Timescales and genetic linkage explain the variable impact of defense systems on horizontal gene transfer

Yang Liu[1], João Botelho[1], Jaime Iranzo[1,2,3]

[1] Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Madrid, Spain.

[2] Centro de Astrobiología (CAB), CSIC-INTA, Madrid, Spain.

[3] Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain.

Competing interests: The authors declare no competing financial interests.

Abstract

Prokaryotes have evolved a wide repertoire of defense systems to prevent invasion by mobile genetic elements (MGE). However, because MGE are vehicles for the exchange of beneficial accessory genes, defense systems could consequently impede rapid adaptation in microbial populations. Here, we study how defense systems impact horizontal gene transfer (HGT) in the short and long terms. By combining comparative genomics and phylogeny-aware statistical methods, we quantified the association between the presence of 7 widespread defense systems and the abundance of MGE in the genomes of 196 bacterial and 1 archaeal species. We also calculated the differences in the rates of gene gain and loss between lineages that possess and lack each defense system. Our results show that the impact of defense systems on HGT is highly species- and system-dependent. CRISPR-Cas stands out as the defense system that most often associates with a decrease in the number of MGE and reduced gene acquisition. Timescale analysis reveals that defense systems must persist in a lineage for a relatively long time in order exert an appreciable

negative impact on HGT. In contrast, at short evolutionary times, defense systems, MGE, and gene gain rates tend to be positively correlated. Based on these results and given the high turnover rates experienced by defense systems, we propose that the inhibitory effect of most defense systems on HGT is masked by recent co-transfer events involving MGE.

Keywords: defense system, horizontal gene transfer, mobile genetic element, CRISPR-Cas, phage, plasmid, comparative genomics.

Introduction

Gene exchange plays a key role in the adaption of microbes to changing environments, facilitating the spread of antibiotic resistance, pathogenicity factors, metabolic genes, and other accessory functions (1). Over the last decade, there has been an increasing interest in assessing the ecological and genetic factors that control horizontal gene transfer (HGT) and determine the outcome of newly acquired genes in microbial populations (2-4). HGT is often mediated by mobile genetic elements (MGE), such as phages, integrative and conjugative elements, and plasmids, against which bacteria have evolved an elaborate repertoire of defense systems (5-9). As a result, large-scale patterns of HGT are shaped by an interplay of ecological and genetic variables that underlie cross-strain and cross-species differences in susceptibility to MGE (10).

Recent studies have highlighted the role of CRISPR-Cas as widespread adaptive immunity systems that protect archaea and bacteria against viruses and other MGE (11, 12). While in vitro experiments have provided supportive evidence for this function (13-15), questions remain about the extent to which CRISPR-Cas systems constrain gene exchange in nature (16-21). More generally, there is a paucity of research on how other defense systems, such as restriction-modification (RM), abortive infection (Abi), and an expansive repertoire of recently identified gene systems (e.g., DRT, BASS, and Gabija) affect HGT in prokaryotes.

At odds with the naïve expectation that defense systems should reduce the rates of HGT by interfering with the propagation of MGE, several empirical and theoretical observations suggest that the relation between defense systems and HGT might be more complex. First, the selection pressure to maintain defense systems in a population (and consequently their prevalence) generally increases with the exposure to MGE (22). Second, defense systems are mobilized by MGE, which could lead to a trivial positive association with HGT rates. In a less trivial manner, whole defense systems or parts of them are often encoded by MGE (23-25), turning the latter into addictive modules with beneficial side-effects in cellular defense (26, 27).

Here, we applied a phylogeny-aware statistical framework to 19,323 high-quality genomes from 197 species, revealing a species- and time-dependent pattern of positive and negative associations between 7 widespread defense systems (RM, CRISPR-Cas, CBASS, DMS, DRT, Gabija, and Abi) and HGT rates. Comparative data analysis singles out CRISPR-Cas as the system that most often presents a negative impact on MGE abundance and gene exchange. Our results shed light on the nuanced consequences of the interplay between defense systems and MGE on bacterial evolution across time scales.

Results

Association between defense systems and MGE abundance is MGE- and species-dependent

We used species-wise phylogenetic generalized mixed-effect models (PGMM) to study the association between the 7 most prevalent defense systems in the dataset (RM, DMS, Abi, CRISPR-Cas, Gabija, DRT, and CBASS), genome size, and the number of MGE per genome (Fig. 1). Notably, the sign and magnitude of the associations are strongly species-, system-, and MGE-dependent. Out of 197 species included in the analysis, around 20-30 (depending on the defense system) displayed a statistically significant ($p < 0.05$) positive association between the presence of the defense system and genome size. In contrast, statistically significant negative associations

84 were only observed in 5-15 species. Because differences in the number of genomes
85 per species could bias comparisons based on p-values, we also explored an
86 alternative criterion based on effect sizes to determine the number of positive and
87 negative associations (see Methods). Regardless of the criterion, CRISPR-Cas stood
88 out as the defense system that most often displays negative associations with
89 genome size. Other defense systems do not show a clear trend or, in the case of
90 CBASS and DRT, positively correlate with larger genomes in most of the species. The
91 analysis also reveals differences regarding the association between defense systems
92 and distinct types of MGE. In all defense systems, negative associations with
93 prophages are more frequent than negative associations with plasmids. Once again,
94 the clearest trend corresponds to CRISPR-Cas, whose presence tends to correlate
95 with a reduction in the number of prophages, but with higher numbers of
96 transposable elements and plasmids.

97 A more detailed analysis at the level of functional categories reveals that the
98 presence of defense systems is most often associated with changes in the number of
99 genes from COG categories X (mobilome), L (replication, recombination and repair),
100 U (intracellular trafficking and secretion), K (transcription), D (cell cycle control, cell
101 division and chromosome partitioning), and V (defense) (Fig. 1 and S1). Genes from
102 these functional categories are typically present in MGE, suggesting that correlations
103 (both positive and negative) between defense systems and genome size are
104 primarily due to differences in the abundance of MGE. We confirmed that by
105 masking genomic regions that correspond to MGE and rerunning the statistical
106 analysis. As expected, 82% of the significant associations disappeared after masking
107 MGE (Fig S2). Although a few significant associations persisted, a closer inspection
108 revealed that those were often related to genes from prophages and other MGE
109 (such as phage satellites) that had not been originally identified as such and
110 remained unmasked.

111 The sign of the association between defense systems and MGE does not follow a
112 clear taxonomic trend (Fig. 2, S3 and S4), with opposite signs sometimes found in
113 closely related species (see, for example, the differences for CRISPR-Cas in

114 *Phocaeicola vulgatus* and *P. dorei*). Moreover, the same species often display
115 opposite trends for different defense systems. For example, in *Pseudomonas*
116 *aeruginosa*, genomes with CRISPR-Cas contain fewer MGE, whereas genomes with
117 CBASS, Gabija, and RM systems are significantly enriched in MGE. Negative
118 associations between CRISPR-Cas and MGE are more abundant in the genus
119 *Acinetobacter* (5 out of 8 species), the phylum Bacteriodetes (negative association in
120 8 species, positive association in a single species) and the class Clostridia, the latter
121 especially affecting prophages (negative association in 11 species, positive
122 association in 2 species; Fig. S4). The genus *Acinetobacter* is also enriched in negative
123 associations involving Abi (5 species) and CBASS (4 species). Furthermore, three
124 almost non-overlapping groups of streptococci display negative associations for
125 different defense systems. These encompass *S. pyogenes*, *S. gordonii*, *S. anginosus*,
126 *S. mutans*, and *S. salivarius* in the case of CRISPR-Cas; *S. anginosus*, *S. oralis*, *S.*
127 *intermedius*, and *S. suis* in the case of RM and DMS; and *S. pyogenes*, *S. dysgalactiae*,
128 *S. equi*, and *S. uberis* in the case of Gabija.

129

130 <u>Associations between defense systems and MGE arise from differences in the rate of</u>
131 <u>gene acquisition</u>

132 To investigate if correlations between defense systems and MGE involve cross-strain
133 differences in genome plasticity, we built high-resolution strain trees and identified
134 clades that contain defense systems (DEF^+). Then, we inferred the rates of gene gain
135 and loss associated with different classes of MGE in DEF^+ clades and in their
136 respective sister groups lacking the defense system (DEF^-) using the phylogenomic
137 reconstruction tool GLOOME. We quantified the relative differences in gene gain and
138 loss by dividing the rates observed in DEF^+ clades by those from sister DEF^- clades.
139 Finally, we compared the resulting DEF^+/DEF^- gain and loss ratios between species in
140 which MGE abundances and defense systems are positively and negatively
141 correlated. We found that DEF^+ and DEF^- clades often differ in their gain rates but
142 not in their loss rates (Fig. 3). Specifically, in species in which defense systems are
143 associated with increased MGE abundance, gene gain rates are typically higher in

144 clades that contain the defense system. The opposite (that is, a reduction of gene
145 gain rates in DEF^+ clades) is observed in species in which defense systems are
146 associated with lower MGE abundance. Among the latter, the biggest reductions in
147 plasmid acquisition occur in association with RM, DRT, DMS and Abi, whereas
148 significant drops in prophage gain are observed for DRT, CBASS and CRISPR-Cas.
149 Taken together, these results confirm that correlations between defense systems
150 and MGE arise from cross-strain differences in gene gain rather than loss, as
151 expected given the role of MGE as facilitators of horizontal gene transfer.

152

153 Timescales and linkage determine the sign of associations between defense systems
154 and HGT

155 Because prokaryotic defense systems are often located within or adjacent to MGE,
156 we hypothesized that positive associations between defense systems and MGE
157 abundance could be explained, at least in part, by recent co-transfer events
158 (henceforth, we term this the "linkage hypothesis"). As a result, the negative effects
159 of defense systems on HGT should be easier to detect at longer timescales or under
160 evolutionary conditions that weakened the physical linkage of defense systems and
161 MGE.

