ABSTRACT

Title of dissertation:	Investigating the Distribution of CRISPR Adaptive Immune Systems Among Prokaryotes
	Jake L. Weissman Doctor of Philosophy, 2019
Dissertation directed by:	Professor Philip L.F. Johnson, Department of Biology Professor William F. Fagan, Department of Biology

Just as larger organisms face the constant threat of infection by pathogens, so too do bacteria and archaea. In response, prokaryotes employ a diverse set of strategies to simultaneously cope with their viral and physical environments.

Here I explore the ecology and evolution of the CRISPR adaptive immune system, a powerful form of protection against viruses that is the only known example of adaptive immunity in prokaryotes. CRISPR systems are widespread across diverse bacterial and archaeal lineages, suggesting that CRISPR effectively defends against viruses in a broad array of environments. Nevertheless, this defense system is nearly absent in many bacterial groups, and in many environments. I focus on understanding these patterns in CRISPR incidence and the ecological drivers behind them.

First, I identify the ecological conditions that favor the adoption of a CRISPR-based defense strategy. I develop a phylogenetically-conscious machine learning approach to build a predictive model of CRISPR incidence using data on over 100 phenotypic traits across over 2600 species and discovered a strong but hitherto-unknown negative interaction between CRISPR and aerobicity.

I then consider the multiplicity of CRISPR arrays on a genome, testing whether or not selection favors redundancy in immunity. I use a comparative genomics approach, looking across all prokaryotes to demonstrate that on average, organisms are under selection to maintain more than one CRISPR array. I then explain this surprising result with a theoretical model demonstrating that a trade-off between memory span and learning speed could select for paired ‘long-term memory” and “short-term memory” CRISPR arrays.

Finally, I provide a theoretical examination of the phenomenon of immune loss, specifically in the context of CRISPR immunity. In doing so, I propose an additional mechanism to answer the perennial question: “How do bacteria and bacteriophage coexist stably over long time-spans?” I show that the regular loss of immunity by the bacterial host can produce host-phage coexistence more reliably than other mechanisms, pairing a general model of immunity with an experimental and theoretical case study of CRISPR-based immunity.

Investigating the Distribution of CRISPR Adaptive Immune Systems Among Prokaryotes

by

Jake L. Weissman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Philip L.F. Johnson, Chair/Advisor
Professor William F. Fagan, Co-Advisor
Professor Stephanie A. Yarwood
Professor Pierre-Emmanuel Jabin
Professor Charles F. Delwiche

Preface

This dissertation contains an overview (Chapter 1), three research chapters in manuscript form (Chapters 2, 3, and 4), and appendices to the chapters which include all supplemental information (text, tables, and figures) for the publications on which these chapters are based. A single bibliography is provided at the end for literature cited throughout the dissertation.

This dissertation is based on the following publications:

Chapter 2: **Jake L. Weissman**, Rohan M.R. Laljani, William F. Fagan, and Philip L.F. Johnson. Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *The ISME journal*, 2019. https://doi.org/10.1038/s41396-019-0411-2

Chapter 3: **Jake L. Weissman**, William F. Fagan, and Philip L.F. Johnson. Selective maintenance of multiple CRISPR arrays across prokaryotes. *The CRISPR Journal*, 1(6):405-413, 2018. http://doi.org/10.1089/crispr.2018.0034

Chapter 4: **Jake L. Weissman**, Rayshawn Holmes, Rodolphe Barrangou, Sylvain Moineau, William F. Fagan, Bruce Levin, and Philip L.F. Johnson. Immune loss as a driver of coexistence during host-phage coevolution. *The ISME Journal*, 12(2):585-597, February 2018. https://doi.org/10.1038/ismej.2017.194

Acknowledgments

Many individuals and organizations contributed to the work before you, and to my graduate education more generally. First I would like to thank my sources of funding, whose generosity provided me with flexibility in pursuing my research goals. I have been supported by a COMBINE Network Science Fellowship (NSF award DGE-1632976) as well as a GAANN Fellowship (U.S. Department of Education). I have also been supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-14-1-0490.

Next, my advisor Philip L.F. Johnson and co-advisor William F. Fagan have been instrumental in advancing my training as a scientist. They have given me the freedom to explore and follow independent research paths, as well as the support I needed to develop successful projects. Any line of research inevitably faces setbacks, but Philip taught me to even celebrate “sideways progress”.

I thank my committee, Stephanie A. Yarwood, Pierre-Emmanuel Jabin, and Charles F. Delwiche, for their guidance and careful critique throughout my PhD. Special thanks to Stephanie for her course on microbial ecology that helped initiate my transition into this field. I am also grateful to Michelle Girvan and Daniel Serrano for their mentorship as part of the COMBINE program.

A number of collaborators have been involved in this work. Specifically, Rayshawn Holmes, under the supervision of Bruce Levin, generated the experimental data described in Chapter 4. Rodolphe Barrangou and Sylvain Moineau also contributed sequencing data for that chapter, as well as guidance while writing the final manuscript. Rohan M.R. Laljani contributed to the analysis of restriction modification systems in Chapter 2 as an undergraduate working under my supervision.

My lab-mates study everything that I don't, from T-cells to turtles. I thank Silvia Alvarez, Nina Attias, Nicole Barbour, Noelle Beckman, Sharon Bewick, Eleanor Brush, Fabian Casas-Arenas, Jeff Demers, Xianghui Dong, Andy Foss-Grant, Eliezer Gurarie, Allison Howard, Kumar Mainali, Julie M. Mallon, Shauna Rasband, Phillip Staniczenko, Anshuman Swain, Wei Xiao, Hao Y. Yiu, and Jenny Zambrano for constantly exposing me to new ideas and for occasionally forcing me to also think about eukaryotes. Special thanks to Hao for creating the cartoon elements that are the basis for Fig 1.1, and also for collaborating on outreach efforts [1].

The BEES community has been constantly supportive during my time at the University of Maryland. Our student group, BEESst, has steadily worked to build camaraderie among students and create new opportunities for student-faculty interaction. I thank the members of this community at large, and especially those who have taken on leadership roles and dedicated their time to this community's improvement.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BIC	Bayesian Information Criterion
bp	Base Pair
BREX	Bacteriophage Exclusion
Cas	CRISPR-Associated
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRDetect	CRISPR Detection Algorithm
crRNA	CRISPR-RNA
CV	Cross-Validation
DISARM	Defense Island System Associated With Restriction-Modification
E-value	Expect Value
FTP	File Transfer Protocol
HMM	Hidden Markov Model
HGT	Horizontal Gene Transfer
KEGG	Kyoto Encyclopedia of Genes and Genomes
MINT sPLS-DA	Multivariate Integrative sPLS-DA
ML	Maximum Likelihood
MOI	Multiplicity of Infection
MSE	Mean Squared Error
NB	Negative Binomial
NCBI	National Center for Biotechnology Information
NHEJ	Non-Homolohous End Joining
OG	Orthologous Group
PAM	Protospacer Adjacent Motif
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
Pfam	Protein Family Database
PLS	Partial Least Squares Regression
ProTraits	Prokaryotic Traits Database

REBASE	Restriction Enzyme Database
RefSeq	NCBI Reference Sequence Database
RM	Restriction Modification
RF	Random Forest
rRNA	Ribosomal RNA
sPLS-DA	Sparse Partial Least Squares Discriminant Analysis
t-SNE	t-Distributed Stochastic Neighbor Embedding

Chapter 1: Introduction

1.1 Prokaryotic Antiviral Defense Systems

Viruses of bacteria and archaea severely impact their hosts' population and evolutionary dynamics [2, 3]. In an ecological context, these viruses lead to the release of important nutrients back into the environment [4] and may play a role in maintaining microbial diversity [5, 6, 7]. In an evolutionary context, viruses drive the evolution of host immune strategy, often leading to iterative co-evolutionary dynamics [8, 9]. In the microbial world these two contexts are not distinct, with demographic and genetic changes occurring at similar rates, making any separation of scales infeasible. This is especially true at the interface of virus-host interactions, where the set of host defense systems is diverse and fast-evolving [10].

Prokaryotic antiviral defense systems range in complexity from crude membrane modifications that prevent viral attachment, all the way to CRISPR (clustered regularly interspaced short palindromic repeats) adaptive immune systems, which are able to record "memories" of past infections in order to specifically target those viruses in the future [10, 11]. In between these extremes lies a great deal of strategic diversity, including restriction-modification and prokaryotic Argonaute systems which degrade viral DNA, altruistic abortive-infection systems which kill the host cell upon infection, and an ever-growing set of recently-discovered, mechanistically diverse novel defense-systems [12, 13]. This strategic diversity is combinatorial, where many organisms employ multiple strategies [14], or even multiple seemingly-redundant versions of the same strategy [15]. Additionally, microbes can rapidly swap-out strategy sets when defense-related genes are lost and gained via horizontal gene transfer. These transfers occur frequently, making defense genes one of the most labile classes of genes [16], and often leading to a great deal of immune diversity even among closely related sets of strains [6, 17].

The apparent complexity of microbial antiviral defenses immediately leads us to a number of pressing questions: In the face of such incredible functional diversity, is there any discernible link between an organism's ecology and its immune strategy? Why would an organism employ more than one defense system, especially in cases where those systems appear to be redundant? Considering how frequently defense genes are gained and lost, what are the implications of these evolutionary dynamics on the ecological dynamics of host and virus communities? These questions form the core of this dissertation, motivating, respectively, Chapters 2, 3, and 4. As my focus I take the ecology and evolution of CRISPR immunity, described in detail below, expanding to antiviral defense systems in general where appropriate.

1.2 CRISPR, What is it?

CRISPR is a powerful form of protection against viruses and other mobile genetic elements that is the only known example of adaptive immunity in prokaryotes.

CRISPR systems can rapidly acquire novel and highly specific immune "memory" and then use this memory to degrade viral genetic material [18, 19]. In general, these systems are composed of two parts:

- *The CRISPR array* serves as a repository for the immune memories acquired from viruses. These genomic loci consist of variable numbers of short ($\sim$ 30bp) conserved repeat sequences ("repeats") interspaced with short variable regions ($\sim$ 30bp) called "spacers" [18, 20], preceded by a "leader" sequence which is important for array transcription and the integration of novel spacers [19, 20, 21, 22, 23, 24]. Each spacer is an individual immune memory, corresponding to a matching target on a viral genome or some other mobile genetic element (the "protospacer"; although self-targeting is also possible, e.g., [25, 26, 27]). In fact, the first hints that CRISPR might be an adaptive immune system were that many of these spacers matched known viral sequences [18]. Importantly, CRISPR immune memory is encoded on the host genome, meaning it will be vertically transferred along a host lineage [28, 29].

- *The CRISPR-associated (*cas*) genes* serve as the machinery used to acquire spacers and subsequently target viruses and are typically located adjacent to the CRISPR array on the genome [20, 30].

Immunity proceeds in three stages:

1. During acquisition Cas proteins cut out a piece of viral DNA and integrate this short sequence into the host genome at the leading end of the CRISPR array as a novel spacer [19, 23]. Spacers are inserted progressively at the

leader end, creating a linear history of infection along the array as the host encounters novel viral species [24, 31]. All CRISPR systems share the same core acquisition genes, *cas*1 and *cas*2, though the acquisition process may differ in some details between systems (with some systems using additional acquisition proteins [30], and some even aacquiring spacers from RNA [32]).

2. Arrays are then transcribed and processed into short CRISPR-RNA (crRNA) molecules, which associate with the Cas targeting machinery and surveil the cell for their corresponding protospacer [33, 34, 35, 36, 37].

3. Finally, if the crRNA-Cas complex finds its matching protospacer, the target is degraded [19, 34, 35].

For a caricature of this process see Fig 1.1. More details on the subtleties of CRISPR biology will be provided in each chapter as needed, but this general picture should suffice for the time being. For the CRISPR initiate or layperson, I recommend our review of CRISPR immunity for Frontiers Young Minds, which is targeted towards young readers (ages 8-12) but should be comprehensible to a general audience [1].

Finally, I must note that the discovery of CRISPR, now over a decade ago [18, 38], has generated great interest among biologists, who have repurposed the programmable targeting specificty of CRISPR-associated (Cas) proteins to create novel genome-editing tools [39, 40]. Here I will focus on the natural distribution and evolution of these systems, happily avoiding any discussions of applications for the remainder of this dissertation (fair warning for those who opened this text looking for genome editing wisdom– there is none to be found here).

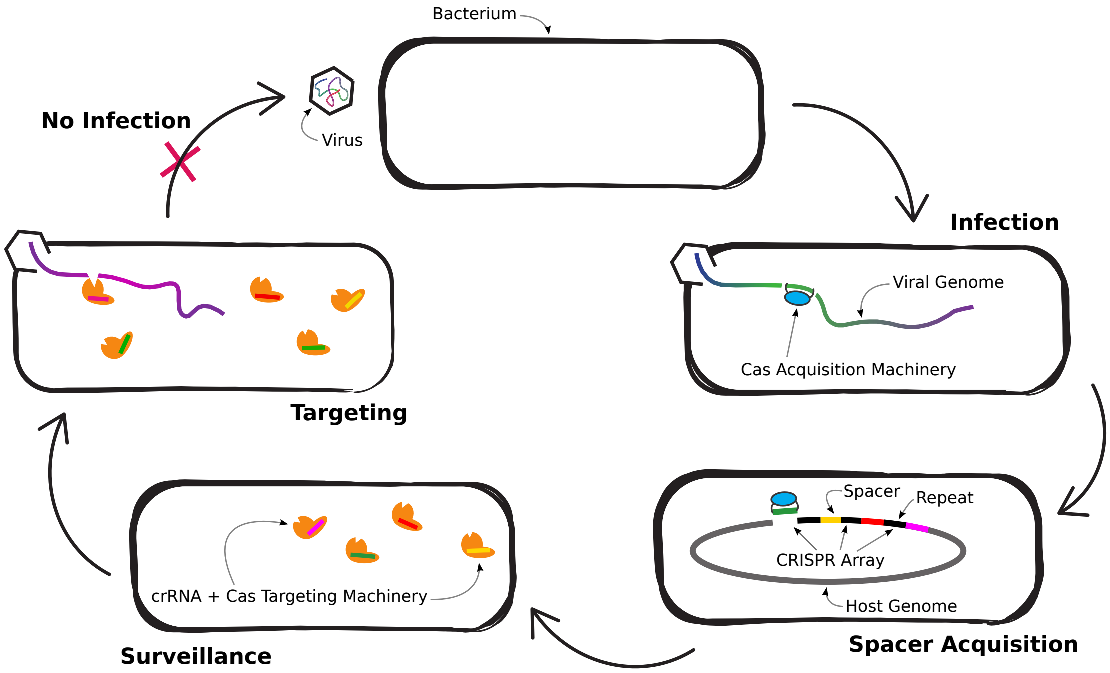


Figure 1.1: Outline of CRISPR immunity. Spacers are acquired from viral genetic material and then used to guide proteins to degrade those target sequences in the future. I note that many details are ommitted in this simple cartoon, and some CRISPR systems work somewhat differently [30]. I provide further details on the more subtle subtle aspects of CRISPR immunity as needed in each individual chapter. This figure is adapted from Weissman et al. [1].

1.3 The Distribution of CRISPR Among Prokaryotes

For the biologist interested in studying the ecology and evolution of microbial immune strategy using a comparative framework, CRISPR exists in something of a sweet-spot. Some defense systems are extremely common among microbes, such as restriction-modification systems which are nearly ubiquitous [41], while others are extremely rare, such as the BREX and DISARM systems which are present in $< 10\%$ of sequenced prokaryotic genomes [42, 43]. CRISPR, on the other hand is present in about half of sequenced bacterial genomes ($\sim 40\%$, though much more common in archaea; [20, 44, 45, 46, 47, 48, 49]), and because it is frequently horizontally transferred and lost [16, 50, 51], its distribution among species likely cannot be explained by shared evolutionary history alone. In fact, CRISPR is found across the prokaryotic tree (Fig A.1), in both bacteria and archaea and even in members of the Cadidate Phyla Radiation who were thought to largely disfavor this immune system [52, 53, 54]. Additionally, organisms differ greatly in both the number of CRISPR systems they encode on their genomes [15, 55] and the number of spacers included in any given CRISPR array [56, 57, 58], implying that the relative importance of CRISPR as a primary line of defense against viruses varies greatly between organisms. Thus we might glean some insight into what factors drive CRISPR's distribution by comparing the characteristics of taxa that tend to favor or disfavor CRISPR immunity.

More broadly, CRISPR provides a tractable model for the evolution of memory where memories are discrete, observable objects (spacers). The heterogeneous incidence of CRISPR across species suggests that memory is not always adaptive. In fact, the core questions of my dissertation can be re-framed in the context of the evolution of memory: Under what environmental conditions can we expect memory to evolve (Chapter 2)? What processes result in the evolution of short or long-term memory, together or in isolation (Chapter 3)? What dynamics occur in an antagonistic system (i.e., host-virus) when memory is lost (Chapter 4)? Incidentally, I was first drawn to this system while thinking about spatial memory in the context of animal movement, realizing that the immediately observable states of CRISPR memory are easily represented with theoretical treatments, in contrast to the many abstractions required to model animal memory.

What, then, is known about the distribution of CRISPR systems among prokaryotes? CRISPR incidence varies consistently along certain environmental gradients. For example, surveys of public genomic databases show that CRISPR systems are far more prevalent in thermophiles than in mesophiles [44, 45, 46, 47, 48, 49]. Very recently, a survey of CRISPR immune diversity in the oceans revealed that the total number of immune "memories" associated with CRISPR increases along a depth gradient [59], which agrees with my own observation that the incidence of CRISPR immunity increases with decreasing metabolic oxygen requirement (see Chapter 2; [49]).

What mechanisms can explain such environmental trends? Theoretical work has supported the hypothesis that if the local viral community is very diverse, then a memory-based system like CRISPR will not be advantageous [47, 48], because an individual cell is unlikely to encounter the same virus twice. Similarly, the presence of viral anti-CRISPR proteins will strongly adversely affect the adaptive advantage of having this system [60].

It is also possible that having an active CRISPR system is simply too costly in some environments. CRISPR incurrs costs via self-targeting (i.e., autoimmunity; [25, 27, 61]), expression [62], and lost opportunities for beneficial horizontal gene transfer (e.g., of antibiotic resistance genes [51]). The frequency of viral infection can influence the favorability of CRISPR by altering cost structures, as expression of these systems is often very costly but inducible, unlike membrane modifications which are generally constitutive but possibly less costly [63]. Costs will also vary based on the competitive environment, with recent work showing that CRISPR is favored in competitive environments that constrain the evolution of cell surface molecules (making intracellular immunity the only option for antiviral defense; [64]).

Alternatively, I and others have suggested that the abiotic environment may impact selection for or against CRISPR immunity more directly (via negative interference with certain DNA repair pathways [49] or via constraints on membrane evolution [65, 66, 67]). Thus we are left with a complex set of potential drivers of immune strategy, where the abiotic environment, local viral community, and host community may all play a role. Determining the ecological drivers of microbial immune strategy, it seems, is no easy task.

1.4 Outline of Dissertation

Just as larger organisms face the constant threat of infection by pathogens, so too do bacteria and archaea. This work focuses on these interactions, and how prokaryotes employ a diverse set of strategies to simultaneously cope with their viral and physical environments.

More specifically, my dissertation explores the ecology and evolution of the CRISPR adaptive immune system, a powerful form of protection against viruses that is the only known example of adaptive immunity in prokaryotes [18, 19]. CRISPR rapidly incorporates novel and highly specific immune "memory" and then uses this memory to degrade selfish genetic elements such as bacteriophage and plasmids [19]. CRISPR systems are widespread across diverse bacterial and archaeal lineages, suggesting that CRISPR effectively defends against viruses in a broad array of environments [30, 45, 68]. Nevertheless, this defense system is nearly absent in many bacterial groups [52], and in many environments [44, 45, 46, 47, 48, 49, 59]. In this dissertation I focus on understanding these patterns in CRISPR incidence and the ecological drivers behind them.

In my second chapter I identify the ecological conditions that favor the adoption of a CRISPR-based defense strategy [49]. I develop a phylogenetically conscious machine learning approach to build a predictive model of CRISPR presence/absence across over 2600 species using a large microbial trait database. I find evidence for a strong negative interaction between CRISPR and DNA repair processes in the cell. It seems that tradeoffs may constrain the evolution of memory in microbes, which contrasts with other work that implicates the pathogenic environment and local competition as determinants of the adaptiveness of CRISPR immune memory.

My third chapter focuses on the multiplicity of CRISPR arrays on a genome, testing whether or not selection favors redundancy in immunity [15]. Around 20% of sequenced prokaryotic genomes have more than one CRISPR array. While immune diversity likely reduces the chance of pathogen evolutionary escape, it remains puzzling why many prokaryotes also have multiple, seemingly redundant, copies of the same type of immune system. Adapting population genetic models to build a neutral model of gene content evolution, I demonstrate that on average, prokaryotes are under selection to maintain more than one CRISPR array. I explain this surprising result with a theoretical model demonstrating that trade-offs between memory span and learning speed can favor paired "long-" and "short-term memory" arrays.

In my fourth chapter I provide a theoretical examination of the phenomenon of immune loss, where it has been shown that CRISPR systems can lose functionality at a high rate [69]. In doing so, I propose an additional mechanism to answer the perennial question: "How do bacteria and bacteriophage coexist stably over long time-spans?". In well-mixed host-phage systems we typically expect to see a runaway evolutionary arms race, ultimately leading to the extinction of one species. Nevertheless, in many systems, host and pathogen coexist with minimal coevolution. I show that the regular loss of immunity by the bacterial host could produce host-phage coexistence more reliably than other tradeoff-based mechanisms, pairing a general model of immunity with an experimental and theoretical case study of CRISPR-based immunity.

Chapter 2: Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy

2.1 Abstract

Bacteria and archaea are locked in a near-constant battle with their viral pathogens. Despite previous mechanistic characterization of numerous prokaryotic defense strategies, the underlying ecological drivers of different strategies remain largely unknown and predicting which species will take which strategies remains a challenge. Here, we focus on the CRISPR immune strategy and develop a phylogenetically-corrected machine learning approach to build a predictive model of CRISPR incidence using data on over 100 traits across over 2600 species. We discover a strong but hitherto-unknown negative interaction between CRISPR and aerobicity, which we hypothesize may result from interference between CRISPR associated proteins and non-homologous end-joining DNA repair due to oxidative stress. Our predictive model also quantitatively confirms previous observations of an association between CRISPR and temperature. Finally, we contrast the environmental associations of different CRISPR system types (I, II, III) and restriction

modification systems, all of which act as intracellular immune systems.

2.2 Introduction

In the world of prokaryotes, infection by viruses poses a constant threat to continued existence (e.g., [70]). In order to evade viral predation, bacteria and archaea employ a range of defense mechanisms that interfere with one or more stages of the viral life-cycle. Modifications to the host's cell surface can prevent viral entry in the first place. Alternatively, if a virus is able to enter the host cell, then intracellular immune systems, such as the clustered regularly inter-spaced short palindromic repeat (CRISPR) adaptive immune system or restriction-modification (RM) innate immune systems, may degrade viral genetic material and thus prevent replication [11, 17, 18, 19, 38, 71]. Despite our increasingly in-depth understanding of the mechanisms behind each of these defenses, we lack a comprehensive understanding of the factors that cause selection to favor one defense strategy over another.

Here we focus on the CRISPR adaptive immune system, which is a particularly interesting case study due to its uneven distribution across prokaryotic taxa and environments. Previous analyses have shown that bacterial thermophiles and archaea (both mesophilic and thermophilic) frequently have CRISPR systems ($\sim$ 90%), whereas less than half of mesophilic bacteria have CRISPR ($\sim$ 40%; [44, 45, 46, 47, 48]). Environmental samples have revealed that many uncultured bacterial lineages have few or no representatives with CRISPR systems, and that the apparent lack of CRISPR in these lineages may be linked to an obligately symbiotic lifestyle and/or a highly reduced genome [52]. Nevertheless, no systematic exploration of the ecological conditions that favor the evolution and maintenance of CRISPR immunity has been made. Additionally, though these previous results appear broadly true [72], no explicit accounting has been made for the potentially confounding effects of phylogeny in linking CRISPR incidence to particular traits.

What mechanisms might shape the distribution of CRISPR systems across microbes? Some researchers have emphasized the role of the local viral community, suggesting that when viral diversity and abundance is high CRISPR will fail, and thus be selected against [47, 48, 63]. Others have focused on the tradeoff between constitutively expressed defenses like membrane modification and inducible defenses such as CRISPR [63]. Yet others have noted that hot, and possibly other extreme environments can constrain membrane evolution, necessitating the evolution of intracellular defenses like CRISPR or RM systems [65, 66, 67]. Many have observed that since CRISPR prevents horizontal gene transfer, it may be selected against when such transfers are beneficial (e.g. [51, 73]). More recently it has been shown that at least one CRISPR-associated (Cas) protein can suppress non-homologous end-joining (NHEJ) DNA repair, which may lead to selection against having CRISPR in some taxa [74]. In order to determine the relative importances of these different mechanisms, we must first identify the habitats and microbial lifestyles associated with CRISPR immunity.

Here we aim to expand on previous analyses of CRISPR incidence in three ways: (1) by drastically expanding the number of environmental and lifestyle traits considered as predictors using the combination of a large prokaryotic trait database and machine learning approaches, (2) by incorporating appropriate statistical corrections for non-independence among taxa due to shared evolutionary history, which has not always been done, and (3) by simultaneously looking for patterns in RM systems, which will help us untangle the difference between environments that specifically favor CRISPR adaptive immunity versus DNA-degrading intracellular immune systems in general (RM and CRISPR).

2.3 Methods

2.3.1 Data

For a schematic outlining the entire data compilation process see Fig A.2. For a list of all visualizations, predictive models, and statistical tests see Appendix A.1.

2.3.1.1 Trait Data

We downloaded the ProTraits microbial traits database [75] which describes 424 traits in 3046 microbial species. These traits include metabolic phenotypes, preferred habitats, and specific behaviors like motility, among many others. ProTraits was built using a semi-supervised text-mining approach, drawing from several online databases and the literature. All traits are binary, with categorical traits split up into dummy variables (e.g. oxygen requirement listed as “aerobic”, “anaerobic”, and “facultative”). For each trait in each species, two “confidence scores” in the range $[0, 1]$, are given, corresponding to the confidence of the text mining approach that a particular species does (c_+) or does not (c_-) have a particular trait.

We derived a single score (p) that captured the confidences both that a species does and does not have a particular trait. Assuming we want our score to lay in the interval $[0, 1]$, such a score should be zero when we are completely confident that a species does not have a trait, one when we are completely confident that a species has a trait, and 0.5 when we are completely uncertain whether or not a species has a trait (i.e., equally confident that it does and does not have the trait). In the following formula, $\frac{c_+}{c_+ + c_-}$ captures the relative confidence that a species does rather than does not have a trait, which we then scale by the overall maximal confidence (so that as overall confidence decreases the score shrinks towards 0.5)

$$p = \frac{1}{2} + \left(\frac{c_+}{c_+ + c_-} - \frac{1}{2} \right) \times \max(c_+,\, c_-). \tag{2.1}$$

Many of the scores are missing for particular species-trait combinations (18%), indicating situations in which the text mining approach was unable to make a trait prediction. Our downstream analyses do not tolerate missing data, and so we imputed missing values using a random forest approach (R package missForest; [76]). There is a set of summary traits in the ProTraits dataset that were created de-novo using a machine learning approach, as well as a number of traits describing the growth substrates a particular species can use. We removed both summary and substrate traits from the dataset for increased interpretability (post-imputation; 174 traits remaining).

We note that the authors of ProTraits also used genomic data to help them infer trait scores, though we found that the exclusion of this data does not affect our overall outcome (Appendix A.2).

2.3.1.2 Genomic Data and Immune Systems

For each species listed in the ProTraits dataset we downloaded a single genome from NCBI's RefSeq database, with a preference for completely assembled reference or representative genomes. See Appendix A.3 for a confirmation that our results are robust to the resampling of genomes. A number of species (333) had no genomes available in RefSeq, or only had genomes that had been suppressed since submission, and we discarded these species from the ProTraits dataset.

CRISPR incidence in each genome was determined using CRISPRDetect [77]. Additionally, data on the number of CRISPR arrays found among all available RefSeq genomes from a species were taken from Weissman et al. ([15]).

We downloaded the REBASE Gold database of experimentally verified RM proteins and performed blastx searches of our genomes against this database [78, 79]. The distribution of E-values we observed was bimodal, providing a natural cutoff ($E < 10^{-19}$).

To assess the ability of a microbe to perform non-homologous end-joining (NHEJ) DNA repair we used hmmsearch to search the HMM profile of the Ku protein implicated in NHEJ against all RefSeq genomes (E-value cutoff of 10^{-2}/number of genomes; Pfam PF02735; [80, 81, 82]). We also used the annotated number of 16s rRNA genes in each downloaded RefSeq genome as a proxy for growth rate and the annotated *cas3*, *cas9*, and *cas10* genes as indicators of system type [83]. Where available as meta-data from NCBI, we also downloaded the oxygen (1949 records) and temperature requirements (1094 records) for the biosample record associated with each RefSeq genome. The NCBI trait data was used exclusively for building Fig 2.4 and the analyses implicating Ku in the CRISPR versus oxygen association.

2.3.2 Phylogeny

We used PhyloSift to locate and align a large set of marker genes (738) found broadly across microbes, generally as a single copy [84, 85]. Of these marker genes, 67 were found in at least 500 of our genomes, and we limited our analysis to just this set. Additionally, eight genomes had few (< 20) representatives of any marker genes and were excluded from further analysis. We concatenated the alignments for these 67 marker genes and used FastTree (general-time reversible and CAT options; [86]) to build a phylogeny (Fig A.3). In order to analyze the effect of tree uncertainty on our phylogenetic regressions, we bootstrapped our dataset using seqboot and built a new tree from each replicate.

2.3.3 Visualizing CRISPR/RM Incidence

The size of the ProTraits dataset, both in terms of number of species and number of traits, and the probable complicated interactions between variables necessitate techniques that can handle complex, large scale data. To visualize the structure of microbial trait space and the distribution of immune strategies within that space we made use of two unsupervised machine learning techniques, principal component analysis (PCA, prcomp() function in R) and t-distributed stochastic neighbor embedding (t-SNE, perplexity = 50 and 5000 iterations using Rtsne() function in Rtsne R package, otherwise default parameters, perplexity varied in Fig A.4; [87, 88]).

PCA is a well-used technique in ecology that allows us to reduce the dimensionality of a dataset for effective visualization in two-dimensional space. Essentially, we collapse our trait dataset into two or three composite traits and observe whether species with a particular immune strategy tend to vary systematically in terms of where they fall in this “trait space”. A newer variant of this approach, t-SNE, performs a similar process, but unlike PCA allows for non-linear transformations of trait space. Therefore, local structure and non-linear interactions between traits in high dimensional space are preserved by t-SNE but often not captured by PCA [87]. On the other hand, t-SNE axes are less easily interpreted precisely because they represent non-linear rather than linear combinations of variables.

2.3.4 CRISPR/RM Prediction from ProTraits

In order to predict the distribution of CRISPR and RM systems, we applied a number of supervised machine learning approaches to our dataset (see Fig A.5 for a flow-chart describing the logic behind our model choices). In order to obtain accurate estimates of model performance, we initially set aside a portion of the data as a test set to be used exclusively in model assessment after all models were constructed (no fitting to this set). Because of the underlying evolutionary relationships in the data, we chose a test set that is phylogenetically independent of our training set. Alternatively, if we were to draw a test set at random from the microbial species we would risk underestimating our prediction errors due to non-independence of the training and test sets [89]. We chose the Proteobacteria as a test set because they are well-represented in the dataset (1139 species), ecologically diverse, and highly heterogeneous in terms of CRISPR incidence (Fig A.1). The remaining phyla were used to train our models.

First we built a series of linear models to classify species by immune strategy (CRISPR present or absent) using logistic regression. We had a large number of predictor variables (100+), which necessitated a model-selection approach in order to build a reasonably (and optimally) sized model. We used a forward selection algorithm to select the optimal set of predictors for each model size, with mean squared error under cross validation (CV) as our optimality criterion. We then selected model size by comparing BIC among these optimal models (i.e., selecting the model with the lowest score).

Similar to choosing a test set, care must be taken when performing CV on phylogenetically-structured data. CV assumes that when the data is partitioned into folds, each of these folds is independent of the others. If we draw species at random from a phylogeny, this assumption is violated, since the same hierarchical tree-structure will underlay each fold. Therefore, it is better to perform "blocked" CV than random CV [89], wherein folds are chosen based on divergent groups on the tree (e.g. phyla). If each group has diverged far enough in the past from the others, we can consider these folds to be essentially evolutionarily independent in terms of trait evolution (see Fig A.6 for a conceptual example). Therefore blocked CV is essentially a non-parametric method (i.e., no explicit evolutionary model) to account for the non-independence arising from the shared evolutionary history between species. We use both random and blocked CV to build models. We clustered the data into blocked folds using the pairwise distances between tips on our tree (partitioning around mediods, pam() function in R package cluster, five folds so that $k = 5$; [90, 91]). A key assumption we make here is that our folds can be taken as independent from one another (i.e. no effect of shared evolutionary history). Since these clusters correspond roughly to Phylum-level splits, and since CRISPR and other prokaryotic immune systems are rapidly gained and lost over evolutionary time [16], we are comfortable making this assumption. We also repeated this analysis using phylogenetic logistic regression to more formally correct for phylogeny (R package phylolm; [92, 93]). Phylogenetic logistic regression is a more powerful method since it fits an explicit model of trait evolution, although it relies on the assumption that traits evolve according to the chosen model and can give misleading results otherwise.

Stepwise methods for variable selection, such as those used above (i.e., forward subset selection), are simple, computationally feasible, and easy to implement and interpret, but perform poorly when variables in the dataset covary with one another (i.e. multicollinearity; [94, 95]). As it so happens, the trait data used here exhibit strong multicollinearity (R package mctest; [96, 97]). Therefore, we sought out methods that deal well with this type of data, specifically partial least squares regression (PLS; [94]). Briefly, PLS combines features of PCA and linear regression to find the linear combination of predictors that maximizes the variance of the data in the space of outcome variables. We use a variant of PLS, sparse partial least squares discriminant analysis (sPLS-DA), where the “sparse” refers to a built-in variable selection process in the model-fitting algorithm and “discriminant analysis” refers to the fact that we are focused on a classification problem (i.e., presence vs. absence of a particular immune strategy; we used tune.splsda() perform 5-fold cross validation, repeated 50 times, to select the optimal number of components n to include and splsda() to perform variable selection and model selection simultaneously given n as an input; functions in R package mixOmics; [98, 99]).

We also attempt to ameliorate the effects of shared evolutionary history on our PLS model by using a philosophically similar approach to our blocked CV method above. Multivariate integrative (MINT) sPLS-DA is a variant of PLS that can account for systematic variation between groups of data when those groupings are known (e.g., our phylogenetically-blocked folds from above). It was originally developed for use in situations where multiple experiments testing the same hypothesis could show systematic biases from one another. In our case, the history of prokaryotic evolution is our experiment, and deep branching lineages are our replicates. We apply MINT sPLS-DA to the data, using the same blocked folds we used for CV (we used tune.mint.splsda() to perform 5-fold blocked cross validation to select the optimal number of components n to include and mint.splsda() to perform variable selection and model selection simultaneously given n as an input; functions in R package mixOmics; [99, 100]).

While regression provides easily interpretable trait weights and is computational efficient, in order to capture higher-order relationships between microbial traits we needed more powerful methods. Random forests (RF) are an attractive choice for our aims since they produce a readily-interpretable output and can incorporate nonlinear relationships between predictor variables [101]. We built an RF classifier on our training data from 5000 trees (otherwise default settings in R package randomForest so that the number of variables tried at each split is the square root of the total number of predictors; [102]). To prevent fitting to phylogeny, we took an ensemble approach which was similar in philosophy to our blocked CV and MINT sPLS-DA approaches above. Using the phylogenetically blocked folds defined above we fit five individual forests, each leaving out one of the five folds. We then weighted these forests by their relative predictive ability on the respective fold excluded during the fitting process (measured as Cohen's κ; [103]). We predicted using our ensemble of forests by choosing the predicted outcome with the greatest total weight.

2.4 Results

Below, we associate specific microbial immune strategies with a diverse list of microbial traits. The traits span a range of scales including aspects of habitat (e.g. "aquatic"), morphology (e.g., "coccus"), and physiology (e.g., "heterotroph") [75]. While this variety of scales poses a modeling challenge to traditional approaches including linear regression, machine learning algorithms provide an elegant means of integrating such multi-scale traits in a statistically rigorous predictive framework. In particular, we apply algorithms that excel at identifying both linear and non-linear combinations of traits with high predictive ability. For a systematic comparison of the output of our predictive models, discussed individually below, please see Figs A.7 and A.8.

2.4.1 Visualizing CRISPR Incidence in Trait Space

We visualized CRISPR incidence in microbial trait space using two unsupervised algorithms to collapse high-dimensional data (174 binary traits assessed in 2679 species; see Methods) into fewer dimensions. Both methods revealed clear differences between the placement of CRISPR-encoding and CRISPR-lacking organisms in trait space, despite the fact that no explicit information about CRISPR was included.

First, principal components analysis (PCA) of the trait data reveals several previously recognized patterns of microbial lifestyle choice and CRISPR incidence. The first principal component (17% variance explained) corresponds broadly to an axis running from host-associated to free-living microbes (Table 2.1), as observed by others [104, 105]. CRISPR-encoding and CRISPR-lacking microbes are not differentiated along this axis (Fig A.9). We see CRISPR-encoding and CRISPR-lacking organisms beginning to separate along the second (10% variance explained) and third (7% variance explained) principal components (Fig 2.1). The second component roughly represents a split between extremophilic species typically living in low-productivity environments and mesophilic, plant-associated species (Table 2.1).

Optimal growth temperature appears to be an important predictor of CRISPR incidence, as previously noted by others [47, 48]. The third component is not as easy to interpret, but appears to indicate a spectrum from group living microbes (e.g. biofilms) to microbes that tend to live as lone, motile cells (Table 2.1). That CRISPR is possibly favored in group-living microbes is not entirely surprising, considering the increased risk of viral outbreak at high population density, and that some species up-regulate CRISPR during biofilm formation [106].

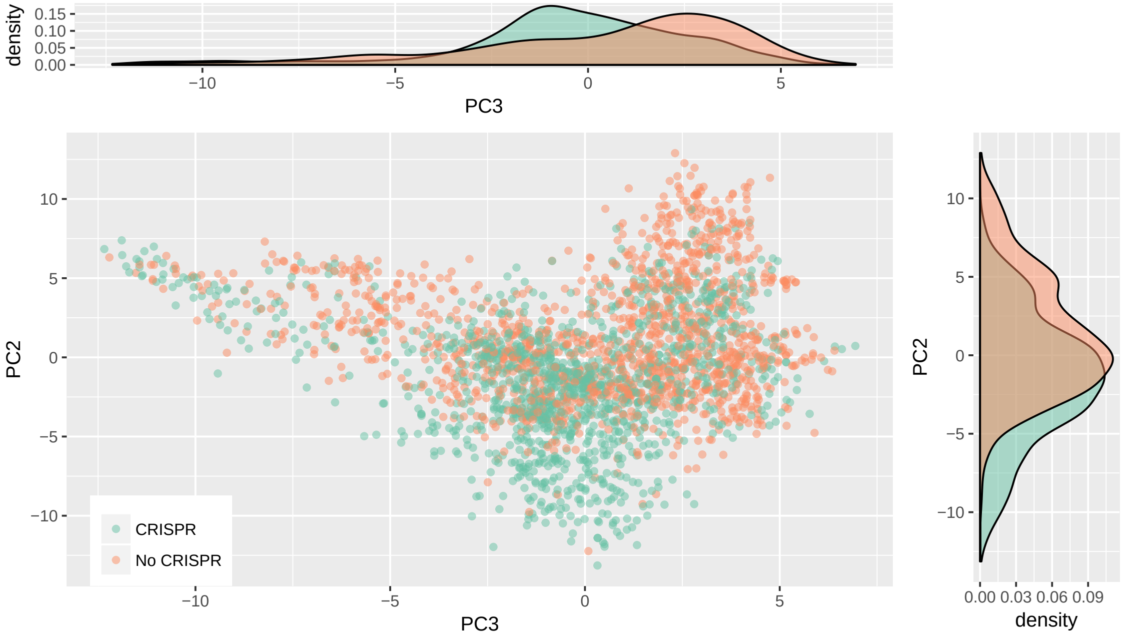


Figure 2.1: Organisms with CRISPR separate from those without in trait space. The second and third components from a PCA of the microbial traits dataset are shown, where each point is a single species. CRISPR incidence is indicated by color (green with, orange without), but was not included when constructing the PCA. Notice the separation of organisms with and without CRISPR along both components. Marginal densities along each component are shown to facilitate interpretation. See Fig A.9 for the first component.

PC1	Weight	PC2	Weight	PC3	Weight
ecosystemcategory_human	-0.16	temperaturerange_mesophilic	0.19	growth_in_groups	-0.24
specificecosystem_sediment	0.16	temperaturerange_thermophilic	-0.19	gram_stain_positive	-0.24
ecosystem_environmental	0.16	oxygenreq_strictanaero	-0.19	cellarrangement_singles	0.21
knownhabitats_host	-0.15	temperaturerange_hyperthermophilic	-0.18	cellarrangement_filaments	-0.20
ecosystemsubtype_intertidalzone	0.15	knownhabitats_hotspring	-0.17	sporulation	-0.20
ecosystem_hostassociated	-0.15	exosystemtype_rhizoplane	0.17	energysource_chemoorganotroph	-0.19
habitat_hostassociated	-0.15	habitat_specialized	-0.16	cellarrangement_clusters	-0.18
habitat_freeliving	0.15	metabolism_methanogen	-0.16	shape_tailed	-0.18
ecosystemtype_digestivesystem	-0.14	ecosystemcategory_plants	0.15	habitat_terrestrial	-0.18
specificecosystem_fecal	0.14	ecosystemtype_thermalsprings	-0.15	motility	0.17

Table 2.1: Top 10 variable loadings on the first three principal components of the PCA performed on the microbial traits dataset, shown in Figs 2.1 and A.9. These three components explain 17%, 10%, and 7% of the total variance, respectively.

Second, we visualized the trait data using *t*-distributed stochastic neighbor embedding (t-SNE), which is a nonlinear method that can often detect more subtle relationships in a dataset (Fig 2.2; [87]). This method reveals a clustering of CRISPR-encoding microbes in trait space, further emphasizing that microbial immune strategy is influenced by ecological conditions. Because the axes of t-SNE plots are not easily interpretable, we mapped the top weighted traits from the PCA above (Table 2.1) onto the t-SNE reduced data (Fig A.10). Surprisingly, the most clearly aligned trait with CRISPR-incidence is having an obligately anaerobic metabolism.

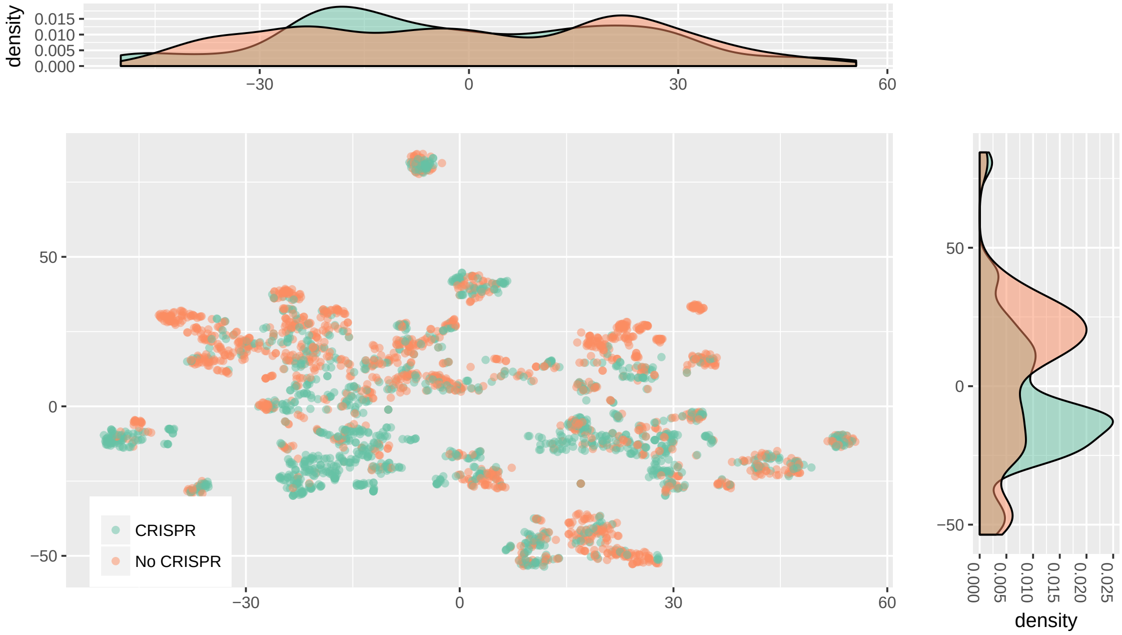


Figure 2.2: Organisms with CRISPR partially cluster in trait space away from those without. Two dimensional output of t-SNE dimension reduction of the microbial traits dataset are shown, where each point is a single species (same dataset as in Fig 2.1). CRISPR incidence is indicated by color (green with, orange without), but was not included when performing dimension reduction. The axes of t-SNE plots have no clear interpretation due to the non-linearity of the transformation.

2.4.2 Predicting CRISPR Incidence

The above unsupervised approaches (i.e. uninformed about the outcome variable, CRISPR) revealed that CRISPR incidence appears to be impacted by other microbial traits. In order to more formally characterize these patterns, and exploit them for their predictive ability, we applied several supervised prediction methods (i.e. trained with information about CRISPR incidence) methods to the complete trait dataset.

Unlike traditional statistical techniques focused on assigning p-values to particular input variables, with our machine learning approach we assessed model performance in terms of predictive ability. For unbiased error estimates, we chose an independent "test" set to withhold during the model fitting process and to be used only during model assessment. We consider effective prediction of CRISPR incidence in this independent dataset as support that our model encodes real information about how different microbial traits influence the ecological advantages of the CRISPR system. We then examined the structure of these models, and which variables play an outsize role in their performance, in order to select candidate traits associated with CRISPR incidence. Importantly, we chose the Proteobacteria as our test set because they represent a phylogenetically-independent group from our training set (see Methods).

All models we implemented showed improved predictive ability over a null model only accounting for the relative frequency of CRISPR among species (Cohen's $\kappa > 0$; Table 2.2), indicating that there is some ecological signal in CRISPR incidence, though overall predictive performance was not overwhelming. Of these models the random forest (RF) model ranked highest, and did reasonably well ($\kappa = 0.241$). The percent incidences of CRISPR in the training (56%) and test sets (36%) are considerably different, which may have been difficult for these models to overcome. It is also possible that the Proteobacteria vary systematically from other phyla in terms of ecology and immune strategy, making them a particularly difficult (and thus conservative) test set. Nevertheless, the trait data clearly held some information about CRISPR incidence. We will primarily focus here on the RF model since it performed best, but see Appendix A.4 for further discussion of the performance of our other models.

