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 prokaryotes—remains poorly understood. Using a dataset of 52,909 plasmids from isolates of bacteria and archaea [10], we found that their 16,547 Plasmid Taxonomic Units (PTUs) [11,12] exhibit a distribution with a fat 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, consistent with recent findings for species abundance distributions across several taxonomic groups [13]. The host distribution also presents a heavy 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 taxa levels, from species to phyla. Yet, plasmids from common PTUs are more mobile, likely explaining their success. The large group of rare PTUs constitutes a much more diverse reservoir of genetic material 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. Therefore, this work opens new paths into plasmid research.

Keywords: Rare biosphere; Plasmid, Antimicrobial resistance, Virulence genes

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. 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 a source of bacterial innovation

and may contain 'hidden' accessory genes, such as those involved in virulence and antibiotic resistance.

To study the distribution of PTUs, we analysed the RefSeq database containing 52,909 complete plasmids from bacterial and archaeal isolates[10], spanning 16,547 PTUs.

Rank-abundance and relative abundance distributions of PTUs

The rank-abundance distribution of the 16,547 PTUs is fat-tailed, revealing a pronounced disparity: a few PTUs are represented by many plasmids, and most PTUs (72%) are represented by only a single plasmid (**Fig. 1a**). In a log-log plot, the distribution of plasmid counts per PTU follows a straight line, suggesting a power-law with a slope close to -0.74 (or -0.91 if we exclude PTUs with a single plasmid) (**Fig. 1b**).

Notably, the most abundant PTU contains 650 plasmids, whereas 11,897 PTUs are singletons (i.e., represented by just a single plasmid). How should these numbers be interpreted? If a given PTU has a single representative in the dataset, it means that, among the 52,909 plasmids, no other plasmid in the dataset is similar enough to be included in the same taxonomic unit. We do not know, however, whether this plasmid is very frequent or rare in the habitat (e.g, the gut of an animal) where it was found. We can envisage that this plasmid and the corresponding PTU could be common in a particular habitat, yet no other plasmid of the same taxonomic unit was sampled in the other thousands of habitats [22].

We also analyse relative abundance distributions, commonly referred to in ecological studies as "species abundance distributions" (SADs), and by analogy, we will refer to them as "PTU abundance distributions" (PADs). These distributions represent, on the x-axis, abundance measured by number of plasmids per PTU, typically shown on a base-2 logarithmic scale (octaves: one plasmid, two or three plasmids, four to seven plasmids, eight to 15 plasmids, *et seq.*), and on the y-axis, the number of PTUs with a given abundance (number of PTUs with one plasmid, with two or three plasmids, and so on). There is an extensive body of ecological literature on fitting SADs [23,24]. Following recent work [13], we fit the PAD using two statistical distributions: the negative binomial (also called Poisson gamma) and the Poisson log-normal distributions, both via maximum likelihood methods [25]. As can be visually assessed (**Fig. 1c**), the Poisson log-normal distribution provides an excellent fit, in contrast to the negative binomial (see **Supplementary Table S1**, which includes the Akaike information criterion (AIC) values), and **Supplementary Table S2** for the estimated parameters of the distributions).

We also plot the PAD in double-logarithmic scales, which further highlights the superior fit of the Poisson log-normal compared to the alternative distribution **(Fig. 1d**).

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

Are PTUs rare or common because their hosts are also rare or common, respectively?

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

Of the 7,715 hosts, 896 have no defined species (they are identified as 'Genus sp.'). As we intend to perform analyses of the rarity of the host species, and since within a sp. category there may be several species for which we have no information on whether they are common or rare, we have excluded these genomes from the analyses. Therefore, for the host-related analyses, we removed 3,593 plasmids that were identified in hosts with undefined species.

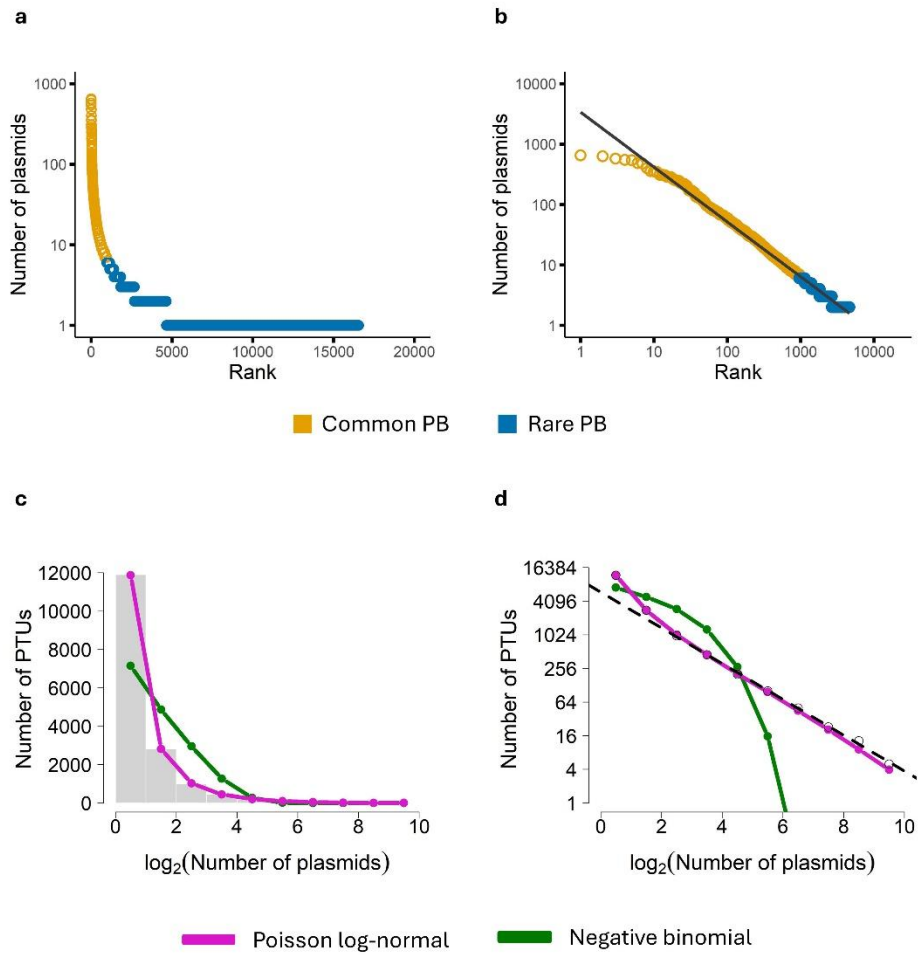

Fig. 1| Rank-abundance and relative abundance distributions of PTUs. a, the vertical axis (in $\log_{10}$ scale) represents the number of plasmids representing each PTU. The horizontal axis shows the ranking of PTUs, from the most common, with 650 plasmids, to the rarer, with a single plasmid. Yellow: 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).

