation [3,5,21,22]. If a rare plasmid biosphere also exists, it is likely to constitute a source of bacterial

69 innovation and contain 'hidden' accessory genes, such as virulence or antibiotic-
70 resistance genes.
71 To study the distribution of PTUs, we analysed the RefSeq database containing
72 52,909 complete plasmids from bacterial and archaea isolates[10], spanning 16,547 PTUs.
73
74 **Rank-abundance and relative abundance distributions of PTUs**
75 The rank-abundance distribution of the 16,547 PTUs is fat-tailed, revealing a
76 pronounced disparity: a few PTUs are represented by many plasmids and most PTUs
77 (72%) are represented by only a single plasmid (**Fig. 1a**). In a log-log plot, the distribution
78 of plasmid counts per PTU follows a straight line, suggesting a power-law with a slope
79 close to -0.74 (or -0.91 if we exclude PTUs with a single plasmid) (**Fig. 1b**).
80 Notably, the most abundant PTU contains 650 plasmids, whereas 11,897 PTUs
81 are singletons (i.e., represented by just a single plasmid). How should these numbers be
82 interpreted? If a given PTU has a single representative in the dataset, it means that, among
83 the 52,909 plasmids, no other plasmid of the same taxonomic unit is present in the dataset.
84 We do not know, however, whether this plasmid was very frequent or rare in the habitat
85 (e.g, the gut of an animal) where it was found. We can envisage that this plasmid and the
86 corresponding PTU could be common in a particular habitat, yet no other plasmid of the
87 same taxonomic unit was collected in the other thousands of habitats [22].
88 We also analyse relative abundance distributions, commonly referred to in
89 ecological studies as "species abundance distributions" (SADs), and by analogy, we will
90 refer to them as "PTU abundance distributions" (PADs). These distributions represent,
91 on the x-axis, abundance measured by number of plasmids per PTU, typically shown on
92 a base-2 logarithmic scale (octaves: one plasmid, two or three plasmids, four to seven
93 plasmids, eight to 15 plasmids, *et seq.*), and on the y-axis, the number of PTUs with a
94 given abundance (number of PTUs with one plasmid, with two or three plasmids, and so
95 on). There is an extensive body of ecological literature on fitting SADs [23,24]. Following
96 recent work [13], we fit the PAD using three statistical distributions: the log-series, the
97 negative binomial (also called Poisson gamma) and the Poisson log-normal distributions,
98 all via maximum likelihood methods [25]. As can be visually assessed (**Fig. 1c**), the Poisson
99 log-normal provides an excellent fitting, in contrast with the log-series and the negative
100 binomial (see **Supplementary Table S1** with the Akaike information criterion (AIC)
101 values). The log-series and the negative binomial yield nearly indistinguishable fits,
102 which was expected, as the log-series is a limiting case of the negative binomial when its

103 shape parameter approaches zero, which applies to our data (**Supplementary Table S2**).
104 We also plot the PAD in double-logarithmic scales, which further highlights the superior
105 fit of the Poisson log-normal compared to the alternative distributions **(Fig. 1d**).

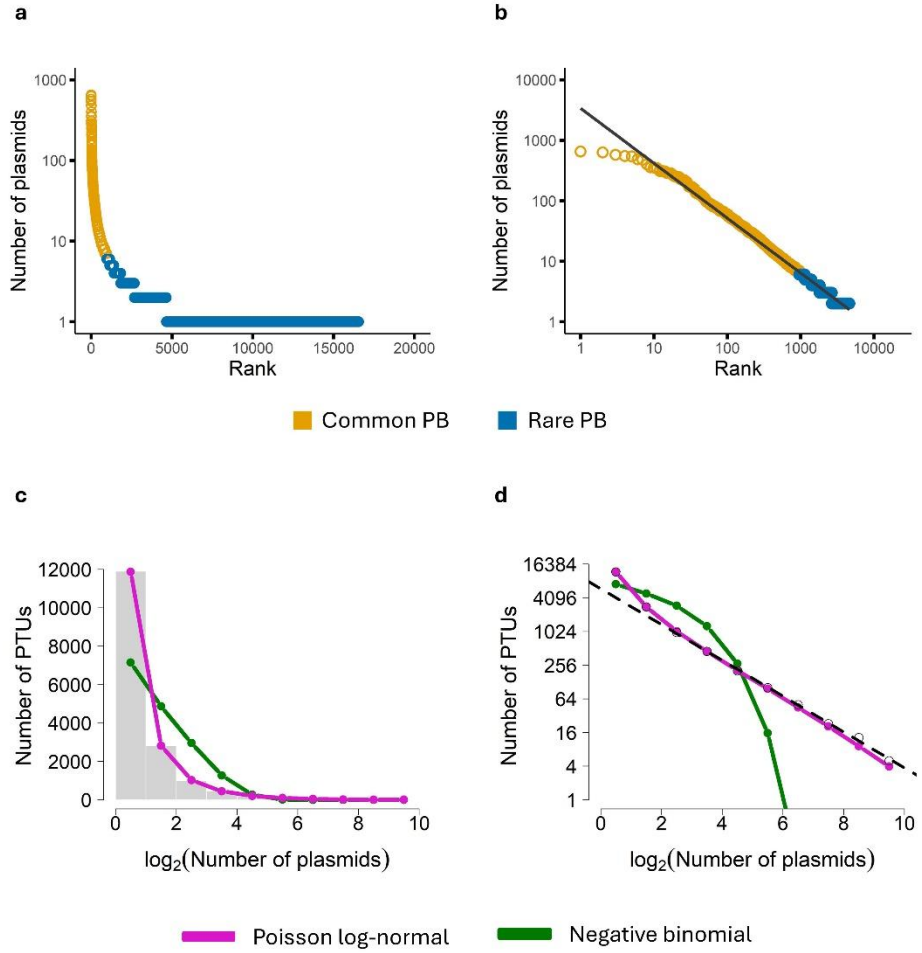

Fig. 1| Rank-abundance and relative abundance distributions of PTUs. a, the vertical axis (in $\log_{10}$ scale) represents the number of plasmids representing each PTU. The horizontal axis shows the ranking of PTUs, from the most common, with 650 plasmids, to the rarer, with a single plasmid. Yellow: represents the 961 most abundant PTUs, defined as the PTUs with more than 1% of the most abundant PTU (see main text) (Common Plasmid Biosphere); Blue: represents the 15,586 rarer PTUs (Rare Plasmid Biosphere). **b**, the same data as in a, but with horizontal axis in $\log_{10}$ scale. The straight line (slope = -0.91 and r^2 = 0.97) does not consider the plasmids with a single PTU. For this case, the number of plasmids in each rank is given, approximately, by the power-law $3{,}404.\text{rank}^{-0.91}$. Including all plasmids, slope = -0.74 and r^2 = 0.89. **c,** shows the PTU abundance distribution (PAD) along with the fits provided by the Poisson log-normal (pink) and the negative binomial (green) distributions. The

horizontal axis is in $\log_2$ scale. **d,** PAD on double logarithmic axes, where we also included a regression line (dashed) fitted to the tail of the distribution, with an estimated slope of approximately -1.063 (excluding the two leftmost points).

106

107 The fat-tailed distribution implies the existence of many rare PTUs, and the next
108 section will compare these rare PTUs with the common ones. To this aim, we define a
109 PTU as rare if its relative abundance is below a certain threshold. This threshold value
110 implies that the dataset is divided into two 'plasmid biospheres'(PBs): a Common PB (the
111 collection of PTUs with many plasmids) and a Rare PB (the collection of PTUs with few
112 plasmids). The concept of rarity has varied thresholds across studies [3,21]. Because the most
113 abundant PTU has 650 plasmids, we define a PTU as rare if its relative abundance is
114 below 1% - this threshold means that PTUs with fewer than 650x1% = 6.50 plasmids,
115 i.e., with six or fewer plasmids, belong to the Rare PB. Otherwise, they belong to the
116 Common PB. With this threshold, there are 961 PTUs in the Common PB comprising
117 30,428 plasmids and 15,586 PTUs in the Rare PB comprising 22,481 plasmids.

