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

Yang Liu[1], João Botelho[1], Jaime Iranzo[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.

Running title: Defense systems and horizontal gene transfer

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

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 taxon- 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 to 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.

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 (Arnold et al. 2022). 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 (Soucy et al. 2015; Hall et al. 2020; Lee et al. 2022). 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 (Doron et al. 2018; Botelho et al. 2023; Georjon and Bernheim 2023; Mayo-Munoz et al. 2023; Shaw et al. 2023). 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 (Haudiquet et al. 2022).

Recent studies have highlighted the role of CRISPR-Cas as widespread adaptive immunity systems that protect archaea and bacteria against viruses and other MGE (van Vliet et al. 2021; Watson et al. 2024). While in vitro experiments have provided supportive evidence for this function (Marraffini and Sontheimer 2008; O'Hara et al. 2017; Watson et al. 2018), questions remain about the extent to which CRISPR-Cas systems constrain gene exchange in nature (Gophna et al. 2015; O'Meara and Nunney 2019; Shehreen et al. 2019; Westra and Levin 2020; Wheatley and MacLean 2021; Pursey et al. 2022). 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 including Gabija, CBASS (cyclic oligonucleotide-based antiphage signaling system), DMS (DNA modification-based systems), and DRT (defense-associated reverse transcriptases) affect HGT in prokaryotes (Tesson et al. 2022; Costa et al. 2024).

In contrast with the expectation that defense systems restrict 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 (Oliveira et al. 2016; Meaden et al. 2022). 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 (Makarova et al. 2011; Pinilla-Redondo et al. 2022; Botelho 2023), promoting the retention of the latter for their beneficial side-effects in cellular defense (Koonin et al. 2020; Rocha and Bikard 2022).

Here, we investigated the association between 7 widespread defense systems and HGT rates in 197 prokaryotic species. By combining high-quality genomic data, phylogenomic methods, and phylogeny-aware statistical inference, our study aimed to 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 taxon-dependent

We used species-wise phylogenetic generalized linear mixed models (PGLMM) to study the association between the presence or absence of the 7 most prevalent defense systems in the dataset (RM, DMS, Abi, CRISPR-Cas, Gabija, DRT, and CBASS), genome size (measured as the total number of genes), and the number of MGE per genome (Fig. 1, Tables S1-S4). Notably, the sign and magnitude of the associations are strongly taxon-, 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 were only observed in 5-15 species (Fig. 1a, top row). Because differences in the number of genomes per species could bias comparisons based on p-values, we also explored an alternative criterion based on effect sizes to determine the number of positive and negative associations (see Methods). Regardless of the criterion, CRISPR-Cas stood out as the defense system that most often displays negative associations with genome size. In contrast, other defense systems positively correlate with larger genomes in most of the species. The analysis also reveals differences regarding the association between defense systems and distinct types of MGE (Fig. 1a, middle rows). In all defense systems, negative associations with prophages are more frequent than negative associations with plasmids. Once again, the clearest trend corresponds to CRISPR-Cas, whose presence often correlates with a reduction in the number of prophages, but with higher numbers of transposable elements and plasmids. Significant associations, when detected,

affect a sizeable fraction of the accessory genome, with 20-40% differences in MGE content between genomes that do and do not harbor the defense system (Fig. 1b).

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

The sign of the association between defense systems and MGE does not follow a clear taxonomic trend (Fig. 2, S3 and S4), with opposite signs sometimes found in closely related species (see, for example, the differences for CRISPR-Cas in *Phocaeicola vulgatus* and *P. dorei*). Moreover, the same species often display opposite trends for different defense systems. For example, in *Pseudomonas aeruginosa*, genomes with CRISPR-Cas contain fewer MGE, whereas genomes with CBASS, Gabija, and RM systems are significantly enriched in MGE. Negative associations between CRISPR-Cas and MGE are more abundant in the genus *Acinetobacter* (5 out of 8 species), the phylum *Bacteroidota* (negative association in 8 species, positive association in a single species) and the class *Clostridia*, the latter especially affecting prophages (negative association in 11 species, positive association in 2 species; Fig. S4). The genus *Acinetobacter* is also enriched in negative associations involving Abi (5 species) and CBASS (4 species). Furthermore, three almost non-overlapping groups of streptococci

display negative associations for different defense systems. These encompass *S. pyogenes*, *S. gordonii*, *S. anginosus*, *S. mutans*, and *S. salivarius* in the case of CRISPR-Cas; *S. anginosus*, *S. oralis*, *S. intermedius*, and *S. suis* in the case of RM and DMS; and *S. pyogenes*, *S. dysgalactiae*, *S. equi*, and *S. uberis* in the case of Gabija.

The only archaeal species in the study (*Methanococcus maripaludis*) did not show any remarkable trend, other than a relatively strong (but not statistically significant) positive association between prophages and DMS/RM, and between transposons and Gabija, and a moderate negative association between prophages and Gabija.

Associations between defense systems and MGE arise from differences in the rate of gene acquisition

To investigate if correlations between defense systems and MGE involve cross-strain differences in genome plasticity, we built high-resolution strain trees and identified clades that contain defense systems (DEF^+). Then, we inferred the rates of gene gain and loss associated with different classes of MGE in DEF^+ clades and in their respective sister groups lacking the defense system (DEF^-) using the phylogenomic reconstruction tool GLOOME. We quantified the relative differences in gene gain and loss by dividing the rates observed in DEF^+ clades by those from sister DEF^- clades. Finally, we compared the resulting DEF^+/DEF^- gain and loss ratios between species in which MGE abundances and defense systems are positively and negatively correlated. We found that DEF^+ and DEF^- clades often differ in their gain rates but not in their loss rates (Fig. 3). Specifically, in species in which defense systems are associated with increased MGE abundance, gene gain rates are typically higher in clades that contain the defense system. The opposite (that is, a reduction of gene gain rates in DEF^+ clades) is observed in species in which defense systems are associated with lower MGE abundance. Among the latter, the biggest reductions in plasmid acquisition occur in association with RM, DRT, DMS and Abi, whereas significant drops in prophage gain are observed for DRT, CBASS and CRISPR-Cas. Taken together, these results confirm that correlations between defense systems and

MGE arise from cross-strain differences in gene gain rather than loss, as expected given the role of MGE as facilitators of HGT.

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

Because prokaryotic defense systems are often located within or adjacent to MGE, we hypothesized that positive associations between defense systems and MGE abundance could be explained, at least in part, by recent co-transfer events (henceforth, we term this the "linkage hypothesis"). Uder this hypothesis, positive associations would simply result from defense systems travelling together with MGE.

To test the linkage hypothesis, we first verified that defense systems tend to be co-transferred with MGE. We used ancestral reconstruction methods to identify the branches in which defense systems and MGE were gained and lost along strain trees (Table S5). We found that >95% of defense acquisition events occurred in branches in which MGE were also gained, and >90% of defense losses occurred in branches in which MGE were also lost (Fig. 4a). The random expectation given the rates of MGE gain and loss would be 71% and 63%, respectively (deviations with respect to these expectations are statistically significant with $p < 10^{-8}$, binomial test). We also quantified the effect of MGE gain and loss on the per-branch probability of gaining or losing defense systems. The probability of acquiring a defense system is around 50-fold higher in branches in which MGE are gained than in other branches (Fig. 4b-c; Fisher exact test $p < 10^{-8}$ for all defense systems). Similarly, the probability of losing a defense system is 10- to 50-fold higher if MGE are also lost in the same branch (Fig. 4b-c; Fisher exact test $p < 10^{-20}$ for all defense systems). These trends are observed even in very short branches, spanning an evolutionary time of 10^{-6} substitutions per site in core genes (Fig. S5). To rule out the possibility that these associations were due to the presence of incomplete genomes, we separately considered terminal and internal tree branches. Because the dataset only includes complete and nearly complete genomes, the absence of a small number of missing genes, if relevant, would only affect the inference of gene gain and loss in terminal branches (those leading from the

163 immediate ancestor to each incomplete genome). Despite some quantitative differences, strong co-
164 acquisition (and co-loss) of MGE and defense occurs in both terminal and internal branches (Fig. S5),
165 confirming that these associations are genuine.

166 Although co-gain of MGE and defense systems along the same tree branch does not necessarily imply
167 a single event of joint gain, the strength of the association, the extremely short timespans, and the
168 fact that similar trends are observed for gene losses strongly suggest that concurrent gain and loss of
169 MGE and defense systems involve genetic linkage. (Alternative explanations in terms of strong
170 selective pressure to quickly acquire defense mechanisms upon exposure to MGE and lose them
171 once the MGE disappear might produce similar trends at the population level but not at the single-
172 genome level, whereas episodic increases in the overall rate of HGT leading to separate but
173 correlated acquisition of MGE and defense systems would not explain correlated losses.)