	Phylogenetic Correction			Performance		
Model Type	Non-Parametric	Parametric	Model Size	Accuracy	κ	TPR
Log. Reg.	No	No	18	66.1%	0.152	0.233
Log. Reg.	Yes	No	9	67.5%	0.168	0.209
Log. Reg.	No	Yes	10	67.7%	0.188	0.246
Log. Reg.	Yes	Yes	6	67.4%	0.160	0.294
sPLS-DA	No	No	[7, 159, 4, 169, 50] (5 comp.)	68.4%	0.190	0.219
MINT sPLS-DA	Yes	No	32 (1 comp.)	60.5%	0.173	0.538
RF	No	No	-	68.8%	0.241	0.327
RF Ensemble	Yes	No	-	68.6%	0.240	0.332

Table 2.2: Predictive ability of models of CRISPR incidence on the Proteobacteria test set. Model size refers to number of variables chosen overall, or per-component in the case of the partial least squares models. Accuracy is measured as the total number of correct predictions over the total attempted and κ is Cohen's κ, which corrects for uneven class counts that can inflate accuracy even if discriminative ability is low. Roughly, κ expresses how much better the model predicts the data than one that simply knows the frequency of different classes ($\kappa = 0$ being no better, $\kappa > 0$ indicating improved predictive ability). The true positive rate (TPR) is the number of correctly identified genomes having CRISPR divided by the total number of genomes having CRISPR in the test set. The non-parametric correction for phylogeny refers to our phylogenetically blocked folds, whereas the parametric correction refers to our use of phylogenetic logistic regression [92]. Observe that the RF model appears to perform best at prediction in general.

While each of our models revealed a distinct set of top predictors of CRISPR incidence, there was broad agreement overall (Table A.1 and Figs 2.3, A.11, and A.12). Keywords indicating a thermophilic lifestyle (e.g. thermophilic, hot springs, hyperthermophilic, thermal springs) appeared across all models as either the most important or second most important predictor of CRISPR incidence. Keywords relating to oxygen requirement (e.g. anaerobic, aerobic) also appeared across nearly all models as top predictors, excluding only the two worst performing models (Table A.1). In the case of the RF and sPLS-DA models, oxygen requirement was always one of the top three predictors, and often the top predictor of CRISPR incidence (Figs 2.3, A.11, A.12, and A.13). Other predictors that frequently appeared across model types included termite hosts (host_insectstermites), the degradation of polycyclic aromatic hydrocarbons (PAH; metabolism_pahdegrading), freshwater habitat (knownhabitats_freshwater), and growth as filaments (shape_filamentous). In general, the sPLS-DA, MINT sPLS-DA, RF, and RF ensemble models agreed with each other rather closely. Finally, we built an RF model using only traits related to temperature range, oxygen requirement, and thermophilic lifestyle (hot springs, thermal springs, hydrothermal vents). This temperature- and oxygen-only RF model outperformed all non-RF models ($\kappa = 0.191$). These traits alone appear to hold the majority of information about CRISPR incidence in the dataset.

As an additional check that these candidate traits versus CRISPR associations are real and not due to some irregularity in our dataset, we downloaded metadata available from NCBI. We were able to reproduce the result that thermophiles strongly prefer CRISPR (92% with CRISPR as opposed to 49% in mesophiles, Fig

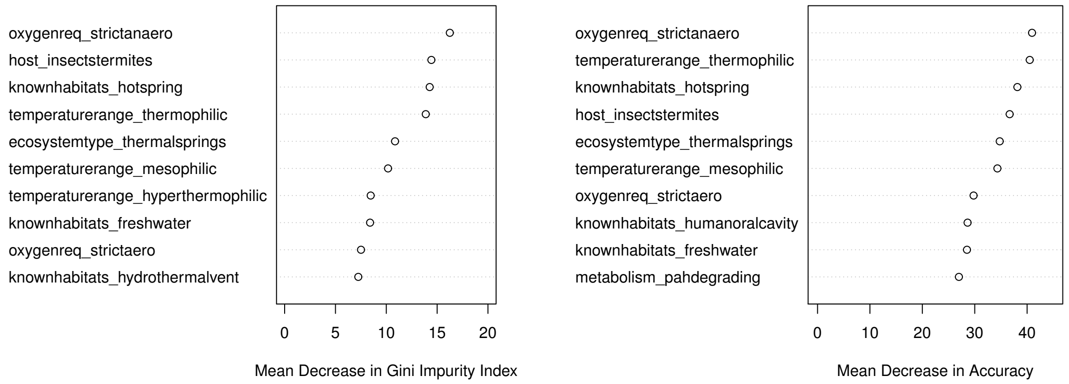


Figure 2.3: Importance of top ten predictors in the RF model of CRISPR incidence using the ProTraits predictors. The mean decrease in accuracy measures the reduction in model accuracy when a variable is randomly permuted in the dataset. The Gini impurity index is a common score used to measure the performance of decision-tree based models (e.g. RF models). Briefly, when a decision tree is built the Gini impurity index measures how well separated the different classes of outcome variable are at the terminal nodes of the tree (i.e., how “pure” each of the nodes is). The mean decrease in Gini impurity measures the estimated reduction in impurity (increase in purity) when a given variable is added to the model. These importance scores are useful to rank variables as candidates for further study, but in themselves should not be taken as statistical support or effect sizes similar to those seen in linear regression. RF models may include non-linear combinations of variables, and therefore the contribution of any one variable is not as easily interpreted as with a linear model, a drawback of this approach. See Fig A.14 for all predictor importances.

2.4a; [47, 48]). Though we have too few genomes categorized as psychrotolerant (35) or psychrophilic (14) to make any strong claims, these genomes seem to lack CRISPR most of the time, suggesting that CRISPR incidence decreases continuously as environmental temperatures decrease [46]. We were also able to confirm that, in agreement with our visualizations and predictive modeling, aerobes disfavor CRISPR immunity (34% with CRISPR) while anaerobes favor CRISPR immunity (67% with CRISPR, Fig 2.4b). This is true independent of growth temperature, with mesophiles showing a similarly strong oxygen-CRISPR link (Fig A.15). Overall, both oxygen ($\chi^2 = 254.04$, $p < 2.2 \times 10^{-16}$, categories with < 10 observations excluded) and temperature ($\chi^2 = 98.86$, $p < 2.2 \times 10^{-16}$, categories with < 10 observations excluded) had significant effects on incidence (for breakdown see Fig 2.4).

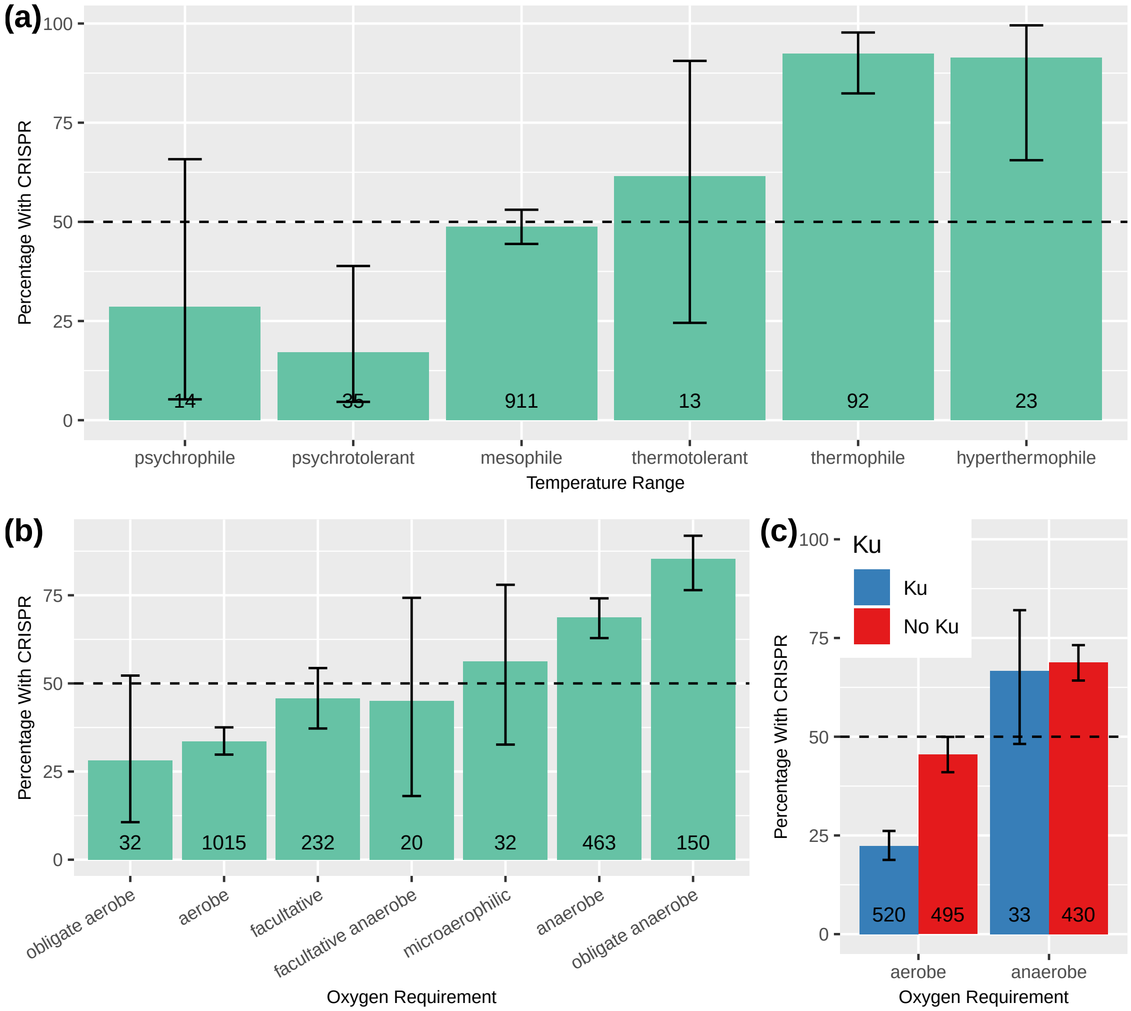


Figure 2.4: Temperature range and oxygen requirement are strong predictors of CRISPR incidence. Trait data taken from NCBI. (a) Thermophiles strongly favor CRISPR immunity, while mesophiles appear ambivalent. (b) Anaerobes favor CRISPR immunity, while aerobes tend to lack CRISPR and facultative species fall somewhere in between. (c) CRISPR and the Ku protein are negatively associated in aerobes but not anaerobes. Error bars are 99% binomial confidence intervals (non-overlapping intervals can be taken as evidence for a statistically significant difference at the $p < 0.01$ level). Total number of genomes in each trait category shown at the bottom of each bar. Categories represented by fewer than 10 genomes were omitted.

Following previous suggestions that CRISPR incidence might be negatively associated with host population density and growth rate [47, 48, 63], and that this could be driving the link between CRISPR incidence and optimal temperature range, we sought to determine if growth rate was a major determinant of CRISPR incidence. The number of 16s rRNA genes in a genome is an oft-used, if imperfect, proxy for microbial growth rates and an indicator of copiotrophic lifestyle in general [107, 108, 109]. While CRISPR-encoding genomes had slightly more 16s genes than CRISPR-lacking ones (3.1 and 2.9 on average, respectively), the 16s rRNA gene count in a genome was not a significant predictor of CRISPR incidence (logistic regression, $p = 0.05248$), although when correcting for phylogeny 16s gene count does seem to be significantly positively associated with CRISPR incidence (phylogenetic logistic regression, $m = 0.06277$, $p = 6.651 \times 10^{-5}$), the opposite of what we would expect if growth rate were driving the CRISPR-temperature relationship (though the effect was not consistent across bootstrapped trees; Table A.2).

As a secondary confirmation of the link between oxygen and CRISPR, we examined metagenomic data from the Tara Oceans Project [110], and found that across a large set of ocean metagenome samples CRISPR prevalence was inversely related to environmental oxygen concentration (Appendix A.5).

We also attempted to predict the number of CRISPR arrays in a genome given that that genome had at least one array, though this attempt was entirely unsuccessful (Appendix A.6).

2.4.3 Predicting CRISPR Type

Each CRISPR system type is associated with a signature *cas* targeting gene unique to that type (*cas3*, *cas9*, and *cas10* for type I, II, and III systems respectively). There are many species in the dataset with *cas3* (605), but relatively few with *cas9* (160) and *cas10* (222), suggesting that the traits correlated with CRISPR incidence probably correspond primarily to type I systems (the dominance of type I systems has been noted previously [30]). We mapped the incidence of each of these genes onto the PCA we constructed earlier (see Fig A.9 and Table 2.1), and found that *cas9* separates from *cas3* and *cas10* along the first component (Fig 2.5a). Broadly, this indicates that type II systems are more commonly found in host-associated than free-living microbes, the opposite of the other two system types.

We built an RF model of *cas9* incidence, with the Proteobacteria as the test set. Because our training set had so few cases of *cas9* incidence (10% of set), we performed stratified sampling during the RF construction process to ensure representative samples of organisms with and without *cas9*. Surprisingly, despite the extremely small number of organisms with *cas9* in the training and test sets (160 and 58 respectively), this model was accurately able to predict type II CRISPR incidence and had some discriminative ability (Accuracy = 93.0%, $\kappa = 0.164$), though it missed many of the positive cases (TPR = 0.172). This model also suggested that a host-associated lifestyle seems to be a major factor influencing the incidence of type II systems, with many of the top-ranking variables in terms of importance corresponding to keywords having to do with the split between host associated and

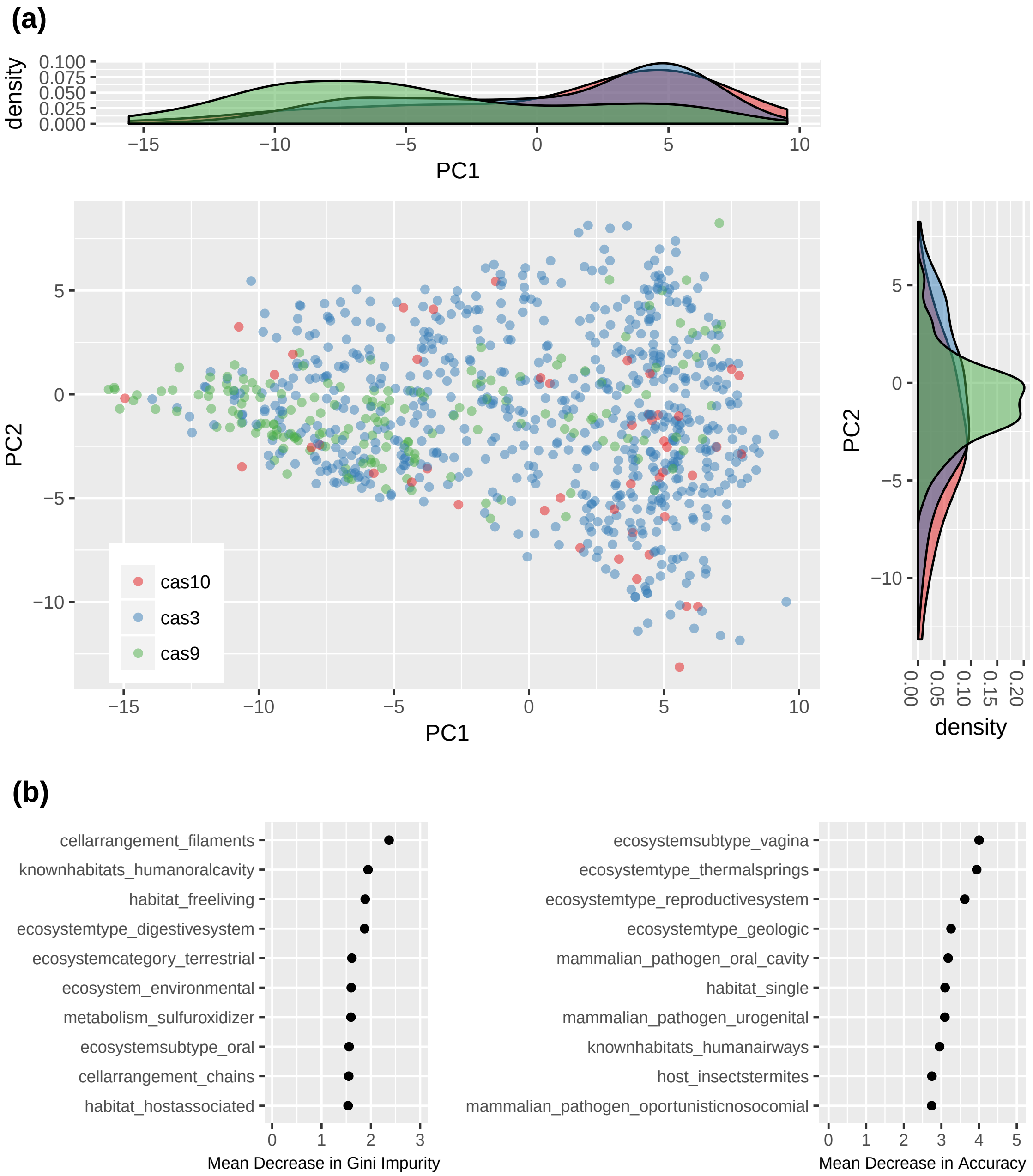


Figure 2.5: Type II CRISPR systems appear to be more prevalent in host-associated microbes. (a) The cas targeting genes associated with type I, type II, and type III systems (*cas3*, *cas9*, and *cas10* respectively) mapped onto the PCA in Fig A.9. Organisms without any targeting genes were omitted from the plot for readability. Recall from Table 2.1 that PC1 roughly corresponds to a spectrum running from host-associated to free-living microbes. (2) A variable importance plot from an RF model of *cas9* incidence. Observe that keywords related to a host-associated lifestyle appear many times.

free-living organisms (Fig 2.5b).

2.4.4 NHEJ, CRISPR, and Oxygen

Recently, Bernheim et al. [74] demonstrated that the type II-A CRISPR system interferes with the NHEJ DNA repair pathway, leading to an inverse relationship between the presence of type II-A systems and the NHEJ pathway in microbial genomes. We hypothesized that this negative relationship between CRISPR and NHEJ might be more widespread across system types. We also hypothesized that this could explain the negative relationship between CRISPR and aerobicity we observe, since reactive oxygen species produced during aerobic respiration can induce double-strand breaks, thus selecting for the presence of NHEJ repair in aerobic organisms [111, 112]. We use the presence of Ku protein as a proxy for the NHEJ pathway, since this protein is central to the pathway.

There was a clear interaction between the presence of Ku and aerobicity on the incidence of CRISPR (Fig 2.4c, using aerobicity meta-data from NCBI for this and below analyses). Using our full set of RefSeq genomes, we found a weak negative association between CRISPR and Ku incidence overall (Pearson's correlation, $\rho = -0.012$; $\chi^2 = 15.015$, $p = 1.067 \times 10^{-4}$), but restricting only to aerobes the negative association between Ku and CRISPR was much stronger (Pearson's correlation, $\rho = -0.250$, $p = 9.109 \times 10^{-16}$), whereas in anaerobes it was nonexistent ($\rho = -0.023$, $p = 0.704$). This pattern was consistent when correcting for phylogeny (Appendix A.7), and was true for both type I and III systems individually, though was not significant for type II systems of which there were fewer in the dataset Fig A.19.

Similar to our CRISPR analysis, we used PCA and an RF model to find if and where Ku-possessing organisms clustered in trait space. We found that the NHEJ pathway clusters strongly in trait space (Fig A.17), and is favored in soil-dwelling, spore-forming, aerobic microbes, consistent with expectations of where NHEJ will be most important (Fig A.18; [111, 112]).

2.4.5 Predicting RM Incidence

So far, our analyses have not distinguished if temperature and oxygen predict whether a microbe has an intracellular immune system that degrades DNA in general, or whether these traits are specific to CRISPR adaptive immunity. We tested these two possibilities by building an RF model of restriction enzyme incidence using the same stratified sampling approach that we used for CRISPR system type. This model showed decent predictive ability ($\kappa = 0.317$). However, the correlation between variable importance scores for the CRISPR and restriction enzyme RF models was low (Fig 2.3 vs. Fig A.21; Pearson's correlation, $\rho = 0.169$ for mean decrease in Gini Impurity Index, $\rho = -0.0487$ for mean decrease in accuracy; also Figs A.7 and A.8). This result implies that RM systems have different traits determining their incidence than do CRISPR systems (also note PCA plot, Fig A.20). When we directly tested for an association with temperature and oxygen we also found that the number of restriction enzymes was, unlike CRISPR incidence, negatively associated with an anaerobic lifestyle ($m = -4.53877$, $p = 2 \times 10^{-16}$, phylogenetic linear regression), and only marginally significantly associated with a thermophilic lifestyle ($m = 1.51063$, $p = 0.03779$, phylogenetic linear regression). These results were consistent across bootstrapped trees (Table A.3).

2.5 Discussion

We detected a clear association between microbial traits and the incidence of the CRISPR immune system across species. We found that two predictors were especially important for predicting CRISPR incidence, thermophilicity and aerobicity. The links between these two traits and CRISPR were confirmed with annotations from NCBI, and in the case of aerobicity with metagenomic data from the Tara Oceans Project (Appendix A.5; [110]). The relationship between temperature and CRISPR is well known [44, 45, 46], but we lend further support here by formally correcting for shared evolutionary history in our statistical analyses using both parametric and non-parametric approaches.

Previous theoretical models predict that CRISPR will be selected against in environments with dense and diverse viral communities [47, 48], since hosts are less likely to repeatedly encounter the same virus in such environments. These models in turn predict that in high-density host communities CRISPR will not be adaptive, since high host density leads to high viral diversity [47, 48], and that this might explain why potentially slow-growing thermophiles favor CRISPR immunity (as opposed to copiotrophic mesophiles). Our results show a marginal positive association between growth rate and CRISPR incidence, and that group-living microbes seem to favor CRISPR immunity, calling these prior viral diversity and density based explanations into question. Additionally, our analysis suggests that psychrophilic and psychrotolerant species disfavor CRISPR more strongly than mesophiles, which is not clearly explained or predicted by hypotheses based on host density.

We suspect that another factor could be affecting the degree of viral diversity that a host encounters, so that viral diversity is high in colder environments and low in hotter ones. Differences in dispersal limitation among viruses could lead to lower immigration rates in hot environments, as viral decay rates may be low at lower temperatures and high at higher temperatures [113], though this is highly speculative. We note that host dispersal rates are unlikely to affect the viral diversity seen by a host on average unless most of the host population is dispersing, an unrealistic expectation.

Surprisingly, we find that oxygen requirement appears to be just as important of a predictor of CRISPR incidence as temperature, and that this pattern is independent of any effect of temperature. Possibly, this association can be explained by inhibitory effects of CRISPR on NHEJ DNA repair. Type II-A CRISPR systems have been shown to directly interfere with the action of the NHEJ DNA repair pathway in prokaryotes [74]. Reactive oxygen species are produced during aerobic metabolism and can cause DNA damage [111], making NHEJ potentially particularly important in aerobes. Thus, if CRISPR interferes with the NHEJ repair pathway, and this pathway is important in aerobes, we would expect CRISPR incidence to be inversely related to the presence of oxygen. Our data showed a clear interaction between aerobicity and the NHEJ machinery in determining CRISPR incidence that suggests that the link between CRISPR and aerobicity may be mediated by the presence of the NHEJ pathway (Fig 2.4c). The Cas proteins share many structural similarities with proteins implicated in DNA repair, and in some cases prefer to associate with DSBs, and it is perhaps unsurprising that they appear to broadly inhibit the NHEJ pathway whose proteins may be competing for substrate [114]. Nevertheless, the evidence supporting this hypothesis is only preliminary. The negative interaction between CRISPR and Ku should be experimentally confirmed in type I and type III systems. Additionally, our repair versus immunity tradeoff hypothesis could be tested using an experimental evolution setup in which organisms with CRISPR are exposed to DNA damage.

The link that we propose between aerobic metabolism and NHEJ repair is somewhat tenuous. Reactive oxygen species are thought to directly produce single strand breaks which are most often converted to double strand breaks during cell growth, the precise time when repair may be possible via homologous recombination due to the presence of multiple genome copies. That being said, reactive oxygen species can lead to double strand breaks during stationary phase when damage is spatially clustered on the genome [115, 116], when cells experience specific types of starvation that lead to vulnerable single-stranded DNA gaps [117, 118], or when ROS occurs in conjunction with other damaging agents including cyanide [119] and irradiation [120, 121, 122]. Furthermore, while NHEJ certainly will be important during stationary phase, its relevance during growth is unknown. The pathway itself does appear to be more prevalent in environments with oxygen (Figs A.17 and A.18). Nevertheless, we have no ability to assess causality presently, and the strong interaction between Ku and aerobicity on CRISPR incidence we observed could be the result of some other, as yet unrevealed driver. For example, NHEJ is thought to be important for desiccation resistance [123, 124], and many organisms facing this specific threat are likely to be aerobic.

As an alternative to our NHEJ hypothesis, could patterns in viral diversity explain the relationship between aerobicity and CRISPR incidence? The viral-decay hypothesis we proposed to explain the enrichment of thermophiles with CRISPR does not make sense in this context, since we might expect viruses to decay more readily in the presence of oxygen rather than under anoxic conditions. It is unclear to us why the viruses of anaerobes would be more dispersal limited. Nevertheless, if the viral communities infecting anaerobes were shown to be less diverse than those infecting aerobes this could also explain the increased incidence of CRISPR among these organisms.

We found no strong link between the incidence or number of RM systems on a genome and a thermophilic or anaerobic lifestyle, suggesting that the major drivers of CRISPR incidence are indeed CRISPR specific, consistent with our viral-diversity and NHEJ-inhibition hypotheses.

We were also able to show that CRISPR types vary in in terms of the environments they are found in, with type II systems appearing primarily in host-associated microbes. This phenomenon could be due in part to phylogenetic biases in the dataset, but our use of a phylogenetically independent test set lends credence to the overall trend. We have no clear mechanistic understanding of why *cas9* containing microbes tend to favor a host-associated lifestyle. Nevertheless this result may have practical implications for CRISPR genome editing, since it has recently been found that humans frequently have a preexisting adaptive immune response to variants of the Cas9 protein [125]. We note that type I and III systems do not appear to have a strong link to host-associated lifestyles.

While our dataset spanned a broad phylogenetic range (with some notable exceptions such as the Candidate Phyla Radiation [126]), we had a limited number of microbial traits, which may have obscured some important CRISPR-trait associations. With the number of microbial genomes in public databases constantly expanding, so too should efforts to provide metadata about each of the organisms represented by those genomes. At least part of the problem lies in the lack of a universally accepted controlled vocabulary for microbial traits (similar to that provided by the Gene Ontology Consortium [127]), although some admirable attempts have been made [128, 129]. This would both facilitate the construction of more expansive trait databases, and would help deal with the issue of comparing traits that span many different scales.

The ecological drivers of microbial immune strategy are likely as diverse as the ever-increasing number of known prokaryotic defense systems [13, 42]. The exploratory, database-centered approach we take here can be complemented by targeted studies examining shifts in immune strategy across environmental gradients (e.g., Appendix A.5) to provide a more fine-grained understanding of how microbial populations adapt to their local pathogenic and abiotic environments. Ultimately, experimental manipulations will provide the power to fully validate proposed mechanisms behind ecological patterns in immune strategy.

Chapter 3: Selective maintenance of multiple CRISPR arrays across prokaryotes

3.1 Abstract

Prokaryotes are under nearly constant attack by viral pathogens. To protect against this threat of infection, bacteria and archaea have evolved a wide array of defense mechanisms, singly and in combination. While immune diversity in a single organism likely reduces the chance of pathogen evolutionary escape, it remains puzzling why many prokaryotes also have multiple, seemingly redundant, copies of the same type of immune system. Here, we focus on the highly flexible CRISPR adaptive immune system, which is present in multiple copies in a surprising 28% of the prokaryotic genomes in RefSeq. We use a comparative genomics approach looking across all prokaryotes to demonstrate that, on average, organisms are under selection to maintain more than one CRISPR array. Given this surprising conclusion, we consider several hypotheses concerning the source of selection and include a theoretical analysis of the possibility that a tradeoff between memory span and learning speed could select for both "long-term memory" and "short-term memory" CRISPR arrays.

3.2 Introduction

Just as larger organisms must cope with the constant threat of infection by pathogens, so too must bacteria and archaea. To defend themselves in a given pathogenic environment, prokaryotes may employ a range of different defense mechanisms, and oftentimes more than one [11, 12, 17]. While having multiple types of immune systems may decrease the chance of pathogen evolutionary escape [130], having multiple instances of the same type of system is rather more puzzling. Here we explore this apparent redundancy in the context of CRISPR-Cas immunity.

The CRISPR-Cas immune system is a powerful defense mechanism against mobile genetic elements such as viruses and plasmids, and is the only known example of adaptive immunity in prokaryotes [45, 131]. This system allows prokaryotes to acquire specific immune memories, called "spacers", in the form of short viral genomic sequences which they store in CRISPR arrays in their own genomes [18, 19, 38]. These sequences are then transcribed and processed into short RNA fragments that guide CRISPR-associated (Cas) proteins to degrade matching foreign DNA or RNA [19, 132, 133]. Thus the CRISPR array is the genomic location in which memories are recorded, while the Cas proteins act as the machinery of the immune system.

CRISPR systems appear to be widespread across diverse bacterial and archaeal lineages, with previous analyses of genomic databases indicating that $\sim 40\%$ of bacteria and $\sim 80\%$ of archaea have at least one CRISPR system [53, 68, 134]. These systems vary widely in *cas* gene content and targeting mechanism, although the *cas1* and *cas2* genes involved in spacer acquisition are universally required for a system to be fully functional [19, 68]. Such prevalence suggests that CRISPR systems effectively defend against phage in a broad array of environments. The complete story seems to be more complicated, with recent analyses of environmental samples revealing that some major bacterial lineages almost completely lack CRISPR systems and that the distribution of CRISPR systems across prokaryotic lineages is highly uneven [52]. Other studies suggest that particular environmental factors can be important in determining whether or not CRISPR immunity is effective (e.g., in thermophilic environments [47, 48]). While previous work has focused on the presence or absence of CRISPR across lineages and habitats, little attention has been paid to the number of systems in a genome.

In fact, the multiplicity of CRISPR systems per individual genome varies greatly, with many bacteria having multiple CRISPR arrays and some having multiple sets of *cas* genes as well (e.g., [56, 58]). CRISPR and other immune systems are horizontally transferred at a high rate relative to other genes in bacteria [16], meaning that any apparent redundancy of systems may simply be the result of the selectively neutral accumulation of systems within a genome. Alternatively, some microbes may experience selection for multiple sets of *cas* genes or CRISPR arrays.

We suspected that prokaryotes may be under selection to maintain multiple CRISPR arrays, given that it is common for organisms across lineages to have multiple systems (as detailed below) and, in some clades, these appear to be conserved over evolutionary time (e.g. [135, 136]). Because microbial genomes have a deletion bias [137, 138], we would expect extraneous systems to be removed over time. Here we construct a test of neutral CRISPR array accumulation via horizontal transfer and loss. Using publicly available genome data we show that the number of CRISPR arrays in a wide range of prokaryotic lineages deviates from this neutral expectation by approximately two arrays. Thus we conclude that, on average, prokaryotes are under selection to have multiple CRISPR arrays. We go on to discuss several hypotheses for why having multiple arrays might be adaptive. Finally, we suggest that a tradeoff between the rate of acquisition of immune memory and the span of immune memory could lead to selection for multiple CRISPR arrays.

3.3 Methods

3.3.1 Dataset

All available completely sequenced prokaryotic genomes (all assembly levels, bacteria and archaea) were downloaded from NCBI's non-redundant RefSeq database FTP site ([139]) on December 23, 2017. Genomes were scanned for the presence of CRISPR arrays using the CRISPRDetect v2.2 software [77]. We used default settings except that we did not take the presence of *cas* genes into account in the scoring algorithm (to avoid circularity in our arguments), and accordingly used a quality score cutoff of three, following the recommendations in the CRISPRDetect documentation. CRISPRDetect also identifies the consensus repeat sequence and determines the number of repeats for each array. Presence or absence of *cas* genes were determined using genome annotations from NCBI's automated genome annotation pipeline for prokaryotic genomes [83]. We discarded genomes that lacked a CRISPR array in any known members of their species. In this way we only examined genomes known to be compatible with CRISPR immunity.

3.3.2 Test for selection maintaining multiple arrays

We detect selection by comparing non-functional (i.e., neutrally-evolving) and functional (i.e., potentially-selected) CRISPR arrays. Since all known CRISPR systems require the presence of *cas1* and *cas2* genes in order to acquire new spacers, we use the presence of both genes on a genome as a marker for functionality of arrays on that genome and the absence of one or both genes as a marker for non-functionality (validated in Appendix B.1). This differentiation allows us to consider the probability distributions of the number of CRISPR arrays i in non-functional (N_i) and functional (F_i) genomes, respectively.

We start with our null hypothesis that in genomes with functional CRISPR systems possession of a single array is highly adaptive (i.e. viruses are present and will kill any susceptible host) but additional arrays provide no additional advantage. Thus these additional arrays will appear and disappear in a genome as the result of a neutral birth/death horizontal transfer and loss process, where losses are assumed to remove an array in its entirety. This hypothesis predicts that the non-functional distribution will look like the functional distribution shifted by one (S_i):

$$H_0 : N_i \approx S_i = F_{i+1} / \sum_{j=1}^{\infty} F_j \tag{3.1}$$

for $i \geq 0$ (S_i renormalized to account for loss of 0-array category).

We take two approaches to testing this hypothesis: one parametric from first principles and one non-parametric with less power but fewer assumptions. In our parametric approach, we construct a stochastic model of neutral array accumulation and find that both N_i and S_i should fit a negative binomial distribution at equilibrium (see Appendix B.2 for derivation). We calculate point maximum likelihood estimates of the means of these fitted distributions ($\hat{\mu}_N$ and $\hat{\mu}_S$). We expect that $\hat{\mu}_S > \hat{\mu}_N$ if more than one array is selectively maintained, and we bootstrap confidence intervals on these estimates by resampling with replacement from our functional and non-functional array count distributions in order to determine whether the effect is significant.

We also construct a non-parametric test for selection by determining at what shift s the mismatch between $F_{i+s}/\sum_{j=s}^{\infty} F_j$ and N_i, measured as the sum of squared differences between the distributions, is minimized:

$$s^\star = \underset{s}{\operatorname{argmin}} \sum_{i=0}^{\infty} \left(N_i - F_{i+s} / \sum_{j=s}^{\infty} F_j \right)^2 . \tag{3.2}$$

Under our null hypothesis $s^\star = 1$, and a value of $s^\star > 1$ implies that selection maintains more than one array. Our parametric test is superior to $s^\star$ because it can detect if selection maintains more than one array across the population on average, but not in all taxa, so that the optimal shift is fractional.

We note that the array accumulation process underlying these methods assumes that CRISPR arrays are primarily lost all-at-once (e.g. due to recombination between flanking insertion sequences [50, 140]) rather than through a process of gradual decay due to spacer loss. Experimental evidence supports spontaneous loss of the entire CRISPR array [51], as do comparisons between closely related genomes [50]. We discuss this assumption and provide evidence supporting spontaneous loss in Appendix B.3.

3.3.3 CRISPR spacer turnover model

We develop a simple deterministic model of the spacer turnover dynamics in a single CRISPR array of a bacterium exposed to n viral species (i.e., disjoint protospacer sets; Appendix B.4). This model allows us to specify the strength of priming (i.e., if a CRISPR array has a spacer targeting a particular viral species, the rate of spacer acquisition towards that species is increased; [141, 142]) and a functional form for spacer loss over time.

Using this model we can determine the optimal spacer acquisition rate given a particular pattern of pathogen recurrence in the environment. If the optima for distinct recurrence patterns do not overlap, it indicates that multiple arrays would be required to simultaneously combat viral species with these distinct recurrence patterns. For model analysis see Appendix B.4.

We consider two functional forms for spacer loss based on known features of CRISPR biology. (1) The rate of per-spacer loss increases linearly with locus length. This form is based on the observation that spacer loss appears to occur via homologous recombination between repeats [31, 143, 144], which becomes more likely with increasing numbers of spacers (and thus repeats). (2) The length of an array is capped at some fixed "effective" number of spacers. This form is based on evidence

Genome Set	With CRISPR array	> 1 CRISPR arrays	> 1 signature *cas* genes	> 1 type of signature *cas* gene
Full dataset	44%	28%	5%	2%
Subsampled	40%	24%	9%	5%

Table 3.1: CRISPR array and *cas* multiplicity across prokaryotic genomes.

that mature crRNA transcripts from the leading end of the CRISPR array are far more abundant than those from the trailing end, and that this decay over the array happens quickly (most transcripts are from the first few spacers; [145, 146, 147]). We analyze both models (Appendix B.4), though they give qualitatively similar results, and so we focus on case (1) in the Results.

3.4 Results

3.4.1 Having more than one CRISPR array is common

Almost half of the prokaryotic genomes in the RefSeq database have at least one CRISPR array, and around a quarter have multiple CRISPR arrays (Table 3.1). In contrast to this result, having more than one set of *cas* targeting genes is not nearly as common. We counted the number of signature targeting genes diagnostic for type I, II, and III systems in each genome (*cas3*, *cas9*, and *cas10* respectively [30]). Only 5% of all genomes have more than one targeting gene. Of these cases, about half correspond to cases of multiple types of targeting genes in the same genome (Table 3.1).

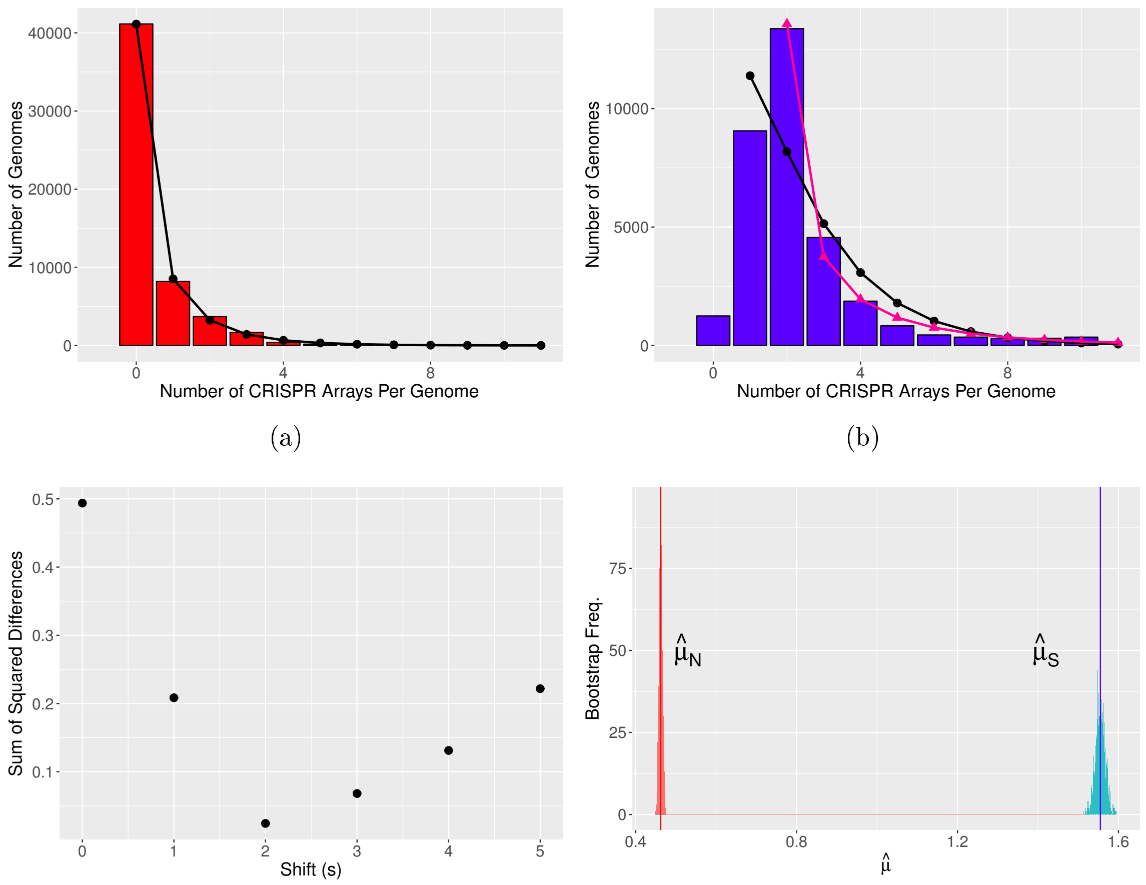
Some species are overrepresented in RefSeq (e.g. because of medical relevance), and we wanted to avoid results being driven by just those few particular species.

We controlled for this bias by randomly sub-sampling 10 genomes from each species with more than 10 genomes in the database and found broadly similar results (Table 3.1).

3.4.2 Selection maintains multiple CRISPR arrays

We leveraged the difference between functional and non-functional genomes, within each of which the process of CRISPR array accumulation should be distinct (Fig 3.1 and Table B.1). Non-functional CRISPR arrays should accumulate neutrally in a genome following background rates of horizontal gene transfer and gene loss (see Methods). We constructed two point estimates of this background accumulation process using our parametric model to infer the distribution of the number of arrays. One estimate came directly from the non-functional genomes ($\hat{\mu}_N$, Fig 3.1(a)). The other came from the functional genomes, assuming that having one array is adaptive in these genomes, but that additional arrays accumulate neutrally ($\hat{\mu}_S$, Fig 3.1(b)). If selection maintains multiple (functional) arrays, then we should find that $\hat{\mu}_N < \hat{\mu}_S$. We found this to be overwhelmingly true, with about two arrays on average seeming to be evolutionarily maintained across prokaryotic taxa ($\Delta\mu = \hat{\mu}_S - \hat{\mu}_N = 1.09 \pm 0.03$). We bootstrapped 95% confidence intervals of our estimates by resampling genomes (Table B.1) and found that the bootstrapped distributions did not overlap, indicating a highly significant result (Fig 3.1(d)). To control for the possibility that multiple sets of *cas* genes in a small subset of genomes could be driving this selective signature, we restricted our dataset only to genomes with one or fewer signature targeting genes (*cas3*, *cas9*, or *cas10* [30, 68]) and one or fewer copies each of the genes necessary for spacer acquisition (*cas1* and *cas2*). Even in this restricted set selection maintains more than one (functional) CRISPR array, though the effect size is smaller ($\Delta\mu = 0.61 \pm 0.02$, Fig B.1).

In order to further confirm our results we (1) subsampled overrepresented taxa in the dataset, (2) performed phylogenetically-corrected tests to account for possible evolutionary correlation in rates of horizontal gene transfer (HGT), (3) considered the effects of potential physical linkage between *cas* genes and CRISPR arrays, (4) looked for artifacts as a factor of genome assembly level, (5) considered the potential effects of CRISPR immunity on rates of HGT [148], and, finally, (6) merged arrays with identical repeats to account for the potential formation of neo-CRISPR arrays by off-target spacer integration [149] as well as other array duplication events. In all cases our qualitative result of selection ($\Delta\mu > 0$) holds (Appendices B.5 and B.6). Additionally, we explored the possibility that the CRISPR detection algorithm we used could be biased and/or suffering from a high rate of false positives, and found our qualitative result did not change when using a higher score cutoff, restricting to arrays with experimentally verified repeat sequences, or using an alternative algorithm (Appendix B.7).



(a)

(b)

(c)

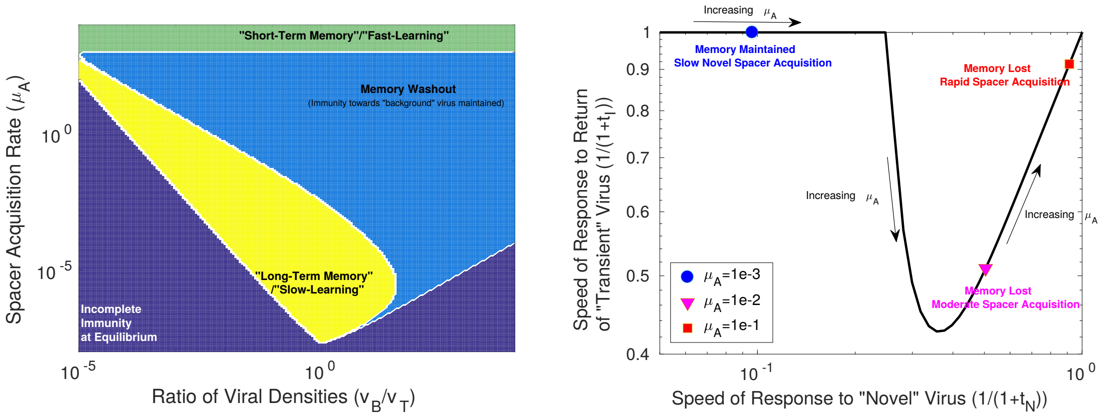
(d)

Figure 3.1: Selection maintains more than one CRISPR array on average across prokaryotes. (a-b) Distribution of number of arrays per genome in (a) genomes with non-functional CRISPR immunity and (b) genomes with putatively functional CRISPR immunity. The tails of these distributions are cut off for ease of visual comparison (24 genomes with > 10 arrays in (a) and 498 genomes with > 10 arrays in (b)). In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and pink triangles to the double-shifted distribution ($s = 2$). Note that the fit to the double-shifted distribution (pink triangles in b) visually resembles the distribution of non-functional arrays shown in (a). (c) We formally quantify the difference between the non-functional/shifted function distributions and find an optimal shift of $s^{\star} = 2$. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show no overlap with 1000 bootstrap replicates.

3.4.3 A tradeoff between memory span and acquisition rate could select for multiple arrays in a genome

We built a simple model of spacer turnover dynamics in a single CRISPR array. We consider three patterns of viral residency in the environment corresponding to the major threats prokaryotes are likely to face: (1) "background" viruses that coexist with their hosts over long time periods [150], (2) periodic outbreaks of a particular "transient" virus that enters and leaves the system [151], and (3) "novel" viruses that a host has not previously encountered (see Methods and Appendix B.4). For very high spacer acquisition rates, a host will be able to effectively defend against all three types of viral species simultaneously, because the acquisition of immunity will be nearly instantaneous ("short-term memory"/"fast-learning" in Figs 3.2 and B.2). Such high rates are unrealistic due to physical constraints on the speed of adaptation as well as the evolutionary constraint of autoimmunity (Appendix B.8, [25, 27, 61, 152, 153]). CRISPR adaptation is rapid, but it is not instantaneous, and infected but susceptible hosts will often perish before a spacer can be acquired [154]. This is precisely why the memory-like quality of CRISPR immunity is advantageous.

Our analysis also reveals a region of parameter space with low spacer acquisition rates in which immunity is maintained towards both background and transient viruses ("long-term memory"/"slow-learning" in Fig 3.2(a)). The "long-term memory"/"slow-learning" region of parameter space is separated from the "short-term memory"/"fast-learning" region of parameter space by a "memory-washout" region in which spacer turnover is high so that memory is lost but acquisition is



(a)

(b)

Figure 3.2: Immune memory is maximized at intermediate and low spacer acquisition rates, creating a tradeoff with the speed of immune response to novel threats. (a) Phase diagram of the behavior of our CRISPR array model with two viral species, a constant "background" population and a "transient" population that leaves and returns to the system at some fixed interval. The yellow region indicates that immunity towards both viral species was maintained. The green region indicates where immune memory was lost towards the transient phage species, but reacquired almost immediately upon phage reintroduction ($t_I < 10^{-5}$, where t_I is the time to first spacer acquisition after the return of the species to the system following an interval of absence). The light blue region indicates that only immunity towards the background species was maintained (i.e., immune memory was rapidly lost and $t_I > 10^{-5}$). Dark blue indicates where equilibrium spacer content towards one or both species did not exceed one despite both species being present in the system (Appendix B.4). (b) The tradeoff between memory span and learning speed. The speed of immune response to the transient virus is plotted against the speed of response to a novel virus to which the system has not been previously exposed (so that there are no spacers targeting this virus), over a range of spacer acquisition rates ($\mu_A \in [10^{-3.5}, 1]$), and letting the densities of transient and background viruses be equal. The speed of immune response to a virus is defined as $1/(1+t)$ where t is time to first spacer acquisition ($t = 0$ if memory is maintained). The speed of response to the novel virus is therefore $1/(1+t_N)$ where t_N is the time to first spacer acquisition towards this virus. For specifics on calculating t_I and t_N see Appendix B.4. Note that μ_A is the number of spacers expected to be acquired per viral particle adsorbed to the host cell.

not rapid enough to quickly re-acquire immunity towards the transient virus. This sets up a tradeoff between the ability of a host to defend against both transient and novel viruses, since the response time towards novel threats in the "long-term memory"/"slow-learning" region of parameter space is slow (Fig 3.2(b)), but memory of transient threats is lost if spacer acquisition rates are increased. Thus, in order to maximize novel spacer acquisition and memory span simultaneously, a two-system solution will be required.