118

119 **Are PTUs rare or common because their hosts are rare or common,**
120 **respectively?**

121

122 Plasmids are extra-chromosomal genetic elements, but they can only replicate and
123 remain inside host cells (contrary, for example, to most viruses). Therefore, it is possible
124 that a rare PTU is rare just because its host species are rare, or that a common PTU is
125 common because its host is common. To test this hypothesis, we looked at the rank-
126 abundance distribution of host species.

127 Of the 7,715 hosts, 896 have no defined species (they are identified as 'Genus
128 sp.'). As we intend to perform analyses of the rarity of the host species, and since within
129 a sp. category there may be several species that we do not know if they are common or
130 rare, we excluded these genomes from the analyses. Therefore, for the host-related
131 analyses, we removed 3,593 plasmids that were identified in hosts with undefined species.

132 We used two approaches to address the rank-abundance distribution of host
133 species. In the first approach, we consider the number of plasmids sequenced in each
134 prokaryotic species. According to this approach, the most frequent host species in the
135 dataset is *E. coli*, harbouring as many as 9,343 plasmids. It is followed by *Klebsiella*
136 *pneumoniae* (7,977 plasmids), *Salmonella enterica* (2,247 plasmids), *Acinetobacter*

137 *baumannii* (1,204 plasmids), and 2,359 other prokaryotic species. In the second approach,
138 we count the number of sequenced genomes with plasmids of that species present in the
139 database. According to this approach, the most common species is also *E. coli* (3,068
140 genomes), followed by *K. pneumoniae* (2,328 genomes), *S. enterica* (1,210 genomes),
141 *Staphylococcus aureus* (807 genomes), and 2,359 other prokaryote species.

142 Strikingly, the rank abundance distribution of host species is well fitted by power-
143 laws using both approaches (**Fig. 2a, b**). We then partitioned the hosts' distribution as we
144 did for plasmids. For both ways of counting hosts, we used a threshold of 1% of the most
145 common host species. Using the first approach to define species frequency (based on the
146 number of plasmids found in a species), the most common host species is *E. coli*, with
147 9,343 plasmids identified. Therefore, with the threshold of 1%, common hosts species are
148 those with more than 93 plasmids; accordingly, there are 58 common and 2,305 rare host
149 species. For the second approach, a threshold of 1% means that a species is common if
150 the database contains at least 30 sequenced genomes, because the most frequent host
151 species in the database is *E. coli* with 3,068 genomes with plasmids. Accordingly, there
152 are 68 common and 2,295 rare.

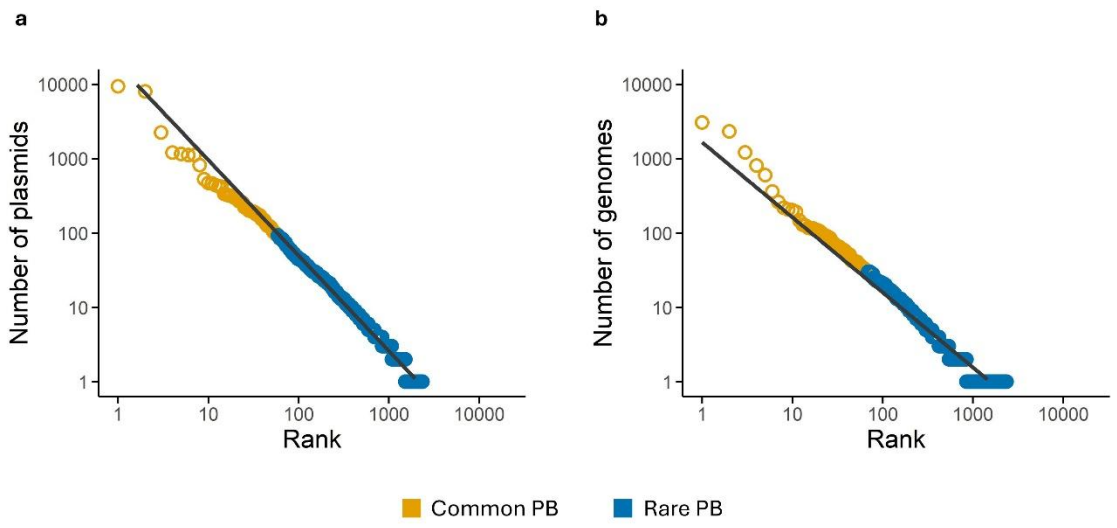

Fig. 2| – | Rank-abundance of host species. a, Rank of host species according to the number of plasmids found in the species (first approach – see main text). The first point on the left (rank = 1) represents *E. coli* with its 9,343 plasmids. Both scales are in Log_{10}. The straight line represents the power-law with slope -1.28, $R^2 = 0.98$. **b,** Rank of host species according to the number of genomes with plasmids of each species in the database

(second approach – see main text). The first point (rank = 1) on the left represents the 3,068 *E. coli* genomes with plasmids. Both scales are in Log_{10}. The straight line represents the power-law with slope -1.01, $R^2 = 0.94$.

153

154 According to the first approach, there are 13,534 plasmids hosted by rare hosts:
155 3,035 plasmids belong to the Common PB and 10,499 to the Rare PB. Among the 35,782
156 plasmids found in common hosts, 26,887 plasmids belong to the Common PB and 8,895
157 to the Rare PB. The association is strong and significant ($\chi^2 = 11{,}435$, d. f. = 1, $p < 2.2\text{x}10^{-}$
158 ^{16}, Cramér's V = 0.482). According to the second approach, there are 13,529 plasmids
159 hosted by rare hosts: 3,218 plasmids belong to the Common PB and 10,311 to the Rare
160 PB. Among the 35,787 plasmids found in common hosts, 26,704 plasmids belong to the
161 Common PB and 9,083 to the Rare PB. The association is strong and significant ($\chi^2 =$
162 10,630, d. f. = 1, $p < 2.2\text{x}10^{-16}$, Cramér's V ≈ 0.464). These analyses show an association
163 between PTU frequency and host frequency, but they do not inform us how far
164 plasmids/hosts are from full randomness. If plasmids of the Rare and Common PBs were
165 randomly assigned to rare and common host species, 57% of the plasmids would not
166 match the biosphere (rare or common) of their hosts. The observed percentage is ca. 25%
167 (using either of the two approaches for defining host numbers: 11,930 or 12,301
168 plasmids), which is between a perfect correspondence (0%) and total randomness (57%).
169 The above assessment overlooks what happens within each species. We examined
170 the association between the total number of plasmids found in a species (first approach)
171 or between the total number of sequenced genomes of a host species (second approach)
172 and the percentage of plasmids of the Rare PB in that species (**Supplementary Fig. 1**). If
173 plasmids from the Rare PB are associated with rarer species, both correlations should be
174 negative: most plasmids found in rare species would belong to the Rare PB, while the
175 percentage of plasmids from the Rare PB would be low in more common species. The
176 correlation is indeed negative, but the association is moderate only, and the widely
177 dispersed data points suggest that other factors may play a more significant role in
178 determining the relationship between PTU and host species abundances (**Supplementary**
179 **Fig. 1**). Both approaches for defining rare and common host species yielded essentially
180 the same pattern – qualitatively identical and quantitatively very similar. Therefore, we
181 present only the results based on the first approach in the sections that follow for
182 readability of the text.

