Restriction-modification systems have shaped the evolution and distribution of plasmids across bacteria

Liam P. Shaw[1,2], Eduardo P. C. Rocha[3], R. Craig MacLean[1]

[1] Department of Biology, University of Oxford, Oxford, UK

[2] Department of Biosciences, University of Durham, Durham, UK

[3] Institut Pasteur, Université Paris Cité, CNRS UMR 3525, Microbial Genomics Unit, Institut Pasteur

Correspondence: liam.shaw@biology.ox.ac.uk

Abstract

Many novel traits such as antibiotic resistance are spread by plasmids between species. Yet plasmids have different host ranges. Restriction-modification systems (R-M systems) are by far the most abundant bacterial defense system and therefore represent one of the key barriers to plasmid spread. However, their effect on plasmid evolution and host range has been neglected. Here we analyse the avoidance of targets of the most abundant R-M systems (Type II) for complete genomes and plasmids across bacterial diversity. For the most common target length (6 bp) we show that target avoidance is strongly correlated with the taxonomic distribution of R-M systems and is greater in plasmid genes. We find stronger avoidance of R-M targets in plasmids which are smaller and have a broader host range. Our results suggest two different evolutionary strategies for plasmids: small plasmids primarily adapt to R-M systems by tuning their sequence composition, and large plasmids primarily adapt through the carriage of additional genes protecting from restriction. Our work provides systematic evidence that R-M systems are important barriers to plasmid transfer and have left their mark on plasmids over long evolutionary time.

Introduction

When DNA enters a bacterial cell from the world outside, it is a potential threat. If transcribed into RNA then translated into protein by the cell's own molecular machinery it may produce disaster. Mobile genetic elements (MGEs) such as lytic phage attempt to hijack cellular machinery to their own advantage: the transcription of phage DNA leads to copies of phage being produced at the expense of the bacterial host, followed by lysis and cell death. For this reason, bacteria have evolved many 'defense systems' which offer protection against external DNA. Defense systems impair or block infection by MGEs. Their evolution is closely linked to MGEs (Koonin, Makarova, and Wolf 2017) and they help to shape routes of gene flow between bacteria (Haudiquet et al. 2022). The majority of prokaryotic genomes contain at least one R-M system (83%) making them by far the most abundant defense systems – over twice as abundant as CRISPR-Cas (Tesson et al. 2022). R-M systems recognise specific DNA motifs and are grouped into four broad types I-IV (Loenen et al. 2014).

Within R-M systems, Type II are the most abundant, present in 39.2% of bacterial genomes (Tesson et al. 2022) with a mean of ~0.5 systems per genome (Oliveira, Touchon, and Rocha 2014). Type II R-M systems consist of two enzyme activities: a restriction endonuclease which cuts double-stranded DNA (dsDNA) at targets and a methyltransferase which modifies targets to protect them from cleavage. These enzymes are typically encoded by separate genes located close together in the genome. The targets of restriction are short sequences of 4-8 bases which are usually palindromic i.e. they are equal to their own reverse complement (Pingoud and Jeltsch 2001) due of the symmetrical subunits of the protein multimers that recognize the target (Arber and Linn 1969; Smith and Wilcox 1970). Any occurrences of the restriction target in the cell's own DNA should be protected from restriction by the methyltransferase. In contrast, DNA originating from a different species or strain should lack this methylation at target sites and will be cleaved by the restriction endonuclease when it enters the cell.

R-M systems are the most-studied class of defense systems and have been heavily investigated since their discovery in the 1960s (Arber 1965; Roberts 2005). Their widespread prevalence across bacteria suggests they provide an important defense against MGEs, which implies a strong selective pressure on MGEs to evade their targeting. Work on the first sequenced phage genomes in the 1980s showed evidence of selection against restriction targets (Sharp, 1986) which was backed up by subsequent research (Burge, Campbell, and Karlin 1992; Gelfand and Koonin 1997; Rocha, Danchin, and Viari 2001; Rusinov et al. 2018). By providing an innate or 'first-line' immunity, R-M systems can impair incoming MGEs prior to the activation of other 'second-line' defense systems. They are compatible with CRISPR-Cas (Dupuis et al. 2013) and restriction endonuclease cleavage of viral DNA can stimulate the subsequent adaptive CRISPR response (Maguin et al. 2022).

As well as functioning as defense systems, R-M systems can also be viewed as selfish elements that serve to propagate themselves. Because the methyltransferase decays more quickly than the endonuclease, a Type II R-M system can function as an addiction system to ensure its own persistence (Ichige and Kobayashi 2005; Kusano et al. 1995), similar to toxin-antitoxin systems (Mruk and Kobayashi 2014). This addictive quality may contribute to their occasional occurrence on MGEs such as plasmids: around 10.5% of plasmids carry R-M systems (Oliveira, Touchon, and Rocha 2014) and experiments have shown R-M system carriage can lead to increased plasmid stability in cells (Kusano et al. 1995).

Despite the different interpretations of the evolutionary role of R-M systems, it is clear that they shape pathways of gene flow between populations. In line with this, bacteria possessing cognate R-M systems have higher rates of horizontal gene flow between them (Oliveira, Touchon, and Rocha 2016). One major route of this gene flow is plasmid transfer. Plasmids are vehicles for novel traits that are beneficial across species (Lehtinen, Huisman, and Bonhoeffer 2021) including antibiotic resistance (MacLean and San Millan 2019). However, plasmid transfer is often constrained by taxonomic boundaries. The host range of a plasmid is defined as the range of different bacteria it can infect, with plasmids traditionally divided into 'narrow' or 'broad' host range. It has been suggested that plasmids with narrower host ranges tend to have a similar sequence composition to their host chromosomes due to ameliorative adaptation, which could include adaptation to defense systems (Suzuki et al. 2010).

More recent large-scale analyses of plasmids have quantified host range by grouping similar plasmids into clusters (Acman et al. 2020; Redondo-Salvo et al. 2020). These studies suggest many plasmids have a limited observed host range: considering only clusters with at least four plasmids, single-species plasmid clusters make up 45% of plasmid taxonomic units (PTUs) (Redondo-Salvo et al. 2020) or 52% of plasmid cliques (Acman et al. 2020). As barriers to the spread of dsDNA MGEs, R-M systems contribute to shaping the possible routes of plasmid transfer (Thomas and Nielsen 2005). Yet existing studies of R-M systems and plasmids are experimental and mostly limited to transfer within a single species – for example, in *Helicobacter pylori* (Ando et al. 2000) or *Enterococcus faecalis* (Price et al. 2016).

Over fifty years ago Arber and Linn (1969) speculated that because 'transferable plasmids have a fair chance of alternating rather frequently among hosts of various specificity...[we should] expect that with relatively small DNA molecules many original sites for the specificities of the most common hosts have been lost'. Yet despite the detailed characterisation of R-M systems compared to other defense systems (Roberts et al. 2015) and their ubiquity across bacteria, we still do not know whether this hypothesis holds true across plasmids. As such, we lack a systematic understanding of the role of R-M systems in shaping plasmid transfer routes across known bacterial diversity.

Here we investigate the avoidance of Type II restriction targets in plasmids, using a dataset of 8,552 complete genomes from 72 species containing 21,814 plasmids, as well as a separate dataset of plasmids with information on host range. Our results confirm that avoidance of restriction targets is a general feature of plasmids. By analysing the taxonomic distribution of Type II R-M systems and plasmids together, we show that avoidance patterns are associated with a plasmid's size and host range: small and broad host range plasmids show greater avoidance of R-M targets. Our findings suggest that Type II R-M systems are important drivers of plasmid evolution and shape routes of plasmid transfer in bacterial populations.

Results

Avoidance of 6-bp palindromes is stronger in plasmid genes than in core genes

The pangenome of a species consists of all the gene families found in the species as a whole (McInerney, McNally, and O'Connell 2017; Shapiro 2017). MGEs are important contributors to the accessory component of the pangenome – genes which are variably present or absent in different members of the species. As defense systems, Type II R-M systems should exert a selective pressure within a pangenome for avoidance of their short targets, which are often palindromic and 4-6bp in length. Older studies have shown that both phage and bacteria avoid short palindromes (Rocha, Danchin, and Viari 2001; Sharp 1986), and one study on the 49kb backbone of the broad host range IncP-1 plasmid found an under-representation of 6-bp palindromes (Wilkins et al. 1996).

We hypothesised that the plasmid-borne components of the pangenome should show stronger avoidance of R-M targets than core genes. To test this hypothesis, we assembled a dataset of high-quality reference genomes for species from NCBI RefSeq (n=72 species with >25 genomes). Within each species, we separated genes into three pangenome components: genes where >99% of genomes in the species had exactly one copy ('core'), other genes on the chromosome ('non-core'), and all genes carried on other replicons ('plasmid'). As an initial proxy for restriction targets, we first analysed the avoidance of short palindromes in each pangenome component for k=4 and k=6 (DNA palindromes require k to be even).