174 A major consequence of the co-transfer of MGE and defense systems is that the negative effects of
175 defense systems on HGT should be easier to detect at longer timescales or under evolutionary
176 conditions that weakened genetic linkage. To test that prediction, we studied how relative
177 differences in the rates of gene gain between DEF^+ and DEF^- clades depend on the depth of their last
178 common ancestor in the strain tree (note that the depth of the last common ancestor serves as an
179 upper limit for the time that the defense system has been retained in a lineage). As expected,
180 positive outliers (with much higher gene gain rate in the DEF^+ clade than in the DEF^- clade) almost
181 invariably correspond to very recent lineages (less than 10^{-5} substitutions per bp in nearly universal
182 core genes), indicating a very recent acquisition of the defense system (Fig. 5a and Table S6). In
183 contrast, negative outliers (with much lower gene gain rate in the DEF^+ clade than in the DEF^- clade)
184 generally correspond to deeper lineages (at least 0.001 substitutions per bp). We quantified these
185 trends by calculating the skewness of the distribution of DEF^+/DEF^- log-transformed gain ratios in
186 very recent and older lineages (Fig. 5b and Table S7). All the distributions show significant positive
187 skewness in recent lineages and significant negative skewness in older lineages (all $p < 0.001$ except

188 recent RM lineages, with $p = 0.023$; d'Agostino test for skewness). This quantitative analysis confirms
189 that differences in timescale affect all defense systems, with positive associations between defense
190 systems and gene gain being dominant in the short term and negative associations becoming more
191 frequent in the long term.

192 A second prediction of the linkage hypothesis is that the net effect of defense systems on HGT is not
193 only species specific, but also lineage specific. That is, defense systems may display positive, zero, or
194 negative association with HGT in different lineages depending on when the defense system was
195 acquired and how tight is the linkage to MGE. A more detailed study of gene gain rates in individual
196 species confirms that the effect of CRISPR-Cas is, indeed, lineage-specific, with the same species
197 encompassing recent CRISPR-Cas$^+$ lineages that display increased gene gain rates and older CRISPR-
198 Cas$^+$ lineages with reduced gene gain rates (Fig. 5c). This observation, combined with the very recent
199 acquisition of CRISPR-Cas in most lineages, could explain why many species show non-significant or
200 positive associations between CRISPR-Cas and MGE abundances. Indeed, of the six representative
201 species shown in Fig. 5c, only *Streptococcus anginosus* and *Acinetobacter radioresistens* show a net
202 negative association between CRISPR-Cas and MGE abundance in the PGLMM analysis.

203 The findings described so far indicate that, of the seven defense systems considered in this study,
204 CRISPR-Cas is the one that most often induces a net reduction in HGT rates. According to the linkage
205 hypothesis, this could be due to a comparatively weaker physical association between CRISPR-Cas
206 and MGE. To evaluate that possibility, we calculated the percentage of genes from CRISPR-Cas
207 located inside MGE and compared that to other defense systems (Fig. 5d). Consistent with the
208 linkage hypothesis, CRISPR-Cas is the defense system that is least often encoded by MGE (4.97% vs
209 8.12%, $p < 10^{-8}$, chi-squared test), followed by Gabija and DMS.

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

211 Anti-CRISPR proteins (Acr) have the potential to suppress the possible negative effect of CRISPR-Cas
212 on MGE-driven gene transfer (Mahendra et al. 2020). This, in turn, could contribute to explaining the

sign of the association between CRISPR-Cas and MGE in different lineages. To assess that possibility, we searched all the genomes in the dataset for known Acr, finding 16,058 proteins. Then, we compared the prevalence of Acr in species in which the presence of CRISPR-Cas is positively or negatively correlated with the MGE content, separately considering genomes that do and do not harbor CRISPR-Cas. Because Acr are generally encoded by MGE (Pinilla-Redondo et al. 2020) and the prevalence of MGE systematically varies across groups acting as a confounding factor, we restricted our comparisons to genomes that contain at least one prophage. Our results indicate that genomes with and without CRISPR-Cas differ in the prevalence of Acr (Fig. S6). More importantly, these differences have opposite directions in species in which CRISPR-Cas is positively and negatively associated with MGE abundance. In the former, Acr are more prevalent in genomes that contain CRISPR-Cas (p = 0.0012, chi-squared test). In the latter, Acr are less prevalent in genomes with CRISPR-Cas ($p < 10^{-6}$, chi-squared test). In both groups, the prevalence of Acr in genomes without CRISPR-Cas is similar. These results suggest that negative associations between CRISPR-Cas and HGT could be dependent on (or at least facilitated by) low prevalence of Acr in the genome.

Discussion

Defense systems could have a significant impact on microbial evolution by effectively blocking the transfer of MGE, reducing gene flow and limiting the spread of accessory genes. However, because defense systems are often carried by MGE and these are the main vehicles of HGT, a net positive association between defense systems and gene exchange cannot be ruled out *a priori*. We assessed the relative weight of these two opposite scenarios by quantifying the association between defense systems, MGE abundance, and gene acquisition rates in a phylogeny-aware comparative study of 197 prokaryotic species.

Our results shed light on previous, apparently contradictory findings concerning the effect of CRISPR-Cas on genome evolution and diversification (Gophna et al. 2015; O'Meara and Nunney 2019;

Shehreen et al. 2019; Wheatley and MacLean 2021; Pursey et al. 2022). A pioneering study conducted in 2015 found no evidence to support an overall association between CRISPR-Cas activity and reduced gene acquisition via HGT at evolutionary time scales (Gophna et al. 2015). Such lack of association was explained by several factors, including the high mobility of CRISPR-Cas systems, that limits their long-term impact on host genomes, and the possibility that HGT is mediated by MGE that escape (or are not targeted by) CRISPR-Cas immunity.

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

Besides this general picture, we identified some species in which the presence of defense systems (especially CRISPR-Cas) significantly correlates with smaller genome sizes and MGE abundances. Some of those associations had been previously described in *P. aeruginosa* and *Klebsiella pneumoniae* (Wheatley and MacLean 2021; Botelho et al. 2023). And yet, these species represent special cases rather than the rule, even in the context of host-associated bacteria. In fact, our results underline that the association between defense systems and HGT is strongly system- and lineage-dependent. This conclusion confirms and extends previous findings that showed that the impact of CRISPR-Cas on the spread of antibiotic resistance is highly variable across species and its sign cannot be easily explained by simple ecological, environmental, or genomic variables (Shehreen et al. 2019).

Among the 7 defense systems included in this study, CRISPR-Cas stands out for being the most recurrently associated with reduced genome sizes and lower MGE (especially prophage) abundances. In contrast, DMS, Abi, DRT, and CBASS are more often associated with higher numbers of MGE and accessory genes. This finding is fully consistent with a recent study that compared 73 defense systems in 12 bacterial species (Kogay et al. 2024). We propose that what makes CRISPR-Cas systems different is their weaker (though still substantial) linkage with MGE. Compared to fully functional CRISPR-Cas systems, other defense systems like Abi, Gabija, CBASS, DMS, and RM are more frequently located within or next to MGE (Makarova et al. 2011; Benler et al. 2021; Rousset et al. 2022; Botelho 2023) and may have alternative functions related to MGE propagation. For example, Abi systems have been identified in PICIs as accessory genes that facilitate their parasitic lifecycle (Ibarra-Chavez et al. 2021). The narrow specificity of some defense systems may be another reason why those systems do not significantly interfere with HGT. For instance, the GmrSD type IV RM system selectively targets phages with glucosylated hydroxymethylcytosine (Bair and Black 2007) and the Thoeris defense system only appears to be effective against myoviruses (Doron et al. 2018). This caveat extends to any defense system based on epigenetic modifications, such as RM and DMS, whose overall effect on HGT depend on the repertoire of epigenetic markers in the host and MGE populations (Oliveira et al. 2016). Although selective targeting is often viewed as an outcome of phage-host coevolution, it is tempting to speculate that it could have been evolutionarily favored by the need to fight harmful genetic parasites while maintaining sufficiently high rates of HGT to prevent population-level gene loss (Iranzo et al. 2016).