183 Because prokaryote species may host plasmids both from the Common and Rare
184 PBs, we conducted a binomial test for each common species to assess whether the
185 proportion of common plasmids, (number of plasmids from the Common PB)/(total
186 number of plasmids), is significantly different from the proportion of common plasmids
187 in the sample, namely 30428/52909 = 0.58. Among the 58 common species, 18 host more
188 plasmids from the Rare PB than expected (binomial test, $p < 0.05$), and 11 common
189 species do not have more plasmids of one plasmid biosphere than expected (binomial test,
190 $p > 0.05$). The other 29 common species contain more plasmids from the Common PB
191 than expected (binomial test, $p < 0.05$). Yet, even among these 29 species, there are many
192 plasmids belonging to the Rare PB (**Supplementary Fig. 2**). We then performed a similar
193 analysis with the rare species (hence with 93 plasmids or fewer, following the first
194 approach). There are 2,305 species in these conditions: 21 contain more plasmids of the
195 Common PB than expected (binomial test, $p < 0.05$), 1,686 do not contain more plasmids
196 of any biosphere than expected (binomial test, $p > 0.05$). In 598 rare species, there are
197 more plasmids from the Rare PB than expected (binomial test, $p < 0.05$), but even in these
198 species, there are plasmids of the Common PB. In the particular case of PTUs with just
199 one plasmid (9,692 singletons), 6,307 (approximately 2/3 of singletons) were found in
200 rare host species and 3,385 (~ 1/3) in common host species. Notably, every host species
201 classified as common harboured at least one plasmid assigned to the Rare PB. In fact, any
202 common host species contains at least three different plasmids from the Rare PB. On the
203 other hand, there are 59 rare host species with plasmids only from the Common PB.
204 Overall, these analyses suggest that plasmids of common PTUs are not common
205 just because they were found in common host species, as we found so many plasmids of
206 the Rare PB in the common hosts. Likewise, one finds plasmids of the Common PB in
207 rare host species. These analyses indicate that plasmids from the Rare PB tend to be
208 associated with rare prokaryotic species, but this trend is only moderate as so many
209 plasmids of the Rare PB are found in common species (**Supplementary Figs. 1-3**).
210
211 **The Rare Plasmid Biosphere is Richer, Has More Distinct Virulence Genes**
212 **and Fewer Antibiotic Resistance Genes**
213
214 PTUs from the Common PB are more successful than from the Rare PB, in the
215 sense that they are much more numerous. To understand their success, we tested the
216 hypothesis that plasmids pertaining to bigger PTUs are also longer, eventually containing

217 more genes and enabling them to thrive more easily. However, the size (number of base
218 pairs) of plasmids in the Rare and Common PB is not different **(SupplementaryFig. 4a)**.
219 There are fewer coding DNA sequences (CDS) clusters in the Rare PB than in the
220 Common PB, but the strength of the association is irrelevant, so the statistical significance
221 is driven by a large sample size **(SupplementaryFig. 4b)**. Therefore, data does not
222 support the hypothesis that plasmids from the Common PB are more abundant because
223 they contain more CDS.

224 The two biospheres contain 875,346 CDS clusters in total, however the Common
225 PB has only 76,491 CDS clusters (8.74%), while the Rare PB has 826,373 (94.41%), a
226 significant difference (the combined percentages exceed 100% because some CDS
227 clusters are present in both plasmid biospheres). Another way to see the biological
228 significance of this difference is that, while 48,973 CDS clusters are specific to the
229 Common PB (i.e., absent from the Rare PB), 798,855 CDS clusters are only found in the
230 Rare PB. The difference in the number of CDS clusters between the two plasmid
231 biospheres is significant. Moreover, although there are fewer plasmids in the Rare PB,
232 they have significantly more CDS clusters than the Common PB (**Fig. 3**).

233 Among the 4,814 plasmids containing virulence genes (VGs), there are only 546
234 distinct VGs, so some of them are found more than once across plasmids. The Common
235 PB contains 4,053 plasmids with VGs but only 281 distinct VGs, and the Rare PB
236 includes 761 plasmids with VGs but 423 distinct VGs. Therefore, only 51.5% ($CI_{95\%}$:
237 [47.2%, 55.7%]) of distinct VGs are in the Common PB, whereas 77.5% ($CI_{95\%}$: [73.7%,
238 80.9%]), are in the Rare PB, with an odds ratio of 0.31 ($CI_{95\%}$: [0.24, 0.40]), indicating a
239 significantly higher prevalence of these genes in the Rare PB. Hence, the diversity of
240 virulence genes is higher in the Rare PB than in the Common PB (**Fig. 3**). The biological
241 significance of this difference becomes particularly evident when considering the
242 following: of the 546 distinct VGs identified, the Rare PB includes 265 that are absent
243 from the Common PB. As such, failing to detect the Rare PB would result in overlooking
244 48.5% of the total VGs. In contrast, only 123 VGs (22.5%) found in the Common PB are
245 not present in the Rare PB.
246

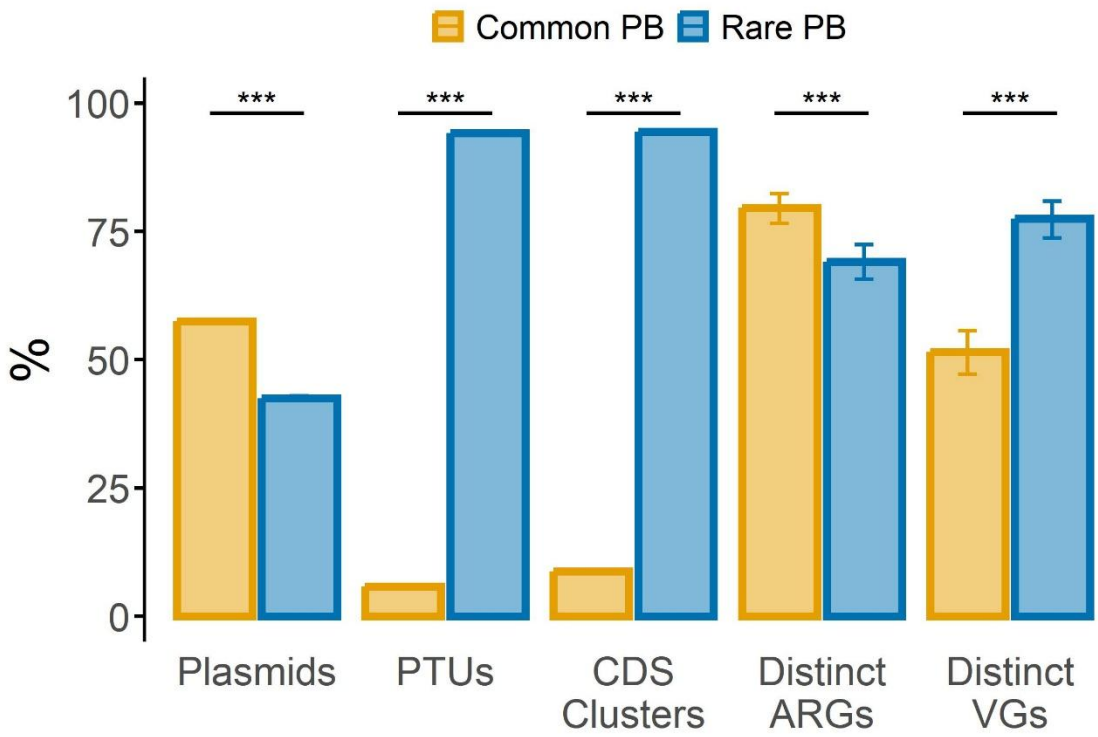

Fig. 3| Characteristics of the Common and Rare Plasmid Biospheres. Plasmids: The Rare PB includes fewer plasmids than the Common PB ($\chi^2(1) = 2387$, $p < 2.2\text{x}10^{-16}$, Cramér's V = 0.150, $CI_{95\%}$: [0.145, 0.157]). The number of PTUs is higher in the Rare PB than in the Common PB ($\chi^2(1) = 25849$, $p < 2.2\text{x}10^{-16}$, Cramér's V = 0.884, $CI_{95\%}$: [0.879, 0.889]). The number of CDS clusters is also higher in the Rare PB than in the Common PB (χ^2 (1) = 1286069, $p < 2.2\text{x}10^{-16}$; Cramér's V = 0.857, $CI_{95\%}$: [0.856, 0.858]). The number of distinct VGs is higher in the Rare PB than in the Common PB (χ^2 (1) = 80, $p < 2.2\text{x}10^{-16}$, Cramér's $V = 0.272$, $CI_{95\%}$: [0.208, 0.332]). The number of distinct ARGs is higher in the Common PB than in the Rare PB (χ^2 (1) = 22, $p = 3.5\text{x}10^{-6}$, Cramér's V = 0.120, $CI_{95\%}$: [0.069, 0.167]).

