



Emergence of networks of shared restriction-modification systems in phage–bacteria ecosystems

RASMUS SKYTTE ERIKSEN^{1,2}, NITISH MALHOTRA³, ASWIN SAI NARAIN SESHASAYEE³,
KIM SNEPPEN^{1*} and SANDEEP KRISHNA^{4*} 

¹Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

²Statens Serum Institut, Copenhagen, Denmark

³National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, India

⁴Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru, India

*Corresponding authors (Emails, sneppen@nbi.dk; sandeep@ncbs.res.in)

MS received 7 December 2021; accepted 18 March 2022

Restriction-modification (RM) systems are the most ubiquitous bacterial defence systems against bacteriophages. Using genome sequence data, we showed that RM systems are often shared among bacterial strains in a structured way. Examining the network of interconnections between bacterial strains within genera, we found that many strains share more RM systems than expected compared with a suitable null model. We also found that many genera have a larger than expected number of bacterial strains with unique RM systems. We used population dynamics models of closed and open phage–bacteria ecosystems to qualitatively understand the selection pressures that could lead to such network structures with enhanced overlap or uniqueness. In our models, we found that the phages impose a selection pressure that favours bacteria with greater number of RM systems, and higher overlap of RM systems with other strains, but in bacteria-dominated states, this is opposed by the increased cost-to-growth rate of these bacteria. Similar to what we observed in the genome data, we found that two distinct bacterial strategies emerge – strains either have a greater overlap than expected, or, at the other extreme, have unique RM systems. The former strategy appears to dominate when the repertoire of available RM systems is smaller but the average number of RM systems per strain is larger.

Keywords. Bacteriophage; evolution of RM systems; networks; restriction-modification systems

1. Introduction

The lives of bacteria can be harsh, with several species competing for limited resources in an environment that is often toxic or contains other stressors. In addition to these already bleak conditions, bacteria must also defend against a ubiquitous parasite – the bacteriophage – which wreaks havoc on bacterial populations. These bacterial viruses outnumber the bacteria 10 to 1

and are a major contributor to bacterial death, for instance, in the ocean, where phages are responsible for ~20% of the bacterial deaths (Suttle 2007; Knowles *et al.* 2016; Bar-On *et al.* 2018). As a consequence, bacteria have evolved strategies for phage evasion and defence, while the phages in turn have evolved counter strategies. This co-evolutionary ‘arms race’ between phages and bacteria is likely a major selective force shaping the bacterial genome (Forterre and Prangishvili 2009; Stern and Sorek 2011; Koonin and Wolf 2012; Makarova *et al.* 2013).

This article is part of the Topical Collection: Emergent dynamics of biological networks.

Supplementary Information: The online version contains supplementary material available at <https://doi.org/10.1007/s12038-022-00274-7>.

One bacterial defence mechanism is the restriction-modification (RM) system, which is the focus of this article. RM systems constitute a simple mechanism for a bacterium to distinguish self from non-self. An RM system consists of two parts: a methyltransferase, which methylates (modifies) specific recognition sites in the genome, and an endonuclease, which cleaves (restricts) the genome at the recognition sites when they are not methylated. When a phage infects the bacterium, its genetic material will then be restricted by the RM systems unless it is methylated at the corresponding recognition sites. In rare cases, the phage genome will, by chance, become methylated by the methyltransferase at all its recognition sites before the endonuclease of the RM system acts. When this happens, the phage genome is said to have ‘escaped’ restriction, and can now freely continue its infection cycle inside the bacterium, commonly leading to lysis and the release of new phage particles, all of which are now methylated at the recognition sites (schematically depicted in figure 1). Thus, RM systems are not perfect defences, with the probability of escape ranging from $\sim 10^{-2}$ to 10^{-8} (Arber 1965; Moineau *et al.* 1993; Dupuis *et al.* 2013; Pleška *et al.* 2016). RM systems are remarkably widespread among bacteria, being present in $\sim 90\%$ of all bacterial strains (Vasu and Nagaraja 2013) and occupying on average $\sim 0.3\%$ of bacterial genomes (Makarova *et al.* 2013). However, since the protection is not perfect, bacteria seemingly need to invest in carrying several distinct RM systems (Vasu and Nagaraja 2013).

Bacteria have mechanisms for sharing their genetic material (e.g. by transduction and conjugation), and as a result, a given RM system is not likely to be unique to its bacterial strain but may be present in other bacterial strains in the surrounding environment. Indeed, sequencing data of bacteria of the same genus reveal that the composition of RM systems within these strains share many common RM systems among each other (Roer *et al.* 2016; Fullmer *et al.* 2019). In figure 1 we show schematically what the sharing or *overlap* of RM systems between related bacterial strains means for their ability to defend against different epigenetic phage variants. A phage infecting a particular bacterium may already be methylated against some of the RM systems present in that bacterium, rendering these RM systems ineffective. However, by sharing RM systems between related bacterial strains, the bacteria, on average, increase the number of effective RM systems against any given phage, which under some conditions may improve their chances of surviving a

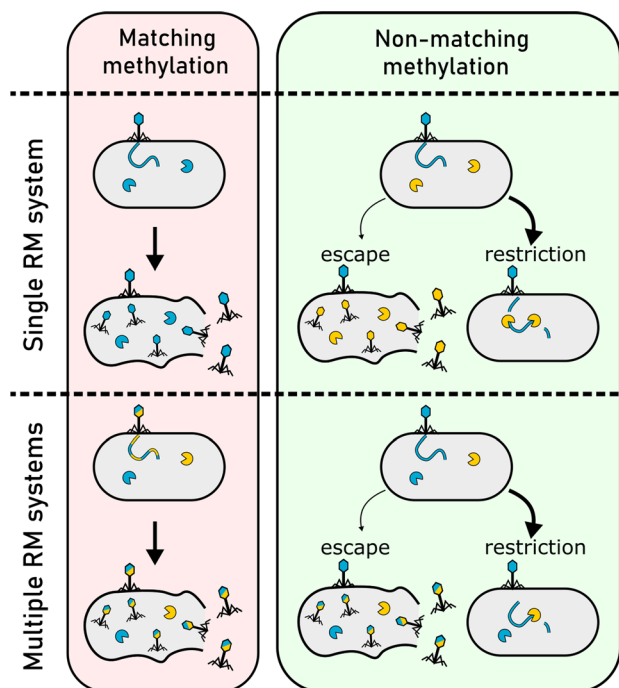


Figure 1. The RM defence mechanism. When a phage infects a bacterium with an RM system, by which it has previously been methylated (indicated by a matching colour), then it will be invisible to the RM system and will lyse the host, producing a burst of offspring phage. If the phage lacks the appropriate methylations, the RM system will, in a majority of times, restrict the phage and thereby prevent it from replicating. However, a phage will sometimes escape the restriction and successfully replicate despite lacking the appropriate methylations, and thus produce offspring that carry the methylation pattern that protects it from the host RM system. With multiple RM systems, the phage must carry the combined methylation pattern against *all* RM systems in the host to reliably infect and lyse the host. If it does not, the RM systems which the phage is not protected against will each attempt to restrict the phage. The probability of escaping restriction is then dependent on the efficacy of each of these RM systems.

phage encounter. We do not yet understand under which conditions such sharing or overlap will be advantageous, and under which it will be disadvantageous for bacterial strains.

This is complicated by the microbial diversity of ecosystems. With several bacterial genera coexisting in the same local environment, the community interactions between genera will further shape the selection pressures on any given bacterial strain. To begin unravelling these selection pressures, we limit our focus to the selection pressures acting within a single ecological niche. Our models thus consist of a set of bacterial strains that grow on the same limiting

resource, can all be infected by the same phage parasite, and have a common repertoire of RM systems they can acquire for defence against this phage (for example, through horizontal gene transfer from other strains or free DNA in their local environment).

For simplicity, whenever we use the term ‘ecosystem’, we refer to such a set of strains. In principle, strains that inhabit a particular ecological niche and satisfy the above three properties may be quite diverse, spanning multiple genera. Thus, our models do not assume that all strains are from a single genus. However, when comparing with data, in the absence of specific information about which strains across which genera are subject to infection by specific phage variants, we stratify our sequencing data at the genus level. This assumption should, of course, be revisited as more detailed information about cross-genus interactions in bacterial ecosystems becomes available.

An ecosystem consisting of multiple bacterial strains with a diverse collection of RM systems can be thought of as a *network* of strains interconnected by the RM systems they share. Our main interest in this article is to understand how such networks of bacterial strains and their RM systems behave and how they are shaped by evolution: Do these networks have non-random structures? If so, how does the structure dictate the population dynamics? Relatedly, what kind of selection pressures lead to the emergence of these non-random structures? For instance, do certain network structures enhance the coexistence of strains or increase the net population of the ecosystem?

In this article, we first examine, in section 2, the RM systems present in publicly available bacterial genomes in the NCBI RefSeq database (O’Leary *et al.* 2016). Specifically, across 42 bacterial genera, we analyse the overall features of RM system use, such as the distribution of the number of RM systems and the degree to which RM systems are shared across strains.

In section 3, we then examine the structures of the networks formed by bacterial strains within each genus and the RM systems they contain. We compare these networks with ensembles of randomized networks obtained from a suitable null model and demonstrate that the real networks often have a higher than expected sharing of RM systems, and many also exhibit a larger than expected number of strains with unique RM systems. Then, in section 4, we develop a population dynamics model that extends previous approaches to include the sharing of RM systems among bacteria. The model allows us to examine the dynamical effects of RM system sharing in closed and open phage–bacteria ecosystems, and thereby to qualitatively

understand what kind of selection pressures may result in the emergence of these non-random network structures. We find, in these models, that two kinds of strategies seem to work well for bacterial strains – they either have multiple RM systems which they share with other strains, or they have fewer but unique RM systems – signatures of which are also found in our analysis of bacterial genomes.

2. Distribution of RM systems from genome sequencing data

Our data set consisted of the complete genomes of 1417 unique bacterial strains across 42 genera which we analysed for the presence and absence of 333 different type II RM systems (see [Methods](#) and supplementary section 1).