162 To test this hypothesis, we studied how relative differences in the rates of overall
163 gene gain between DEF^+ and DEF^- clades depend on the depth of their last common
164 ancestor in the strain tree (note that the depth of the last common ancestor serves
165 as an upper limit for the time that the defense system has been retained in a
166 lineage). As predicted, positive outliers (with much higher gene gain rate in the DEF^+
167 clade than in the DEF^- clade) almost invariably correspond to very recent lineages
168 ($< 10^{-5}$ substitutions per bp in nearly universal core genes), indicating a very recent
169 acquisition of the defense system (Fig. 4a). In contrast, negative outliers (with much
170 lower gene gain rate in the DEF^+ clade than in the DEF^- clade) generally correspond
171 to deeper lineages (> 0.001 substitutions per bp). We quantified these trends by
172 calculating the skewness of the distribution of DEF^+/DEF^- log-transformed gain ratios

173 in very recent and older lineages (Fig. 4b and Table S1). All the distributions show
174 significant positive skewness in recent lineages and significant negative skewness in
175 older lineages (all $p < 0.001$ except recent RM lineages, with $p = 0.023$; d'Agostino
176 test for skewness). This quantitative analysis confirms that differences in timescale
177 affect all defense systems, with positive associations between defense systems and
178 gene gain being dominant in the short term and negative associations becoming
179 more frequent in the long term.

180 A second prediction of the linkage hypothesis is that the net effect of defense
181 systems on HGT is not only species, but also lineage specific. That is, defense systems
182 may display positive, zero, or negative association with HGT in different lineages
183 depending on when the defense system was acquired and how tight is the linkage to
184 MGE. A more detailed study of gene gain rates in individual species confirms that the
185 effect of CRISPR-Cas is, indeed, lineage-specific, with the same species encompassing
186 recent CRISPR-Cas$^+$ lineages that display increased gene gain rates and older CRISPR-
187 Cas$^+$ lineages with reduced gene gain rates (Fig. 4c). This observation, combined with
188 the very recent acquisition of CRISPR-Cas in most lineages, could explain why many
189 species show non-significant or positive associations between CRISPR-Cas and MGE
190 abundances. Indeed, of the 6 representative species shown in Fig. 4c, only
191 *Streptococcus anginosus* and *Acinetobacter radioresistens* show a net negative
192 association between CRISPR-Cas and MGE abundance in the PGLMM analysis.

193 The findings described so far indicate that, of the 7 defense systems considered in
194 this study, CRISPR-Cas is the one that most often induces a net reduction in HGT
195 rates. According to the linkage hypothesis, this could be due to a comparatively
196 weaker physical association between CRISPR-Cas and MGE. To evaluate that
197 possibility, we calculated the percentage of genes from CRISPR-Cas located inside
198 MGE and compared that to other defense systems. Consistent with the linkage
199 hypothesis, CRISPR-Cas is the defense system that is least often encoded by MGE,
200 followed by Gabija and DMS (Fig. 4d). (4.97% vs 8.12%, $p < 10^{-8}$, chi-squared test).

201

202 Anti-CRISPR proteins modulate associations between CRISPR-Cas and HGT

203 Anti-CRISPR proteins (Acr) have the potential to suppress the possible negative
204 effect of CRISPR-Cas on MGE-driven gene transfer (28). This, in turn, could contribute
205 to explaining the sign of the association between CRISPR-Cas and MGE in different
206 lineages. To assess that possibility, we searched all the genomes in the dataset for
207 known Acr, finding 16,058 proteins. Then, we compared the prevalence of Acr in
208 species in which the presence of CRISPR-Cas is positively or negatively correlated
209 with the MGE content, separately considering genomes that do and do not harbor
210 CRISPR-Cas. Because Acr are generally encoded by MGE (29) and the prevalence of
211 MGE systematically varies across groups acting as a confounding factor, we
212 restricted our comparisons to genomes that contain at least one prophage. Our
213 results indicate that genomes with and without CRISPR-Cas differ in the prevalence
214 of Acr (Fig. S5). More importantly, these differences have opposite directions in
215 species in which CRISPR-Cas is positively and negatively associated with MGE
216 abundance. In the former, Acr are more prevalent in genomes that contain CRISPR-
217 Cas (p = 0.0012, chi-squared test). In the latter, Acr are less prevalent in genomes
218 with CRISPR-Cas ($p < 10^{-6}$, chi-squared test). Interestingly, the prevalence of Acr in
219 genomes without CRISPR-Cas is similar in both groups. These results suggest that
220 negative associations between CRISPR-Cas and HGT could be dependent on (or at
221 least facilitated by) low prevalence of Acr in the genome.

222

223 **Discussion**

224 Defense systems could have a significant impact on microbial evolution by effectively
225 blocking the transfer of MGE, reducing gene flow and limiting the spread of
226 accessory genes. However, because defense systems are often carried by MGE and
227 these are the main vehicles of HGT, a net positive association between defense
228 systems and gene exchange cannot be ruled out *a priori*. We assessed the relative
229 weight of these two opposite scenarios by quantifying the association between

230 defense systems, MGE abundance, and gene acquisition rates in a phylogeny-aware
231 comparative study of 197 prokaryotic species.

232 Our results shed light on previous, apparently contradictory findings concerning the
233 effect of CRISPR-Cas on genome evolution and diversification (16-20). A pioneering
234 study conducted in 2015 found no evidence to support an overall association
235 between CRISPR-Cas activity and reduced gene acquisition via HGT at evolutionary
236 time scales (16). Such lack of association was explained by several factors, including
237 the high mobility of CRISPR-Cas systems, that limits their long-term impact on host
238 genomes, and the possibility that HGT is mediated by MGE that escape (or are not
239 targeted by) CRISPR-Cas immunity.

240 Our analyses support the general conclusion that CRISPR-Cas and other defense
241 systems have little overall impact on HGT in most bacterial species. Specifically,
242 differences in the rates of gene acquisition in lineages that do and do not harbor
243 defense systems are centered around zero. That said, we identified significant
244 opposite trends at very short and intermediate evolutionary time scales. In
245 particular, positive associations between defense systems and HGT are more
246 frequent at very short time scales, whereas negative associations become dominant
247 at longer time scales. These opposite trends suggest that the actual effects of
248 defense systems on gene exchange may be obscured by recent co-transfer events
249 involving MGE. As a result, the possible negative effects of defense systems on HGT
250 only become detectable if the defense system is maintained for long enough periods
251 of time.

252 Besides this general picture, we identified some species in which the presence of
253 defense systems (especially CRISPR-Cas) significantly correlates with smaller genome
254 sizes and MGE abundances. Some of those associations had been previously
255 described in *Pseudomonas aeruginosa* and *Klebsiella pneumoniae* (9, 20). And yet,
256 these species represent special cases rather than the rule, even in the context of
257 host-associated bacteria. In fact, our results underline that the association between
258 defense systems and HGT is strongly system- and species-dependent. This conclusion
259 confirms and extends previous findings that showed that the impact of CRISPR-Cas

260 on the spread of antibiotic resistance is highly variable across species and its sign
261 cannot be easily explained by simple ecological, environmental, or genomic
262 variables(19).

263 Among the 7 defense systems included in this study, CRISPR-Cas stands out for being
264 the most recurrently associated with reduced genome sizes and lower MGE
265 (especially prophage) abundances. In contrast, DMS, Abi, DRT, and CBASS are more
266 often associated with higher numbers of MGE and accessory genes. This finding is
267 fully consistent with a recent study that compared 73 defense systems in 12
268 bacterial species (30). We propose that what makes CRISPR-Cas systems different is
269 their weaker (though still substantial) linkage with MGE. Compared to fully
270 functional CRISPR-Cas systems, other defense systems like Abi, Gabija, CBASS, DMS,
271 and RM are more frequently located within or next to MGE (23, 24, 31, 32) and may
272 have alternative functions related to MGE propagation. For example, Abi systems
273 have been identified in PICIs as accessory genes that facilitate their parasitic lifecycle
274 (33). The narrow specificity of some defense systems may be another reason why
275 those systems do not significantly interfere with HGT. For instance, the GmrSD type
276 IV RM system selectively targets phages with glucosylated hydroxymethylcytosine
277 (34) and the Thoeris defense system only appears to be effective against myoviruses
278 (6). Although selective targeting is often viewed as an outcome of phage-host
279 coevolution, it is tempting to speculate that it could have been evolutionarily
280 favored by the need to fight harmful genetic parasites while maintaining sufficiently
281 high rates of HGT to prevent population-level gene loss (35).

282 All in all, we showed that some defense systems, especially CRISPR-Cas, can
283 significantly reduce HGT, although the effect is often masked by the fact that these
284 systems travel together with MGE. Beyond the possible functional connections
285 briefly discussed above, the linkage between defense systems and MGE is a
286 fundamental consequence of the arms race between parasites and hosts. Indeed,
287 because defense systems are costly and their efficacy drops as parasites evolve, they
288 are subject to rapid turnover and must rely on HGT for long-term persistence in
289 microbial populations (36-39). As a result, it is extremely challenging to disentangle

290 the impact of defense systems on gene flow from the causes that lead to their
291 presence or absence, especially at short evolutionary time scales. We expect that
292 future research on defense systems and MGE, fueled by the unprecedented
293 availability of genomic and metagenomic data, will help overcome these difficulties
294 by producing precise quantitative estimates of their linkage and better null models
295 for testing their role in microbial adaptation at different time scales.

296

297 **Methods**

298 Genome collection and identification of defense systems

299 We parsed the Genome Taxonomy Database (GTDB) release 202 (40) to identify all
300 high-quality genomes (according to the MIMAG criteria (41)) with completeness
301 >99%, contamination <1%, and contig count <500. The 82,595 genomes that passed
302 these filters were downloaded from the NCBI FTP site. CRISPR-Cas systems were
303 identified with CRISPRCasTyper v1.2.4 (42) using default parameters. We classified a
304 genome as CRISPR-Cas$^+$ if it contains at least one high-confidence Cas operon and
305 CRISPR-Cas$^-$ otherwise. Only the species with >10 genomes and at least 5 CRISPR-
306 Cas$^+$ genomes were further considered. To reduce the computational cost, we only
307 considered a maximum of 500 genomes per species. Species with >500 genomes
308 were randomly subsampled to keep at most 350 CRISPR-Cas$^+$ and 150 CRISPR-Cas$^-$
309 genomes. Other defense systems were identified with Padloc v1.1.0 (db v1.4.0) (43).
310 Our analysis focused on the most prevalent defense systems: restriction-
311 modification (RM), DMS, Abi, CRISPR-Cas, Gabija, DRT, and CBASS.