We used two approaches to address the rank-abundance distribution of host species. In the first approach, we consider the number of plasmids sequenced in each of the prokaryotic species. According to this approach, the most frequent host species in the dataset is *E. coli*, harbouring as many as 9,343 plasmids. It is followed by *Klebsiella pneumoniae* (7,977 plasmids), *Salmonella enterica* (2,247 plasmids), *Acinetobacter baumannii* (1,204 plasmids), and 2,359 other prokaryotic species. In the second approach, we count the number of sequenced genomes with plasmids of that species present in the database. According to this approach, the most common species is also *E. coli* (3,068 genomes), followed by *K. pneumoniae* (2,328 genomes), *S. enterica* (1,210 genomes), *Staphylococcus aureus* (807 genomes), and 2,359 other prokaryotic species.

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

We next asked whether rare host species preferentially carry plasmids from rare PTUs and common host species carry plasmids from common PTUs. If plasmids were randomly distributed between hosts, 57% of them would inhabit a host whose biosphere category (rare or common) does not match their own category. Using the first classification approach, we identified 13,534 plasmids in rare hosts: 3,035 from the Common PB and 10,499 from the Rare PB. Of the 35,782 plasmids in common hosts, 26,887 belong to the Common PB and 8,895 to the Rare PB. Thus, 3,035 + 8,895 = 11,930 plasmids (24.2% of the total) do not match the biosphere category of their hosts. With the second approach, 13,529 plasmids were associated with rare hosts (3,218 from the Common PB; 10,311 from the Rare PB) and 35,787 with common hosts (26,704 from the Common PB; 9,083 from the Rare PB), yielding 3,218 + 9,083 = 12,301 mismatched plasmids, or 24.9%. Hence, in both approaches, roughly one-quarter of plasmids reside

168 in hosts of the opposite biosphere category, far fewer than the 57% expected under
169 random assignment, but well above a perfect correspondence (0%).
170
171

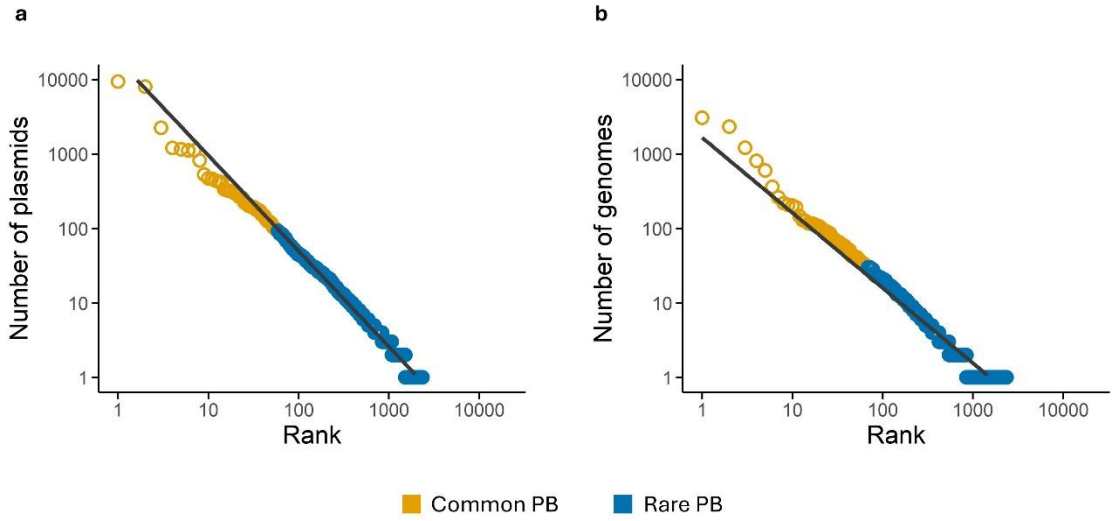

Fig. 2| – | Rank-abundance of host species. a, Rank of host species according to the number of plasmids found in the species (first approach – see main text). The first point on the left (rank = 1) represents *E. coli*, which has 9,343 plasmids. Both scales are in Log_{10}. The straight line represents 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, $R^2 = 0.94$.

172 The above assessment overlooks what happens within each species. We examined
173 the association between the total number of plasmids found in a species (first approach)
174 or between the total number of sequenced genomes of a host species (second approach)
175 and the percentage of plasmids of the Rare PB in that species. If plasmids from the Rare
176 PB are associated with rarer species, both correlations should be negative: most plasmids
177 found in rare species would belong to the Rare PB, while the percentage of plasmids from
178 the Rare PB would be low in more common species. The correlation is indeed negative,

179 but the association is moderate only, and the widely dispersed data points suggest that
180 other factors may play a more significant role in determining the relationship between
181 PTU and host species abundances (**Supplementary Fig. 1**).
182 Both approaches for defining rare and common host species yielded essentially
183 the same pattern – qualitatively identical and quantitatively very similar. Therefore, the
184 following sections present only the results based on the first approach.
185 Because prokaryote species may host plasmids both from the Common and Rare
186 PBs, we conducted a binomial test for each common species to assess whether the
187 proportion of common plasmids, (number of plasmids from the Common PB) / (total
188 number of plasmids), is significantly different from the proportion of common plasmids
189 in the sample, namely 30428/52909 = 0.58. Among the 58 common species, 18 host more
190 plasmids from the Rare PB than expected (binomial test, $p < 0.05$), and 11 common
191 species do not have more plasmids of one plasmid biosphere than expected (binomial test,
192 $p > 0.05$). The other 29 common species contain more plasmids from the Common PB
193 than expected (binomial test, $p < 0.05$). Yet, even among these 29 species, there are many
194 plasmids belonging to the Rare PB (**Supplementary Fig. 2**). We then performed a similar
195 analysis with the rare species (hence with 93 plasmids or fewer, following the first
196 approach). There are 2,305 species in these conditions: 21 contain more plasmids of the
197 Common PB than expected (binomial test, $p < 0.05$), and 1,686 do not contain more
198 plasmids of any biosphere than expected (binomial test, $p > 0.05$). In 598 rare species,
199 there are more plasmids from the Rare PB than expected (binomial test, $p < 0.05$), but
200 even in these species, there are plasmids of the Common PB. In the particular case of
201 PTUs with just one plasmid (9,692 singletons), 6,307 (approximately 2/3 of singletons)
202 were found in rare host species and 3,385 (~ 1/3) in common host species. Notably, every
203 host species classified as common harboured at least three plasmids assigned to the Rare
204 PB. On the other hand, 59 rare host species harbour plasmids only from the Common PB.
205 Overall, these analyses suggest that plasmids of common PTUs are not common
206 just because they were found in common host species, as we found so many plasmids of
207 the Rare PB in the common hosts and plasmids of the Common PB in rare host species.
208 These analyses indicate that plasmids from the Rare PB tend to be associated with rare
209 prokaryotic species, but this trend is only moderate, as so many plasmids of the Rare PB
210 are found in common species (**Supplementary Figs. 1-3**).
211