There are two important considerations when testing the avoidance of targets across bacterial diversity. First, when testing evidence of avoidance of a specific target it is important to account for differences in sequence composition; for example, a GC-rich sequence should *a priori* contain fewer occurrences of an AT-rich target. To do so, we used a maximal Markov model to calculate an exceptionality score for each k-mer (Schbath 1997). This exceptionality score is based on the deviation between the actual occurrences of the k-mer from the null expectation of occurrences one would expect given the distribution of (k-1)-mers. Positive values of the exceptionality score for a k-mer (>0) indicate evidence of over-representation and negative values (<0) indicate avoidance. To ensure that exceptionality scores had the same statistical power for comparisons between components, we also subsampled pangenome components to fixed lengths (see Methods). Genes in all three components clearly avoided palindromes (exceptionality score < 0, k=6 Fig. 1a, for k=4 see Fig. S1). We found a hierarchy of avoidance, with plasmid genes avoiding 6-bp palindromes significantly more on average than core and non-core chromosomal genes ($p<0.001$ two-sided Wilcoxon paired test, Fig. 1a). There was a significant correlation at the species level for palindrome avoidance in core and plasmid genes (Fig. 1b).

Second, genome composition is correlated with phylogeny and public databases are unevenly sampled, making overall findings about 'average' effects from comparative studies potentially misleading. Phylogenetically controlled analyses are required to draw reliable conclusions (Stone, Nee, and Felsenstein 2011; Hadfield and Nakagawa 2010). We controlled for phylogenetic signal in our analysis by modelling palindrome avoidance in pangenome components with generalized linear mixed models (GLMMs) (Hadfield 2022), controlling for phylogeny and number of genomes (see Methods and Table S1). For 6-bp palindromes, plasmid genes showed an overall greater avoidance than core genes despite variability between species (Fig. 1c; R^2=7.4%, Table S1b) with strong phylogenetic signal (>40%; Table S1b), suggesting that the taxonomic distribution of R-M systems may be an important

contributor to these phylogenetically clustered patterns of target avoidance. Notably, variation in palindrome avoidance was much greater in plasmid genes than core genes (Fig. 1c, inset panel) consistent with the expectation that plasmids seen within a species may have diverse evolutionary histories. This greater variability suggests the importance of considering differences between individual plasmids.

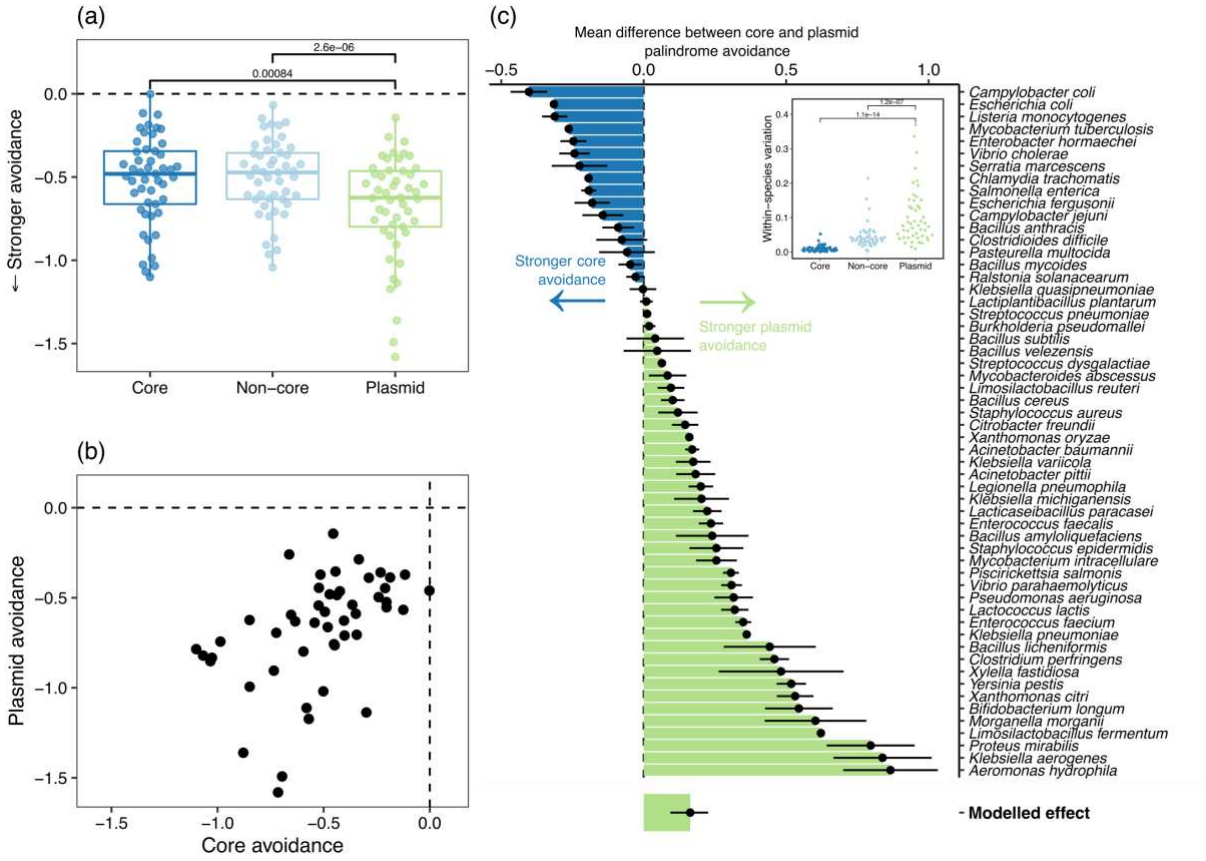

Fig. 1. Avoidance of short palindromes (k=6) is stronger but more variable in plasmids. (a) A hierarchy of 6-bp palindrome avoidance scores, with significantly greater avoidance in plasmid genes compared to core and non-core chromosomal genes ($p<0.001$, two-sided Wilcoxon paired test). (b) Mean avoidance is strongly structured by species, with a strong correlation between avoidance in core and plasmid genes (Spearman's $\rho=0.55$, $p<0.001$). (c) Relative palindrome avoidance for species for core vs. plasmid genes (>0 denotes greater avoidance in plasmid genes). Points are mean, error bars show standard error. The modelled effect was computed using a phylogenetically-controlled GLMM (see Methods). Data shown are mean avoidance scores of 6-bp palindromes ($4^3=64$) calculated with R'MES after pangenome construction then subsampling each per-isolate pangenome compoment to 50kbp i.e. only genomes with at least 50kbp are included (3,912 isolate genomes across 44 species). The inset panel shows within-species variation in mean palindrome avoidance score for each pangenome component. Only species with at least 3 genomes meeting these criteria are shown. For 4-bp palindromes, there was no significant difference between plasmid and core genes (Fig. S1) and mean avoidance was uncorrelated with 6-bp palindrome avoidance (Spearman's $\rho=0.005$, Fig. S2). Notably, Wilkins et al. (1996) found that 4-bp palindromes were not strongly avoided in the IncP-1 backbone and suggested that R-M systems with 6-bp targets were a stronger selective pressure, in line with our findings here.

The taxonomic distribution of Type II R-M systems correlates with target avoidance

Our genomic dataset spanned a wide range of bacterial diversity (Fig. S3). We hypothesised that selective pressure for avoidance of a target should correlate with the frequency of encounter with an R-M system targeting it. Reliable prediction of targets for novel sequences is only possible for Type II R-M systems where restriction and methylation are carried out by different enzymes (Oliveira, Touchon, and Rocha 2016) (see Methods). We developed a pipeline ('rmsFinder') to predict both the presence and targets of Type II R-M systems in our dataset using the curated REBASE database of known R-M enzymes. We produced a presence-absence matrix of k-mers targeted by Type II R-M systems across species in our dataset: when we detected a system with a target *t* in a genome from species *s*, we classed *t* as a within-species restriction target of *s*. In turn, we used this presence-absence matrix to produce a taxonomic dictionary of targets for each species (Fig. 2a), ranging from within-species to within-phylum targeting based on the detected presence of R-M systems across our dataset. We detected 8,469 Type II R-M systems where we could confidently predict their target in 2,740 genomes (32.0%). Of these systems, 7,734 (91.3%) were carried on the chromosome. R-M systems targeted 103 known REBASE motifs. Accounting for ambiguous bases, R-M systems targeted 278 specific *k*-mers (Table 1). Since motifs of *k*=7 and 9 were not prevalent (only observed in 66 genomes) we analysed targets for *k*=4,5,6 (98/103 motifs; Table 1) across our pangenome dataset. Type II R-M systems for these targets showed a highly variable presence/absence distribution across species (Fig. S4-S6 for different *k*).

For all pangenome components and all *k*, avoidance of targets was strongly correlated with the taxonomic distribution of the associated R-M systems (*k*=6 Fig. 2d-e; k=4 Fig. S7 and *k*=5 Fig. S8). Species pangenomes have the greatest avoidance of targets of the R-M systems found within that species. Core and non-core chromosomal genes had highly similar avoidance patterns. We found that 6-bp targets within the same taxonomic family were avoided more by plasmid genes at nearby taxonomic levels (species to family), with this difference decreasing at higher taxonomic orders (class, phylum) and with no difference when considering avoidance of all observed R-M targets within the dataset (kingdom). Selective pressure from R-M systems has imposed selection for plasmids to avoid R-M targets, and the strength of this avoidance is proportional to their frequency of encounter. This is consistent with the hypothesis that R-M systems are closely connected with taxonomic boundaries and plasmid host range.