Defense systems sometimes exhibit synergistic interactions (Dupuis et al. 2013; Wu et al. 2024), which could contribute to the heterogeneity of effects reported in this study. The vast number of potential interactions and the limited availability of high-quality genomes made it unfeasible to systematically account for the effect of interactions with the methods developed in this work. Determining if and to what extent synergy among defense systems affects HGT remains a subject for future investigation, possibly focused on a small set of experimentally validated interactions in highly

sequenced species. Another open question concerns which levels of detail, both taxonomic and functional, best capture the effect of defense systems on HGT. From a taxonomic perspective, working at or below the species level is a natural choice because species represent genetically cohesive units (Bobay and Ochman 2017; Konstantinidis 2023; Conrad et al. 2024) and, as a result, uncontrolled confounding factors are less likely to affect within-species than cross-species comparisons. In contrast, more complex multi-level approaches would be required to detect trends at higher taxonomic ranks. From a functional perspective, we grouped defense systems based on their mechanism of action, under the assumption that functionally similar systems produce similar effects on HGT. Though reasonable, this grouping criterion may not be optimal in systems in which subtypes markedly differ in their eco-evolutionary dynamics and linkage with MGE. Moreover, fine-grain dissection of highly abundant systems, such as RM and DMS, could help improve the sensitivity of statistical tests by producing more balanced sets of strains with and without the subtypes of interest.

All in all, we showed that some defense systems, especially CRISPR-Cas, can significantly reduce HGT, although the effect is often masked by the fact that these systems travel together with MGE. Beyond possible functional connections, the linkage between defense systems and MGE is an inevitable consequence of the arms race between parasites and hosts. Because defense systems are costly and their efficacy drops as parasites evolve, they are subject to rapid turnover and depend on HGT for long-term persistence in microbial populations (van Houte et al. 2016; Iranzo et al. 2017; Koonin et al. 2017; Puigbo et al. 2017). As a result, it is extremely challenging to disentangle the impact of defense systems on gene flow from the causes that lead to their presence or absence, especially at short evolutionary time scales. As more and more genomic data become available, we expect that future research will overcome this challenge by quantifying the linkage between defense systems and MGE, developing more realistic null models, and testing the role of defense systems on microbial adaptation at different time scales.

Methods

Genome collection and identification of defense systems

We parsed the Genome Taxonomy Database (GTDB, https://gtdb.ecogenomic.org) release 202 (Parks et al. 2020) to identify all high-quality genomes (according to the MIMAG criteria (Bowers et al. 2017)) with completeness >99%, contamination <1%, and contig count <500. The 82,595 genomes that passed these filters were downloaded from the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov). CRISPR-Cas systems were identified with CRISPRCasTyper v1.2.4 (Russel et al. 2020) using default parameters. We classified a genome as CRISPR-Cas$^+$ if it contains at least one high-confidence Cas operon and CRISPR-Cas$^-$ otherwise. Only the species with >10 genomes and at least 5 CRISPR-Cas$^+$ genomes were further considered. To reduce the computational cost, we only considered a maximum of 500 genomes per species. Species with >500 genomes were randomly subsampled to keep at most 350 CRISPR-Cas$^+$ and 150 CRISPR-Cas$^-$ genomes. Other defense systems were identified with Padloc v1.1.0 (db v1.4.0) (Payne et al. 2022). Our analysis focused on the most prevalent defense systems: restriction-modification (RM), DMS, Abi, CRISPR-Cas, Gabija, DRT, and CBASS.

After applying these criteria, 19,323 genomes belonging to 196 bacterial and 1 archaeal species (*sensu* GTDB) were included in the analysis (Tables S1 and S2). Of those, 2,964 correspond to complete genomes and the rest to high-quality, nearly complete ones.

Gene prediction and annotation.

Open reading frames (ORF) were predicted with Prodigal v2.6.3, using codon table 11 (prokaryotic genetic code) and "single" mode, as recommended for finished and draft quality genomes (Hyatt et al. 2010). Orthologous ORF were then separately clustered for each species with Roary v3.13.0 (Page et al. 2015) setting an 80% identity threshold for initial clustering followed by synteny-based refinement (options '-t 11 -i 80'). The resulting gene clusters were functionally annotated by selecting

a representative sequence, arbitrarily chosen among those with length between 0.95 and 1.05 times the average length of all sequences in the cluster. Representative sequences were functionally annotated by mapping them to in-house profiles of the Clusters of Orthologous Genes (COG) database (2020 release) (Galperin et al. 2021) with HMMER v3.1b2 (e-value < 0.001) (http://hmmer.org). The 26 major prokaryotic functional categories defined in the COG database were assigned to the annotated genes. Some functional categories (A, RNA processing and modification; B, chromatin structure and dynamics; W, extracellular structures; T, signal transduction; and Z, cytoskeleton) were excluded since they rarely or never occur in prokaryotic genomes. The case-insensitive keywords "phage", "plasmid" and "transpos*" in the COG gene annotations were used to identify genes associated with prophages, plasmids, and transposons, respectively, and the resulting gene lists were manually curated to minimize false assignments. We based our statistical analyses on marker gene counts rather than full MGE counts because the latter are more susceptible to technical artifacts (e.g., different heuristics to deal with nested MGE will affect the number of MGE, but not the number of MGE marker genes). Gene counts per genome and functional category are listed in Table S8. Anti-defense proteins, including anti-CRISPR, were identified by running HMMER v3.1b2 (e-value < 10^{-10}) against the dbAPIS database (Yan et al. 2024).

Identification of genomic regions containing mobile genetic elements

Prophages were detected with Phispy v4.2.21 using default options (Akhter et al. 2012). Short transposons were identified based on the presence of isolated or paired genes annotated as transposases. In the latter case, we allowed for up to one additional gene between two transposon-related genes to account for the genetic architecture of some insertion sequences (Gomez et al. 2014). Due to the inherent difficulty to discriminate among plasmids, ICE, and IME (Botelho et al. 2023), we restricted the search for these elements to complete genomes. Plasmids were identified as extra-chromosomal replicons that contain the "plasmid" label in their NCBI description lines. After removing those replicons, ICEfinder (Liu et al. 2019) was used to identify ICE and IME. Finally, we ran

Phispy v4.2.21 with option '-phage_genes 0' to identify any other MGE, integrons, pathogenicity islands, and fragments of MGE that could have been missed by the previous approaches. The MGE identified through these approaches were masked from complete genomes to produce the gene counts in Table S9 and the results shown in Fig S2.

Species trees

Phylogenetic trees were separately built for each species based on the set of 120 prokaryotic marker genes (122 in the case of Archaea) proposed by the GTDB r202. For each species, only those marker genes with prevalence >80% were used for phylogenetic reconstruction. We aligned the amino acid sequences of each marker gene with mafft-linsi (L-INS-I algorithm, default options, MAFFT v7.475) (Katoh and Standley 2013) and back-translated the amino acid alignments to nucleotide alignments with pal2nal.pl v14 (Suyama et al. 2006) using codon table 11. After concatenating all nucleotide alignments, we built preliminary trees with FastTree v2.1.10 (options '-gtr -nt -gamma -nosupport -mlacc 2 -slownni') (Price et al. 2010). The tree topologies produced by FastTree were subsequently provided to RaxML v8.2.12 (Stamatakis 2014) for branch length optimization (raxmlHPC with options '-f e -m GTRGAMMA'). The final trees are included in Supplementary File S1.

To visualize trends across species (Fig. 2, S3, and S4), we used the online tool iTOL (Letunic and Bork 2021) and the multispecies tree from GTDB r202.

Phylogenetic generalized linear mixed models (PGLMM)

For each genome in the dataset, we collected the following response variables: the total number of genes, the number of genes belonging to each functional category, and the number of genes associated with prophages, plasmids, and transposons (Table S8). Genes that belong to the 7 defense systems of interest were excluded when computing these values. Then, for each response variable, we fitted a PGLMM with Poisson distribution and canonical link function, using the presence or absence of each defense system as predictors and the species trees as guides to generate the covariance matrix.

For each species, the PGLMM assumes that the response variable, Y_i, follows a Poisson distribution with mean μ_i, that is, $Y_i \sim \mathrm{Poisson}(\mu_i)$. The expected gene abundance, μ_i, is modeled as $\log \mu_i = \beta_0 + \sum \beta_j X_{ij} + \epsilon_i$, where β_0 is the intercept, β_j is the coefficient associated with the defense system j, and $X_{ij} \in \{0,1\}$ denotes the absence or presence of defense system j in genome i. The random effects ϵ_i follow a multivariate normal distribution, $\epsilon \sim \mathrm{Gaussian}(0, \sigma^2_{phy}\mathbf{C})$, where σ^2_{phy} is the strength of the phylogenetic signal and $\mathbf{C}$ is a covariance matrix derived from the phylogenetic tree under the assumption of Brownian motion evolution.