212 **The Rare Plasmid Biosphere is Richer, Has More Distinct Virulence Genes**
213 **and Fewer Antibiotic Resistance Genes**
214
215 Because each common PTU contains more plasmids than every rare PTU, PTUs
216 in the Common PB can be viewed as the more "successful" lineages. To test whether this
217 success reflects larger plasmid size, and hence a greater gene repertoire, we compared
218 plasmid lengths (in base pairs) between the Common and Rare PBs. The size of plasmids
219 does not differ between the two groups, indicating that the greater abundance of common
220 PTUs cannot be explained by longer genomes **(Supplementary Fig. 4a)**. Moreover, there
221 are fewer coding DNA sequences (CDS) clusters in the Rare PB than in the Common PB,
222 but the strength of the association is irrelevant, so the statistical significance is driven by
223 a large sample size **(Supplementary Fig. 4b)**. Therefore, data does not support the
224 hypothesis that plasmids from the Common PB are more abundant because they contain
225 more CDS clusters.
226 The two biospheres together account for 875,346 CDS clusters, but the Common
227 PB has only 76,491 CDS clusters (8.74%) while the Rare PB has 826,373 (94.41%), a
228 significant difference (the combined percentages exceed 100% because some CDS
229 clusters are present in both plasmid biospheres). Another way to see the biological
230 significance of this difference is that, while 48,973 CDS clusters are specific to the
231 Common PB (i.e., absent from the Rare PB), 798,855 CDS clusters are only found in the
232 Rare PB. The difference in the number of CDS clusters between the two plasmid
233 biospheres is significant. Moreover, although there are fewer plasmids in the Rare PB,
234 they have significantly more CDS clusters than the Common PB (**Fig. 3**).
235 Among the 4,814 plasmids containing virulence genes (VGs), there are only 546
236 distinct VGs, so some of them are found more than once across plasmids. The Common
237 PB contains 4,053 plasmids encoding VGs but only 281 distinct VGs. The Rare PB, on
238 the other hand, includes only 761 plasmids with VGs, but 423 distinct VGs. Therefore,
239 only 51.5% ($CI_{95\%}$: [47.2%, 55.7%]) of distinct VGs are in the Common PB, whereas
240 77.5% ($CI_{95\%}$: [73.7%, 80.9%]), are in the Rare PB, with an odds ratio of 0.31 ($CI_{95\%}$:
241 [0.24, 0.40]), indicating a significantly higher prevalence of these genes in the Rare PB.
242 Hence, the diversity of virulence genes is higher in the Rare PB than in the Common PB
243 (**Fig. 3**). The biological significance of this difference becomes particularly evident when
244 considering the following: of the 546 distinct VGs identified, the Rare PB includes 265
245 that are absent from the Common PB. As such, failing to detect the Rare PB would result

246 in overlooking 48.5% of the total VGs. In contrast, only 123 VGs (22.5%) found in the
247 Common PB are not present in the Rare PB.
248

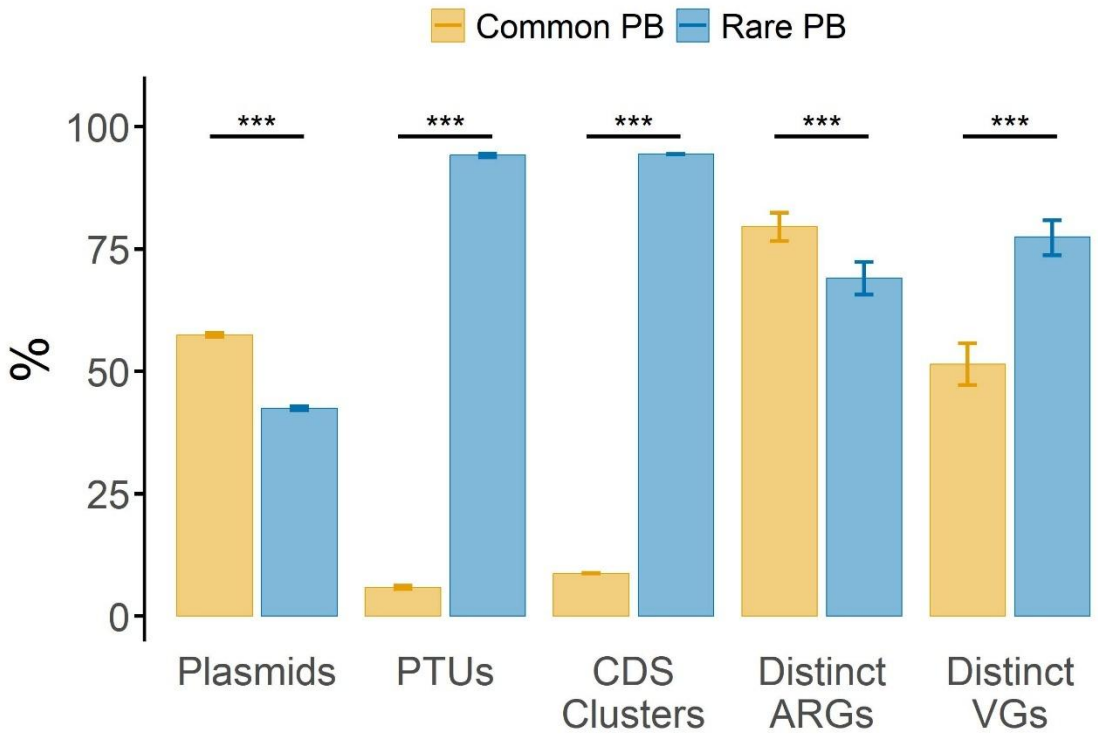

Fig. 3| Characteristics of the Common and Rare Plasmid Biospheres. The Rare PB includes fewer plasmids than the Common PB ($\chi^2(1) = 2387$, $p < 2.2\text{x}10^{-16}$, Cramér's V = 0.150, $CI_{95\%}$: [0.145, 0.157]). The number of PTUs is higher in the Rare PB than in the Common PB ($\chi^2(1) = 25849$, $p < 2.2\text{x}10^{-16}$, 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]). Bars on columns indicate the $CI_{95\%}$.