247

248 The VFDB database classifies virulence genes into 14 categories. One category –
249 post-translational modification– was absent from all plasmids in the dataset. However, 12
250 categories were identified in both plasmid biospheres (**Supplementary Fig. 5a**). No
251 category was found exclusively in the Common PB, but one category – stress survival –
252 was found exclusively in the Rare PB, represented by the *sodCI* gene. This gene, detected
253 in two plasmids, encodes the superoxide dismutase SodCI, which enhances the virulence

254 of *Salmonella enterica* serovar Typhimurium by protecting against oxidative stress
255 generating the superoxide produced during the macrophage oxidative burst [26].
256 Concerning antibiotic resistance genes (ARGs), we found 79.6% ($CI_{95\%}$: [76.6%,
257 82.4%]) of the 761 different ARGs in the Common PB, whereas 69.1% ($CI_{95\%}$: [65.7%,
258 72.4%]) of the ARGs were present in the Rare PB, with an odds ratio of 1.75 ($CI_{95\%}$:
259 [1.38, 2.21]), indicating that the odds of ARG presence in Common PB were higher than
260 those in the Rare PB (**Fig. 3**). The dataset contains 14,718 plasmids with 761 distinct
261 ARGs, of which 371 are present in both plasmid biospheres, 155 are only present in the
262 Rare PB, and 235 in the Common PB. Therefore, although the Common PB is richer in
263 ARGs than the Rare PB, overlooking the Rare PB would mean ignoring the presence of
264 the 155 ARGs specific to this plasmid biosphere.
265 According to AMRFinderPlus, plasmids in the dataset contain genes conferring
266 resistance to 28 distinct antibiotic categories (**SupplementaryFig. 5b**). Of these, 27 are
267 found in both plasmid biospheres, while one category—Nitroimidazole—is exclusive to
268 the Rare PB. This category includes the genes *nimA* and *nimD* found in plasmids from
269 *Bacteroides fragilis*, and *nimE* found in plasmids from *Bacteroides fragilis*, *Bacteroides*
270 sp. PHL 2737, and *Bacteroides xylanisolvens*.
271
272 **PTUs of the Rare PB Spans Through More Host Species, and, per Plasmid,**
273 **have Broader Host-Range**
274
275 Plasmids of the Rare PB represent only 42.5% of all the 52,909 plasmids analysed in this
276 study. Yet, this section shows that they span more taxa than the plasmids from the
277 Common PB (Table 1). Once again, we removed from the analysis at species level the
278 plasmids identified in genomes with unidentified species (3,593 plasmids, 87 from
279 archaea and 3,506 from bacteria).
280 There are 27 hosts' phyla with plasmids belonging to the Rare PB but just 11
281 hosts' phyla with plasmids of the Common PB. The two most represented phyla for each
282 plasmid biosphere are the same: Pseudomonadota, Bacillota. The third most common host
283 phylum is not the same for the plasmids belonging to the Common PB (Spirochaedota,
284 2.8%) and Rare PB (Actinomycedota, 7.8%). The phylum Spirochaedota (2.8%) is in
285 fourth place in the Rare PB.In general, plasmids from the Rare PB were found across
286 more phyla than plasmids from the Common PB (**Supplementary Fig. 6**).
287

288

Table 1: Taxonomy of host species harbouring plasmids. [1]

	Total nr (hosts)	Hosts with plasmids of the Common PB	Hosts with plasmids of the Rare PB	Hosts with plasmids only of the Common PB	Hosts with plasmids only of the Rare PB
Phylum	27	11	27	0	16
Class	56	15	56	0	41
Order	133	35	133	0	98
Family	302	58	300	2	244
Genus	888	144	880	8	744
Species	2,363	465	2,304	59	1,898

289 [1]Among the 20,580 genomes with plasmids, 218 genomes are from archaea (with 474 plasmids) and 20,362
290 genomes are from bacteria (with 52,435 plasmids).The 52,909 plasmids analysed in this study were found
291 in 2,363 prokaryote species: 2,211 bacterial species and 152 species of archaea. The 218 genomes of
292 archaea hosted 474 plasmids classified in 424 PTUs; seven plasmids belonged to the Common PB, while
293 all remaining archaeal plasmids belonged to the Rare PB. The 20,362 plasmid-bearing bacteria hosted
294 52,435 plasmids (16,123 PTUs), of which 30,421 belonged to the Common PB and 22,014 to the Rare PB.
295

296 Recently, Redondo-Salvo et al. classified [12] host-ranges of PTUs into a scale of
297 six grades, from grade I (restricted to a single host species) to grade VI (able to colonize
298 different phyla). For example, the 32 host species harbouring the 637 plasmids (more 13
299 plasmids from five unidentified species) of the largest PTU belong to a single family
300 (Enterobacteriaceae) and just 12 genera (*Citrobacter*, *Cronobacter*, *Enterobacter*,
301 *Escherichia*, *Klebsiella*, *Kluyvera*, *Leclercia*, *Phytobacter*, *Pseudescherichia*, *Raoultella*,
302 *Salmonella*, *Shigella*). Although this is the largest PTU, it is present in different genera
303 within a single family. Therefore, it is of 'intermediate' host-range, i.e., grade III. In the
304 Common PB, a PTU comprising 24 plasmids is found across three phyla:
305 Campylobacterota, Bacillota (Firmicutes), and Actinomycetota. This is a case of very
306 broad host-range (grade VI). However, the PTU found in more phyla is present in four
307 phyla (Bacillota, Verrucomicrobiota, Actinomycetota and Pseudomonadota) despite
308 belonging to the Rare PB. Therefore, even rare PTUs may be distributed among very
309 distinct taxonomic groups. This was unexpected because, by definition, a PTU belongs
310 to the Rare PB if they have up to six plasmids in the dataset. This specific PTU present
311 in four phyla has only four plasmids. As many as 948 Common PTUs are in a unique

312 phylum. Ten Common PTUs are present in at least 20 host species, and one common PTU
313 has 477 plasmids of which 460 are present in 42 identified species and 17 in unidentified
314 species (1 in *Citrobacter sp*., 1 in *Aeromonas sp*., 3 in *Shewanella sp*. and 12 in
315 Providencia sp.). This means that this PTU is present in at least 46 species. On the other
316 extreme, 325 Common PTUs are in a unique host species, hence very narrow-host-range
317 plasmids (grade I) (**Supplementary Fig. 7**).
318 Next, we analyse and compare the host-ranges of Common and Rare PTUs in a
319 systematic way. It is likely that common PTUs are present in more taxa (e.g., in more
320 species) than rare PTUs simply because common PTUs are, according to our definition,
321 those having more representative plasmids in the dataset (therefore with more
322 opportunities to be found in more taxa). To rule out this confounding factor concerning
323 host-range, we calculated the number of hosts' species per number of plasmids of each
324 PTU. This ratio is significantly higher (Wilcoxon signed-rank, $p < 0.05$) for the PTUs of
325 the Rare PB than of the Common PB **(Fig. 4a).** Since thousands of PTUs contain only
326 one plasmid (singletons), we conducted a similar comparison excluding these PTUs from
327 the analysis. Again, the diversity of hosts' species per plasmid is higher in the Rare PB
328 than in the Common PB **(Fig. 4b)**. These ratios may be too low in the bigger PTUs. For
329 example, the hosts' species/plasmid ratio for a PTU with 30 plasmids present in 5 species
330 (ratio = 5/30) is lower than for a PTU with 5 plasmids in 5 species (ratio = 5/5). To avoid
331 these biases, we compared the success of PTUs with seven plasmids (belonging to the
332 Common PB) with those with six plasmids (belonging to the Rare PB). Again, for this
333 case, the PTUs of the Rare PB are present in more host species, on average, than the PTUs
334 of the Common PB **(Fig. 4c).** A similar analysis for genus, family, order, class and
335 phylum has shown that this ratio is consistently higher in the Rare than in the Common
336 PBs (Wilcoxon signed-rank test $p < 0.05$, large effect sizes for all cases). **Figs. 4 d to f**
337 show these analyses for the phylum level, and **Supplementary Fig. 8** shows these
338 analyses for the genus, family, order, and class levels. For all cases, the relative host range
339 of the Rare PB' PTUs is higher than that of the Common PB. Therefore, other factors
340 must be responsible for the success of Common PTUs.
341