312 After applying these criteria, 19,323 genomes belonging to 196 bacterial and 1
313 archaeal species (*sensu* GTDB) were included in the analysis. Of those, 2,964
314 correspond to complete genomes and the rest to high-quality, nearly complete ones.

315 Gene prediction and annotation.

316 Open reading frames (ORF) were predicted with Prodigal v2.6.3, using codon table
317 11 (prokaryotic genetic code) and "single" mode, as recommended for finished and

318 draft quality genomes (44). Orthologous ORF were then separately clustered for
319 each species with Roary v3.13.0 (45) setting an 80% identity threshold for initial
320 clustering followed by synteny-based refinement (options '-t 11 -i 80'). The resulting
321 gene clusters were functionally annotated by selecting a representative sequence,
322 arbitrarily chosen among those with length between 0.95 and 1.05 times the average
323 length of all sequences in the cluster. Representative sequences were functionally
324 annotated by mapping them to in-house profiles of the Clusters of Orthologous
325 Genes (COG) database (2020 release) (46) with HMMER v3.1b2 (e-value < 0.001)
326 (http://hmmer.org). The 26 major prokaryotic functional categories defined in the
327 COG database were assigned to the annotated genes. Some functional categories (A,
328 RNA processing and modification; B, chromatin structure and dynamics; W,
329 extracellular structures; T, signal transduction; and Z, cytoskeleton) were excluded
330 since they rarely or never occur in prokaryotic genomes. The case-insensitive
331 keywords "phage", "plasmid" and "transpos*" in the COG gene annotations were
332 used to identify genes associated with prophages, plasmids, and transposons,
333 respectively, and the resulting gene lists were manually curated to minimize false
334 assignments. Anti-defense proteins, including anti-CRISPR, were identified by
335 running HMMER v3.1b2 (e-value < 10^{-10}) against the dbAPIS database (47).

336 Identification of genomic regions containing mobile genetic elements

337 Prophages were detected with Phispy v4.2.21 using default options (48). Short
338 transposons were identified based on the presence of isolated or paired genes
339 annotated as transposases. In the latter case, we allowed for up to one additional
340 gene between two transposon-related genes to account for the genetic architecture
341 of some insertion sequences (49). Due to the inherent difficulty to discriminate
342 among plasmids, ICE, and IME (9), we restricted the search for these elements to
343 complete genomes. Plasmids were identified as extra-chromosomal replicons that
344 contain the "plasmid" label in their NCBI description lines. After removing those
345 replicons, ICEfinder (50) was used to identify ICE and IME. Finally, we ran Phispy
346 v4.2.21 with option '-phage_genes 0' to identify any other MGE, integrons,

347 pathogenicity islands, and fragments of MGE that could have been missed by the
348 previous approaches.

349 Species trees

350 Phylogenetic trees were separately built for each species based on the set of 120
351 prokaryotic marker genes (122 in the case of Archaea) proposed by the GTDB r202.
352 For each species, only those marker genes with prevalence >80% were used for
353 phylogenetic reconstruction. We aligned the amino acid sequences of each marker
354 gene with mafft-linsi (L-INS-I algorithm, default options, MAFFT v7.475) (51) and
355 back-translated the amino acid alignments to nucleotide alignments with pal2nal.pl
356 v14 (52) using codon table 11. After concatenating all nucleotide alignments, we
357 built preliminary trees with FastTree v2.1.10 (options '-gtr -nt -gamma -nosupport -
358 mlacc 2 -slownni') (53). The tree topologies produced by FastTree were subsequently
359 provided to RaxML v8.2.12 (54) for branch length optimization (raxmlHPC with
360 options '-f e -m GTRGAMMA').

361 To visualize trends across species (Fig. 2, S3, and S4), we used the online tool iTOL
362 (55) and the multispecies tree from GTDB r202.

363 Phylogenetic generalized linear mixed models (PGLMM)

364 For each species, we collected the total genome size, the number of genes belonging
365 to each functional category, and the number of genes associated with prophages,
366 plasmids, and transposons, in genomes that do and do not harbor the defense
367 system of interest. Genes that belong to the defense system of interest were not
368 counted in any of the groups. For each variable, we fitted a PGLMM with Poisson
369 distribution and canonical link function, with the presence or absence of the defense
370 system as the predictor, the number of genes as the response variable, and the
371 corresponding species tree as the guide to generate the covariance matrix. The
372 PGLMM were fit with the function pglmm_compare() from the R package phyr v1.1.0
373 (56).

374 We used two alternative criteria to identify species in which the presence of a
375 defense system is positively or negative associated with gene numbers. On the one

376 side, we adopted a classical criterion of statistical significance ($p < 0.05$) for the
377 predictor variable in the PGLMM. On the other side, we an alternative criterion
378 based on effect size, aimed at comparing species with different sample sizes in which
379 comparison of p-values could be affected by sensitivity biases. For each variable and
380 defense system, we jointly considered the PGLMM of all the species and determined
381 the effect size threshold as the smallest effect size (in absolute value) that reached
382 statistical significance ($p < 0.05$) in any species.

383 Inference of gene gain and loss rates

384 Gene gains and losses at each branch of each species tree were estimated with
385 Gloome (57), using as inputs the gene presence/absence matrices previously
386 generated by Roary and the species trees. The parameter configuration file was set
387 to optimize the likelihood of the observed phyletic profiles under a genome
388 evolution model with 4 categories of gamma-distributed gain and loss rates and
389 stationary frequencies at the root.

390 Comparison of gene gain and loss rates between DEF^+ and DEF^- clades

391 For each species tree, we defined DEF^+ clades as the narrowest possible clades such
392 that at least 80% of the leaves contain the defense system of interest. Candidate
393 DEF^- clades were defined in an analogous way, referring to leaves without the
394 defense system. Next, we identified pairs of DEF^+ and DEF^- clades that constitute
395 sister groups. Sister $DEF^{+/-}$ pairs were excluded if both clades contained a single
396 genome. For each clade in a valid pair, we computed the overall gene gain and loss
397 rates as the expected number of gene gains (or losses) in that clade divided by the
398 total branch length. Gain and loss rates for different functional categories and MGE
399 we calculated in an analogous way but restricting the sum of gene gains and losses
400 to the genes of interest. When calculating overall and category-wise gain and loss
401 rates, we did not take into account the contribution of species-wise singletons
402 (genes without homologs in other genomes of the same species), as they may
403 represent false gene predictions or genes that are replaced at unusually high rates
404 (58). To account for the non-negative nature of gain and loss rates and their heavy-

405 tailed distributions, comparisons between DEF^+ and DEF^- sister branches were done
406 based on log-transformed rate estimates. To calculate the skewness of the
407 distributions and their statistical significance, we use the method proposed by
408 D'Agostino (59) as implemented by the scipy.stats.skewtest() function in Python (60).

409

410 **Funding**

411 Y.L. is supported by China Scholarship Council (No.202008440425). J.B. is supported
412 by the Maria Zambrano grant of the Spanish Ministry of Universities (Grant No.
413 UP2021-035), and the Severo Ochoa Program for Centres of Excellence in R&D of the
414 Agencia Estatal de Investigación of Spain (Grant No. CEX2020-000999-S (2022–2025)
415 to the CBGP). J.I is supported by the Ramón y Cajal Programme of the Spanish
416 Ministry of Science (Grant No. RYC-2017–22524); the Agencia Estatal de
417 Investigación of Spain (Grant No. PID2019-106618GA-I00), the Severo Ochoa
418 Programme for Centres of Excellence in R&D of the Agencia Estatal de Investigación
419 of Spain (Grant No. SEV-2016–0672 (2017–2021) to the CBGP); and the Comunidad
420 de Madrid (through the call Research Grants for Young Investigators from
421 Universidad Politécnica de Madrid, Grant No. M190020074JIIS).

422

423 **Figure legends**

424 **Figure 1:** Association between 7 widespread defense systems, genome size, and
425 MGE abundance. The length of the bars indicates the number of species displaying
426 positive or negative associations in a phylogenetic generalized linear mixed effects
427 model (see Methods) according to two alternative criteria: $p < 0.05$ and absolute
428 effect size greater than the smallest significant effect size (|ES|>SSES), separately
429 computed for each response variable and defense system. The association with MGE
430 was calculated based on marker genes for prophages (Ph), plasmids (Pl), and
431 transposons (Tr). Abbreviations of functional categories: X (mobilome), L (replication,
432 recombination and repair), U (intracellular trafficking and secretion), K

433 (transcription), D (cell cycle control, cell division and chromosome partitioning), and
434 V (defense).

435 **Figure 2:** Taxonomic distribution of species displaying positive and negative
436 associations between the presence of defense systems and the number of genes
437 from the mobilome (based on COG annotations). Taxa discussed in the text are
438 labeled: A, *Acinetobacter*; Pa, *Pseudomonas aeruginosa*; Pv, *Phocaeicola vulgatus*;
439 Pd, *Phocaeicola dorei*; S1, *Streptococcus gordonii*, *S. anginosus*, *S. mutans*, and *S.*
440 *salivarius*; S2, *S. oralis*, *S. intermedius*, and *S. suis*; S3, *S. pyogenes*, *S. dysgalactiae*, *S.*
441 *equi*, and *S. uberis*. See figure S3 for a larger version including all species names.