To extend the PGLMM to multiple species, we considered that the effect of defense systems on gene content can be species-dependent (this was a reasonable assumption *a priori* and later confirmed by the analysis). To account for that, the multi-species model must include an interaction term "defense × species", whose coefficients are relevant per se, and, accordingly, modeled as fixed effects. Moreover, because HGT rates are highly variable among species (Puigbò et al. 2014; Iranzo et al. 2019), it makes little sense to extend the phylogenetic correction beyond single species or assume that the strength of the phylogenetic signal is the same for all species. These considerations are captured by a phylogenetic covariance matrix with block-diagonal structure (one block per species), and species-wise values of the phylogenetic coefficient (one for each block of the phylogenetic covariance matrix). In practice, fitting a multi-species model with these specifications is formally equivalent to fitting independent models for each species. The latter approach, that we adopted, has the advantage of being more suitable for parallelization and requiring fewer computational resources. Thus, for each species and response variable, we fitted a PGLMM with the function pglmm_compare(response_variable ~ Abi + CBASS + CRISPR + DMS + DRT + Gabija + RM, family = "poisson", data = SpData, phy = SpTree) from the R package phyr v1.1.0 (Li et al. 2020). In the formula, Abi, CBASS, CRISPR, and so on, are binary variables representing the presence (1) or absence (0) of each defense system. Table S4 presents the coefficients, p-values, and goodness of fit of the model.

412 The model described above includes nine coefficients (intercept, seven defense systems, and the
413 phylogenetic signal), which could lead to overfitting in species in which the number of available
414 genomes is limited. As an alternative, we also fitted seven separate PGLMM, one for each defense
415 system, involving a single predictor and the phylogenetic random effect (Table S3). This approach
416 does not account for correlations among defense systems, but it has the advantage of not being
417 affected by overfitting. The figures in the manuscript are based on this second set of models,
418 although, in practice, both approaches produce very similar quantitative results.

419 For each class of PGLMM, we also fitted non-phylogenetic models in which the covariance matrix $\mathbf{C}$ of
420 the random effect was replaced by the identity matrix. The PGLMM were compared to their non-
421 phylogenetic counterparts using the conditional Akaike Information Criterion (cAIC) as previously
422 described (Greven and Kneib 2010; Säfken et al. 2021). Based on the cAIC, PGLMM performed better
423 than non-phylogenetic models in 65% of the species-response-defense triplets and were close to
424 non-phylogenetic models (ΔcAIC < 2) in another 25% of the triplets (in most of those cases, the
425 strength of the phylogenetic signal was close to zero, which made both phylogenetic and non-
426 phylogenetic models equivalent).

427 To account for the possibility that other (less abundant) defense systems could explain part of the
428 variability in the results, we explored a more complex set of models that included the total number
429 of other defense systems as an additional predictor. These models generally performed worse than
430 their simpler variants (ΔcAIC > 0 in 83% of the species) and were not considered for further analysis.

431 Large differences in sample size among species and defense systems translate into unequal precision
432 in the estimation of the model coefficients. To deal with that limitation, we used two different
433 criteria to identify species in which the presence of a defense system is positively or negatively
434 associated with gene numbers. For one option, we adopted a classical criterion of statistical
435 significance ($p < 0.05$) for the predictor variable in the PGLMM. For the other option, we applied an
436 alternative criterion based on effect size, aimed at comparing species with different sample sizes in

which p-values are not commensurable. Specifically, for each variable and defense system, we jointly considered the PGLMM of all the species and determined the smallest effect size (in absolute value) that reached statistical significance ($p < 0.05$) in any species. Then we used the smallest significant effect size (SSES) as a threshold to classify associations as positive, negative, or null.

Inference of gene gain and loss

Gene gains and losses at each branch of each species tree were estimated with Gloome (Cohen and Pupko 2010), using as inputs the gene presence/absence matrices previously generated by Roary and the species trees. The parameter configuration file was set to optimize the likelihood of the observed phyletic profiles under a genome evolution model with 4 categories of gamma-distributed gain and loss rates and stationary frequencies at the root.

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

For each species tree, we defined DEF^+ clades as the narrowest possible clades such that at least 80% of the leaves contain the defense system of interest. Candidate DEF^- clades were defined in an analogous way, referring to leaves without the defense system. Next, we identified pairs of DEF^+ and DEF^- clades that constitute sister groups. Sister $DEF^{+/-}$ pairs were excluded if both clades contained a single genome. For each clade in a valid pair, we computed the overall gene gain and loss rates as the expected number of gene gains (or losses) in that clade divided by the total branch length. Gain and loss rates for different functional categories and MGE we calculated in an analogous way but restricting the sum of gene gains and losses to the genes of interest. When calculating overall and category-wise gain and loss rates, we did not take into account the contribution of species-wise singletons (genes without homologs in other genomes of the same species), as they may represent false gene predictions or genes that are replaced at unusually high rates (Wolf et al. 2016). To account for the non-negative nature of gain and loss rates and their heavy-tailed distributions, comparisons between DEF^+ and DEF^- sister branches were done based on log-transformed rate estimates. To calculate the skewness of the distributions and their statistical significance, we use the

462 method proposed by D'Agostino (D'Agostino et al. 1990) as implemented by the
463 scipy.stats.skewtest() function in Python (Virtanen et al. 2020).

464 **Data access**

465 All the data generated or analyzed during this study are included in this published article and its
466 supplementary information files.

467 **Competing interests**

468 The authors declare no competing financial interests.

469 **Acknowledgements**

470 Y.L. is supported by China Scholarship Council (No.202008440425). J.B. is supported by the Maria
471 Zambrano grant of the Spanish Ministry of Universities (Grant No. UP2021-035), and the Severo
472 Ochoa Program for Centres of Excellence in R&D of the Agencia Estatal de Investigación of Spain
473 (Grant No. CEX2020-000999-S (2022–2025) to the CBGP). J.I is supported by the Ramón y Cajal
474 Programme of the Spanish Ministry of Science (Grant No. RYC-2017–22524); the Agencia Estatal de
475 Investigación of Spain (Grant Nos. PID2019-106618GA-I00 and CNS2023-145430), the Severo Ochoa
476 Programme for Centres of Excellence in R&D of the Agencia Estatal de Investigación of Spain (Grant
477 No. SEV-2016–0672 (2017–2021) to the CBGP); and the Comunidad de Madrid (through the call
478 Research Grants for Young Investigators from Universidad Politécnica de Madrid, Grant No.
479 M190020074JIIS).

480 We thank Jorge Calle-Espinosa for helpful discussions.

481

482 **References**

483 Akhter S, Aziz RK, Edwards RA. 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that
484 combines similarity- and composition-based strategies. *Nucleic acids research* **40**: e126.

485 Arnold BJ, Huang IT, Hanage WP. 2022. Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev*
486 *Microbiol* **20**: 206-218.
487 Bair CL, Black LW. 2007. A type IV modification dependent restriction nuclease that targets glucosylated
488 hydroxymethyl cytosine modified DNAs. *J Mol Biol* **366**: 768-778.
489 Benler S, Faure G, Altae-Tran H, Shmakov S, Zheng F, Koonin E. 2021. Cargo Genes of Tn7-Like Transposons
490 Comprise an Enormous Diversity of Defense Systems, Mobile Genetic Elements, and Antibiotic
491 Resistance Genes. *mBio* **12**: e0293821.
492 Bobay LM, Ochman H. 2017. Biological species are universal across Life's domains. *Genome biology and*
493 *evolution* **9**: 491-501.
494 Botelho J. 2023. Defense systems are pervasive across chromosomally integrated mobile genetic elements and
495 are inversely correlated to virulence and antimicrobial resistance. *Nucleic acids research* **51**: 4385-
496 4397.
497 Botelho J, Cazares A, Schulenburg H. 2023. The ESKAPE mobilome contributes to the spread of antimicrobial
498 resistance and CRISPR-mediated conflict between mobile genetic elements. *Nucleic acids research* **51**:
499 236-252.
500 Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR,
501 Eloe-Fadrosh EA et al. 2017. Minimum information about a single amplified genome (MISAG) and a
502 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature biotechnology* **35**: 725-
503 731.
504 Cohen O, Pupko T. 2010. Inference and characterization of horizontally transferred gene families using
505 stochastic mapping. *Molecular biology and evolution* **27**: 703-713.
506 Conrad RE, Brink CE, Viver T, Rodriguez-R LM, Aldeguer-Riquelme B, Hatt JK, Venter SN, Amann R, Rossello-
507 Mora R, Konstantinidis KT. 2024. Microbial species exist and are maintained by ecological
508 cohesiveness coupled to high homologous recombination. *bioRxiv* doi:10.1101/2024.05.25.595874:
509 2024.2005.2025.595874.
510 Costa AR, van den Berg DF, Esser JQ, Muralidharan A, van den Bossche H, Bonilla BE, van der Steen BA,
511 Haagsma AC, Fluit AC, Nobrega FL et al. 2024. Accumulation of defense systems in phage-resistant
512 strains of Pseudomonas aeruginosa. *Sci Adv* **10**: eadj0341.