Figure 2 shows the distribution of RM systems across our sequenced genera in two different ways. At the genus level, it appears that the majority of genera have few RM systems on average, but with a substantial fraction containing more than one RM system on average (figure 2A). Figure 2B shows the same distribution at the level of individual strains within a genus, aggregated across all genera (the contribution of each genus is weighted by the number of samples within our data set; see figure 2F). As expected, the distribution of RM systems among strains has a longer tail than the same distribution at the genus level, due to averaging in the latter. The corresponding distributions for each genus are separately shown in supplementary figure 2. These distributions are unimodal with varying widths. Most are peaked at a small number, 1–2, of RM systems per strain. However, there are a few genera, such as *Helicobacter* and *Neisseria*, which are peaked at a much higher number of RM systems per strain and contribute to the long tail in figure 2B.

Figure 2C and D shows the complementary distributions describing how many genera or strains a given RM system is found within. supplementary figure 2 shows the corresponding distributions within each genus separately. In some genera, such as *Bacillus* or *Staphylococcus*, each RM system is found in a small number of strains. In other genera, the distribution is more skewed, where one or two RM systems are widespread among the bacterial strains in that genus and the remaining RM systems are rare (e.g. *Mycobacterium* or *Phaeobacter*). One also can see genera with a broader distribution, such as *Helicobacter* or *Neisseria*. It is noteworthy that a substantial number of RM systems are found both in multiple

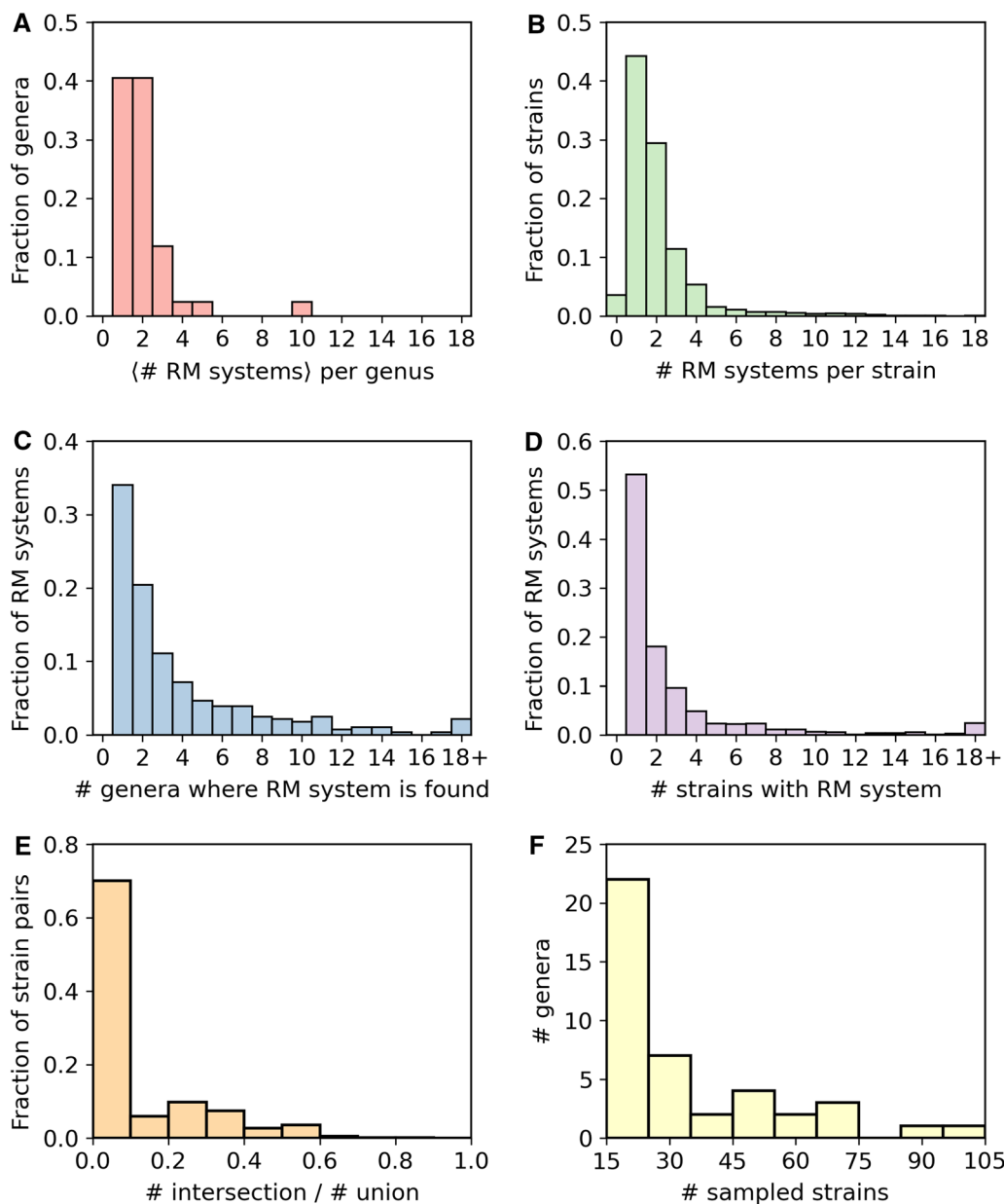


Figure 2. Analysis of the RM system distribution in the data set consisting of 279 unique RM systems and 1417 unique bacterial strains (see [Methods](#)). **(A)** The average number of RM systems per strain $\sum_{r,s} Q_{rs}^g / S(g)$, for each genus g , where $S(g)$ is the number of strains in genus g (see **F**) and Q_{rs}^g is the presence–absence matrix whose elements are assigned 1 if an RM system r is found in the bacterial strain s of genus g , and 0 otherwise. **(B)** The number of RM systems present in each strain s within a given genus is $\sum_r Q_{rs}^g$. The average of all these histograms is shown, weighted by $1/S(g)$ to account for the differences in sample sizes between genera (see supplementary figure 2 (green bars) for genus-level distributions). **(C)** Distribution of RM systems across genera, i.e. a plot of $\sum_g P_{rg}$, where P_{rg} is a genus-level presence–absence matrix whose elements are 1 if an RM system r is found in the genus g , and 0 otherwise. **(D)** The number of strains each RM system is found in is $\sum_s Q_{rs}^g$. The average of all these histograms is shown, weighted by $1/S(g)$ (see supplementary figure 2 (purple bars) for genus-level distributions). **(E)** The overlap of RM systems for each strain pair (i, j) in a given genus is defined as the ratio of the number of shared RM systems, labelled ‘intersection’, to the total number of distinct RM systems, labelled ‘union’, in that pair (see [Methods](#)). An unweighted average of these histograms is shown (see supplementary figure 2 (orange bars) for genus-level distributions). **(F)** Distribution of samples across genera, $S(g)$.

genera as well as in multiple strains within a genus. This naturally leads one to ask how often these RM systems are shared across strains. We define the RM system ‘overlap’ between a pair of strains to be the ratio of the number of RM systems shared by the two strains to the total number of unique RM systems across both strains, which will be a number ranging from 0 (no overlap) to 1 (complete overlap). Figure 2E shows the distribution of this measure of overlap, computed first for all pairs within the same genus, and then summed over all genera. Most pairs of bacterial strains do not have any common RM systems. However, there are a substantial fraction of pairs that have a non-zero overlap. One may wonder whether larger a greater overlap occurs only in situations where each member of the pair has a very small number of RM systems with one or two in common (e.g. a pair with two RM systems each and one in common would give an overlap of 0.33). In fact, in supplementary section 5, we show that there are many pairs with a large number of RM systems each of which also have many in common, in particular, within *Helicobacter*. Examining the overlap distribution within each genus separately (see supplementary figure 2), we find that most distributions have a high peak at zero overlap, like the aggregated distribution in figure 2E, except for *Helicobacter* and *Phaeobacter*, where the distribution peaks at an overlap of 0.25.

In summary, the distributions in figure 2 and supplementary figure 2 show that while many bacterial strains have RM systems they do not share with other strains, a substantial fraction of RM systems are found in multiple strains of the same genus. This feature was observed to varying degrees in the genera we analysed.

3. Non-random networks of interconnected bacterial strains due to shared RM systems

A natural representation of an ecosystem containing multiple strains of bacteria and their RM systems, some of which are shared, is provided by a *network* of interconnected bacterial strains and RM systems (Pavlopoulos *et al.* 2018). In such a representation, bacterial strains are linked to other strains via the RM systems they share. More precisely, a network (or graph) of such a system consists of a set of nodes connected by links. Figure 3 shows three examples of such a network representation from our data set of the genera *Lactococcus*, *Xanthomonas* and *Mycoplasma*. These were chosen as examples of bacterial genera with increasing average RM abundance, i.e., networks

that range from quite disconnected to well-connected. Each bacterial strain is shown as a blue node and each RM system is shown as a red node. If a specific RM system is present within a particular bacterial strain, this is represented by a link connecting those two nodes. Thus, these networks are bipartite: there are no links between two red or two blue nodes, and the ‘presence–absence matrix’ described in figure 2 specifies the links that exist between red and blue nodes. In supplementary section 3 we show the network structures for all 42 genera in our data set. For genera where each RM system is found in a small number of strains, such as *Bacillus* or *Staphylococcus*, the corresponding networks consist of multiple disconnected pieces. Where one or two RM systems are widespread among the bacterial strains in that genus and the remaining RM systems are rare, such as *Mycobacterium* or *Phaeobacter*, the networks have a star-like structure, with many strains connected to one or two RM systems. More densely connected networks are the ones with a broader distribution, such as *Helicobacter* or *Neisseria*.