442 **Figure 3:** Relative differences in the rates of gene gain and loss between sister clades
443 that do and do not harbor defense systems (DEF^+ and DEF^-, respectively). The
444 boxplots represent the distribution of the DEF^+/DEF^- ratio of gene gain (or loss) rates
445 for each class of MGE, calculated for every pair of sister clades. Values greater (or
446 smaller) than 1 indicate increased (or reduced) gene flux in lineages that contain the
447 defense system. (a) Species that show a negative association between the defense
448 system and the number of marker genes for each class of MGE (PGLMM with
449 smallest significant effect size criterion). (b) Species that show a positive association
450 between the defense system and the number of marker genes for each class of MGE.
451 In the boxplots, the central line indicates the median, the box limits correspond to
452 the 25 and 75 percentiles, and the whiskers extend to the largest and smallest values
453 not classified as outliers. P-values are based on Wilcoxon test with log-ratio = 0 as
454 null hypothesis.

455 **Figure 4:** Association between defense systems and gene gain rates depends on the
456 time scale. (a) Ratio of overall gene gain rates between sister clades that do and do
457 not harbor defense systems (DEF^+ and DEF^-, respectively). Values greater (or smaller)
458 than 1 indicate increased (or reduced) gene flux in lineages that contain the defense
459 system. (b) Density distribution of the DEF^+/DEF^- gain ratios, comparing clades that
460 represent recent acquisitions of the defense system (depth < 10^{-5} substitutions per
461 site in core genes) and clades that have retained the defense system for longer times
462 (depth > 10^{-3} substitutions per site). The probability density function (y-axis) is

463 represented in logarithmic scale to facilitate the visualization the tails. The skewness
464 of all distributions is statistically significant (see Table S1), with positive values for
465 recent clades and negative values for deeper clades (c) Representative examples of
466 gene gain ratios in shallow and deeper sister groups from the same species. (d)
467 Percentage of genes from different defense systems located within known MGE.
468 Whiskers represent 95% confidence intervals based on the binomial distribution.

469

470 **Supplementary Figures**

471 **Figure S1:** Association between 7 widespread defense systems and the number of
472 genes from different functional categories (based on COG annotations). The length
473 of the bars indicates the number of species displaying positive or negative
474 associations in a phylogenetic generalized linear mixed effects model (see Methods)
475 according to two alternative criteria: $p < 0.05$ and absolute effect size greater than
476 the smallest significant effect size (|ES|>SSES), separately computed for each
477 response variable and defense system. Abbreviations of functional categories, C:
478 energy production and conversion; D: cell cycle control, cell division, chromosome
479 partitioning; E: amino acid transport and metabolism; F: nucleotide transport and
480 metabolism; G: carbohydrate transport and metabolism; H: coenzyme transport and
481 metabolism; I: lipid transport and metabolism; J: translation, ribosomal structure and
482 biogenesis; K: transcription; L: replication, recombination and repair; M: cell
483 wall/membrane/envelope biogenesis; N: cell motility; O: posttranslational
484 modification, protein turnover, chaperones; P: inorganic ion transport and
485 metabolism; Q: secondary metabolites biosynthesis, transport and catabolism; R:
486 general function prediction only; S: function unknown; T: signal transduction
487 mechanisms; U: intracellular trafficking, secretion, and vesicular transport; V:
488 defense mechanisms; X: mobilome: prophages, transposons.

489 **Figure S2:** Association between 7 widespread defense systems and the number of
490 genes from different functional categories before and after masking MGE from
491 complete genomes. The length of the bars indicates the number of species displaying

492 significant ($p < 0.05$) positive or negative associations in a phylogenetic generalized
493 linear mixed effects model. Abbreviations of functional categories: X (mobilome), L
494 (replication, recombination and repair), U (intracellular trafficking and secretion), K
495 (transcription), D (cell cycle control, cell division and chromosome partitioning), and
496 V (defense).

497 **Figure S3:** Taxonomic distribution of species displaying positive or negative
498 associations between the presence of defense systems and the number of genes
499 from the mobilome (based on COG annotations).

500 **Figure S4:** Taxonomic distribution of species displaying positive or negative
501 associations between the presence of defense systems and the number of genes in
502 the genome (left) or the number of marker genes for prophages (right).

503 **Figure S5:** Prevalence of anti-CRISPR proteins (Acr) in genomes that do and do not
504 contain CRISPR-Cas systems, separately calculated for species that display a positive
505 or negative association between the presence of CRISPR-Cas and the number of
506 genes from the mobilome (PGLMM, smallest significant effect size criterion). To
507 control for the confounding effect of MGE prevalence on Acr, only genomes with
508 prophages were considered. Whiskers represent 95% confidence intervals based on
509 the binomial distribution.

510

511 **Supplementary Tables**

512 **Table S1:** Skewness and associated p-values for the distributions of DEF^+/DEF^- log-
513 transformed gain ratios in recent (depth $< 10^{-5}$) and deep (depth $> 10^{-3}$) lineages.

514

515 **References**

516 1. Arnold BJ, Huang IT, Hanage WP. Horizontal gene transfer and adaptive
517 evolution in bacteria. Nat Rev Microbiol. 2022;20(4):206-18.
518 2. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web
519 of life. Nature reviews Genetics. 2015;16(8):472-82.

520 3. Hall RJ, Whelan FJ, McInerney JO, Ou Y, Domingo-Sananes MR. Horizontal
521 Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes. Front
522 Microbiol. 2020;11:1569.
523 4. Lee IPA, Eldakar OT, Gogarten JP, Andam CP. Bacterial cooperation through
524 horizontal gene transfer. Trends Ecol Evol. 2022;37(3):223-32.
525 5. Georjon H, Bernheim A. The highly diverse antiphage defence systems of
526 bacteria. Nat Rev Microbiol. 2023;21(10):686-700.
527 6. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic
528 discovery of antiphage defense systems in the microbial pangenome. Science.
529 2018;359(6379).
530 7. Shaw LP, Rocha EPC, MacLean RC. Restriction-modification systems have
531 shaped the evolution and distribution of plasmids across bacteria. Nucleic acids
532 research. 2023;51(13):6806-18.
533 8. Mayo-Munoz D, Pinilla-Redondo R, Birkholz N, Fineran PC. A host of armor:
534 Prokaryotic immune strategies against mobile genetic elements. Cell Rep.
535 2023;42(7):112672.
536 9. Botelho J, Cazares A, Schulenburg H. The ESKAPE mobilome contributes to
537 the spread of antimicrobial resistance and CRISPR-mediated conflict between mobile
538 genetic elements. Nucleic acids research. 2023;51(1):236-52.
539 10. Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. Selfish, promiscuous and
540 sometimes useful: how mobile genetic elements drive horizontal gene transfer in
541 microbial populations. Philos Trans R Soc Lond B Biol Sci. 2022;377(1861):20210234.
542 11. Watson BNJ, Capria L, Alseth EO, Pons B, Biswas A, Lenzi L, et al. CRISPR-Cas
543 in Pseudomonas aeruginosa provides transient population-level immunity against
544 high phage exposures. The ISME journal. 2024.
545 12. van Vliet AHM, Charity OJ, Reuter M. A Campylobacter integrative and
546 conjugative element with a CRISPR-Cas9 system targeting competing plasmids: a
547 history of plasmid warfare? Microb Genom. 2021;7(11).
548 13. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene
549 transfer in staphylococci by targeting DNA. Science. 2008;322(5909):1843-5.

550 14. Watson BNJ, Staals RHJ, Fineran PC. CRISPR-Cas-Mediated Phage Resistance
551 Enhances Horizontal Gene Transfer by Transduction. mBio. 2018;9(1).
552 15. O'Hara BJ, Barth ZK, McKitterick AC, Seed KD. A highly specific phage defense
553 system is a conserved feature of the Vibrio cholerae mobilome. PLoS genetics.
554 2017;13(6):e1006838.
555 16. Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. No evidence
556 of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales.
557 The ISME journal. 2015;9(9):2021-7.
558 17. O'Meara D, Nunney L. A phylogenetic test of the role of CRISPR-Cas in limiting
559 plasmid acquisition and prophage integration in bacteria. Plasmid. 2019;104:102418.
560 18. Pursey E, Dimitriu T, Paganelli FL, Westra ER, van Houte S. CRISPR-Cas is
561 associated with fewer antibiotic resistance genes in bacterial pathogens. Philos Trans
562 R Soc Lond B Biol Sci. 2022;377(1842):20200464.
563 19. Shehreen S, Chyou TY, Fineran PC, Brown CM. Genome-wide correlation
564 analysis suggests different roles of CRISPR-Cas systems in the acquisition of antibiotic
565 resistance genes in diverse species. Philos Trans R Soc Lond B Biol Sci.
566 2019;374(1772):20180384.
567 20. Wheatley RM, MacLean RC. CRISPR-Cas systems restrict horizontal gene
568 transfer in Pseudomonas aeruginosa. The ISME journal. 2021;15(5):1420-33.
569 21. Westra ER, Levin BR. It is unclear how important CRISPR-Cas systems are for
570 protecting natural populations of bacteria against infections by mobile genetic
571 elements. Proceedings of the National Academy of Sciences of the United States of
572 America. 2020;117(45):27777-85.
573 22. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, et al.
574 High viral abundance and low diversity are associated with increased CRISPR-Cas
575 prevalence across microbial ecosystems. Curr Biol. 2022;32(1):220-7 e5.
576 23. Makarova KS, Wolf YI, Snir S, Koonin EV. Defense islands in bacterial and
577 archaeal genomes and prediction of novel defense systems. Journal of bacteriology.
578 2011;193(21):6039-56.