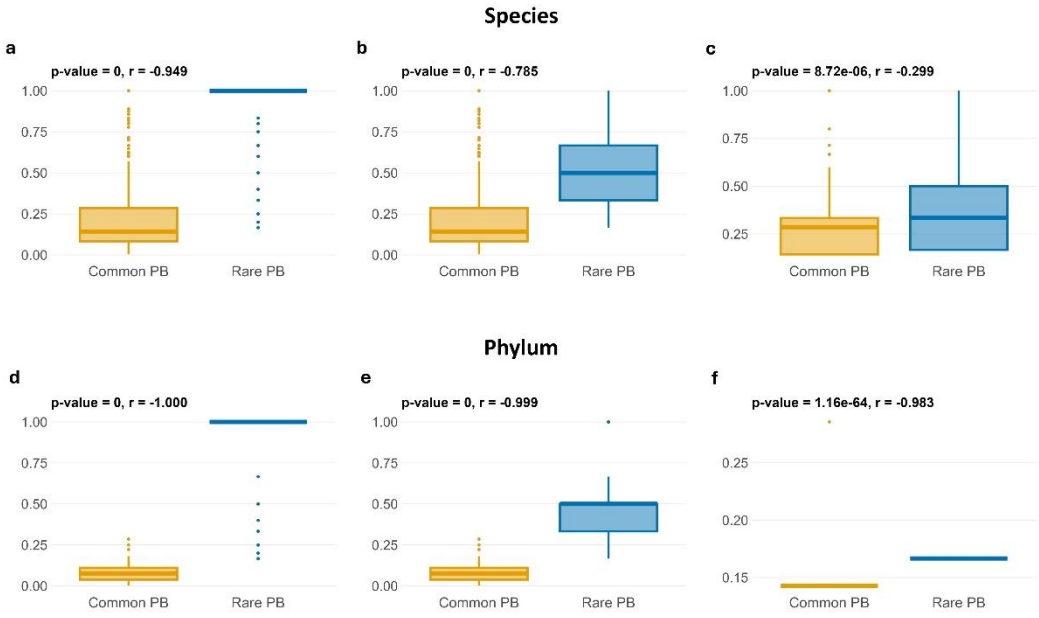

Fig. 4| The relative host-ranges of PTUs of the Rare and Common PB. The relative success of each PTU, measured as the number of hosts' taxa per number of plasmids in that PTU. From left to right: all plasmids, not including singletons, including only PTUs with seven (Common) or six (Rare) plasmids, respectively. **a, b, c**, species level. **d, e, f**, phylum level. Host-range is higher for the PTUs of the Rare PB than that of the Common PB for both levels (Wilcoxon signed-rank test $p < 0.05$, large effect size in all cases, except for the species level, where the effect size is moderate, r = -0.299, when comparing PTUs with six versus seven plasmids).

342

343 **Plasmid mobility is higher among plasmids of the Common PB**

344 With the success of the most common PTUs still unexplained, we looked for their
345 transfer ability. Plasmid mobility falls into four main classes. Conjugative plasmids
346 (pCONJ) encode all the necessary genes for self-transfer by conjugation, namely a
347 relaxase, an origin of transfer (*ori*T) and a type IV secretion system. Mobilizable plasmids
348 (pMOB) retain only part of this machinery—typically a relaxase and an origin of transfer,
349 *ori*T—and "hitch-hike" on the conjugation systems of co-resident plasmids. Some
350 plasmids, pOriT, encode just *ori*T; these plasmids typically need to "hitch-hike" on more
351 than one plasmid, using the conjugation system of one plasmid and the relaxase of a third
352 plasmid. Finally, some plasmids are non-transmissible (pNT), lacking all transfer genes

353 and being inherited only through cell division [14,27,28]. The proportion of each class in the
354 Common and Rare PBs tend to be significantly different from the expected values if they
355 were randomly distributed through the PBs (Table 2).
356

Table 2| Plasmids' transfer ability in the Common and Rare PBs.

		N	Obs (%)	$CI_{95\%}$	E (%)[1]	$CI_{95\%}$
Common PB	pCONJ	10928	35.9	[35.4, 36.4]	28.5	[28.0, 29.0]
	pMOB	6438	21.2	[20.7, 21.6]	22.5	[22.0, 23.0]
	pOriT	5871	19.3	[18.8, 19.7]	16.5	[16.1, 16.9]
	pNT	7191	23.6	[23.1, 24.1]	32.5	[31.9, 33.0]
Rare PB	pCONJ	4174	18.6	[18.1, 19.1]	28.5	[27.9, 29.1]
	pMOB	5460	24.3	[23.7, 24.8]	22.5	[21.9, 23.0]
	pOriT	2860	12.7	[12.3, 13.2]	16.5	[16.0, 17.0]
	pNT	9987	44.4	[43.8, 45.1]	32.5	[31.8, 33.1]

357 [1] Within either PB, the expected percentage composition mirrors the global proportions
358

359 The percentage of pCONJ is higher among Common PTUs than Rare PTUs.
360 Together, the proportion of pMOB and pOriT plasmids is also higher in the Common PB
361 (40.4 %, $CI_{95\%}$ = [39.9, 41.0]) than in the Rare PB (37.0%, $CI_{95\%}$ = [36.4, 37.6])
362 (**Supplementary Table 3**). In contrast, pNT plasmids are more frequent in the Rare PB
363 (44.4 %) than in the Common PB (23.6%). These results suggest that transfer ability plays
364 a fundamental role in the success of Common PTUs, although 55.6% of the plasmids of
365 the Rare PB are transferable and 23.6% of the plasmids of the Common PB are non-
366 transmissible.
367

368 **Are rare hosts receiving plasmids from phylogenetically close common hosts?**

369 We have seen that several rare host species harbour plasmids from common PTUs.
370 For example, there are 59 rare host species with plasmids only from the Common PB.
371 How can we explain this observation? One possible explanation is that these hosts
372 received the plasmid from phylogenetically close common hosts. This is the case of
373 *Vibrio owensii*, a rare species with eight plasmids from three PTUs. It belongs to the
374 genus Vibrio, which includes one common species and 53 rare ones, and to the family
375 Vibrionaceae, which includes one common and 64 rare species. The common species
376 within this genus and family shares all three PTUs with *Vibrio owensii*. However, this
377 case of *Vibrio owensii* is far from general.

To investigate the generality that rare hosts harbour common plasmids due to their taxonomic proximity to common hosts with the same PTUs, we performed the following analysis. For each of 59 rare species, we looked at plasmids to check whether their PTUs were also found in common host species of the same genus or family. For 46 species (78%), there was no common host species within the same genus harbouring plasmids of any of the common PTUs found in the rare species. At the family level, 40 species (68%) had no common host species within the same family with plasmids of any of the same PTUs. For example, *Chlamydia trachomatis* is a rare species with 33 plasmids, all belonging to the same PTU. This species belongs to the genus Chlamydia and the family Chlamydiaceae. Both the genus and family include 14 other species, all of which are also rare. Overall, these results suggest that rare hosts carry only common plasmids generally not owing to the taxonomic proximity to common host species.

Discussion

It has been proposed that rare microbial species within ecosystems act as reservoirs of genetic resources, which may become accessible under favourable conditions, thereby potentially enhancing ecological stability through insurance effects [1]. Our results show that, also in the case of plasmids, some taxonomic units are very abundant while many are very rare and that rare PTUs harbour a strong diversity of genes, including virulence genes. The main caveat of this work is certainly the fact that the plasmids analysed in this study were not derived from a single ecosystem, so caution is warranted when interpreting the broader implications of our findings [10]. Yet, several findings concerning this dataset present striking parallels with important findings using data from ecosystems regarding plants, animals or microorganisms [1,4,5,13].