The network structure will affect the ability of phages to infect bacterial strains. A genus with a dense network will have strains with many overlapping RM systems, which means that phages successfully infecting one strain will often have the methylation pattern to avoid some RM systems in a different strain, whereas disconnected bacterial strains in a sparser network will generally be immune to phages that arise by lysing another strain. We, therefore, expect such networks to have a non-random structure reflecting the selection pressures that shape the evolution of bacterial strains and their RM systems in the presence of phage. We are most interested in the following question: given a repertoire of RM systems and given the constraints that determine how many RM systems each strain contains, what are the selection pressures that may lead to enhanced *sharing* of RM systems between strains? Therefore, we have compared each network with a null model, consisting of the same number of bacterial strains and available RM systems, with each bacterial strain retaining the same number of RM systems, but where the composition of RM systems is randomized. In other words, for each network, we create an ensemble of randomized networks, where each blue node (bacterial strain) has the same degree (number of links connected to it) as in the real network but is now connected to randomly chosen red nodes (RM systems). This corresponds to a ‘neutrally evolving’ system where each strain independently loses and gains random RM systems from a common repertoire while

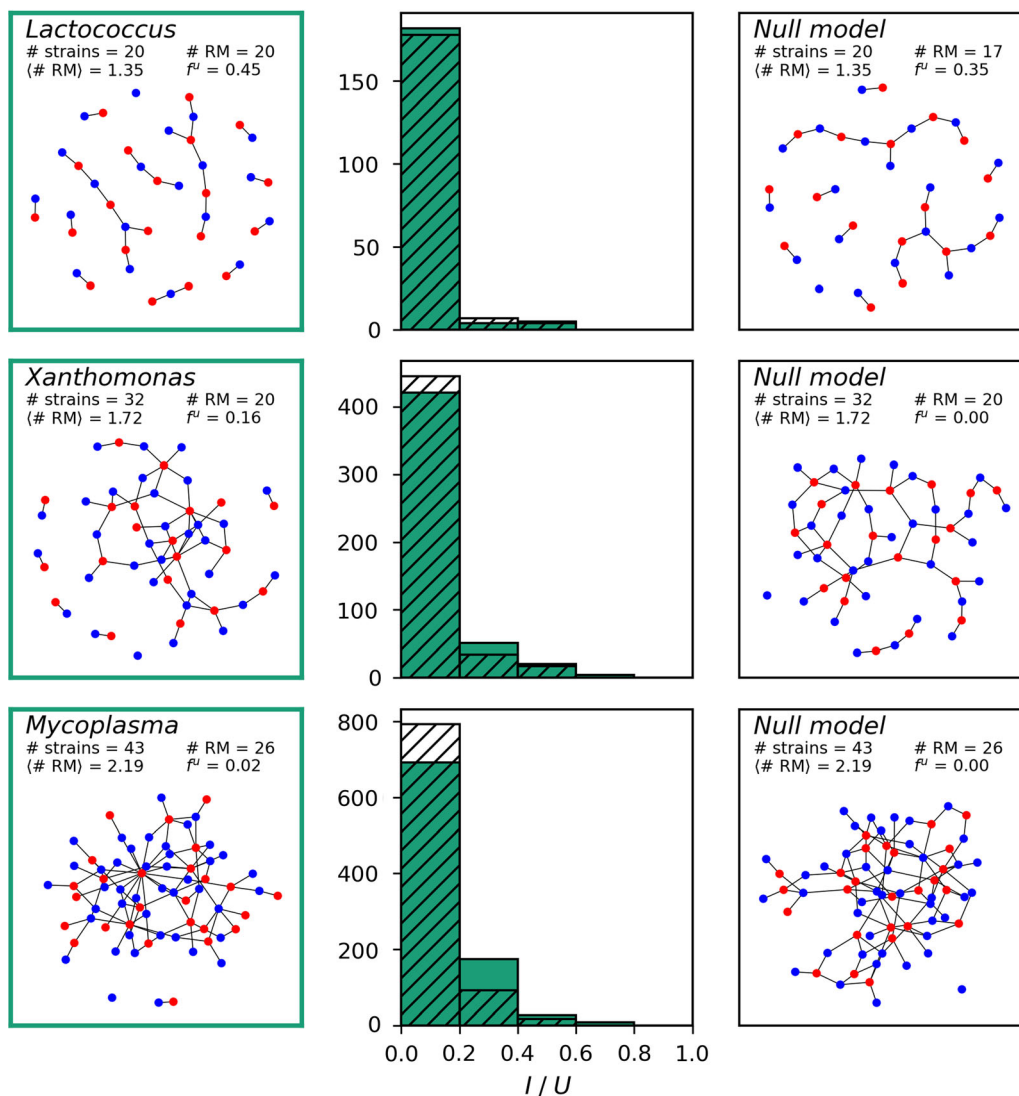


Figure 3. Networks of RM systems. The left column shows bipartite network representations of the distribution of the RM systems among the bacterial strains for three different genera: *Lactococcus*, *Xanthomonas* and *Mycoplasma*. Each bacterial strain is shown as a blue node and each RM system is shown as a red node, with links between the two corresponding to the RM system being found within the bacterial strain. The right column shows, for each case, a null model network where the RM systems are randomized for each strain (see [Methods](#)). The distributions of the overlap in the real (solid green bars) and random network (hatched bars) are shown in the central column. Overlap between a pair of bacterial strains is defined as the ratio of the number of RM strains common to both (denoted as ‘intersection’, or I) to the total number of RM strains across both strains (denoted ‘union’, or U).

keeping the number of RM systems it has constant. In such a neutrally evolving scenario there is no selection pressure for or against *sharing* of RM systems.

Comparing networks from this null model with networks generated from the sequenced data allows us to investigate the influence of selection by a phage parasite on the *overlap* and *uniqueness* of RM systems across bacterial strains.

The right column in figure 3 shows one randomized network from the ensemble corresponding to each of the three real networks. There are notable differences

in the distribution of the RM system overlap between real and random networks, as can be observed visually (notice the relatively fewer red nodes with just one link in the random networks) as well as quantitatively in the histograms in the middle panels. Figure 4 demonstrates this more rigorously by comparing real networks against an ensemble of 100 random networks, for all the genera. Figure 4A shows that the average overlap within each genus (as defined previously) differs significantly from the null expectation for many genera. The genera are arranged in order of

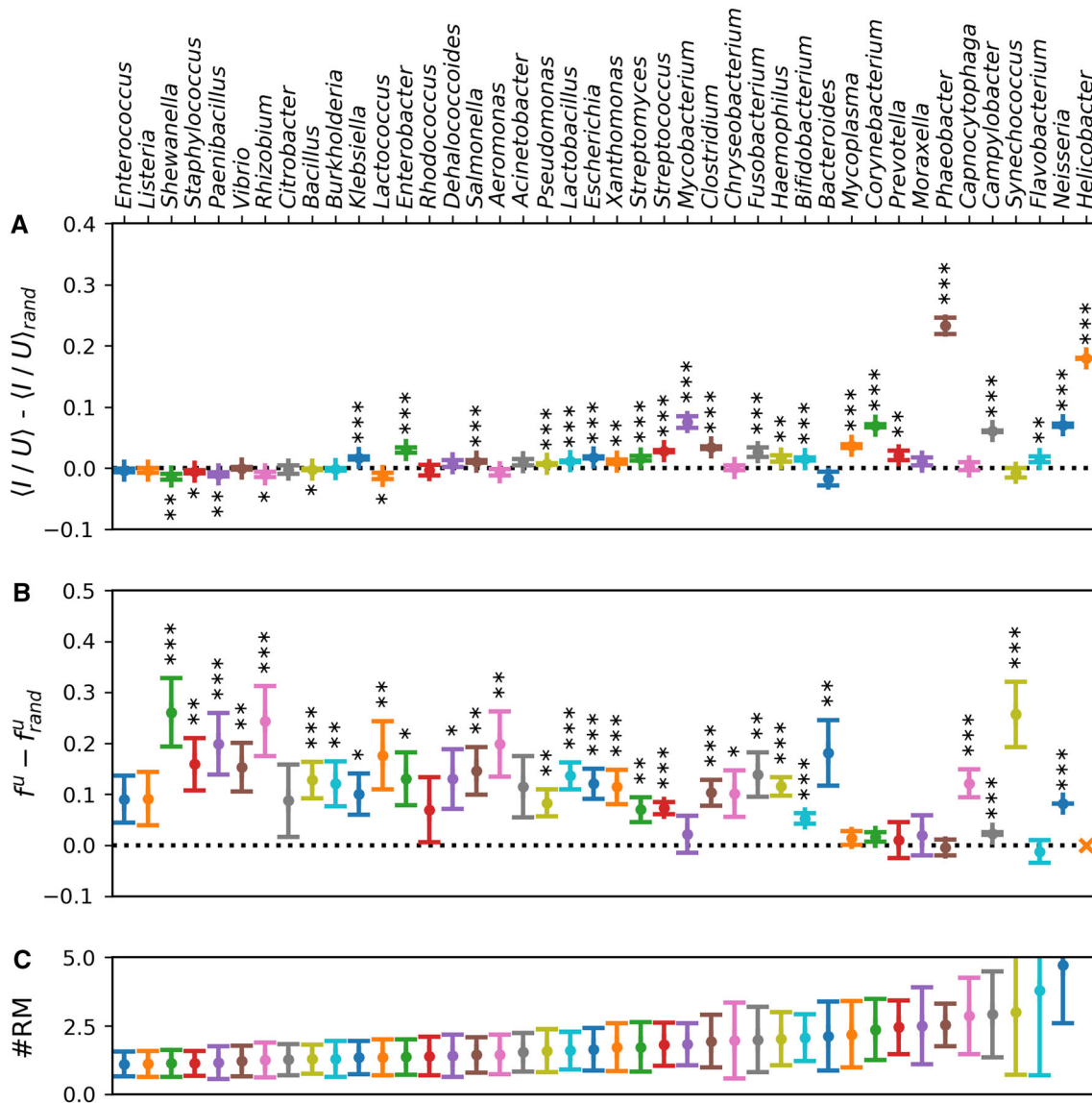


Figure 4. Differences between genome data and the null model. For each of our 42 genera, we generated 100 samples of the corresponding null model network (see [Methods](#)). For each genus, the plots of (A) the difference in the overlap (averaged over all pairs of strains in that genus) between the real and randomized networks, (B) the fraction of bacterial strains with unique RM systems compared with the average of this quantity across the randomized networks for that genus, and (C) the average number and standard deviation of RM systems per strain, are shown. The y-axis limits were chosen to enhance clarity and do not show the data for *Helicobacter*, whose strains have 9.8 RM systems on average, with a standard deviation of 2.9. (A–B) Points indicate averages over the 100 randomized networks, error bars indicate standard deviation, and asterisks denote that the quantities shown are different from zero with significance levels of 95%, 99% and 99.9%, respectively.

increasing average RM systems per strain, and interestingly, the departure from the null expectation is stronger for genera with a larger average number of RM systems. The visual observation that the real networks seem to have more strains with a unique set of RM systems (unconnected clusters with a single blue node and one or more red nodes) is quantified in figure 4B. We define the measure f^u to be the fraction

of bacterial strains that share none of their RM systems with other strains. We compute f^u for each real network (i.e. for each genus) and for each network in the corresponding randomized ensemble. Except for a few cases, all genera appear to have significantly more strains with unique RM systems than the random expectation. Note, however, that this measure is less stable to undersampling than the overlap measure, and

so we might be overestimating these values (see supplementary figure 1).