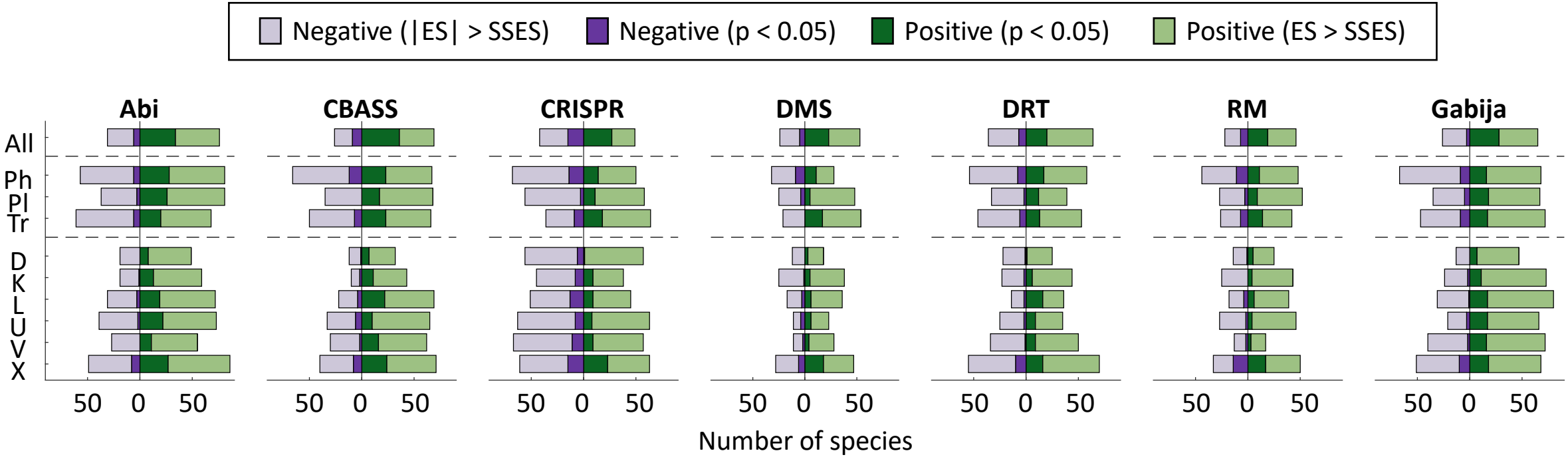
579 24. Botelho J. Defense systems are pervasive across chromosomally integrated
580 mobile genetic elements and are inversely correlated to virulence and antimicrobial
581 resistance. Nucleic acids research. 2023;51(9):4385-97.
582 25. Pinilla-Redondo R, Russel J, Mayo-Munoz D, Shah SA, Garrett RA, Nesme J, et
583 al. CRISPR-Cas systems are widespread accessory elements across bacterial and
584 archaeal plasmids. Nucleic acids research. 2022;50(8):4315-28.
585 26. Rocha EPC, Bikard D. Microbial defenses against mobile genetic elements and
586 viruses: Who defends whom from what? PLoS Biol. 2022;20(1):e3001514.
587 27. Koonin EV, Makarova KS, Wolf YI, Krupovic M. Evolutionary entanglement of
588 mobile genetic elements and host defence systems: guns for hire. Nature reviews
589 Genetics. 2020;21(2):119-31.
590 28. Mahendra C, Christie KA, Osuna BA, Pinilla-Redondo R, Kleinstiver BP, Bondy-
591 Denomy J. Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer.
592 Nature microbiology. 2020;5(4):620-9.
593 29. Pinilla-Redondo R, Shehreen S, Marino ND, Fagerlund RD, Brown CM,
594 Sorensen SJ, et al. Discovery of multiple anti-CRISPRs highlights anti-defense gene
595 clustering in mobile genetic elements. Nature communications. 2020;11(1):5652.
596 30. Kogay R, Wolf YI, Koonin EV. Defense systems and horizontal gene transfer in
597 bacteria. bioRxiv. 2024:2024.02.09.579689.
598 31. Benler S, Faure G, Altae-Tran H, Shmakov S, Zheng F, Koonin E. Cargo Genes
599 of Tn7-Like Transposons Comprise an Enormous Diversity of Defense Systems,
600 Mobile Genetic Elements, and Antibiotic Resistance Genes. mBio.
601 2021;12(6):e0293821.
602 32. Rousset F, Depardieu F, Miele S, Dowding J, Laval AL, Lieberman E, et al.
603 Phages and their satellites encode hotspots of antiviral systems. Cell Host Microbe.
604 2022;30(5):740-53 e5.
605 33. Ibarra-Chavez R, Hansen MF, Pinilla-Redondo R, Seed KD, Trivedi U. Phage
606 satellites and their emerging applications in biotechnology. FEMS Microbiol Rev.
607 2021;45(6).

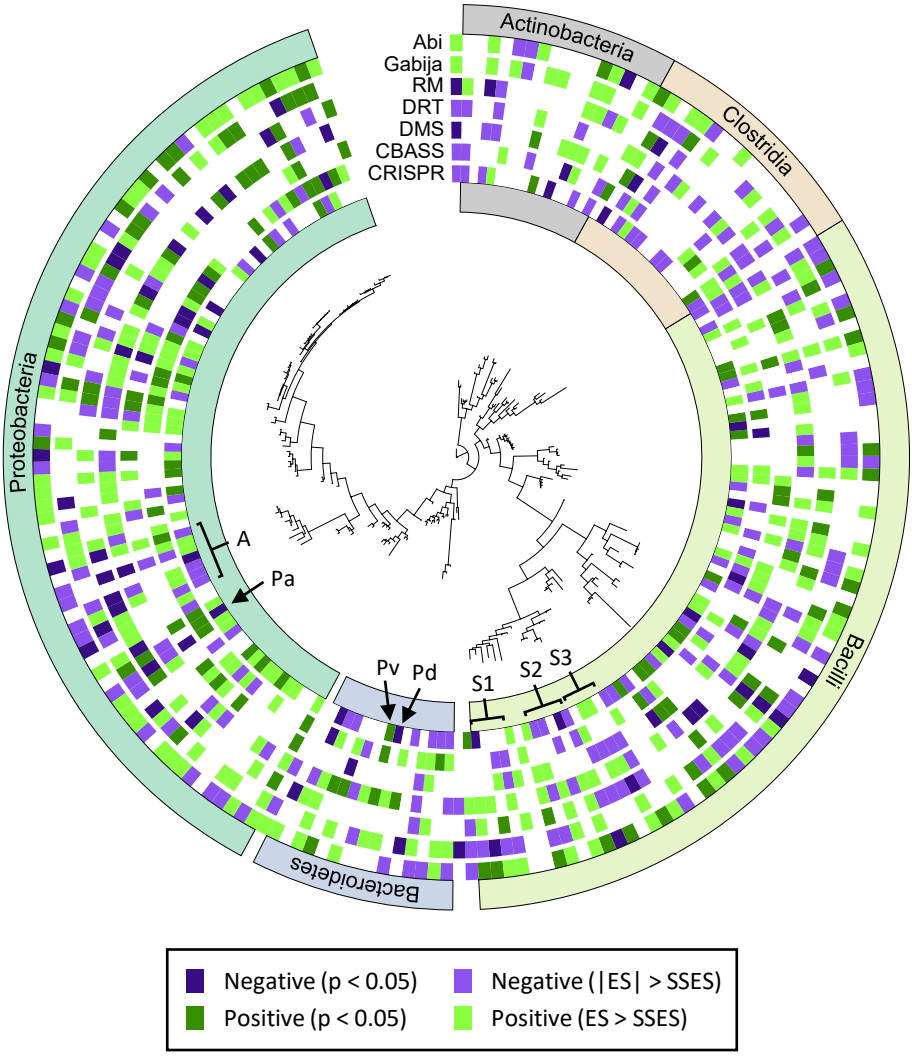
608 34. Bair CL, Black LW. A type IV modification dependent restriction nuclease that
609 targets glucosylated hydroxymethyl cytosine modified DNAs. J Mol Biol.
610 2007;366(3):768-78.
611 35. Iranzo J, Puigbo P, Lobkovsky AE, Wolf YI, Koonin EV. Inevitability of Genetic
612 Parasites. Genome biology and evolution. 2016;8(9):2856-69.
613 36. van Houte S, Buckling A, Westra ER. Evolutionary Ecology of Prokaryotic
614 Immune Mechanisms. Microbiology and molecular biology reviews : MMBR.
615 2016;80(3):745-63.
616 37. Puigbo P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV. Reconstruction of
617 the evolution of microbial defense systems. BMC evolutionary biology.
618 2017;17(1):94.
619 38. Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. Disentangling the
620 effects of selection and loss bias on gene dynamics. Proceedings of the National
621 Academy of Sciences of the United States of America. 2017;114(28):E5616-E24.
622 39. Koonin EV, Makarova KS, Wolf YI. Evolutionary Genomics of Defense Systems
623 in Archaea and Bacteria. Annual review of microbiology. 2017;71:233-61.
624 40. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A
625 complete domain-to-species taxonomy for Bacteria and Archaea. Nature
626 biotechnology. 2020;38(9):1079-86.
627 41. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy
628 TBK, et al. Minimum information about a single amplified genome (MISAG) and a
629 metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature
630 biotechnology. 2017;35(8):725-31.
631 42. Russel J, Pinilla-Redondo R, Mayo-Munoz D, Shah SA, Sorensen SJ.
632 CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-
633 Cas Loci. CRISPR J. 2020;3(6):462-9.
634 43. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, et al. PADLOC:
635 a web server for the identification of antiviral defence systems in microbial genomes.
636 Nucleic acids research. 2022;50(W1):W541-W50.