Length (*k*)	REBASE motifs	*k*-mer targets	Palindromes	Genomes*	Species
4	12	12	10 of 12	690	33/72
5	30	46	-	1430	60/72
6	56	128	45 of 64	1423	53/72
7	4	28	-	61	
9	1	64	-	5	

Table 1: Detected Type II R-M targets across the dataset of 8,552 genomes.
* Number of genomes with at least 1 R-M system targeting a target of length *k*.

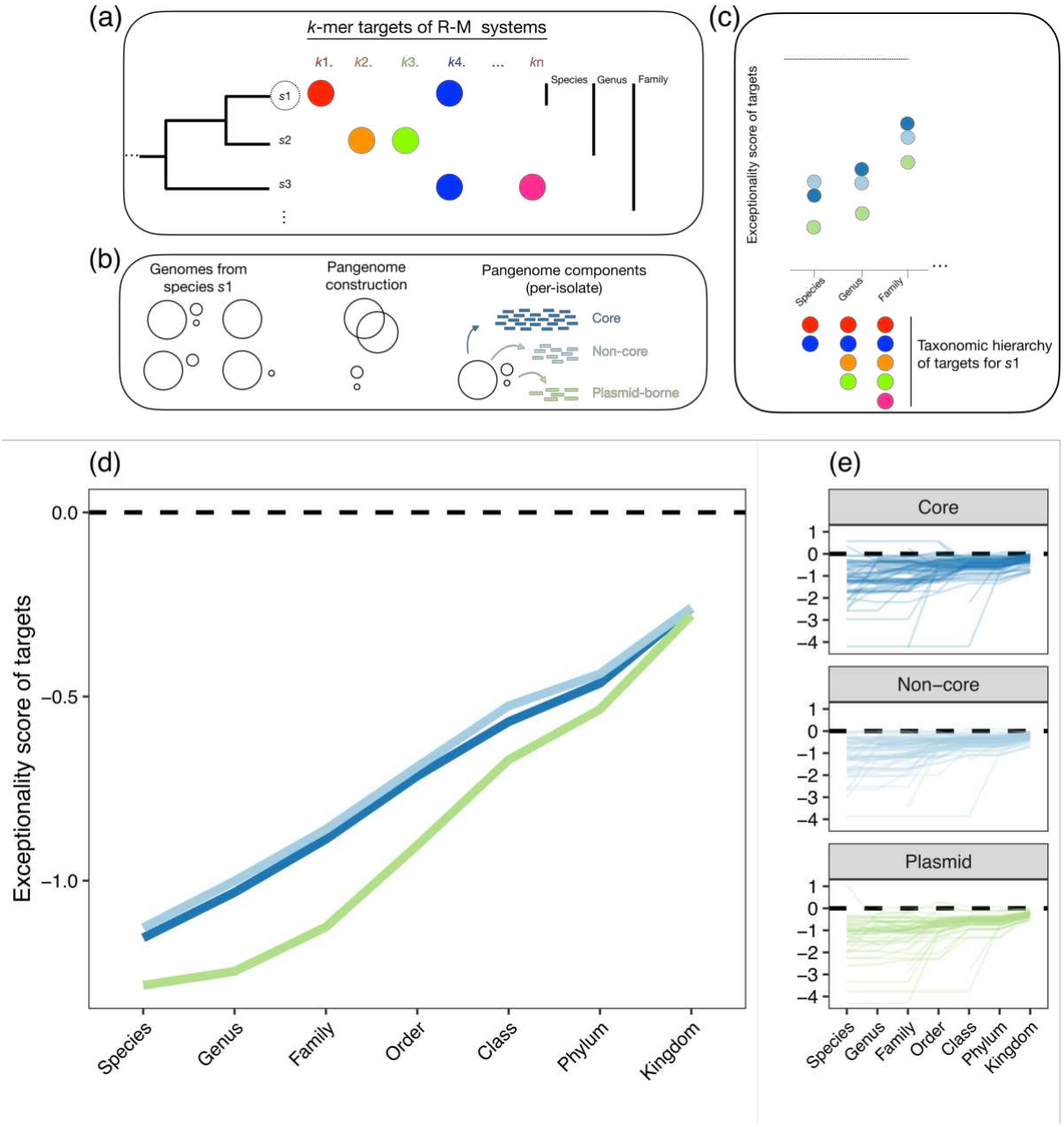

Fig. 2. The taxonomic distribution of R-M systems correlates with avoidance of their targets. (a-c) Methodological approach to connect Type II R-M system distribution to target avoidance: (a) We search for Type II R-M systems in n=8,552 genomes from 72 species, detecting complete systems with confident prediction of targets them in 2,740 genomes (Table 1). From these hits, we created a taxonomic hierarchy of their targets across a set of species. (b) We construct a pangenome for each species in our dataset, then separate each individual isolate into genes in three pangenome components: core, non-core and plasmid. (c) We subsample pangenome components to a fixed size and use R'MES to calculate exceptionality scores for fixed-length *k*-mers for *k*=4,5,6 for each species, using the taxonomic hierarchy of R-M targets to correlate exceptionality scores with R-M distribution. (d-e) Exceptionality scores for 6-mers by pangenome component as a function of the taxonomic hierarchy of R-M targets: (d) averaged over all species and (e) for individual species. Subsampling is to 50kbp for each within-isolate pangenome component. Other subsampling lengths show the same pattern (see github repository).

The density of within-species R-M targets increases with plasmid size

It is the actual number of occurrences of a R-M target within a plasmid that determines the extent to which it will be restricted by the associated R-M system. The expected number of target occurrences increases linearly with the size of the plasmid: for a plasmid of length *L*, the probability of containing a given *k*-mer scales as ~ $L/4^k$. For a random *k*-mer, one should expect a constant mean density. However, when we examine plasmids from the most

prevalent species in our genomic dataset, *Escherichia coli*, the density of R-M targets increases with plasmid size: larger plasmids have a disproportionate number of targets (Fig. 3a-b). This pattern is consistent across species (Fig. 3c-e).

From an evolutionary perspective, this result is consistent with the way that selective pressure from R-M systems acts at the whole-plasmid level. The efficiency of R-M systems in restricting sequences should increase with target frequency, although some systems can restrict sequences with only a single target and others require two targets to function (Embleton, Siksnys, and Halford 2001; Bath et al. 2002). R-M systems thus exert a selective pressure for target depletion: without other avoidance mechanisms, to avoid restriction a plasmid must lose the restriction targets from its sequence. The number of targets, and thus the number of mutations required to lose them, increases with plasmid length.

By way of an example, consider the case of a target of length k=6. Each extra 5kb of sequence will, on average, add ~1 more occurrence of the target (4^6=4,096). At one extreme, for a small 5kb plasmid, losing its only copy of the target requires only one mutation. This mutation will carry a large fitness advantage. However, larger plasmids will require many more mutations to become target-free: a 100kb plasmid will contain ~20 copies. While the final target-free sequence will have a large fitness advantage relative to its initial state, it must be reached gradually. Each mutational step will likely have only a weakly positive advantage compared to the previous step. Therefore, the larger a plasmid gets, the less evolutionarily accessible the mutational route to evade R-M systems becomes. The clear increase we find in the density of R-M targets with plasmid size across thousands of plasmids suggests that larger plasmids need other mechanisms of avoiding restriction.

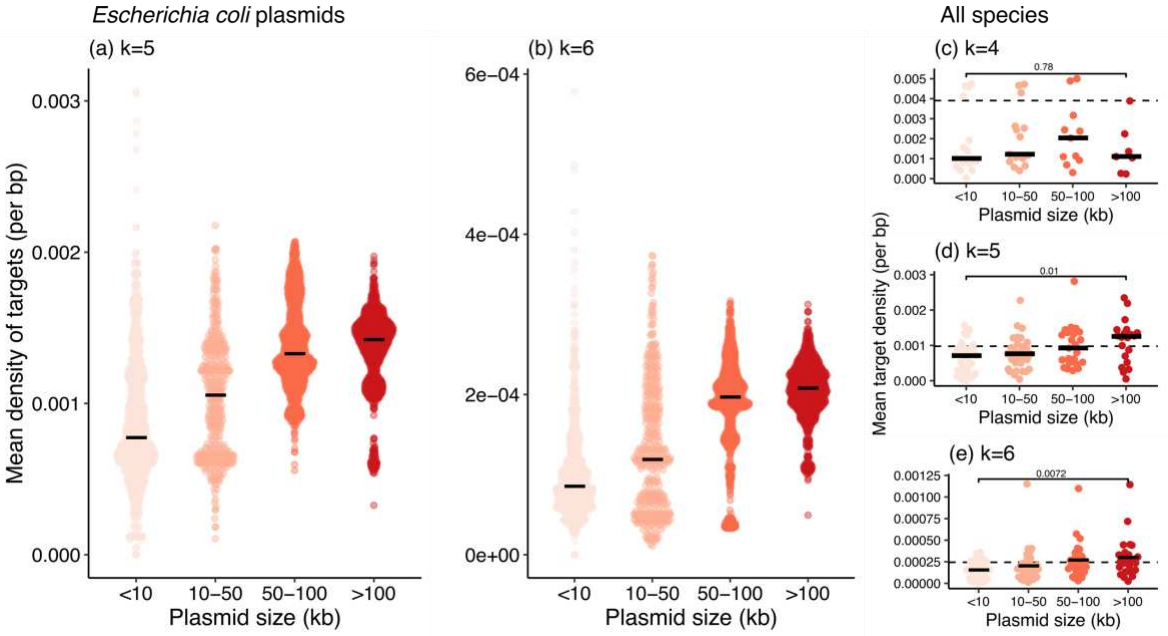

Fig. 3. Larger plasmids have a higher density of the targets of within-species R-M systems. (a-b) Results for the best-sampled species in our genomic dataset, *Escherichia coli*, for the mean density of within-species R-M targets of length (a) k=5 (4 targets) and (b) k=6 (33 targets). Each point is the mean density of targets within a single plasmid (no deduplication), black lines show median for each category. (c-e) Results for at a per-species level for different values of k. Species without R-M systems with targets of length k are omitted. Each point represents the median of the mean densities of within-species R-M targets for plasmids in that species, including only size/species combinations with >5 plasmids. Dashed lines shows the expected) density of a random k-mer in a random sequence (4^{-k}). Comparisons between the largest (>100kb) and smallest (<10kb) plasmid categories are significant ($p<0.05$) for k=5 and 6 but not for k=4.