249
250 The Virulence Factor Database (VFDB) classifies virulence genes into 14
251 categories [26]. One category – post-translational modification – was absent from all

252 plasmids in the dataset. However, 12 categories were identified in both plasmid
253 biospheres (**Supplementary Fig. 5a**). No category was found exclusively in the Common
254 PB, but the stress survival category was found exclusively in the Rare PB, represented by
255 the *sodCI* gene. This gene, typically present in the chromosome of *Salmonella enterica*
256 serovar Typhimurium, was detected in two plasmids. The *sodCI* gene encodes the
257 superoxide dismutase SodCI, which enhances the virulence of these bacteria by protecting
258 them against macrophages[27].

259 Concerning antibiotic resistance genes (ARGs), we found 79.6% ($CI_{95\%}$: [76.6%,
260 82.4%]) of the 761 different ARGs in the Common PB, whereas 69.1% ($CI_{95\%}$: [65.7%,
261 72.4%]) of the ARGs were present in the Rare PB, with an odds ratio of 1.75 ($CI_{95\%}$:
262 [1.38, 2.21]), indicating that the odds of ARG presence in the Common PB were higher
263 than those in the Rare PB (**Fig. 3**). The dataset contains 14,718 plasmids with 761 distinct
264 ARGs, of which 371 are present in both plasmid biospheres, 155 are only present in the
265 Rare PB, and 235 in the Common PB. Therefore, although the Common PB is richer in
266 ARGs than the Rare PB, overlooking the Rare PB would mean ignoring the presence of
267 the 155 ARGs specific to this plasmid biosphere.

268 According to AMRFinderPlus, plasmids in the dataset contain genes conferring
269 resistance to 28 distinct antibiotic categories (**Supplementary Fig. 5b**) [28]. Of these, 27
270 are found in both plasmid biospheres, while one category—Nitroimidazole—is exclusive
271 to the Rare PB. This category includes the genes *nimA* and *nimD* found in plasmids from
272 *Bacteroides fragilis*, and *nimE* found in plasmids from *Bacteroides fragilis*, *Bacteroides*
273 sp. PHL 2737, and *Bacteroides xylanisolvens*.

274

275 **PTUs of the Rare PB Spans Through More Host Species, and, *per* Plasmid,**
276 **have Broader Host-Range**

277

278 Plasmids of the Rare PB represent only 42.5% of the 52,909 plasmids analysed in
279 this study. Yet, they span more taxa than the plasmids from the Common PB (**Table 1**).
280 For example, there are 27 hosts' phyla with plasmids belonging to the Rare PB, but just
281 11 hosts' phyla with plasmids of the Common PB. (**Supplementary Fig. 6**). Moreover,
282 there are no hosts' phyla harboring plasmids only of the Common PB, but there are 16
283 hosts' phyla with plasmids only of the Rare PB. This pattern repeats at all taxa levels until
284 the species level (**Table 1**). At this level, there are 2,304 host species with plasmids of
285 the Rare PB, but only 465 with plasmids of the Common PB. Also, 744 host species

harbour plasmids only of the Rare PB but only eight harbour plasmids only of the Common PB (**Table 1**; as explained before, we removed from the analysis at the species level the plasmids identified in genomes with unidentified species: 3,593 plasmids, 87 from archaea, and 3,506 from bacteria).

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

	Number of 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

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

Recently, Redondo-Salvo et al. classified [12] host-ranges of PTUs into a scale of six grades, ranging from I (restricted to a single host species) to VI (able to colonize different phyla). For example, the 32 host species harbouring the 637 plasmids (more 13 plasmids from five unidentified species) of the largest PTU belong to a single family (Enterobacteriaceae) and just 12 genera (*Citrobacter*, *Cronobacter*, *Enterobacter*, *Escherichia*, *Klebsiella*, *Kluyvera*, *Leclercia*, *Phytobacter*, *Pseudescherichia*, *Raoultella*, *Salmonella*, *Shigella*). Although this is the largest PTU, it is present in different genera within a single family. Therefore, it is of 'intermediate' host-range, i.e., grade III. In the Common PB, a PTU comprising 24 plasmids is found across three phyla: Campylobacterota, Bacillota (Firmicutes), and Actinomycetota. This is a case of a very

309 broad host-range (grade VI). However, the PTU found in more phyla is present in four
310 phyla (Bacillota, Verrucomicrobiota, Actinomycetota and Pseudomonadota) despite
311 belonging to the Rare PB. Therefore, even rare PTUs may be distributed among very
312 distinct taxonomic groups. This was unexpected because, by definition, a PTU belongs
313 to the Rare PB if it has up to six plasmids in the dataset. This specific PTU present in four
314 phyla has only four plasmids. On the other extreme, 325 Common PTUs are in a unique
315 host species, hence very narrow host-range plasmids (grade I) (**Supplementary Fig. 7**).
316 Next, we analyse and compare the host-range of Common and Rare PTUs in a
317 systematic way. It is likely that common PTUs are present in more taxa (e.g., in more
318 species) than rare PTUs simply because common PTUs are, according to our definition,
319 those having more representative plasmids in the dataset (therefore with more
320 opportunities to be found in more taxa). To rule out this confounding factor concerning
321 host-range, we calculated the number of hosts' species per number of plasmids of each
322 PTU. This ratio is significantly higher (Wilcoxon signed-rank, $p < 0.05$) for the PTUs of
323 the Rare PB than of the Common PB **(Fig. 4a).** Since thousands of PTUs contain only
324 one plasmid (singletons), we conducted a similar comparison excluding these PTUs from
325 the analysis. Again, the diversity of hosts' species per plasmid is higher in the Rare PB
326 than in the Common PB **(Fig. 4b)**. These ratios may be too low in the bigger PTUs. For
327 example, the hosts' species/plasmid ratio for a PTU with 30 plasmids present in 5 species
328 (ratio = 5/30) is lower than for a PTU with 5 plasmids in 5 species (ratio = 5/5). To avoid
329 these biases, we compared the success of PTUs with seven plasmids (belonging to the
330 Common PB) with those with six plasmids (belonging to the Rare PB). Again, for this
331 case, the PTUs of the Rare PB are present in more host species, on average, than the PTUs
332 of the Common PB **(Fig. 4c).** A similar analysis for genus, family, order, class and
333 phylum has shown that this ratio is consistently higher in the Rare PB than in the Common
334 PB (Wilcoxon signed-rank test $p < 0.05$, large effect sizes for all cases) (**Figs. 4d, 4e, and**
335 **4f,** and **Supplementary Fig. 8** show these analyses for the genus, family, order, and class
336 levels). Therefore, other factors must be responsible for the success of Common PTUs.
337