513 D'Agostino RB, Belanger A, D'Agostino Jr RB. 1990. A Suggestion for Using Powerful and Informative Tests of
514 Normality. *The American Statistician* **44**: 316-321.
515 Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018. Systematic discovery of
516 antiphage defense systems in the microbial pangenome. *Science* **359**.
517 Dupuis ME, Villion M, Magadan AH, Moineau S. 2013. CRISPR-Cas and restriction-modification systems are
518 compatible and increase phage resistance. *Nature communications* **4**: 2087.
519 Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. 2021. COG database update: focus
520 on microbial diversity, model organisms, and widespread pathogens. *Nucleic acids research* **49**: D274-
521 D281.
522 Georjon H, Bernheim A. 2023. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol* **21**:
523 686-700.
524 Gomez MJ, Diaz-Maldonado H, Gonzalez-Tortuero E, Lopez de Saro FJ. 2014. Chromosomal replication
525 dynamics and interaction with the beta sliding clamp determine orientation of bacterial transposable
526 elements. *Genome biology and evolution* **6**: 727-740.
527 Gophna U, Kristensen DM, Wolf YI, Popa O, Drevet C, Koonin EV. 2015. No evidence of inhibition of horizontal
528 gene transfer by CRISPR-Cas on evolutionary timescales. *The ISME journal* **9**: 2021-2027.
529 Greven S, Kneib T. 2010. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*
530 **97**: 773-789.
531 Hall RJ, Whelan FJ, McInerney JO, Ou Y, Domingo-Sananes MR. 2020. Horizontal Gene Transfer as a Source of
532 Conflict and Cooperation in Prokaryotes. *Front Microbiol* **11**: 1569.
533 Haudiquet M, de Sousa JM, Touchon M, Rocha EPC. 2022. Selfish, promiscuous and sometimes useful: how
534 mobile genetic elements drive horizontal gene transfer in microbial populations. *Philos Trans R Soc*
535 *Lond B Biol Sci* **377**: 20210234.
536 Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition
537 and translation initiation site identification. *BMC Bioinformatics* **11**: 119.
538 Ibarra-Chavez R, Hansen MF, Pinilla-Redondo R, Seed KD, Trivedi U. 2021. Phage satellites and their emerging
539 applications in biotechnology. *FEMS Microbiol Rev* **45**.

Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. 2017. Disentangling the effects of selection and loss bias on gene dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **114**: E5616-E5624.

Iranzo J, Puigbo P, Lobkovsky AE, Wolf YI, Koonin EV. 2016. Inevitability of Genetic Parasites. *Genome biology and evolution* **8**: 2856-2869.

Iranzo J, Wolf YI, Koonin EV, Sela I. 2019. Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. *Nature communications* **10**: 5376.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**: 772-780.

Kogay R, Wolf YI, Koonin EV. 2024. Defense systems and horizontal gene transfer in bacteria. *bioRxiv* doi:10.1101/2024.02.09.579689: 2024.2002.2009.579689.

Konstantinidis KT. 2023. Sequence-discrete species for prokaryotes and other microbes: A historical perspective and pending issues. *mLife* **2**: 341-349.

Koonin EV, Makarova KS, Wolf YI. 2017. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annual review of microbiology* **71**: 233-261.

Koonin EV, Makarova KS, Wolf YI, Krupovic M. 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nature reviews Genetics* **21**: 119-131.

Lee IPA, Eldakar OT, Gogarten JP, Andam CP. 2022. Bacterial cooperation through horizontal gene transfer. *Trends Ecol Evol* **37**: 223-232.

Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research* **49**: W293-W296.

Li D, Dinnage R, Nell LA, Helmus MR, Ives AR. 2020. phyr: An r package for phylogenetic species-distribution modelling in ecological communities. *Methods in Ecology and Evolution* **11**: 1455-1463.

Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou HY. 2019. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic acids research* **47**: D660-D665.

Mahendra C, Christie KA, Osuna BA, Pinilla-Redondo R, Kleinstiver BP, Bondy-Denomy J. 2020. Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. *Nature microbiology* **5**: 620-629.

567 Makarova KS, Wolf YI, Snir S, Koonin EV. 2011. Defense islands in bacterial and archaeal genomes and
568 prediction of novel defense systems. *Journal of bacteriology* **193**: 6039-6056.
569 Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by
570 targeting DNA. *Science* **322**: 1843-1845.
571 Mayo-Munoz D, Pinilla-Redondo R, Birkholz N, Fineran PC. 2023. A host of armor: Prokaryotic immune
572 strategies against mobile genetic elements. *Cell Rep* **42**: 112672.
573 Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, Fineran PC. 2022. High viral abundance
574 and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems.
575 *Curr Biol* **32**: 220-227 e225.
576 Morris JA, Gardner MJ. 1988. Calculating confidence intervals for relative risks (odds ratios) and standardised
577 ratios and rates. *Br Med J (Clin Res Ed)* **296**: 1313-1316.
578 O'Hara BJ, Barth ZK, McKitterick AC, Seed KD. 2017. A highly specific phage defense system is a conserved
579 feature of the Vibrio cholerae mobilome. *PLoS genetics* **13**: e1006838.
580 O'Meara D, Nunney L. 2019. A phylogenetic test of the role of CRISPR-Cas in limiting plasmid acquisition and
581 prophage integration in bacteria. *Plasmid* **104**: 102418.
582 Oliveira PH, Touchon M, Rocha EP. 2016. Regulation of genetic flux between bacteria by restriction-
583 modification systems. *Proceedings of the National Academy of Sciences of the United States of*
584 *America* **113**: 5658-5663.
585 Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015.
586 Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**: 3691-3693.
587 Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species
588 taxonomy for Bacteria and Archaea. *Nature biotechnology* **38**: 1079-1086.
589 Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, Jackson SA. 2022. PADLOC: a web server for the
590 identification of antiviral defence systems in microbial genomes. *Nucleic acids research* **50**: W541-
591 W550.
592 Pinilla-Redondo R, Russel J, Mayo-Munoz D, Shah SA, Garrett RA, Nesme J, Madsen JS, Fineran PC, Sorensen SJ.
593 2022. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids.
594 *Nucleic acids research* **50**: 4315-4328.

595 Pinilla-Redondo R, Shehreen S, Marino ND, Fagerlund RD, Brown CM, Sorensen SJ, Fineran PC, Bondy-Denomy
596 J. 2020. Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic
597 elements. *Nature communications* **11**: 5652.
598 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments.
599 *PLoS One* **5**: e9490.
600 Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of
601 genome dynamics in prokaryote supergenomes. *BMC biology* **12**: 66.
602 Puigbo P, Makarova KS, Kristensen DM, Wolf YI, Koonin EV. 2017. Reconstruction of the evolution of microbial
603 defense systems. *BMC evolutionary biology* **17**: 94.
604 Pursey E, Dimitriu T, Paganelli FL, Westra ER, van Houte S. 2022. CRISPR-Cas is associated with fewer antibiotic
605 resistance genes in bacterial pathogens. *Philos Trans R Soc Lond B Biol Sci* **377**: 20200464.
606 Rocha EPC, Bikard D. 2022. Microbial defenses against mobile genetic elements and viruses: Who defends
607 whom from what? *PLoS Biol* **20**: e3001514.
608 Rousset F, Depardieu F, Miele S, Dowding J, Laval AL, Lieberman E, Garry D, Rocha EPC, Bernheim A, Bikard D.
609 2022. Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe* **30**: 740-753
610 e745.
611 Russel J, Pinilla-Redondo R, Mayo-Munoz D, Shah SA, Sorensen SJ. 2020. CRISPRCasTyper: Automated
612 Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J* **3**: 462-469.
613 Säfken B, Rügamer D, Kneib T, Greven S. 2021. Conditional Model Selection in Mixed-Effects Models with
614 cAIC4. *Journal of Statistical Software* **99**: 1 - 30.
615 Shaw LP, Rocha EPC, MacLean RC. 2023. Restriction-modification systems have shaped the evolution and
616 distribution of plasmids across bacteria. *Nucleic acids research* **51**: 6806-6818.
617 Shehreen S, Chyou TY, Fineran PC, Brown CM. 2019. Genome-wide correlation analysis suggests different roles
618 of CRISPR-Cas systems in the acquisition of antibiotic resistance genes in diverse species. *Philos Trans*
619 *R Soc Lond B Biol Sci* **374**: 20180384.
620 Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nature reviews*
621 *Genetics* **16**: 472-482.