We conclude that for a substantial number of genera, selection pressures have led to the emergence of non-random network structures. Specifically, many strains of the same genus, even if they have multiple RM systems, more often than expected contain RM systems that are unique to that strain and not shared with other strains of that genus. However, there also remains a substantial overlap in RM systems between strains, indicating that any selection pressure to have unique RM systems is not so strong as to entirely eliminate the overlap of RM systems in the ecosystem.

In the rest of this article, we use dynamical models of a simplified phage–bacteria ecosystem with shared RM systems to obtain a qualitative understanding of the kinds of selection pressures that could lead to such non-random network features.

4. Evolution of shared RM systems in a mathematical model of phage–bacteria ecosystems

We studied the dynamics of phage and bacteria with RM systems in a well-mixed system using equations based on the models of Frank (1994), Sneppen *et al.* (2015), and Eriksen and Krishna (2020b), which we generalized to include sharing of RM systems across bacterial strains (see Methods). Our model implemented sharing of RM systems similarly to Pagie (2000), which, in contrast to our model, considers a spatially extended, not well-mixed, system. Each RM system is assumed to have a corresponding efficacy ω_r and growth rate penalty described by the parameter γ_r . For each bacterial strain, its overall growth rate is assumed to be the product of the γ_r s of all RM systems it contains, while the overall efficacy of the RM defence can be calculated from the methylation status of the incoming phage (see Methods). First, we describe a few simple scenarios to build some intuition about the advantages and disadvantages of sharing RM systems. We term these ‘closed’ ecosystems because the number and kind of strains and RM systems are fixed, although their populations may change with time. We then examine a model of an evolving ‘open’ ecosystem where new strains of bacteria invade, sometimes introducing novel RM systems, and others go extinct. We use these models to understand what selection pressures may lead to networks features similar to those we observed in sections 2 and 3.

4.1 Costs and benefits of overlapping RM systems in simple closed ecosystems

4.1.1 Sharing of RM systems may help increase individual or net population: In figure 5A, we show the case where there is no overlap of RM systems between two bacterial strains. Here, as previously shown (Frank 1994; Sneppen *et al.* 2015; Eriksen and Krishna 2020b), due to the uniqueness of the RM systems, phage variants have little effect on bacteria that are not from their parent strain. As a consequence, the population density of each bacterial strain is limited primarily by its corresponding phage variant, and all strains reach roughly the same population level irrespective of their intrinsic growth rates. This result generalizes to any number of strains with non-overlapping RM systems, up to the limit imposed by the phage burst size.

Allowing for overlap between the RM systems changes the situation, even in a simplified example where just two distinct RM systems exist. Here, as shown in figure 5B, there are three possible combinations to consider, each with its own epigenetic phage variant. The phage that is methylated at the recognition sites of both RM systems can easily infect all hosts in the system, while the other two phages are primarily limited to their corresponding hosts. Despite the decreased growth rate of the host that carries both RM systems, it may reach a higher level than the other bacteria due to the increased effectiveness of its corresponding phage variant. In other words, due to the overlap, a strain with multiple RM systems may compete better against other strains which have fewer RM systems.

4.1.2 Sharing of RM systems may leave a strain vulnerable to immune phages: However, carrying several RM systems is not always the best strategy, even when they impose only a small cost to the bacterial growth rate. In figure 5C we show what happens if the phage develops immunity against one of the RM systems, e.g. by restriction site avoidance (Vasu and Nagaraja 2013). In this example, the phage is immune to the yellow RM system, and the strain carrying two RM systems now has an RM system that is ineffective against the phage. This, in turn, prevents the strain from creating a unique epigenetic phage variant. As a consequence, the strain is now in direct competition with the bacterial strain which has not invested in the now ineffective RM system. This competition results in the competitive exclusion of the strain with two RM systems.

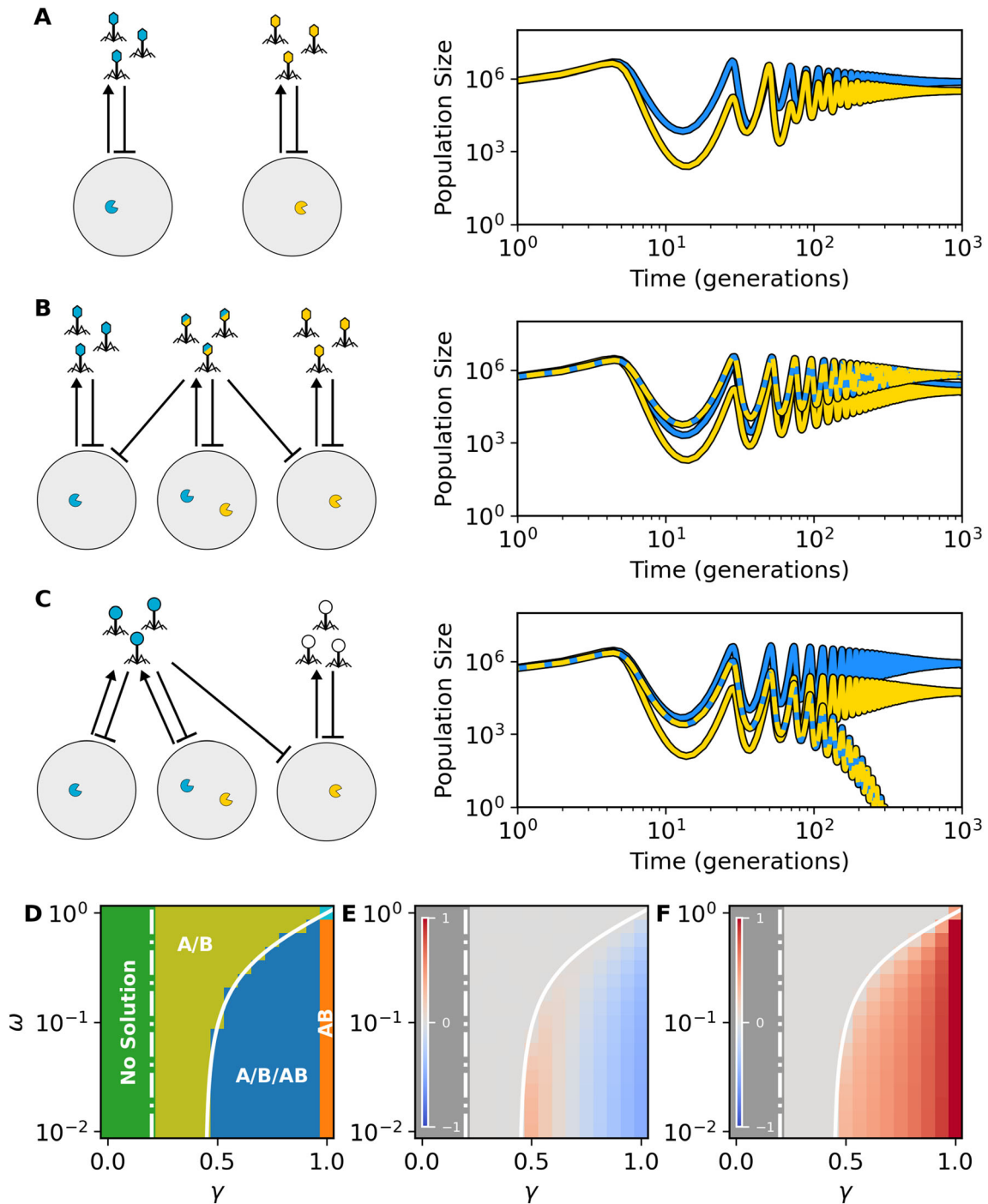


Figure 5. The population dynamics of closed ecosystems. (A) When the bacteria have unique RM systems, all bacterial strains achieve high population densities determined by the efficacy of their respective RM systems. Each phage strongly suppresses their respective hosts (barred arrows) and produce new phages (regular arrows). (B) When the RM systems are shared between different hosts, the cross-couplings become more important as some phages will freely attack several hosts, but this in turn does not produce more phages of that epigenetic variant. The bacterial strains that are now preyed upon by multiple phage variants are accordingly less abundant. (C) If a phage has native immunity against an RM system (yellow in this case), it cannot be methylated by that RM system and only an unmethylated (white) and a blue phage variant will exist. Any host which has invested in the ineffective blue RM system is now at a disadvantage and likely to be competitively excluded. (D) Regions of possible coexistence solutions for the system in (B) where all RM systems have identical efficacies and costs. (E) Relative gain and loss of population size for this system compared with the system in (A). (F) Relative fraction of b_{AB} to $b_A + b_B$ for this system. (A–F: Simulations use default parameters values, except for $\beta = 25$.)

Note that for the phage, developing immunity to the RM systems can be a strong selective advantage. In supplementary section 8, we show that in the scenario where a non-immune and immune strain of phages compete in the above case (i.e. we test the combined scenario of figure 5B–C), the immune strain will out-compete the non-immune strain if it has no fitness cost associated with immunity.