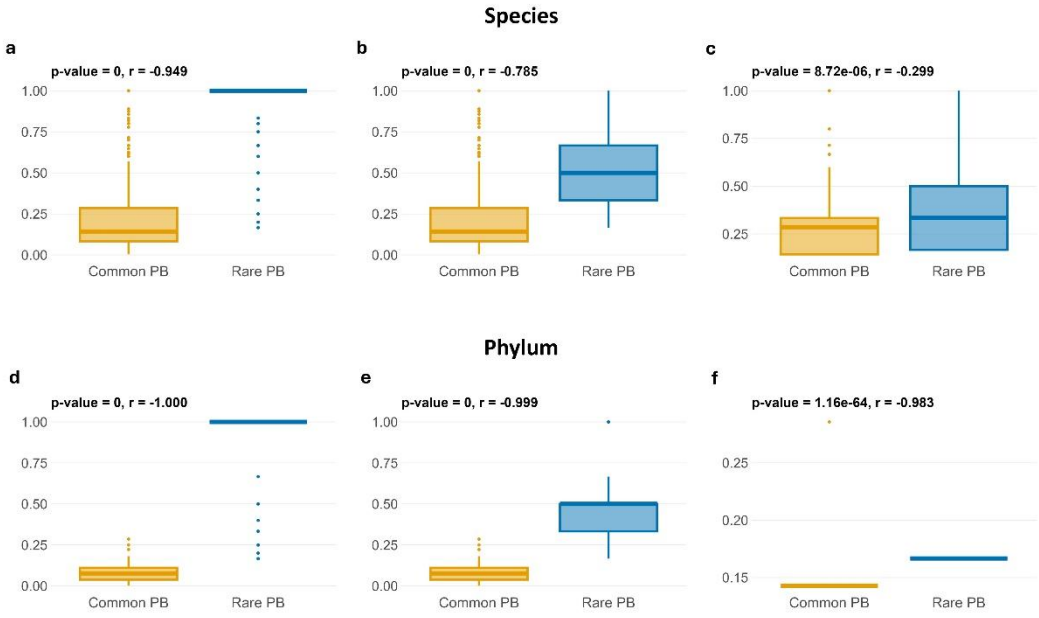

Fig. 4| The relative host-range of PTUs of the Rare and Common PB. The relative success of each PTU, measured as the number of hosts' taxa per number of plasmids in that PTU. From left to right: all plasmids, not including singletons, including only PTUs with seven (Common) or six (Rare) plasmids, respectively. **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).

338

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

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

349 division or hitchhiking mechanisms such as conduction (co-integration in a transferable
350 plasmid), transduction (with the help of a phage), or transformation [14,29–33]. The
351 proportions of each class in the Common and Rare PBs tend to be significantly different
352 from the expected values if they were randomly distributed through the PBs (Table 2).
353

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]

354 [1] Within either PB, the expected percentage composition mirrors the global proportions
355

356 The percentage of pCONJ is higher among Common PTUs than Rare PTUs.
357 Together, the proportion of pMOB and pOriT plasmids is also higher in the Common PB
358 (40.4 %, $CI_{95\%}$ = [39.9, 41.0]) than in the Rare PB (37.0%, $CI_{95\%}$ = [36.4, 37.6])
359 (**Supplementary Table 3**). In contrast, pNT plasmids are more frequent in the Rare PB
360 (44.4 %) than in the Common PB (23.6%). These results suggest that transfer ability plays
361 a fundamental role in the success of Common PTUs, although 55.6% of the plasmids of
362 the Rare PB are transferable, and 23.6% of the plasmids of the Common PB are non-
363 transmissible.
364

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

366 We have seen that several rare host species harbour plasmids from common PTUs.
367 For example, there are 59 rare host species with plasmids only from the Common PB.
368 How can we explain this observation? One possible explanation is that these hosts
369 received the plasmid from phylogenetically close common hosts. That is likely the case
370 of *Vibrio owensii*, a rare species with eight plasmids from three PTUs. *V. owensii* belongs
371 to the genus Vibrio, which includes one common species and 53 rare ones, and to the
372 family Vibrionaceae, which includes one common and 64 rare species. The common

species within this genus and family shares all three PTUs with *Vibrio owensii*. However, this case of *Vibrio owensii* is far from general.

We have seen that, at least in the case of the *Vibrio* genus, a common host species harbouring plasmids from common PTUs likely contributes to the presence of these plasmids in rare host species of the same genus. 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 the 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. As an example, we investigate the case of *Bacteroides cellulosilyticus*, a rare species, carrying three plasmids from two common PTUs: (i) one from a certain PTU and (ii) two plasmids from another PTU. In the first case, plasmids of the same PTU are also present in *Pseudomonas aeruginosa* and *Escherichia coli*, two common host species – both evolutionary distant from *Bacteroides cellulosilyticus* – sharing the same phylum, but not the same family, order, or class. This means that this PTU has a broad host-range (grade V [12]). Therefore, this plasmid of a common PTU found in the *Bacteroides cellulosilyticus*, a rare host species, may have been transferred from *Escherichia coli* or *Pseudomonas aeruginosa*, two common hosts. In the case (ii), the two plasmids belong to a PTU that contains 20 plasmids in the dataset, 13 from the genus Bacteroides, six from the genus Parabacteroides, and one from the genus Phocaeicola, always from rare species. Hence, these two common plasmids found in the rare species *Bacteroides cellulosilyticus* do not appear to have been transferred from a common species (neither evolutionary close nor distant from *Bacteroides cellulosilyticus*).

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 undoubtedly the fact that the

407 plasmids analysed in this study were not derived from a single ecosystem, so caution is
408 warranted when interpreting the broader implications of our findings [10]. Yet, several
409 findings concerning this dataset present striking parallels with important findings using
410 data from ecosystems regarding plants, animals or microorganisms [1,4,5,13].

411 Recognizing the existence of rare biospheres within bacterial communities has
412 been highly significant, as these bacterial species may constitute extensive reservoirs of
413 genomic innovation [3,5,21,22,22]. They have a great potential for dynamic adaptation because
414 individual cells from low-abundance bacterial taxa can proliferate rapidly through clonal
415 replication when environmental conditions shift to favour their growth [4,5]. If this ability
416 of clonal amplification is relevant for bacterial cells, it is even more so in the case of
417 plasmids. These genetic elements may hitchhike the clonal expansion of their host, but
418 they can also expand horizontally through bacterial communities, even between bacterial
419 taxa, if the conditions are appropriate [14,34,35]. Moreover, plasmids may even help other
420 plasmids or be helped by other plasmids to transfer to other cells [30,32,36–38]. These
421 plasmids' characteristics might annihilate the fat-tailed distributions frequently observed
422 in bacterial communities, but our findings show that that was not the case.