Additionally, priming expands the "washout" region of parameter space, because high spacer uptake from background viruses will crowd out long term immune memory (Fig B.3). This suggests that priming strengthens the learning vs. memory tradeoff and makes a two-array solution more likely.

3.4.4 Selection varies between taxa and system types

A handful of species in the dataset were represented by a large number of genomes (> 1000), with at least one each of functional and non-functional genomes. We performed our test for selection on each of these species individually and found a large amount of variation between species (Table B.2). Notably, genomes of *Campylobacter jejuni*, *Escherichia coli*, and *Salmonella enterica* show evidence for selection against having a functional CRISPR array (negative $\Delta\mu$), indicating that CRISPR immunity is selected against on average in some groups of organisms. Previous work has shown that CRISPR in *E. coli* and *S. enterica* appears to be non-functional as an immune system under natural conditions [155, 156]. We had relatively few archaeal genomes ($< 1\%$ of dataset), but they showed a clear signal of selection maintaining multiple arrays ($\Delta\mu = 1.05 \pm 0.56$, Fig B.4).

While we do not have direct information on system type for the majority of arrays in our dataset, we can subdivide genomes into those containing the signature *cas* targeting genes for type I, II, or III CRISPR systems (*cas3*, *cas9*, and *cas10* respectively) as a proxy for system type [30]. The number of arrays per genome differed significantly among system types (Fig B.5), and the largest difference was between genomes with class I targeting proteins which had around 2 arrays on average (type I and type III, 2.10 and 1.96 respectively) and class II targeting proteins which only had one array on average (type II, 1.05). We excluded genomes with multiple types of targeting genes for this analysis.

We cannot run our test for selection directly on these subsets of the data, since they exclude genomes without arrays or *cas* genes. Instead we classified species into types if the only observed targeting gene type among all representatives of that species corresponded to a a particular type. Thus we can test for our signature of selection among species that "favor" a particular type of CRISPR system. All types showed a signature of multi-array selection ($\Delta\mu = 1.09 \pm 0.05$, 0.62 ± 0.02, 1.79 ± 0.06 respectively). In particular type III "species" had an exceptionally strong signal, and organisms in this group may be under selection to maintain three arrays.

3.5 Discussion

3.5.1 Selection maintains multiple CRISPR arrays across prokaryotes

On average, prokaryotes are under selection to maintain more than one CRISPR array. The number of CRISPR arrays in a genome appears to follow a negative binomial distribution quite well (Figs 3.1, B.1, B.6, and B.7), consistent with our theoretical prediction. We note that, due to the large size of this dataset, formal goodness-of-fit tests to the negative binomial distribution always reject the fit due to small but statistically significant divergences from the theoretical expectation.

Our test for selection is conservative to the miscategorization of arrays as "functional" or "non-functional." Miscategorizations could occur for several reasons because preexisting spacers may continue to confer immunity, some CRISPR arrays may be conserved for non-immune purposes (e.g. [155, 157]), and intact acquisition machinery is no guarantee of system functionality. Our test is conservative precisely because of such miscategorizations, as they should drive $\hat{\mu}_N$ and $\hat{\mu}_S$ closer to each other. Selection against having a CRISPR array in non-functional genomes could produce a false signature of multi-array selection, but this is unlikely because non-functional arrays probably carry extremely low or nonexistent associated costs [158].

Our test for selection is also robust to false positive or negative array discovery rates because it relies on relative differences between array counts in functional and non-functional genomes, not their absolute values. The only problem could arise if the discovery error rates were different between the two categories; however, the array detection process did not take functionality into account and we found only a marginal difference in CRISPRDetect confidence scores between the two groups (Fig B.8). We further confirmed this robustness to peculiarities of the detection algorithm by changing our CRISPRDetect score threshold and comparing to the distribution of arrays per genome in the independently-generated CRISPR Database (Appendix B.7; [159]).

Finally, we note that $\hat{\mu}_N$ and $\hat{\mu}_S$ take on a range of values depending on what subset of taxa/genomes is considered. This is to be expected as each set of species will occupy a distinct environment in terms of both the rate of horizontal gene transfer and the usefulness of CRISPR immunity. Nevertheless, our qualitative signature of selection is robust to this quantitative variability.

3.5.2 Why have multiple CRISPR-Cas systems?

Possibly, multiple arrays could be selectively maintained even in the absence of any fitness advantage if, by chance, each array acquired complementary spacer content towards distinct viral targets. In type I and II systems, if arrays share acquisition machinery then such complementarity is unlikely because priming will ensure both arrays contain spacers towards any target encountered, meaning that the content of the two arrays will be largely redundant [141]. Type III systems are unprimed and have slow spacer acquisition rates [160], and therefore may be maintained via spacer complementarity, perhaps explaining why species favoring type III systems appear to experience selection maintaining three rather than just two CRISPR arrays. Even in type I and II systems, if each array is associated with a separate set of spacer acquisition machinery, then cross-priming will be less likely and complementarity could arise. Nevertheless, this does not explain the multi-array conservation we see in genomes with only a single set of *cas* genes.

Therefore we are left with two broadly defined reasons why having multiple CRISPR arrays might be adaptive: (1) multiple similar systems could lead to improved immunity through redundancy and (2) multiple dissimilar systems could allow specialization towards distinct types of threats.

In the case of similar systems, immunity could be improved by (a) an increased spacer acquisition rate, (b) an increased rate of targeting, or (c) a longer time to expected loss of immunity. Duplication of *cas* genes could increase uptake (a) and targeting rates (b), but again this could not explain our results with a single set of *cas* genes. Alternatively, duplication of CRISPR arrays could increase targeting (b) by producing a larger number of crRNA transcripts or increase memory duration (c) through spacer redundancy. However, the effectiveness of crRNA may actually decrease in the presence of competing crRNAs [158, 161, 162], and spacer redundancy across multiple arrays has little advantage over redundancy within a single array (Appendix B.9). At a larger scale, redundancy of either arrays or *cas* genes might be a form of bet-hedging against mutation-induced loss of functionality of the CRISPR system [51, 69].

Alternatively, dissimilar systems could help defend against diverse threats. Diverse *cas* genes may allow hosts to evade broadly-acting anti-CRISPR proteins encoded by some viruses [60, 163]. Indeed, promiscuous type III Cas proteins are often encoded alongside type I systems and can cooperate to target phages that have mutated to escape type I targeting [164]. Empirically, we see the inclusion of genomes with multiple *cas* targeting genes increases the effect size of our test for selection, suggesting these factors may play a role. However, these *cas*-diversity hypotheses cannot explain the signature for multi-array adaptiveness observed among genomes with only a single set of targeting proteins. We note that we observed our signature of selection on multiple arrays both when limiting our analyses to arrays with identical (Appendix B.10) and dissimilar (Appendix B.6) repeat sequences. Therefore selective maintenance of multiple arrays does not appear to be isolated to genomes with arrays of the same type or different types, but rather to be a much more general phenomenon. Additionally, given the very small number of genomes with multiple types of *cas* targeting genes in our dataset, it is unlikely that selection for multiple types of systems is particularly widespread even if it does exist in some cases.

We develop a hypothesis that diversity in spacer acquisition rate among arrays could lead to selection for multiple arrays. Our theoretical model illustrates how factors intrinsic to the mechanism of CRISPR immunity could create a tradeoff between memory span and learning speed. Either the physical loss of spacers due to homologous recombination or the effective loss of spacers due to differential transcription along the array leads to a qualitatively similar result. In both cases, rapid spacer uptake causes rapid spacer loss (either physical or effective), producing the aforementioned tradeoff. A low acquisition rate system is unlikely to pick up a spacer from a single viral exposure, but, over a long time-frame, it may acquire spacers from viruses that periodically reappear in the system. Additionally, recombination between arrays [165] could potentially facilitate the passage of memories between "fast" and "slow" arrays, allowing short-term memories to become long-term ones.

While we do not have empirical evidence that rate variation drives the observed signature of selection of multiple arrays, this hypothesis remains attractive since it can explain the signature even in the absence of multiple sets of *cas* genes. Acquisition rates vary between arrays, even on the same genome [63, 150], and even when those arrays share *cas* genes and have an identical or nearly identical repeat sequence [166, 167]. We found no clear link between the diversity of repeat sequences and a proxy for spacer acquisition rates (Appendix B.10). Further, we found indications of selection even when restricting to arrays with identical repeats (Appendix B.10). Thus the factors influencing acquisition rate appear to be idiosyncratic, perhaps related to the genomic position of the CRISPR array.

When partial spacer-target matches exist, variability in spacer acquisition rates among arrays will be largely irrelevant because priming will ensure rapid acquisition of new spacers. On the other hand, when no match exists, either due to spacer loss or the introduction of a truly novel viral species into the environment, primed spacer uptake will not occur. Thus the rate at which a host encounters novel threats will determine the importance of the baseline spacer acquisition rate. In environments where novel viruses are frequently encountered, small differences in acquisition rate can be important for host fitness, whereas in environments where host and virus pairs consistently coevolve over time priming will be the more important phenomenon.

Finally, our examination of immune configuration is likely relevant to the full range of prokaryotic defense mechanisms. In contrast to previous work focusing on mechanistic diversity (e.g. [48, 63, 130, 152]), we emphasize the importance of the multiplicity of immune systems in the evolution of host defense. As we suggest, a surprising amount of strategic diversity may masquerade as simple redundancy.

Chapter 4: Immune Loss as a Driver of Coexistence During Host-Phage Coevolution

4.1 Abstract

Bacteria and their viral pathogens face constant pressure for augmented immune and infective capabilities, respectively. Under this reciprocally imposed selective regime, we expect to see a runaway evolutionary arms race, ultimately leading to the extinction of one species. Despite this prediction, in many systems host and pathogen coexist with minimal coevolution even when well-mixed. Previous work explained this puzzling phenomenon by invoking fitness tradeoffs, which can diminish an arms race dynamic. Here we propose that the regular loss of immunity by the bacterial host can also produce host-phage coexistence. We pair a general model of immunity with an experimental and theoretical case study of the CRISPR-Cas immune system to contrast the behavior of tradeoff and loss mechanisms in well-mixed systems. We find that, while both mechanisms can produce stable coexistence, only immune loss does so robustly within realistic parameter ranges.

4.2 Introduction

While the abundance of bacteria observed globally is impressive [126, 168, 169], any apparent microbial dominance is rivaled by the ubiquity, diversity, and abundance of predatory bacteriophages (or "phages"), which target these microbes [170, 171, 172, 173, 174]. As one might expect, phages are powerful modulators of microbial population and evolutionary dynamics, and of the global nutrient cycles these microbes control [168, 170, 172, 173, 175, 176, 177, 178, 179, 180]. Despite this ecological importance, we still lack a comprehensive understanding of the dynamical behavior of phage populations. More specifically, it is an open question what processes sustain phages in the long term across habitats.

Bacteria can evade phages using both passive forms of resistance (e.g. receptor loss, modification, and masking) and active immune systems that degrade phages (e.g. restriction-modification systems, CRISPR-Cas) [71]. These defenses can incite an escalating arms race dynamic in which host and pathogen each drive the evolution of the other [8, 9]. However, basic theory predicts that such an unrestricted arms race will generally be unstable and sensitive to initial conditions [181]. Additionally, if phages have limited access to novel escape mutations, an arms race cannot continue indefinitely [182, 183, 184]. This leads to an expectation that phage populations will go extinct in the face of host defenses [183].

While typically this expectation holds [e.g. 185], phages sometimes coexist with their hosts, both in natural [e.g. 186, 187] and laboratory settings [e.g. 150, 181, 183, 188, 189, 190, 191, 192]. These examples motivate a search for mechanisms to explain the deescalation and eventual cessation of a coevolutionary arms race dynamic, even in the absence of any spatial structure to the environment. Previous authors have identified (1) fluctuating selection and (2) costs of defense as potential drivers of coexistence in well-mixed systems. Here we propose (3) the loss of immunity, wherein the host defense mechanism ceases to function, as an additional mechanism. We focus on intracellular *immunity* (e.g., CRISPR-Cas) in which immune host act as a sink for phages rather than extracellular *resistance* (e.g., receptor modifications), since the former poses more of an obstacle for phages and thus more of a puzzle for explaining long-term coexistence.

Under a fluctuating selection dynamic, frequencies of immune and infective alleles in the respective host and phage populations cycle over time [193, 194, 195, 196]. That is, old, rare genotypes periodically reemerge because the dominant host or pathogen genotype faces negative frequency dependent selection. Fluctuating selection is likely in situations where host immune and phage infectivity phenotypes match up in a one-to-one “lock and key” type manner [195], and there is evidence that arms races do give way to fluctuating selection in some host-phage systems [184]. Fluctuating selection cannot always proceed, though. When novel phenotypes correspond to increased generalism we do not expect past phenotypes to recur [195, 196] since they will no longer be adaptive. Such expanding generalism during coevolution has been seen in other host-phage systems [197]. Thus the relevance of fluctuating selection depends on the nature of the host-phage immune-infective phenotype interaction.

Another possible driver of coexistence are costs incurred by tradeoffs between growth and immunity (for host) or host range and immune evasion (for phage) [190, 198, 199, 200]. A tradeoff between immunity and growth rate in the host can lead to the maintenance of a susceptible host population on which phages can persist [183, 189, 190, 199, 201, 202]. Tradeoffs often imply a high cost of immunity that does not always exist [e.g. 181], particularly in the case of intracellular host immunity, as we show later.

Finally, in large host populations typical of bacteria, even low rates of immune loss could produce a substantial susceptible host subpopulation, which, in turn, could support phage reproduction and coexistence. Such loss of function in the host defenses could be due to either mutation or stochastic phenotypic changes. Delbrück [203] initially described this hypothesis of loss of defense via back-mutation in order to challenge the evidence for lysogeny. Lenski [204] reiterated this hypothesis in terms of phenotypic plasticity and noted that conditioning the production of a susceptible host population on a resistant one could lead to very robust, host-dominated coexistence. More recently, Meyer *et al.* [205] presented an empirical example of a system in which stochastic phenotypic loss of resistance leads to persistence of a coevolving phage population.

We hypothesize that coexistence equilibria will be more robust under an immune loss mechanism than under a tradeoff mechanism [204]. We build a general mathematical model to demonstrate this point and then use a combination of experimental evidence and simulation-based modeling to apply this result to the coevolution of *Streptococcus thermophilus* and its lytic phage 2972 in the context of CRISPR immunity.

4.3 General Immune Loss Model

We begin with a general model that considers two populations of host ("defended" with a functional immune system; "undefended" without) and one population of pathogen. Starting from classical models of bacteria-phage dynamics [198, 206], we add key terms to capture the effects autoimmunity (i.e., a tradeoff), immune loss, and the implicit effects of coevolution. This relatively simple model allows us to analyze steady states and parameter interactions analytically. Later, we examine the CRISPR-Cas immune system in detail and build a model with explicit coevolutionary dynamics.

We examine the chemostat system with resources:

$$\dot{R} = w(A - R) - \frac{evR}{z + R}(D + U) \tag{4.1}$$

defended host:

$$\dot{D} = D\left(\frac{vR}{z + R} - \delta\phi_d P - \alpha - \mu - w\right), \tag{4.2}$$

undefended host:

$$\dot{U} = U\left(\frac{vR}{z + R} - \delta\phi_u P - w\right) + \mu D, \tag{4.3}$$

and phage:

$$\dot{P} = P\left(\delta U(\phi_u\beta - 1) + \delta D(\phi_d\beta - 1) - w\right), \tag{4.4}$$

where parameter definitions and values can be found in Table 4.1 and rationale/references for parameter values in Appendix C.1. However, we describe here the parameters of direct relevance to coexistence.

First, we allow for defended host to come with the tradeoff of autoimmunity (α), which applies naturally to the CRISPR-Cas system examined later. While autoimmunity could either decrease the host growth rate [207] or be lethal, we focus on the latter as lethality will increase the stabilizing effect of this tradeoff [26, 207, 208]. However, we also find similar general results when applying a penalty to the resource affinity or maximum growth rate of the defended host (Appendix C.2, Figs C.1-C.8).

Second, we add flow from the defended to undefended host populations representing loss of immunity at rate μ.

Finally, we model the effect of coevolution by allowing a fraction of even the defended host population to remain susceptible ($0 < \phi_d \leq 1$). In a symmetric fashion, even nominally undefended host may have secondary defenses against phage ($0 < \phi_u \leq 1$).

Symbol	Definition	Value
R	Resources	$R_0 = 350\,\mu$g/mL
D	Defended Host	$D_0 = 10^6$ cells/mL
U	Undefended Host	$U_0 = 10^2$ cells/mL
P	Phage	$P_0 = 10^6$ particles/mL
e	Resource consumption rate of growing bacteria	$5 \times 10^{-7}\,\mu$g/cell
v	Maximum bacterial growth rate	1.4 divisions/hr
z	Resource concentration for half-maximal growth	$1\,\mu$g/mL
A	Resource pool concentration	$350\,\mu$g/mL
w	Flow rate	0.3mL/hr
δ	Adsorption rate	10^{-8} mL per cell per phage per hr
β	Burst Size	80 particles per infected cell
ϕ_u	Degree of susceptibility of undefended host	1
ϕ_d	Degree of susceptibility of defended host	0
α	Autoimmunity rate	2.5×10^{-5} deaths per individual per hr
μ	Rate of immune inactivation/loss	5×10^{-4} losses per individual per hr

Table 4.1: Definitions and oft used values/initial values of variables, functions, and parameters for the general mathematical model

We analyze our model analytically as well as numerically to verify which equilibria are reachable from plausible (e.g., experimental) starting values (Appendix C.3).

Assuming no phage coevolution ($\phi_d = 0$), this model has a single analytic equilibrium in which all populations coexist (Table C.1). In Fig 4.1, we explore model behavior under varying rates of autoimmunity (α) and immune loss (μ). Clearly when autoimmunity and loss rates surpass unity, defended host go extinct in the face of excessive immune loss and autoimmune targeting. At the opposite parameter extreme, we see coexistence disappear from the numeric solutions (Fig 4.1b) as phage populations collapse. This leads to a band of parameter space where coexistence is possible, stable, and robust. In this band, autoimmunity and/or immune loss occur at high enough rates to ensure maintenance of coexistence, but not so high as to place an excessive cost on immunity. Crucially, this band is much more constrained in the α-dimension, with autoimmunity restricted to an implausibly high and narrow region of parameter space. This suggests a greater robustness of coexistence under an immune loss mechanism even at low loss rates (Fig 4.1, Figs C.2-C.8). To assess more directly the degree of robustness of each driver of coexistence we can perturb our system and see its response. We move our system away from equilibrium $\tilde{X}$ so that $X' = \tilde{X} \exp(\gamma(Y - \frac{1}{2}))$ where $Y \sim$ Uniform$[0, 1]$, and then solve numerically using X' as our initial condition. Under increasing levels of perturbation the system is less likely to reach stable coexistence, specifically in the α-dimension, indicating that autoimmunity produces a far less robust coexistence regime (Fig 4.1c-e, Figs C.2-C.8).

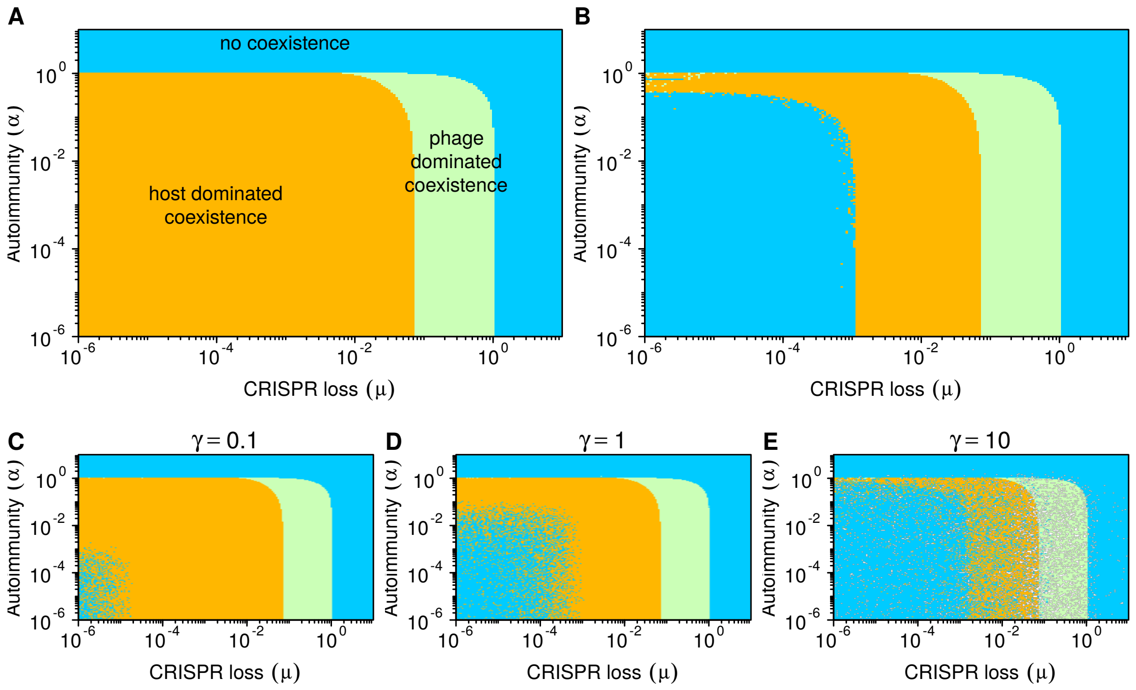


Figure 4.1: Model behavior under variations in the rates of autoimmunity (α) and CRISPR-Cas system loss (μ). Equilibria (Table C.1) derived from Equations 4.1-4.4 are shown in (a) where orange indicates a stable equilibrium with all populations coexisting and defended host dominating phage populations, green indicates that all populations coexist but phages dominate, and blue indicates that defended bacteria have gone extinct but phages and undefended bacteria coexist. In (b) we find numerical solutions to the model at 80 days using realistic initial conditions more specific to the experimental setup ($R(0) = 350$, $D(0) = 10^6$, $U(0) = 100$, $P(0) = 10^6$). In this case orange indicates coexistence at 80 days with defended host at higher density than phages, green indicates a phage-dominated coexistence at 80 days, and blue indicates that coexistence did not occur. Numerical error is apparent as noise near the orange-blue boundary. We neglect coevolution and innate immunity in this analysis ($\phi_u = 1$, $\phi_d = 0$). (c-e) Phase diagrams with perturbed starting conditions. Numerical simulations with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X} \exp(\gamma(Y - \frac{1}{2}))$ where $Y \sim U[0,1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. A single simulation was run for each parameter combination.

If we add large amounts of innate immunity to undefended host ($\phi_u < 0.5$), we find phage-dominated coexistence for a wider range of α (Fig C.10). This result is in line with the counterintuitive suggestion that higher immunity may increase phage density by allowing the host population to increase in size [48]. However, secondary defense has minimal effects for more plausible levels of protection (ϕ_u closer to 1).

In the case of phage coevolution ($\phi_d > 0$), the equilibria still have closed forms, but are not easily representable as simple equations and so are not written here. When $\phi_d > \frac{1}{\beta}$, defended host contribute positively to phage growth, eventually shifting the coexistence equilibrium from host to phage dominance (Fig C.9).

4.4 A Case Study: CRISPR-Phage Coevolution

The CRISPR (Clustered Regularly Inter-spaced Short Palindromic Repeats) prokaryotic adaptive immune system incorporates specific immune memory in the form of short sequences of DNA acquired from foreign genetic elements ("spacers") and then uses this memory to target the corresponding sequences ("protospacers") during subsequent infections [18, 19, 38, 209]. CRISPR can lead to rapidly escalating arms races between bacteria and phages [150, 210, 211], in which evolutionary and population dynamics occur on the same timescale [150, 212, 213, 214].

CRISPR-Cas can quickly drive phages extinct in an experimental setting [185], but in some cases long-term CRISPR-phage coexistence has been observed [150]. Previous theoretical and limited experimental work has explained short-term coexistence through tradeoffs and spacer loss [215], and long-term coexistence by invoking continued coevolution via fluctuating selection [214] or tradeoffs with host switching to a constitutive defense strategy such as surface receptor modification [63, 216].

However, these previous hypotheses are insufficient to explain simple coevolution experiments with *Streptococcus thermophilus* (type II-A CRISPR-Cas system) and its lytic phage 2972 resulting in long-term coexistence [26, 150]. In these experiments, bacteria are resource-limited and appear immune to phages, implying they have "won" the arms race and that phages are persisting on a small susceptible subpopulation of hosts. Deep sequencing of the same experimental system shows dominance by a few spacers that drift in frequency over time, inconsistent with a fluctuating selection dynamic [26]. Specifically, these results contradict the coexistence regime seen in the Childs et al. [214, 217] model, wherein host are phage-limited and the system undergoes a fluctuating selection dynamic. Thus either (1) costs associated with CRISPR immunity or (2) the loss of CRISPR immunity is playing a role in maintaining susceptible host subpopulations on which phages can persist.

In this system, the primary cost of a functional CRISPR-Cas system is autoimmunity via the acquisition of self-targeting spacers. It is unclear how or if bacteria distinguish self from non-self during the acquisition step of CRISPR immunity [25, 27, 61, 152, 153]. In *S. thermophilus*, experimental evidence suggests that there is no mechanism of self vs. non-self recognition and that self-targeting spacers are acquired frequently [27], which implies that autoimmunity may be a significant cost.

Outright loss of CRISPR immunity at a high rate could also lead to coexistence. The bacterium *Staphylococcus epidermidis* loses phenotypic functionality in its CRISPR-Cas system, either due to wholesale deletion of the relevant loci or mutation of essential sequences (i.e. the leader sequence or *cas* genes), at a rate of 10^{-4}-10^{-3} inactivation/loss events per individual per generation [51]. Functional CRISPR loss has been observed in other systems as well [143, 218].

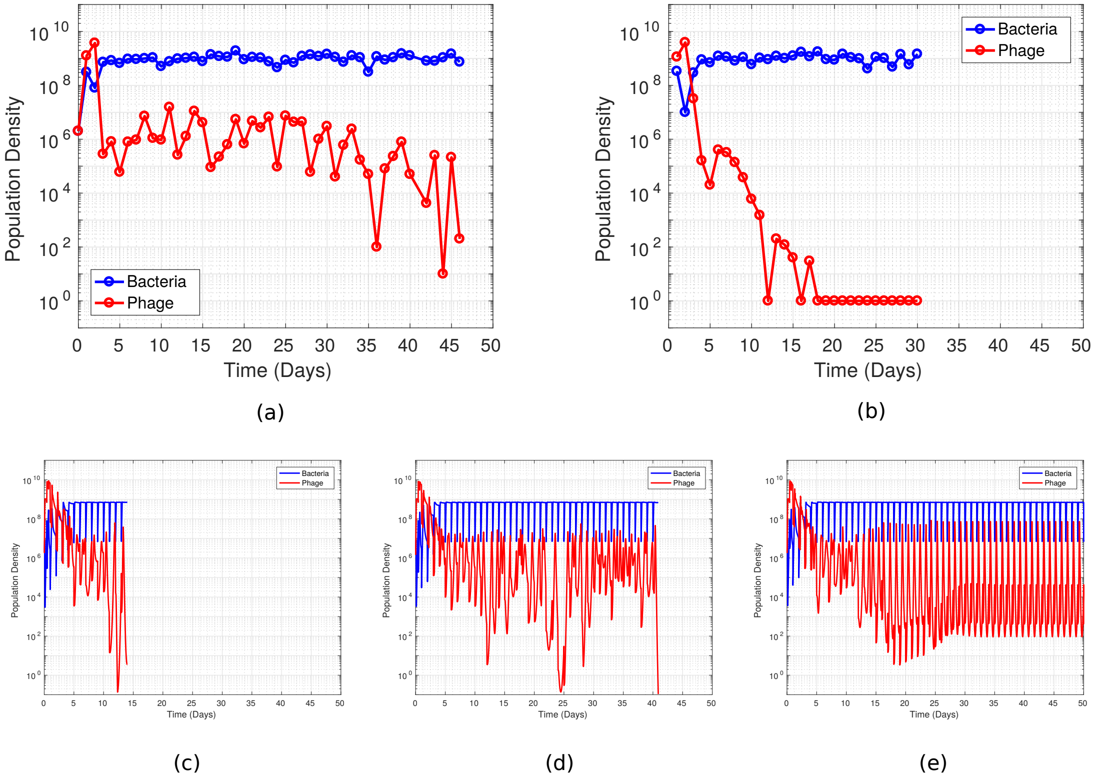
Below we replicate the serial-transfer coevolution experiments performed by Paez-Espino et al. [26, 150] and develop a simulation-based coevolutionary model to explain the phenomenon of coexistence.

4.4.1 Experiments

We performed long-term daily serial transfer experiments with *S. thermophilus* and its lytic phage 2972 in milk, a model system for studying CRISPR evolution (see Appendix C.4 for detailed methods). We measured bacteria and phage densities on a daily basis. Further, on selected days we PCR-amplified and sequenced the CRISPR1 and CRISPR3 loci, the two adaptive CRISPR loci in this bacterial strain.

From the perspective of density, phages transiently dominated the system early on, but the bacteria quickly took over and by day five appeared to be resource-limited rather than phage-limited (Fig 4.2a,b). This switch to host-dominance corresponded to a drop in phage populations to a titer two to three orders of magnitude below that of the bacteria. Once arriving at this host-dominated state, the system either maintained quasi-stable coexistence on an extended timescale (over a month and a half), or phages continued to decline and went extinct relatively quickly (Fig 4.2a,b). We performed six additional replicate experiments which confirmed this dichotomy between either extended coexistence (4 lines quasi-stable for > 2 weeks) or quick phage extinction (2 lines < 1 week) (Fig C.11).

Sequencing of the CRISPR1 and CRISPR3 loci revealed the rapid gain of a single spacer (albeit different spacers in different sequenced clones) in CRISPR1 followed by minor variation in spacer counts with time (Fig C.12), with CRISPR1 being more active than CRISPR3. We tracked the identity of the first novel spacer in the CRISPR1 array over time. We found a cohort of four spacers that persisted over time and were repeatedly seen despite a small number of samples taken at each time point (less than 10 per time point; Table 4.2). Other spacers were sampled as well, but this small cohort consistently reappeared while other spacers were only found at one or two timepoints, indicating this cohort was dominating the system (Table C.2). Such a pattern is inconsistent with a fluctuating selection hypothesis. Further, we did not observe frequent spacer loss in the CRISPR1 or CRISPR3 arrays.



(a)

(b)

(c)

(d)

(e)

Figure 4.2: Serial transfer experiments carried out with *S. thermophilus* and lytic phage 2972 Bacteria are resource-limited rather than phage-limited by day five and phages can either (a) persist at relatively low density in the system on long timescales (greater than 1 month) or (b) collapse relatively quickly. These results agree with those of Paez-Espino [150] where coexistence was observed in *S. thermophilus* and phage 2972 serially transferred culture for as long as a year. Experiments were initiated with identical starting populations and carried out following the same procedure. In (c-e) we show that our simulations replicate the qualitative patterns seen in the data, with an early phage peak, followed by host-dominated coexistence that can either be (c) stable, (d) sustained but unstable, or (e) short-lived. Each plot is a single representative simulation and simulations were ended when phages went extinct. Note that experimental data has a resolution of one time point per day, preventing conclusions about the underlying population dynamics (e.g., cycling), whereas simulations are continuous in time.

	Spacer ID						
Time	**D**	**E**	**F**	**G**	**Total in Cohort**	**Total Sampled Sequences**	**Percent Samples in Cohort**
1	0	0	0	0	0	3	0
2	1	1	1	0	3	4	75
3	2	0	1	1	4	5	80
4	0	0	0	0	0	1	0
5	0	0	2	1	3	7	43
11	1	2	0	2	5	7	71
15	1	1	1	0	3	7	43
25	0	5	0	0	5	8	63
35	0	1	0	2	3	6	50
40	0	0	0	2	2	9	22

Table 4.2: Sequencing data shows four first-order spacers that persist as a high-frequency cohort over time. Samples identified by the first novel spacer added to the array as compared to the wild-type. See Table C.2 for complete spacer dynamics.

4.4.2 CRISPR-phage Coevolutionary Model

We next built a hybrid deterministic/stochastic lineage-based model similar to an earlier model by Childs et al. [214, 217] that explicitly models the coevolutionary dynamics of the CRISPR-phage system wherein bacteria acquire spacers to gain immunity and phages escape spacers via mutations. Our simulations also replicate the resource dynamics of a serial dilution experiment, wherein the system undergoes large daily perturbations.

We model phage mutations only in the protospacer adjacent motif (PAM) region, which is the dominant location of CRISPR escape mutations [150] to prevent the possibility of spacer re-acquisition. This approach differs from previous models which considered mutations in the protospacer region itself [e.g. 47, 48, 214] and thus allowed for the possibility of spacer re-acquisition. We justify modeling only PAM mutations with three arguments. First, the probability of spacer re-acquisition will be quite low if there are many protospacers. Second, re-acquired spacers will already have undergone selection for escape mutation by phage, and, assuming that there are therefore diverse escape mutations in the phage population, these spacers will thus provide limited benefit to the host. Third, as we move away from the PAM along the protospacer sequence, more substitutions are tolerated by the CRISPR matching machinery [219], meaning that mutations farther away from the PAM will be less effective at escaping immunity [220].

We model population dynamics using differential equations for resources:

$$\dot{R} = \frac{-evR}{z+R}\left(U + \sum_i D_i\right) \tag{4.5}$$

CRISPR-enabled bacteria with spacer set X_i:

$$\dot{D}_i = D_i\left(\frac{vR}{z+R} - \delta\left(\sum_j (1 - M(X_i, Y_j))P_j\right) - \alpha - \mu_L\right) \tag{4.6}$$

a pool of undefended bacteria with a missing or defective CRISPR-Cas system:

$$\dot{U} = U\left(\frac{vR}{z+R} - \delta\sum_i P_i\right) + \mu_L \sum_i D_i \tag{4.7}$$

and phages with protospacer set Y_i :

$$\dot{P}_i = \delta P_i\left(U(\beta_i - 1) + \sum_j D_j(\beta_i(1 - M(X_j, Y_i)) - 1)\right), \tag{4.8}$$

and stochastic events occur according to a Poisson process with rate λ:

$$\lambda = \sum_i \lambda_{B_i} + \sum_i \lambda_{P_i} + \sum_i \lambda_{K_i} \tag{4.9}$$

which is a sum of the total per-strain spacer-acquisition rates:

$$\lambda_{B_i} = \mu_b \delta D_i \sum_j P_j \tag{4.10}$$

total per-strain PAM mutation rates:

$$\lambda_{P_i} = \mu_p \beta_i \delta P_i \left(U + \sum_j (1 - M(X_i, Y_j)) D_i \right) \tag{4.11}$$

and total per-strain PAM back mutation rates:

$$\lambda_{Q_i} = \mu_q \beta_i \delta P_i \left(U + \sum_j (1 - M(X_i, Y_j)) D_i \right) . \tag{4.12}$$

In this way each unique CRISPR genotype (X_i), defined as a set of linked spacers sharing the same array, is modeled individually, as is each phage genotype (Y_i). As new spacers are added and new PAMs undergo mutation, new pairs of genotypes and equations are added to the system. Host that have undergone immune loss are modeled separately (U), as if they have no CRISPR-Cas system.

The function $M(X_i, Y_j)$ is a binary matching function between (proto)spacer content of bacterial and phage genomes that determines the presence or absence of immunity. We refer to the "order" of a host or phage strain, which is the number of evolutionary events that strain has undergone, $|X_i|$ or $n_s - |Y_i|$ respectively. The PAM back mutation rate μ_q describes the rate at which we expect a mutated PAM to revert to its original sequence (assuming the mutation is a substitution). While back mutation is not required to generate stable host-dominated coexistence, it greatly expands the relevant region of parameter space because it allows phages to avoid the cost we will impose on PAM mutations, discussed below, when those immune escape mutations are no longer beneficial. Recombination among viral strains could have a similar effect by providing another route to an un-mutated or less mutated genome.

Páez-Espino [150] suggest that recombination can produce stable host-dominated coexistence, although we reject such diversity-driven hypotheses [e.g. 214] based on our sequencing data.

We assume that the number of PAM mutations in a single phage genome is constrained by a tradeoff with phage fitness, as this is necessary to prevent the total clearance of protospacers from a single strain at high mutation rates. Increases in host breadth at the species level generally come at a cost for viruses due to pleiotropic effects [221]. More broadly, mutations tend to be deleterious on average [e.g. 222]. It is reasonable to speculate that phages have evolved under pressure to lose any active PAMs on their genomes, and thus that the persisting PAMs may have been preserved because their loss is associated with a fitness cost.

The function

$$\beta_i = -\frac{c\beta_{\text{base}}}{n_s}|Y_i| + \beta_{\text{base}} \tag{4.13}$$

incorporates a linear cost of mutation into the phage burst size. See Table 4.3 for further definitions of variables, functions, and parameters in Equations 4.5-4.13. Simulation procedures and rationale for parameter values, including phage genome size, are detailed in Appendix C.3.

Symbol	**Definition**	**Value**
R	Resource concentration	350 μg/mL
B_i	Population size of CRISPR$^+$ bacterial strain i	10^6
P_i	Population size of phage strain i	10^6
B_u	Population size of CRISPR$^-$ bacteria	10^2
λ_{B_i}	Mutation rate of bacterial strain i	n/a
λ_{P_i}	PAM mutation rate of phage strain i	n/a
λ_{Q_i}	PAM back mutation rate of phage strain i	n/a
λ	Total rate of mutation events occurring in model	n/a
$M(X_i, Y_j)$	Matching function between spacer set of bacterial strain i and protospacer set of phage strain j	no matches initially
$\beta(\|Y_i\|)$	Burst size as a function of the order of phage strain i	$\beta(0) = 80$
$\|X_i\|$	Order of bacterial strain i	0
$\|Y_i\|$	Order of phage strain i	0
e	Resource consumption rate of growing bacteria	5×10^{-7} μg
v	Maximum bacterial growth rate	1.4/hr
z	Resource concentration for half-maximal growth	1 μg
δ	Adsorption rate	10^{-8} mL per cell per phage per hr
β_{base}	Maximum burst size	80 particles per infected cell
n_s	Size of phage genome	10 protospacers
c	Cost weight per PAM mutation	3
μ_L	Per individual per generation CRISPR inactivation/loss rate	5×10^{-4}
α	Rate of autoimmunity	$50\mu_b$ deaths per individual per hr
μ_b	Spacer acquisition rate	5×10^{-7} acquisitions per individual per hr
μ_p	Per-protospacer PAM mutation rate	5×10^{-8} mutations per spacer per individual per hr
μ_q	PAM back mutation rate	5×10^{-9} mutations per spacer per individual per hr

Table 4.3: Definitions and oft used values/initial values of variables, functions, and parameters for the simulation model

4.4.2.1 Stable Host-Dominated Coexistence

Simulations with immune loss reliably produce extended coexistence within a realistic region of the parameter space (Fig 4.3) thus replicating our experimental results (Fig 4.2), and confirming our qualitative results from the simpler deterministic model (Fig 4.1). We observed no simulations in which autoimmunity alone produced stable coexistence. This agrees with our earlier numerical results from the general model where unrealistically high rates of autoimmunity were required to produce coexistence.

Similar to our experimental results, for a single set of parameters this model can stochastically fall into either stable coexistence or a phage-free state (Fig 4.3). The relative frequencies with which we see each outcome, as well as the distribution of times that phages are able to persist, depend on the specific set of parameters chosen. In particular, increasing the PAM back mutation rate will increase the probability of the coexistence outcome (Fig 4.4), although even in the absence of back mutation the system will occasionally achieve stable coexistence. This dependence on back mutation is caused by the combined effects of the cumulative cost we impose on PAM mutations and the inability of phages to keep up with host in a continuing arms race. In the early stages of the arms race it is optimal for phages to continue undergoing PAM mutations as the most abundant available hosts are high-order CRISPR variants, whereas once hosts are able to pull sufficiently ahead of phages in the arms race it becomes optimal for phages to feed on the lower-density but consistently available CRISPR-lacking host population (Fig C.13).

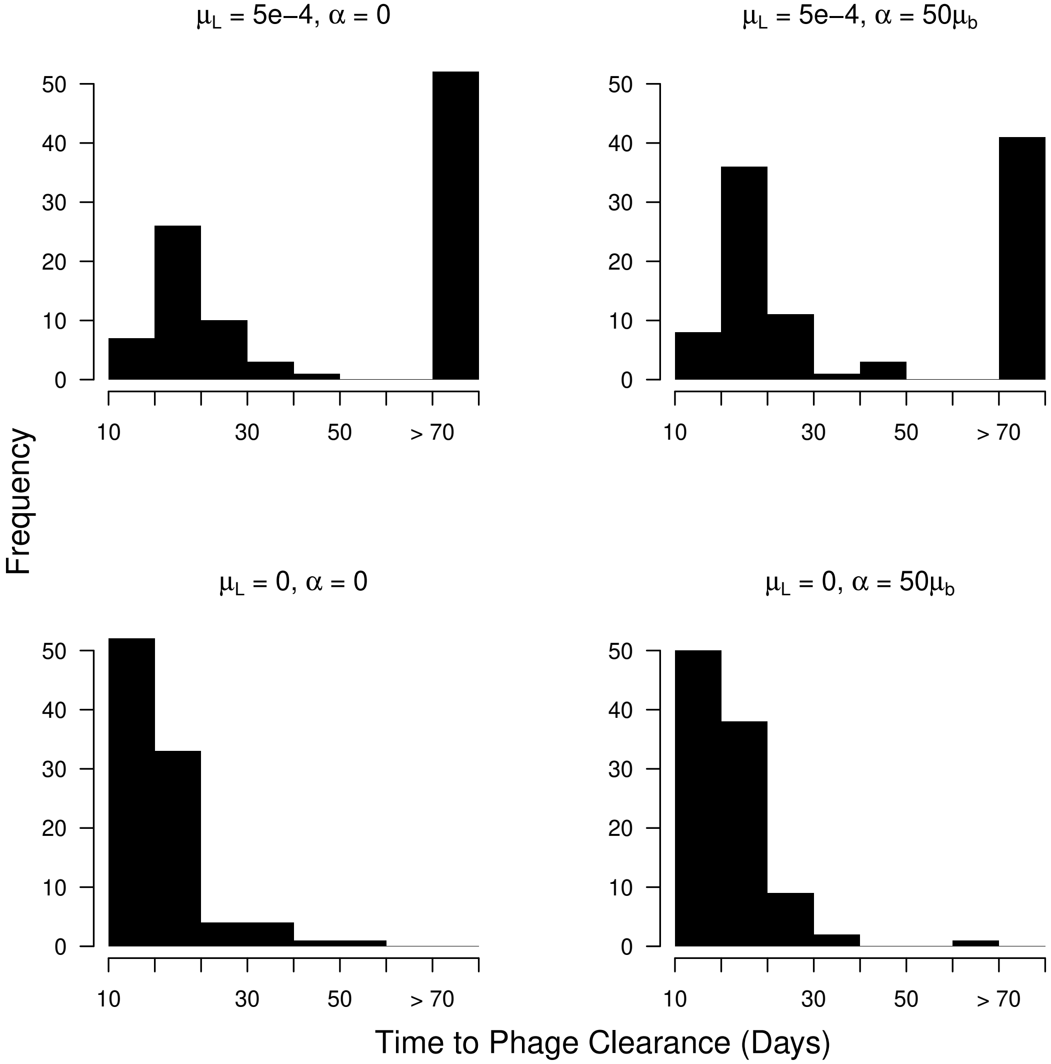


Figure 4.3: Distribution of phage extinction times in bacterial-dominated cultures with different possible combinations of coexistence mechanisms. The peak at ≥ 75 corresponds to what we call stable coexistence (simulations ran for a maximum of 80 days). There is no significant difference between the top two panels in the number of simulations reaching the 80 day mark ($\chi^2 = 2.8904$, $df = 1$, $p-$value $= 0.08911$). Back mutation was set at $\mu_q = 5 \times 10^{-9}$.

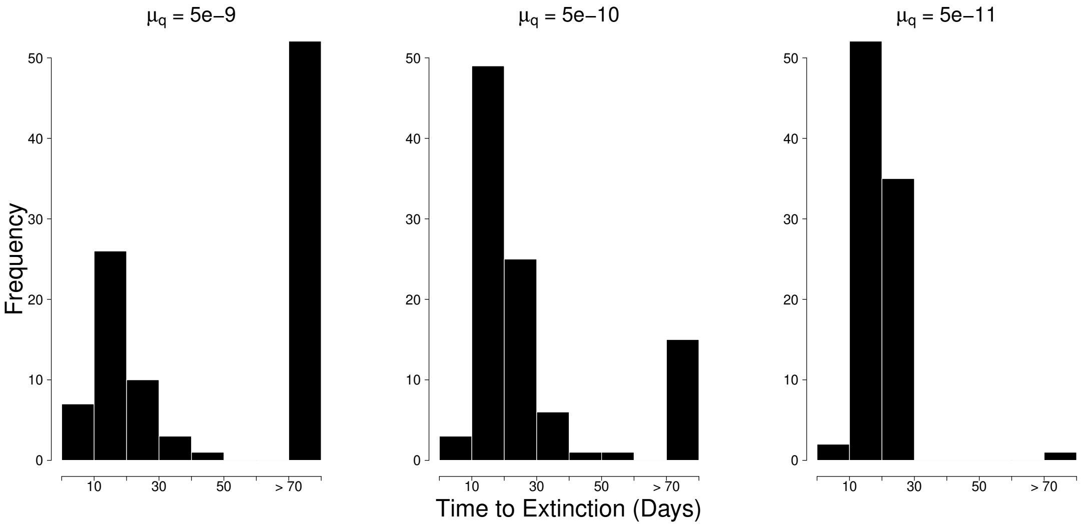


Figure 4.4: Distribution of phage extinction times in bacterial-dominated cultures with different rates of PAM back mutation in phages (μ_q). The peak at 80 corresponds to what we call stable coexistence (simulations ran for a maximum of 80 days). These results are shown for a locus-loss mechanism only ($\mu_L = 5 \times 10^{-4}$, $\alpha = 0$). The histogram for $\mu_q = 5 \times 10^{-8}$ is omitted as it is nearly identical to that for $\mu_q = 5 \times 10^{-9}$, indicating that the height of the coexistence peak saturates at high back mutation.

The adsorption rate, on a coarse scale, has an important effect on how the model behaves (Fig C.14). At high values of δ where we would expect phages to cause host extinction in the absence of CRISPR immunity ($\delta = 10^{-7}$) we see that long-term coexistence occurs rarely, and is negatively associated with the phage back mutation rate. In this case phages will rapidly consume the susceptible host population and crash to extinction unless they have undergone PAM mutations that lower their growth rate. This causes a reversal in the previous trend seen with back mutation where the ability of phages to escape the costs of PAM mutation was essential to their persistence. A decrease in the adsorption rate to a very low value ($\delta = 10^{-9}$) leads to most simulations persisting in host-dominated coexistence until the 80 day cutoff. Because both evolutionary and demographic dynamics occur much more slowly in this case, long term persistence does not necessarily imply actual stability, as suggested by our and previous [150] experimental results in which coexistence eventually ends. In general, lower adsorption rates lead to longer periods of host-dominated coexistence and reduce the chance of phage extinction.