Recognizing the existence of rare biospheres within bacterial communities has been highly significant, as these bacterial species may constitute extensive reservoirs of genomic innovation [3,5,21,22,22]. They have a great potential for dynamic adaptation because individual cells from low-abundance bacterial taxa can proliferate rapidly through clonal replication when environmental conditions shift to favour their growth [4,5]. If this ability of clonal amplification is relevant for bacterial cells, it is even more so in the case of plasmids. These genetic elements may hitchhike the clonal expansion of their host, but they can also expand horizontally through bacterial communities, even between bacterial taxa, if the conditions are appropriate [14,29,30]. Moreover, plasmids may even help other

411 plasmids or be helped by other plasmids to transfer to other cells [28,31–34]. These plasmids'
412 characteristics might annihilate the fat-tailed distributions frequently observed in
413 bacterial communities, but our findings show that that was not the case.

414 The plasmids of this study have very diverse origins. Nevertheless, we observed
415 the same striking distribution pattern as seen in microbial communities from a given
416 ecosystem or habitat: characterized by a few highly abundant PTUs and a fat-tail of rare
417 ones (**Fig. 1 a, b**). The analysis of the PTU Abundance Distribution (PAD) curve revealed
418 that the Poisson log-normal distribution provided an excellent fit (**Fig. 1 c, d**). This result
419 aligns with recent findings [13] and suggests a compelling similarity between plasmid
420 abundance patterns and well-established results from other biodiversity studies. This
421 observation is noteworthy, as the shape of the PAD may offer insight into the underlying
422 processes of plasmid diversification, as well as the emergence and persistence of rare
423 plasmids. For instance, under Hubbell's Neutral Theory of Biodiversity and
424 Biogeography [24], a monotonically decreasing function, such as the ones we observed,
425 implies that new species arise with very few individuals. The PAD in our dataset peaks
426 at singletons, likely reflecting a process of "speciation" in which novel PTUs emerge at
427 initially very low abundance. This interpretation is consistent with the well-known ability
428 of plasmids to change [12,35–38]. These questions should be addressed in future research.

429 An important biological implication of the power-law is that some PTUs are
430 highly successful. What is the cause of their success? If PTU's success was totally
431 attributed to that of their hosts, one should expect 0% of mismatches, that is, all plasmids
432 belonging to common PTUs should be hosted by common species and all plasmids
433 belonging to rare PTUs should be hosted by rare species. If PTUs were randomly assigned
434 as rare or common, 57% plasmids would not match the assignment of their hosts. We
435 observed an intermediate number: 25% of the plasmids belong to a rare or a common
436 species, despite belonging to a common or rare PTU, respectively. We conclude that some
437 plasmids and the corresponding hosts share an evolutionary success. Yet, they are far
438 from the majority. For example, the dataset includes 9,343 plasmids from *E. coli*—one of
439 the most extensively studied bacterial species. Despite being the most common host,
440 1,649 of these plasmids (18%) belong to rare PTUs, and thus to the Rare PB.
441 *Lactiplantibacillus plantarum*, the eighth most common host species, shows an even
442 greater proportion: 77.4% of its plasmids (632 out of 816) belong to the Rare PB, with
443 only 184 in the Common PB, or *Levilactobacillus brevis*, a common species with 104 of
444 its 112 (92.86%) plasmids in the Rare PB. These examples also suggest that the success

445 of plasmids from common PTUs is not necessarily driven by the success of their common
446 hosts, nor are these hosts responsible for the success of their plasmids from the common
447 PTUs. The same for plasmids from rare PTUs.

448 Intriguingly, the long list of rare PTUs is associated with exceptional diversity in
449 both coding sequence clusters and virulence genes (**Fig. 3**). Therefore, gene diversity is
450 not responsible for the success of the common PTUs. In contrast, the diversity of
451 antibiotic resistance genes (ARGs) is higher among the plasmids from the common PTUs
452 than in the rare ones. One possible explanation for such contrast is that the sequencing of
453 RefSeq genomes may have been motivated by antibiotic resistance. Yet, even if antibiotic
454 resistance genes are less diverse in the Rare PB, 155 out of 761 genes (20.4%) are only
455 present in the Rare PB. For example, resistance genes of the Nitroimidazole antibiotic
456 class were found only in the Rare PB (**Supplementary Fig. 5**). Let us suppose that future
457 studies confirm that most antibiotic resistance genes are indeed more often present in the
458 Common PB. That might suggest that resistance genes in the Rare PB are not as relevant
459 as, e.g., virulence genes. Yet this interpretation should be approached with caution. If new
460 antibiotics are developed — particularly those with novel modes of action — potential
461 resistance genes are more likely to emerge from CDS that have not yet been identified as
462 ARGs, rather than from those already known. Our results, which reveal exceptionally
463 high CDS cluster diversity among plasmids in the Rare PB, suggest that this group
464 represents a vast gene pool from which novel resistance genes may evolve. As such, these
465 findings serve as a warning: resistance to future antibiotics may arise from currently
466 uncharacterized gene sequences within the Rare PB.

467 A recent study has shown that ARGs are preferentially acquired and carried by
468 broad-host-range plasmids and that these plasmids are more 'plastic' (e.g., higher
469 recombination rates or higher mobility) [39]. We have found that one PTU belonging to the
470 Rare PB is present in hosts from as many as four different phyla (**Supplementary Fig.**
471 **7**). Future studies should look for recombinases in these plasmids and study rates of allelic
472 exchanges and changes of gene repertoires, as well as their mobility [39]. This PTU is a
473 peculiar case of an important characteristic of rare PTUs concerning host ranges: per
474 number of plasmids in each PTU, the host range of rare PTUs is higher than common
475 PTUs. Strikingly, we observed the same pattern at all taxonomic levels, from species to
476 phyla (**Fig. 4** and **Supplementary Fig. 8**). One could argue that the overall broader host
477 range of rare PTUs is irrelevant because these PTUs have very few plasmids. However,
478 these PTUs are very numerous, representing 94% of the PTUs and comprising almost half

479 of the plasmids in the dataset. Thus, these PTUs likely play critical roles in bacterial
480 evolution when hosted by different prokaryote species/genus/…/phyla, eventually co-
481 inhabiting the same cell with other plasmids, from both plasmid biospheres.

482 In bacterial communities, it has been proposed that rare taxa tend to remain rare
483 because they are less visible to predators, such as viruses or protists [4,40]. Yet, our results
484 strongly suggest that this explanation does not apply to plasmids. If it did, rare host
485 species should exclusively harbour plasmids from the Rare PB, while common hosts
486 should lack plasmids from this pool; this scenario is not supported by our data.

487 We have seen that the very good fit of a Poisson log-normal distribution to the
488 PTU abundance distribution suggests that many new rare PTUs likely arise from genetic
489 changes of the already existing pool of plasmids (from common or rare PTUs). Therefore,
490 a potential mechanism underlying the emergence and/or persistence of plasmids in the
491 Rare PB could be the acquisition of new genes, also likely to transform plasmids from a
492 common PTU into a rare PTU. Nevertheless, this mechanism seems unlikely because
493 plasmids belonging to Rare and Common PBs do not differ significantly in size (base-
494 pairs), and plasmids in the Common PB contain only a slightly higher number of CDS
495 (**Supplementary Fig. 4**).

496 Finally, plasmid mobility may explain the success of common PTUs. About
497 76.4% of the plasmids in the Common PB were transferable (pCONJ, pMOB, and pOriT
498 plasmids), in contrast with the 55.7% transferable plasmids from the Rare PB.
499 Consequently, the proportion of non-transmissible plasmids (pNT), is almost the double
500 (44.4%) in the Rare PB than in the Common PB (23.6%) (**Table 2**). It is intriguing,
501 however, that so many pNT plasmids are rare. There are some possible explanations, for
502 example: (i) they may be classified as pNT, but future research may find *OriT*'s sequences
503 on them [14,41]; (ii) they belong to a PTU that includes transferable plasmids; (iii) or they
504 were recently pOriT, pMOB or pCONJ, but they lost relevant transfer genes [37,42].

505 The dataset contained 20,580 genomes, 218 of archaea and 20,362 of bacteria,
506 harbouring plasmids from 16,547 PTUs. Yet, none of these PTUs is present both in
507 archaea and bacteria – even if one of the PTUs of archaea belong to the Common PB.
508 This suggests that inter-domain plasmid transfers are rare events (possibly because their
509 habitats and selective pressure do not coincide) [43,44].