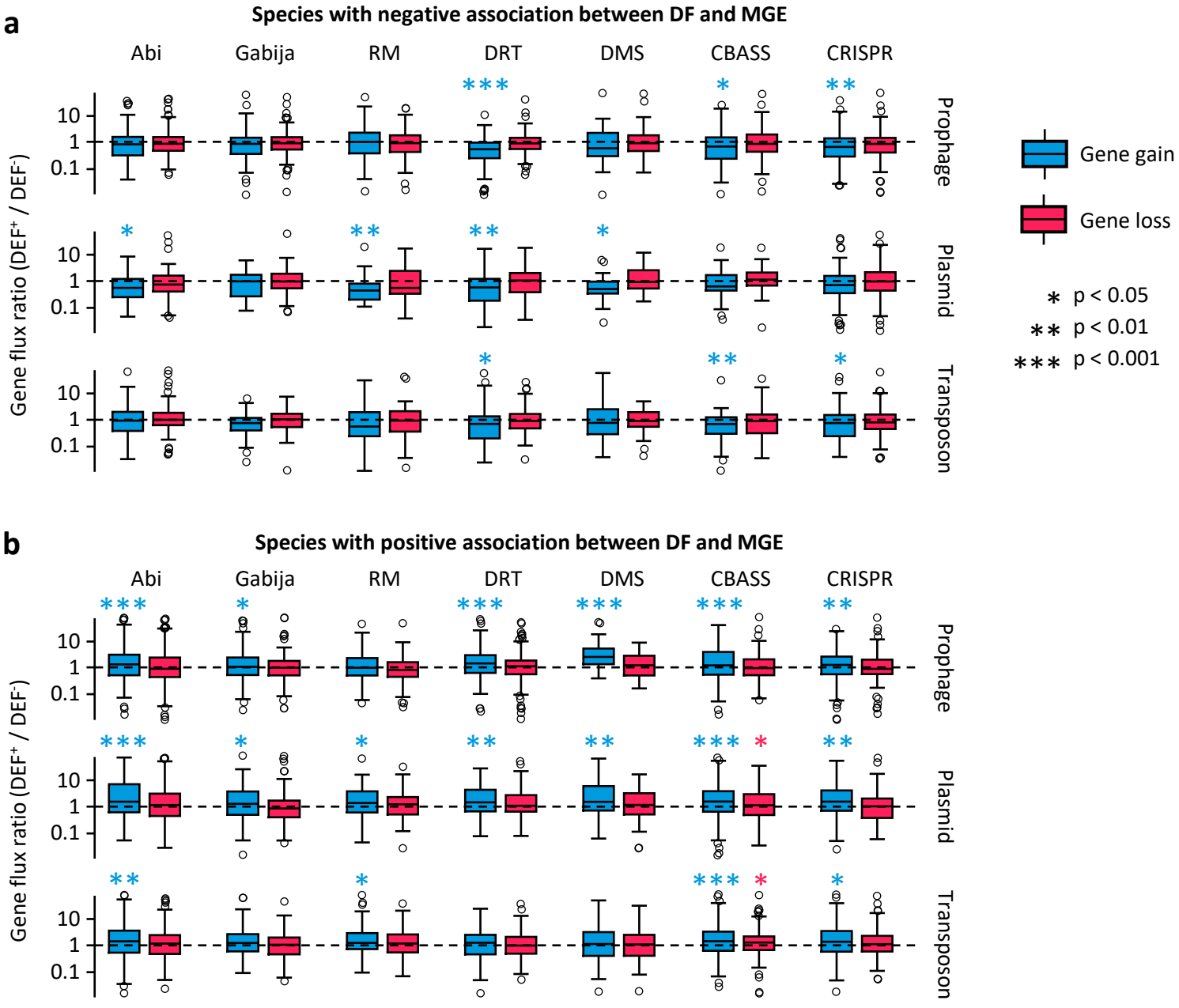
637 44. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal:
638 prokaryotic gene recognition and translation initiation site identification. BMC
639 Bioinformatics. 2010;11:119.
640 45. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary:
641 rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-
642 3.
643 46. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV.
644 COG database update: focus on microbial diversity, model organisms, and
645 widespread pathogens. Nucleic acids research. 2021;49(D1):D274-D81.
646 47. Yan Y, Zheng J, Zhang X, Yin Y. dbAPIS: a database of anti-prokaryotic immune
647 system genes. Nucleic acids research. 2024;52(D1):D419-D25.
648 48. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding
649 prophages in bacterial genomes that combines similarity- and composition-based
650 strategies. Nucleic acids research. 2012;40(16):e126.
651 49. Gomez MJ, Diaz-Maldonado H, Gonzalez-Tortuero E, Lopez de Saro FJ.
652 Chromosomal replication dynamics and interaction with the beta sliding clamp
653 determine orientation of bacterial transposable elements. Genome biology and
654 evolution. 2014;6(3):727-40.
655 50. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, et al. ICEberg 2.0: an updated database of
656 bacterial integrative and conjugative elements. Nucleic acids research.
657 2019;47(D1):D660-D5.
658 51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version
659 7: improvements in performance and usability. Molecular biology and evolution.
660 2013;30(4):772-80.
661 52. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein
662 sequence alignments into the corresponding codon alignments. Nucleic acids
663 research. 2006;34(Web Server issue):W609-12.
664 53. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood
665 trees for large alignments. PLoS One. 2010;5(3):e9490.

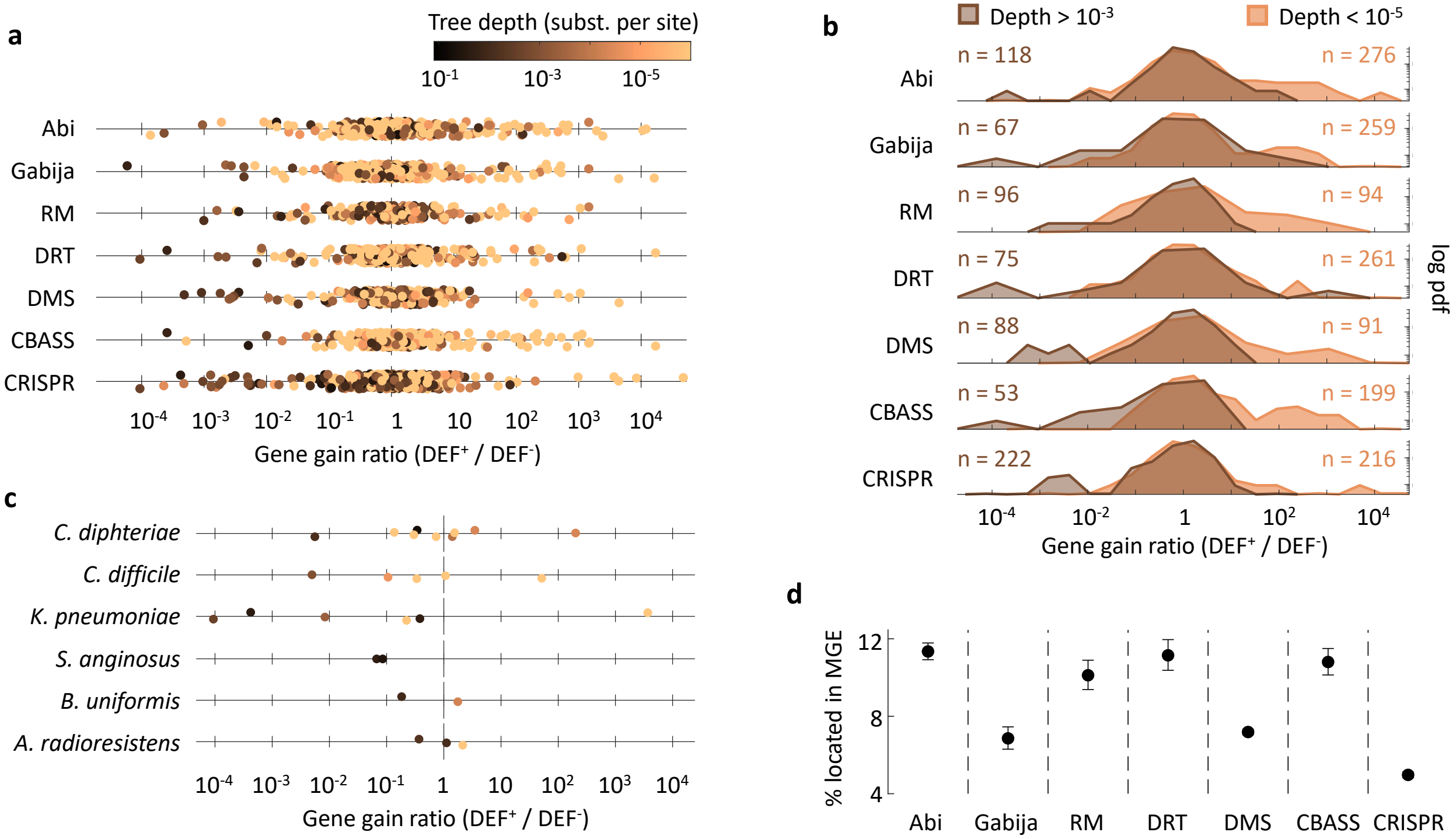
666 54. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-
667 analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
668 55. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for
669 phylogenetic tree display and annotation. Nucleic acids research.
670 2021;49(W1):W293-W6.
671 56. Li D, Dinnage R, Nell LA, Helmus MR, Ives AR. phyr: An r package for
672 phylogenetic species-distribution modelling in ecological communities. Methods in
673 Ecology and Evolution. 2020;11(11):1455-63.
674 57. Cohen O, Pupko T. Inference and characterization of horizontally transferred
675 gene families using stochastic mapping. Molecular biology and evolution.
676 2010;27(3):703-13.
677 58. Wolf YI, Makarova KS, Lobkovsky AE, Koonin EV. Two fundamentally different
678 classes of microbial genes. Nature microbiology. 2016;2:16208.
679 59. D'Agostino RB, Belanger A, D'Agostino Jr RB. A Suggestion for Using Powerful
680 and Informative Tests of Normality. The American Statistician. 1990;44(4):316-21.
681 60. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D,
682 et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat
683 Methods. 2020;17(3):261-72.