4.1.3 Robustness to parameter variation: In the special case where all RM systems have identical parameters, i.e. they incur the same growth-rate penalty ($\gamma_r = \gamma$) and have the same efficacy ($\omega_r = \omega$), we can derive the full solution analytically (see supplementary section 9.1). In particular, for the system shown in figure 5B, we can compute when each strain will survive or go extinct, as shown in figure 5D. When the growth rate, γ , is less than the dilution rate, $\alpha = 0.2$, no bacteria will survive in the long term. When $\alpha < \gamma < f(\omega)$ (where $f(\omega)$ is a particular function which we derive in supplementary section 9.1) bacteria with single RM (A or B) begin to survive in the long term, but the strain containing both RM systems (AB) is not yet able to overcome its growth rate deficit. When $\gamma > f(\omega)$, the host with both RM systems now also survives and all three hosts coexist. At $\gamma = 1$, the bacteria with single RM systems lose their growth advantage and only the bacteria with both RM systems persist. However, when $\gamma = \omega = 1$, the system is fully degenerate and all solutions are possible.

Panels E and F in figure 5 show how the population sizes change with the inclusion of the double RM strain (AB). Figure 5E shows the total population size in the ecosystem compared to the scenario with the A/B solutions alone. When the growth rate is slightly larger than $f(\omega)$, there is an increase in the overall population, but as the growth rates increase further, the net population size decreases due to the presence of the double RM strain. Figure 5F shows that the population size of the double RM strain is always greater than the population size of the single RM strains, which highlights the individual benefit of having several RM systems.

In supplementary section 9.2, we also consider the high-diversity limit, that is, a scenario where a population of D bacterial strains in which there are T ‘triplets’ that share RM systems in the $A/B/AB$ motif (‘hierarchical triplets’) while the other $D - 3T$ strains have unique RM systems. In this case, the range of parameters where the ecosystem can support the double RM system strains is dependent on the number of triplets T present in the system. As more and more triplets are added, the range of parameters that support the

triplets increase, suggesting that having more RM systems is an invasive strategy. This is further substantiated by considering triplets of the $AB/BC/AC$ motif (‘looped triplets’), where the range of parameters that support the triplets not only increases faster as T increases than for the hierarchical triplets but also covers a larger range of parameters.

4.2 Evolution of bacterial networks in an evolving open ecosystem

The above dynamics in model closed ecosystems provides some intuition about the advantages and disadvantages of sharing RM systems across strains and the kinds of selection pressures that may result in the gain or loss of RM systems. We next consider a more relevant open ecosystem where we periodically add new bacterial strains with potentially new RM systems (and their corresponding phage variants) while removing those that go extinct. As a consequence, we obtain a large open ecosystem that consists of a changing complex interconnected network of bacterial strains, linked by their RM systems. The new strains that are being added have the same average RM abundance as the existing strains but the RM systems are chosen randomly from a set of K possible RM systems. In figure 6A–C we show an example simulation with $K = 50$. Here, the number of available RM systems is relatively low, and thus the dynamics are more affected by the sharing of RM systems. However, since any particular bacterial strain can have any combination of these RM systems, there are roughly 10^{15} possible distinct strains for this value of K . In supplementary section 10 we compare examples with higher values of K .

4.2.1 Phage-dominated and bacteria-dominated states: In the $K = \infty$ limit, the system reduces exactly to that studied by Sneppen *et al.* (2015) and Eriksen and Krishna (2020b), where each strain has unique RM systems and there is zero overlap. Even with a more limited repertoire of available RM systems at finite K , and the presence of overlap, some aspects of the dynamics remain the same:

- Early on, competition between bacteria is weak and the phage pressure is high. Due to the relative lack of competition, the diversity (i.e., the number of strains with a non-zero population) increases and the phage pressure decreases (see figure 6A). Under these conditions, the ability to defend against

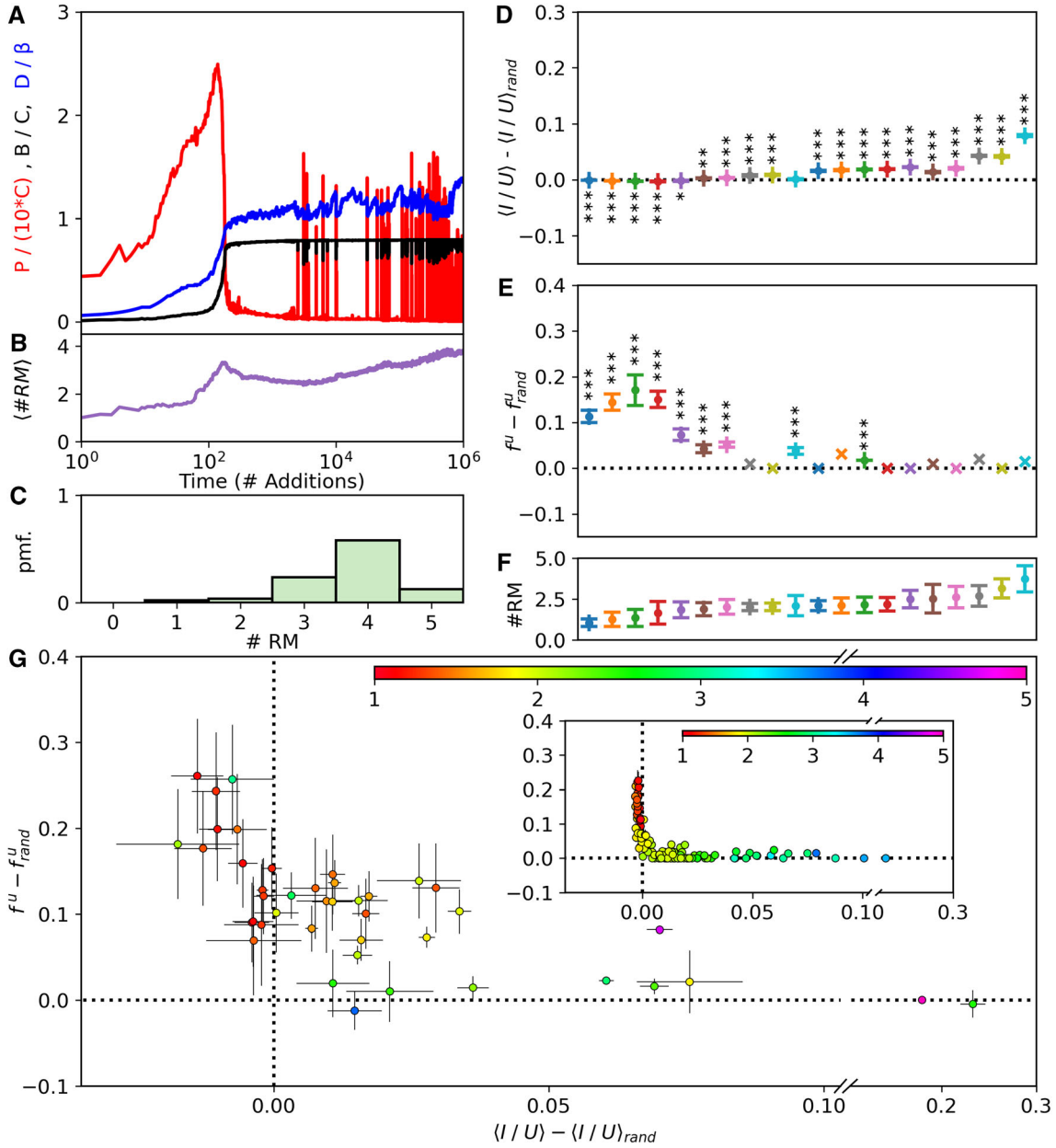


Figure 6. Two bacterial strategies in genome data and open ecosystem simulations. **(A)** Example of the population dynamics of open ecosystems. Here, every $T = 10^3$ generation, we add a new bacterial strain which carries a novel combination of $K = 50$ possible RM systems (see [Methods](#)). The plot shows the total bacterial density, B , in units of the carrying capacity C (black curve), the total phage density P (red curve) in units of $10 \times C$, and the diversity D (blue curve) in units of the phage burst size β . **(B)** The average number of RM systems per strain $\langle \#RM \rangle$ in the system over time. **(C)** Distribution of the number RM systems each bacterial strain carries at the end of the simulation. **(D–F)** We run our simulation with $K = 800, 400, 200, 100$ and 50 and sample the network after the addition of $10^3, 10^4, 10^5$, and 10^6 new strains. For each sampled network, we compute **(D)** the difference in overlap with 25 realizations of the corresponding randomized network, **(E)** the difference in the fraction of strains with unique RM systems compared with random expectation, and **(F)** the average number of RM systems per strain for each case. Error bars indicate standard deviation. Asterisks denote significance levels of 95%, 99% and 99.9%, respectively. **(G)** Scatter plot of the two key network characteristics for each genus, obtained by comparing genome data with the null model (from figure 4). The colour of each dot indicates the average number of RM systems per strain in that genus. **(Inset)** The same as **(G)** but for the simulations of the open ecosystems **(D)** and **(E)**. Here, we have included the data for 6 repeats of the simulations. Both the data and simulations indicate the presence of roughly two distinct bacterial strategies. Strains either invest in multiple RM systems with relatively large overlap between strains, or in unique RM systems that are not shared by other strains. In the model, the strains transition from the second strategy to the first at later times when the average number of RM systems is larger.

phages is more important than the ability to grow fast, and therefore, bacteria with several RM systems (which typically have a smaller growth rate) are highly competitive and the average RM abundance increases (see figure 6B).