Plasmid host range correlates with stronger avoidance of R-M targets

Previous work by Redondo-Salvo et al. (2020) clustered 10,634 plasmids based on their sequence similarity, defining 276 plasmid taxonomic units (PTUs) with at least four member plasmids (3,725 plasmids). They defined a host range for each PTU using its observed hosts, ranging from I-VI (from species to phylum). Under the hypothesis that R-M systems are a significant barrier to plasmid transfer, we would expect PTUs with a greater host range to have experienced more recent selection from a wider variety of R-M systems and therefore to have greater avoidance of R-M targets.

Using 6-bp palindromes as a proxy for Type II R-M targets, we find that host range is correlated with avoidance (Fig. 4). Interestingly, the avoidance of 6-bp palindromes in plasmids that are not members of an assigned PTU suggests that they are most similar to PTUs with a within-species host range in terms of palindrome avoidance. Many singleton plasmids (those detected only once) are probably indeed restricted to single species, although notably there is a long tail of more negative exceptionality scores, which suggests some may have broader host ranges and/or be more recent entrants into the pangenome of that species, so still have more avoidance of targets of R-M systems seen outside the species.

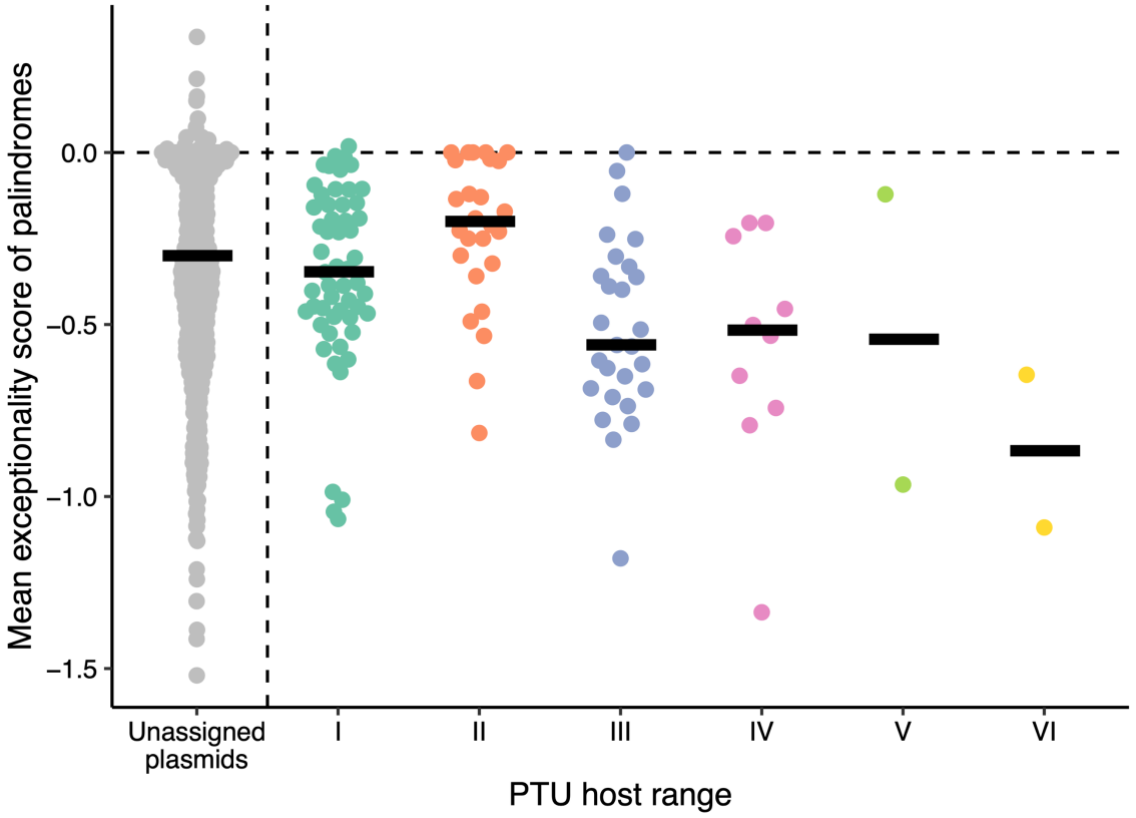

Fig. 4. PTU host range is associated with greater avoidance of 6-bp R-M targets. Avoidance of 6-bp palindromes in PTUs >10kbp correlates with PTU host range. Each point is one PTU (mean exceptionality score) apart from unassigned plasmids (those not classified into a clasdsed , lines show within-category median. Unassigned plasmids which are not members of a PTU show a mean avoidance of palindromes most similar to PTUs with a host range at the family level (III). (b) The most-sampled *Enterobacterales* species avoid 6-bp R-M targets correlated with their host range. PTU host range as assigned by Redondo-Salvo et al.: I (species) to VI (phylum). R-M target category based on observed distribution of Type II R-M systems in the whole genome dataset, from within-species to within-phylum (categories are hierarchically inclusive), and also shown are palindromic k-mers for comparison. Only plasmids >10kbp are included (subsampling to 10kbp for exceptionality score calculation).

We then modelled the avoidance of R-M targets using our taxonomic hierarchy in 4,000 PTUs seen in the same species as our dataset of complete genomes (see Methods). Linear models for exceptionality scores of 6-bp R-M targets in PTUs showed that the host range of plasmids was consistently associated with stronger avoidance of targets (Fig. 5a). In contrast, plasmid length was associated with weaker avoidance (Fig. 5b), a finding recapitulated for other values of *k*, confirming that small plasmids show greater signatures of mutational adaptation to evade R-M systems (*k*=4 Fig. S9; *k*=5 Fig. S10).

The magnitude of coefficient estimates decreased in magnitude for R-M targets from progressively wider taxonomic distributions (Fig. 5a-b), consistent with avoidance patterns being signatures of plasmid adaptation to their hosts within taxonomic boundaries. The number of plasmids within a PTU did not affect its average avoidance patterns (Fig. 5c). Models explained more variance at lower taxonomic levels of R-M target distribution (Fig. 5d), with the most variance explained for PTU avoidance of R-M targets from the same order as the plasmid. Taken together, these modelling results provide strong evidence that PTUs of small size and broad-host range have greater avoidance of R-M targets. Furthermore, these effects are most noticeable for R-M targets from nearby taxonomic levels. Evading R-M targeting through mutation is an important adaptive route for small, broad host range plasmids – raising the question of how larger plasmids evade R-M systems.

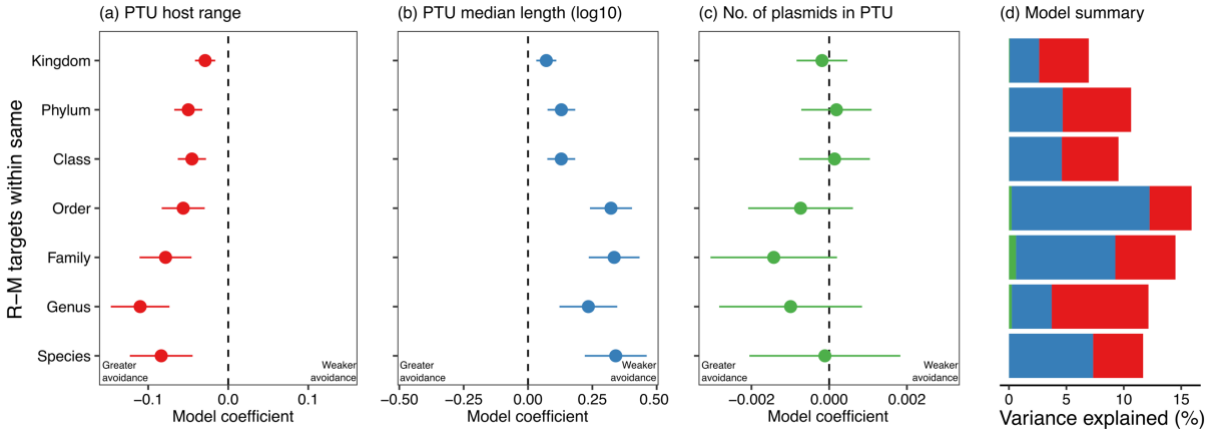

Fig. 5. Small and broad host range PTUs have stronger avoidance of R-M targets. (a-c) Coefficients in linear models (mean estimates with standard error shown by errorbars) for the exceptionality score of R-M targets. A different model was run for each possible of level of R-M targets within the taxonomic hierarchy, from R-M targets of R-M systems within-species to within-kingdom, with three variables for each PTU: host range, median length, and number of plasmids. (a) PTU host range, converted to a numeric variable for modelling where larger values denote broader host range, is negatively associated with exceptionality score of R-M targets i.e. broader host range PTUs have stronger avoidance. (b) Median length of plasmids within PTU (log10 for modelling) is positively associated with exceptionality score of R-M targets i.e. larger plasmids have weaker avoidance. (c) Number of plasmids within the PTU has no significant effect. (d) Total variance explained by each model, with colours denoting the three different variables (red: host range, blue: length, green: number of plasmids).