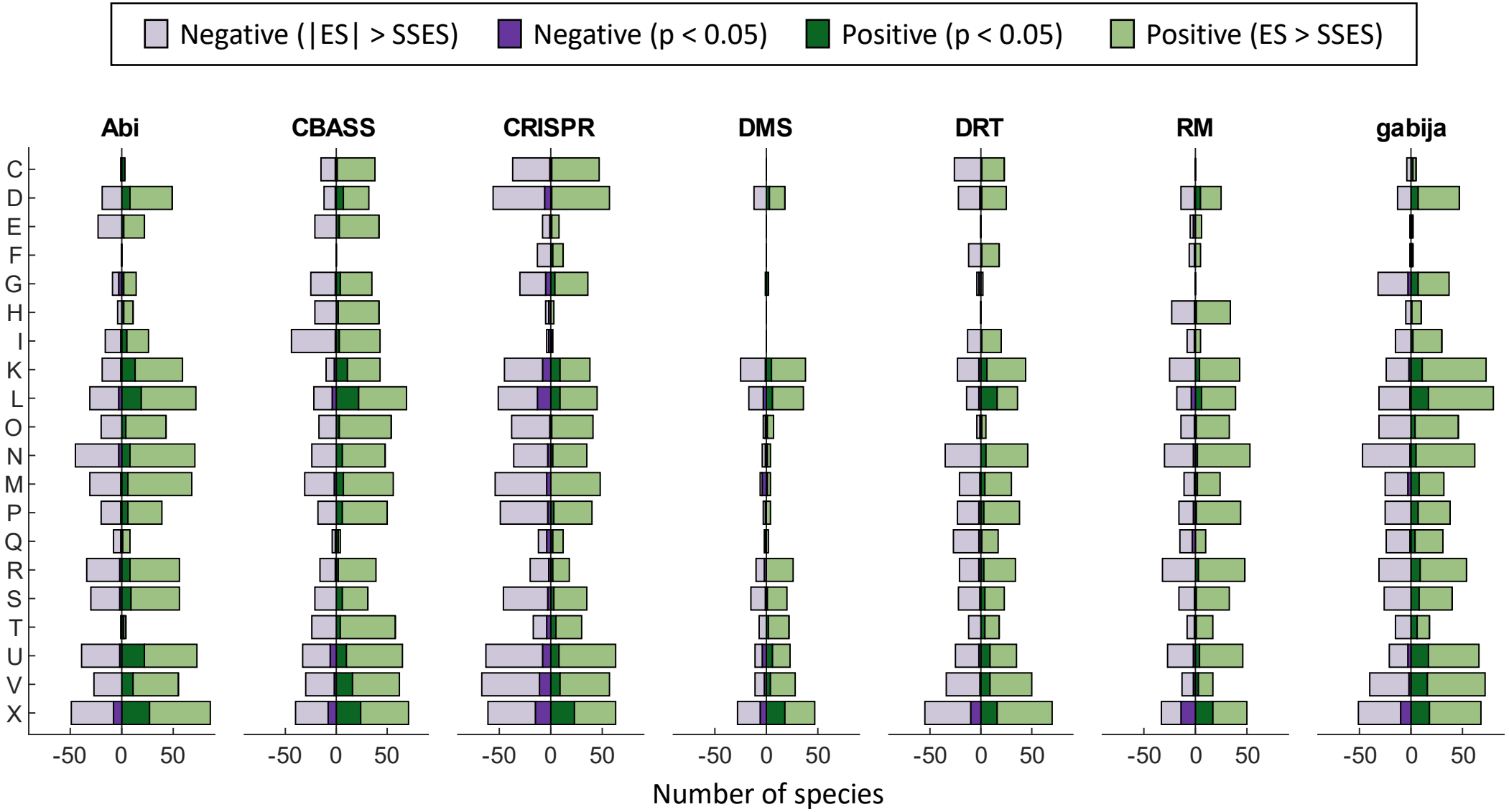
684

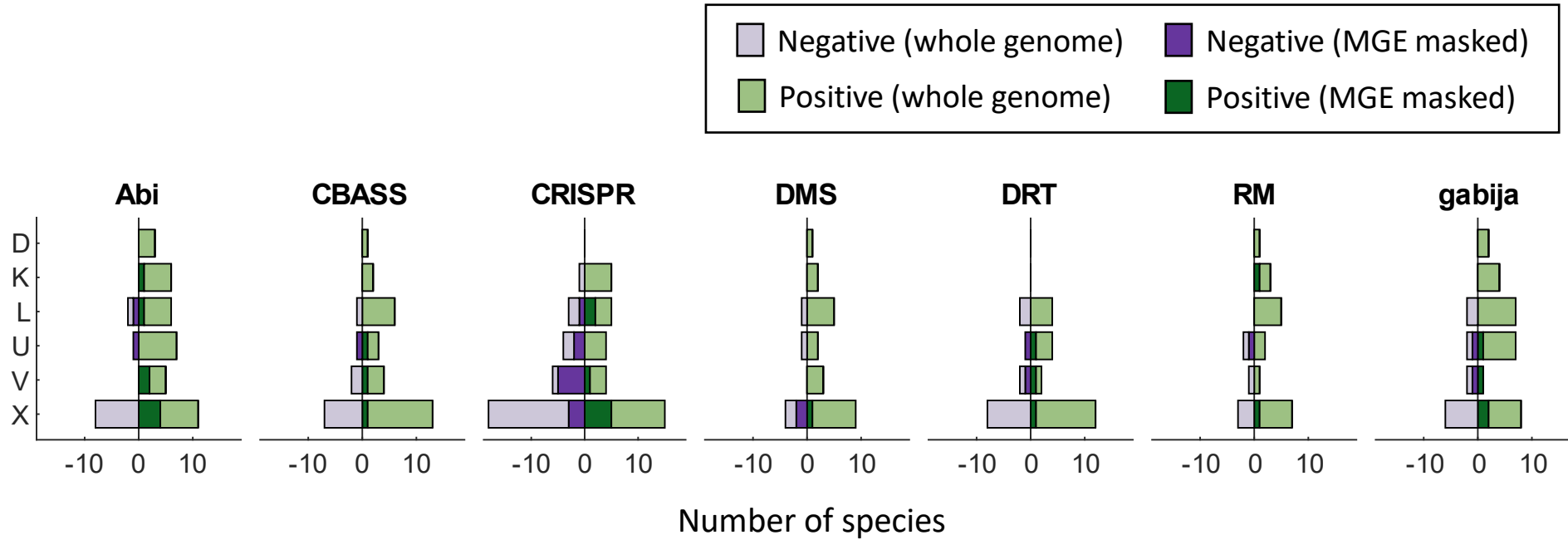

Negative (|ES| > SSES)
Negative (p < 0.05)
Positive (p < 0.05)
Positive (ES > SSES)
Abi
CBASS
CRISPR
DMS
DRT
RM
Gabija
All
Ph
Pl
Tr
D
K
L
U
V
X
50 0 50
Number of species

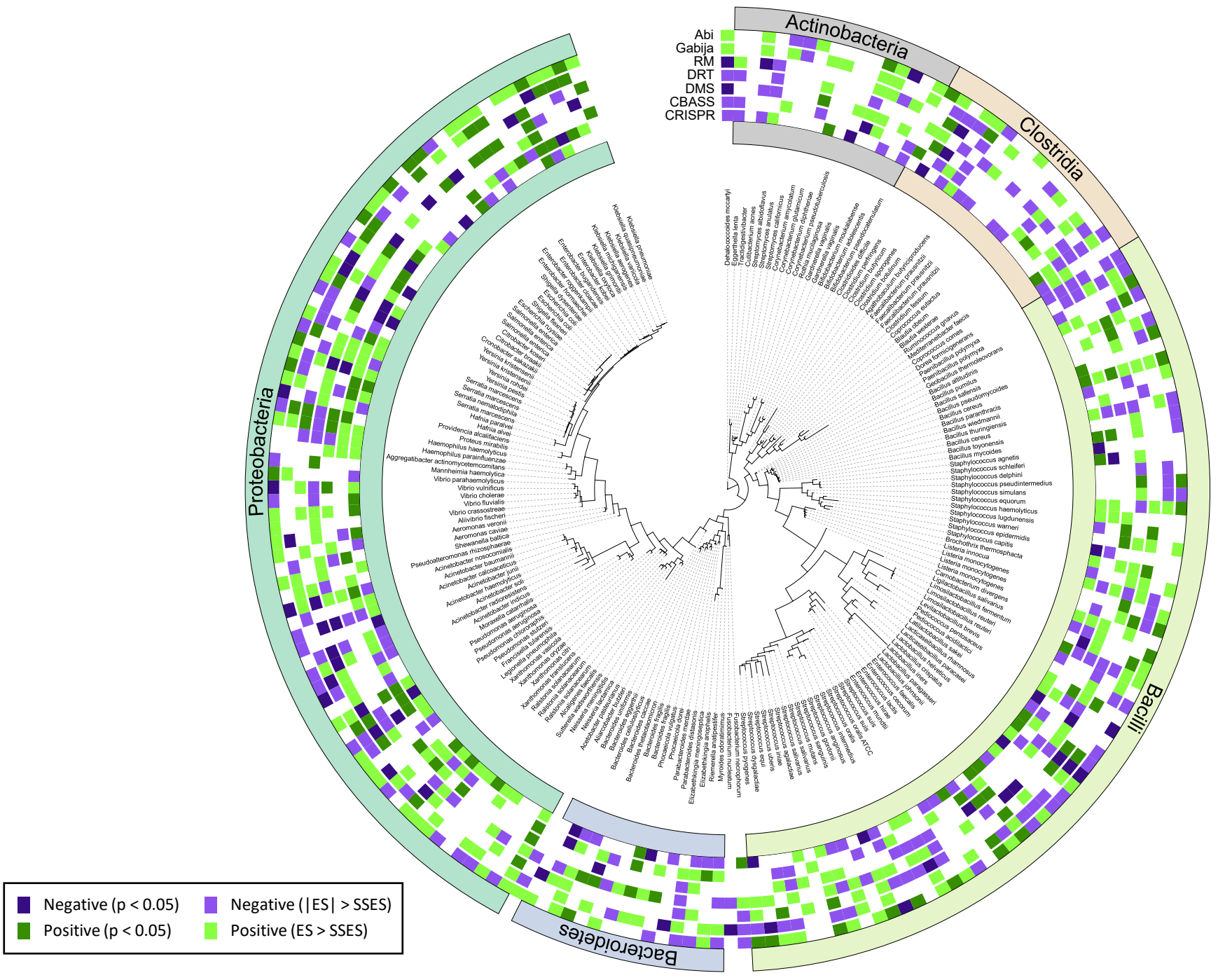

Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Actinobacteria
Clostridia
Bacilli
Bacteroidetes
Proteobacteria
A
Pa
Pv
Pd
S1
S2
S3
Negative (p < 0.05)
Positive (p < 0.05)
Negative (|ES| > SSES)
Positive (ES > SSES)