510 Our study has several limitations. The first limitation was already mentioned
511 throughout this study: plasmids (and their hosts, bacteria and archaea) studied here do not
512 belong to a single ecosystem/habitat. For example, *E. coli* is certainly not the most

abundant bacterial species. Future studies should focus on plasmids from a single habitat. That would imply sequencing all plasmids present in a single microbiome. A second limitation is the very low number of plasmids from archaea. Future studies should study their plasmids in this context. This limitation is related to the first one, as bacterial species have been more interesting, e.g., from the clinical or industrial point of views. Third, the dataset used in this study is very likely biased towards clinical bacteria. Different datasets have different characteristics, e.g. the number of plasmids carrying antibiotic resistance genes [45]. Yes, despite these limitations, the excellent fit of a power-law to the rank abundance or the Poisson log-normal to the PTU abundance distribution is remarkable. That implies that PTUs are strongly unbalanced, leading to the definition of a rare and a common plasmid biosphere and the results shown here.

Our results strongly suggest that the rare plasmid biosphere may play critical roles in microbial ecosystems. Plasmids of rare PTUs act as reservoirs of diverse coding sequences, including genetic resources linked to virulence and antibiotic resistance, potentially providing an 'insurance effect' in response to selective pressures such as antibiotic use, as previously hypothesized for microorganisms [1,46]. However, future research is needed to address key questions within specific habitats: (1) Do rare plasmid biospheres exist at the level of single habitats? (2) In which habitats or ecosystems are rare plasmid biospheres present? (3) Is the metabolic activity of cells harbouring rare or common plasmids different?

Our finding has at least two main practical implications. First, researchers can look carefully at rare PTUs to identify possible candidate genes that confer resistance to antibiotics under clinical trials. Second, by studying plasmids of rare PTUs, we can also find relevant new genes, such as those involved in nutrient cycling, bioremediation, or pollutant degradation [47–49].

Acknowledgments

João S. Rebelo and Célia P. F. Domingues acknowledge FCT-Fundação para a Ciência e a Tecnologia, IP for their fellowships (PhD grants SFRH/BD/04631/2021, and UI/BD/153078/2022, respectively). FCT also supports our research through the strategic funding to cE3c (DOI:10.54499/UIDB/00329/2020) and the associate lab CHANGE (LA/P/0121/2020). The authors thank Eduardo P.C. Rocha, Lounès Chichi and João. A. Gama for discussions and critical reading.

546

547 **Author contributions**

548 Conceptualization, J.S.R., C.P.F.D., L.B.A., and F.D; Software, J.S.R., C.P.F.D. and
549 L.B.A; Methodology, J.S.R., C.P.F.D., L.B.A., T.N., and F.D.; Investigation, J.S.R. and
550 C.P.F.D.; Formal Analysis, J.S.R., C.P.F.D., L.B.A., T.N. and F.D.; Writing – Original Draft,
551 J.S.R., C.P.F.D., L.B.A., and F.D.; Resources, T.N.; Visualization, J.S.R., C.P.F.D., and L.B.A;
552 Supervision, T.N. and F.D.

553

554 **References**

555

556 1. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems
557 strings. *ISME J* **11**, 853–862 (2017).
558 2. Litchman, E. *et al.* Refocusing the microbial rare biosphere concept through a
559 functional lens. *Trends in Ecology & Evolution* **39**, 923–936 (2024).
560 3. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat*
561 *Rev Microbiol* **13**, 217–229 (2015).
562 4. Pedrós-Alió, C. The Rare Bacterial Biosphere. *Annual Review of Marine Science* **4**,
563 449–466 (2012).
564 5. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare
565 biosphere". *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12115–12120 (2006).
566 6. Jørgensen, B. B., Dunker, R., Grünke, S. & Røy, H. Filamentous sulfur bacteria,
567 Beggiatoa spp., in arctic marine sediments (Svalbard, 79 degrees N). *FEMS*
568 *Microbiol Ecol* **73**, 500–513 (2010).
569 7. Mallon, C. A. *et al.* Resource pulses can alleviate the biodiversity-invasion
570 relationship in soil microbial communities. *Ecology* **96**, 915–926 (2015).
571 8. van der Gast, C. J. *et al.* Partitioning core and satellite taxa from within cystic
572 fibrosis lung bacterial communities. *ISME J* **5**, 780–791 (2011).

9. van Elsas, J. D. *et al.* Microbial diversity determines the invasion of soil by a bacterial pathogen. *Proceedings of the National Academy of Sciences* **109**, 1159–1164 (2012).
10. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**, D733–D745 (2016).
11. Garcillán-Barcia, M. P., Redondo-Salvo, S. & de la Cruz, F. Plasmid classifications. *Plasmid* **126**, 102684 (2023).
12. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature Communications* **11**, 3602 (2020).
13. Callaghan, C. T., Borda-de-Água, L., van Klink, R., Rozzi, R. & Pereira, H. M. Unveiling global species abundance distributions. *Nat Ecol Evol* **7**, 1600–1609 (2023).
14. Ares-Arroyo, M., Coluzzi, C. & Rocha, E. P. C. Origins of transfer establish networks of functional dependencies for plasmid transfer by conjugation. *Nucleic Acids Research* **51**, 3001–3016 (2023).
15. Wang, Y. & Dagan, T. The evolution of antibiotic resistance islands occurs within the framework of plasmid lineages. *Nat Commun* **15**, 4555 (2024).
16. Donlan, R. M. Biofilms: Microbial Life on Surfaces. *Emerg. Infect. Dis.* **8**, 881–890 (2002).
17. Johnson, T. J. & Nolan, L. K. Pathogenomics of the Virulence Plasmids of *Escherichia coli*. *Microbiol Mol Biol Rev* **73**, 750–774 (2009).
18. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* **31**, e00088-17 (2018).

19. Heuer, H. & Smalla, K. Plasmids foster diversification and adaptation of bacterial populations in soil. *Fems Microbiol Rev* **36**, 1083–1104 (2012).
20. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).
21. Pascoal, F., Costa, R. & Magalhães, C. The microbial rare biosphere: current concepts, methods and ecological principles. *FEMS Microbiology Ecology* **97**, fiaa227 (2021).
22. Jia, X., Dini-Andreote, F. & Falcão Salles, J. Community Assembly Processes of the Microbial Rare Biosphere. *Trends in Microbiology* **26**, 738–747 (2018).
23. McGill, B. J. *et al.* Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecol Lett* **10**, 995–1015 (2007).
24. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography | Princeton University Press*. (Princeton University Press, 2001).
25. Prado, P. I., Miranda, M. D. & Chalomm, A. sads: Maximum likelihood models for species abundance distributions. R package version 0.4.2. (2018).
26. Golubeva, Y. A. & Slauch, J. M. Salmonella enterica Serovar Typhimurium Periplasmic Superoxide Dismutase SodCI Is a Member of the PhoPQ Regulon and Is Induced in Macrophages. *Journal of Bacteriology* **188**, 7853–7861 (2006).
27. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & De La Cruz, F. Mobility of Plasmids. *Microbiol Mol Biol Rev* **74**, 434–452 (2010).

618 28. Ares-Arroyo, M., Coluzzi, C., Sousa, J. A. M. de & Rocha, E. P. C. Hijackers,
619 hitchhikers, or co-drivers? The mysteries of mobilizable genetic elements. *PLOS*
620 *Biology* **22**, e3002796 (2024).
621 29. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F.
622 Mobility of Plasmids. *Microbiol Mol Biol R* **74**, 434-+ (2010).
623 30. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San
624 Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial
625 evolution. *Nat Rev Microbiol* **19**, 347–359 (2021).
626 31. Gama, J. A., Zilhao, R. & Dionisio, F. Multiple plasmid interference - Pledging
627 allegiance to my enemy's enemy. *Plasmid* **93**, 17–23 (2017).
628 32. Gama, J. A., Zilhão, R. & Dionisio, F. Co-resident plasmids travel together. *Plasmid*
629 **93**, 24–29 (2017).
630 33. Gama, J. A., Zilhão, R. & Dionisio, F. Conjugation efficiency depends on intra and
631 intercellular interactions between distinct plasmids: Plasmids promote the
632 immigration of other plasmids but repress co-colonizing plasmids. *Plasmid* **93**, 6–16
633 (2017).
634 34. Gama, J. A., Zilhão, R. & Dionisio, F. Plasmid Interactions Can Improve Plasmid
635 Persistence in Bacterial Populations. *Front. Microbiol.* **11**, 2033 (2020).
636 35. Lang, A. S., Buchan, A. & Burrus, V. Interactions and evolutionary relationships
637 among bacterial mobile genetic elements. *Nat Rev Microbiol* **23**, 423–438 (2025).
638 36. San Millan, A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical
639 Context. *Trends in Microbiology* **26**, 978–985 (2018).
640 37. Hanke, D. M., Wang, Y. & Dagan, T. Pseudogenes in plasmid genomes reveal past
641 transitions in plasmid mobility. *Nucleic Acids Research* **52**, 7049–7062 (2024).