Broad host range plasmids carry more methyltransferases

The carriage of anti-restriction genes can help MGEs to evade restriction even when they carry sites recognized by the host (Spoerel, Herrlich, and Bickle 1979). Most of these sytems remain poorly described. Yet, a well-characterised way to evade restriction is by encoding a solitary Type II MTase. Such 'orphan' MTases are present in many prokaryotes and likely have functions linked to genome regulation (Blow et al. 2016), but they can also provide a plasmid with effective protection against restriction against multiple R-M targets (Fomenkov et al. 2020). Our hypothesis about the necessity of adaptation through gene carriage for large

plasmids suggests that solitary MTases should be frequently carried by larger plasmids and particularly those with a broader host range.

We searched all 10,634 plasmids in the Redono-Salvo dataset for MTases: 1,444 carried at least one Type II MTase with a predicted target (13.6% of which 243 carried >1 MTase), of which 789 had an MTase with a 4-6bp target (173 plasmids had >1 MTase). Larger plasmids within PTUs with a broad host range were more likely to carry MTases (Fig. 6). Analysing at the level of PTUs and subsetting based on their size, large PTUs (>100kbp) had both a greater proportion of their members carrying MTases and a greater normalised density of MTases (Fig. S11). We modelled MTase carriage as a function of PTU median length (log10) and host range. Both size and host range were associated with MTase carriage. When only considering large PTUs (>100kbp; n=61 PTUs), host range was strongly associated with a greater per-base density of Type II MTases ($p<0.001$, adjusted $R^2=34.3\%$). Though carriage of MTases could also be linked to modulation of host chromosome gene expression, these patterns are consistent with the expected differential responses to selective pressure from R-M systems. Small plasmids rarely carry MTases but can still have a broad host range despite this because of adaptive mutations. In contrast, most large plasmids with a broad host range carry MTases.

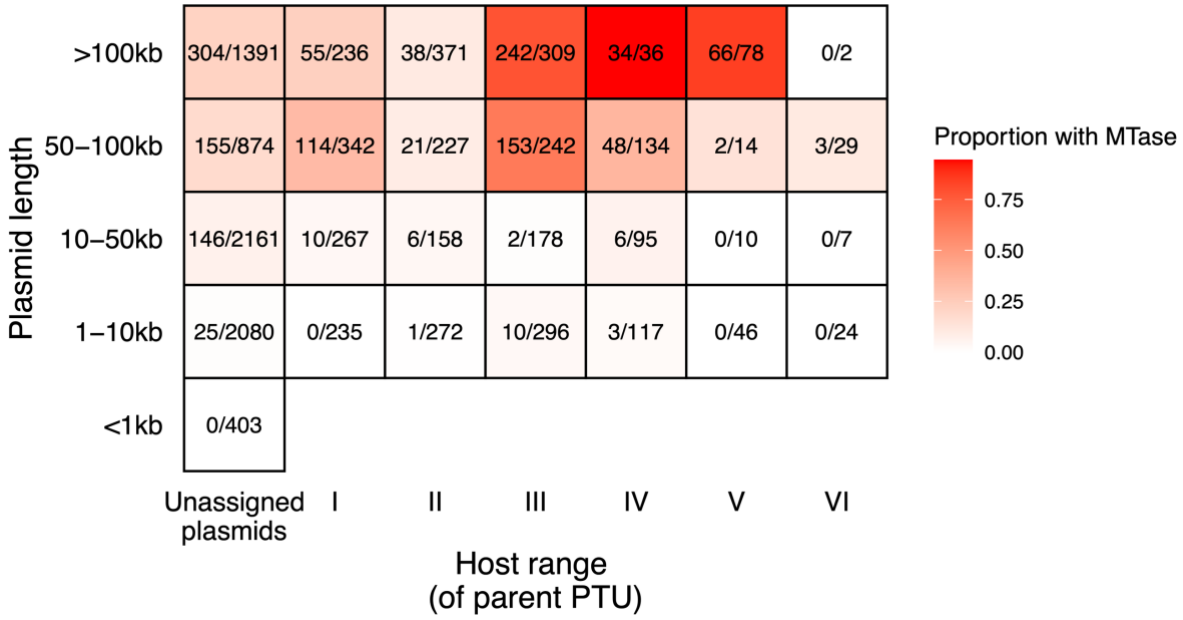

Figure 6. Large plasmids with a broad host range are more likely to carry MTases. Numbers show the number of plasmids in that category with at least one MTase out of the total number of plasmids.

Discussion

In human history, trade routes such as the Silk Road have been shaped by geography and politics; they have played an important long-term role in the movement of people, goods, and ideas. In bacterial evolution, routes of horizontal gene transfer between species have been shaped by defense systems. Here, by analysing the taxonomic distribution of the most prevalent of these defense systems – Type II R-M – we show that they have shaped the evolution and host range of plasmids. Our findings are consistent with a fifty-year-old hypothesis of Arber and Linn (1969) that small plasmids should avoid R-M targets in relation to their frequency of encounter.

The avoidance of short palindromes, assumed to be a proxy for Type II R-M targets, has previously been reported as a generic feature of bacterial genomes. However, these analyses have been limited in scope and not phylogenetically controlled. We have verified that this avoidance persists when accounting for phylogeny across a wide range of bacteria. Furthermore, we have gone beyond examining palindromes alone, and shown that the taxonomic distribution of R-M systems is correlated with avoidance of their targets. We found that plasmid genes show greater avoidance than core genes. They also show much greater variation, consistent with their diversity of evolutionary histories. We found that the host range of plasmid taxonomic units (PTUs) was associated with greater avoidance, suggesting that an interplay between R-M systems and plasmid host range. Models of R-M target avoidance explained the most variance for targets of systems seen within the same taxonomic order, which coincides with the observation that only 2.5% of PTUs have wider host ranges (Redondo-Salvo et al. 2020). Our findings are understandable from the perspective of an evolutionary arms race between bacteria and plasmids.

We found that small plasmids had a greater avoidance of R-M targets. We argued this is consistent with the greater evolutionary 'accessibility' of target removal by mutation compared to large plasmids: small plasmids need fewer mutations to become target-free, and each of these mutations has a strong fitness advantage. Furthermore, smaller plasmids tend to exist at higher plasmid copy number per cell. Since multi-copy plasmids can accelerate adaptive evolution by providing a greater mutational supply (San Millan et al. 2016) and avoidance of restriction is likely to be adaptive, this may contribute to an even greater depletion of restriction targets. Phage avoidance of R-M targets is greater for non-temperate phage, which have a lifestyle more dependent on horizontal transmission (Rusinov et al. 2018). Small multi-copy plasmids may be more ‘phage-like’ in this sense.

Plasmids have a highly bimodal size distribution: a strong peak at 5kb, very few plasmids at around 20kb, and a broad peak around 100kb (Smillie et al. 2010). But their fitness costs do not seem to be correlated with their size, at least when considering (Vogwill and MacLean 2015). The bimodal distribution is so widely recognised, yet it presents a puzzle: if adding genes to plasmids is cheap, why do so many plasmids remain small? Physical considerations of horizontal gene transfer must play a role. First, the apparatus of conjugation and transfer machinery has a minimum size (~10 kb), giving larger conjugative plasmids a minimum size. Second, there may be selection for mobilisable plasmids that are able to exploit phage mechanisms for horizontal transfer, giving small plasmids a maximum size of ~40kb (Humphrey et al. 2021). As is often the case in biology, there are likely multiple contributing factors. We suggest one that may have been overlooked is the role of R-M systems.

First, we found that 6-bp targets were the most common Type II R-M system. The first peak in plasmid size at 5kb is the length at which the expectation of a given 6-mer is ~1 (4^6=4,096), making it possible to evade any 6-mer targeting system through a single mutation (for 7-mer targets, the corresponding size is ~16.3kb). Second, species with many and diverse Type II R-M systems appear to have smaller plasmids, suggesting that R-M systems drive small plasmids to remain small. The relative absence of intermediate plasmids could be explained by this factor. Third, increasing plasmid size has a larger R-M-associated cost for smaller plasmids: the difference between zero and one or two copies of a target is a large one. It should be noted that some R-M systems interact with two recognition sites to cleave DNA, and more targets will probably increase the efficiency of restriction (Embleton, Siksnys, and Halford 2001; Bath et al. 2002). However, once plasmids have many copies of an R-M target in their sequence, having an additional target present is unlikely to be as great a proportional burden as the first few targets. Instead, because mutational adaptation becomes increasingly difficult with plasmid size, carrying additional genes becomes the main route of adaptation: genes which allow the evasion of R-M systems (single MTases, or anti-restriction enzymes) or other genes that confer benefits on the host to increase the likelihood of vertical inheritance after breakthrough infection. Most pairs of plasmids with 95% identical relaxases have fewer than 50% of homologs (Coluzzi et al. 2022), demonstrating that gene gain and loss are rapid. For this reason, larger plasmids are able to expand in size if required. R-M systems can therefore simultaneously drive small plasmids to become smaller and large plasmids to become larger. A similar logic applies to all defense systems targeting small DNA motifs.