The failure of autoimmunity to produce coexistence warrants further investigation. Upon closer examination, it is clear that in the early stages of the arms race where CRISPR-enabled bacteria have not yet obtained spacers or been selected for in the host population, phages are able to proliferate to extremely high levels and greatly suppress the CRISPR-lacking host. Because autoimmunity as a mechanism of coexistence relies on the continued presence of immune-lacking host, it may not be able to function in the face of this early phage burst if susceptible host are driven extinct. There is a possibility that very low locus loss rates that reintroduce CRISPR-lacking bacteria but do not appreciably contribute to their density combined with high rates of autoimmunity could maintain high enough density susceptible host populations to sustain phage. To investigate this possibility we imposed a floor of $U > 1$ and ran further simulations. Even with very high rates of autoimmunity based on an upper limit of likely spacer acquisition rates ($\alpha = 50\mu_b$, $\mu_b = 10^{-5}$) the susceptible host population does not grow quickly enough to sufficiently high levels to sustain phage (Fig C.15). Thus it is not early dynamics that rule out autoimmunity but the insufficiency of the mechanism itself for maintaining large enough susceptible host populations.

4.4.2.2 Transient Coexistence with Low Density Phage

While we do not observe stable coexistence in any case where there is not loss of the CRISPR-Cas immune system, we did observe prolonged phage persistence in some cases where $\mu_L = \alpha = 0$ (Fig 4.3) and in cases with autoimmunity only ($\mu_L = 0$). Phages were able to persist at very low density ($\sim 10-100$ particles/mL) for as long as two months in a host-dominated setting without the presence of a CRISPR-lacking host subpopulation (Fig 4.3, Fig C.16). It appears that in these cases phages are at sufficiently low density as to have a minimal effect on their host population and thus that host strain is selected against very slowly. Because the phages have undergone many PAM mutations at this point they are unable to proliferate rapidly enough between dilution events to have an easily measurable impact on the host population. Essentially, phages delay their collapse by consuming their host extremely slowly (Fig C.16). However, with an active locus loss mechanism (i.e., $\mu_L > 0$), we did not see this sustained but unstable coexistence occur, likely because the undefended hosts would have driven the phage population to higher levels and increased selection on the susceptible CRISPR variants.

4.5 Discussion

We paired a general model of immunity with a case study of the CRISPR immune system to characterize and contrast the potential drivers of long-term host-phage coexistence in well-mixed systems. We found that a tradeoff mechanism does not lead to a robust coexistence equilibirum in the case of intracellular host immunity. We also ruled out coevolutionary drivers of coexistence in the *S. thermophilus*-phage 2972 system based on a combination of our own sequencing data and previous work on the same system [26]. Since some mechanism(s) must be producing susceptible hosts on which phages can replicate, we are left with an immune loss hypothesis as the best remaining explanation for our empirical results. Our simulations showed that the addition of early coevolutionary dynamics alongside immune loss replicates key features of our experimental results, including stochastic switching between the possible outcomes of long term coexistence and rapid phage clearance. Therefore we predict that that the CRISPR-Cas immune system is lost at a nontrivial rate in *S. thermophilus* in addition to *S. epidermidis* [51], and possibly other species.

With regards to CRISPR, while our experiments do not speak to the relative importance of locus loss versus costly autoimmunity, our theoretical results reject autoimmunity as a realistic mechanism of phage persistence. Our experimental setup was in serial dilution, which subjects the culture to large daily perturbations, ruling out any mechanism that does not produce a robust coexistence regime.

We emphasize that CRISPR immunity, and immunity in general, is still likely costly [62]. Nevertheless, in cases of intracellular host immunity those costs are insufficient to drive continued phage persistence in the environment. Intracellular immunity destroys phages rather than simply preventing phage replication. Thus the threshold density of susceptible host for phage persistence is higher than in systems where hosts have an extracellular defense strategy (i.e. receptor/envelope modification), meaning the cost of immunity must be higher. When hosts escape phage predation via receptor modifications, a growth-resistance tradeoff may lead to coexistence.

Our sequencing results in the *S. thermophilus* system reject coevolutionary mechanisms for coexistence. We can directly reject an arms race dynamic since it predicts the rapid, continued accumulation of spacers, which does not occur in our data. A fluctuating selection dynamic makes the more subtle prediction that the frequencies of spacers in the population should cycle over time, with different spacers dominating at different times. Even with relatively small sample sizes (<10 CRISPR loci sequenced per timepoint), we see a small cohort of spacers increase in frequency early in the experiment and continue to be detected at later timepoints (Table 4.2). These results are consistent with those of Paez-Espino et al. [26] who performed deep sequencing with the same phage-host system and observed dominant spacers that drifted in frequency over time. This continued detection and dominance of particular spacers rules out strong fitness differences between these spacers, which, in turn, contradicts the expectation of fluctuating selection that fitnesses change over time. Our stochastic simulations agree, with coevolutionary dynamics in the absence of loss or cost most often yielding rapid phage extinction and only occasionally showing coexistence for over a month – but never exhibiting sustained coexistence (Fig 4.3).

A similar model by Childs et al. [217] found that a fluctuating selection dynamic could lead to long term coexistence in a CRISPR-phage system when arrays were “saturated”, in the sense that they were filled to some preset maximum capacity with spacers, which we do not observe in our experimental data. The fact that we see little expansion of the array suggests that hosts are completely immune to phages, as rapid phage genome degradation inside the CRISPR-immune cells can prevent further uptake of spacers [223].

While we conclude that immune loss plays a key role in our system, it is not immediately clear why bacterial immune systems would lose functionality at such a high rate. Our sequencing of the *S. thermophilus* CRISPR loci did not reveal pervasive spacer loss events, indicating that immune loss is at the system rather than spacer level. Perhaps in the case of CRISPR there is some inherent instability of the locus, leading to high rates of horizontal transfer [50, 143, 218, 224, 225]. Jiang et al. [51] propose that CRISPR loss is a bet-hedging strategy that allows horizontal gene transfer to occur in stressful environments (e.g., under selection for antibiotic resistance). This proposal is consistent with evidence that CRISPR does not inhibit horizontal gene transfer on evolutionary timescales [226]. A high rate of CRISPR loss and inactivation could produce pressure for bacteria to frequently acquire new CRISPR-Cas systems through horizontal gene transfer, perhaps explaining why strains with multiple CRISPR-Cas systems are frequently observed, including *S. thermophilus* [56, 58]. This is consistent with a broader view in which prokaryotic defense systems appear to be labile, having higher rates of gain and loss than other genetic content [16].

While some clear anecdotes of immune loss exist [51, 205], other examples of this phenomenon may have been missed because it is difficult to detect. Phages will quickly destroy any evidence of loss, and loss rates can be low while still affecting population dynamics. Jiang et al. [51] go to great lengths to demonstrate loss in their system. Particularly with complex systems like CRISPR-Cas, a mutation in any number of components can lead to inactivation, making loss hard to detect from genetic screens. Phenotypic screens like those of Jiang et al. [51] require the engineering of CRISPR spacer content and/or plasmid sequence as well as an otherwise competent host.

Other paths to sustained coexistence between CRISPR-enabled hosts and phages may also exist. There is a great diversity of CRISPR-Cas system types and modes of action [30] and the particular mechanism of each system may lead to distinct host-phage dynamics. That being said, our model of CRISPR evolutionary dynamics is rather general, and we recovered similar qualitative results over a wide range of parameter values apart from the *S. thermophilus*-specific parameter space.

Finally, our results show that the regular loss of immunity can sustain a viable phage population, leading to the maintenance of selective pressure and thus keeping immunity prevalent in the population overall. Even though long-term coexistence with phages may not affect overall host population density, we suggest that, counterintuitively, the periodic loss of individual immunity may drive the maintenance of a high population immune prevalence.

Appendix A: Supplemental Information For: "Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy"

A.1 Outline of Analyses

Visualizations

- CRISPR Incidence
  - PCA (Figs 2.1 and A.9, Table 2.1)
  - *t*-SNE (Figs 2.2, A.10, and A.4)
- CRISPR Type
  - PCA (Fig 2.5)
- RM Incidence
  - PCA (Fig A.20)
- Ku Incidence
  - PCA (Fig A.17)

Predictive Models (Proteobacteria Test Set)

Comparison of all predictive models in Figs A.7 and A.8

- CRISPR Incidence (Table 2.2, Appendix A.4, Figs A.7 and A.8)
  - Logistic Regression with Forward Subset Selection and Random CV (Table A.1)
  - Logistic Regression with Forward Subset Selection and Blocked CV (Table A.1)
  - Phylogenetic Logistic Regression with Forward Subset Selection and Random CV (Table A.1)
  - Phylogenetic Logistic Regression with Forward Subset Selection and Blocked CV (Table A.1)
  - sPLS-DA (Fig A.11)
  - MINT sPLS-DA (Fig A.12)
  - Random Forest (Figs 2.3 and A.14)
  - Random Forest Ensemble (A.13)
  - Random Forest, no genetic information (Appendix A.2, Fig A.16)
- CRISPR Incidence With only Temperature and Oxygen as Predictors to Train Model
  - Random Forest

- Number of CRISPR Systems
  - Random Forest (regression; Appendix A.6)
- Type II CRISPR Incidence (*cas9*)
  - Random Forest (Fig 2.5)
- RM Incidence
  - Random Forest (Fig A.21)
- Ku Incidence
  - Random Forest (Fig A.18)

Phylogenetic Regressions

- CRISPR vs 16s rRNA count (also on bootstrapped trees; Table A.2)
- CRISPR vs Ku and Oxygen (also on bootstrapped trees, oxygen use from NCBI metadata; Appendix A.7, Table A.4)
- Number of Restriction Enzymes vs Temperature (also on bootstrapped trees; Table A.3)
- Number of Restriction Enzymes vs Oxygen (also on bootstrapped trees; Table A.3)

Metagenomic Data

- *cas1,2,3,9,10* Coverage vs Dissolved Oxygen (Appendix A.5, Fig A.23)

Other

- CRISPR vs Temperature and Oxygen (NCBI metadata)
  - Binomial confidence intervals (Figs 2.4 and A.15)
  - χ^2 test
- Correlation between CRISPR and Ku
- Resampling genomes for CRISPR incidence (Appendix A.3, Fig A.22)

A.2 ProTraits Without Genomic Data

The ProTraits database, from which we take our trait data, combines various "sources" of text-based and genomic information to make trait predictions [75]. While the inclusion of genomic sources of information considerably improves the trait confidence scores, some of these sources explicitly consider gene presence/absence, and we worried it may lead to circularity in our arguments (e.g. if *cas* gene presence were used to predict a trait, which was then used to predict CRISPR incidence). Therefore we repeated our predictive analyses excluding the "phyletic profile" and "gene neighborhood" sources in ProTraits. We took the maximum confidence scores for having and lacking a trait respectively across all other sources in the database to produce a negative and positive trait score. We integrated these into a single score as described in Methods. We then built an RF model of CRISPR incidence, as this was the highest performing model on the complete dataset. This model had comparable predictive ability ($\kappa = 0.243$). We also found similar predictors to when the full dataset was used (Fig A.16). A notable change is that termite host and PAH degradation no longer appear as important predictors in the model.

A.3 Resampling Genomes

For our main analysis we sampled one genome from RefSeq per species to assess CRISPR incidence, with a preference for completely assembled reference and representative genomes. Often, CRISPR is lost by members of a species [51], and incidence can vary among strains [227]. Therefore, we attempted to determine the potential effects of our sampling process. In general, it is better to sample single genomes to assess CRISPR incidence rather than averaging across all genomes for a given species, since species are unevenly represented in RefSeq and thus the variances in incidence between species will not be equal. A drawback of sampling is that it throws away information, although strong trends should still be apparent if species have consistent tendencies to either possess or lack CRISPR. In fact, for 84% of the species in our trait dataset, the available genomes either all have or all lack CRISPR (no within species variability). Thus, sampling should have relatively minor effects on our outcome.

We verified this by repeatedly resampling CRISPR incidences from the set of all RefSeq genomes (previously determined in [15]). First we randomly resampled a new genome with known CRISPR incidence for each species in the dataset, then we split the data into training and test sets (using Proteobacteria again as the test set) and built an RF model as in the main text. This process was repeated 1000 times, and the resulting κ values and top predictors are reported in Fig A.22. The results of this analysis were consistent with our analysis in the main text.

A.4 Additional Models and Phylogenetic Corrections

In addition to the RF models described in the main text, we built several other models (described in Methods). Here we discuss their performance. For the logistic regression models, taking phylogeny into consideration, both via blocked cross validation (CV) ($\kappa = 0.168$) and an explicit evolutionary model of trait evolution ($\kappa = 0.188$), improved predictive ability relative to the phylogenetically-uninformed models. When combined these two corrections appeared to conflict with one another ($\kappa = 0.160$). This is to be expected based on the different ways these two approaches deal with the problem of shared evolutionary history. Blocked cross validation prevents overfitting to the underlying tree by leaving out contiguous portions of the data during the fitting process (see Methods). Phylogenetic logistic regression assumes an explicit model of trait evolution and attempts to fit that model to the data using a provided phylogeny. Because blocked CV leaves out chunks of the tree, phylogenetic logistic regression is unable to fit to those missing pieces of the tree, and thus the method's performance is reduced. In other words, blocked CV and phylogenetic logistic regression can both improve model performance when working with phylogenetically structured data, but combining these two approaches is unlikely to work well.

Moving on to our partial-least squares models (see Methods), sPLS-DA performed better than all logistic regression models, indicating that multicollinearity was likely a significant hurdle for logistic regression with subset selection (even more so than phylogenetic structure). Our cluster-based approach to phylogenetic correction (MINT) for sPLS-DA reduced overall predictive ability, but dramatically improved the true positive rate of the prediction (TPR = 0.538), at the cost of an increased false positive rate. In general there is always a tradeoff between false positive and false negative rates, but it is unclear to us why MINT sPLS-DA set its threshold for detection so low in this case. This is possibly an artefact from the differences in CRISPR incidence between our training and test sets, where MINT sPLS-DA learned to predict CRISPR presence at too low a threshold due to an overly CRISPR-enriched training set.

The RF and phylogenetically-informed RF ensemble models had nearly identical performance. We note though, that the ensemble approach gave a much more reliable estimate of predictive ability on the training set (mean $\kappa = 0.258$ predicting on excluded clusters) than the internal estimate automatically generated by the global RF model (out-of-bag estimate, $\kappa = 0.441$). In general, with phylogenetically structured data the internal error estimates generated by an RF model will be misleading, and the blocked cross-validation approach we employ is one way to correct these estimates.

A.5 CRISPR in the Tara Oceans Data

An alternative, and complementary approach to the one we took here is to directly measure the change in prevalence of a particular immune strategy (e.g. CRISPR) across environmental gradients using metagenomics. This approach has its own pitfalls and will require its own solutions. For example, in complex communities it may be difficult to link CRISPR proteins to particular genomes or organisms, meaning that it will be difficult to differentiate between changes in CRISPR prevalence due to differential gene content in the same set of organisms and changes in prevalence due to shifts in community composition. The situation is further complicated by the fact that many organisms have multiple CRISPR systems, or conversely have partial and non-functional systems, and that CRISPR and other defense systems are extremely labile, being gained and lost frequently [15, 16, 51]. This makes the metagenome assembly process significantly more difficult with respect to correctly mapping CRISPR to host. We also note that our current dataset integrates microbial traits across many scales, whereas a metagenomic approach will only link CRISPR prevalence to the coarsest scale of environmental parameters. Even considering oxygen, in many environments there is a possibility for extremely fine-grained variation that allows aerobes and anaerobes to live in close proximity (e.g. anoxic sediments in wetlands aerated by plant roots). In other words, our approach in the main text labels microbes as “is this”, whereas relating environmental gradients to metagenomic data labels microbes as “lives here”, where “here” is by necessity an average across the sample. A metagenomics approach links immune strategy to microbial traits indirectly via environment.

Nevertheless, metagenomics is an attractive alternative because it allows us to analyze strategy shifts actually occurring in the environment. While it is beyond the scope of the current study to perform extensive analyses of metagenomic datasets, we wish to provide an encouraging example to motivate future work in what we think is an exciting area. We used data from the Tara Oceans project [110], a global study of the microbial communities in earth's oceans, as our case study. The dataset consisted of a set of functional profiles provided by Tara, in which reads were mapped to particular orthologous groups (OG) using the KEGG orthologous groups database, as well as environmental metadata for each sample [110]. We identified the OGs for *cas1* and *cas2* (universal CRISPR marker genes; K15342 and K09951), *cas3* (type I marker; K070012), *cas9* (type II; K09952), and *cas10* (type III; K07016). We then normalized the coverage of each OG by total coverage in a given sample, and paired this data with the dissolved oxygen concentration for each sample.

Similar to our results based on ProTraits, we found a negative association between oxygen and CRISPR (Fig A.23). We found a significant negative correlation between oxygen and *cas1* (Pearson's product moment correlation, $\rho = -0.1757433$, $p = 0.00668$), *cas2* ($\rho = -0.2254487$, $p = 0.0004696$), *cas3* ($\rho = -0.1939399$, $p = 0.002714$), and *cas10* ($\rho = -0.4018567$, $p = 1.304 \times 10^{-10}$). The relationship between oxygen and *cas9* was not significant ($\rho = -0.03446256$, $p = 0.5976$). We note that this data doesn't strictly represent an oxygen "gradient" since dissolved oxygen content appears to be bimodal, with peaks corresponding to oxic and anoxic conditions (Fig A.23).

A.6 Number of CRISPR Arrays

Many prokaryotes have multiple CRISPR arrays, and this multiplicity is potentially maintained by selection [15]. We sought to assess whether we could predict the multiplicity of CRISPR arrays on a genome using our trait data. CRISPRDetect identifies individual arrays, so that our original dataset already included information about array multiplicity as well as incidence. We excluded all species lacking CRISPR so as not to confound the question of incidence (who has CRISPR?) with multiplicity (how many CRISPR arrays do they have?). Random forests can be used on continuous outcome variables (regression), and so we built a RF model using the same procedure as in the main text, but with multiplicity rather than incidence as the outcome variable. This model performs extremely poorly, with essentially no predictive ability (MSE $= 4.26$, $R^2 = 0.008$). The predicted and actual values on the test set were barely significantly correlated (Pearson's correlation, $\rho = 0.09$, $p = 0.048$). This is not entirely surprising, as regression is generally more difficult that classification. In other words, it is harder to predict whether an organism has one, two, or three CRISPR arrays than it is to predict if it has CRISPR at all.

A.7 NHEJ-Oxygen Model

Using our annotations for Ku and the NCBI annotations for oxygen requirement (aerobes and anaerobes only, facultative organisms excluded) we compiled a set of 1473 genomes for which both pieces of information were available. We built a phylogeny for these genomes using the method described in the main text. We then built a phylogenetically corrected linear model with CRISPR incidence as the binary outcome variable, Ku and aerobicity as binary predictors, as well as an interaction term (phylogenetic logistic regression, phylolm R package; [92, 93]). Ives and Garland [92] recommend that when categories have small sample sizes (as does our anaerobe with Ku category at 33 genomes) that p-values for phylogenetic logistic regression are obtained via bootstrapping, although this method is more computationally intensive. We performed 1000 bootstrap replicates (the 'boot' option in the phyloglm() function) to assess the statistical significance of each term in the model. We repeated this analysis with the *cas3*, *cas9*, and *cas10* genes, which are diagnostic of CRISPR system type, in order to see if any Ku-oxygen-CRISPR interaction was type-specific.

Our bootstrapped p values for both Ku and Oxygen, as well as their interaction, were all below 0.001 (all bootstrapped coefficients differed from zero in a consistent direction across all replicates). These p-values differed from the maximum likelihood estimates generated from the phylogenetic logistic regression model (notably, the interaction between Ku and oxygen was not significant using these estimates, at $p = 0.088164$, though the effects of Ku 1.53×10^{-5}, and oxygen 0.001183 remained significant), though this should not be surprising as the behavior of these estimates are not well characterized at low sample sizes and bootstrap estimates are generally the favored approach [92].

For type I and III systems, the results were generally consistent. In the case of type I systems all model terms were significant under bootstrapping (Ku $p < 0.001$, oxygen $p = 0.016$, interaction $p < 0.001$) but when using p-values from the ML estimate oxygen was not a significant predictor of *cas3* incidence (Ku $p = 4.164 \times 10^{-10}$, oxygen $p = 0.1138959$, interaction $p = 0.0004477$). The same was true for type III systems in terms of bootstrapped p-values (Ku $p < 0.001$, oxygen $p = 0.039$, interaction $p = 0.002$) and those from the ML estimates (Ku $p = 0.0002035$, oxygen $p = 0.3048236$, interaction $p = 0.0014942$). For type II systems only the effects of Ku were significant, and only in the bootstrapped case (Ku $p = 0.005$, oxygen $p = 0.052$, interaction $p = 0.0546$), not for the ML estimates (Ku $p = 0.1176$, oxygen $p = 0.2542$, interaction $p = 0.7550$).

For all of these phylogenetic regressions, results were consistent on 10 bootstrapped trees (Table A.4).

Logistic Regression	
Random CV	Blocked CV
temperaturerange_thermophilic (+)	temperaturerange_thermophilic (+)
mammalian_pathogen_oral_cavity (+)	mammalian_pathogen_oral_cavity (+)
knownhabitats_freshwater (+)	metabolism_carbondioxidefixation (+)
ecosystemtype_marine (-)	host_insectstermites (+)
pathogenic_in_mammals (-)	ecosystemtype_geologic (+)
knownhabitats_hydrothermalvent (+)	energysource_autotroph (+)
metabolism_carondioxidefixation (+)	ecosystemsubtype_vagina (+)
host_insectstermites (+)	metabolism_sulfuroxidizer (-)
shape_tailed (+)	habitat_terrestrial (-)
knownhabitats_soil (-)	
knownhabitats_creosotecontaminatedsoil (-)	
energysource_heterotroph (-)	
cellarrangement_tetrads (-)	
ecosystemsubtype_vagina (+)	
knownhabitats_insectendosymbiont (-)	
ecosystemtype_thermalsprings (+)	
habitat_hostassociated (+)	
cellarrangement_singles (-)	

Phylogenetic Logistic Regression	
Random CV	Blocked CV
knownhabitats_hotspring (+)	knownhabitats_hotspring (+)
mammalian_pathogen_oral_cavity (+)	mammalian_pathogen_oral_cavity (+)
host_insectstermites (+)	host_insectstermites (+)
shape_filamentous (+)	shape_filamentous (+)
oxygenreq_strictaero (-)	oxygenreq_strictaero (-)
ecosystemtype_reproductivesystem (+)	energysource_heterotroph (-)
mammalian_pathogen_respiratory_lundisease (-)	
ecosystemtype_marine (-)	
knownhabitats_hydrothermalvent (+)	
ecosystemcategory_plants (-)	

Table A.1: Predictors added to each logistic regression model during forward selection (top to bottom in order of addition). Plus and minus signs indicate whether a variable is positively or negatively associated with CRISPR incidence.

Outcome Variable	Bootstrap	β_{16s}	p_{16s}
CRISPR	1	0.05444265	0.0004987372
CRISPR	2	0.06871256	1.650863E-05
CRISPR	3	0.05602856	0.0003348601
CRISPR	4	-0.06244074	4.65824E-05
CRISPR	5	-0.06051718	7.066252E-05
CRISPR	6	-0.0656118	1.96243E-05
CRISPR	7	-0.06858516	9.275297E-06
CRISPR	8	-0.06523327	2.200228E-05
CRISPR	9	-0.06482068	2.414822E-05
CRISPR	10	0.06773283	1.521424E-05

Table A.2: Phylogenetic logistic regression of CRISPR incidence as predicted by 16s rRNA count on 10 bootstrapped trees.

Outcome Variable	Bootstrap	$\beta_{\text{Temperature}}$	$p_{\text{Temperature}}$
No. R Enzymes	1	1.46992099	0.04000844
No. R Enzymes	2	1.47619604	0.0395859
No. R Enzymes	3	0.05938679	0.67987639
No. R Enzymes	4	1.47619604	0.0395859
No. R Enzymes	5	1.49642946	0.03825504
No. R Enzymes	6	1.46992099	0.04000844
No. R Enzymes	7	1.46196112	0.04055124
No. R Enzymes	8	1.51134694	0.0373039
No. R Enzymes	9	0.0593766	0.67990236
No. R Enzymes	10	1.51134694	0.0373039

Outcome Variable	Bootstrap	β_{O^2}	p_{O^2}
No. R Enzymes	1	-4.5032905	4.775284E-35
No. R Enzymes	2	-4.5236085	3.343288E-35
No. R Enzymes	3	-0.9838951	1.133434E-08
No. R Enzymes	4	-4.5262046	3.194396E-35
No. R Enzymes	5	-4.540566	2.482726E-35
No. R Enzymes	6	-4.5216758	3.458622E-35
No. R Enzymes	7	-4.5195994	3.586961E-35
No. R Enzymes	8	-4.5446124	2.312513E-35
No. R Enzymes	9	-0.9838037	1.135213E-08
No. R Enzymes	10	-4.5419952	2.421222E-35

Table A.3: Phylogenetic regression of number of restriction enzymes as predicted by temperature or oxygen on 10 boostrapped trees.

Outcome Variable	Bootstrap	β_{Ku}	p_{Ku}	β_{O^2}	p_{O^2}	$\beta_{\text{Interaction}}$	$p_{\text{Interaction}}$
CRISPR	1	-0.7776371	0.001	0.66532	0.001	0.8076997	0.001
CRISPR	2	-0.7762393	0.001	0.6650369	0.001	0.7553189	0.001
CRISPR	3	-0.7543548	0.001	0.6523729	0.001	0.7743154	0.002
CRISPR	4	-0.7475297	0.001	0.6470698	0.001	0.7970853	0.001
CRISPR	5	-0.7542887	0.001	0.6499091	0.001	0.7606847	0.001
CRISPR	6	-0.7750393	0.001	0.6560871	0.001	0.7523748	0.001
CRISPR	7	-0.752069	0.001	0.6479241	0.001	0.8017921	0.001
CRISPR	8	-0.7570104	0.001	0.6461722	0.001	0.7978345	0.001
CRISPR	9	-0.7931716	0.001	0.676406	0.001	0.7285855	0.001
CRISPR	10	-0.7782145	0.001	0.6634043	0.001	0.7725446	0.002
cas3	1	-1.309984	0.001	0.3429482	0.009	1.586328	0.001
cas3	2	-1.339081	0.001	0.3289939	0.007	1.582142	0.001
cas3	3	-1.304121	0.001	0.3257619	0.017	1.581483	0.001
cas3	4	-1.311048	0.001	0.2938582	0.011	1.590015	0.001
cas3	5	-1.333989	0.001	0.3291178	0.023	1.586987	0.001
cas3	6	-1.315037	0.001	0.3253995	0.008	1.585864	0.001
cas3	7	-1.308	0.001	0.2901924	0.014	1.58649	0.001
cas3	8	-1.325072	0.001	0.3139974	0.023	1.59048	0.001
cas3	9	-1.351763	0.001	0.3449322	0.007	1.590153	0.001
cas3	10	-1.328922	0.001	0.3384614	0.014	1.584191	0.001
cas9	1	-0.6166104	0.001	0.3386759	0.05	-0.3086753	0.536
cas9	2	-0.5713407	0.011	0.4218407	0.019	-0.3035605	0.538
cas9	3	-0.6371981	0.002	0.380116	0.04	-0.3139595	0.545
cas9	4	-0.2427449	0.05	1.0055809	0.025	-0.3538297	0.451
cas9	5	-0.598588	0.005	0.3807694	0.04	-0.304255	0.556
cas9	6	-0.6344803	0.006	0.4011055	0.02	-0.3178936	0.565
cas9	7	-0.6162547	0.01	0.3864871	0.017	-0.2999418	0.555
cas9	8	-0.6261813	0.002	0.3399736	0.059	-0.3096002	0.564
cas9	9	-0.6106022	0.006	0.4121876	0.011	-0.3016357	0.552
cas9	10	-0.5917813	0.003	0.381879	0.051	-0.3111323	0.523
cas10	1	-2.78319	0.001	0.338453	0.048	3.0924286	0.001
cas10	2	-2.76847	0.001	0.3371324	0.028	3.0689795	0.001
cas10	3	-0.735233	0.002	0.2797778	0.039	0.7382052	0.028
cas10	4	-1.209831	0.001	0.3202466	0.027	1.2494469	0.014
cas10	5	-2.927175	0.001	0.3800277	0.047	3.1038217	0.001
cas10	6	-2.750308	0.001	0.3269329	0.05	3.080121	0.001
cas10	7	-2.706733	0.001	0.3376516	0.032	2.9835339	0.001
cas10	8	-2.839044	0.001	0.3164233	0.064	3.116515	0.001
cas10	9	-1.944355	0.001	0.3521944	0.026	2.244925	0.002
cas10	10	-2.891364	0.001	0.3741208	0.03	3.1046525	0.001

Table A.4: Phylogenetic logistic regression of CRISPR incidence as predicted by Ku and oxygen on 10 boostrapped trees. Bootstrapped p-values shown as discussed in Appendix A.7

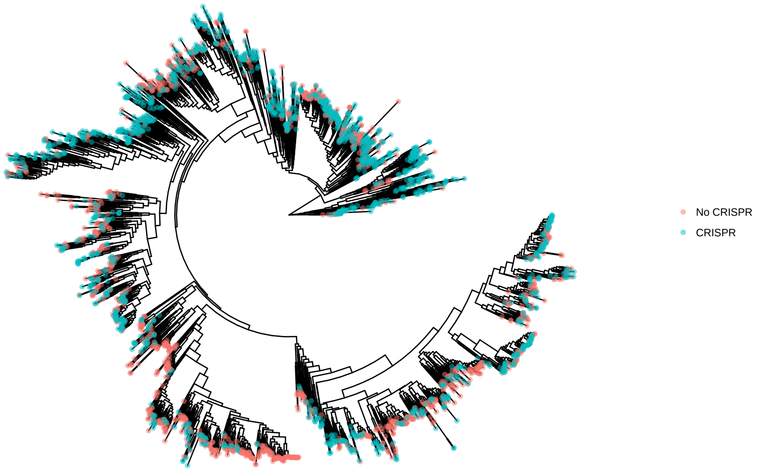


Figure A.1: Phylogeny generated from PhyloSift marker genes (as in Fig A.3). Color indicates CRISPR incidence.

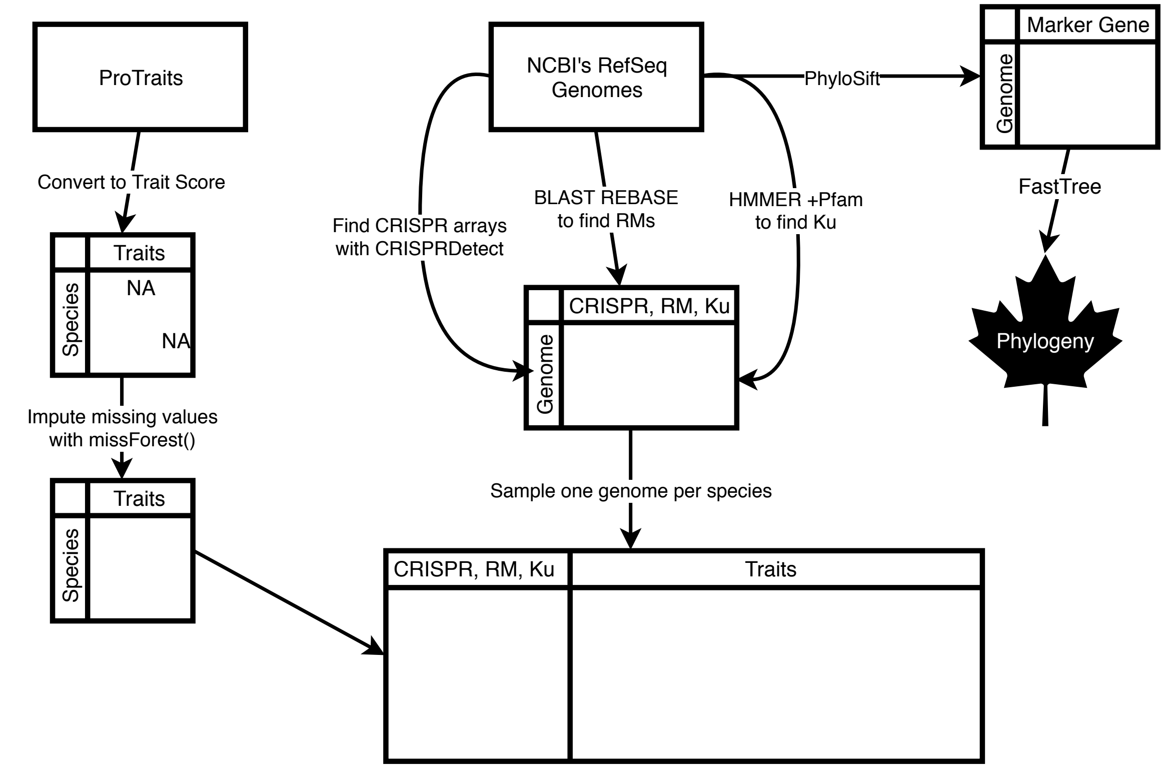


Figure A.2: Pipeline for generating trait and immunity dataset and matching phylogeny. See Methods for details.

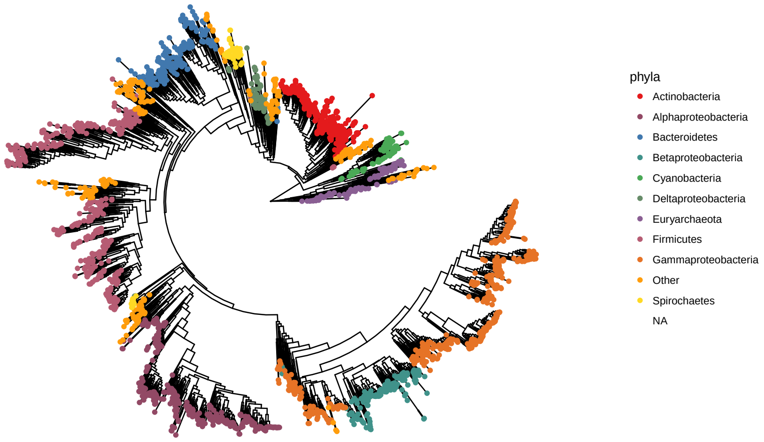


Figure A.3: Phylogeny generated from PhyloSift marker genes. Phylum indicated by color, with taxonomic classifications taken from NCBI.

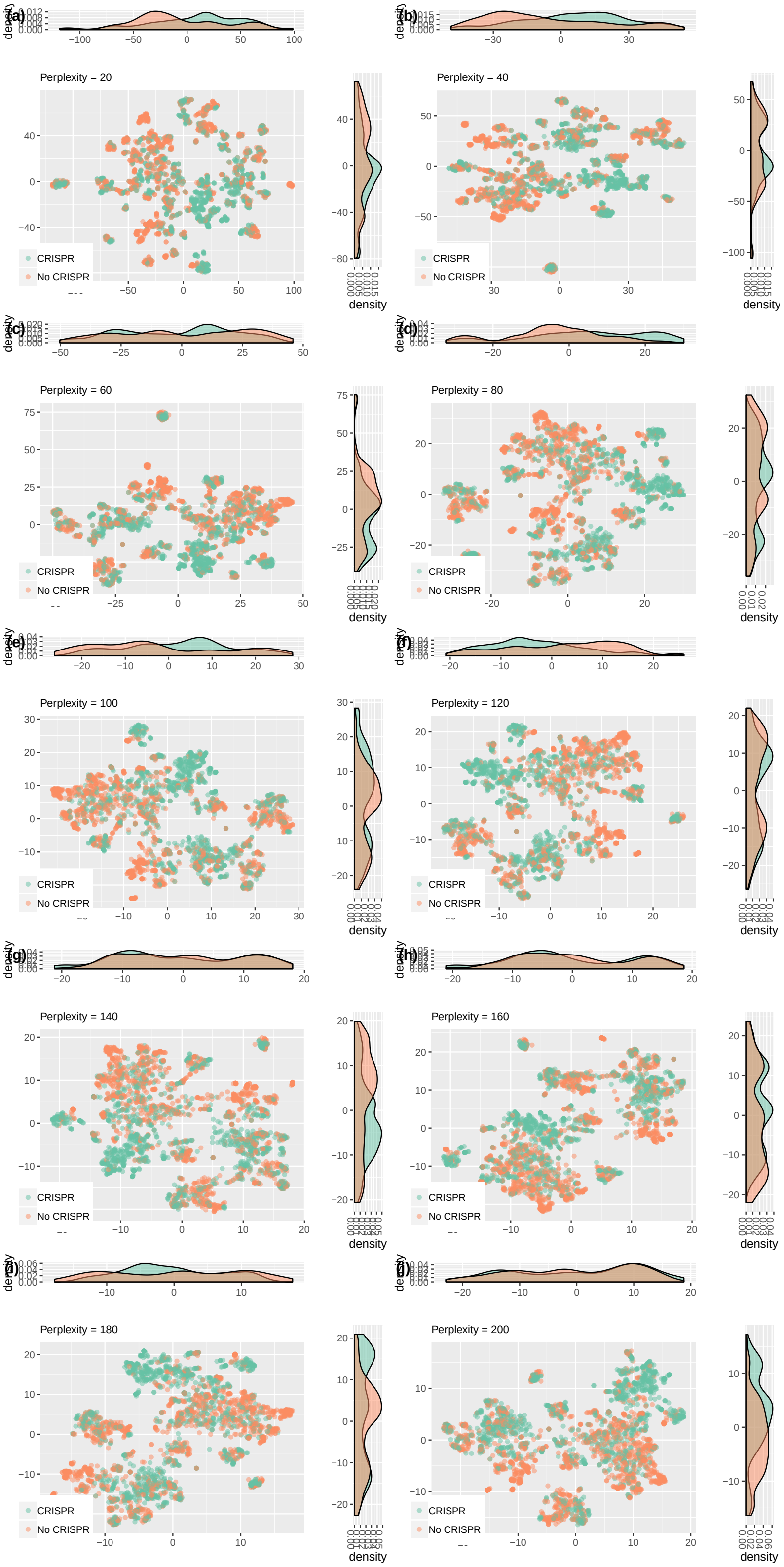


Figure A.4: Repeated t-SNE decomposition of ProTraits data with CRISPR incidence visualized for varied perplexity values. The CRISPR versus no-CRISPR separation is somewhat less apparent for very high perplexity values.

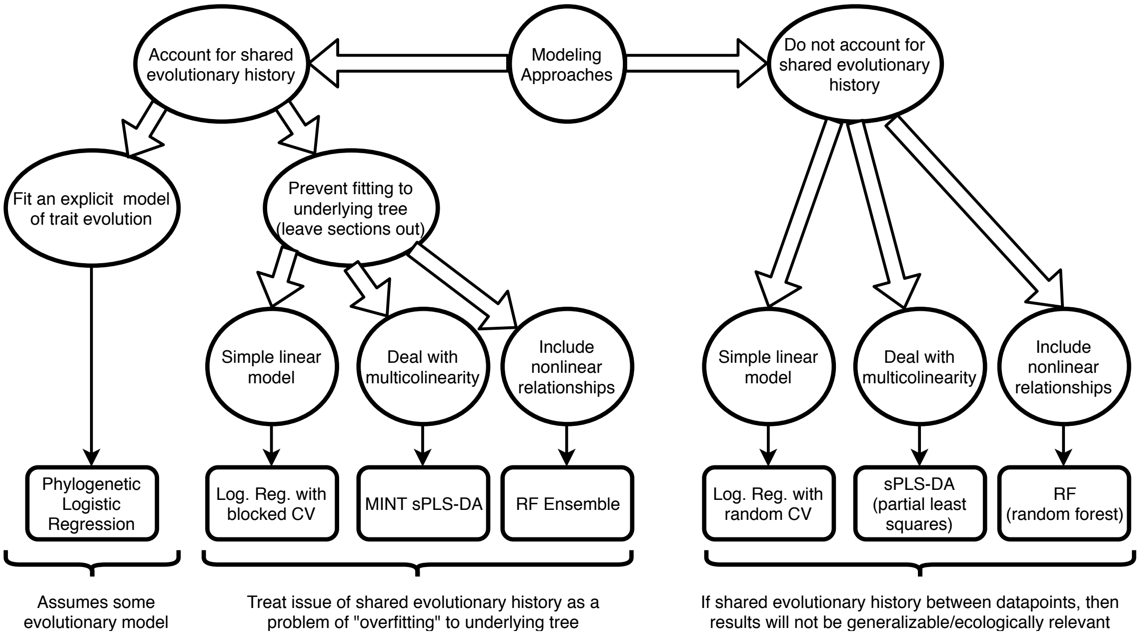
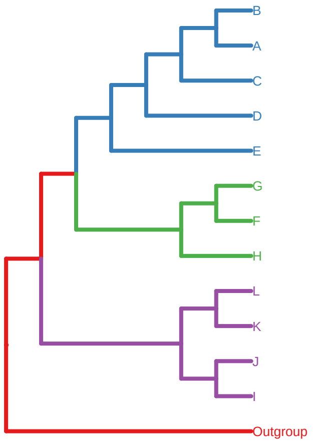

Modeling Approaches
Account for shared evolutionary history
Do not account for shared evolutionary history
Fit an explicit model of trait evolution
Prevent fitting to underlying tree (leave sections out)
Simple linear model
Deal with multicolinearity
Include nonlinear relationships
Simple linear model
Deal with multicolinearity
Include nonlinear relationships
Phylogenetic Logistic Regression
Log. Reg. with blocked CV
MINT sPLS-DA
RF Ensemble
Log. Reg. with random CV
sPLS-DA (partial least squares)
RF (random forest)
Assumes some evolutionary model
Treat issue of shared evolutionary history as a problem of "overfitting" to underlying tree
If shared evolutionary history between datapoints, then results will not be generalizable/ecologically relevant

Figure A.5: Flowchart showing the decision-making process that would lead to the various modeling approaches used here. Major considerations for each approach are noted. See Methods for details on each approach.



	Blocked Folds	Random Folds
Fold 1	A, B, C, D, E	D, H, K, L
Fold 2	F, G, H	A, B, G, I
Fold 3	I, J, K, L	C, E, F, J

Figure A.6: A conceptual example of the differences between blocked and random folds for cross validation. Cross validation (CV) relies on the assumption that folds are independent from one another, but when species share an evolutionary history this assumption is violated. By instead choosing folds based on phylogenetic groups that have diverged from each other sufficiently far in the past, we can better avoid the inclusion of phylogenetic signal in our model fit. In other words, in blocked CV we attempt to choose "evolutionarily independent" folds.

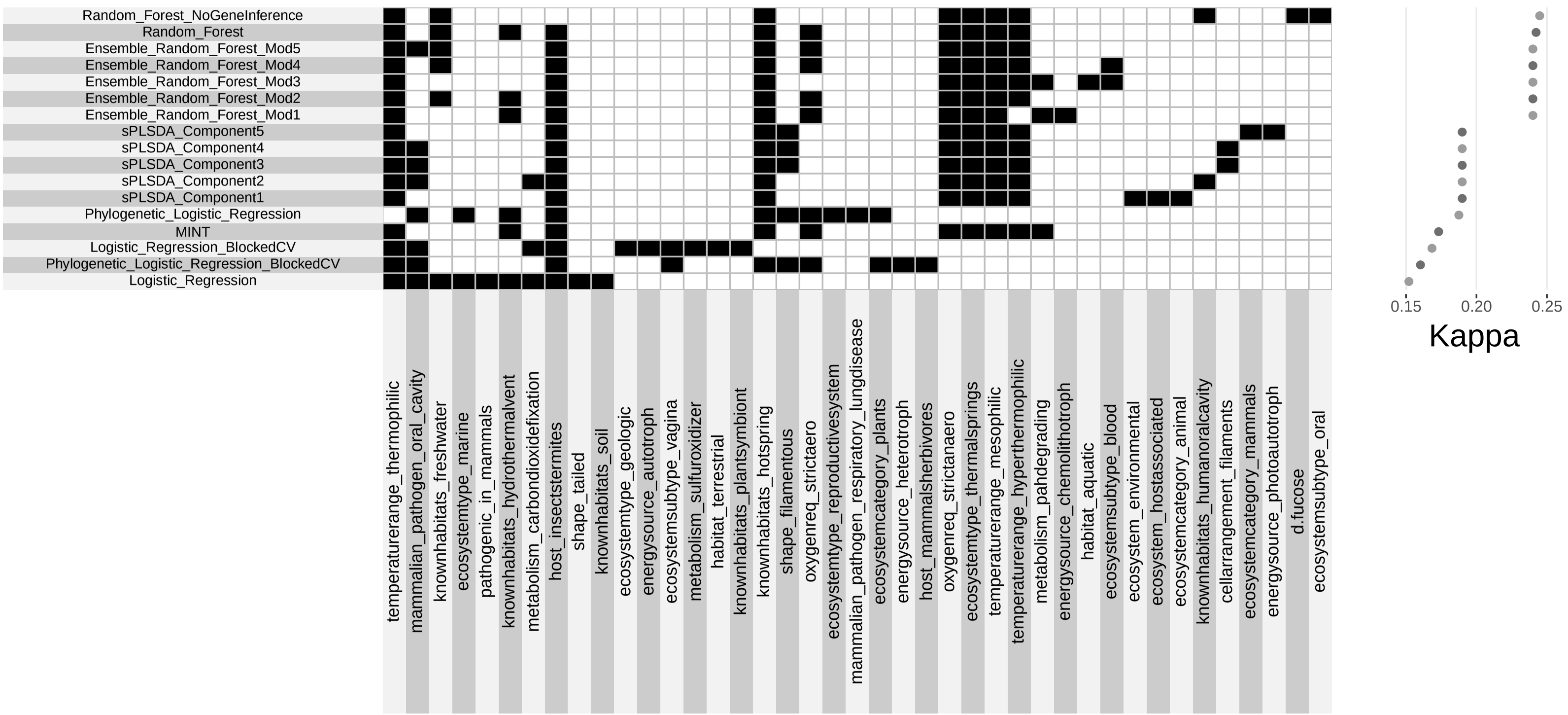


Figure A.7: Comparison of variable importance across predictive models. The top 10 most important variables (columns) for each predictive model of CRISPR incidence (rows) are indicated by filled black cells. Models are ordered top to bottom in order of decreasing performance in terms of Cohen's κ (shown at right). Note that for the high performing models, temperature variables and oxygen (anaerobe and aerobe) are consistently found in the top 10 predictors. All models incorporate temperature as an important predictor, and the only models without oxygen as a top predictor are the two logistic regression models that were not formally corrected for phylogeny (and were low-performing). In general, moderate and high performing models are largely in agreement about a core set of important variables. "NoGeneInference" corresponds to the model built in Appendix A.2.