- As the total bacterial population increases and approaches the carrying capacity, the ecosystem goes from a phage-dominated state to a bacteria-dominated state. Bacteria with fewer RM systems (i.e., bacteria with typically higher growth rates) begin to become competitively viable, which is why the average number of RM systems per strain initially reduces in this phase. However, this does not, in turn, drive the phage pressure back up since the diversity continues to increase and the phage now has a decreasing probability of hitting vulnerable bacteria. In this high-diversity bacteria-dominated state, there are thus two opposing selection pressures on the number of RM systems in each strain. For the same reason as seen in the closed ecosystems, there is an advantage in having a larger number of RM systems, and this is opposed by the corresponding cost to the growth rate. The net result is that the average number of RM systems slowly rises and will eventually stabilize when these opposing forces balance (see supplementary section 10).
- However, the distribution of RM systems per bacterial strain is quite broad (see figure 6C). supplementary section 10 shows that this distribution widens with decreasing diversity of available RM systems, i.e., with decreasing K .
- In this bacteria-dominated state, the diversity increases slightly beyond the limit set by the phage burst size β (Sneppen *et al.* 2015). Eriksen and Krishna (2020b) showed that, in the absence of overlap, this happens because of a trade-off between the growth rates of bacterial strains and the strength of their RM systems. In supplementary section 10.3, we show that this explanation also applies to this work, where we allowed for sharing of RM systems across bacterial strains.

4.2.2 Dynamics of overlap of RM systems and intermittent resurgence of the phage: We now focus on the dynamical behaviour that differs from that seen by Eriksen and Krishna (2020b), namely, the dynamics of the overlap of RM systems and the structure of the bacteria–RM networks. In the bacteria-dominated phase, the competition between bacteria results in a strong selection for RM systems with small costs (large γ_r). This restricts the number of viable RM systems and increases the

sharing of RM systems. Thus, the overlap between pairs of strains rises. We find that the number of hierarchical and looped triplets, discussed in the context of closed ecosystems in section 5, increases with time (see supplementary section 10.1). The increase of overlap as time passes, in turn, increases the chance that the phages are partially protected against the RM systems of the bacterial strains in the ecosystem. When this overlap of RM systems becomes sufficiently large, the phages sometimes experience short resurgences and the dynamics become increasingly noisy, as can be seen in figure 6A. These phage-dominated states are not very stable and typically disappear with the next addition of a bacterial strain (see supplementary section 10.2). The timing and duration of these intermittent resurgences are not understood, but seem to depend on the number of available RM systems K (see supplementary section 10).

4.2.3 Dynamics of the network structure: two distinct bacterial strategies: Figure 6D–F shows an analogue of the plots in figure 4 for our simulated open ecosystem. Qualitatively similar to what is seen in figure 4, we observe that the networks that emerge in our simulations have a larger overlap than random expectation when they contain a larger average number of RM systems. The pattern is more distinct in our simulations – networks that have a lower average number of RM systems more clearly have a lower than expected overlap. We also observe that strains in our evolved networks are more likely to have unique RM systems than the random expectation. Here again, there is a more distinct pattern in our simulations as a function of the number of RM systems, but the trend is the same. The genome data also shows another interesting pattern when we plot, for each genus, the fraction of strains with unique RM systems compared with random expectation ($f^u - f_{rand}^u$) vs. the average overlap for that genus compared to random expectation ($\langle I / U \rangle - \langle I / U \rangle_{rand}$). As shown in figure 6G, these quantities are anti-correlated, with a higher than expected uniqueness being seen when the overlap is lower than expected and vice versa. This pattern is reproduced in an even starker form in networks from our simulated open ecosystem, as shown in the inset of figure 6G. In the simulated ecosystem, there appear to be two fairly distinct strategies – strains within a genus either have a larger overlap than expected but almost the same number of strains have unique RM systems as a random network, or they have the same or lower overlap than the random expectation but there are more strains in that genus with unique RM systems. In the simulations, the former strategy appears to dominate at

later times as the average number of RM systems and the number of hierarchical and looped triplets increases. It also dominates more in our simulations when the repertoire of available RM systems is smaller, i.e., when K is smaller. In the genome data, there is a similar but weaker correlation with the average number of RM systems, and the two strategies are not so distinctly separated although the trend is qualitatively similar.

5. Methods

5.1 Distribution of RM systems

Updated sequences and target site information of Type II restriction enzymes and methyltransferases as of March 2020 were obtained from the REBASE database (Roberts *et al.* 2015). Sequences of all annotated proteins across complete bacterial genomes as of April 2019 were obtained from the NCBI RefSeq database (O’Leary *et al.* 2016). Sequence homologs for R and M proteins were identified according to the method used by Seshasayee *et al.* (2012). To simplify the analysis, an RM pair recognising multiple target sequences (e.g., GANT) was considered as a separate pair from another RM pair even if they had overlapping target site sequences (e.g., GACT).

In total, we extracted the presence and absence of 333 RM systems across 12388 bacterial strains. In our analysis, we grouped these bacterial strains on the level of genera rather than species to increase the statistical power of the samples. However, phage strains are highly host-specific and only a subset of phage strains will be able to successfully prey on multiple hosts within a specific genus (Koskella and Meaden 2013). In our data set, we did observe a tendency for the RM systems to be clustered at the genus level, which suggests interaction with phage strains at the genus level (see supplementary section 6 for details). When repeating our analysis with grouping at the level of bacterial species, our data set contained fewer samples but the results mirrored the results at the genus level (see supplementary section 7).

At the level of genera, almost 4 out of 5 bacterial strains in our data set have an identical composition of RM systems to another strain. For our purposes, we considered two bacteria to be identical if they carried the same set of RM systems.

After filtering, 1021 genera in our data had less than 15 strains with the remaining 42 genera containing 1417 strains in total. These strains contained 279 out of the 333 known RM systems. Only these genera with 15 or more samples were included in the analysis.

In summary, our data set included the presence and absence of 279 RM sequences for 1417 bacterial genomes across 42 different genera, and was represented by a 279×1417 ‘presence–absence matrix’, Q_{rs} , with 1 in the k th row and j th column if RM system k was present in strain j , and 0 otherwise (see supplementary figure 6 for a visual representation). Such presence–absence matrices have been previously studied for *Helicobacter* (Fullmer *et al.* 2019) and *Salmonella* (Roer *et al.* 2016) (see supplementary section 4 for a comparison).

From these presence–absence matrices, we can determine: (i) the number of RM systems per strain, (ii) the overlap in RM systems between any two bacterial strains of the same genera, (iii) the fraction of strains with a unique set of RM systems and (iv) the network of shared RM systems.

- (i) The number of RM systems per strain can be readily computed from the presence–absence matrix by summing the RM systems present in each column.
- (ii) To measure the overlap of RM systems, we first described each bacterial strain i by a set S_i containing the IDs of its RM systems. With these list, we defined the overlap between strain i and strain j as the ratio of the number of shared RM systems, $I = |S_i \cap S_j|$, to the number of unique RM systems, $U = |S_i \cup S_j|$, across both strains. This measure ranged from 0 for no overlap between the RM systems of the pair to 1 when the two strains had identical RM systems.
- (iii) To determine the fraction of strains with unique RM systems, f^u , we compared the set of RM systems from each strain S_i to the set containing all of the RM systems in the remaining strains: $S_{-i} = \bigcup_{j \neq i} S_j$. If these sets had no intersection ($S_i \cap S_{-i} = \emptyset$), the strain contributes 1 to the average while it contributes 0 otherwise. That is, for each genus $f^u \equiv \langle S_i \cap S_{-i} == \emptyset \rangle_i$, with the average being over all strains i in that genus (the $==$ operator is assumed to return 1 if the condition is true and 0 if false). The special case where a strain contains no RM systems ($S_i = \emptyset$) also contributes with 0.
- (iv) Finally, we constructed graphs where RM systems and bacterial strains were represented as nodes and where edges between two nodes signified the RM system is found in the bacterial strain.

5.2 Null model networks

To better quantify the network characteristics of the RM distribution, we developed a null model to compare against.

For each genus, g , we had the presence and absence matrix of the RM systems (i.e. presence–absence matrices). From this data, we then generated corresponding random networks that conform to the following rules:

1. Each bacterial strain in the random network must have the same number of RM systems as the corresponding one in the real network.
2. These RM systems, for each strain, were chosen randomly from the K_g available RM systems in the presence–absence matrix (i.e. the strain could contain only RM systems that were present in the sequencing data for that genus).
3. Each randomly generated bacterial strain must be unique in its RM composition (as per our filtering requirement).

Notably, this means that while our null model networks had the same RM abundance distribution (and therefore the same average number of RM systems per bacterial strain), *which* RM systems were present in each strain was randomized. This means that not all of the K_g possible RM systems were necessarily present in the null model networks, and the degree distribution of the RM system nodes was altered.

5.3 Population dynamics of a phage–bacteria ecosystem without overlap of RM systems

We based our model on previous models for an ecosystem of N bacterial strains which were being preyed upon by a single strain of phage (see figure 7 for a schematic representation). The bacteria were all valid hosts for the phage, but each bacterial strain carried a set of unique RM systems that protects against the phage. When a phage escapes restriction in a bacterial strain, its offspring will emerge with the methylation corresponding to this ‘parent’ bacterial strain. Subsequently, these phages with this particular methylation pattern can freely infect their parent strain. Eventually, phage will escape restriction from all N bacterial strains, giving rise to N corresponding epigenetic phage variants. In a well-mixed ecosystem, the equations governing the dynamics of the N bacterial strains and the N phage variants take the form (Frank 1994; Sneppen et al. 2015; Eriksen and Krishna 2020b):

$$\dot{b}_i = \gamma_i b_i (1 - B/C) - \eta_i b_i p_i - \eta_i \omega_i b_i (P - p_i) - \alpha b_i \quad (1)$$

$$\dot{p}_i = \eta_i \beta_i b_i p_i + \eta_i \omega_i \beta_i b_i (P - p_i) - \eta_i p_i B - \delta_i p_i \quad (2)$$

These equations describe the densities of each bacterial strain, b_i , and the corresponding epigenetic phage

variant, p_i . The bacteria grow logistically at a maximal rate γ_i until the total bacteria density, $B = \sum b_i$, reaches the carrying capacity $C = 10^8$ of the ecosystem. The growth rates, γ_i , are measured in units of the maximal attainable growth rate of the bacterial strains, and thus are in units of the minimal generation time τ . This is the timescale that we measured our parameters relative to. Each of the $P = \sum_i p_i$ phages adsorbs to the bacteria at a rate $\eta_i = 10^{-8}$ in units of $1/\tau$. For bacterial strain i , only the corresponding phage variant p_i can ignore the defence granted by the RM system of that strain. The remaining phage variants successfully prey on the strain with probability ω_i (see figure 7 for an example illustration of the model).