423 The plasmids of this study have very diverse origins. Nevertheless, we observed
424 the same striking distribution pattern as seen in microbial communities from a given
425 ecosystem or habitat: characterized by a few highly abundant PTUs and a fat-tail of rare
426 ones (**Figs. 1 a, b**). One possible explanation for this finding is that the PTUs have not
427 been sampled exhaustively, a phenomenon referred to as the "veil line" in biodiversity
428 studies. The veil line is an imaginary boundary that shifts progressively to the left of the
429 distribution as sampling effort increases [13]. As this shift occurs, a peak may emerge in
430 intermediate abundance classes, rather than in the singleton class, as observed in the
431 present data.

432 The fitting of the PTU Abundance Distribution (PAD) curve revealed that the
433 Poisson log-normal distribution provided an excellent fit (**Figs. 1 c, d**). This result aligns
434 with recent findings [13] and suggests a compelling similarity between plasmid abundance
435 patterns and well-established results from other biodiversity studies. This observation is
436 noteworthy, as the shape of the PAD may offer insight into the underlying processes of
437 plasmid diversification, as well as the emergence and persistence of rare plasmids. If the
438 PADs are genuinely monotonically decreasing, this would be a noteworthy result, as the
439 PAD shape could provide valuable insight into the underlying processes of plasmid

440 diversification, as well as the emergence and persistence of rare plasmids. For instance,
441 under Hubbell's Neutral Theory of Biodiversity and Biogeography [24], a monotonically
442 decreasing function, such as the ones we observed, implies that new species arise with
443 very few individuals. The PAD in our dataset peaks at singletons, likely reflecting a
444 process of "speciation" in which novel PTUs emerge at initially very low abundance. This
445 interpretation is consistent with the well-known ability of plasmids to change [12,39–42].
446 These questions should be addressed in future research.

447 An important biological implication of the power-law is that some PTUs are
448 highly successful (hence belonging to the Common PB). What is the cause of their
449 success? If PTU's success was attributed to that of their hosts, one should expect that all
450 plasmids belonging to common PTUs should be hosted by common species and all
451 plasmids belonging to rare PTUs should be hosted by rare species (0% mismatches). In
452 contrast, if PTUs were randomly assigned as rare or common, 57% plasmids would not
453 match the assignment of their hosts. We observed an intermediate number: 25% of the
454 plasmids belong to a rare or a common species, despite belonging to a common or rare
455 PTU, respectively. We conclude that some plasmids and the corresponding hosts share
456 an evolutionary success. Yet, they are far from the majority. For example, the dataset
457 includes 9,343 plasmids from *E. coli*—one of the most extensively studied bacterial
458 species. Despite being the most common host, 1,649 of these plasmids (18%) belong to
459 rare PTUs, and thus to the Rare PB. *Lactiplantibacillus plantarum*, the eighth most
460 common host species, shows an even greater proportion: 77.4% of its plasmids (632 out
461 of 816) belong to the Rare PB, with only 184 in the Common PB, or *Levilactobacillus*
462 *brevis*, a common species with 104 of its 112 (92.86%) plasmids in the Rare PB. These
463 examples also suggest that the success of plasmids from common PTUs is not necessarily
464 driven by the success of their common hosts, nor are these hosts responsible for the
465 success of their plasmids from the common PTUs. The same applies to plasmids from
466 rare PTUs.

467 Intriguingly, the long list of rare PTUs is associated with exceptional diversity in
468 both coding sequence clusters and virulence genes (**Fig. 3**). Therefore, gene diversity is
469 not responsible for the success of the common PTUs. In contrast, the diversity of
470 antibiotic resistance genes (ARGs) is higher among the plasmids from the common PTUs
471 than in the rare ones. One possible explanation for such a contrast is that the sequencing
472 of RefSeq genomes may have been motivated by antibiotic resistance. Yet, even if
473 antibiotic resistance genes are less diverse in the Rare PB, 155 out of 761 genes (20.4%)

474 are only present in the Rare PB. For example, resistance genes of the Nitroimidazole
475 antibiotic class were found only in the Rare PB (**Supplementary Fig. 5**). Let us suppose
476 that future studies confirm that most antibiotic resistance genes are indeed more often
477 present in the Common PB. That might suggest that resistance genes in the Rare PB are
478 not as relevant as, e.g., virulence genes. Yet this interpretation should be approached with
479 caution. If new antibiotics are developed — particularly those with novel modes of action
480 — potential resistance genes are more likely to emerge from CDS that have not yet been
481 identified as ARGs, rather than from those already known. Our results, which reveal
482 exceptionally high CDS cluster diversity among plasmids in the Rare PB, suggest that
483 this group represents a vast gene pool from which novel resistance genes may evolve. As
484 such, these findings serve as a warning: resistance to future antibiotics may arise from
485 currently uncharacterized gene sequences within the Rare PB.

486 A recent study has shown that ARGs are preferentially acquired and carried by
487 broad-host-range plasmids and that these plasmids are more 'plastic' (e.g., higher
488 recombination rates or higher mobility) [43]. We have found that one PTU belonging to the
489 Rare PB is present in hosts from as many as four different phyla (**Supplementary Fig.**
490 **7**). Future studies should look for recombinases in these plasmids and study rates of allelic
491 exchanges and changes of gene repertoires, as well as their mobility [43]. This PTU is a
492 peculiar case of an important characteristic of rare PTUs concerning host ranges: per
493 number of plasmids in each PTU, the host range of rare PTUs is higher than common
494 PTUs. Strikingly, we observed the same pattern at all taxonomic levels, from species to
495 phylum (**Fig. 4** and **Supplementary Fig. 8**). One could argue that the overall broader host
496 range of rare PTUs is irrelevant because these PTUs have very few plasmids. However,
497 these PTUs are very numerous, representing 94% of the PTUs and comprising almost half
498 of the plasmids in the dataset. Thus, these PTUs likely play critical roles in bacterial
499 evolution when hosted by different prokaryotic species/genera/…/phyla, eventually co-
500 inhabiting the same cell with other plasmids, from both plasmid biospheres.

501 In bacterial communities, it has been proposed that rare taxa tend to remain rare
502 because they are less visible to predators, such as viruses or protists [4,44]. Yet, our results
503 strongly suggest that this explanation does not apply to plasmids. If it did, rare host
504 species should exclusively harbour plasmids from the Rare PB, while common hosts
505 should lack plasmids from this pool; this scenario is not supported by our data.