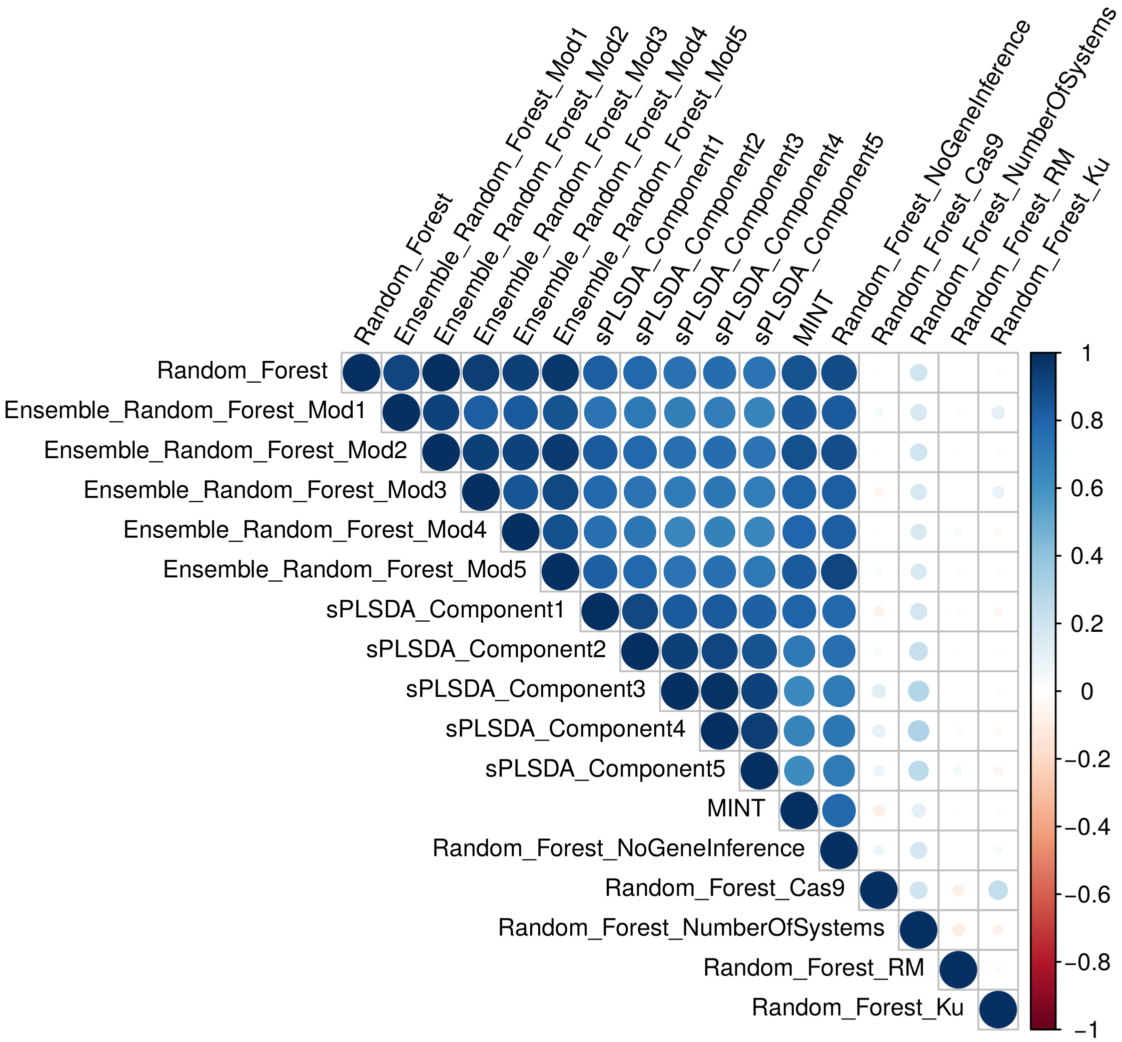


Figure A.8: Comparison of variable importance across predictive models. Pearson's correlation between variable importance scores for all predictive models (CRISPR incidence and otherwise) in the paper measured as % increase in node purity for random forests and variable importance projections for PLS models; logistic regression models were excluded because importance is measured as rank - i.e. what order the variable was added to the model. Note the high agreement between the models predicting CRISPR incidence, and some agreement with the model predicting number of CRISPR systems. Also note that models predicting the incidence of RM systems and Ku appear to have distinct predictors (these models performed well at prediction tasks in the main text). "NoGeneInference" corresponds to the model built in Appendix A.2.

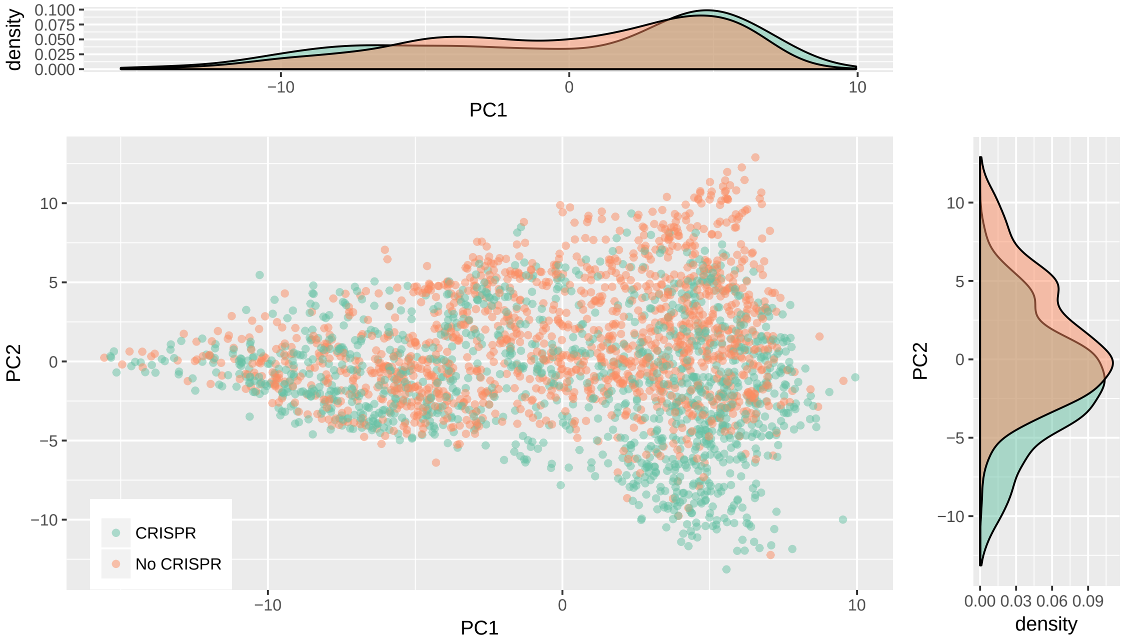
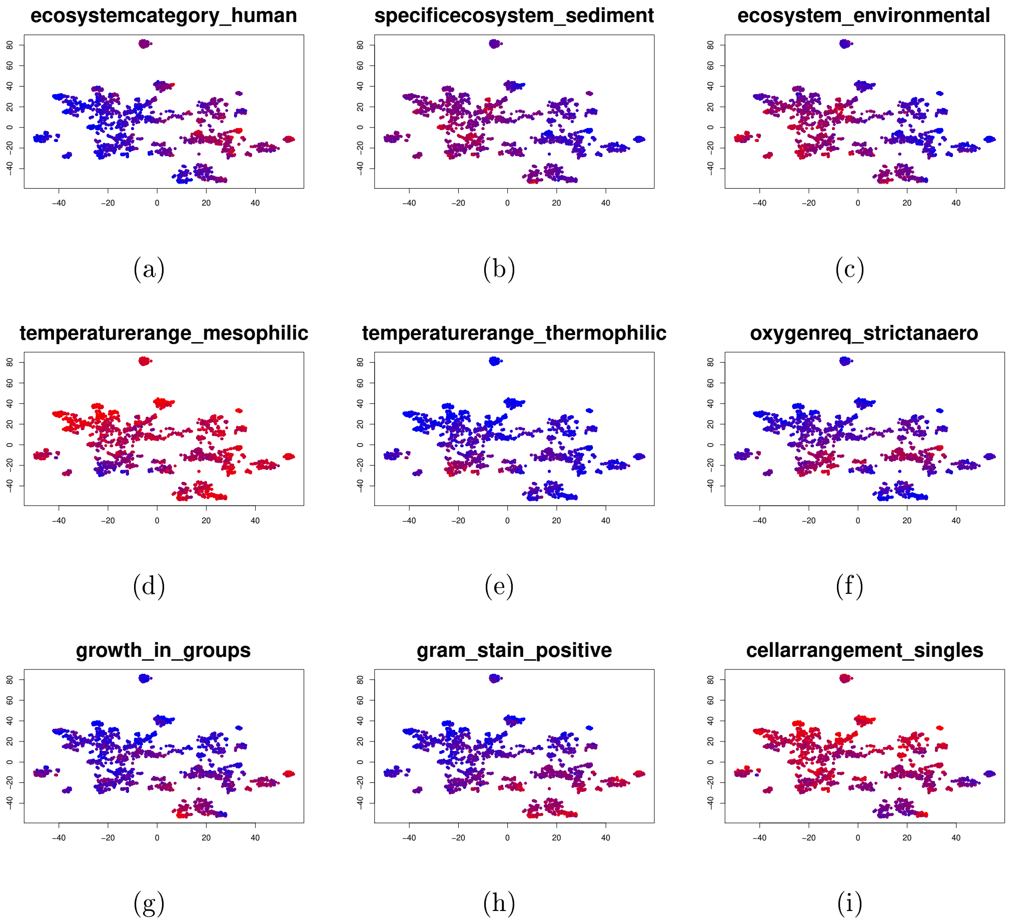


Figure A.9: Organisms with CRISPR do not separate from those without along the first principal component of trait space. The first and second components from a PCA of the microbial traits dataset are shown. CRISPR incidence is indicated by color (green with, orange without), but was not included when constructing the PCA. Marginal densities along each component are shown to facilitate interpretation. See Fig 2.1 for the third component.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure A.10: Trait distributions over t-SNE reduced dataset. Each point is an organism mapped onto our t-SNE decomposition of trait space. Instead of coloring points by presence/absence of CRISPR as shown in Fig 2.2, we color each organism by its score for selected microbial traits in our trait dataset (set of traits shown chosen because they were highly weighted in our PCA). Recall that scores range from zero (blue) to one (red). We note that, in a general sense, the region occupied by anaerobic microbes appears to correspond to the densest regions of CRISPR incidence in Fig 2.2.

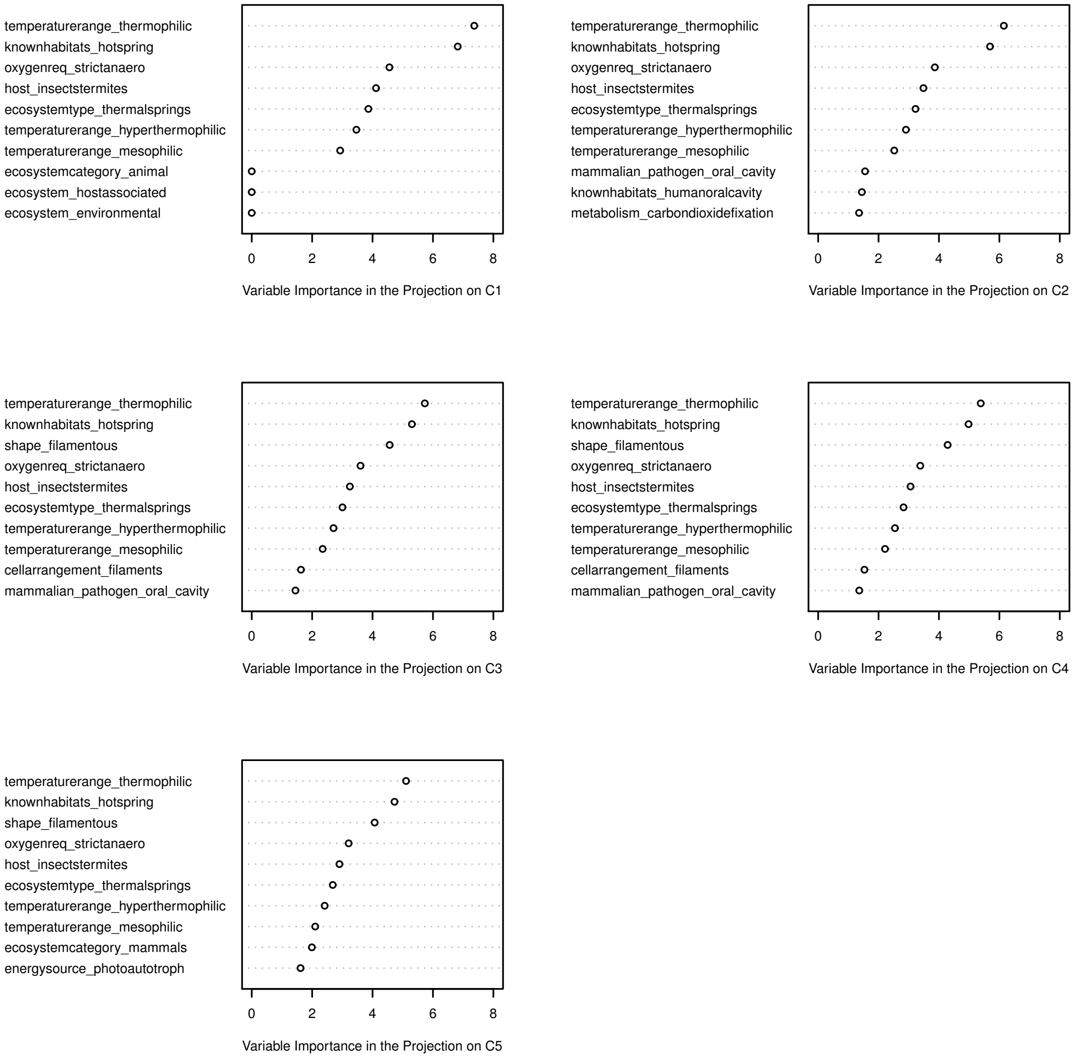


Figure A.11: Variable importance scores from sPLS-DA model for top 10 predictors on the 5 components included in model. Variable importance scores generated by the vip() function in the mixOmics package for R.

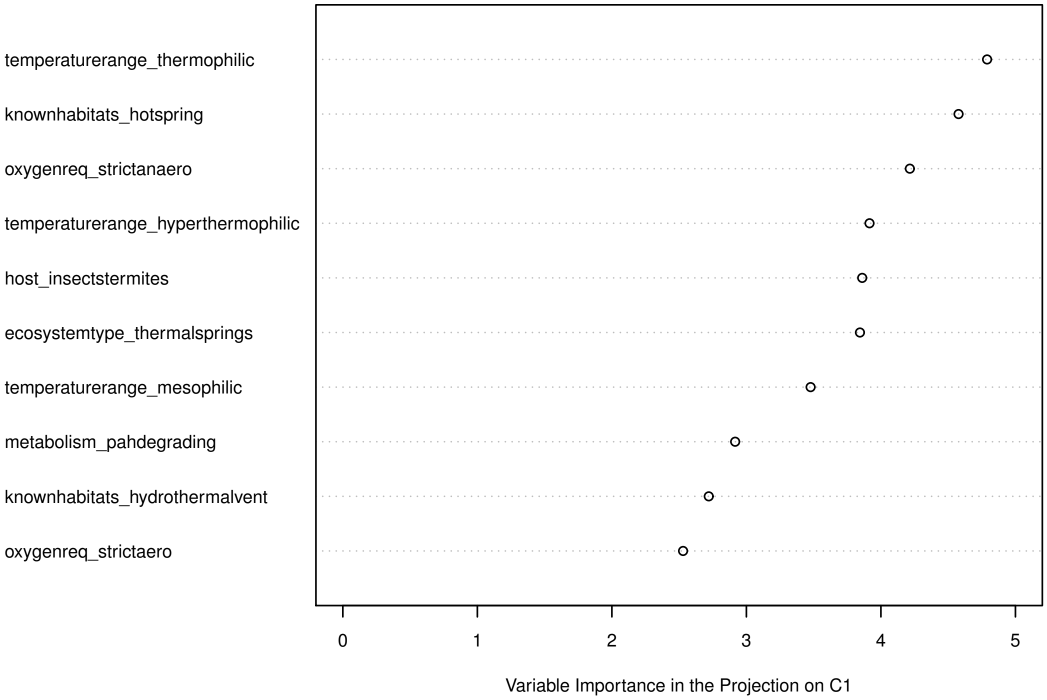


Figure A.12: Variable importance scores from MINT sPLS-DA model for top 10 predictors on the single component included in model. Variable importance scores generated by the vip() function in the mixOmics package for R.

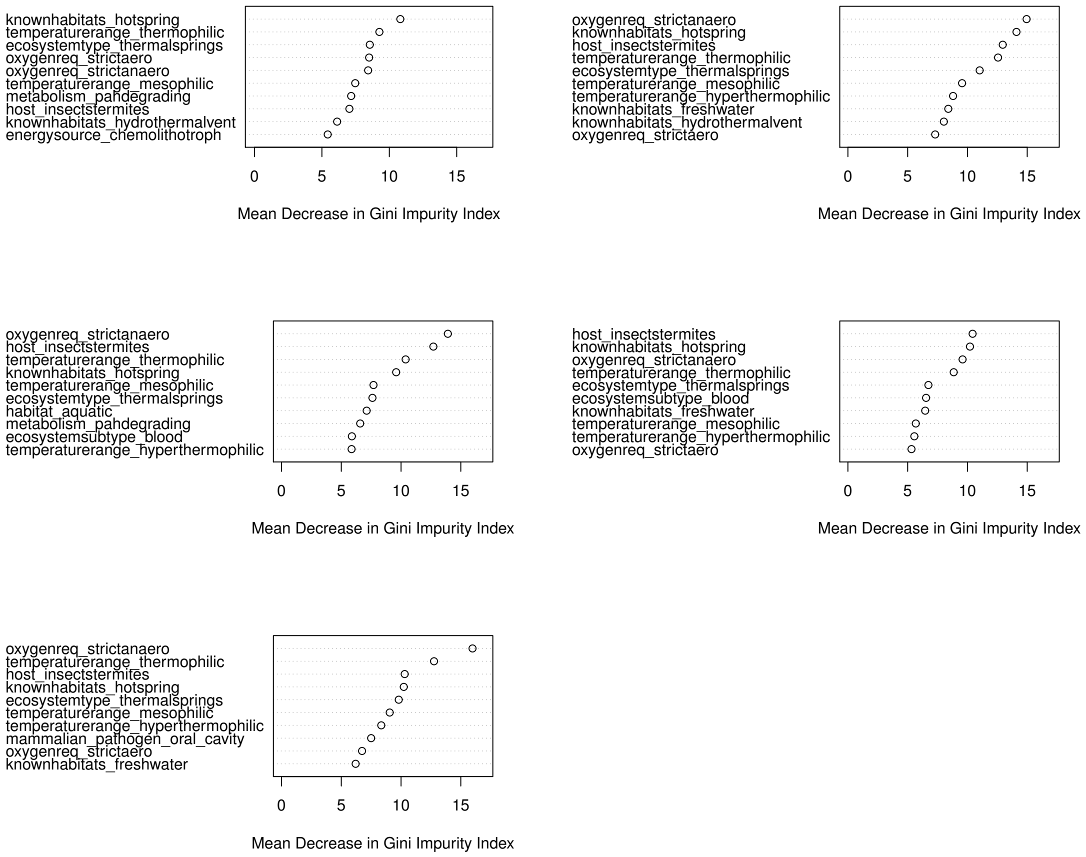
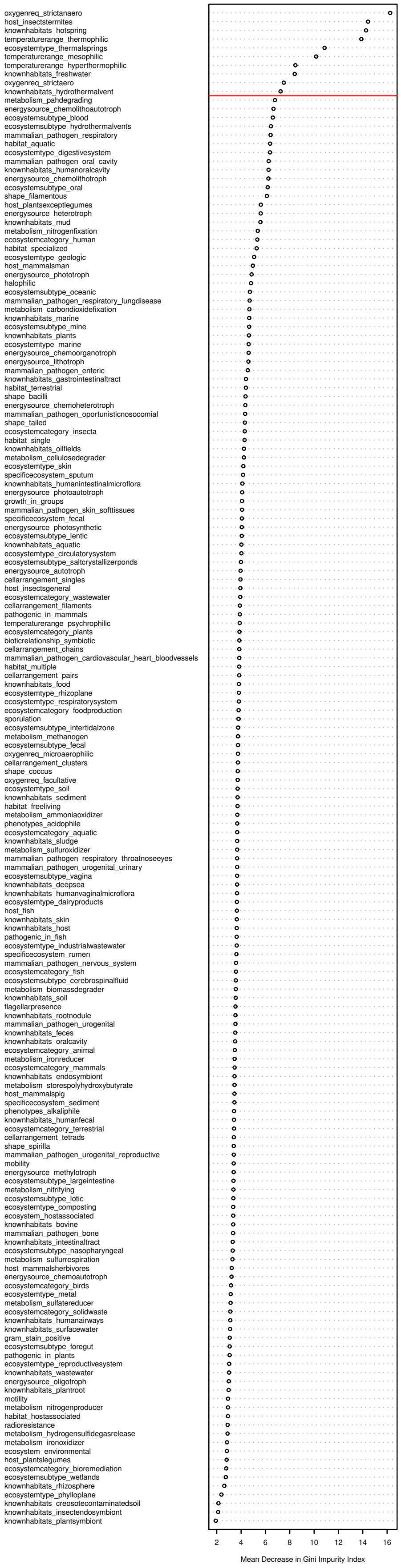


Figure A.13: Importance of top ten predictors in each of the five forests included in the RF ensemble model, as measured by the mean decrease in the Gini impurity index or accuracy when that variable is excluded from the respective forest. The relative ranking of the top ten predictors does vary somewhat over the five forests, but the set of top predictors is largely consistent across the forests.


oxygenreq_strictanaero
host_insectstermites
knownhabitats_hotspring
temperaturerange_thermophilic
ecosystemtype_thermalsprings
temperaturerange_mesophilic
temperaturerange_hyperthermophilic
knownhabitats_freshwater
oxygenreq_strictaero
knownhabitats_hydrothermalvent
metabolism_pahdegrading
energysource_chemolithoautotroph
ecosystemsubtype_blood
ecosystemsubtype_hydrothermalvents
mammalian_pathogen_respiratory
habitat_aquatic
ecosystemtype_digestivesystem
mammalian_pathogen_oral_cavity
knownhabitats_humanoralcavity
energysource_chemolithotroph
ecosystemsubtype_oral
shape_filamentous
host_plantsexceptlegumes
energysource_heterotroph
knownhabitats_mud
metabolism_nitrogenfixation
ecosystemcategory_human
habitat_specialized
ecosystemtype_geologic
host_mammalsman
energysource_phototroph
halophilic
ecosystemsubtype_oceanic
mammalian_pathogen_respiratory_lungdisease
metabolism_carbondioxidefixation
knownhabitats_marine
ecosystemsubtype_mine
knownhabitats_plants
ecosystemtype_marine
energysource_chemoorganotroph
energysource_lithotroph
mammalian_pathogen_enteric
knownhabitats_gastrointestinaltract
habitat_terrestrial
shape_bacilli
energysource_chemoheterotroph
mammalian_pathogen_oportunisticnosocomial
shape_tailed
ecosystemcategory_insecta
habitat_single
knownhabitats_oilfields
metabolism_cellulosedegrader
ecosystemtype_skin
specificecosystem_sputum
knownhabitats_humanintestinalmicroflora
energysource_photoautotroph
growth_in_groups
mammalian_pathogen_skin_softtissues
specificecosystem_fecal
energysource_photosynthetic
ecosystemsubtype_lentic
knownhabitats_aquatic
ecosystemtype_circulatorysystem
ecosystemsubtype_saltcrystallizerponds
energysource_autotroph
cellarrangement_singles
host_insectsgeneral
ecosystemcategory_wastewater
cellarrangement_filaments
pathogenic_in_mammals
temperaturerange_psychrophilic
ecosystemcategory_plants
bioticrelationship_symbiotic
cellarrangement_chains
mammalian_pathogen_cardiovascular_heart_bloodvessels
habitat_multiple
cellarrangement_pairs
knownhabitats_food
ecosystemtype_rhizoplane
ecosystemtype_respiratorysystem
ecosystemcategory_foodproduction
sporulation
ecosystemsubtype_intertidalzone
metabolism_methanogen
ecosystemsubtype_fecal
oxygenreq_microaerophilic
cellarrangement_clusters
shape_coccus
oxygenreq_facultative
ecosystemtype_soil
knownhabitats_sediment
habitat_freeliving
metabolism_ammoniaoxidizer
phenotypes_acidophile
ecosystemcategory_aquatic
knownhabitats_sludge
metabolism_sulfuroxidizer
mammalian_pathogen_respiratory_throatnoseeyes
mammalian_pathogen_urogenital_urinary
ecosystemsubtype_vagina
knownhabitats_deepsea
knownhabitats_humanvaginalmicroflora
ecosystemtype_dairyproducts
host_fish
knownhabitats_skin
knownhabitats_host
pathogenic_in_fish
ecosystemtype_industrialwastewater
specificecosystem_rumen
mammalian_pathogen_nervous_system
ecosystemcategory_fish
ecosystemsubtype_cerebrospinalfluid
metabolism_biomassdegrader
knownhabitats_soil
flagellarpresence
knownhabitats_rootnodule
mammalian_pathogen_urogenital
knownhabitats_feces
knownhabitats_oralcavity
ecosystemcategory_animal
metabolism_ironreducer
ecosystemcategory_mammals
knownhabitats_endosymbiont
metabolism_storespolyhydroxybutyrate
host_mammalspig
specificecosystem_sediment
phenotypes_alkaliphile
knownhabitats_humanfecal
ecosystemcategory_terrestrial
cellarrangement_tetrads
shape_spirilla
mammalian_pathogen_urogenital_reproductive
mobility
energysource_methylotroph
ecosystemsubtype_largeintestine
metabolism_nitrifying
ecosystemsubtype_lotic
ecosystemtype_composting
ecosystem_hostassociated
knownhabitats_bovine
mammalian_pathogen_bone
knownhabitats_intestinaltract
ecosystemsubtype_nasopharyngeal
metabolism_sulfurrespiration
host_mammalsherbivores
energysource_chemoautotroph
ecosystemcategory_birds
ecosystemtype_metal
metabolism_sulfatereducer
ecosystemcategory_solidwaste
knownhabitats_humanairways
knownhabitats_surfacewater
gram_stain_positive
ecosystemsubtype_foregut
pathogenic_in_plants
ecosystemtype_reproductivesystem
knownhabitats_wastewater
energysource_oligotroph
knownhabitats_plantroot
motility
metabolism_nitrogenproducer
habitat_hostassociated
radioresistance
metabolism_hydrogensulfidegasrelease
metabolism_ironoxidizer
ecosystem_environmental
host_plantslegumes
ecosystemcategory_bioremediation
ecosystemsubtype_wetlands
knownhabitats_rhizosphere
ecosystemtype_phylloplane
knownhabitats_creosotecontaminatedsoil
knownhabitats_insectendosymbiont
knownhabitats_plantsymbiont
2
4
6
8
10
12
14
16
Mean Decrease in Gini Impurity Index

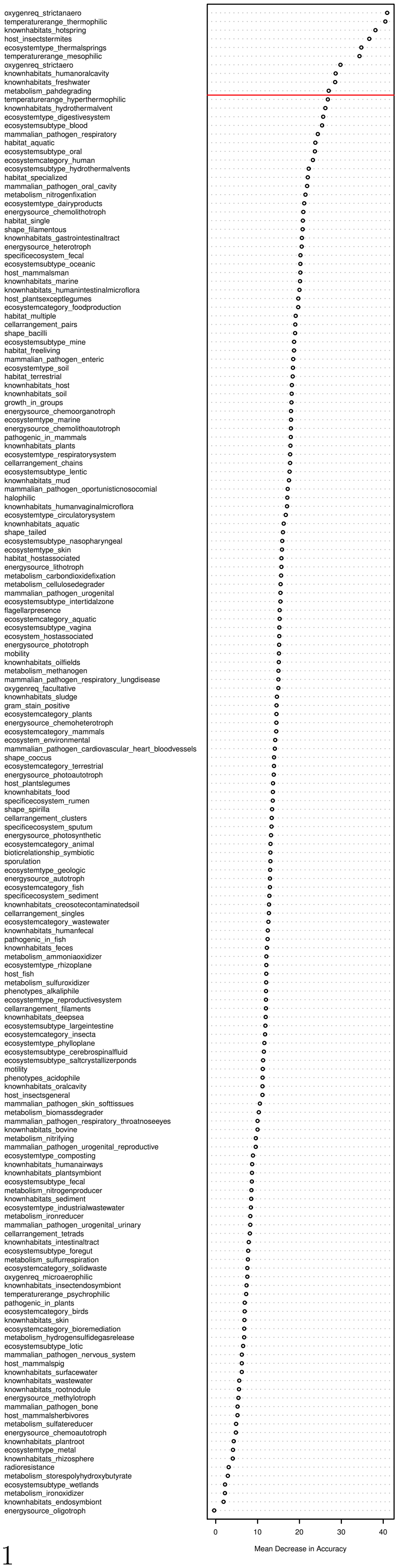

oxygenreq_strictanaero
temperaturerange_thermophilic
knownhabitats_hotspring
host_insectstermites
ecosystemtype_thermalsprings
temperaturerange_mesophilic
oxygenreq_strictaero
knownhabitats_humanoralcavity
knownhabitats_freshwater
metabolism_pahdegrading
temperaturerange_hyperthermophilic
knownhabitats_hydrothermalvent
ecosystemtype_digestivesystem
ecosystemsubtype_blood
mammalian_pathogen_respiratory
habitat_aquatic
ecosystemsubtype_oral
ecosystemcategory_human
ecosystemsubtype_hydrothermalvents
habitat_specialized
mammalian_pathogen_oral_cavity
metabolism_nitrogenfixation
ecosystemtype_dairyproducts
energysource_chemolithotroph
habitat_single
shape_filamentous
knownhabitats_gastrointestinaltract
energysource_heterotroph
specificecosystem_fecal
ecosystemsubtype_oceanic
host_mammalsman
knownhabitats_marine
knownhabitats_humanintestinalmicroflora
host_plantsexceptlegumes
ecosystemcategory_foodproduction
habitat_multiple
cellarrangement_pairs
shape_bacilli
ecosystemsubtype_mine
habitat_freeliving
mammalian_pathogen_enteric
ecosystemtype_soil
habitat_terrestrial
knownhabitats_host
knownhabitats_soil
growth_in_groups
energysource_chemoorganotroph
ecosystemtype_marine
energysource_chemolithoautotroph
pathogenic_in_mammals
knownhabitats_plants
ecosystemtype_respiratorysystem
cellarrangement_chains
ecosystemsubtype_lentic
knownhabitats_mud
mammalian_pathogen_oportunisticnosocomial
halophilic
knownhabitats_humanvaginalmicroflora
ecosystemtype_circulatorysystem
knownhabitats_aquatic
shape_tailed
ecosystemsubtype_nasopharyngeal
ecosystemtype_skin
habitat_hostassociated
energysource_lithotroph
metabolism_carbondioxidefixation
metabolism_cellulosedegrader
mammalian_pathogen_urogenital
ecosystemsubtype_intertidalzone
flagellarpresence
ecosystemcategory_aquatic
ecosystemsubtype_vagina
ecosystem_hostassociated
energysource_phototroph
mobility
knownhabitats_oilfields
metabolism_methanogen
mammalian_pathogen_respiratory_lungdisease
oxygenreq_facultative
knownhabitats_sludge
gram_stain_positive
ecosystemcategory_plants
energysource_chemoheterotroph
ecosystemcategory_mammals
ecosystem_environmental
mammalian_pathogen_cardiovascular_heart_bloodvessels
shape_coccus
ecosystemcategory_terrestrial
energysource_photoautotroph
host_plantslegumes
knownhabitats_food
specificecosystem_rumen
shape_spirilla
cellarrangement_clusters
specificecosystem_sputum
energysource_photosynthetic
ecosystemcategory_animal
bioticrelationship_symbiotic
sporulation
ecosystemtype_geologic
energysource_autotroph
ecosystemcategory_fish
specificecosystem_sediment
knownhabitats_creosotecontaminatedsoil
cellarrangement_singles
ecosystemcategory_wastewater
knownhabitats_humanfecal
pathogenic_in_fish
knownhabitats_feces
metabolism_ammoniaoxidizer
ecosystemtype_rhizoplane
host_fish
metabolism_sulfuroxidizer
phenotypes_alkaliphile
ecosystemtype_reproductivesystem
cellarrangement_filaments
knownhabitats_deepsea
ecosystemsubtype_largeintestine
ecosystemcategory_insecta
ecosystemtype_phylloplane
ecosystemsubtype_cerebrospinalfluid
ecosystemsubtype_saltcrystallizerponds
motility
phenotypes_acidophile
knownhabitats_oralcavity
host_insectsgeneral
mammalian_pathogen_skin_softtissues
metabolism_biomassdegrader
mammalian_pathogen_respiratory_throatnoseeyes
knownhabitats_bovine
metabolism_nitrifying
mammalian_pathogen_urogenital_reproductive
ecosystemtype_composting
knownhabitats_humanairways
knownhabitats_plantsymbiont
ecosystemsubtype_fecal
metabolism_nitrogenproducer
knownhabitats_sediment
ecosystemtype_industrialwastewater
metabolism_ironreducer
mammalian_pathogen_urogenital_urinary
cellarrangement_tetrads
knownhabitats_intestinaltract
ecosystemsubtype_foregut
metabolism_sulfurrespiration
ecosystemcategory_solidwaste
oxygenreq_microaerophilic
knownhabitats_insectendosymbiont
temperaturerange_psychrophilic
pathogenic_in_plants
ecosystemcategory_birds
knownhabitats_skin
ecosystemcategory_bioremediation
metabolism_hydrogensulfidegasrelease
ecosystemsubtype_lotic
mammalian_pathogen_nervous_system
host_mammalspig
knownhabitats_surfacewater
knownhabitats_wastewater
knownhabitats_rootnodule
energysource_methylotroph
mammalian_pathogen_bone
host_mammalsherbivores
metabolism_sulfatereducer
energysource_chemoautotroph
knownhabitats_plantroot
ecosystemtype_metal
knownhabitats_rhizosphere
radioresistance
metabolism_storespolyhydroxybutyrate
ecosystemsubtype_wetlands
metabolism_ironoxidizer
knownhabitats_endosymbiont
energysource_oligotroph
0
10
20
30
40
Mean Decrease in Accuracy

Figure A.14: Importance of all predictors in CRISPR RF model, as measured by the mean decrease in the Gini impurity index or accuracy when that variable is excluded from the respective forest. Note the elbow in the Gini importance ranking after the first ten predictors.

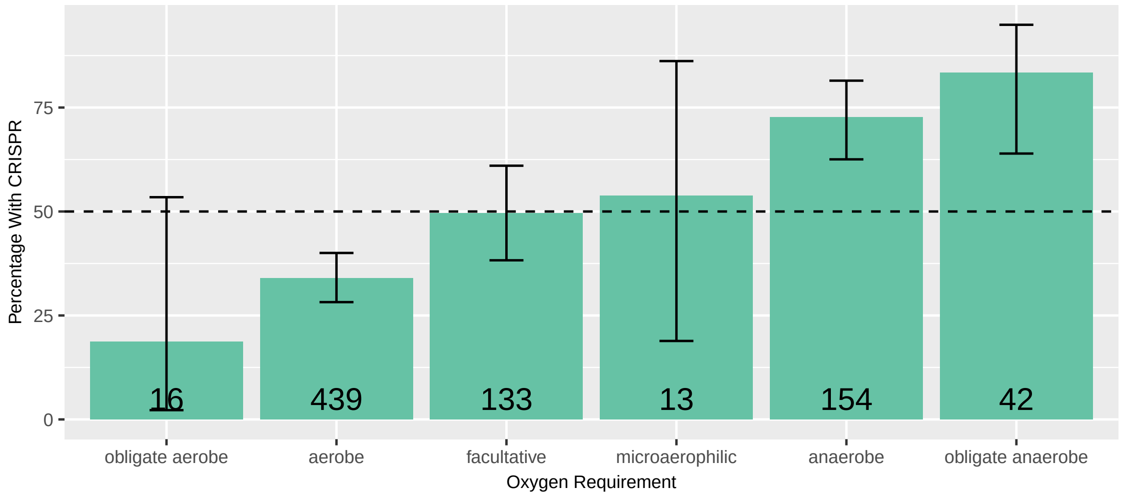


Figure A.15: The link between oxygen requirement and CRISPR incidence is apparent even when sub-setting to only mesophiles. Error bars are 99% binomial confidence intervals. Total number of genomes in each trait category shown at the bottom of each bar. Categories represented by fewer than 10 genomes were omitted.

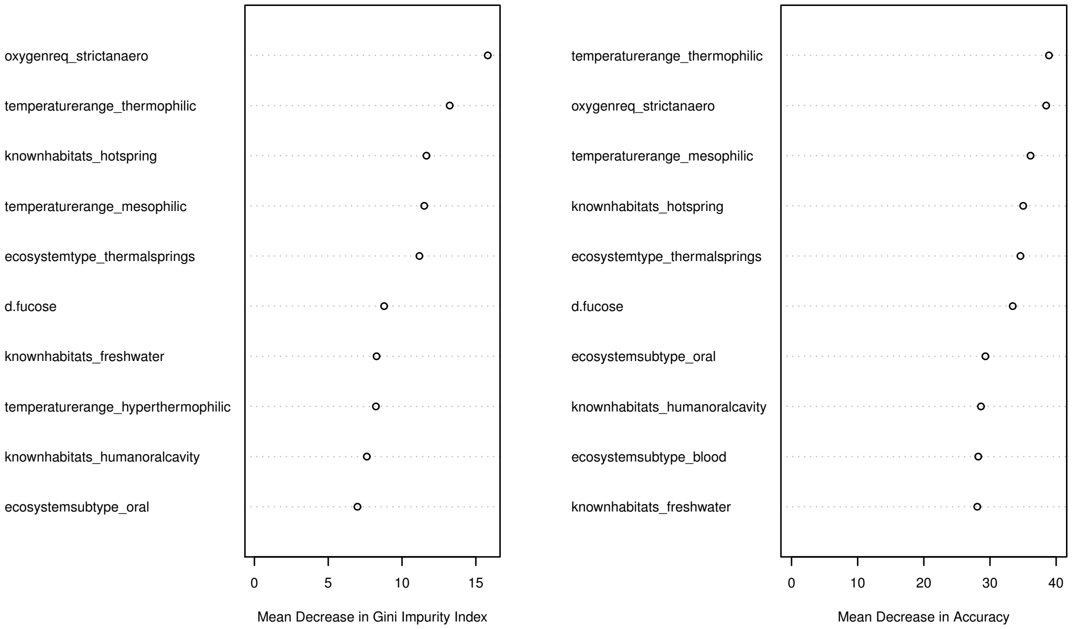
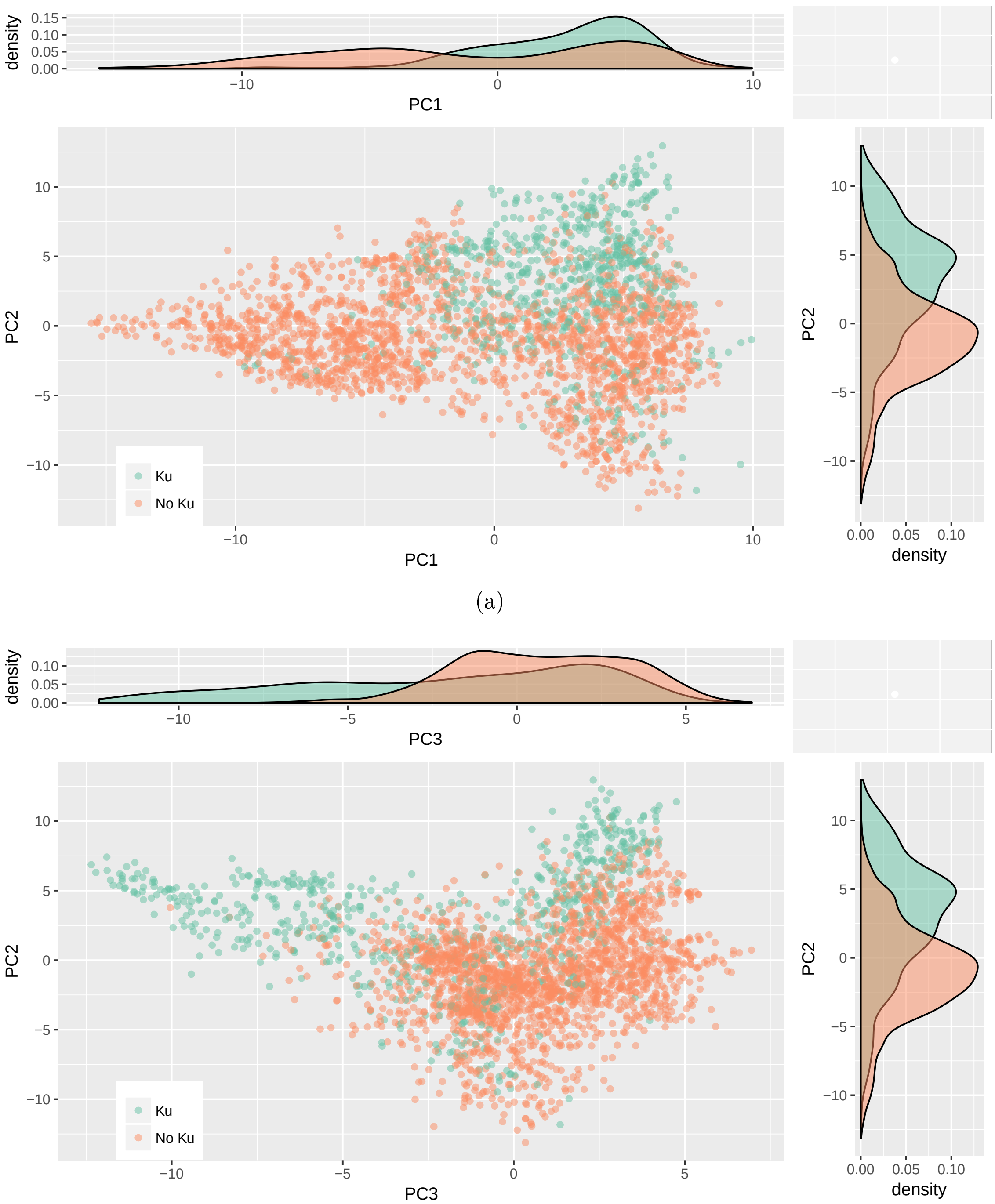


Figure A.16: Importance of top ten predictors in the RF model built excluding the “phyletic profile” and “gene neighborhood” information sources from ProTraits, as measured by the mean decrease in the Gini impurity index or accuracy when that variable is excluded from the model.



(a)

(b)

Figure A.17: The incidence of the Ku protein in trait space. PCA as in Figs 2.1 and A.9.

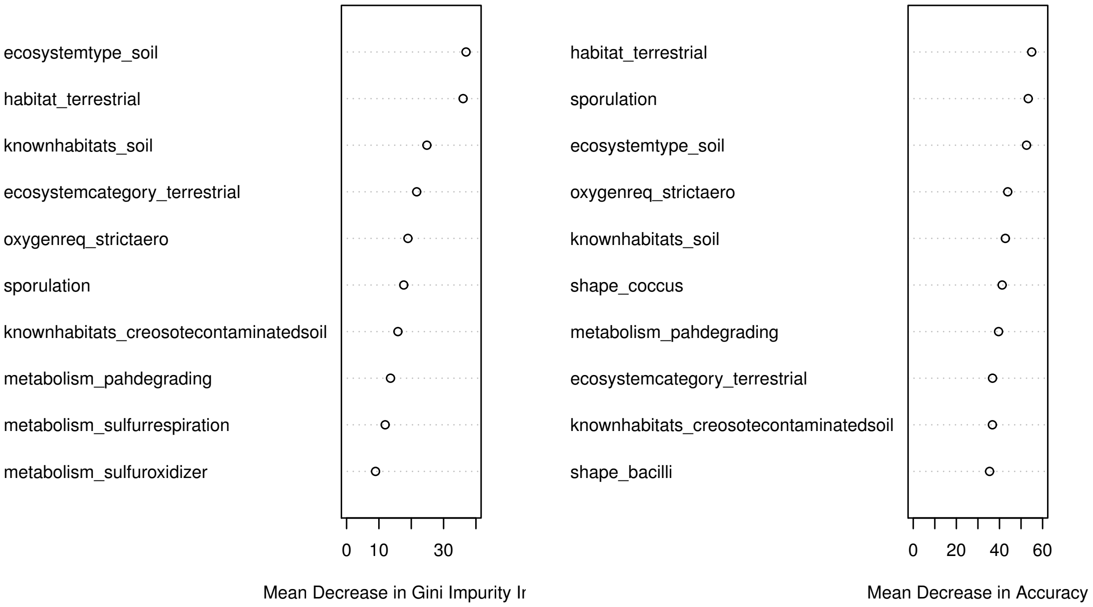
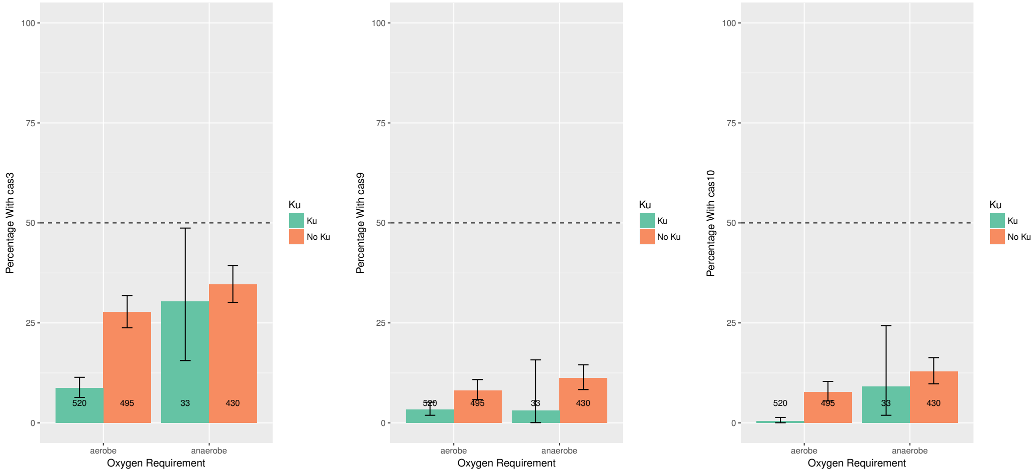
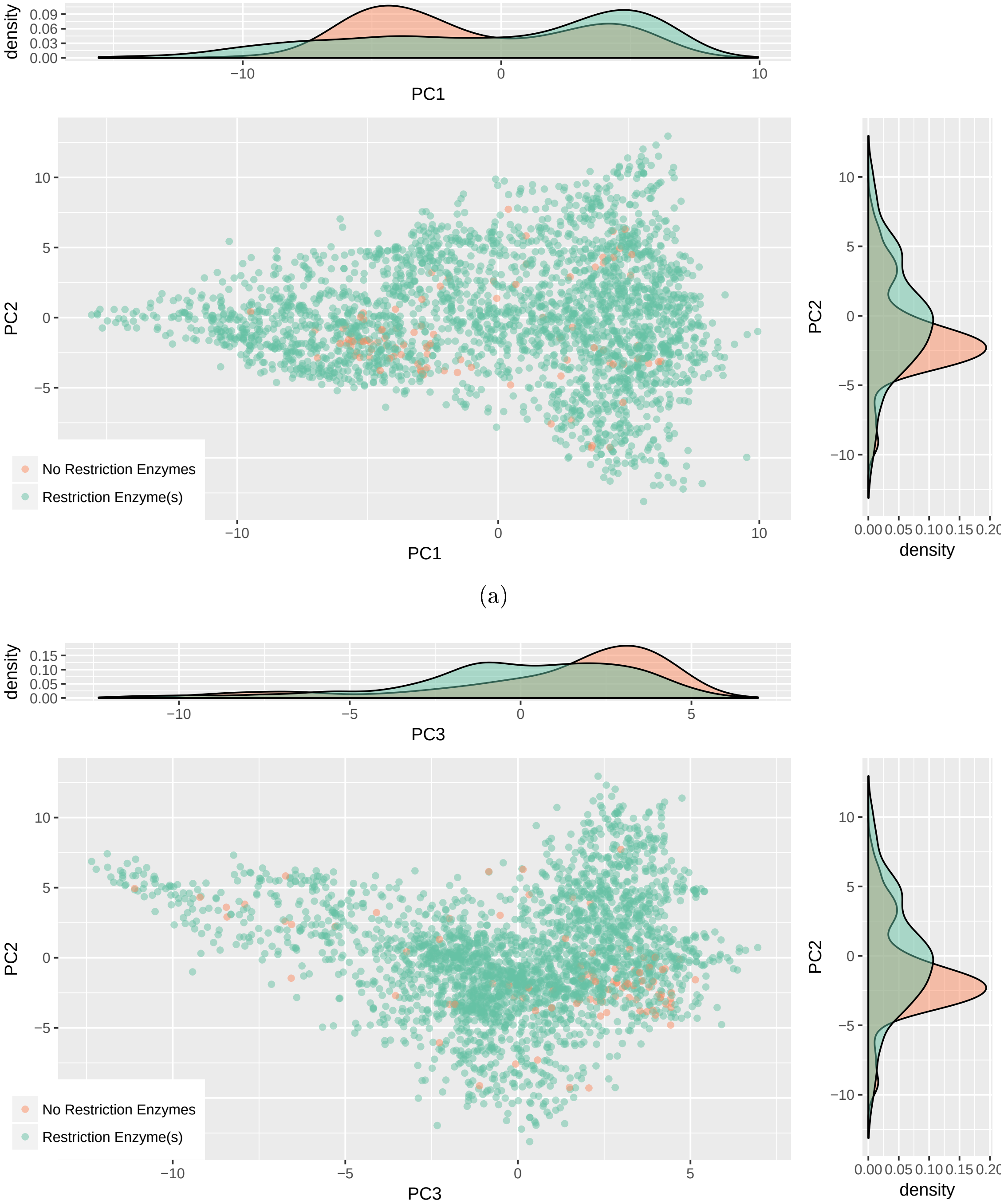


Figure A.18: Importance of top ten predictors in the RF model of Ku incidence. This model had high predictive ability ($\kappa = 0.578$).