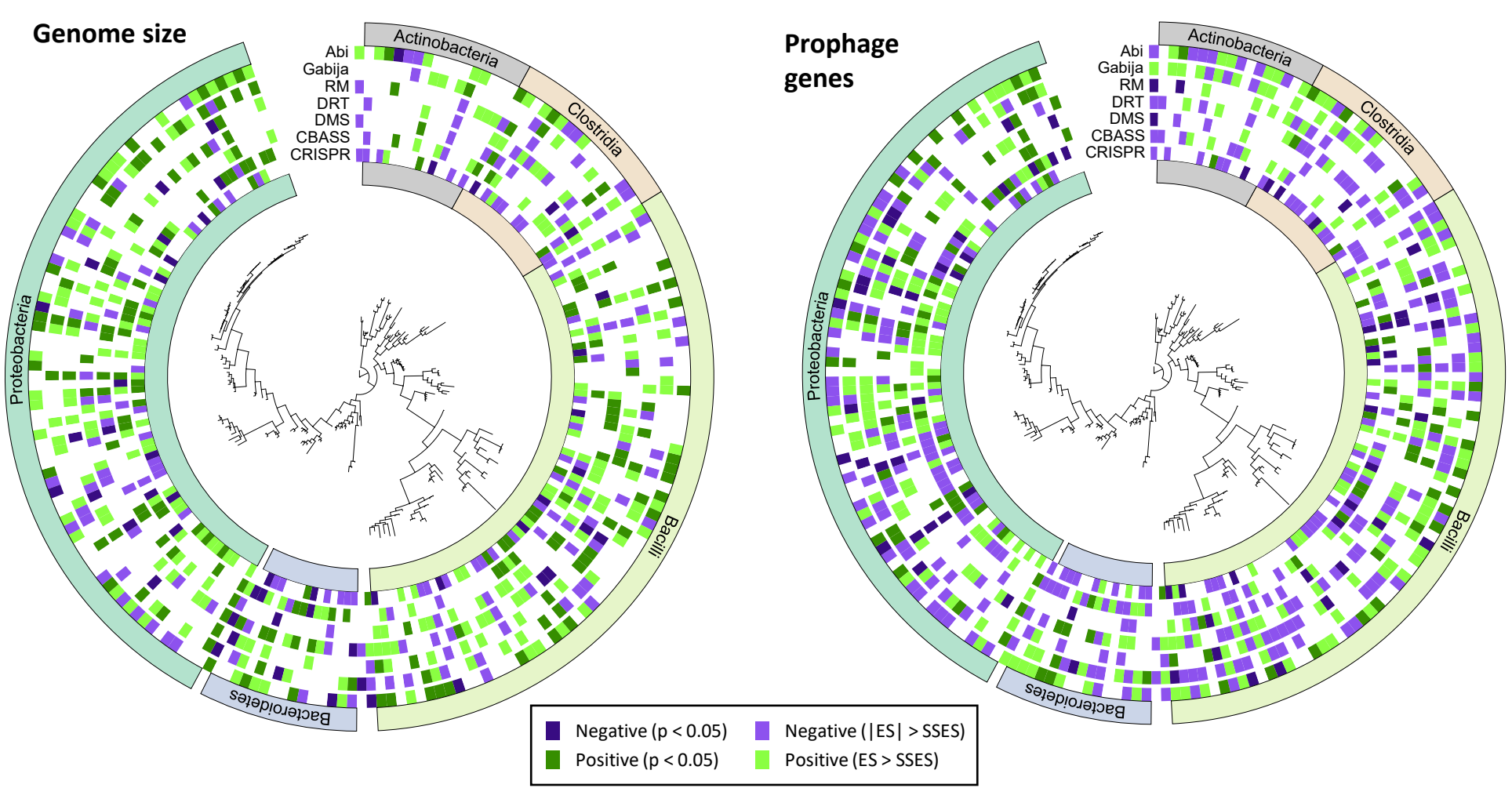

a
Species with negative association between DF and MGE
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Prophage
Plasmid
Transposon
Gene flux ratio (DEF+ / DEF-)
10
1
0.1
Gene gain
Gene loss
* p < 0.05
** p < 0.01
*** p < 0.001
b
Species with positive association between DF and MGE

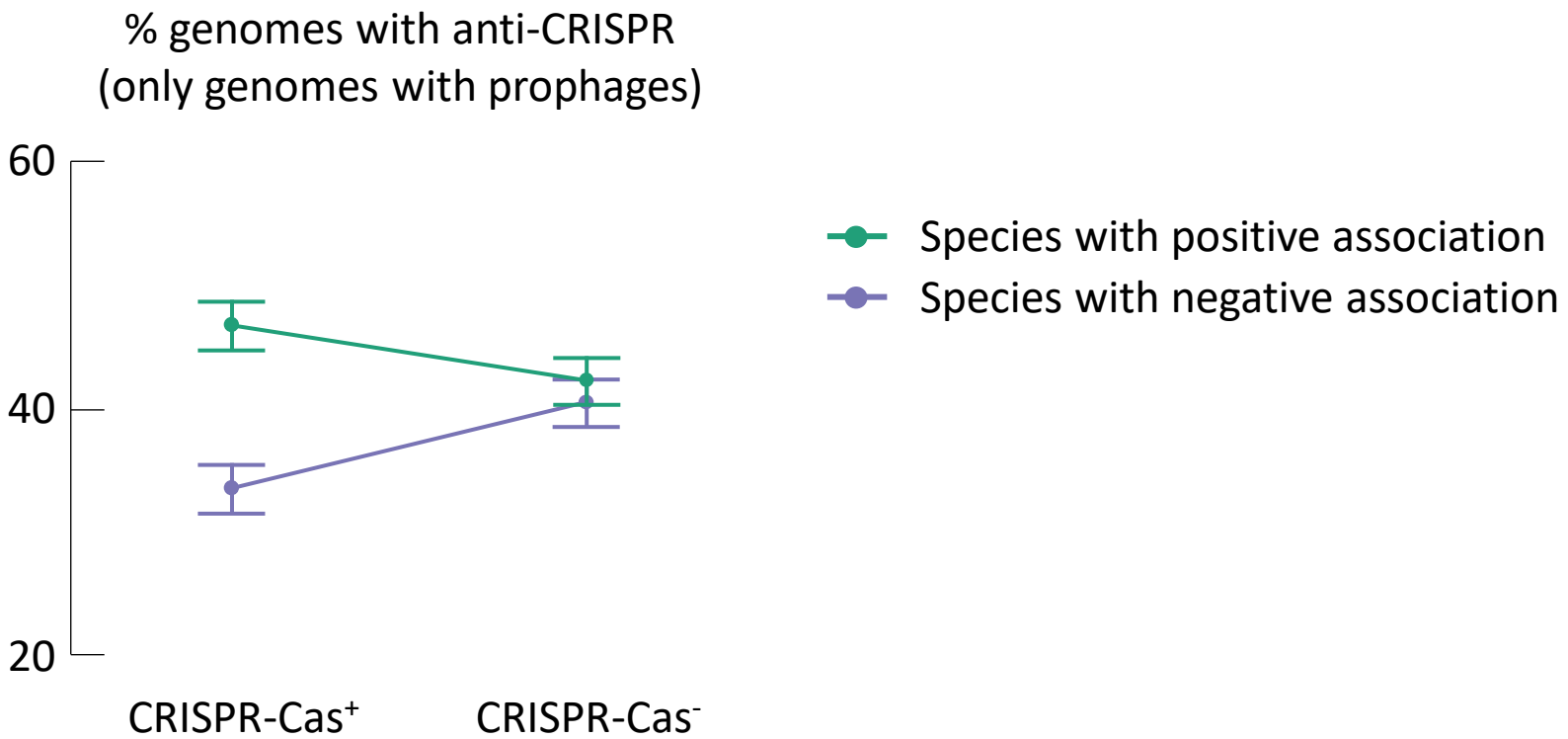

a
Tree depth (subst. per site)
10-1
10-3
10-5
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
10-4 10-3 10-2 10-1 1 10 102 103 104
Gene gain ratio (DEF+ / DEF-)
b
Depth > 10-3
Depth < 10-5
n = 118
n = 276
n = 67
n = 259
n = 96
n = 94
n = 75
n = 261
n = 88
n = 91
n = 53
n = 199
n = 222
n = 216
log pdf
10-4 10-2 1 102 104
Gene gain ratio (DEF+ / DEF-)
c
C. diphteriae
C. difficile
K. pneumoniae
S. anginosus
B. uniformis
A. radioresistens
10-4 10-3 10-2 10-1 1 10 102 103 104
Gene gain ratio (DEF+ / DEF-)
d
% located in MGE
12
8
4
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR


Negative (|ES| > SSES)
Negative (p < 0.05)
Positive (p < 0.05)
Positive (ES > SSES)
Abi
CBASS
CRISPR
DMS
DRT
RM
gabija
C
D
E
F
G
H
I
K
L
O
N
M
P
Q
R
S
T
U
V
X
-50
0
50
Number of species


Negative (whole genome)
Negative (MGE masked)
Positive (whole genome)
Positive (MGE masked)
Abi
CBASS
CRISPR
DMS
DRT
RM
gabija
D
K
L
U
V
X
-10
0
10
Number of species


Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Actinobacteria
Clostridia
Bacilli
Bacteroidetes
Proteobacteria
Negative (p < 0.05)
Positive (p < 0.05)
Negative (|ES| > SSES)
Positive (ES > SSES)


Genome size
Prophage genes
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Actinobacteria
Clostridia
Bacilli
Bacteroidetes
Proteobacteria
Negative (p < 0.05)
Positive (p < 0.05)
Negative (|ES| > SSES)
Positive (ES > SSES)


% genomes with anti-CRISPR
(only genomes with prophages)
60
40
20
CRISPR-Cas+
CRISPR-Cas-
Species with positive association
Species with negative association