We ran our simulations with a phage burst size $\beta = 100$, comparable with what is found for real phages (De Paepe and Taddei 2006). We parametrized the strength of an RM system with the probability ω_i that the phage will bypass the RM system in the bacterial strain i despite not having the right methylation pattern. Whenever the phage successfully infects a bacterium, β new phage particles are produced whose pattern of methylation pattern matches the RM system in the parent bacterium. The model also included a separate decay rate of the bacteria, $\alpha = 0.2 \tau^{-1}$, and the phages, $\delta_i = 0.2 \tau^{-1}$. See supplementary table 3 for the default parameter values.

Our analysis focused on a relatively rich nutritional environment where C is substantially larger than the typical population size of a single bacterial strain being preyed upon by a phage (order of magnitude: $\frac{\delta_i}{\eta_i \beta}$). In this limit, a single bacterial population is not sufficient to consume the limiting nutrients of the local ecosystem. In the closed ecosystem models, the value for C is less important, since the densities can be measured in units of C without changing the dynamics. However, in the open ecosystem models, since we removed strains with less than one member, the value is important because strains with densities below $1/C$ are considered extinct. However, we found that our results only depend weakly on the value of C (see supplementary section 11 for more details).

5.4 Population dynamics with overlapping RM systems

We extended the previous model by considering each bacterial strain to be identified by a unique combination of (one or more) RM systems, some of which may be shared among different bacterial strains. In this case, it is important to keep track of the individual methylation

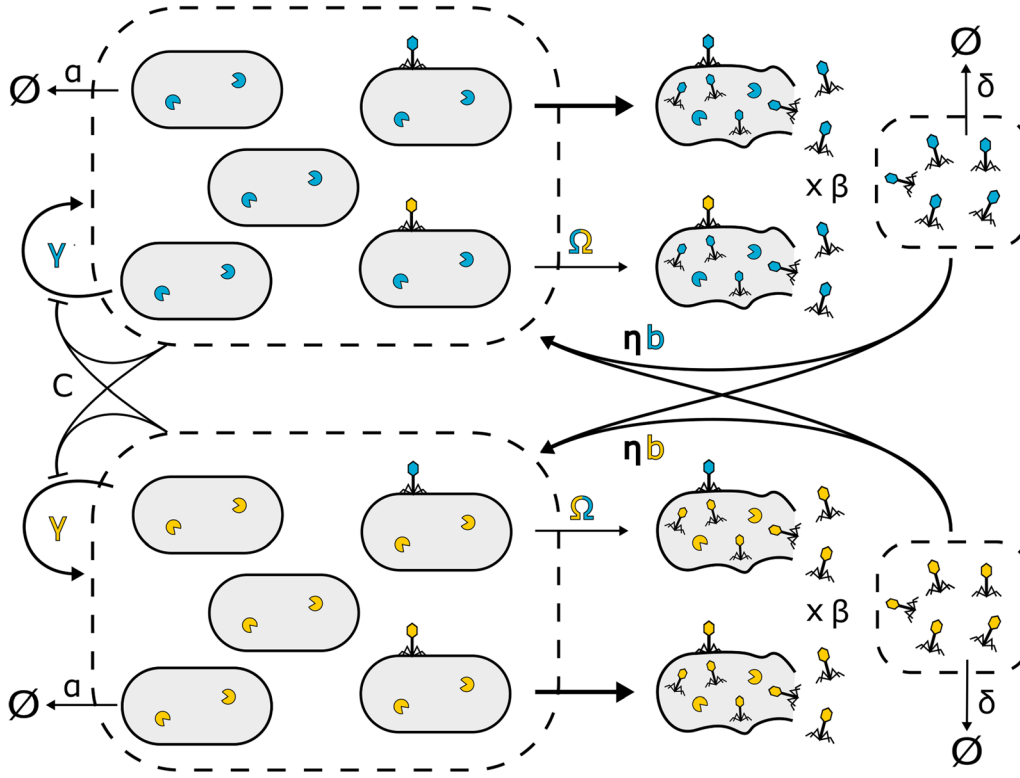


Figure 7. Model schematic. Example of our model with $N = 2$ bacterial strains. Each strain is marked by a colour which denotes its unique RM composition. The bacteria exhibit logistic growth at strain-specific growth rates γ limited by the carrying capacity C . Bacteria are removed from the system at a strain-independent rate of α , representing a general predation or death rate common to all strains in the ecosystem. Phage infection depends on the methylation pattern of the invading phage, which will ignore the RM systems from their originating host and invade new hosts with the probability Ω which depends on the RM composition of the new and originating hosts. These phages produce β offspring on successful infection, all of which have methylation patterns corresponding to the host bacterium. Once reproduced, phages are either removed at a rate of δ or they adsorb to new hosts at a rate of η per bacterium.

patterns on the epigenetic phage variants and the RM systems of the potential hosts. We achieved this by labelling each RM system with a number r , and each bacterial strain by a list of numbers, S_i , corresponding to the RM systems it contains. Formally, we defined the i th bacterial strain by the set of RM systems $S_i \equiv \{r_1^i, r_2^i, \dots, r_{N_i}^i\}$ similar to the method in Pagie (2000). Similarly, the j th phage variant, which arises by lysing a bacterium of the strain j , will have the methylation pattern of its parent bacterium, namely, $S_j \equiv \{r_1^j, r_2^j, \dots, r_{N_j}^j\}$.

In addition, where needed we allowed the phage to have innate immunities against specific RM systems (e.g. via restriction site avoidance; Vasu and Nagaraja 2013). This list of immunities, I , is the same for all epigenetic phage variants since it is a property of the phage itself, independent of its methylation pattern. Combined, this means the efficacy of the RM systems of the i th bacterial strain against the j th phage variant can be defined by the set of effective RM systems:

$$S_{i,j} \equiv \{r \in S_i \mid r \notin (I \cup S_j)\} \quad (3)$$

Essentially, this set, $S_{i,j}$, defines the RM systems in the i th bacterial strain that the j th phage is *not* immune or epigenetically protected against.

Accordingly, the probability of phage j escaping the RM systems in bacterial strain i becomes

$$\Omega_{i,j} = \prod_{r \in S_{i,j}} \omega_r \quad (4)$$

We assumed that the cost, to bacterial strains, for having RM systems, comes solely from the RM systems in the bacterium. The cost of each RM system, r , is encoded in a parameter γ_r , such that the growth rate of a bacterial strain is then

$$\Gamma_i = \prod_{r \in S_i} \gamma_r \quad (5)$$

With these definitions, our extended model reads as follows:

$$\dot{b}_i = \Gamma_i b_i (1 - B/C) - b_i \sum_j \eta_j \Omega_{i,j} p_j - \alpha b_i \quad (6)$$

$$\dot{p}_i = \beta_i b_i \sum_j \eta_j \Omega_{i,j} p_j - \eta_i p_i B - \delta_i p_i \quad (7)$$

This model is a generalization of the simple model without overlap, from the previous section.

5.5 Open ecosystem model

With a fixed number of bacterial strains, the bacterial and phage densities have dynamics with timescales of the order of $\sim 10^3$ bacterial generations (see figure 5 and Sneppen *et al.* 2015). However, over even longer timescales, one may expect that new strains of bacteria (with different, unique RM systems) arise by invasion, mutation or acquisition of RM systems from the environment or other bacteria. As a simplified model of this scenario, we considered an ecosystem consisting of M strains of bacteria and the corresponding M phage variants, described by the equations above. Periodically, at times $T, 2T, 3T, \dots$, we introduced a single bacterium from a new strain with a unique combination of RM systems and a single phage particle with patterns of methylation corresponding to these RM systems. After this addition, we allowed the dynamics to proceed for $T = 10^3$ generations to allow the dynamics time to evolve before the cycle repeated with the addition of a new strain.

Whenever a new strain was added, we also removed all bacterial strains whose densities had fallen below the level corresponding to a single individual. Conversely, the remaining strains were unchanged and the simulations continued with their current population size as the initial conditions for the next cycle. Dynamics that occur on slower timescales than T will thus occur over several simulation cycles, and the value chosen for T only weakly affected the results (see supplementary section 12 for further details).

We initialized the ecosystem with 5 strains each containing a single, randomly chosen RM system. To allow the number of RM systems per bacterium to change over time, the invading bacteria should resemble the existing bacteria but with sufficient variability to allow for the selection of strains. To achieve this, the new bacteria contain m random RM systems, where m is drawn from a Poisson distribution with a mean equal to the median number of RM systems in the current population of bacterial strains.

5.6 Choice of parameter values

To completely specify our open ecosystem model, we must describe how the parameters of the initial strains and each new strain were chosen. In Sneppen *et al.* (2015), the phage parameters η, β and δ were chosen to be the same for all phage variants, whereas ω_i and γ_i were allowed to vary. We did the same (see supplementary table 3 for a full list of parameter values), except that we now chose the ω_r and γ_r values for each RM system r instead of each host i . The values of ω_r were sampled from a log 10 uniform probability distribution between 1 and 10^{-4} (Arber 1965; Moineau *et al.* 1993; Dupuis *et al.* 2013; Pleška *et al.* 2016). The γ_r values were chosen independently from a uniform probability distribution between 0.9 and 1 (Pleška *et al.* 2016) corresponding to an average growth rate penalty of 5% per RM system before selection. Our model was constricted by computational complexity in the need to run the simulations for 10^6 additions of bacteria.