Our work has limitations. Most notably, plasmid sequences are subject to a far greater range of selective pressures than we have explored here. Even considering just other defense systems alone, we have not investigated: the dual-function Type IIG enzymes with combined REase and MTase function (Loenen et al. 2014), the less common but still highly prevalent Type I, III, and IV R-M systems (Roberts et al. 2015) or indeed other ‘antiviral’ systems altogether (Tesson et al. 2022). There is also a growing appreciation that MGEs use 'defense' systems as weapons of intragenomic conflict (Rocha and Bikard 2022). Other pressures apart from defense systems may shape sequence composition: for example, there is some evidence that plasmids are AT-rich compared to chromosomes to reduce their metabolic burden (Dietel et al. 2019). In restricting our analysis to Type II R-M systems we have been deliberately conservative. Although we believe our findings are consistent with their expected action against plasmids, our analysis is only a partial picture of these complex overlapping pressures.

In conclusion, although Type II R-M systems are usually studied through the lens of phage defense, they have also shaped plasmid evolution. The selective pressure from R-M systems manifests differently with different plasmid sizes: small plasmids have the possibility of evading restriction through mutation, but large plasmids must adapt through accessory genes. More generally, our work suggests that avoidance patterns in MGEs contain information on the immune pressures they have endured. At a time when many novel ‘phage defense systems’ are being discovered, analysis of avoidance patterns can elucidate how these systems may have shaped the evolution and spread of other MGEs.

Acknowledgements. Thanks to the curators of REBASE and its many contributors, without which this work would not have been possible. Thanks also to Sophie Schbath for correspondence about R'MES.

Funding. LPS is a Sir Henry Wellcome Postdoctoral Fellow funded by Wellcome (Grant 220422/Z/20/Z). RCM was supported by funding from Wellcome (Grant 106918/Z/15Z). EPCR acklowdges support from the Fondation pour la Recherche Médicale (Grant EQU201903007835) and Laboratoire d'Excellence IBEID : Integrative Biology of Emerging Infectious Diseases (Grant ANR-10-LABX-62-IBEID).

Materials and Methods

Predicting Type II R-M systems. Our analysis approach requires a presence/absence database of R-M systems targeting particular motifs across different species of bacteria. We therefore developed a pipeline 'rmsFinder' to detect Type II R-M systems and then predict their target motifs: (https://github.com/liampshaw/rmsFinder). Previous work (Oliveira, Touchon, and Rocha 2016) determined protein similarity thresholds above which enzymes are likely to have the same target specificity: 50% (REases) and 55% (MTases). We used these as default values to define predicted targets. rmsFinder uses previously published hidden Markov models (HMMs) from either Oliveira, Touchon, and Rocha (2016) (--hmm oliveira) or Tesson et al. (2022) (--hmm tesson) to find putative Type II REases and MTases in a proteome. It then compares these putative enzymes to those enzymes in REBASE (Roberts et al. 2015) which have known or previously predicted targets.

In rmsFinder, we define the presence of a Type II R-M system as the presence of an MTase and REase with a shared predicted target within 4 genes of each other. rmsFinder returns both a list of possible hits to MTases and REases as well as this final prediction of Type II R-M systems with a known target. This final level of prediction can operate using different subsets of REBASE enzymes at decreasing levels of stringency:

- 'gold' - REBASE 'gold standard' proteins for which the biochemical function has been experimentally characterized and the nucleotide sequence coding for the exact protein is known.
- 'nonputative' - REBASE proteins that are known to have biochemical function (i.e. excluding proteins predicted bioinformatically by REBASE based on protein similarity).
- 'all' - all REBASE proteins, including putative protein sequences predicted bioinformatically by REBASE based on similarity to existing proteins.

Other parameters of rmsFinder are available on the github page. Results presented in this manuscript are from the 'all' mode of rmsFinder using REBASE v110 (downloaded 19 October 2021). We use the proteins defined within REBASE as Type II REases or MTases. We investigated the possibility of predicting the targets of Type IIG systems where the restriction and methylation functions are encoded in a single enzyme, but found that this was not reliable (data not shown) and so restricted our analysis only to Type II systems where the REase and MTase are separate enzymes.

Species genomes. We downloaded genomes for all n=104 species with >25 complete genomes in NCBI RefSeq (as of 20 January 2022) then filtered them with PanACoTA v1.3.1 (Perrin and Rocha 2021). After filtering, n=72 species had >25 complete genomes (8,552 genomes in total; 'RefSeq:>25' dataset). For each species, we used PanACoTA v1.3.1 to perform a pangenome analysis. We defined a gene family as 'core' if >99% of genomes had exactly one member (corepers subcommand of PanACoTA with '-t 0.99 -X'). This is a more relaxed definition than a strict core genome where all genomes are required to have exactly one copy of each core gene; such a definition can produce reduced core genomes when using public genomes, because an error in any single assembled genome can remove a gene from the core genome. After annotating to find CDSs, we split each RefSeq genome into three gene components: core genes on the chromosome ('core'), non-core genes on the

chromosome ('non-core'), and genes on other replicons ('plasmid'). Three species in our dataset contained secondary chromosomes: *Burkholderia pseudomallei* (81/91 isolates), *Vibrio cholerae* (57/70) and *Vibrio parahaemolyticus* (43/43). For the purposes of our analysis, we treated genes on these secondary chromosomes as 'plasmid' genes. Excluding them did not change our conclusions (Fig. S12). We analysed target avoidance both for the entire genome and for each pangenome component separately.

Plasmid genomes. We downloaded the dataset of n=10,634 plasmids previously analysed by Redondo-Salvo et al. (Redondo-Salvo et al. 2020). We used their existing classification of these plasmids into plasmid taxonomic units (PTUs). Redondo-Salvo et al. define the host range of a PTU from I-VI based on its observed distribution across taxonomic levels, from narrow (I: within-species) to broad (VI: within-phylum) (see Supp. Dataset 2 of that paper). We filtered the plasmids to n=4,000 plasmids that were seen in species from our RefSeq:>25 dataset (using TaxName in Redondo-Salvo Dataset S2 and disregarding extra specificity after genus and species). For modelling purposes, we note that host range is not strongly correlated with plasmid size (e.g. for k=6 linear model dataset, Spearman's ρ=0.046, p=0.10).

R-M target distribution. We ran rmsFinder on the 8,552 filtered genomes in our dataset of 72 species, of which 2,740 (32.0%) contained a Type II R-M system with a predicted target motif. Of these R-M-containing genomes, 2,035/2,740 (74.3%) contained a single R-M system (range: 0-18 R-M systems; *Helicobacter pylori* genomes accounted for all those with >9 R-M systems). Six species contained no predicted R-M systems (*Bacillus anthracis*, *Chlamydia trachomatis*, *Corynebacterium pseudotuberculosis*, *Limosilactobacillus reuteri* , *Mycobacterium tuberculosis*, *Piscirickettsia salmonis*). R-M systems targeted 104 known REBASE motifs corresponding to 278 unambiguous sequences (hereafter: 'targets') of which the vast majority were between 4 and 6 bases long (Table 1). Where a motif contained ambiguity codes (e.g. ATNNAT) we include all possibilities as independent targets i.e. with equal weighting compared to unambiguous targets. Out of the 98 motifs of 4-6 bases, 26 were targeted by only a single species. On average, a given REBASE motif was targeted by systems found in a median of 3 species (range: 1-28) and 21 genomes (range: 1-790).

We then aggregated these results by species into a binary presence/absence matrix of species against k-mers for k=4,5,6. In this matrix a 1 denotes that a functional R-M system targets the k-mer, and a 0 that no R-M system was observed in the dataset targeting that k-mer. We then used the AMR package in R to generate taxonomic classifications for all species. For a given species, we can therefore define the set of motifs that are targeted by R-M systems observed within-species, within-genus, within-family etc. up to the order of phylum. This 'taxonomic dictionary' allows us to explore how the distribution of R-M systems is linked to avoidance of their associated targets in bacterial genomes and plasmids.

Calculating target avoidance. Sequence composition strongly affects the number of times a short motif appears in a stretch of DNA. For example, one would expect few occurrences of GGCC in an AT-rich genome. We therefore used R'MES v3.1.0 (https://forgemia.inra.fr/sophie.schbath/rmes) to calculate an exceptionality score for all *k*-mers (*k*=4,5,6). R'MES controls for sequence composition by using a Markov chain model to calculate the expected occurrences of a word *W* of length *k* using the observed occurences of shorter words. This gives a null expectation which can be compared with the actual occurences of *W* to produce an exceptionality Z-score. For our analyses, we use the maximal model of order *m*=*k*-2 which uses the observed occurrences of all words with lengths ≤ *k-1* (Schbath 1997). The use of a maximal Markov model has the advantage that when a *k*-mer is observed significantly less than expected under the null model, this is a strong sign of selection against the word itself, rather than against the substrings it contains. Where a *k*-mer has zero observed occurrences and zero expected occurrences, its score as calculated by R'MES is defined as zero. Using the taxonomic dictionary of the presence of systems targeting particular R-M targets (Fig. 2a-c) we then calculated the median exceptionality score for defined groups of targets for each species. For example: assume that for a given species S_a, we detect R-M systems which target k_1, k_2 and k_3. A different species S_b within the same genus has R-M systems targeting k_1, k_4 and k_5. The within-species R-M targets of S_a are {k_1, k_2. k_3} and the within-genus targets are {k_1, k_2, k_3, k_4, k_5}. This logic extends up the taxonomic hierarchy,

up through family, order, class, phylum and finally to kingdom, the set of targets includes all k-mers targeted by any R-M system detected within our dataset. We use only the presence of an R-M system and do not use any prevalence information.