622 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.
623 *Bioinformatics* **30**: 1312-1313.
624 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the
625 corresponding codon alignments. *Nucleic acids research* **34**: W609-612.
626 Tesson F, Herve A, Mordret E, Touchon M, d'Humieres C, Cury J, Bernheim A. 2022. Systematic and quantitative
627 view of the antiviral arsenal of prokaryotes. *Nature communications* **13**: 2561.
628 van Houte S, Buckling A, Westra ER. 2016. Evolutionary Ecology of Prokaryotic Immune Mechanisms.
629 *Microbiology and molecular biology reviews : MMBR* **80**: 745-763.
630 van Vliet AHM, Charity OJ, Reuter M. 2021. A Campylobacter integrative and conjugative element with a
631 CRISPR-Cas9 system targeting competing plasmids: a history of plasmid warfare? *Microb Genom* **7**.
632 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser
633 W, Bright J et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat*
634 *Methods* **17**: 261-272.
635 Watson BNJ, Capria L, Alseth EO, Pons B, Biswas A, Lenzi L, Buckling A, van Houte S, Westra ER, Meaden S.
636 2024. CRISPR-Cas in Pseudomonas aeruginosa provides transient population-level immunity against
637 high phage exposures. *The ISME journal* doi:10.1093/ismejo/wrad039.
638 Watson BNJ, Staals RHJ, Fineran PC. 2018. CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene
639 Transfer by Transduction. *mBio* **9**.
640 Westra ER, Levin BR. 2020. It is unclear how important CRISPR-Cas systems are for protecting natural
641 populations of bacteria against infections by mobile genetic elements. *Proceedings of the National*
642 *Academy of Sciences of the United States of America* **117**: 27777-27785.
643 Wheatley RM, MacLean RC. 2021. CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas
644 aeruginosa. *The ISME journal* **15**: 1420-1433.
645 Wolf YI, Makarova KS, Lobkovsky AE, Koonin EV. 2016. Two fundamentally different classes of microbial genes.
646 *Nature microbiology* **2**: 16208.
647 Wu Y, Garushyants SK, van den Hurk A, Aparicio-Maldonado C, Kushwaha SK, King CM, Ou Y, Todeschini TC,
648 Clokie MRJ, Millard AD et al. 2024. Bacterial defense systems exhibit synergistic anti-phage activity.
649 *Cell Host Microbe* **32**: 557-572 e556.

650 Yan Y, Zheng J, Zhang X, Yin Y. 2024. dbAPIS: a database of anti-prokaryotic immune system genes. *Nucleic*

651 *acids research* **52**: D419-D425.

652

653

Figure legends

Figure 1: Association between 7 widespread defense systems, total number of genes, and MGE abundance. (a) Number of species displaying positive or negative associations in a phylogenetic generalized linear mixed effects model (see Methods) according to two different criteria: $p < 0.05$ and absolute effect size greater than the smallest significant effect size ($|ES| > SSES$), separately computed for each response variable and defense system. The top row (Total) indicates the association with the total number of genes. The association with MGE was calculated based on marker genes for prophages (Ph), plasmids (Pl), and transposons (Tr). Abbreviations of functional categories: X (mobilome), L (replication, recombination and repair), U (intracellular trafficking and secretion), K (transcription), D (cell cycle control, cell division and chromosome partitioning), and V (defense). (b) Effect sizes, measured as relative differences in gene and MGE abundances. Each point corresponds to one species. Species with values beyond the axis limits are collapsed in a single point with size proportional to the number of species. Vertical lines indicate the median over all the species that show positive or negative association according to the p-value and SSES criteria.

Figure 2: Taxonomic distribution of species displaying positive and negative associations between the presence of defense systems and the number of genes from the mobilome (based on COG annotations). Taxa discussed in the text are labeled: A, *Acinetobacter*; Pa, *Pseudomonas aeruginosa*; Pv, *Phocaeicola vulgatus*; Pd, *Phocaeicola dorei*; S1, *Streptococcus pyogenes*, *S. dysgalactiae*, *S. equi*, and *S. uberis*; S2, *S. gordonii*, *S. anginosus*, *S. mutans*, and *S. salivarius*; S3, *S. oralis*, *S. intermedius*, and *S. suis*. The classes *Bacilli* and *Clostridia* are the major components of the Genome Taxonomy Database phyla "Bacillota" and "Bacillota_A". See figure S3 for a larger version including all species names.

Figure 3: Relative differences in the rates of gene gain and loss between sister clades that do and do not harbor defense systems (DEF^+ and DEF^-, respectively). The boxplots represent the distribution of the DEF^+/DEF^- ratio of gene gain (or loss) rates for each class of MGE, calculated for every pair of

679 sister clades. Values greater (or smaller) than 1 indicate increased (or reduced) gene flux in lineages
680 that contain the defense system. (a) Species that show a negative association between the defense
681 system and the number of marker genes for each class of MGE (PGLMM with smallest significant
682 effect size criterion). (b) Species that show a positive association between the defense system and
683 the number of marker genes for each class of MGE. In the boxplots, the central line indicates the
684 median, the box limits correspond to the 25 and 75 percentiles, and the whiskers extend to the
685 largest and smallest values not classified as outliers. P-values are based on Wilcoxon test with log-
686 ratio = 0 as null hypothesis.

687 **Figure 4: Co-occurrence of defense system gains and losses and MGE gains and losses along the**
688 **phylogeny.** (a) Percentage of DEF gains (and losses) that occur in the same branch as an MGE gain (or
689 loss). (b) Conditional probability of gaining (or losing) a defense system provided that an MGE is also
690 gained (or lost) in the same branch, compared to the conditional probabilities when an MGE is not
691 acquired (or lost) in the same branch. (c) Effect of MGE gain (or loss) on the per branch probability to
692 acquire (or loss) a defense system, measured as a risk ratio. Error bars in (a) and (b) correspond to
693 95% confidence intervals based on the binomial distribution. Error bars in (c) indicate the 95%
694 confidence intervals for the relative risk (Morris and Gardner 1988).

695 **Figure 5: Association between defense systems and gene gain rates depends on the time scale.** (a)
696 Ratio of overall gene gain rates between sister clades that do and do not harbor defense systems
697 (DEF^+ and DEF^-, respectively). Values greater (or smaller) than 1 indicate increased (or reduced) gene
698 flux in lineages that contain the defense system. (b) Density distribution of the DEF^+/DEF^- gain ratios,
699 comparing clades that represent recent acquisitions of the defense system (depth < 10^{-5} substitutions
700 per site in core genes) and clades that have retained the defense system for longer times (depth > 10^{-3}
701 substitutions per site). The probability density function (y-axis) is represented in logarithmic scale to
702 facilitate the visualization the tails. The skewness of all distributions is statistically significant (Table
703 S7), with positive values for recent clades and negative values for deeper clades (c) Representative

704 examples of gene gain ratios in shallow and deeper sister groups from the same species. (d)
705 Percentage of genes from different defense systems located within known MGE. Whiskers represent
706 95% confidence intervals based on the binomial distribution.