(a) (b) (c)

Figure A.19: CRISPR and Ku are negatively associated in aerobes but not anaerobes. Percentage of genomes with Cas proteins associated with a particular system type. Error bars are 99% binomial confidence intervals. Total number of genomes in each trait category shown at the bottom of each bar. Of the 1047 genomes represented here 253 have *cas3*, 61 have *cas9*, and 54 have *cas10*.



(a)

(b)

Figure A.20: The incidence of restriction enzymes in trait space. PCA as in Figs 2.1 and A.9.

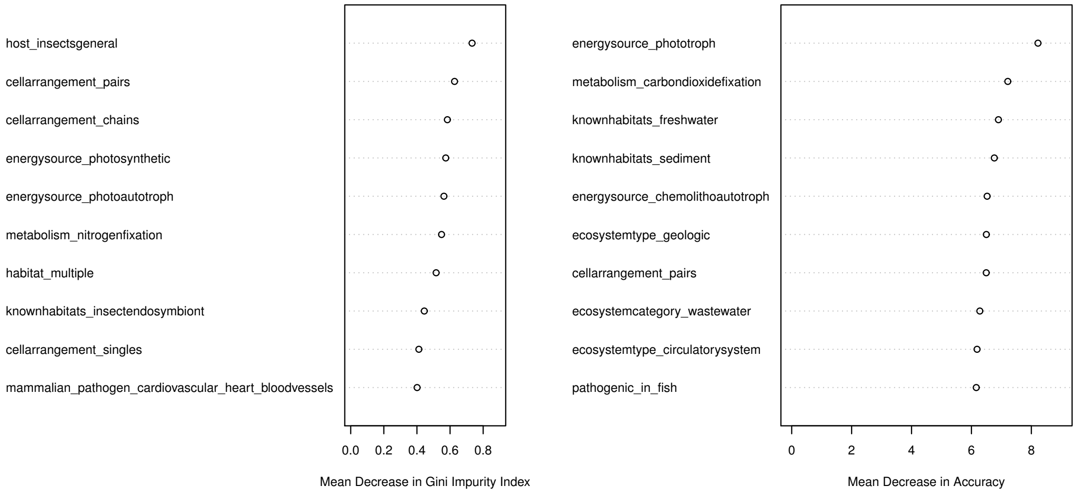


Figure A.21: Importance of top ten predictors in the RF model of restriction enzyme incidence, as measured by the mean decrease in the Gini impurity index or accuracy when that variable is excluded from the model.

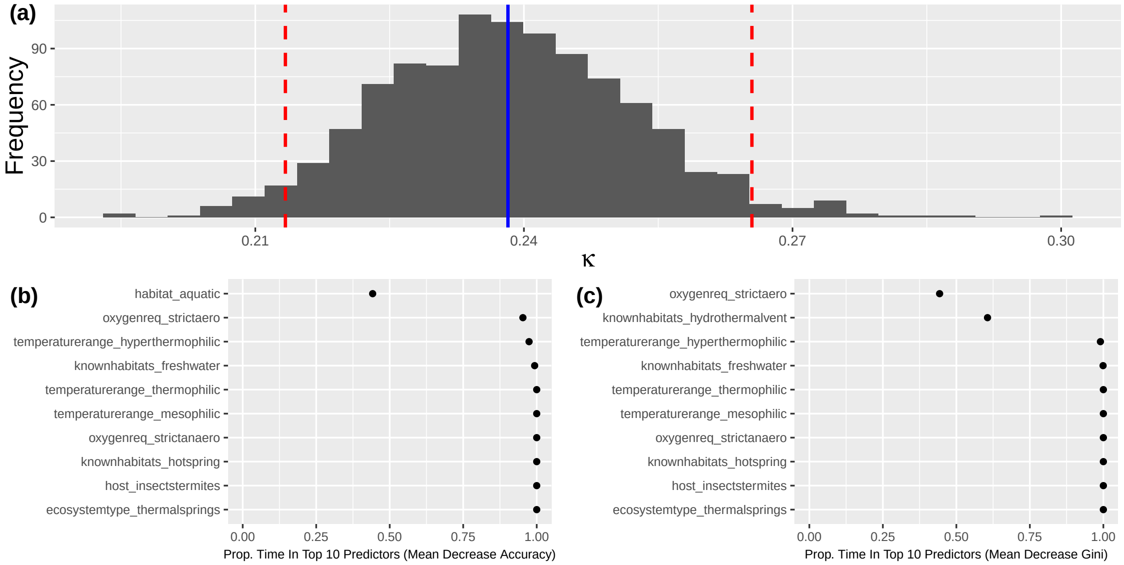


Figure A.22: Resampling genomes has little effect on our overall outcome. (a) Distribution of κ values for 1000 RF models built with resampled datasets. Mean (blue) and 95% CIs (red) indicated with vertical lines. (b-c) The proportion of resampled datasets for which each predictor fell within the set of top 10 predictors based on variable importance scores.

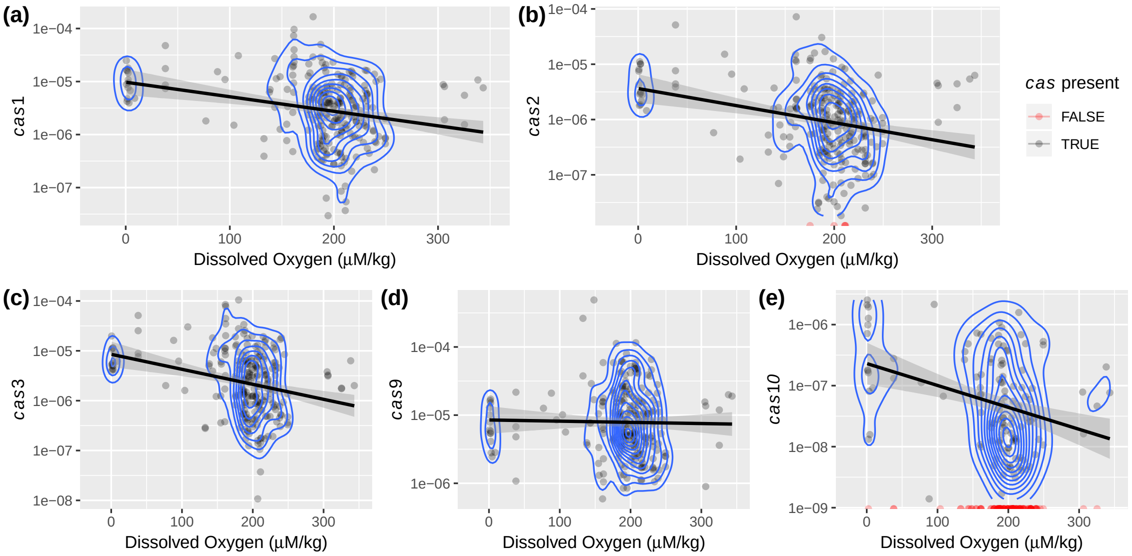


Figure A.23: Functional profiles for *cas* genes from Tara Oceans Project with corresponding oxygen metadata. Values for each *cas* gene shown are the coverage mapping to that orthologous group normalized by the total coverage in the metagenome. Zero values for coverage plotted along x-axis in red (since data plotted on log scale). Trend-lines plotted on log transformed data for ease of interpretation.

Appendix B: Supplemental Information For: "Selective maintenance of multiple CRISPR arrays across prokaryotes"

B.1 Validation of functional / non-functional classification

Our power to detect selection depends critically on our ability to classify genomes as CRISPR functional vs. non-functional. Functional CRISPR arrays should, on average, contain more spacers than non-functional arrays [226]. Thus we compared the number of repeats in CRISPR arrays in genomes with both *cas1* and *cas2* present ("functional", 16.01 repeats on average) to the number of spacers in genomes lacking either or both genes ("non-functional", 12.23 repeats on average) and confirmed that the former has significantly more than the latter ($t = -36.516$, $df = 55340$, $p < 2.2 \times 10^{-16}$; Fig B.9). This difference in length (3.80 repeats) is not as large as one might expect, possibly because some systems are able to acquire or duplicate spacers via homologous recombination [165] and arrays may have been inherited recently from strains with active *cas* machinery. The mean array length across the dataset was 15.12 repeats.

B.2 Deriving the distribution of number of arrays per genome under a neutral accumulation model

Recall our null hypothesis that in genomes with functional CRISPR systems possession of a single array is highly adaptive (i.e. viruses are present and will kill any susceptible host) but additional arrays provide no additional advantage. Thus these additional arrays will appear and disappear in a genome as the result of a neutral birth/death horizontal transfer and loss process, where losses are assumed to remove an array in its entirety. This hypothesis predicts that the non-functional distribution will look like the functional distribution shifted by one (S_i):

$$H_0 : \; N_i \approx S_i = F_{i+1} / \sum_{j=1}^{\infty} F_j \tag{B.1}$$

for $i \geq 0$.

We begin by deriving a functional form for the distribution N_i from first principles following a neutral process. If CRISPR arrays arrive in a given genome at a constant rate via rare horizontal transfer events, then we can model their arrivals using a Poisson process with rate η. Assuming arrays are also lost independently at a constant rate, the lifetime of each array in the genome will be exponentially distributed with rate ν. This leads to a linear birth-death process of array accumulation, which yields a Poisson equilibrium distribution with rate $\lambda = \frac{\eta}{\nu}$. While this rate might be constant for a given taxon, it will certainly vary across taxa due to different intrinsic (e.g. cell wall and membrane structure) and extrinsic factors (e.g. density of neighbors, environmental pH and temperature) [16]. We model this variation by allowing genome j to have rate $\lambda_j = \frac{\eta_j}{\nu_j}$ and assuming $\lambda_j \sim \text{Gamma}(\alpha, \beta)$, which we pick for its flexibility and analytic tractability. This combination of gamma and Poisson distributions leads to the number of arrays i in a random genome following a negative binomial distribution $N_i = \text{NB}(r, p)$ where $r = \alpha$ and $p = \frac{\beta}{1+\beta}$.

Now we can fit this distribution to data to find maximum likelihood estimates of r and p for the distribution of array counts in both the set of non-functional genomes (N_i) and the set of functional genomes as shifted under our null hypothesis (S_i). This allows us to construct a parametric test of selection. We expect that $\hat{r}_N \approx \hat{r}_S$ and $\hat{p}_N \approx \hat{p}_S$ under our null hypothesis (where subscripts correspond to the distribution to which the parameters were fit). When our null hypothesis is violated it should shift the means of these distributions. Therefore we estimate and compare these means $\mu_x = \frac{p_x r_x}{1-p_x}$, $x \in \{N, S\}$. We expect that $\hat{\mu}_S > \hat{\mu}_N$ if more than one array is selectively maintained, and we bootstrap confidence intervals on these estimates by resampling with replacement from our functional and non-functional array count distributions in order to determine whether the effect is significant.

B.3 Instantaneous array loss vs. gradual decay

There are two possible routes to complete CRISPR array loss: (1) an all-at-once loss of the array (e.g. due to recombination between flanking insertion sequences [50, 140]) and (2) gradual decay due to spacer loss. Previous experimental evidence supports route (1) spontaneous loss of the entire CRISPR array [51], as do comparisons between closely related genomes [50]. The distinction above is important, because if CRISPR array loss were to occur primarily via route (2) gradual decay, then functional genomes would have an intrinsically lower rate of array loss than non-functional genomes. This is because in functional genomes spacer acquisition would counteract spacer loss, reducing the rate of array decay, whereas this compensation would not occur in non-functional genomes. This could lead us to spuriously accept a result of selection maintaining multiple arrays.

If arrays were primarily lost via gradual decay, we would expect our data to show a positive relationship between the number of arrays in a genome and the average array length in a genome, because arrays experiencing more decay (either due to increased spacer loss rates or reduced acquisition rates) should be shorter and prone to eventual deletion. In functional genomes with the complete spacer acquisition machinery (*cas1* and *cas2*) this trend would be due to the higher probability of stochastically reaching a 0-spacers state in shorter arrays, and arrays will in general be shorter in genomes with lower spacer acquisition rates. In non-functional genomes that lack the complete spacer acquisition machinery, this trend would result from differences in time since loss of acquisition machinery, where genomes that had lost that machinery farther in the past would have both shorter arrays and fewer arrays on average.

Overall we see no relationship between mean array length and array count in a genome (linear regression, $m = -0.001$, $p = 0.109$, $R^2 = 5.55 \times 10^{-5}$). Surprisingly, in functional genomes we find a slightly negative linear relationship between the number of arrays in a genome and mean array length in that genome ($m = -0.0081$,

$p < 2 \times 10^{-16}$, $R^2 = 0.0032$). In non-functional genomes we see a slightly positive relationship ($m = 0.0054$, $p = 7.23 \times 10^{-10}$, $R^2 = 0.0026$). While both of these relationships are significant, they are extremely weak and probably spurious. This lack of a clear relationship suggests that gradual decay is not the primary cause of array loss. Nevertheless, because rates of spacer acquisition and loss and array acquisition (via HGT) and loss are two somewhat easily confounded processes, this evidence is not conclusive.

As an additional, more direct test of whether array decay or instantaneous loss drives CRISPR array loss dynamics, we consider the difference in array length between putatively "functional" and "non-functional" arrays. On average the presence of the *Cas* acquisition machinery leads to the addition of about 4 repeats to a given array (Fig B.9). We expect short arrays, of approximately this length, to be the most rapidly lost upon loss of array functionality. Thus, if array loss occurs primarily through gradual decay, then our result of selection maintaining multiple arrays would be driven primarily by short, functional arrays. By removing any functional arrays below some threshold length from the dataset, we can test this hypothesis. Even upon removal of all functional arrays less than 10 spacers long from the dataset, we still observe a substantial signature of selection maintaining multiple arrays (Figs B.10 and B.11). By design, this signature does weaken slightly as the threshold length for functional arrays is increased, but this is to be expected as the removal of arrays from the functional dataset must also decrease the difference in mean number of arrays between the functional and non-functional datasets (since no removals are made from the non-functional set). It is notable though, that this decrease in signal is small (especially when removing arrays < 6 spacers long), demonstrating that the selective signal is not being driven primarily by short, functional arrays.

B.4 Model Analysis

We develop a simple deterministic model of the spacer turnover dynamics in a single CRISPR array of a bacterium exposed to n viral species (i.e., disjoint protospacer sets). Let C_i be the number of spacers in the array that target viral species i:

$$\frac{dC_i}{dt} = \underbrace{a_i(t, C_i)}_{\text{Acquisition}} - \underbrace{\mu_L l_i(C_1, ..., C_n)}_{\text{Loss}}, \tag{B.2}$$

$$a_i(t, C_i) = \mu_A v_i f_i(t) g(C_i), \tag{B.3}$$

where μ_L is the per-spacer loss rate parameter, $l_i(C_1, ..., C_n)$ is a function describing how spacer loss depends on the array length, μ_A is the per-infection spacer acquisition rate, v_i is a composite parameter describing the infection intensity in the environment (viral density times adsorption rate), $f_i(t) \in [0, 1]$ is a function describing the fluctuations of viral population i over time, and

$$g(C_i) = \begin{cases} 1 & C_i < 1 \\ p & C_i \geq 1 \end{cases} \tag{B.4}$$

is a function determining whether or not the system is "primed" towards viral species i (i.e., if a CRISPR array has a spacer targeting a particular viral species, the rate of spacer acquisition towards that species is increased; [141, 142]), where $p > 1$ is the degree of priming (see Table C.1 for summary of model parameters and units).

Using this model we can determine the optimal spacer acquisition rate given a particular pattern of pathogen recurrence in the environment, $f_i(t)$. If the optima for distinct recurrence patterns do not overlap, it indicates that multiple arrays would be required to simultaneously combat viral species with these distinct recurrence patterns.

We analyze the ability of the array to respond to three types of viral threats: (1) "background" species representing the set of all viruses persisting over time in the environment ($f_B(t) = 1$), "transient" species that leave the system and return after some interval of time ($f_T(t) \in \{0, 1\}$), and "novel" species that have not been previously encountered. Thus we can compare how CRISPR balances the need for consistent immunity, long-term memory, and rapid adaptation (full analysis below).

We consider two forms for the function l_i based on known features of CRISPR biology. (1) The rate of per-spacer loss increases linearly with locus length. This form is based on the observation that spacer loss appears to occur via homologous recombination between repeats [31, 143, 144], which becomes more likely with increasing numbers of spacers (and thus repeats). (2) The length of an array is capped at some fixed "effective" number of spacers. This form is based on evidence that mature crRNA transcripts from the leading end of the CRISPR array are far more abundant than those from the trailing end, and that this decay over the array happens quickly (most transcripts are from the first few spacers; [145, 146, 147]). We analyze both models below, though they give qualitatively similar results, and so we focus on case (1) in the main text Results.

Primary Model: Array-Length-Dependent Spacer Loss

First we consider a version of the model above where the per-spacer loss rate increases linearly with the total number of spacers in the array (i.e., physical loss via homologous recombination):

$$\frac{dC_i}{dt} = \underbrace{a_i(t, C_i)}_{\text{Acquisition}} - \underbrace{\mu_L C_i \sum_{j=1}^{n} C_j}_{\text{Loss}}, \tag{B.5}$$

$$a_i(t, C_i) = \mu_A v_i f_i(t) g(C_i), \tag{B.6}$$

$$g(C_i) = \begin{cases} 1 & C_i < 1 \\ \\ p & C_i \geq 1 \end{cases}. \tag{B.7}$$

That is, $l_i(C_1, ..., C_n) = C_i \sum_{j=1}^{n} C_j$. Parameter values can be found in Table C.1.

For the purposes of simplifying our analysis we consider the case where there are two viral species in the system, one that persists at some background level ("background" B, $f_B(t) = 1$) and another that returns in sharp bursts at periodic intervals ("transient" T, $f_T(t) \in \{0, 1\}$).

This two-virus situation captures the conflict between the ability to maintain immune memory of periodically recurring infections but also defend against ever-present persisting viral enemies. We are interested in the time lag between the return of virus T to the system and the development of host immunity. In the case where memory is maintained during the absence of virus T the lag will be zero, otherwise it will depend on the base spacer uptake rate (μ_A) and the relative densities of the two viral populations.

We examine the time to reacquisition of immunity towards virus T using the following procedure (Fig B.12). Note that for simplicity we let time have units of virus return time (so that one unit of time equals the interval in which T is absent from the system). Our method is as follows: (1) start the system at its equilibrium state with both viral species present, (2) remove species T for a single unit of time and track the decay of C_T, and (3) return T to the system and calculate how long it takes after this return for C_T to exceed one (time to reacquisition of immunity, t_I). Let $t_I = 0$ if C_T remains above one despite the absence of T (i.e. no loss of immune memory). In more detail:

1. Find the equilibrium of our system ($\tilde{C}_B, \tilde{C}_T$) when T is present, assuming that at the equilibrium state $\tilde{C}_B, \tilde{C}_T \geq 1$. We have that

   $$\tilde{C}_B = \frac{v_B\sqrt{\mu_A p}}{\sqrt{\mu_L(v_B + v_T)}} \tag{B.8}$$

   and

   $$\tilde{C}_T = \frac{v_T\sqrt{\mu_A p}}{\sqrt{\mu_L(v_B + v_T)}}. \tag{B.9}$$

2. Use this equilibrium as an initial condition for a system where virus T has been removed (where $f_T(t) = 0$, so $\frac{dC_T}{dt} = -\mu_L C_T (C_T + C_B)$) and solve (numerically) for the state of the system at time $t = 1$ assuming that T remains absent ($C_T(1)$). In practice we use the ode15s solver available in MATLAB (2016a, MathWorks, Inc., Natick, Massachusetts, United States).

3. Find the time to reacquisition of immunity (t_I such that $C_T (t_I + 1) = 1$) by solving the unprimed system where we ignore loss (since no spacers yet exist to be lost, loss should not be important). Define $C^*_T(t)$ so that

$$C^*_T(t) = \mu_A v_T t + C_T(1). \tag{B.10}$$

We are interested in t_I where $C^*_T(t_I) = 1$, so that

$$t_I = \frac{1 - C_T(1)}{\mu_A v_T} \tag{B.11}$$

where $C_T(1)$ is the solution from step (2). This equation only holds if $C_T(1) < 1$ (let $t_I = 0$ otherwise). This time to reacquisition is our measure of fitness, with a lower t_I indicating lower fitness of the host. We define some τ such that if $t_I < \tau$ we say immunity is maintained. Ideally, τ should be shorter than the time it takes viruses to cause irreparable damage to an infected bacterium after initial infection.

We can also use Eq B.11 to find the time of immune acquisition towards a novel viral species (N) arriving in the environment. When no spacers towards that species are already contained in the CRISPR array ($C_N(0) = 0$), the time to novel acquisition is

$$t_N = \frac{1}{\mu_A v_N}. \tag{B.12}$$

Note that we assume here that the spacer loss rate does not effect $\frac{dC_N}{dt}$ when $C_N < 1$ in order to simplify our analysis. During analysis we let $v_N = v_T$ for simplicity.

Alternative Model: Leader-End crRNA Processing

We are interested in having a hard cutoff on the array length in this version of the model (corresponding to a fixed cap on the number of transcribed spacers). Therefore, we consider an "effective" array of fixed length L and let C_i be the proportion of the array taken up by spacers targeting viral species i (n total viral species, $\sum_{i=1}^{n} C_i = 1$). This model generally follows a form similar to that in Eqs B.2-B.4:

$$\frac{dC_i}{dt} = a_i(t, C_i) - C_i \sum_j a_j(t, C_j) \tag{B.13}$$

where,

$$a_i(t, C_i) = \mu_A v_i f_i(t) g(C_i). \tag{B.14}$$

We modify the definition of the priming function g to match the new scenario so that

$$g(C_i) = \begin{cases} 1 & C_i < \frac{1}{L} \\ p & C_i > \frac{1}{L} \end{cases} \tag{B.15}$$

Observe that in Eq B.13 the flow rate into and out of the system match so that the overall number of effective spacers should not change (in this case $l_i(t, C_1, ..., C_n) = C_i \sum_j a_j(t, C_j)$ is also a function of t and we let $\mu_L = 1$).

When $f_i(t) = 1 \, \forall i$ and we assume the system is primed towards all viruses, then we have equilibrium

$$\tilde{C}_i = \frac{v_i}{\sum_j v_j}. \tag{B.16}$$

Let's return to our simple two-virus system with background (B) and transient (T) viral species. We perform a three step analysis similar to the one used for the linear spacer loss model above:

1. Concentrating on an interval where both viruses are present ($f_T = 1$)

$$\frac{dC_i}{dt} = \begin{cases} \mu_A v_i p - p\mu_A C_i \left(v_i + v_{j\neq i}\right) & C_i, C_{j\neq i} > \frac{1}{L} \\ \mu_A v_i - \mu_A C_i \left(v_i + p v_{j\neq i}\right) & C_i < \frac{1}{L}, C_{j\neq i} > \frac{1}{L} \\ \mu_A v_i p - \mu_A C_i \left(p v_i + v_{j\neq i}\right) & C_i > \frac{1}{L}, C_{j\neq i} < \frac{1}{L} \\ \mu_A v_i - \mu_A C_i \left(v_i + v_{j\neq i}\right) & C_i, C_{j\neq i} < \frac{1}{L} \end{cases} \quad i, j \in \{B, T\}. \tag{B.17}$$

We derive the equilibrium spacer content, assuming that the system starts from a doubly primed condition so that

$$\tilde{C}_i = \begin{cases} \frac{v_i}{v_i+v_{j\neq i}} & \frac{1}{L} < \frac{v_i}{v_i+v_{j\neq i}} < 1-\frac{1}{L} \\ \frac{v_i}{v_i+pv_{j\neq i}} & \frac{v_i}{v_i+v_{j\neq i}} < \frac{1}{L} \\ \frac{pv_i}{pv_i+v_{j\neq i}} & \frac{v_i}{v_i+v_{j\neq i}} > 1-\frac{1}{L} \end{cases} \qquad i,j \in \{B,T\}. \tag{B.18}$$

2. We now focus on the dynamics of the system after the transient viral species leaves ($f_T = 0$) assuming the system starts from the equilibrium in Eq B.18. The decay of spacers targeting the transient viral population will follow

   $$\frac{dC_T}{dt} = \begin{cases} -\mu_A v_B p C_T & C_T < 1-\frac{1}{L} \\ -\mu_A v_B C_T & C_T > 1-\frac{1}{L} \end{cases}. \tag{B.19}$$

   We can solve for $C_T(t)$ with the given initial condition so that

   $$C_T(t) = \begin{cases} \left(\frac{v_T}{v_T+v_B}\right) e^{-p\mu_A v_B t} & \tilde{C}_T, \tilde{C}_B > \frac{1}{L} \\ \left(\frac{v_T}{v_T+pv_B}\right) e^{-p\mu_A v_B t} & \tilde{C}_T < \frac{1}{L}, \tilde{C}_B > \frac{1}{L} \\ \left(\frac{pv_T}{pv_T+v_B}\right) e^{-\mu_A v_B t} & t \leq t_B, \tilde{C}_B < \frac{1}{L} \\ \left(\frac{pv_T}{pv_T+v_B}\right) e^{-\mu_A v_B (t_B + p(t-t_B))} & t > t_B, \tilde{C}_B < \frac{1}{L} \end{cases} \tag{B.20}$$

   where

   $$t_B = \frac{-\ln\left(\frac{(pv_T+v_B)(L-1)}{pv_T L}\right)}{\mu_A v_B}. \tag{B.21}$$

3. Let us assume that time is measured in units of viral return intervals so that

the return time of the transient (T) species is $t_I = 1$. Then our goal is to find $C_T(1)$, assess whether this value has dropped below $\frac{1}{L}$ (immune loss), and if so, find the time to immune reacquisition $C_T(t_I - 1) = \frac{1}{L}$. Let us define a new function $C_T^*(t)$ that tracks the re-acquisition of immunity after T returns to the system so that $C_T^*(0) = C_T(1)$ and, assuming no decay of spacers when $C_T^* < \frac{1}{L}$,

$$\frac{1}{L} = C_T^*(t_I) = \mu_A v_T t_I + C_T(1). \tag{B.22}$$

Then

$$t_I = \frac{\frac{1}{L} - C_T(1)}{\mu_A v_T} \tag{B.23}$$

and we have $C_T(1)$ from Eq B.20.

This model gives qualitatively similar results to those found with the primary model (Fig B.2).

B.5 Confirming selection

In order to further confirm our result of selective maintenance of multiple CRISPR arrays, we (1) subsampled overrepresented taxa in the dataset, (2) performed phylogenetically-corrected tests to confirm both that differential rates of horizontal gene transfer (HGT) between species were not driving our results and that the pattern of selection we observed was not isolated to any particular group, (3) showed that potential linkage between *cas* genes and CRISPR arrays was not producing the signature of selection we observed, and (4) demonstrated that the genome assembly level had no effect on our outcome. Additionally, (5) we discuss why the potential effects of CRISPR on rates of HGT cannot account for the selection we see.

(1) Sub-sampling to reduce the influence of overrepresented taxa altered our parameter estimates slightly, but did not change our overall result (sampled 10 genomes from each species with > 10 genomes, $\Delta\mu = 1.13 \pm 0.09$, Fig B.6, Table B.1). To control for the possibility that multiple sets of *cas* genes in a small subset of genomes could be driving this selective signature, we restricted our dataset only to genomes with one or fewer signature targeting genes and one or fewer copies each of the genes necessary for spacer acquisition. Even in this restricted and subsampled set of genomes selection maintains more than one (functional) CRISPR array, though the effect size is smaller ($\Delta\mu = 0.18 \pm 0.08$, Fig B.7).

(2) The number of CRISPR arrays is positively related to the number of *cas* genes in a genome (Fig B.13). Differential rates of horizontal gene transfer (HGT) among species could produce an observed correlation between *cas1* and *cas2* presence and array count in the absence of selection. To control for this potentially confounding effect we performed a species-wise parametric test. For each species k we calculated a species-specific $\Delta\mu_k = \hat{\mu}_{S_k} - \hat{\mu}_{N_k}$ and then bootstrapped the mean of the distribution of these values ($\bar{\Delta\mu}$) to evaluate significance. This test confirms a signature of multi-array selection ($\bar{\Delta\mu} = 0.70 \pm 0.14$) despite low power due to most species having few sequenced genomes. Additionally, in order to determine if the signature of selection was confined to a particular set of clades, we mapped all species-specific $\Delta\mu_k$ values onto the SILVA Living Tree 16s rRNA tree [228]. Of

the 623 species with at least one functional and one non-functional genome, 568 were represented on the tree. Positive and strongly positive (> 1) values of $\Delta\mu_k$ were distributed across the tree, indicating this phenomenon was not isolated to a particular group (Fig B.14). Formal testing revealed no significant phylogenetic signal in the $\Delta\mu_k$ values ($K = 1.88 \times 10^{-9}$, $p = 0.604$; [229, 230]). Briefly, this test for phylogenetic signal involves randomly permuting the underlying phylogeny and comparing the fit of the observed trait data (in this case $\Delta\mu_k$) to these random trees (test developed in [229] and implemented with the "phylosig" function in the phytools R package [230]).

(3) Often CRISPR arrays and *cas* genes are collocated such that loss of one may be linked to loss of the other. At equilibrium, the distribution of array counts per genome will be unaffected by such collocation. To test this assumption directly, we used regression to check if the minimum distance between CRISPR arrays and *cas* genes in a genome drives the species-specific signature of selection, $\Delta\mu_k$ (using only completely assembled genomes). We saw a slight positive relationship between CRISPR-*cas* distance and our signature of multi-array selection, the opposite of what we would expect if linkage were driving our results ($m = 3.163 \times 10^{-7}$, $p = 8.52 \times 10^{-6}$, $R^2 = 0.009937$).

(4) Finally, to confirm that assembly level had no effect on our conclusion, we ran our parametric test restricted to completely assembled genomes in the dataset (6263 genomes, $\Delta\mu = 0.98 \pm 0.09$).

(5) CRISPR immunity may generically increase rates of horizontal gene transfer [148], including transfer of CRISPR arrays themselves. Under our model of

array accumulation, an increase in the rate of the arrival of new arrays in functional genomes would certainly increase the mean of the distribution of arrays per functional genome. Nevertheless, assuming only selection on a single array (our null hypothesis), we would still expect this functional distribution to have negative binomial form shifted by one array (see Chapter 3 Methods and Appendix B.2). It is clear from Fig 3.1(b) that the distribution must be shifted by at least two arrays in order to resemble a negative binomial distribution.

B.6 Neo-CRISPR Arrays

Off-target spacer integration into the genome can spawn novel CRISPR arrays in *E. coli* [149]. This could create a spurious signature of selection maintaining multiple arrays using our test, since the production of "neo-CRISPR arrays" would only occur in functional genomes. A simple way to control for this is to merge all CRISPR arrays with identical consensus repeat sequences in a genome, thus removing any duplicates. Doing this, we find that the signature of multi-array selection remains, albeit being somewhat less strong ($\Delta\mu = 0.46 \pm 0.02$). We were considerably surprised that this signature of selection still remained after merging, since such merging will also remove a large portion of arrays acquired through horizontal transfer, assuming such transfers most often happen between closely related individuals. In any case, while the production of neo-CRISPR arrays may be driving our result in part, it cannot account for the overall signal. It is unclear if neo-CRISPR arrays are commonly produced in bacteria via off-target integration, though [149] found circumstantial evidence it may occur in two other species. The CRISPR system of *E. coli* is not naturally active [231] and requires artificial up-regulation of the spacer acquisition machinery, so that its dynamics may not be representative of CRISPR systems at large. Nevertheless, this mechanism may explain the large number of arrays found in some genomes (e.g., *Clostridium difficile* genomes typically have nine arrays; [135, 136]).

We note that this control also applies to any other potential array duplication process, since repeat sequence should be preserved during duplication.

In both the case of neo-CRISPR arrays and other duplication events, we might expect that the "new" array formed via off-target integration or fragmentation of a larger array would lead to second arrays that were shorter and potentially lower scoring than the first. Comparing arrays in two-array genomes, we do not see this pattern emerge (i.e., no negative correlation between arrays in terms of length or score). In fact, in both cases we see slight positive correlations between arrays, though with little explanatory value (Fig B.15).

B.7 Validation of CRISPRDetect array predictions

We ran our tests for selective maintenance of multiple arrays on the same dataset excluding arrays with a CRISPRDetect score lower than 6 (double the default threshold). We found no qualitative differences in our results when we used this increased detection threshold ($\Delta\mu = 1.00 \pm 0.02$). By default, CRISPRDetect identifies arrays with repeats matching experimentally-verified CRISPR arrays as well as *de novo* repeats. If we restrict to only arrays with a positive hit on this list we again found the same pattern ($\Delta\mu = 0.76 \pm 0.03$).

We also downloaded the set of CRISPR arrays and array-lacking genomes available on the CRISPR Database [159]. This database uses an alternative algorithm for array detection [232] and thus serves as an independent verification of our results. This dataset showed a clear signature of selection maintaining multiple arrays ($\Delta\mu = 1.49 \pm 0.17$).

B.8 Autoimmunity Constrains Unprimed Spacer Acquisition Rates

Empirical evidence suggests that there is no self versus non-self recognition mechanism in the CRISPR systems of *Streptococcus thermophilus*, a popular model system for CRISPR research [27]. Thus any increase in spacer acquisition will also increase the rate of autoimmune targeting. In the absence of viruses or once immunity has been established, we can model the growth of bacteria (B) experiencing autoimmune targeting in a chemostat as

$$\dot{B} = B\left(\frac{vR}{z+R} - \alpha - w\right) \tag{B.24}$$

with resources (R)

$$\dot{R} = w(A - R) - \frac{evR}{R+z}B \tag{B.25}$$

where μ_A is the spacer acquisition rate and $\alpha = 50\mu_A$ is an estimate of the rate of autoimmunity based of the relative genome sizes of *S. thermophilus* and its lytic phage 2972 [38, 233] (other parameters in Table C.2).

As shown in Fig B.16, there is little effect of autoimmunity on the equilibrium density of bacteria for $\mu_A < 10^{-3}$, but after a point around $\mu_A = 10^{-2}$ there is a rapid drop in density. This puts a theoretical cap on the maximum rate of spacer uptake in our system and imposes a severe cost on spacer uptake rates greater than 10^{-2} given the parameters used here, based on the *S. thermophilus* system. Other taxa do seem to exhibit some degree of self versus non-self recognition, but still frequently incorporate self-spacers [25, 61, 153], suggesting that our general result holds across taxa though the threshold is likely to be variable. We note that in Fig B.16 even a 50% reduction in the rate of autoimmunity only shifts the threshold spacer acquisition rate by a small amount. Additionally, although CRISPR may provide a competitive advantage when viruses are present in the system, this advantage cannot help the host overcome the autoimmunity "cap" on acquisition rate after which the population is no longer viable.

B.9 Bet Hedging Against Memory Loss

Spacer loss in the CRISPR array most likely occurs via homologous recombination of repeat sequences [31, 143, 144]. Thus the time to immune loss will increase with the number of arrays targeting a particular viral species. Assuming that immunity towards a given virus in a single array has an exponentially distributed lifetime with expected value τ (i.e., time to loss of all spacers targeting that virus in that array), in the absence of novel acquisitions the expected time to complete immune loss is $\tau \sum_{i=1}^{N} \frac{1}{i}$, where N is the number of arrays that initially target the virus in question. Clearly, the advantage conferred in terms of memory span decreases with each additional array, though this effect is important for the first few added arrays. In fact, it is more appropriate to model the lifetime of individual spacers with an exponential distribution such that the expected time to complete immune loss is $\tau_s \sum_{i=1}^{n} \frac{1}{i}$, where n is the total number of spacers in all arrays and τ_s the expected lifetime of each spacer. Note that we assume here that arrays are of comparable length (so that spacer loss rates remain constant). Thus the relative advantage of multiple arrays is further reduced in the case where each array can have multiple spacers targeting the same virus, assuming that spacer loss rates are similar across arrays (appropriate in the case of identical arrays near some equilibrium length).

If spacers vary in their effectiveness in attacking a viral target then we would expect this to increase the relative payoff of a bet-hedging strategy since it will essentially reduce the number of effective spacers in any given array. There is evidence that spacers vary in their targeting efficiency [234] in some systems. Nevertheless, if a system experiences priming then it is extremely likely that a single array would have many spacers towards the same target, making a bet hedging strategy less likely.

B.10 No evidence for array specialization

In genomes with multiple arrays, the dissimilarity between consensus repeat sequences of arrays in a single genome spanned a wide range of values (Levenshtein

Distance, Figs B.17 and B.18), though the mode was at zero (i.e., identical consensus repeats). When limiting our scope to only genomes with exactly two CRISPR arrays, we saw a bimodal distribution of consensus repeat dissimilarity, with one peak corresponding to identical arrays within a genome and the other corresponding to arrays with essentially randomly drawn repeat sequences except for a few conserved sites between them (S7D Fig). We also observed that among functional genomes, the area of the peak corresponding to dissimilar repeat sequences was significantly higher than among non-functional genomes ($\chi^2 = 61.432$, $df = 1$, $p < 4.582 \times 10^{-15}$, Fig B.17). This suggests that the observed signature of selection may be related to the diversity of consensus repeat sequences among CRISPR arrays in a genome. On the other hand, this enrichment of functional genomes with dissimilar arrays was not observed in an independently-generated dataset, calling this result into question (CRISPRdb [159], AppendixB.7, Fig B.18). Even when looking only at arrays with identical consensus repeats, there is a clear interaction between functionality and having multiple arrays, suggesting that selection maintaining multiple arrays is present even in these cases (Fig B.19).

Finally, we sought to assess if this observed variability in repeat sequences among arrays might have functional implications for CRISPR immunity, even when arrays share a set of *cas* genes. One measure of system functionality is array length, as we expect it to be correlated with the rate of spacer acquisition [226]. Therefore, we determined whether the mean pairwise dissimilarity between array consensus repeat sequences in a genome was associated with the variance of array lengths in that genome. Array length was measured as the number of repeats in an array.

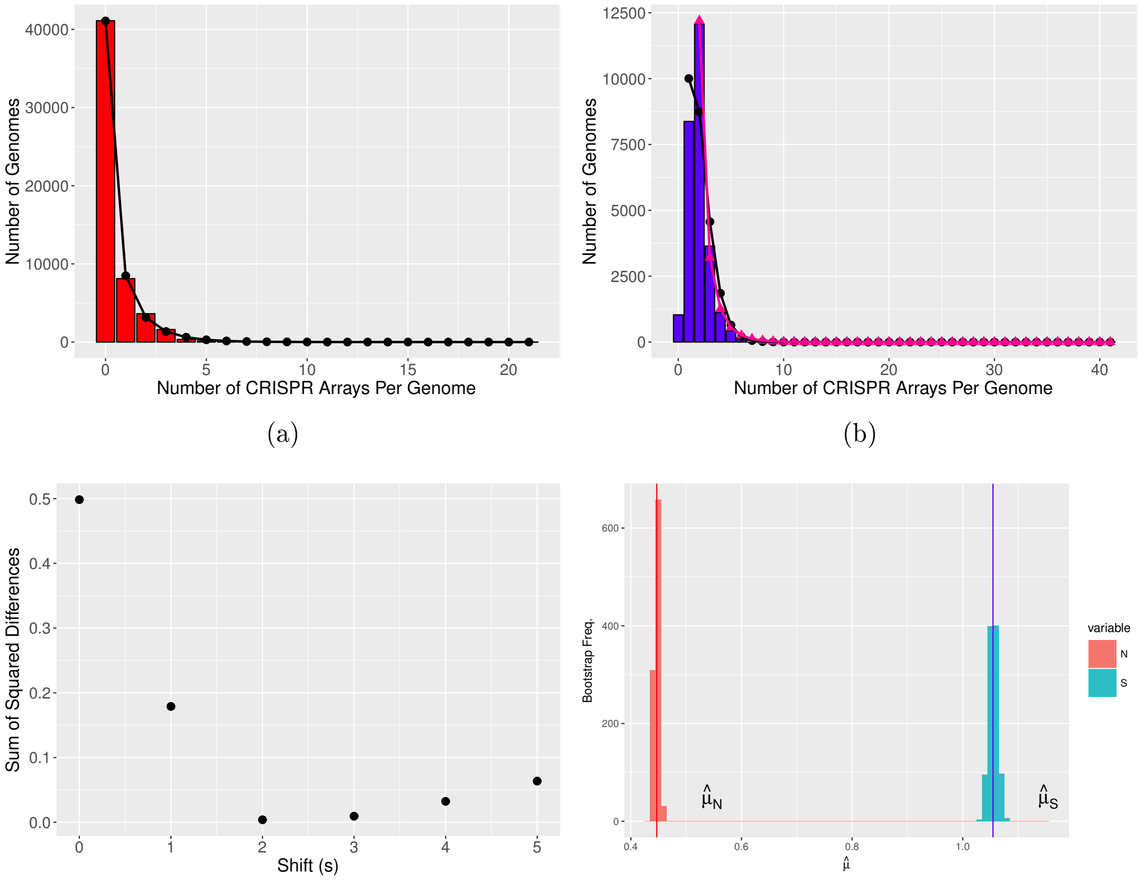
In genomes with exactly two arrays, the mean pairwise distance between consensus repeats within a genome was positively associated with the variance of the number of repeats across arrays in a genome, but this relationship was poorly predictive, not significant considering multiple testing, and likely spurious (linear regression, $R^2 = 0.002557$, $p = 0.0382$).

					Bootstrap		
Only ≤ 1 cas set	Sub-sampled	$\hat{\mu}_S$	$\hat{\mu}_N$	$\Delta\mu$	2.5%	97.5%	$s^\star$
No	No	1.56	0.46	1.09	1.07	1.12	2
No	Yes	2.26	1.13	1.13	1.05	1.22	2
Yes	No	1.05	0.45	0.61	0.59	0.62	2
Yes	Yes	1.26	1.07	0.18	0.11	0.26	1

Table B.1: Test for selection maintaining multiple arrays applied to different subsets of the RefSeq data. See Figs 3.1, B.1, B.6, and B.7.

Species	$\Delta\mu$
Staphylococcus aureus	1.15 ± 0.37
Klebsiella pneumoniae	0.76 ± 0.06
Shigella sonnei	0.72 ± 0.17
Listeria monocytogenes	0.67 ± 0.08
Mycobacterium tuberculosis	0.41 ± 0.05
Pseudomonas aeruginosa	0.35 ± 0.16
Campylobacter jejuni	-0.12 ± 0.05
Escherichia coli	-0.20 ± 0.04
Salmonella enterica	-0.54 ± 0.06

Table B.2: Species specific values of $\Delta\mu$ with bootstrapped 95% CIs.



(a)

(b)

(c)

(d)

Figure B.1: Dataset restricted to genomes with one or fewer sets of *cas* genes (one copy or less each of *cas1*, *cas2*, and a *cas* targeting gene). (a-b) Distribution of number of arrays per genome in (a) non-functional genomes and (b) functional genomes. In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and pink triangles to the double-shifted distribution ($s = 2$). (c) The optimal shift is where the differences between the two distributions is minimized. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show no overlap with $n = 1000$ samples drawn.

Symbol	Definition	Value
μ_A	Unprimed Spacer Acquisition Rate	varied Figs 3.2(a) and B.3, $\frac{\text{spacers}}{\text{infection}}$
μ_L	Per-Spacer Loss Rate	10^{-2} $\frac{1}{\text{spacers}\times\text{time}}$
v_T	Density Virus $T\times$Adsorption	10^2 $\frac{\text{infections}}{\text{time}}$
v_B	Density Virus $B\times$Adsorption	10^2 $\frac{\text{infections}}{\text{time}}$, varied in Fig 3.2(a)
p	Priming Factor	10^4, varied in Fig B.3

Table B.3: Definitions of relevant variables and parameters for CRISPR array model. “Infection” refers to the adsorption and injection of a phage into the host. Time is in units of “viral return intervals” (i.e. the amount of time the transient phage is absent from the system).

Symbol	Definition	Value
e	Resource Consumption Rate of Growing Bacteria	5×10^{-7} μg/mL
v	Maximum Bacterial Growth Rate	1.4 divisions/hr
z	Resource Concentration for Half-Maximal Growth	1 μg/mL
w	Flow Rate	0.3 mL/hr
A	Resource Pool	350 μg/mL

Table B.4: Definitions of relevant variables and parameters for autoimmunity model.

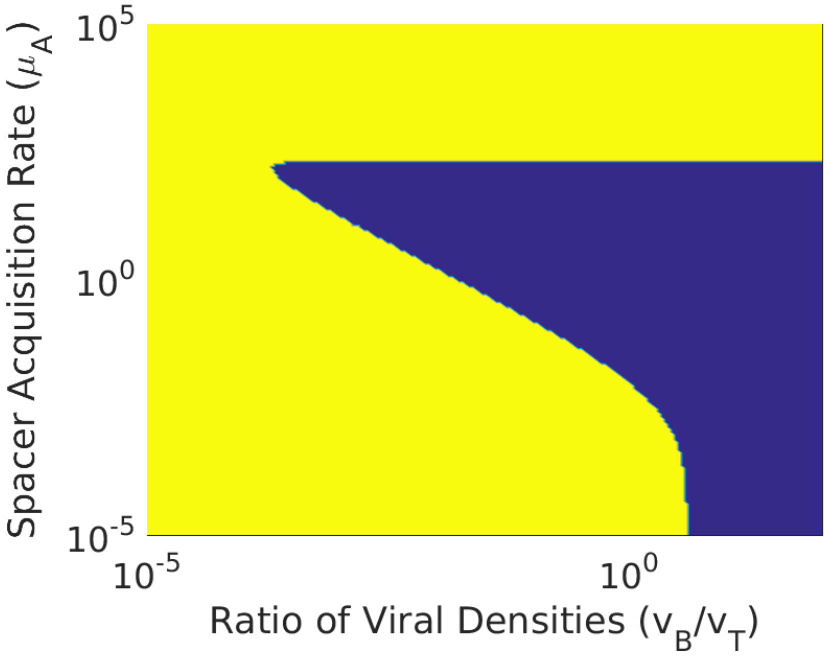


Figure B.2: Alternative model results with array length cap agree qualitatively with those of the primary model ($p = 1$ (no priming), $L = 5$, and $v_T = 100$). Blue signifies memory washout ($t_I \geq 10^{-5}$) and yellow signifies immune maintenance towards the transient virus, T ($t_I < 10^{-5}$).