5.7 Implementation and data availability

Our analyses and models were implemented partly in MATLAB and partly in Python. The code used in our analysis is based on the code (Eriksen and Krishna 2020a) used in Eriksen and Krishna (2020b). We have made the full code and data available in the online repository located here (Eriksen *et al.* 2021): github.com/RasmusSkytte/OverlappingRestrictionModificationSystems/tree/v1.1

6. Discussion

In this study, we investigated the sharing of RM systems between bacterial strains of the same genus and found large variability across different genera. Depending on the genus of the bacterial strains, the distributions of the RM systems range from cases where the RM systems are shared rather uniformly to more skewed distributions, where a few RM systems are widespread among the bacterial strains and the remaining RM systems are rare. The sharing of RM systems connects the strains and RM systems within each genus into a *network*, whose nodes represent strains and RM systems and whose links connect RM systems to strains. Mathematically, these networks are bipartite graphs which, we find, have a non-random structure for many genera. Specifically, we observe that genera with a larger average number of RM systems

per strain have an RM composition that overlaps more with other strains than expected in a similar class of random networks. We also find that these networks typically have more strains with unique RM systems than the random expectation. The uniqueness and overlap, compared with random expectation, are anti-correlated – genera having a larger average number of RM systems tend to have a higher than expected overlap and a similar uniqueness to random networks, and vice versa.

The null model with which we have chosen to compare represents a particular ‘neutrally evolving’ scenario where each bacterial strain independently gains and loses RM systems from a common repertoire while preserving the same number of RM systems that it had in the real network. Other null models are possible, for instance: (i) where we fix the average number of RM systems per strain, but not the precise number of RM systems each strain has, or (ii) where we fix both the number of RM systems each strain has, *and* the number of strains that use each particular RM system. The former is useful for understanding whether the distribution of RM systems per strain differs from the Poisson distribution, while the latter is useful for understanding whether there are specific correlations between strains and RM systems. Our choice, which lies in between, is particularly useful for answering the question we are most interested in – given a repertoire of RM systems and given the constraints that determine how many RM systems each strain can sustain, what are the selection pressures that may lead to enhanced or diminished *sharing* of RM systems between strains? Hence, in our null model, we fixed the number of RM systems each bacterial strain has, but chose these strains randomly from the available repertoire of RM systems.

Extending previous models of ecosystems consisting of a single phage and multiple bacterial strains with RM systems (Frank 1994; Pagie 2000; Sneppen *et al.* 2015; Eriksen and Krishna 2020b) allows us to investigate how such patterns of overlap and uniqueness may shape the dynamics of the microbial ecosystem. In our simulations we found that there are opposing selection pressures – the presence of phage favours bacteria with multiple RM systems, but that potentially comes with a cost of lower growth rate, and therefore, when competition between bacteria is strong, an opposing pressure favours strains with fewer RM systems. In the model, we found an even stronger anti-correlation between the overlap of RM systems between strains and the presence of more strains with unique RM systems than the random expectation.

Effectively, two distinct strategies appear – bacterial strains either invest in multiple RM systems with an overlap significantly larger than random or are more likely than expected to invest in RM systems that are unique and not shared by other strains. In our simulations of an evolving ecosystem, the strains transitioned from the second strategy to the first at later times when the average number of RM systems is larger. The first strategy, of having larger than expected overlap, also seemed to be more prevalent in our model ecosystems when the repertoire of RM systems available to bacteria was smaller. The combination of the phage pressure and the selection pressure towards higher growth rates in bacteria-dominated states drives the transition between strategies.

Existing models of bacteria with RM systems typically lead to only transient dominance by the total phage population over the total bacterial population. When the ecosystem has reached sufficient diversity, the phage can only marginally coexist at a very low relative population (Sneppen *et al.* 2015; Eriksen and Krishna 2020b). Even in a model where RM systems are explicitly allowed to overlap (Pagie 2000), it is predicted that the number of RM systems per bacterium would increase only transiently. In the long run, the phage population collapses and the number of RM systems will subsequently collapse to about one per coexisting host strain. Notably, in our model, the sharing of RM systems seems to allow the phages to intermittently surpass the density of the bacteria, as one observes in real life (Suttle 2007; Knowles *et al.* 2016; Bar-On *et al.* 2018).

Our models in many ways represented a limited view of the complex interplay between the bacterial defence systems and phage predators. In particular, we primarily focused on cases where a single strain of phage preys on bacteria with an ensemble of RM systems. In the real world, a given host only coexists with about one phage per host in any given environment but will be exposed to different phages at different times. Our models assumed an overall well-mixed approximation to a world that has much fewer phage and bacteria coexisting together at any given time and spatial location. Furthermore, since phages also often exhibit restriction site avoidance, a bacterium will therefore need to invest in more RM systems than what our well-mixed model predicts. Finally, we only considered the influence of a single defence system and thus ignored the effects of CRISPR (Payne *et al.* 2018) and abortive infection systems (Berngruber *et al.* 2013). The inclusion of these defence systems may alleviate some of the need for diverse RM systems within each host. Thus,

rather than a precise approximation to the interactions occurring in real phage–bacteria ecosystem, our models should be thought of as a way to qualitatively understand the impact of RM system sharing on the population dynamics of phage and bacteria and to provide examples of the kinds of selection pressures which may, in turn, shape the overlap and uniqueness patterns we observe in the genome data. RM systems are known to serve several functions inside the bacteria (for a review, see Vasu and Nagaraja 2013), and these functions have been suggested as an explanation of their abundance in bacteria. Our analysis suggests that interactions with phages alone may impose a net selection pressure that favours increased investment into RM systems and leads to the emergence of networks of RM systems shared between diverse bacterial strains.

Funding

RSE and KS have received funding for this project from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 740704. NM, ASNS and SK acknowledge the support of the Department of Atomic Energy, India, under Project Identification No. RTI 4006. NM and ASNS are funded by DBT/Wellcome Trust India Alliance Intermediate Fellowship (IA/I/16/2/502711 to ASNS). SK also acknowledges the Simons Foundation (Grant No. 287975) and the Science & Engineering Research Board, Department of Science & Technology, Government of India (Matrics grant MTR/2020/000253) for funding.

References

- Arber W 1965 Host-controlled modification of bacteriophage. *Annu. Rev. Microbiol.* **19** 365–378
- Bar-On YM, Phillips R and Milo R 2018 The biomass distribution on Earth. *Proc. Natl. Acad. Sci. USA* **115** 6506–6511
- Berngruber TW, Lion S and Gandon S 2013 Evolution of suicide as a defence strategy against pathogens in a spatially structured environment. *Ecol. Lett.* **16** 446–453
- De Paep M and Taddei F 2006 Viruses’ life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages. *PLoS Biol.* **4** e193
- Dupuis MÈ, Villion M, Magadán AH and Moineau S 2013 CRISPR–Cas and restriction-modification systems are compatible and increase phage resistance. *Nat. Commun* **4** 2087
- Eriksen RS and Krishna S 2020 Data from: Defense versus growth in a hostile world - Lessons from phage and bacteria. Zenodo online repository. <https://doi.org/10.5281/zenodo.3842682>
- Eriksen RS and Krishna S 2020 Defence versus growth in a hostile world: lessons from phage and bacteria. *R. Soc. Open Sci.* **7** 201118
- Eriksen RS, Malhotra N, Seshasayee ASN, Sneppen K and Krishna S 2021 Data from: Emergence of networks of shared restriction-modification systems in phage–bacteria ecosystems. Zenodo online repository. <https://doi.org/10.5281/zenodo.6210319>
- Forterre P and Prangishvili D 2009 The Great Billion-year War between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. NY Acad. Sci.* **1178** 65–77
- Frank SA 1994 Polymorphism of bacterial restriction-modification systems: the advantage of diversity. *Evolution* **48** 1470–1477
- Fullmer MS, Ouellette M, Louyakis AS, Papke RT and Gogarten JP 2019 The patchy distribution of restriction-modification system genes and the conservation of orphan methyltransferases in Halobacteria. *Genes* **10** 233
- Knowles B, Silveira CB, Bailey BA, et al. 2016 Lytic to temperate switching of viral communities. *Nature* **531** 466–470
- Koonin EV and Wolf YI 2012 Evolution of microbes and viruses: a paradigm shift in evolutionary biology? *Front. Cell. Infect. Microbiol.* **2** 119
- Koskella B and Meaden S 2013 Understanding bacteriophage specificity in natural microbial communities. *Viruses* **5** 806–823
- Makarova KS, Wolf YI and Koonin EV 2013 Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41** 4360–4377
- Moineau S, Pandian S and Klaenhammer TR 1993 Restriction/modification systems and restriction endonucleases are more effective on lactococcal bacteriophages that have emerged recently in the dairy industry. *Appl. Environ. Microbiol.* **59** 197–202
- O’Leary NA, Wright MW, Brister JR, et al. 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44** D733–D745
- Pagie L 2000 Individual- and population-based diversity in restriction-modification systems. *Bull. Math. Biol.* **62** 759–774
- Pavlopoulos GA, Kontou PI, Pavlopoulou A, et al. 2018 Bipartite graphs in systems biology and medicine: a survey of methods and applications. *GigaScience* **7** 1–31

- Payne P, Geyrhofer L, Barton NH and Bollback JP 2018. CRISPR-based herd immunity can limit phage epidemics in bacterial populations. *eLife* **7** 1–33
- Pleška M, Qian L, Okura R, et al. 2016 Bacterial autoimmunity due to a restriction-modification system. *Curr. Biol.* **26** 404–409
- Roberts RJ, Vincze T, Posfai J and Macelis D 2015 REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res.* **43** D298–D299
- Roer L, Hendriksen RS, Leekitcharoenphon P, et al. 2016 Is the Evolution of *Salmonella enterica* subsp. *enterica* linked to restriction- modification systems? *mSystems* **1** e00009–e00016
- Seshasayee ASN, Singh P and Krishna S 2012 Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.* **40** 7066–7073
- Sneppen K, Semsey S, Seshasayee AS and Krishna S 2015 Restriction modification systems as engines of diversity. *Front. Microbiol.* **6** 528
- Stern A and Sorek R 2011 The phage-host arms race: Shaping the evolution of microbes. *BioEssays* **33** 43–51
- Suttle CA 2007 Marine viruses - Major players in the global ecosystem. *Nat. Rev. Microbiol.* **5** 801–812
- Vasu K and Nagaraja V 2013 Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77** 53–72

Corresponding editor: MOHIT KUMAR JOLLY