642 38. Kadibalban, A. S., Landan, G. & Dagan, T. The extent and characteristics of DNA
643 transfer between plasmids and chromosomes. *Current Biology* **34**, 3189-3200.e5
644 (2024).
645 39. Coluzzi, C. & Rocha, E. P. C. The Spread of Antibiotic Resistance Is Driven by
646 Plasmids Among the Fastest Evolving and of Broadest Host Range. *Molecular*
647 *Biology and Evolution* **42**, msaf060 (2025).
648 40. Pedrós-Alió, C. Marine microbial diversity: can it be determined? *Trends in*
649 *Microbiology* **14**, 257–263 (2006).
650 41. Ares-Arroyo, M., Nucci, A. & Rocha, E. P. C. Expanding the diversity of origin of
651 transfer-containing sequences in mobilizable plasmids. *Nat Microbiol* **9**, 3240–3253
652 (2024).
653 42. Coluzzi, C., Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. Evolution of
654 Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids.
655 *Molecular Biology and Evolution* **39**, msac115 (2022).
656 43. Catchpole, R. J. *et al.* A self-transmissible plasmid from a hyperthermophile that
657 facilitates genetic modification of diverse Archaea. *Nat Microbiol* **8**, 1339–1347
658 (2023).
659 44. Dodsworth, J. A. *et al.* Interdomain Conjugal Transfer of DNA from Bacteria to
660 Archaea. *Applied and Environmental Microbiology* **76**, 5644–5647 (2010).
661 45. Domingues, C. P. F., Rebelo, J. S., Dionisio, F. & Nogueira, T. Plasmids Across
662 Datasets: Resistance, Virulence, Mobility, and Host Taxonomy. 2025.01.30.635773
663 Preprint at https://doi.org/10.1101/2025.01.30.635773 (2025).

664 46. Yachi, S. & Loreau, M. Biodiversity and ecosystem productivity in a fluctuating
665 environment: The insurance hypothesis. *Proceedings of the National Academy of*
666 *Sciences* **96**, 1463–1468 (1999).
667 47. Pascoal, F., Magalhães, C. & Costa, R. The Link Between the Ecology of the
668 Prokaryotic Rare Biosphere and Its Biotechnological Potential. *Front. Microbiol.* **11**,
669 (2020).
670 48. Coveley, S., Elshahed, M. S. & Youssef, N. H. Response of the rare biosphere to
671 environmental stressors in a highly diverse ecosystem (Zodletone spring, OK, USA).
672 *PeerJ* **3**, e1182 (2015).
673 49. Wang, Y. *et al.* Quantifying the Importance of the Rare Biosphere for Microbial
674 Community Response to Organic Pollutants in a Freshwater Ecosystem. *Applied and*
675 *Environmental Microbiology* **83**, e03321-16 (2017).
676 50. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**,
677 2068–2069 (2014).
678 51. Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database
679 by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection
680 of Isolates. *Antimicrobial Agents and Chemotherapy* **63**, 10.1128/aac.00483-19
681 (2019).
682 52. Camargo, A. P. *et al.* IMG/PR: a database of plasmids from genomes and
683 metagenomes with rich annotations and metadata. *Nucleic Acids Research* **52**,
684 D164–D173 (2024).
685 53. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing
686 well-connected communities. *Sci Rep* **9**, 5233 (2019).

54. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research* **50**, D20–D26 (2022).

55. Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community Journal* **3**, (2023).

56. Liu, G. *et al.* oriTDB: a database of the origin-of-transfer regions of bacterial mobile genetic elements. *Nucleic Acids Research* **53**, D163–D168 (2025).

57. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

Methods

A total of 54,264 complete plasmids were retrieved from the NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/), last accessed in February 2025 [10]. To exclude potential secondary chromosomes, we removed 1,355 plasmids larger than 500 kb, resulting in a final dataset of 52,909 plasmids for our analysis.

We identified the putative coding sequences (CDS) using Prokka v1.14.6 [50]. We then clustered the CDS using CD-HIT v4.8.1, with a 90% identity threshold.
ARGs were identified using AMRFinderPlus v4.0.19 [51] (database version 2025-03-25), employing default parameters: sequence identity > 90%, e-value < 1e−20, and minimum coverage > 50%. VGs were detected using ABRicate v1.0.1 with minimum identity of 90% against the Virulence Factors Database (VFDB).
To cluster the plasmid sequences from the RefSeq database into Plasmid Taxonomic Units (PTUs) [12], we used the method described by Camargo *et al*. [52]. This approach, similar to the one defined by Redondo-Salvo *et al*. 12, allows for the *de novo* grouping of plasmids based on average nucleotide identity (ANI) and aligned fraction (AF). As a result, all plasmids are assigned to a PTU, including those not previously associated with existing PTUs.
Sequence alignments were performed using blastn (parameters: -outfmt '6 std qlen slen' -max_target_seqs 20000 -task megablast -evalue 1e-5). Afterwards, ANI and AF values were calculated using a Python script available at https://bitbucket.org/berkeleylab/checkv/src/master/scripts/anicalc.py. These values were used to compute the edges (*weight* = *AF* × *ANI*), which were subsequently applied to construct the Leiden graph [53] using the pyLeiden script available at https://**github.com/**apcamargo/pyleiden (parameters: '-n 5 - r 1.2').
Taxonomic information was retrieved from the NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy), which provides curated taxonomic classifications maintained by the National Center for Biotechnology Information (NCBI) [54].

To detect conjugative systems, we used MacSyFinder (v2.1.3) [55] with the parameters "--models CONJScan/Plasmids all --db-type ordered_replicon". If a conjugative element (T4SS) was detected, the plasmid was classified as conjugative (pCONJ). If no

732 conjugative element was found but a mobilisable element (MOB) was detected, the
733 plasmid was classified as mobilisable (pMOB). Plasmids lacking both T4SS and MOB
734 elements were further screened for an origin of transfer (*ori*T). Those containing an *ori*T
735 were classified as pOriT, while the remaining plasmids were considered non-
736 transmissible (pNT). To detect origins of transfer, we downloaded a database of 91
737 experimentally validated *ori*T sequences [56]. A *blastn* (v2.14.0+) search was performed
738 against this database using the parameters "-task blastn-short -evalue 0.01" [57]. We
739 filtered the results with identity and coverage higher than 80%.
740
741
742 **Statistics**
743 The following statistical analyses were conducted in R v4.4.2 [5]. The binomial test was
744 used to calculate the percentage of each characteristic in the respective biosphere.
745 Pearson's chi-squared test, a non-parametric method, was applied to assess the
746 dependence between two categories, with effect size measured using Cramer's V
747 (rcompanion package v2.4.36). Differences in mean values between two groups were
748 tested using a t-test with unequal variance (Welch's t-test), and effect size was
749 calculated using Cohen's d (effectsize package v1.0.0). To compare the prevalence of
750 ARGs and VGs in each biosphere, we used odds ratio (epitools package v0.5.10.1). We
751 used the nonparametric Mann-Whitney U test, also known as the Wilcoxon rank-sum
752 test, to compare the means of two samples, and the effect size was calculated using the
753 Glass rank biserial correlation coefficient (r) (effectsize package v1.0.0). Visualizations
754 were performed with the ggplot2 package v3.5.1.
755
756
757
758
759
760
761
762
763