707

Figure 1

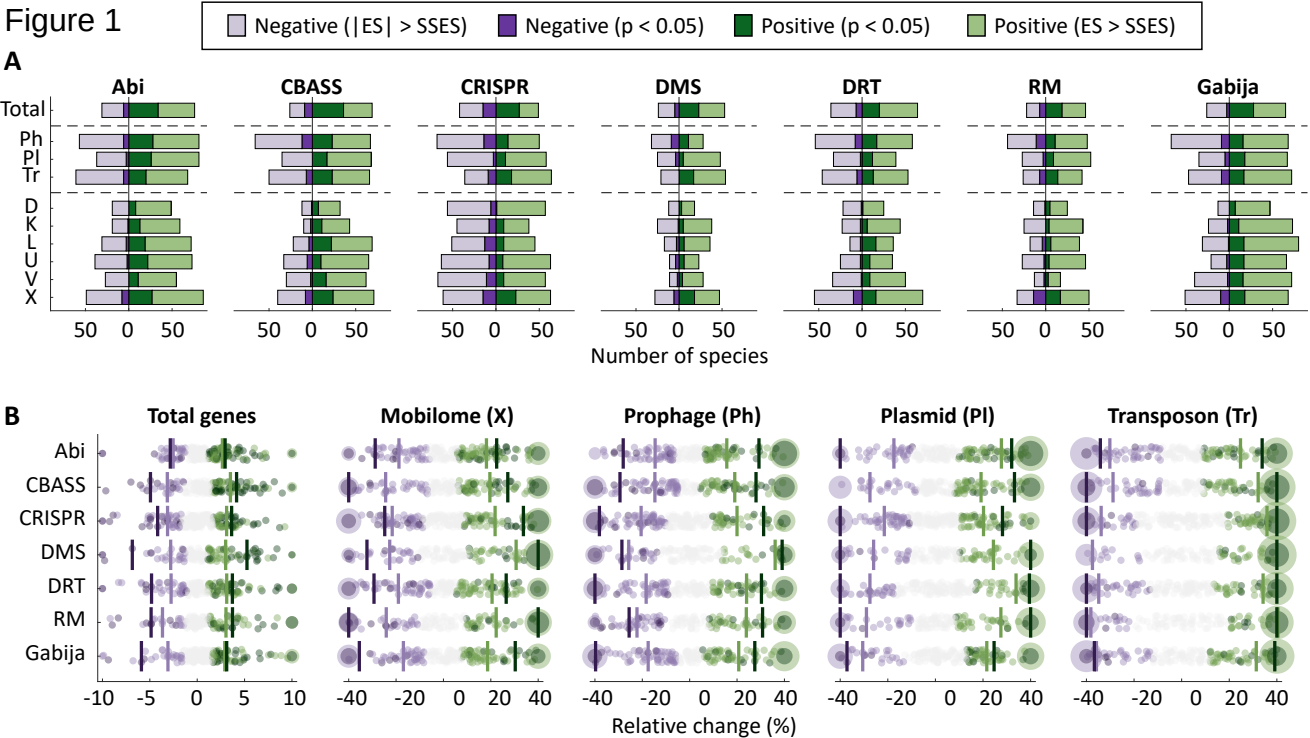

Negative (|ES| > SSES)
Negative (p < 0.05)
Positive (p < 0.05)
Positive (ES > SSES)
A
Abi
CBASS
CRISPR
DMS
DRT
RM
Gabija
Total
Ph
Pl
Tr
D
K
L
U
V
X
50 0 50
Number of species
B
Total genes
Mobilome (X)
Prophage (Ph)
Plasmid (Pl)
Transposon (Tr)
Abi
CBASS
CRISPR
DMS
DRT
RM
Gabija
-10 -5 0 5 10
-40 -20 0 20 40
Relative change (%)

Figure 2

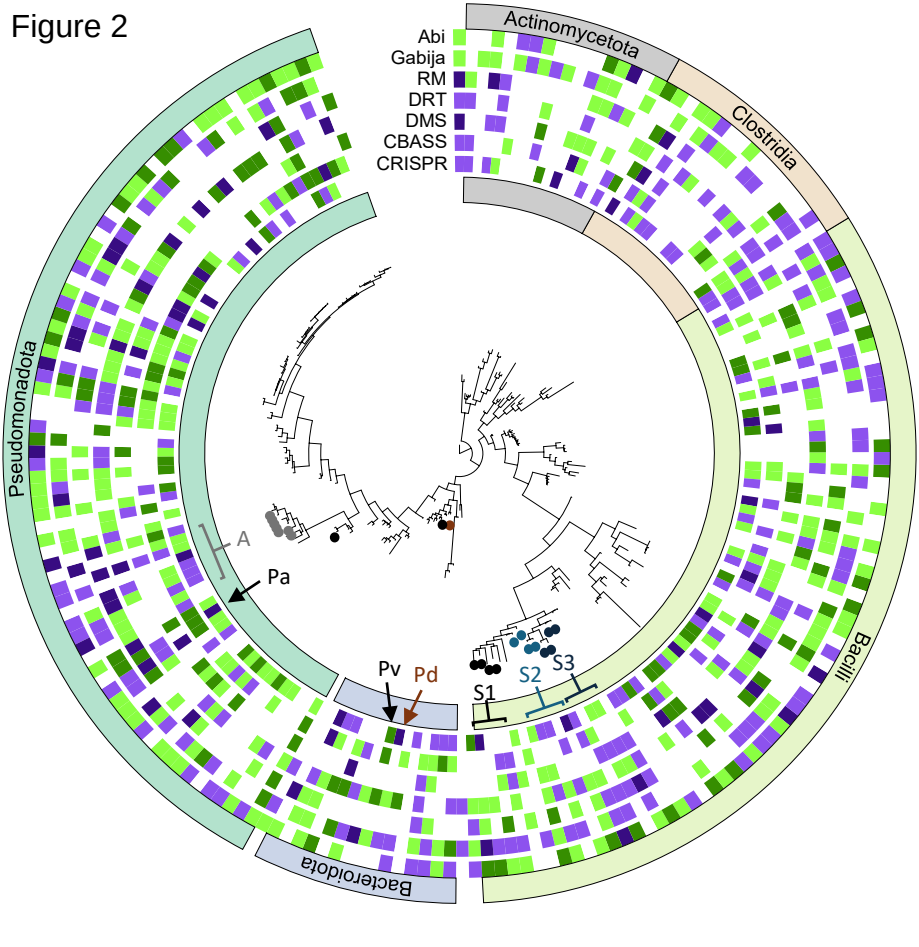

Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Actinomycetota
Clostridia
Bacilli
Bacteriodota
Pseudomonadota
A
Pa
Pv
Pd
S1
S2
S3
Negative (p < 0.05)
Positive (p < 0.05)
Negative (|ES| > SSES)
Positive (ES > SSES)