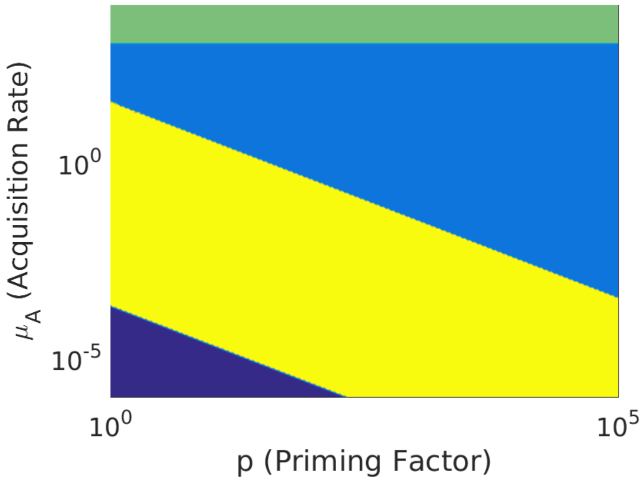
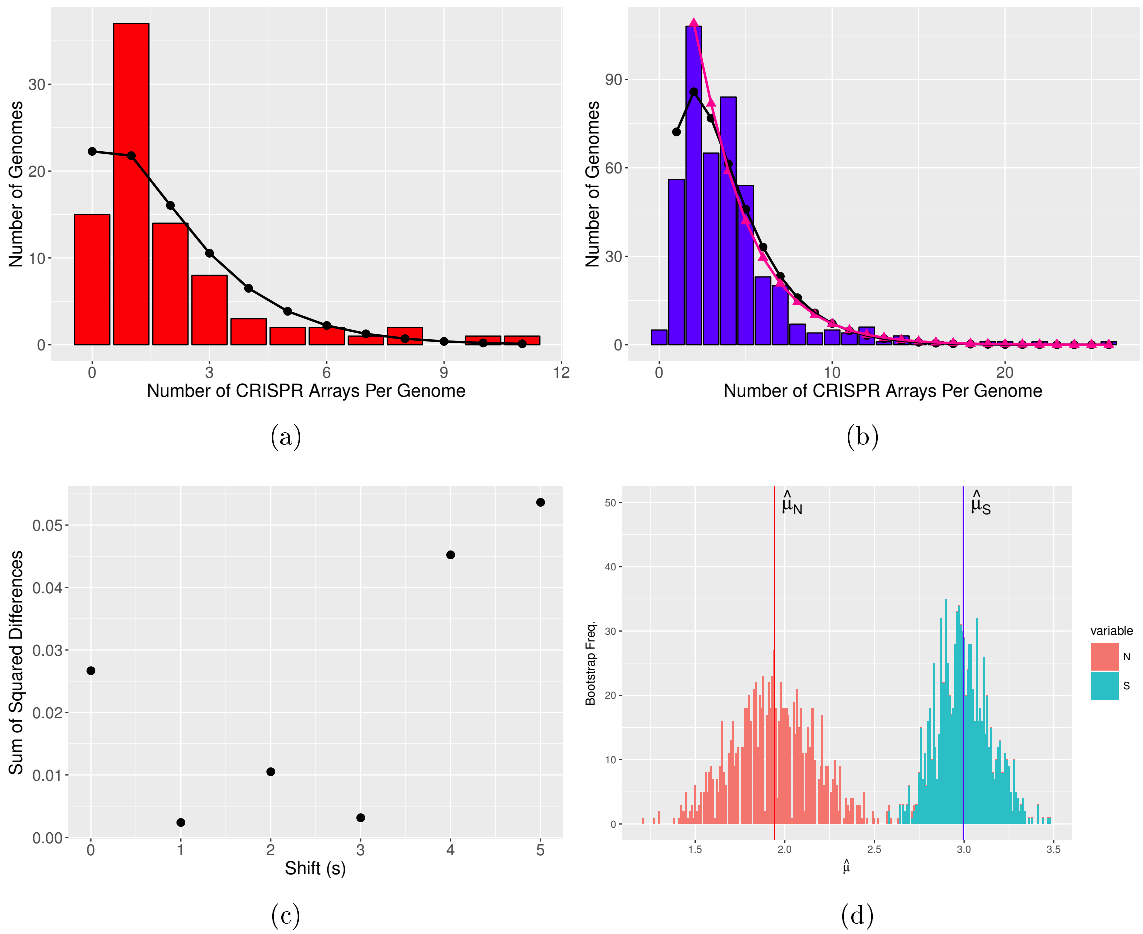


Figure B.3: Priming increases the region of memory washout and thus deepens the memory span versus acquisition rate tradeoff. Phase diagram of the behavior of our CRISPR array model with two viral species, a constant "background" population and a "transient" population that leaves and returns to the system at some fixed interval (Appendix B.4, Fig B.12). The yellow region indicates that immunity towards both viral species was maintained. The green region indicates where immune memory towards the transient viral species was lost, but reacquired almost immediately upon viral reintroduction. The light blue region indicates that only immunity towards the background species was maintained (i.e., immune memory towards the transient viral species was rapidly lost but not rapidly reacquired). Dark blue indicates where equilibrium spacer content towards one or both species did not exceed one despite both species being present in the system. The parameter p is the priming factor by which acquisition rate is increased when spacers towards a given target already exist in the array.



(a)

(b)

(c)

(d)

Figure B.4: Signature of multi-array selection in archaeal genomes. (a-b) Distribution of number of arrays per genome in (a) non-functional genomes and (b) functional genomes. In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and pink triangles to the double-shifted distribution ($s = 2$). (c) The optimal shift is where the differences between the two distributions is minimized. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show significant overlap with $n = 1000$ samples drawn. We note that the large majority of archaeal genomes had CRISPR arrays and were also functional, making our approach less powerful. Further, if those few non-functional genomes lost their *cas* spacer acquisition machinery recently, then our power would be reduced even more because these genomes might still bear the remnants of past selection. In general, as more archaeal genomes become available in public databases we will have more power to search for selection using out test, as currently there is an issue of small sample size overall.

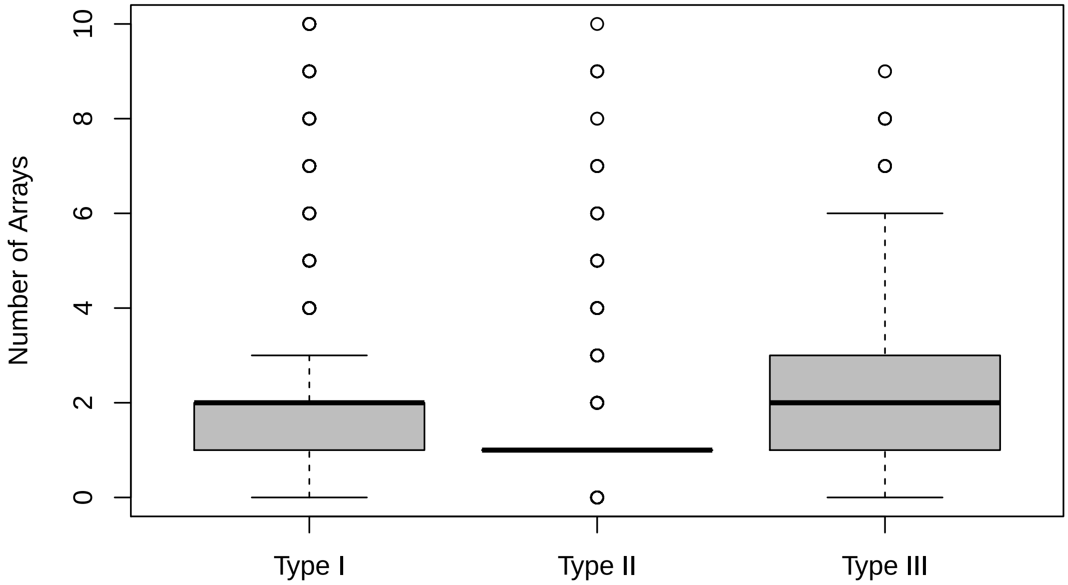
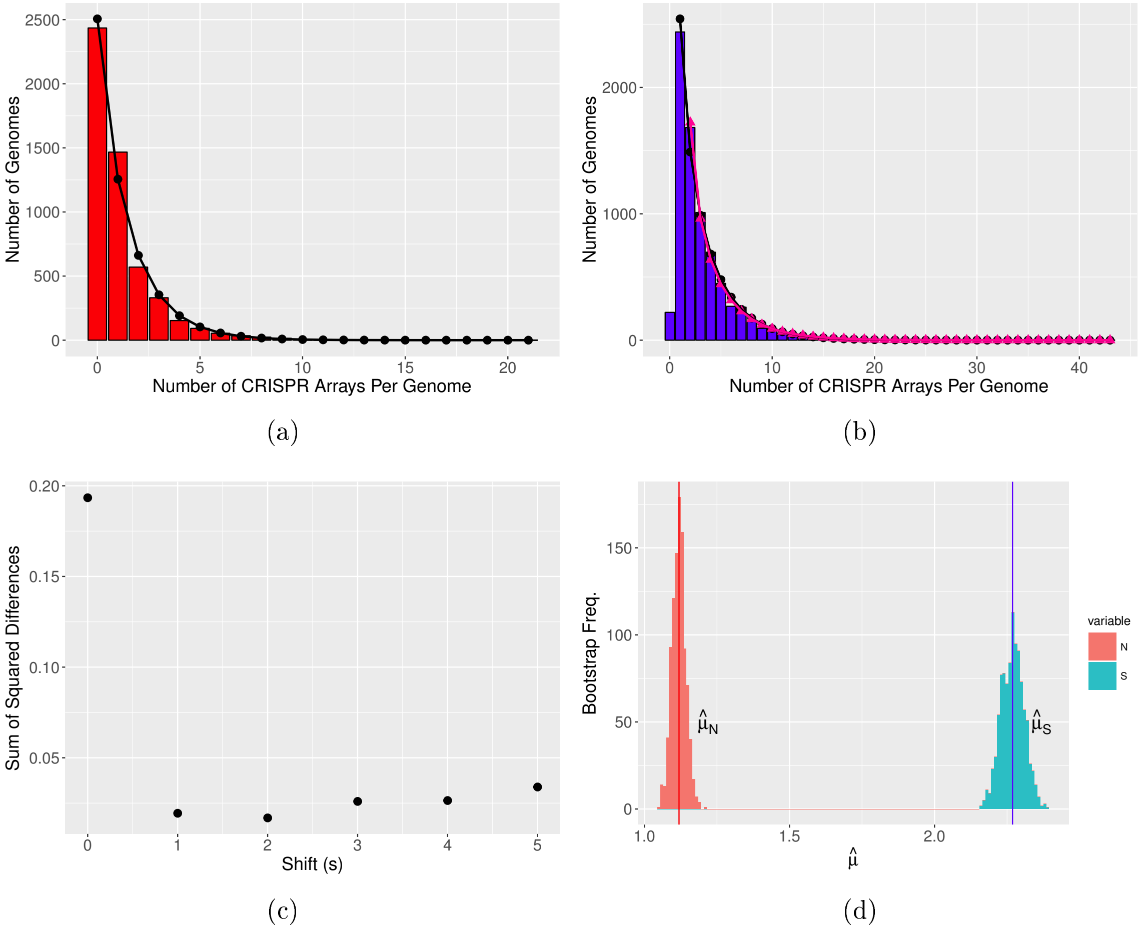


Figure B.5: Boxplots of array counts associated with genomes carrying a particular type of *cas* targeting machinery. System type was determined by the type of *cas* targeting gene found on the genome (genomes with no signature targeting genes or multiple types are excluded). Outlier points for genomes with > 10 arrays not shown for readability.



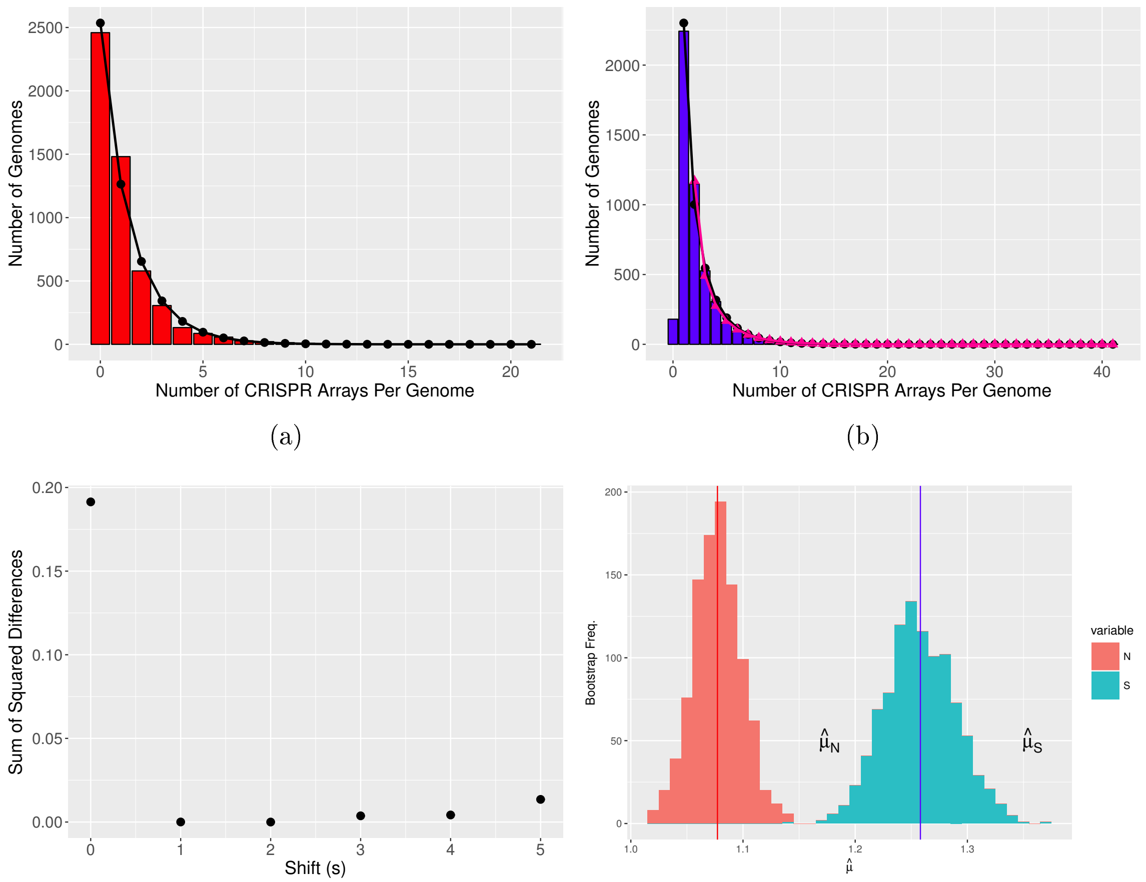
(a)

(b)

(c)

(d)

Figure B.6: Dataset with subsampled genomes of overrepresented taxa. (a-b) Distribution of number of arrays per genome in (a) non-functional genomes and (b) functional genomes. In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and pink triangles to the double-shifted distribution ($s = 2$). (c) The optimal shift is where the differences between the two distributions is minimized. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show no overlap with $n = 1000$ samples drawn.



(a)

(b)

(c)

(d)

Figure B.7: Dataset restricted to genomes with one or fewer sets of *cas* genes (one copy or less each of *cas1*, *cas2*, and a *cas* targeting gene) with subsampled genomes of overrepresented taxa. (a-b) Distribution of number of arrays per genome in (a) non-functional genomes and (b) functional genomes. In (a) the black circles show the negative binomial fit to the distribution of arrays in non-functional genomes. In (b) black circles indicate the negative binomial fit to the single-shifted distribution ($s = 1$) and pink triangles to the double-shifted distribution ($s = 2$). (c) The optimal shift is where the differences between the two distributions is minimized. (d) The bootstrapped distributions of the parameter estimates of $\hat{\mu}_S$ and $\hat{\mu}_N$ show nearly no overlap with $n = 1000$ samples drawn.

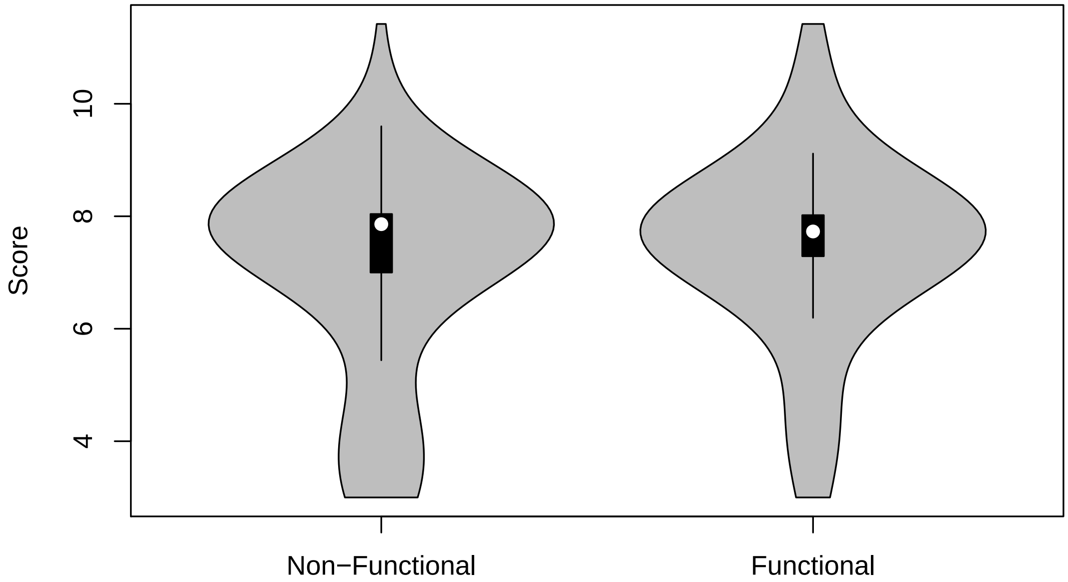
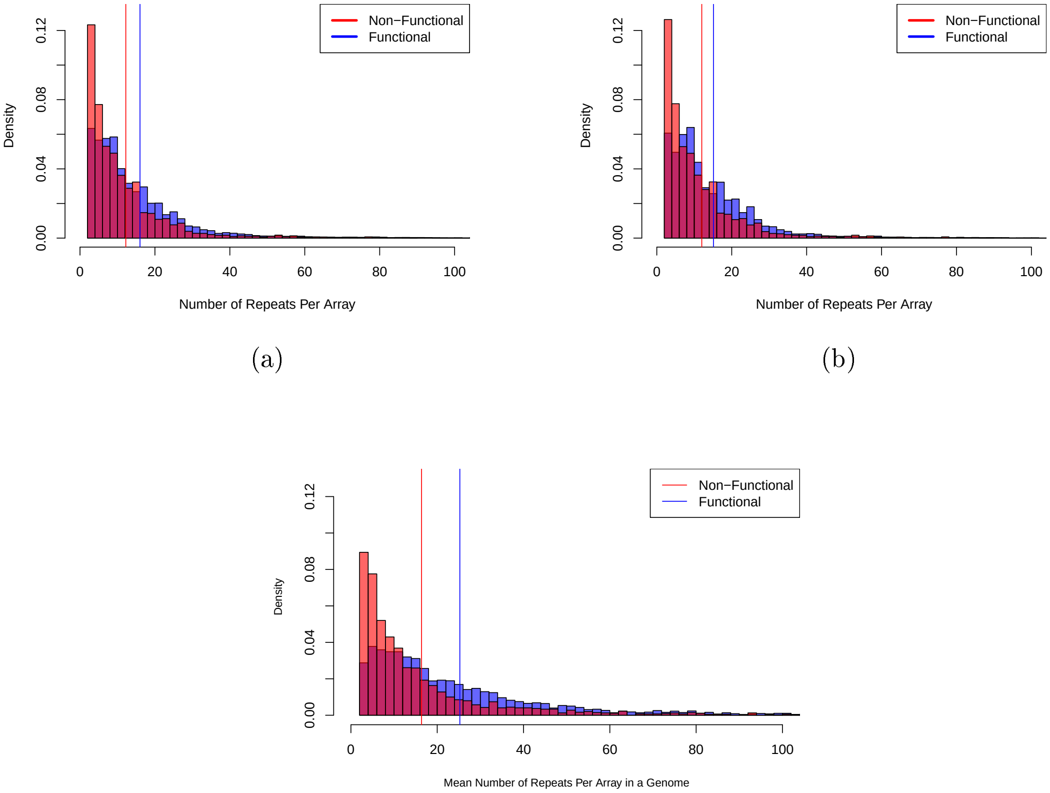


Figure B.8: Similar CRISPRDetect score distributions in non-functional and functional arrays. Non-functional arrays have slightly higher scores (Wilcox rank-sum test, $p = 0.01254$), although the effect size is marginal and statistical significance at this level is questionable after considering the number of tests conducted in this study.

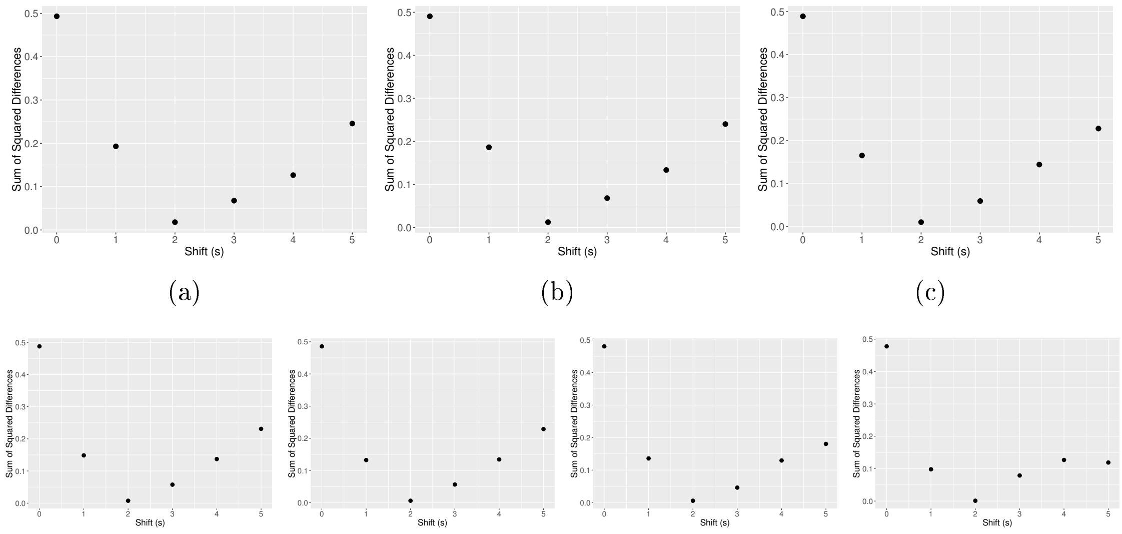


(a)

(b)

(c)

Figure B.9: Arrays in functional genomes are longer on average than arrays in non-functional genomes (t-test, $p < 2 \times 10^{-16}$). (a) Full dataset. (b) Data from genomes with one or fewer sets of *cas*. (c) Functional genomes tend to have more repeats on average in their CRISPR arrays than non-functional genomes (array length first averaged over arrays in each genome, so each datapoint is a genome rather than an array). (a,b,c) In blue is the distribution of mean array length in functional genomes. In red (overlaid) is the distribution of mean array length in non-functional genomes. Vertical lines with corresponding colors indicate the means of these distributions.



(a) (b) (c)

(d) (e) (f) (g)

Figure B.10: Short, functional arrays do not drive the result of our non-parametric test for selection. The results for our non-parametric test for selection when removing all functional arrays shorter than (a) 4 repeats, (b) 5 repeats, (c) 6 repeats, (d) 7 repeats, (e) 8 repeats, (f) 9 repeats, and (g) 10 repeats. Note that in all cases the test indicates selection maintaining two arrays.

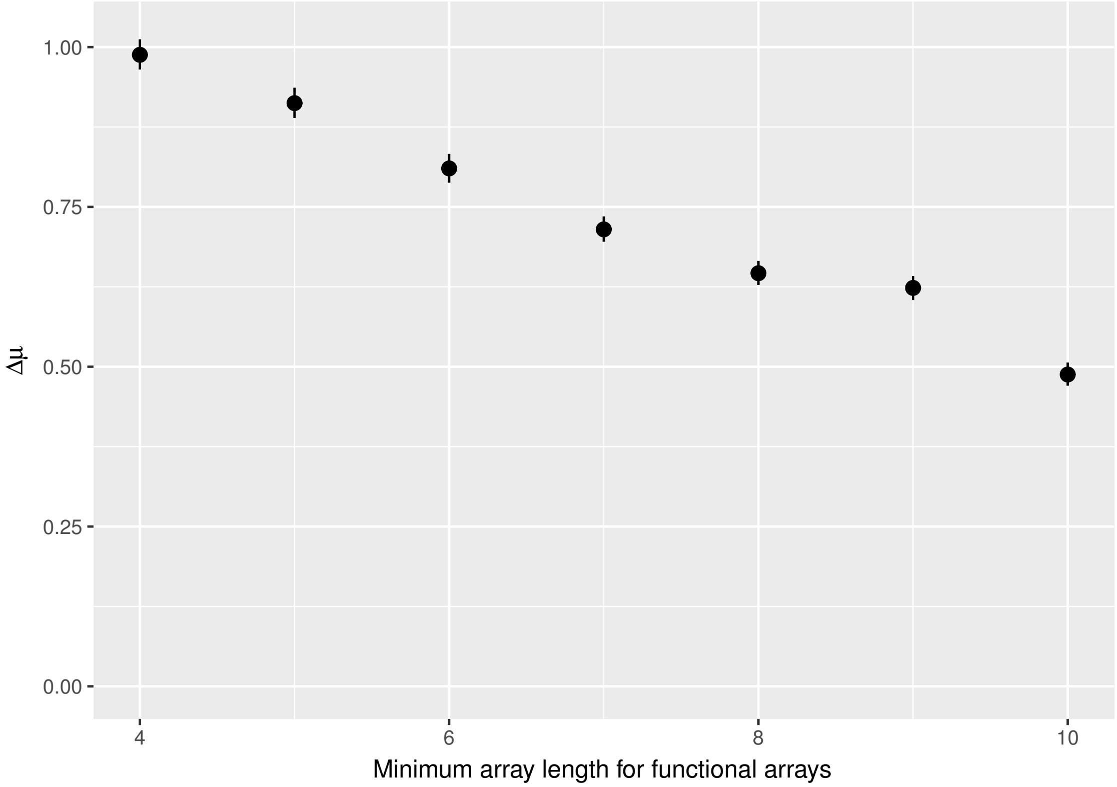


Figure B.11: Short, functional arrays do not drive the result of our parametric test for selection. The results for our parametric test for selection when removing all functional arrays shorter than a given threshold. Note that even when removing all functional arrays less than 10 repeats long we still see a substantial signature of selection maintaining multiple arrays, and that up to a threshold of 6 repeats this signature is rather strong. By the nature of this test, $\Delta\mu$ must decrease monotonically as the threshold increases (Appendix B.3).

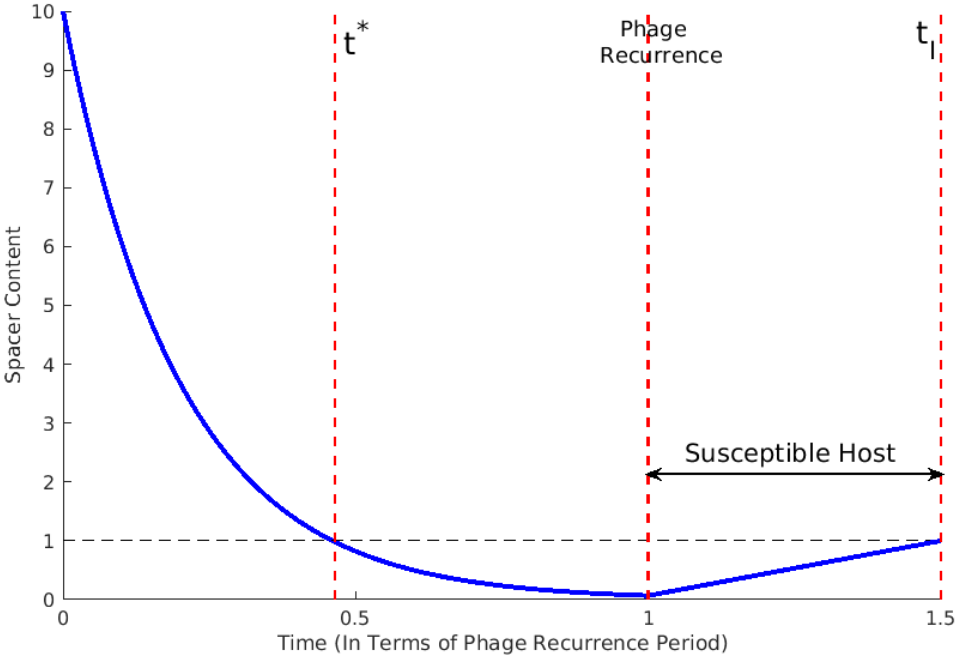


Figure B.12: Outline of model analysis. Hypothetical time-course of C_T during the departure and return of virus T from the system. Virus T leaves at time $t = 0$ and returns at time $t = 1$. C_T exceeds a value of one after viral reintroduction at time $1 + t_I$.

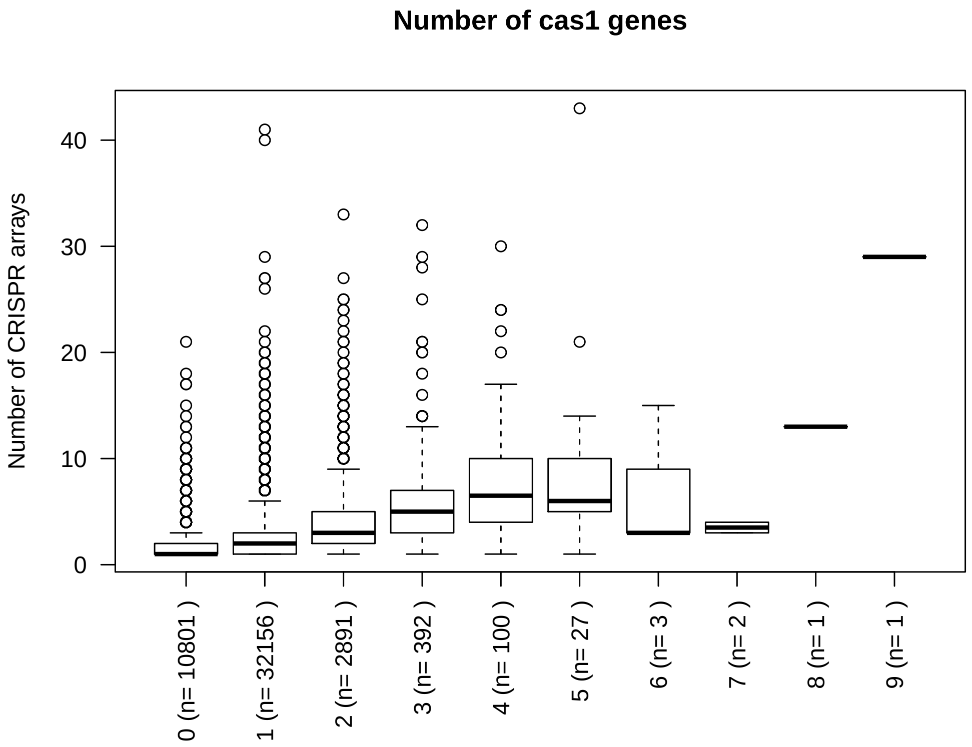


Figure B.13: Relationship between the number of *cas1* genes and the number of CRISPR arrays in a genome.

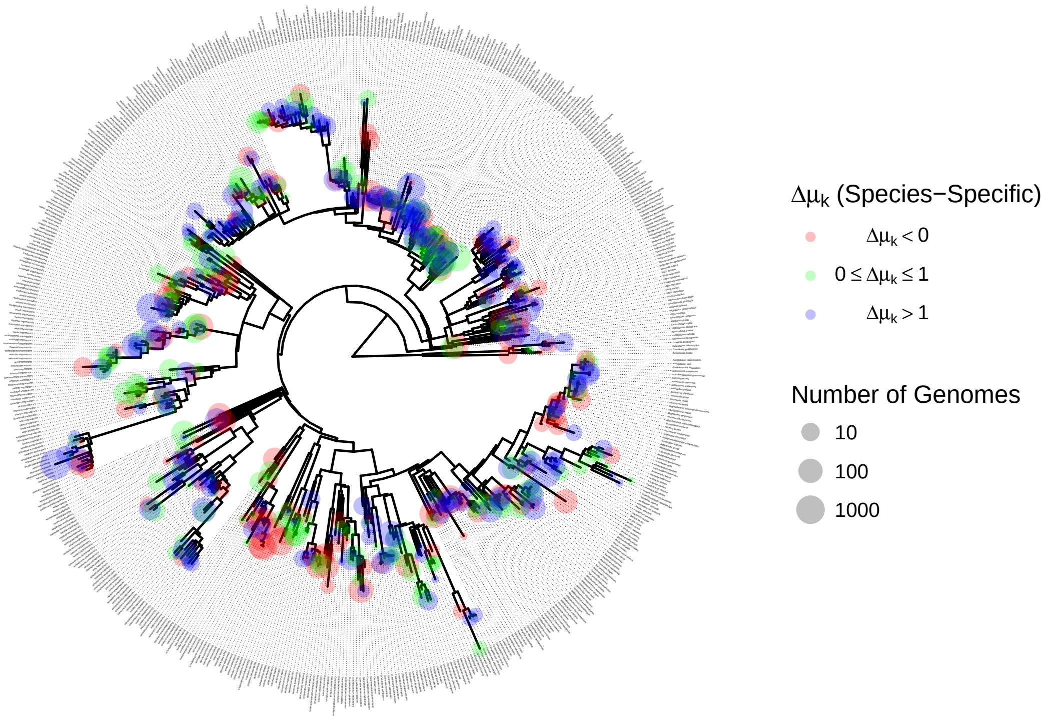
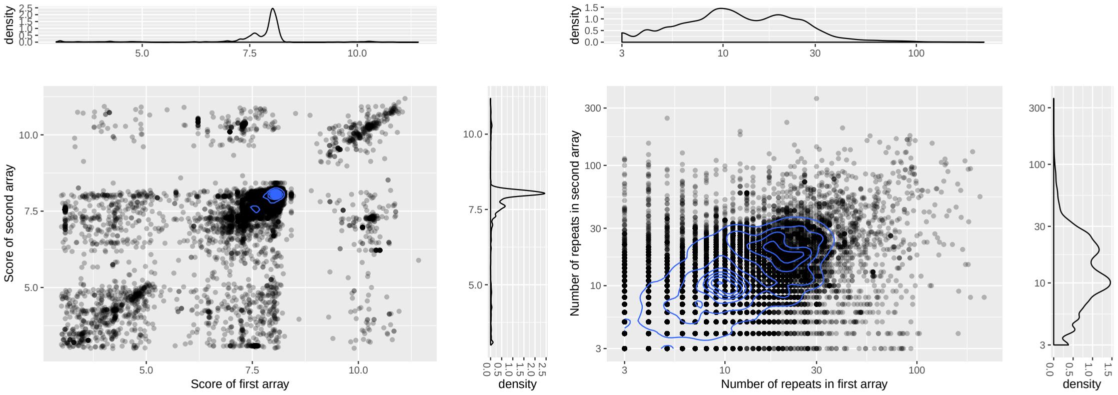


Figure B.14: Species-specific $\Delta\mu_k$ values mapped onto the SILVA Living Tree 16s rRNA Tree. In red are species experiencing apparent selection against having a functional CRISPR array. In blue are species showing a strong signature of selection for multiple arrays. Number of genomes is represented by tip size, and is a rough indicator of power.



(a) (b)

Figure B.15: In two-array genomes there is a slight positive association in both array score and array length between arrays within the same genome. (a) Most array scores are centered around 8, but there is some positive association (linear regression, $m = 0.49$, $p < 2 \times 10^-16$, $R^2 = 0.2387$). (b) A similar relationship holds for length (linear regression, $m = 0.39$, $p < 2 \times 10^-16$, $R^2 = 0.1458$).

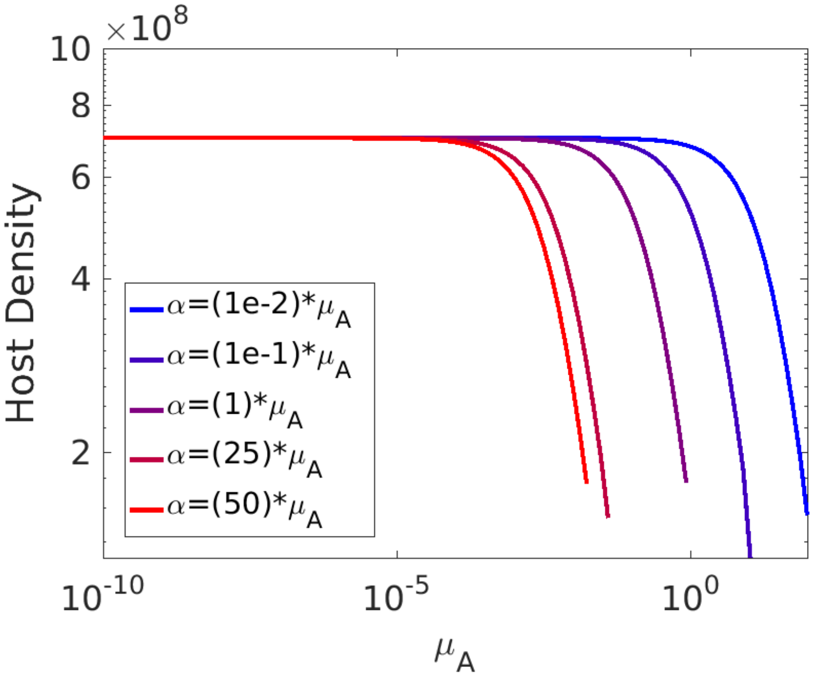
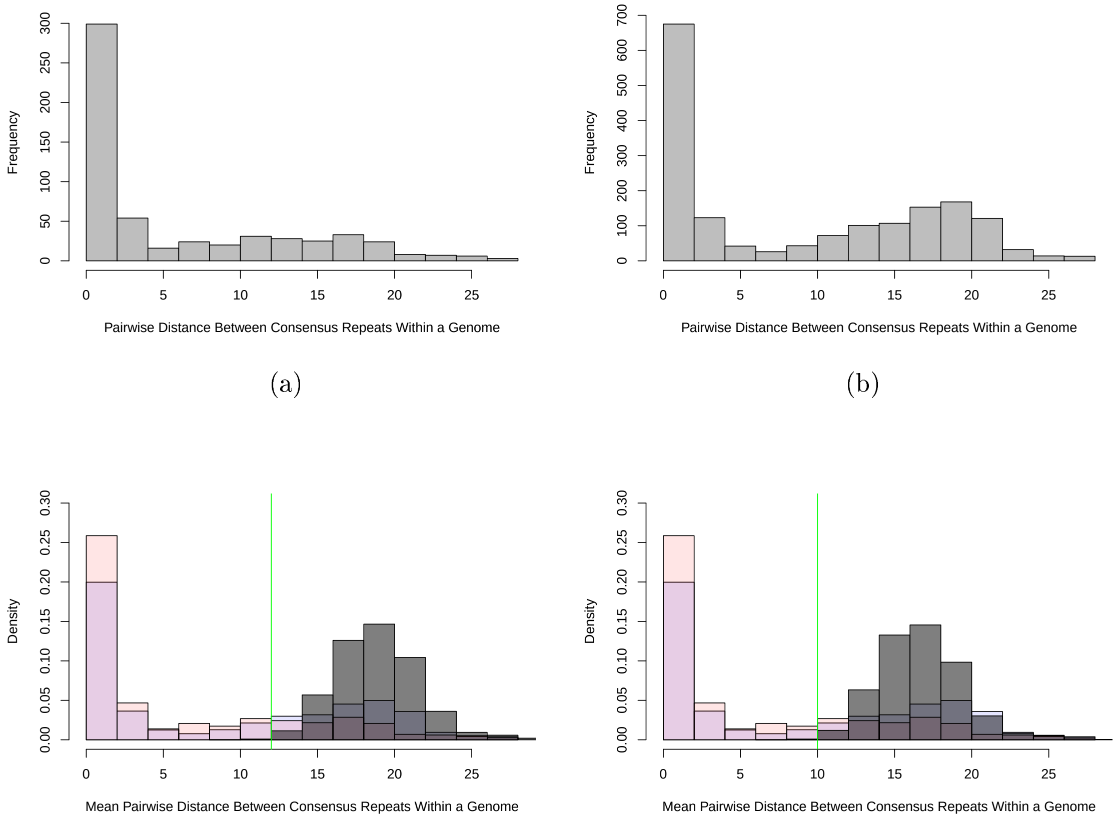


Figure B.16: Equilibrium host density values from autoimmunity model (Appendix B.8) over varying spacer acquisition rates. Curves end because for extremely high α the equilibrium no longer exists if we restrict both host density and resource concentration to be non-negative.



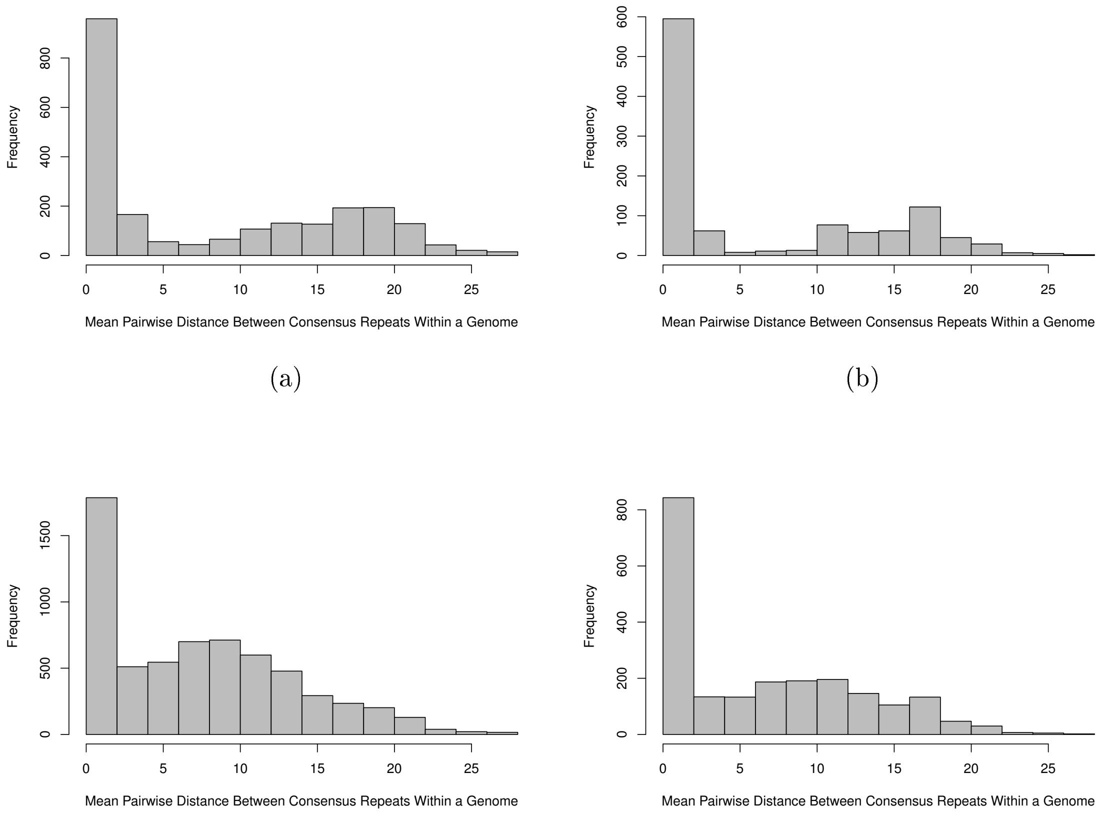
(a)

(b)

(c)

(d)

Figure B.17: Pairwise distance between consensus repeats from arrays within a genome (only genomes with two arrays shown). Distance calculated as Levenshtein Distance between each pair divided by the length of the longest repeat in the pair. (a) Non-functional genomes. (b) Functional genomes. genomes. The functional genomes are enriched with dissimilar arrays ($\chi^2 = 26.406$, $df = 1$, $p = 2.766 \times 10^{-7}$, dissimilarity cutoff at 3 based on median across all two-array genomes). (c) The distributions of pairwise differences between arrays in functional genomes (blue, a) and non-functional genomes (red, b) are overlaid alongside a distribution of distances between random sequences with lengths drawn from the empirical distribution of repeat lengths in the full dataset (gray). The green line indicates the bottom 0.1% of this random (gray) distribution, which can be used as an alternative similarity cutoff ($\chi^2 = 54.653$, $df = 1$, $p = 1.505 \times 10^{-13}$). (d) Same as (c) but with 4 bases held constant across all repeats to simulate some degree of universal sequence conservation at one end of the repeat as observed among type II-A CRISPR systems [235] ($\chi^2 = 64.168$, $df = 1$, $p = 1.142 \times 10^{-15}$). In (c,d) the simulated distribution takes into account the overall frequency of each base across repeat sequences. Histograms drawn using default settings of hist() function in R (right-closed/left-open intervals, except for the first interval which includes the lower bound, i.e. zero).



(a) (b)

(c) (d)

Figure B.18: Consensus repeat diversity across datasets (CRISPRDetect left vs. CRISPRdb right) in two-array genomes (a,b) and all genomes (c,d). (a) Pairwise distance between consensus repeats from arrays within a genome (only genomes with exactly two arrays shown). Distance calculated as Levenshtein Distance between each pair. Histogram drawn using default settings of hist() function in R (right-closed/left-open intervals, except for the first interval which includes the lower bound, i.e. zero). (b) The same as (a) for the CRISPR Database dataset. (c,d) Mean pairwise distance between consensus repeats from all arrays in a genome, including genomes with more than two arrays, for (c) the CRISPRDetect and (d) the CRISPR Database datasets. While the distribution of pairwise-distances between repeat sequences in two-array genomes was approximately the same shape as that we observed for both datasets, the relationship between diversity and functionality was reversed in the CRISPR Database dataset, with non-functional genomes having more diverse consensus repeats among their arrays (χ^2 = 4.3952, df = 1, p = 0.03604). This opposing result calls into question the potential link between selection on multiple arrays and consensus repeat diversity observed in the CRISPRDetect data, though this may be due to the smaller size of the CRISPR Database dataset.

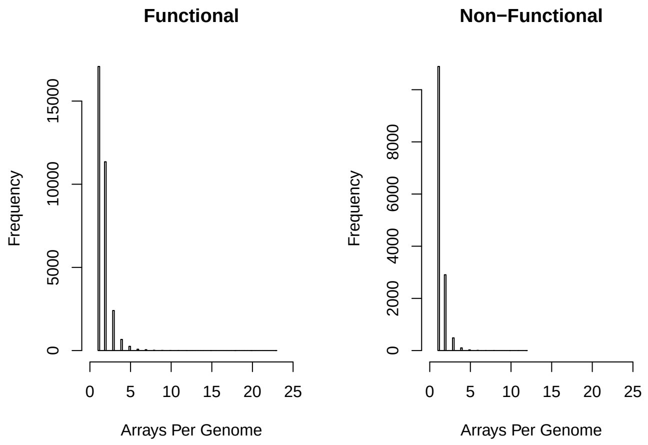


Figure B.19: Even restricting to arrays with identical consensus repeats, functional genomes are more likely to have multiple CRISPR arrays. For each genome with CRISPR we counted the incidence of each unique consensus repeat sequence and retained only arrays with the most common sequence (choosing one at random in the case of a tie, which has no effect on the outcome). We then plotted the frequency of arrays per genome for this dataset, and found a clear excess of 2-array vs. 1-array genomes in the functional category as compared to the non-functional category ($\chi^2 = 1475.9$, $df = 1$, $p < 2.2 \times 10^{-16}$).