Controlling for sequence length. The statistical power to detect significant deviation in the abundance of motifs compared to expectation increases with sequence size. To control for differences in length between genome components, we ran analyses on both whole sequences and also subsampled sequences down to fixed lengths (2.5, 5, 10, 50, and 100 kbp). See github repository for more details.

Controlling for phylogeny. We modelled the difference in R-M target avoidance between plasmid genes and core genes at a within-isolate level, subsampling to 10kbp; n=4,553 genomes across 60 species with at least 10kbp in each of the three pangenome components. Differences between plasmids and chromosomes can be biased by the phylogenetic structure of bacteria. To account for this, we followed the methodology of Dewar et al. (2021), using MCMCglmm (Hadfield 2022) to include phylogeny and number of genomes as random effects in a generalized linear mixed model. For the host species phylogeny, we used a recent tree computed by Zhu et al. (2019) based on 381 marker genes. 17/72 species names were missing from the tree with a simple match. We manually checked these missing species and amended the Zhu phylogeny by either: renaming the taxon (n=5) or adding sister tips to members of the genera already in the tree based on a literature review (n=12), using half of the branch length distance to the nearest other taxon in the tree. We also manually amended the position of *Klebsiella michiganensis* to place it within the *Klebsiella* genus. The tree is provided in supplementary material (Fig. S4).

Modelling for palindrome avoidance. We used MCMCglmm v2.34 (Hadfield 2022) with two fixed effects (phylogeny, number of genomes) for the mean ranks of avoidance. For full details of code, see the github repository.

Software. rmsFinder is written in python. Bioinformatic analysis of genomes and plasmids was carried out using the Biomedical Research Computing (BMRC) facility at the University of Oxford. We conducted downstream analyses in R v4.1.2 and RStudio v2022.07.2 using We following R packages: dplyr v1.0.9, ggplot2 v3.3.6, cowplot v1.1.1, formatR v1.12, ape v5.6-1, ggtree v3.2.1, ggridges v0.5.3, MCMCglmm v2.34, phytools v1.0-3, reshape2 v1.4.4, tidyr v1.2.0, ggrepel v0.9.1, ggbeeswarm v0.6.0. All code is available on github.

Data availability. Genomes analysed are all from public databases (NCBI) and accessions are available via the github repository. Analysis scripts and intermediate data are available online: https://github.com/liampshaw/R-M-and-plasmids.

References

Acman, Mislav, Lucy van Dorp, Joanne M. Santini, and Francois Balloux. 2020. 'Large-Scale Network Analysis Captures Biological Features of Bacterial Plasmids'. *Nature Communications* 11 (1): 2452. https://doi.org/10.1038/s41467-020-16282-w.

Ando, T., Q. Xu, M. Torres, K. Kusugami, D. A. Israel, and M. J. Blaser. 2000. 'Restriction-Modification System Differences in Helicobacter Pylori Are a Barrier to Interstrain Plasmid Transfer'. *Molecular Microbiology* 37 (5): 1052–65. https://doi.org/10.1046/j.1365-2958.2000.02049.x.

Arber, W. 1965. 'Host-Controlled Modification of Bacteriophage'. *Annual Review of Microbiology* 19: 365–78. https://doi.org/10.1146/annurev.mi.19.100165.002053.

Arber, W., and S. Linn. 1969. 'DNA Modification and Restriction'. *Annual Review of Biochemistry* 38: 467–500. https://doi.org/10.1146/annurev.bi.38.070169.002343.

Bath, Abigail J., Susan E. Milsom, Niall A. Gormley, and Stephen E. Halford. 2002. 'Many Type IIs Restriction Endonucleases Interact with Two Recognition Sites before Cleaving DNA *'. *Journal of Biological Chemistry* 277 (6): 4024–33. https://doi.org/10.1074/jbc.M108441200.

Blow, Matthew J., Tyson A. Clark, Chris G. Daum, Adam M. Deutschbauer, Alexey Fomenkov, Roxanne Fries, Jeff Froula, et al. 2016. 'The Epigenomic Landscape of Prokaryotes'. *PLoS Genetics* 12 (2): e1005854. https://doi.org/10.1371/journal.pgen.1005854.

Burge, C, A M Campbell, and S Karlin. 1992. 'Over- and under-Representation of Short Oligonucleotides in DNA Sequences.' *Proceedings of the National Academy of Sciences* 89 (4): 1358–62. https://doi.org/10.1073/pnas.89.4.1358.

Coluzzi, Charles, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, and Eduardo P.C. Rocha. 2022. 'Evolution of Plasmid Mobility: Origin and Fate of Conjugative and Nonconjugative Plasmids'. *Molecular Biology and Evolution* 39 (6): msac115. https://doi.org/10.1093/molbev/msac115.

Dewar, Anna E., Joshua L. Thomas, Thomas W. Scott, Geoff Wild, Ashleigh S. Griffin, Stuart A. West, and Melanie Ghoul. 2021. 'Plasmids Do Not Consistently Stabilize Cooperation across Bacteria but May Promote Broad Pathogen Host-Range'. *Nature Ecology & Evolution* 5 (12): 1624–36. https://doi.org/10.1038/s41559-021-01573-2.

Dietel, Anne-Kathrin, Holger Merker, Martin Kaltenpoth, and Christian Kost. 2019. 'Selective Advantages Favour High Genomic AT-Contents in Intracellular Elements'. *PLOS Genetics* 15 (4): e1007778. https://doi.org/10.1371/journal.pgen.1007778.

Dupuis, Marie-Ève, Manuela Villion, Alfonso H. Magadán, and Sylvain Moineau. 2013. 'CRISPR-Cas and Restriction-Modification Systems Are Compatible and Increase Phage Resistance'. *Nature Communications* 4: 2087. https://doi.org/10.1038/ncomms3087.

Embleton, Michelle L., Virginijus Siksnys, and Stephen E. Halford. 2001. 'DNA Cleavage Reactions by Type II Restriction Enzymes That Require Two Copies of Their Recognition Sites11Edited by J. Karn'. *Journal of Molecular Biology* 311 (3): 503–14. https://doi.org/10.1006/jmbi.2001.4892.

Fomenkov, Alexey, Zhiyi Sun, Iain A Murray, Cristian Ruse, Colleen McClung, Yoshiharu Yamaichi, Elisabeth A Raleigh, and Richard J Roberts. 2020. 'Plasmid Replication-Associated Single-Strand-Specific Methyltransferases'. *Nucleic Acids Research* 48 (22): 12858–73. https://doi.org/10.1093/nar/gkaa1163.

Gelfand, Mikhail S., and Eugene V. Koonin. 1997. 'Avoidance of Palindromic Words in Bacterial and Archaeal Genomes: A Close Connection with Restriction Enzymes'. *Nucleic Acids Research* 25 (12): 2430–39. https://doi.org/10.1093/nar/25.12.2430.

Hadfield, Jarrod. 2022. 'MCMCglmm: MCMC Generalised Linear Mixed Models'. https://CRAN.R-project.org/package=MCMCglmm.

Hadfield, Jarrod, and S. Nakagawa. 2010. 'General Quantitative Genetic Methods for Comparative Biology: Phylogenies, Taxonomies and Multi-Trait Models for Continuous and Categorical Characters'. *Journal of Evolutionary Biology* 23 (3): 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x.

Haudiquet, Matthieu, Jorge Moura de Sousa, Marie Touchon, and Eduardo P. C. Rocha. 2022. 'Selfish, Promiscuous and Sometimes Useful: How Mobile Genetic Elements Drive Horizontal Gene Transfer in Microbial Populations'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377 (1861): 20210234. https://doi.org/10.1098/rstb.2021.0234.

Humphrey, Suzanne, Álvaro San Millán, Macarena Toll-Riera, John Connolly, Alejandra Flor-Duro, John Chen, Carles Ubeda, R. Craig MacLean, and José R. Penadés. 2021. 'Staphylococcal Phages and Pathogenicity Islands Drive Plasmid Evolution'. *Nature Communications* 12 (1): 5845. https://doi.org/10.1038/s41467-021-26101-5.

Ichige, Asao, and Ichizo Kobayashi. 2005. 'Stability of EcoRI Restriction-Modification Enzymes in Vivo Differentiates the EcoRI Restriction-Modification System from Other Postsegregational Cell Killing Systems'. *Journal of Bacteriology* 187 (19): 6612–21. https://doi.org/10.1128/JB.187.19.6612-6621.2005.

Koonin, Eugene V., Kira S. Makarova, and Yuri I. Wolf. 2017. 'Evolutionary Genomics of Defense Systems in Archaea and Bacteria'. *Annual Review of Microbiology* 71 (September): 233–61. https://doi.org/10.1146/annurev-micro-090816-093830.