506 We have seen that the very good fit of a Poisson log-normal distribution to the
507 PTU abundance distribution suggests that many new rare PTUs likely arise from genetic

508 changes of the already existing pool of plasmids (from common or rare PTUs). Therefore,
509 a potential mechanism underlying the emergence and/or persistence of plasmids in the
510 Rare PB could be the acquisition of new genes, also likely to transform plasmids from a
511 common PTU into a rare PTU. Nevertheless, this mechanism seems unlikely because
512 plasmids belonging to Rare and Common PBs do not differ significantly in size (base-
513 pairs), and plasmids in the Common PB contain only a slightly higher number of CDS
514 (**Supplementary Fig. 4**).

515 Finally, plasmid mobility may explain the success of common PTUs. About
516 76.4% of the plasmids in the Common PB were transferable (pCONJ, pMOB, and pOriT
517 plasmids), in contrast with the 55.7% transferable plasmids from the Rare PB.
518 Consequently, the proportion of non-transmissible plasmids (pNT), is almost the double
519 (44.4%) in the Rare PB than in the Common PB (23.6%) (**Table 2**). It is intriguing,
520 however, that so many pNT plasmids are rare. There are some possible explanations, for
521 example: (i) they may be misclassified as pNT, but future research may find *OriT*'s
522 sequences on them [14,45]; (ii) they belong to a PTU that also includes transferable plasmids;
523 (iii) they were recently pOriT, pMOB or pCONJ, but they lost relevant transfer genes
524 [41,46]; or they easily transfer through alternative mechanisms, such as by transformation,
525 or by integrating into the genomes of phages (transduction), or other plasmids
526 (conduction) [31,33].

527 The dataset used in this study contained 20,580 genomes, 218 of archaea and
528 20,362 of bacteria, harbouring plasmids from 16,547 PTUs. Yet, none of these PTUs is
529 present both in archaea and bacteria – even if one of the PTUs of archaea belong to the
530 Common PB. This suggests that inter-domain plasmid transfers are rare events (possibly
531 because their habitats and selective pressure do not coincide) [47,48].

532 Our study has important limitations. The first limitation was already mentioned
533 throughout this study: plasmids (and their hosts, bacteria and archaea) studied here do not
534 belong to a single ecosystem/ habitat. For example, *E. coli* is certainly not the most
535 abundant bacterial species. Future studies should focus on plasmids from a single habitat.
536 That would imply sequencing all plasmids present in a single microbiome. A second
537 limitation concerns the dataset used in this study. The dataset is likely biased towards
538 clinical bacteria. Different datasets have different characteristics, e.g. the number of
539 plasmids carrying antibiotic resistance genes [49]. A third limitation is the very low number
540 of plasmids from archaea. Future studies should study their plasmids in this context. This
541 limitation is related to the second one, as bacterial species have been of greater interest

from the clinical or industrial viewpoints. Yet, 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 [1,50]. 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? (4) We have seen that plasmids of common PTUs have increased transfer ability; could other factors, such as encoding bacteriocins [51], be responsible for the success of the common PTUs?

Our finding opens new avenues in the quantitative biology of plasmids, in parallel with recent discoveries of universal laws relating plasmid size to protein-coding genes and copy-number [52,53]. But this work has at least two main practical implications. First, researchers can carefully look 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 [54–56].

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.

Author contributions

Conceptualization, J.S.R., C.P.F.D., L.B.A., and F.D; Software, J.S.R., C.P.F.D. and L.B.A; Methodology, J.S.R., C.P.F.D., L.B.A., T.N., and F.D.; Investigation, J.S.R. and C.P.F.D.; Formal Analysis, J.S.R., C.P.F.D., L.B.A., T.N. and F.D.; Writing – Original Draft, 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; Supervision, T.N. and F.D.

References

1. Jousset, A. *et al.* Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J* **11**, 853–862 (2017).
2. Litchman, E. *et al.* Refocusing the microbial rare biosphere concept through a functional lens. *Trends in Ecology & Evolution* **39**, 923–936 (2024).
3. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* **13**, 217–229 (2015).
4. Pedrós-Alió, C. The Rare Bacterial Biosphere. *Annual Review of Marine Science* **4**, 449–466 (2012).
5. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12115–12120 (2006).
6. Jørgensen, B. B., Dunker, R., Grünke, S. & Røy, H. Filamentous sulfur bacteria, Beggiatoa spp., in arctic marine sediments (Svalbard, 79 degrees N). *FEMS Microbiol Ecol* **73**, 500–513 (2010).
7. Mallon, C. A. *et al.* Resource pulses can alleviate the biodiversity-invasion relationship in soil microbial communities. *Ecology* **96**, 915–926 (2015).
8. van der Gast, C. J. *et al.* Partitioning core and satellite taxa from within cystic 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. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Res* **50**, D912–D917 (2022).
27. 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).
28. Feldgarden, M. *et al.* Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection

646 of Isolates. *Antimicrobial Agents and Chemotherapy* **63**, 10.1128/aac.00483-19
647 (2019).
648 29. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & De La Cruz, F.
649 Mobility of Plasmids. *Microbiol Mol Biol Rev* **74**, 434–452 (2010).
650 30. Ares-Arroyo, M., Coluzzi, C., Sousa, J. A. M. de & Rocha, E. P. C. Hijackers,
651 hitchhikers, or co-drivers? The mysteries of mobilizable genetic elements. *PLOS*
652 *Biology* **22**, e3002796 (2024).
653 31. Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. The extended mobility of
654 plasmids. *Nucleic Acids Res* **53**, gkaf652 (2025).
655 32. Gama, J. A., Zilhão, R. & Dionisio, F. Co-resident plasmids travel together. *Plasmid*
656 **93**, 24–29 (2017).
657 33. Dionisio, F., Zilhão, R. & Gama, J. A. Interactions between plasmids and other
658 mobile genetic elements affect their transmission and persistence. *Plasmid* **102**,
659 29–36 (2019).
660 34. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & de la Cruz, F.
661 Mobility of Plasmids. *Microbiol Mol Biol R* **74**, 434-+ (2010).
662 35. Rodríguez-Beltrán, J., DelaFuente, J., León-Sampedro, R., MacLean, R. C. & San
663 Millán, Á. Beyond horizontal gene transfer: the role of plasmids in bacterial
664 evolution. *Nat Rev Microbiol* **19**, 347–359 (2021).
665 36. Gama, J. A., Zilhao, R. & Dionisio, F. Multiple plasmid interference - Pledging
666 allegiance to my enemy's enemy. *Plasmid* **93**, 17–23 (2017).
667 37. Gama, J. A., Zilhão, R. & Dionisio, F. Conjugation efficiency depends on intra and
668 intercellular interactions between distinct plasmids: Plasmids promote the

immigration of other plasmids but repress co-colonizing plasmids. *Plasmid* **93**, 6–16 (2017).