Figure 3

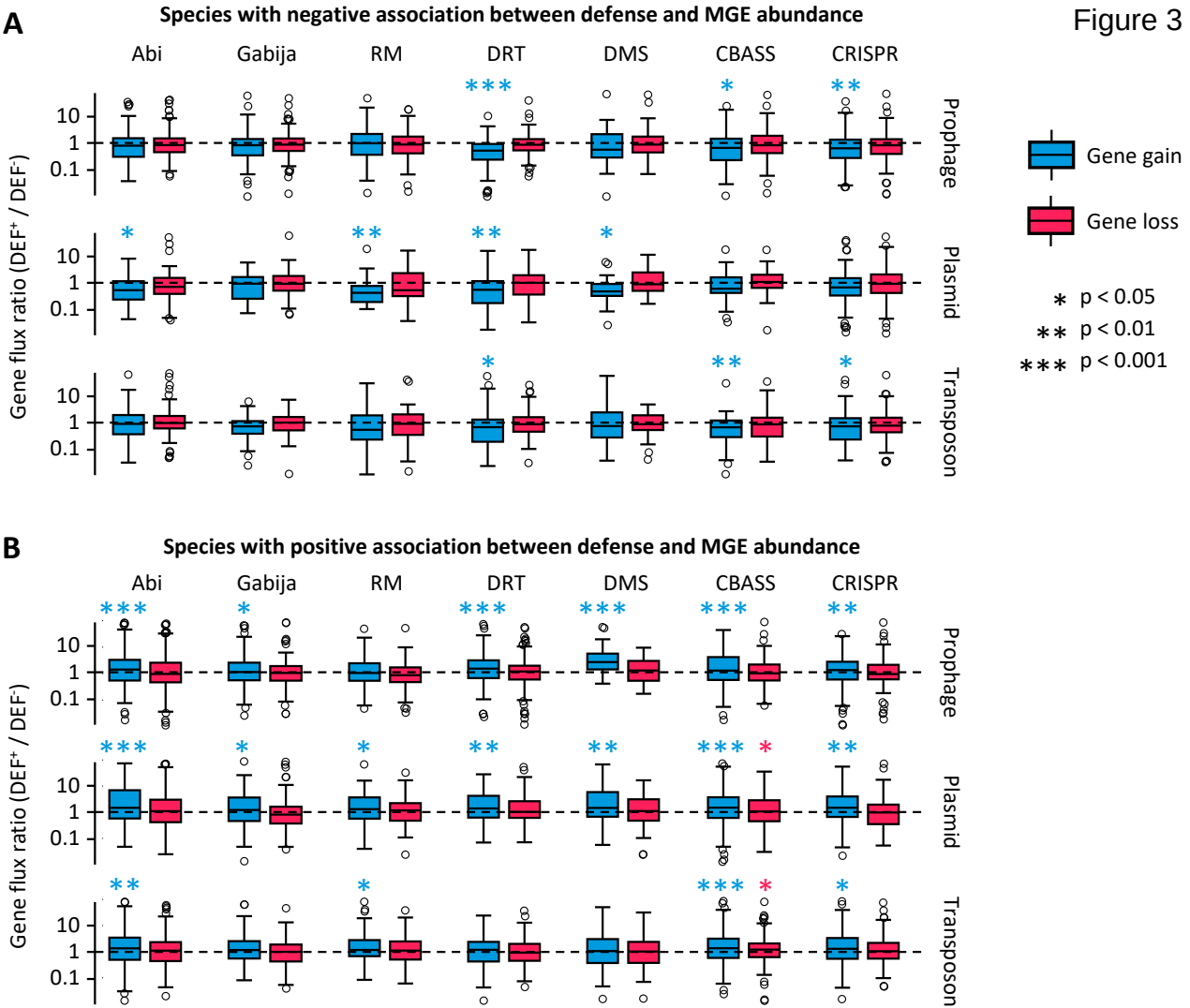

A
Species with negative association between defense and MGE abundance
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Prophage
Plasmid
Transposon
Gene flux ratio (DEF+ / DEF-)
10
1
0.1
B
Species with positive association between defense and MGE abundance
Gene gain
Gene loss
* p < 0.05
** p < 0.01
*** p < 0.001

Figure 4

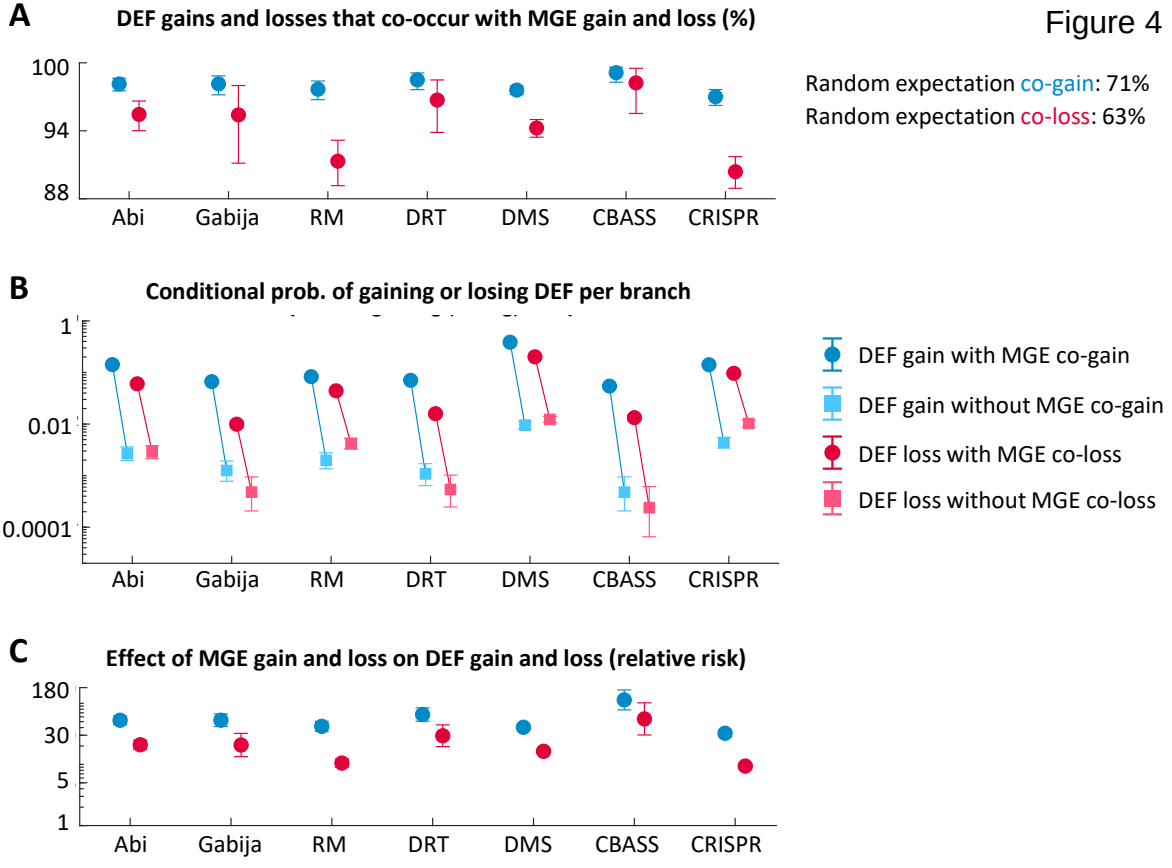
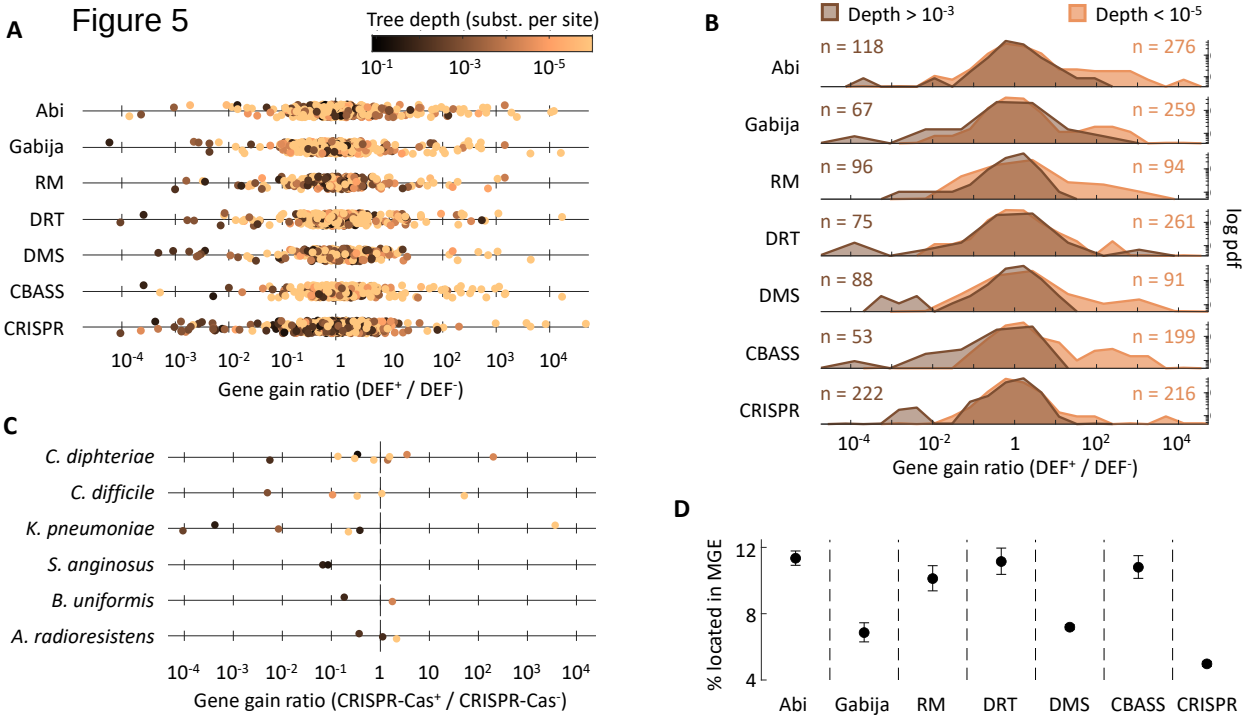

A
DEF gains and losses that co-occur with MGE gain and loss (%)
100
94
88
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
Random expectation co-gain: 71%
Random expectation co-loss: 63%
B
Conditional prob. of gaining or losing DEF per branch
1
0.01
0.0001
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR
DEF gain with MGE co-gain
DEF gain without MGE co-gain
DEF loss with MGE co-loss
DEF loss without MGE co-loss
C
Effect of MGE gain and loss on DEF gain and loss (relative risk)
180
30
5
1
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR

Figure 5


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 10^{2} 10^{3} 10^{4}
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 10^{2} 10^{4}
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 10^{2} 10^{3} 10^{4}
Gene gain ratio (CRISPR-Cas⁺ / CRISPR-Cas⁻)
D
% located in MGE
12
8
4
Abi
Gabija
RM
DRT
DMS
CBASS
CRISPR