Kusano, K, T Naito, N Handa, and I Kobayashi. 1995. 'Restriction-Modification Systems as Genomic Parasites in Competition for Specific Sequences.' *Proceedings of the National Academy of Sciences* 92 (24): 11095–99. https://doi.org/10.1073/pnas.92.24.11095.

Lehtinen, Sonja, Jana S. Huisman, and Sebastian Bonhoeffer. 2021. 'Evolutionary Mechanisms That Determine Which Bacterial Genes Are Carried on Plasmids'. *Evolution Letters* 5 (3): 290–301. https://doi.org/10.1002/evl3.226.

Loenen, Wil A. M., David T. F. Dryden, Elisabeth A. Raleigh, Geoffrey G. Wilson, and Noreen E. Murray. 2014. 'Highlights of the DNA Cutters: A Short History of the Restriction Enzymes'. *Nucleic Acids Research* 42 (1): 3–19. https://doi.org/10.1093/nar/gkt990.

MacLean, R. Craig, and Alvaro San Millan. 2019. 'The Evolution of Antibiotic Resistance'. *Science* 365 (6458): 1082–83. https://doi.org/10.1126/science.aax3879.

Maguin, Pascal, Andrew Varble, Joshua W. Modell, and Luciano A. Marraffini. 2022. 'Cleavage of Viral DNA by Restriction Endonucleases Stimulates the Type II CRISPR-Cas Immune Response'. *Molecular Cell* 82 (5): 907-919.e7. https://doi.org/10.1016/j.molcel.2022.01.012.

McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. 'Why Prokaryotes Have Pangenomes'. *Nature Microbiology* 2: 17040. https://doi.org/10.1038/nmicrobiol.2017.40.

Mruk, Iwona, and Ichizo Kobayashi. 2014. 'To Be or Not to Be: Regulation of Restriction–Modification Systems and Other Toxin–Antitoxin Systems'. *Nucleic Acids Research* 42 (1): 70–86. https://doi.org/10.1093/nar/gkt711.

Oliveira, Pedro H., Marie Touchon, and Eduardo P. C. Rocha. 2014. 'The Interplay of Restriction-Modification Systems with Mobile Genetic Elements and Their Prokaryotic Hosts'. *Nucleic Acids Research* 42 (16): 10618–31. https://doi.org/10.1093/nar/gku734.

———. 2016. 'Regulation of Genetic Flux between Bacteria by Restriction-Modification Systems'. *Proceedings of the National Academy of Sciences of the United States of America* 113 (20): 5658–63. https://doi.org/10.1073/pnas.1603257113.

Perrin, Amandine, and Eduardo P C Rocha. 2021. 'PanACoTA: A Modular Tool for Massive Microbial Comparative Genomics'. *NAR Genomics and Bioinformatics* 3 (1): lqaa106. https://doi.org/10.1093/nargab/lqaa106.

Pingoud, Alfred, and A. Jeltsch. 2001. 'Structure and Function of Type II Restriction Endonucleases'. *Nucleic Acids Research* 29 (18): 3705–27. https://doi.org/10.1093/nar/29.18.3705.

Price, Valerie J., Wenwen Huo, Ardalan Sharifi, and Kelli L. Palmer. 2016. 'CRISPR-Cas and Restriction-Modification Act Additively against Conjugative Antibiotic Resistance Plasmid Transfer in Enterococcus Faecalis'. *MSphere* 1 (3): e00064-16. https://doi.org/10.1128/mSphere.00064-16.

Redondo-Salvo, Santiago, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo P. C. Rocha, M. Pilar Garcillán-Barcia, and Fernando de la Cruz. 2020. 'Pathways for Horizontal Gene Transfer in Bacteria Revealed by a Global Map of Their Plasmids'. *Nature Communications* 11 (1): 3602. https://doi.org/10.1038/s41467-020-17278-2.

Roberts, Richard J. 2005. 'How Restriction Enzymes Became the Workhorses of Molecular Biology'. *Proceedings of the National Academy of Sciences of the United States of America* 102 (17): 5905–8. https://doi.org/10.1073/pnas.0500923102.

Roberts, Richard J., Tamas Vincze, Janos Posfai, and Dana Macelis. 2015. 'REBASE--a Database for DNA Restriction and Modification: Enzymes, Genes and Genomes'. *Nucleic Acids Research* 43 (Database issue): D298-299. https://doi.org/10.1093/nar/gku1046.

Rocha, Eduardo P. C., and David Bikard. 2022. 'Microbial Defenses against Mobile Genetic Elements and Viruses: Who Defends Whom from What?' *PLoS Biology* 20 (1): e3001514. https://doi.org/10.1371/journal.pbio.3001514.

Rocha, Eduardo P. C., Antoine Danchin, and Alain Viari. 2001. 'Evolutionary Role of Restriction/Modification Systems as Revealed by Comparative Genome Analysis'. *Genome Research* 11 (6): 946–58. https://doi.org/10.1101/gr.153101.

Rusinov, I. S., A. S. Ershova, A. S. Karyagina, S. A. Spirin, and A. V. Alexeevski. 2018. 'Avoidance of Recognition Sites of Restriction-Modification Systems Is a Widespread but Not Universal Anti-Restriction Strategy of Prokaryotic Viruses'. *BMC Genomics* 19 (1): 885. https://doi.org/10.1186/s12864-018-5324-3.

San Millan, Alvaro, Jose Antonio Escudero, Danna R. Gifford, Didier Mazel, and R. Craig MacLean. 2016. 'Multicopy Plasmids Potentiate the Evolution of Antibiotic Resistance in Bacteria'. *Nature Ecology & Evolution* 1 (1): 1–8. https://doi.org/10.1038/s41559-016-0010.

Schbath, S. 1997. 'An Efficient Statistic to Detect Over- and under-Represented Words in DNA Sequences'. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 4 (2): 189–92. https://doi.org/10.1089/cmb.1997.4.189.

Shapiro, B. Jesse. 2017. 'The Population Genetics of Pangenomes'. *Nature Microbiology* 2 (12): 1574–1574. https://doi.org/10.1038/s41564-017-0066-6.

Sharp, P M. 1986. 'Molecular Evolution of Bacteriophages: Evidence of Selection against the Recognition Sites of Host Restriction Enzymes.' *Molecular Biology and Evolution* 3 (1): 75–83. https://doi.org/10.1093/oxfordjournals.molbev.a040377.

Smillie, Chris, M. Pilar Garcillán-Barcia, M. Victoria Francia, Eduardo P. C. Rocha, and Fernando de la Cruz. 2010. 'Mobility of Plasmids'. *Microbiology and Molecular Biology Reviews: MMBR* 74 (3): 434–52. https://doi.org/10.1128/MMBR.00020-10.

Smith, H. O., and K. W. Wilcox. 1970. 'A Restriction Enzyme from Hemophilus Influenzae. I. Purification and General Properties'. *Journal of Molecular Biology* 51 (2): 379–91. https://doi.org/10.1016/0022-2836(70)90149-x.

Spoerel, N., P. Herrlich, and T. A. Bickle. 1979. 'A Novel Bacteriophage Defence Mechanism: The Anti-Restriction Protein'. *Nature* 278 (5699): 30–34. https://doi.org/10.1038/278030a0.

Stone, Graham N., Sean Nee, and Joseph Felsenstein. 2011. 'Controlling for Non-Independence in Comparative Analysis of Patterns across Populations within Species'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366 (1569): 1410–24. https://doi.org/10.1098/rstb.2010.0311.

Suzuki, Haruo, Hirokazu Yano, Celeste J. Brown, and Eva M. Top. 2010. 'Predicting Plasmid Promiscuity Based on Genomic Signature'. *Journal of Bacteriology* 192 (22): 6045–55. https://doi.org/10.1128/JB.00277-10.

Tesson, Florian, Alexandre Hervé, Ernest Mordret, Marie Touchon, Camille d'Humières, Jean Cury, and Aude Bernheim. 2022. 'Systematic and Quantitative View of the Antiviral Arsenal of Prokaryotes'. *Nature Communications* 13 (1): 2561. https://doi.org/10.1038/s41467-022-30269-9.

Thomas, Christopher M., and Kaare M. Nielsen. 2005. 'Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria'. *Nature Reviews. Microbiology* 3 (9): 711–21. https://doi.org/10.1038/nrmicro1234.

Vogwill, Tom, and R. Craig MacLean. 2015. 'The Genetic Basis of the Fitness Costs of Antimicrobial Resistance: A Meta-Analysis Approach'. *Evolutionary Applications* 8 (3): 284–95. https://doi.org/10.1111/eva.12202.

Wilkins, B. M., P. M. Chilley, A. T. Thomas, and M. J. Pocklington. 1996. 'Distribution of Restriction Enzyme Recognition Sequences on Broad Host Range Plasmid RP4: Molecular and Evolutionary Implications'. *Journal of Molecular Biology* 258 (3): 447–56. https://doi.org/10.1006/jmbi.1996.0261.

Zhu, Qiyun, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G. Sanders, Pedro Belda-Ferre, et al. 2019. 'Phylogenomics of 10,575 Genomes Reveals Evolutionary Proximity between Domains Bacteria and Archaea'. *Nature Communications* 10 (1): 5477. https://doi.org/10.1038/s41467-019-13443-4.