38. Gama, J. A., Zilhão, R. & Dionisio, F. Plasmid Interactions Can Improve Plasmid Persistence in Bacterial Populations. *Front. Microbiol.* **11**, 2033 (2020).
39. Lang, A. S., Buchan, A. & Burrus, V. Interactions and evolutionary relationships among bacterial mobile genetic elements. *Nat Rev Microbiol* **23**, 423–438 (2025).
40. San Millan, A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends in Microbiology* **26**, 978–985 (2018).
41. Hanke, D. M., Wang, Y. & Dagan, T. Pseudogenes in plasmid genomes reveal past transitions in plasmid mobility. *Nucleic Acids Research* **52**, 7049–7062 (2024).
42. Kadibalban, A. S., Landan, G. & Dagan, T. The extent and characteristics of DNA transfer between plasmids and chromosomes. *Current Biology* **34**, 3189-3200.e5 (2024).
43. Coluzzi, C. & Rocha, E. P. C. The Spread of Antibiotic Resistance Is Driven by Plasmids Among the Fastest Evolving and of Broadest Host Range. *Molecular Biology and Evolution* **42**, msaf060 (2025).
44. Pedrós-Alió, C. Marine microbial diversity: can it be determined? *Trends in Microbiology* **14**, 257–263 (2006).
45. Ares-Arroyo, M., Nucci, A. & Rocha, E. P. C. Expanding the diversity of origin of transfer-containing sequences in mobilizable plasmids. *Nat Microbiol* **9**, 3240–3253 (2024).
46. Coluzzi, C., Garcillán-Barcia, M. P., de la Cruz, F. & Rocha, E. P. C. Evolution of Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids. *Molecular Biology and Evolution* **39**, msac115 (2022).

47. Catchpole, R. J. *et al.* A self-transmissible plasmid from a hyperthermophile that facilitates genetic modification of diverse Archaea. *Nat Microbiol* **8**, 1339–1347 (2023).

48. Dodsworth, J. A. *et al.* Interdomain Conjugal Transfer of DNA from Bacteria to Archaea. *Applied and Environmental Microbiology* **76**, 5644–5647 (2010).

49. Domingues, C. P. F., Rebelo, J. S., Dionisio, F. & Nogueira, T. Plasmids Across Datasets: Resistance, Virulence, Mobility, and Host Taxonomy. 2025.01.30.635773 Preprint at https://doi.org/10.1101/2025.01.30.635773 (2025).

50. Yachi, S. & Loreau, M. Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *Proceedings of the National Academy of Sciences* **96**, 1463–1468 (1999).

51. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in Escherichia coli. *Nat Commun* **16**, 2921 (2025).

52. Ramiro-Martínez, P., de Quinto, I., Lanza, V. F., Gama, J. A. & Rodríguez-Beltrán, J. Universal rules govern plasmid copy number. *Nat Commun* **16**, 6022 (2025).

53. Maddamsetti, R. *et al.* Scaling laws of bacterial and archaeal plasmids. *Nat Commun* **16**, 6023 (2025).

54. Pascoal, F., Magalhães, C. & Costa, R. The Link Between the Ecology of the Prokaryotic Rare Biosphere and Its Biotechnological Potential. *Front. Microbiol.* **11**, (2020).

55. Coveley, S., Elshahed, M. S. & Youssef, N. H. Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (Zodletone spring, OK, USA). *PeerJ* **3**, e1182 (2015).

716 56. Wang, Y. *et al.* Quantifying the Importance of the Rare Biosphere for Microbial
717 Community Response to Organic Pollutants in a Freshwater Ecosystem. *Applied and*
718 *Environmental Microbiology* **83**, e03321-16 (2017).
719 57. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**,
720 2068–2069 (2014).
721 58. Camargo, A. P. *et al.* IMG/PR: a database of plasmids from genomes and
722 metagenomes with rich annotations and metadata. *Nucleic Acids Research* **52**,
723 D164–D173 (2024).
724 59. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing
725 well-connected communities. *Sci Rep* **9**, 5233 (2019).
726 60. Sayers, E. W. *et al.* Database resources of the national center for biotechnology
727 information. *Nucleic Acids Research* **50**, D20–D26 (2022).
728 61. Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify
729 molecular systems in genomes. *Peer Community Journal* **3**, (2023).
730 62. Liu, G. *et al.* oriTDB: a database of the origin-of-transfer regions of bacterial mobile
731 genetic elements. *Nucleic Acids Research* **53**, D163–D168 (2025).
732 63. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**,
733 421 (2009).

734
735
736
737

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 [57]. We then clustered the CDS using CD-HIT v4.8.1, with a 90% identity threshold. ARGs were identified using AMRFinderPlus v4.0.19 [28] (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 a minimum identity of 90% against the Virulence Factor Database (VFDB) [26].

To cluster the plasmid sequences from the RefSeq database into Plasmid Taxonomic Units (PTUs) [12], we used the method described by Camargo *et al.* [58]. 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 [59] 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) [60].

To detect conjugative systems, we used MacSyFinder (v2.1.3) [61] 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 conjugative element was found but a mobilisable element (MOB) was detected, the plasmid was classified as mobilisable (pMOB). Plasmids lacking both

T4SS and MOB elements were further screened for an origin of transfer (*ori*T). Those containing an *ori*T were classified as pOriT, while the remaining plasmids were considered non-transmissible (pNT). To detect origins of transfer, we downloaded a database of 91 experimentally validated *ori*T sequences [62]. A *blastn* (v2.14.0+) search was performed against this database using the parameters "-task blastn-short -evalue 0.01" [63]. We filtered the results with identity and coverage higher than 80%.

Statistics

The following statistical analyses were conducted in R v4.4.2 [5]. The binomial test was used to calculate the percentage of each characteristic in the respective biosphere. Pearson's chi-squared test, a non-parametric method, was applied to assess the dependence between two categories, with effect size measured using Cramer's V (rcompanion package v2.4.36). Differences in mean values between two groups were tested using a t-test with unequal variance (Welch's t-test), and effect size was calculated using Cohen's d (effectsize package v1.0.0). To compare the prevalence of ARGs and VGs in each biosphere, we used odds ratio (epitools package v0.5.10.1). We used the nonparametric Mann-Whitney U test, also known as the Wilcoxon rank-sum test, to compare the means of two samples, and the effect size was calculated using the Glass rank biserial correlation coefficient (r) (effectsize package v1.0.0). Visualizations were performed with the ggplot2 package v3.5.1.