Appendix C: Supplemental Information For: "Immune Loss as a Driver of Coexistence During Host-Phage Coevolution"

C.1 Parameter Values

For both our analytical and simulation models we attempt to constrain parameter values within realistic ranges where possible. Resource uptake parameters (e, v, z) and burst size (β) are taken from Levin et al. [236], although they are rough estimates in some cases. We let $\delta = 10^{-8}$ be our base value for the rate of adsorption. Levin et al. [236] fit a model to data to estimate a value of $\delta = 10^{-7}$, but they also incorporate a lag time, which we approximate by lowering δ tenfold.

C.2 Alternative Costs of Immunity

While autoimmunity represents one class of costs that may be associated with prokaryotic immune systems (i.e. lethality via an additional death term), other cost structures exist that may be applied to growth. Immune host may either suffer from reduced resource affinity (z) or maximal growth rate (v). Thus we can write the chemostat system with resources:

$$\dot{R} = w(A - R) - \frac{evR}{z + R}(D + U) \tag{C.1}$$

defended host:

$$\dot{D} = D\left(\frac{v_d R}{z_d + R} - \delta\phi_d P - \alpha - \mu - w\right), \tag{C.2}$$

undefended host:

$$\dot{U} = U\left(\frac{v_u R}{z_u + R} - \delta\phi_u P - w\right) + \mu D, \tag{C.3}$$

and phage:

$$\dot{P} = P\left(\delta U(\phi_u\beta - 1) + \delta D(\phi_d\beta - 1) - w\right), \tag{C.4}$$

where we let

$$z_d = c_z z_u \tag{C.5}$$

and

$$v_d = \frac{v_u}{c_v}. \tag{C.6}$$

so that the resource affinity penalty, c_z, and growth rate penalty, c_v, describe the costs applied to each aspect of host population growth respectively. It is possible that alternative cost regimes are more capable of producing robust coexistence under realistic parameter ranges than is autoimmunity (α).

Both alternative cost regimes can produce stable coexistence (Fig C.1), although they are applied to a nonlinear term in the growth equations and thus behave differently than autoimmunity (Fig C.2). We see that under realistic initial conditions immune loss can produce coexistence over a wider range of parameter space than the other mechanisms, but all can do so over some range and the range of values for μ and α are not easily comparable with the range for c_z and c_v (Fig C.3). All mechanisms can produce coexistence with initial conditions perturbed away from the equilibrium condition (Figs C.4-C.7), but only immune loss reliably produces coexistence over any part of parameter space when very large perturbations to the system occur (on the order of those expected with serial dilution; Fig C.7). We also note that the condition we call coexistence in Figs C.4-C.7 is that both immune hosts and phages are present at 80 days at a level likely to be detected experimentally (density of 100/mL) which can be taken as the most general requirement for coexistence. If we observe the distribution of outcomes of these simulations in Fig C.8 we see that growth rate and resource affinity costs tend to produce coexistence regimes that are dominated by susceptible hosts even with more mild perturbations (though still severe).

C.2.0.1 Simulation Parameters

The rate of loss of functionality in the CRISPR immune system of *S. epidermidis* has been shown to fall in the range $10^{-4} - 10^{-3}$ losses per individual per generation [51]. We choose to use a value in the middle of this range ($\mu_L = 5 \times 10^{-4}$) for our simulations. Similarly, based on the fact that there appears to be no self vs. non-self recognition in the CRISPR system of *S. thermophilus* [27], and that the genome of S. thermophilus is roughly 50 times the size of its lytic phage 2972

[38, 233], the rate of incorporation of a self-spacer by the CRISPR system should be approximately 50 times the spacer acquisition rate per adsorbed phage. Assuming that incorporation of a self-spacer is instantaneously toxic and leads to unavoidable cell death, we can take this value as our rate of autoimmunity ($\alpha = 50\mu_b$). This gives us a value on the high end of possible α values, as it is possible that there is self recognition that has not been observed experimentally or that the action of autoimmunity can be delayed or avoided through spacer loss or corruption, as has been found in some experiments [26, 51].

We set $n_s = 10$ due to computational constraints, as small increases in cassette length lead to a large increase in computational time. The experiments we compare our models to saw small expansions of the CRISPR cassette (2-3 spacers) and our system reaches either a phage-cleared or stable coexistence state where coevolution is halted well before host have obtained the complete set of spacers. While the phage genome has many protospacers (200+), in the *S. thermophilus*-phage 2972 system the large majority of spacers come from a small subset of possible protospacers on the phage genome (~ 30), and thus our limited protospacer set may be an appropriate approximation of reality [26]. Our value for the cost per PAM mutation (c) is set arbitrarily, but is linked to the value we choose for n_s. Our results are robust to the value of this parameter (Fig C.17).

Childs et al. [214] uses a value of $\mu_p = 5 \times 10^{-7}$ for the protospacer mutation rate in their model of CRISPR-phage coevolution. We choose to introduce newly mutated strains at a population size of 100 individuals to eliminate the effects of drift in our model and thus speed up our simulations, which requires a corresponding decrease in the mutation rate. Additionally, we choose to consider PAM mutations only, and since the PAM region is considerably shorter than the protospacer itself, this also warrants a decrease in mutation rate. Thus we reduce their parameter estimate tenfold for use in our simulations ($\mu_p = 5 \times 10^{-8}$). Similarly we reduce previous estimates of the spacer acquisition rate ($\mu_b = 10^{-6}$) fivefold to account for our introduction of novel strains at a higher population size [19, 210, 211, 214, 236].

We use an initial multiplicity of infection (MOI) of 1 phage per host corresponding to experimental values. We simulated outcomes with an initial MOI of 10 to confirm robustness to initial conditions (Fig C.20), as seen in previous experimental work [150]. Burst size (β) estimates for phage are imprecise. We ran additional simulations at high and low burst sizes to confirm that our qualitative results are robust to changes in this parameter (Fig C.21).

C.2.0.2 Varying adsorption

Because we do not have a good estimate of the rate of adsorption in this system [236], and because we choose to depress our adsorption rate as a compensation for the lack of latent period in our models, we explore the response of our results to large variations in δ (Figs C.14 and C.18).

C.3 Analysis and Simulation Methods

C.3.0.1 Analysis

W found equilibria of our analytical model using Wolfram Mathematica (Version 11.0, Wolfram Research Inc., Champaign, IL, 2016). We assessed stability by linearizing the system around each equilibrium point via the Jacobian. We performed robustness analysis by solving our system numerically using a variable order method for stiff systems (MATLAB 9.0, The MathWorks Inc., Natick, MA, 2016; `ode15s`) at 80 days from inital conditions described in Main Text Fig. 4.1.

C.3.0.2 Simulations

We solve our system numerically using a variable order method for stiff systems (MATLAB 9.0, The MathWorks Inc., Natick, MA, 2016; `ode15s`), pausing the solver to add strains due to spacer acquisition and PAM mutation events, and to perform serial dilutions at 24 hour intervals. When we reach a serial dilution the resource concentration is reset to its initial value and all populations are reduced by a factor of 100. Spacer acquisition and PAM mutation rates are updated at the next strain addition or dilution event or at a preset maximum update interval ($\frac{1}{2}$hr) and used to draw the time of next addition from an exponential distribution.

When an addition event occurs the type of event is drawn with each type's probability being proportional to its calculated rate. We then draw the strain in the population to serve as the base for the new strain, with the probability of choosing each strain proportional to its strain-specific acquisition, mutation, or recombination rate. In the case of spacer acquisition a spacer is drawn with probabilities based on the overall prevalence of each corresponding protospacer in the phage population. In the case of PAM mutation or back mutation a protospacer is drawn uniformly from the chosen strain's genome. In all cases new strains are added at a population size of 100 individuals so that we focus only on strains that are able to establish themselves and neglect drift to speed up computation. Accordingly we set μ_b, μ_p, and μ_q lower than might otherwise be expected (see Appendix C.1). During simulations we dynamically adjust our rates so as to avoid adding strains already present in the system.

C.4 Experimental Methods

Powdered skim milk (Publix) was diluted in distilled water, 10gms/100ml water. The suspension was autoclaved at 110°C for 12 minutes. Three ml of the milk was then put into 13mm x100mm glass tubes. *Streptococcus thermophilus* (DGCC7710) were grown overnight in a broth, LM17 with added calcium [236]. To initiate the serial transfer cultures, 30 μl of the overnight broth was added to the tubes either alone or with the phage from an LM17 Ca lysate. The initial densities of the bacteria and phage were estimated by serial dilution and plating on LM17Ca agar for the bacteria and with LM17Ca soft agar for the phages (see [236]). Each day, the cultures were vortexed to suspend the bacteria and phages (the milk fermented), densities were estimated, and 30 μl of the cultures were transferred to fresh tubes with 3 ml milk. These cultures were serially passaged for the noted number of days, with bacteria and phage densities estimated daily.

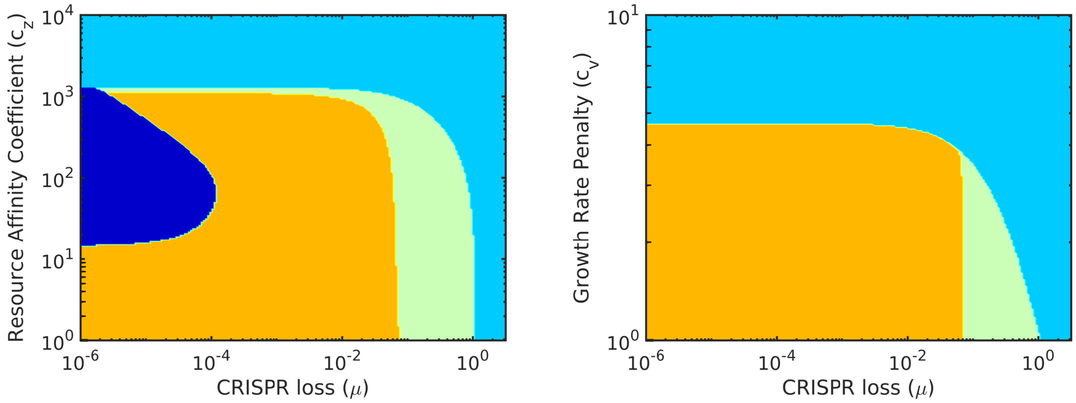
To test for bacteriophage immune mutants, periodically, single colonies from the sampling plates were grown up on LM17 and used as lawns to test for their sensitivity to the original phage and the phage in their respective cultures. For the latter, LM17 lysates were made from single plaques taken from the sampled plates. In this way, we were able to test for CRISPR escape mutants. For example, if the bacteria from the culture appeared immune to the original phage, we would then test for its sensitivity to phage from the serial passage culture. In this way, we were able to follow some of the co-evolution that was occurring in the cultures, via the acquisition of new spacers generating host and phage mutants evolving in these cultures, for more details and more extensive consideration of this co-evolution see [26, 150].

	Eq. 1	Eq. 2	Eq. 3	Eq. 4
$\tilde{R}$	A	$\frac{z(\alpha+\mu+w)}{v-\alpha-\mu-w}$	$\frac{1}{2}\left(A-\frac{ev\tilde{U}}{w}-z+\sqrt{4Az+\left(A-\frac{ev\tilde{U}}{w}-z\right)^2}\right)$	$\frac{zw}{v-w}$
$\tilde{D}$	0	$(\phi_u\beta-1)\tilde{U}-\frac{w}{\delta}$	0	0
$\tilde{U}$	0	$\frac{1}{\phi_u\beta}\left(\frac{w(A-\tilde{R})(z+\tilde{R})}{ev\tilde{R}}+\frac{w}{\delta}\right)$	$\frac{w}{\delta(\phi\beta-1)}$	$\frac{w(A-\tilde{R})(z+\tilde{R})}{ev\tilde{R}}$
$\tilde{P}$	0	$\frac{1}{\phi_u\delta}\left(\frac{v\tilde{R}}{z+\tilde{R}}-w+\mu\left(\frac{\tilde{D}}{\tilde{U}}\right)\right)$	$\frac{1}{\phi\delta}\left(\frac{v\tilde{R}}{z+\tilde{R}}-w\right)$	0

Table C.1: Equilibrium of general model without coevolution. Equilibria for the autoimmunity/locus-loss model where $\mu > 0$ and $\alpha > 0$. Not all substitutions have been made here in the interest of readability, but equilibrium values can be easily computed using the above expressions. The stability of these equilibria can be assessed by linearizing our system around them (i.e., taking the Jacobian) as described in Appendix C.3.

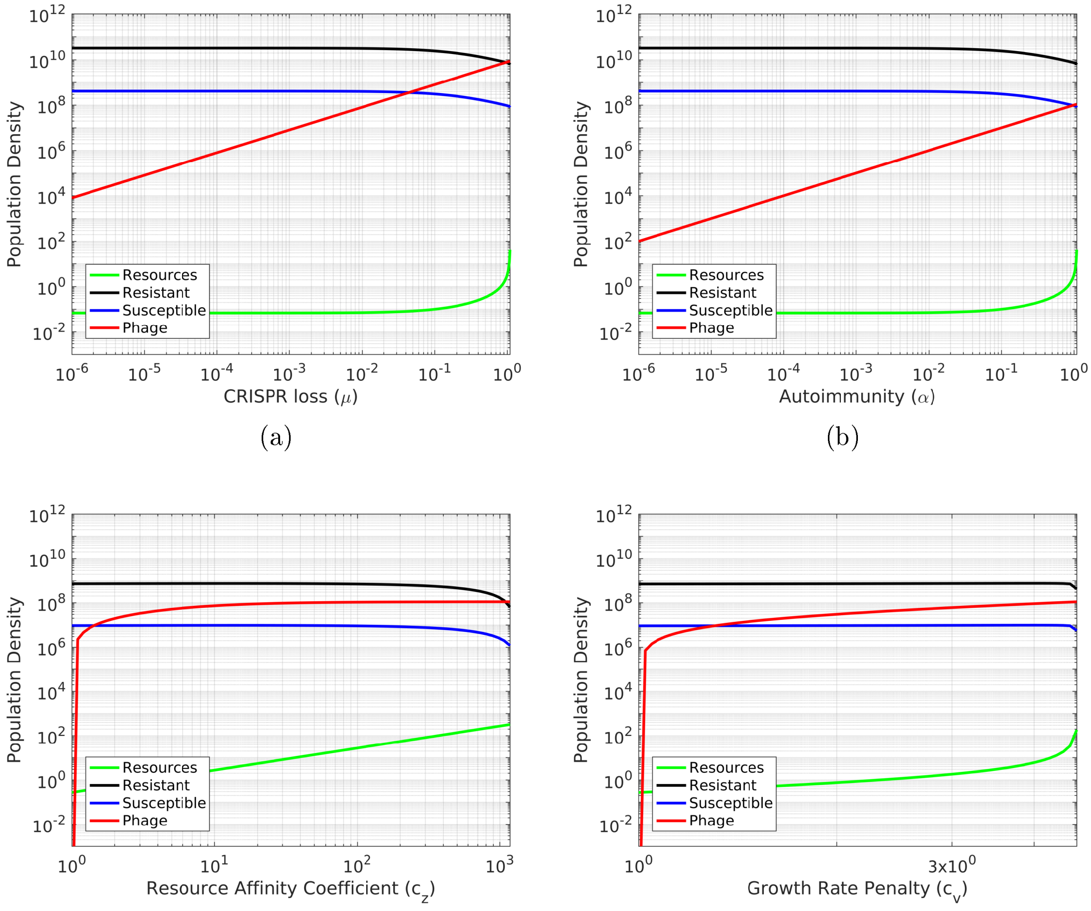
	Spacer ID																					
Transfer	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	1	1	1	1																		
2			1	1	1	1																
3			1	2		1	1															
4								1														
5					2	2	1		1	1												
11				1	1		2				1	1	1									
15				1		1							2	2	1							
25					5				1							1	1					
35					1		2											1	2			
40							2							2						3	1	1

Table C.2: Spacer dynamics for long term coevolution experiment (Experiment 1). Spacers dynamics in the CRISPR1 locus for serial transfer experiment 1. Each letter corresponds to a unique spacer sequence. All sampled sequences shown with the number of times observed at each timepoint.



(a) $c_v = 1$ (b) $c_z = 1$

Figure C.1: Equilibria with alternative costs of immunity. Model behavior under variations in the immune system loss rate and (a) resource affinity coefficient or (b) growth rate penalty. Equilibria derived from our equations in Appendix C.2 are shown where orange indicates a stable equilibrium with all populations coexisting and defended host dominating phage populations, green indicates that all populations coexist but phages dominate, light blue indicates that defended bacteria have gone extinct but phages and undefended bacteria coexist, and dark blue indicates that there is no stable equilibrium. We neglect coevolution and innate immunity in this analysis ($\phi_u = 1$, $\phi_d = 0$) and do not consider the effects of autoimmunity ($\alpha = 0$).



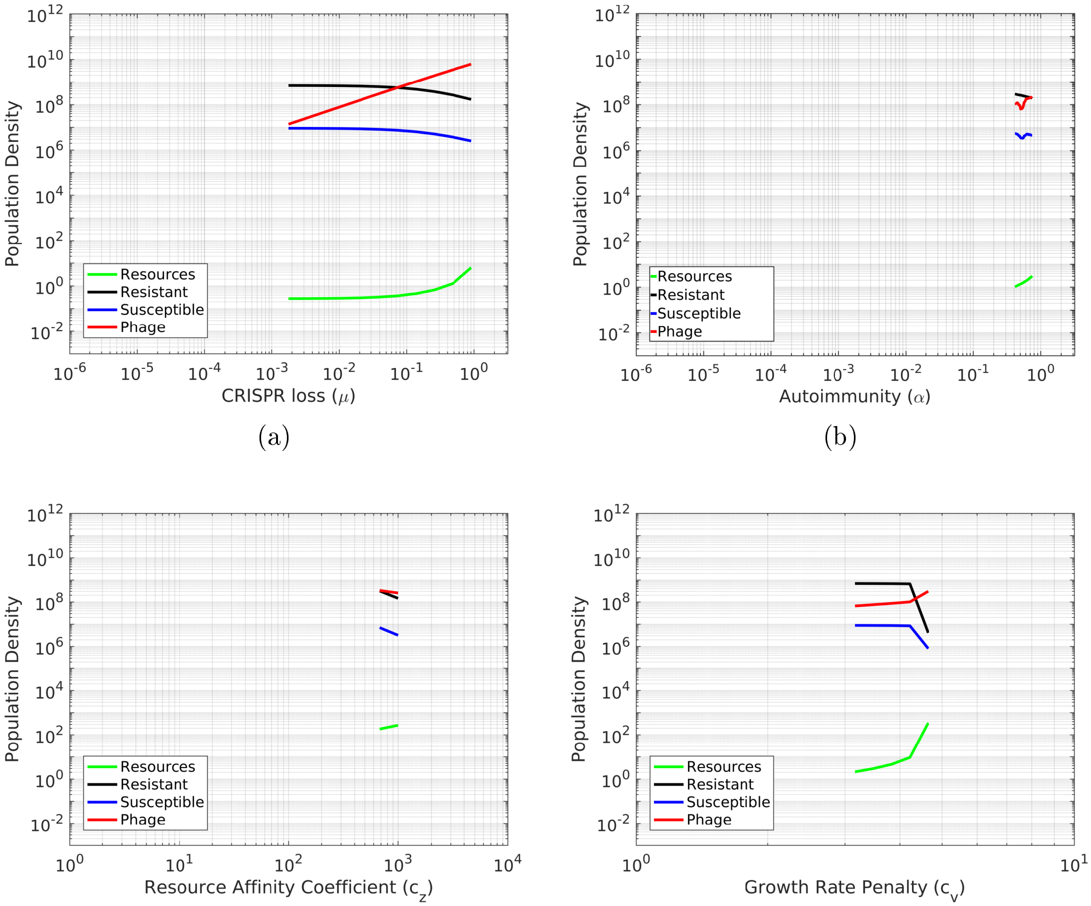
(a)

(b)

(c)

(d)

Figure C.2: Equilibria with each coexistence mechanism in isolation. Behavior of coexistence equilibrium when (a) there is only CRISPR loss without autoimmunity, (b) there is only autoimmunity without CRISPR loss, (c) there is only a cost applied to resource affinity (Appendix C.2), and (d) there is only a cost applied to maximum growth rate (Appendix C.2). Notice that immune loss and autoimmune mechanisms essentially act in the same manner, except that the loss mechanism produces a larger phage population by flushing extra susceptible bacteria into the system. This is consistent with theoretical results showing that increasing resource availability in a host-phage system increases phage rather than host populations [237]. The upper bound of the x-axis in (a-d) represents the upper limit of the cost of immunity, above which coexistence will not occur because immune host cannot survive.



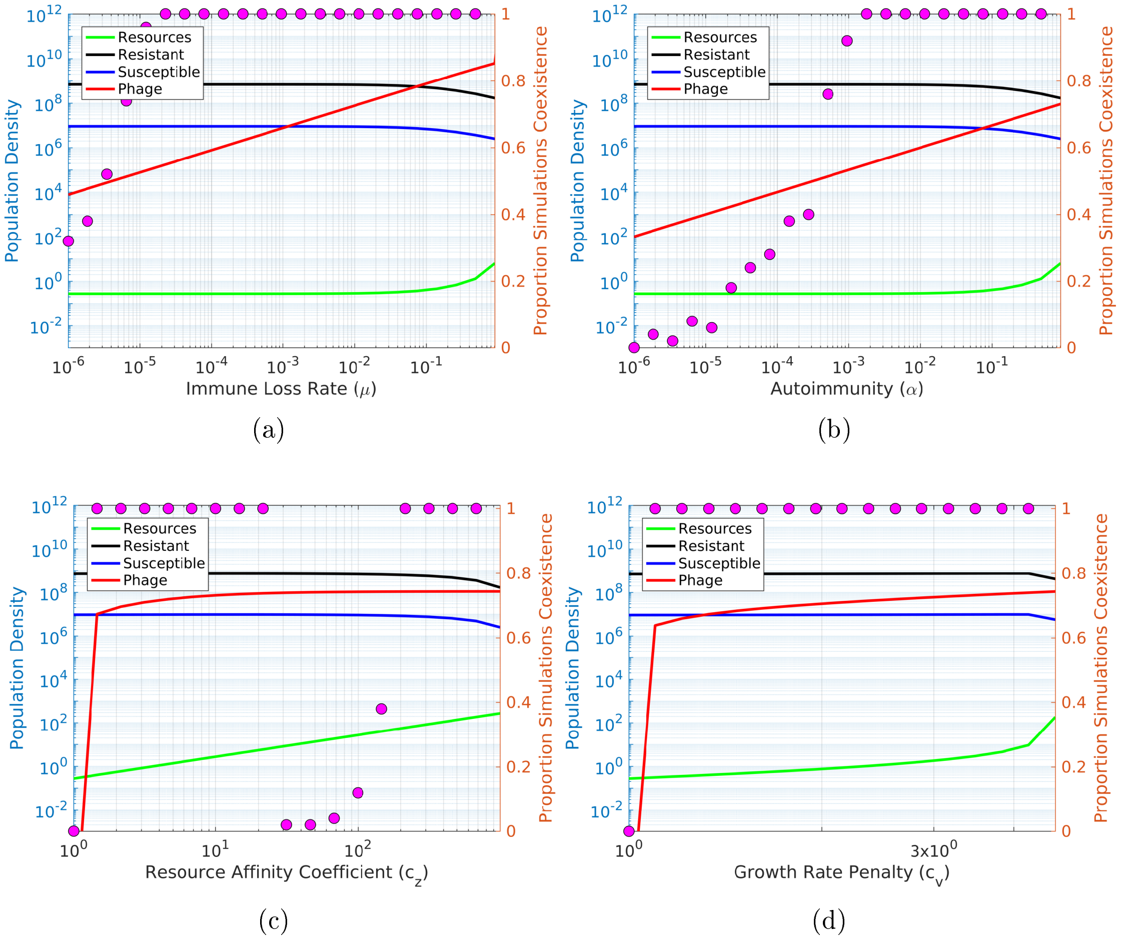
(a)

(b)

(c)

(d)

Figure C.3: Numerical solutions to model at 80 days with realistic initial conditions. Numerical solutions to the alternative cost model (Appendix C.2) at 80 days using realistic initial conditions more specific to the experimental setup ($R(0) = 350$, $D(0) = 10^6$, $U(0) = 100$, $P(0) = 10^6$). Results only shown for cases in which all three populations remained extant. Results in each panel correspond to each mechanism in isolation.



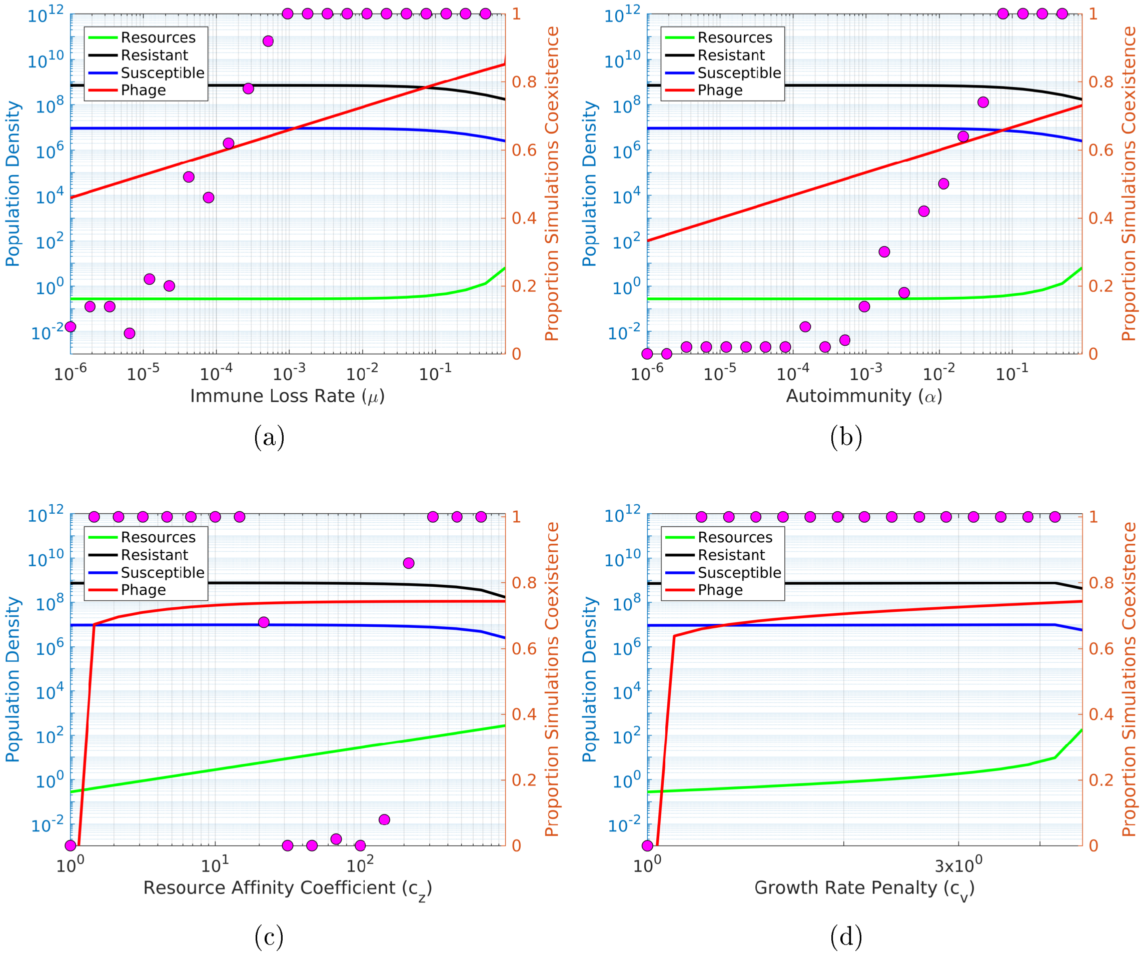
(a)

(b)

(c)

(d)

Figure C.4: Simulations of perturbed starting conditions (small perturbations). We find numerical solutions to the alternative cost model (Appendix C.2) at 80 days with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X}(1 + \gamma Y)$ where $Y \sim U[0, 1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. We ran 50 simulations for each condition. We let $\gamma = 0.1$. Lines correspond to the left axis and purple dots correspond to the right axis. Results in each panel correspond to each mechanism in isolation.



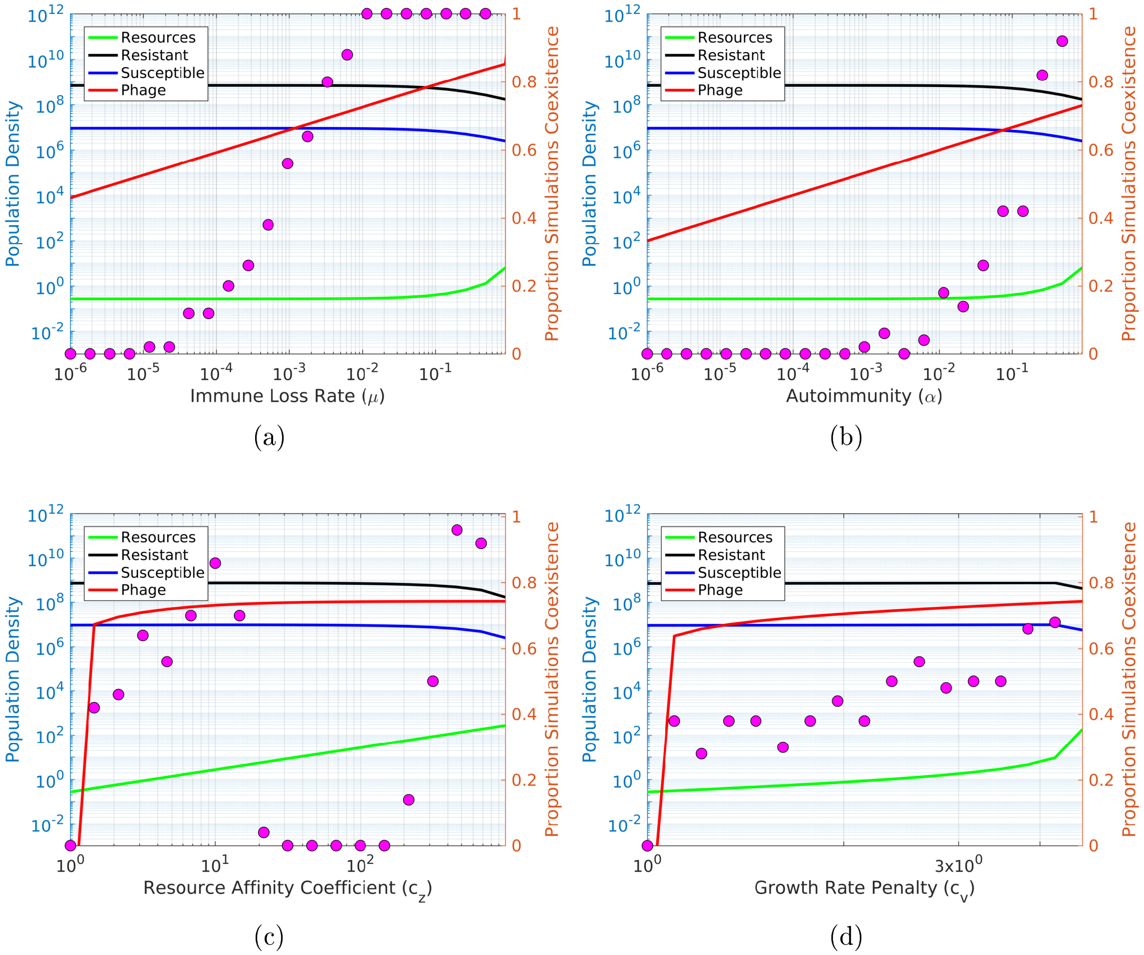
(a)

(b)

(c)

(d)

Figure C.5: Simulations of perturbed starting conditions (intermediate perturbations). We find numerical solutions to the alternative cost model (Appendix C.2) at 80 days with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X}(1 + \gamma Y)$ where $Y \sim U[0, 1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. We ran 50 simulations for each condition. We let $\gamma = 1$. Lines correspond to the left axis and purple dots correspond to the right axis. Results in each panel correspond to each mechanism in isolation.



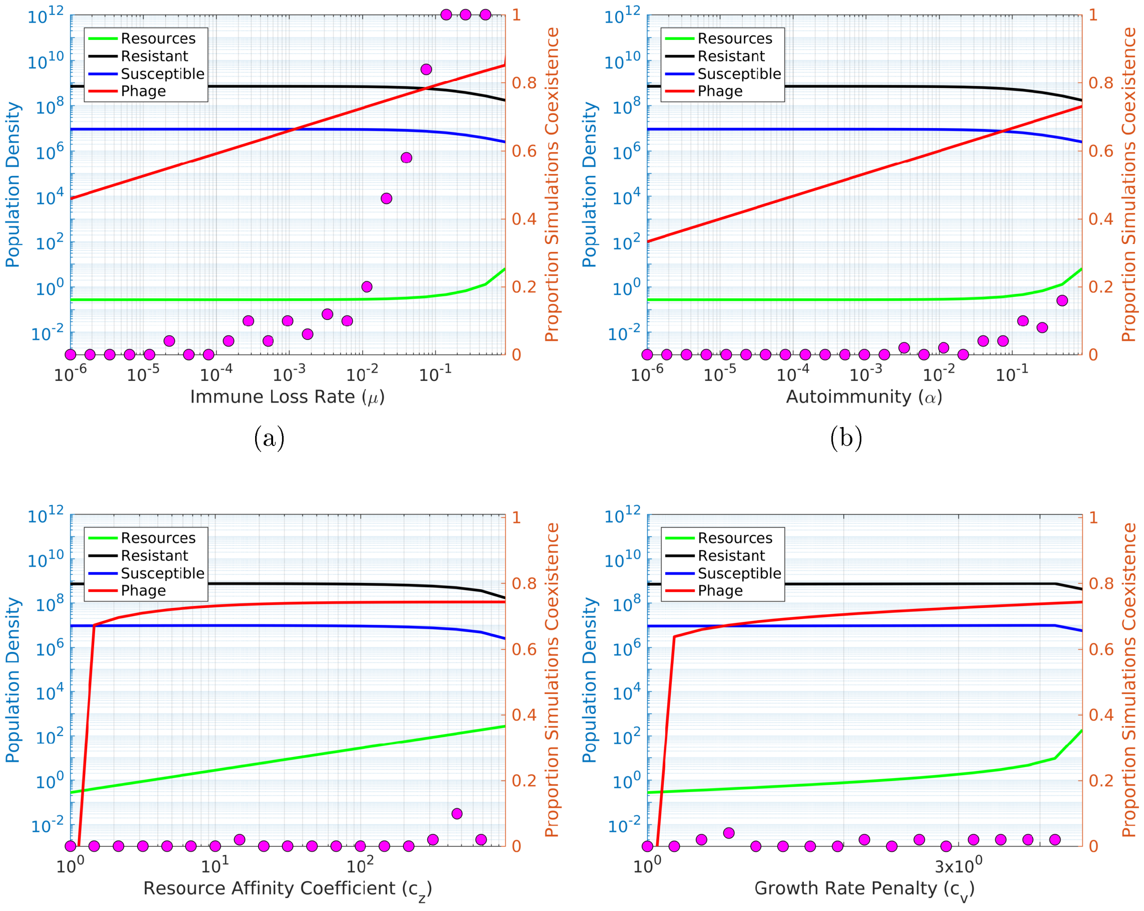
(a)

(b)

(c)

(d)

Figure C.6: Simulations of perturbed starting conditions (large perturbations). We find numerical solutions to the alternative cost model (Appendix C.2) at 80 days with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X}(1 + \gamma Y)$ where $Y \sim U[0, 1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. We ran 50 simulations for each condition. We let $\gamma = 10$. Lines correspond to the left axis and purple dots correspond to the right axis. Results in each panel correspond to each mechanism in isolation.



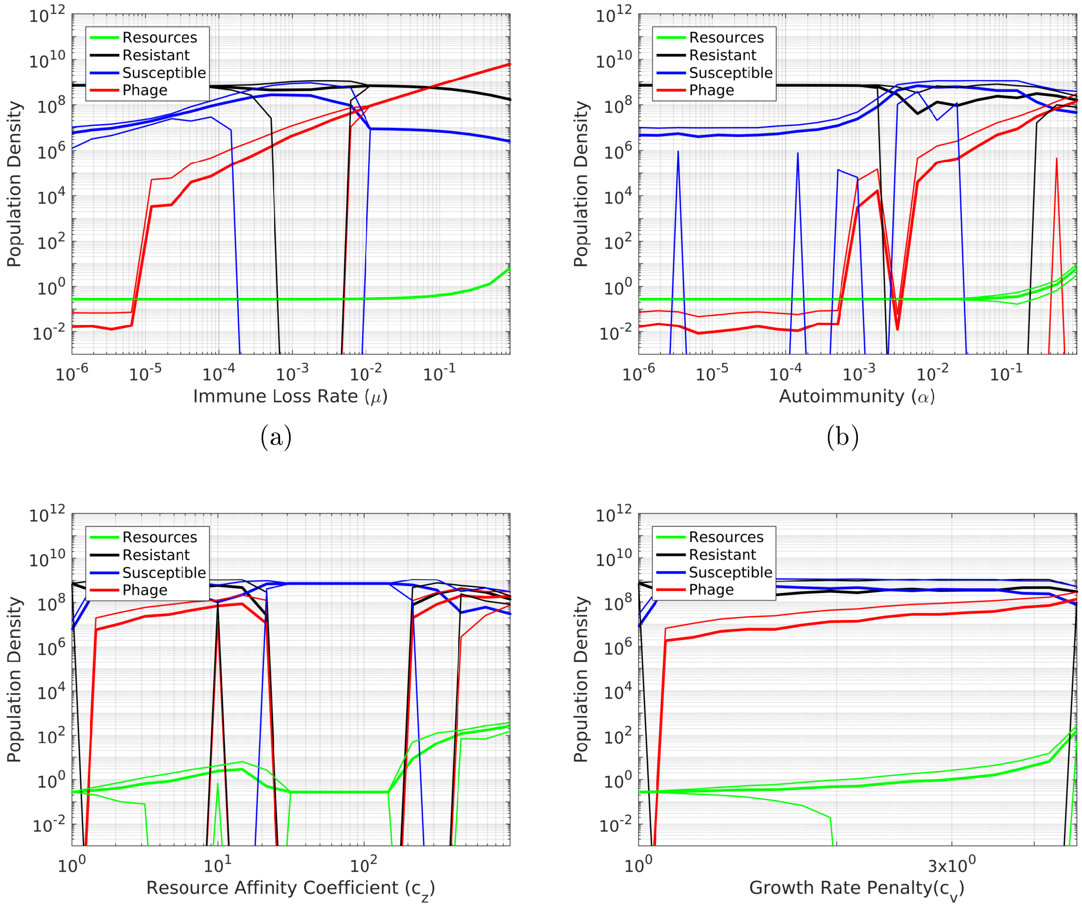
(a)

(b)

(c)

(d)

Figure C.7: Simulations of perturbed starting conditions (very large perturbations). We find numerical solutions to the alternative cost model (Appendix C.2) at 80 days with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X}(1 + \gamma Y)$ where $Y \sim U[0, 1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. We ran 50 simulations for each condition. We let $\gamma = 100$. Lines correspond to the left axis and purple dots correspond to the right axis. Results in each panel correspond to each mechanism in isolation.



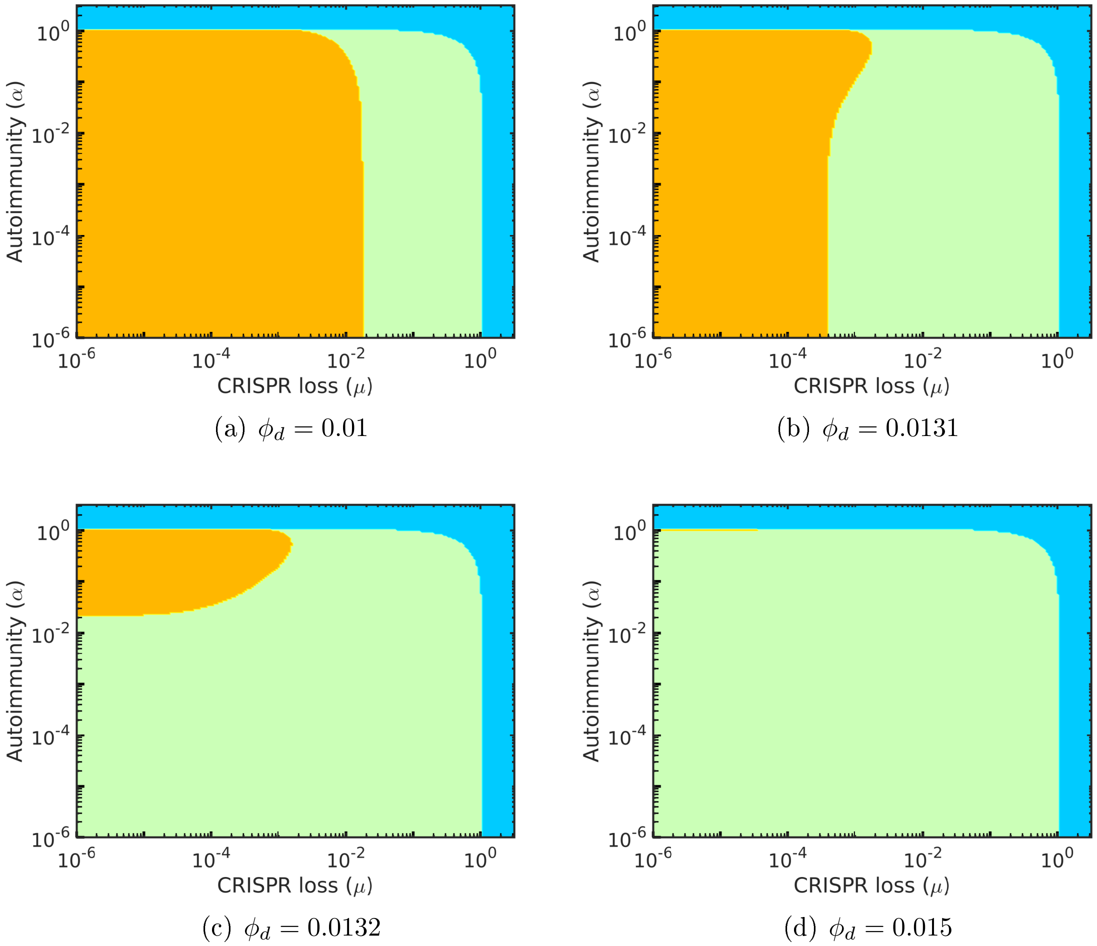
(a)

(b)

(c)

(d)

Figure C.8: Mean population size with perturbed starting conditions (intermediate perturbations). We find numerical solutions to the alternative cost model (Appendix C.2) at 80 days with starting conditions ($X(0) = [R(0), D(0), U(0), P(0)]$) perturbed by a proportion of the equilibrium condition $X(0) = \tilde{X}(1 + \gamma Y)$ where $Y \sim U[0, 1]$ and $\tilde{X}$ signifies an equilibrium value to explore how robust the equilibria are to starting conditions. We ran 50 simulations for each condition. We let $\gamma = 10$. Mean population across all simulations (including cases of phage or host extinction) shown by bold line and two standard deviations away from the mean are represented by the thin lines. Results in each panel correspond to each mechanism in isolation.



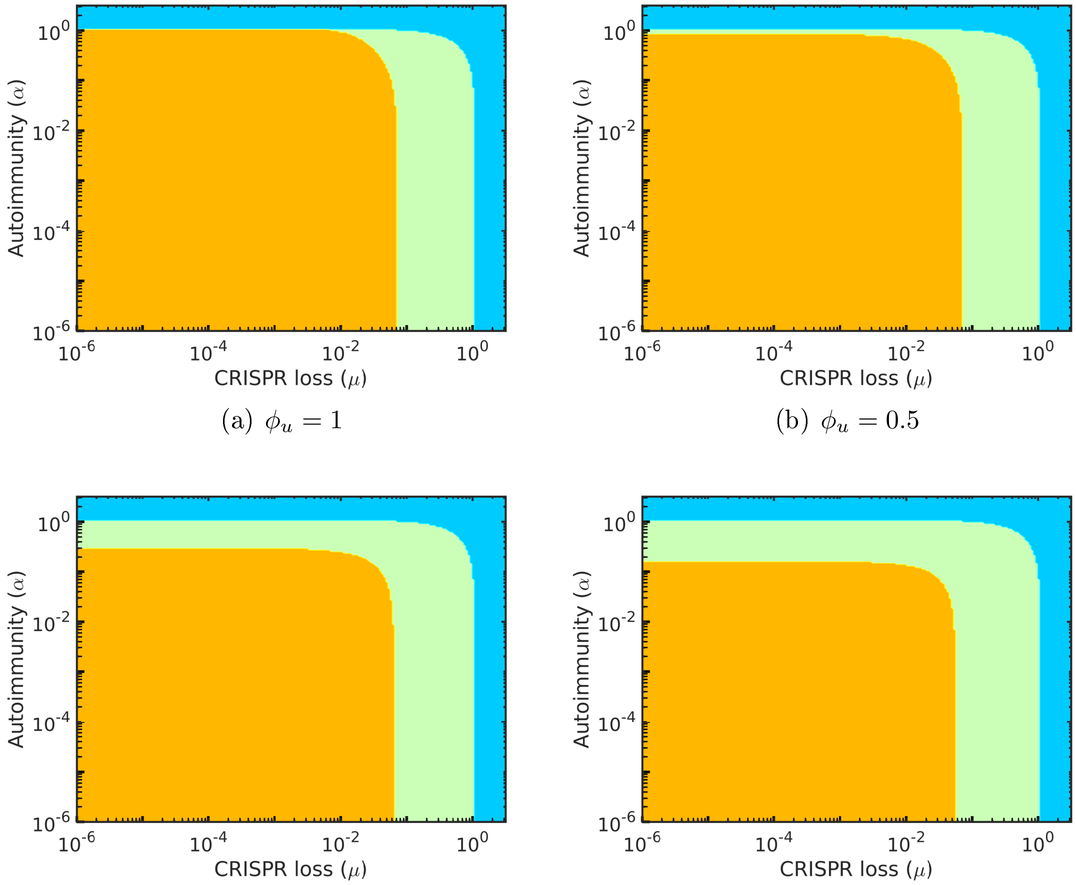
(a) $\phi_d = 0.01$

(b) $\phi_d = 0.0131$

(c) $\phi_d = 0.0132$

(d) $\phi_d = 0.015$

Figure C.9: Phase diagram of general model with phage coevolution. Phase diagrams of simple coevolutionary model behavior under variations in the rates of autoimmunity (α) and CRISPR system loss (μ) over various coevolutionary scenarios (ϕ_d). Values of ϕ_d were chosen so as to demonstrate the rapid shift that occurs from host to phage dominated equilibrium as the infected fraction of defended host increases. Orange indicates a stable equilibrium with all populations coexisting and defended host dominating phage populations, green indicates that all populations coexist but phages dominate, and blue indicates that defended bacteria have gone extinct but phages and undefended bacteria coexist.



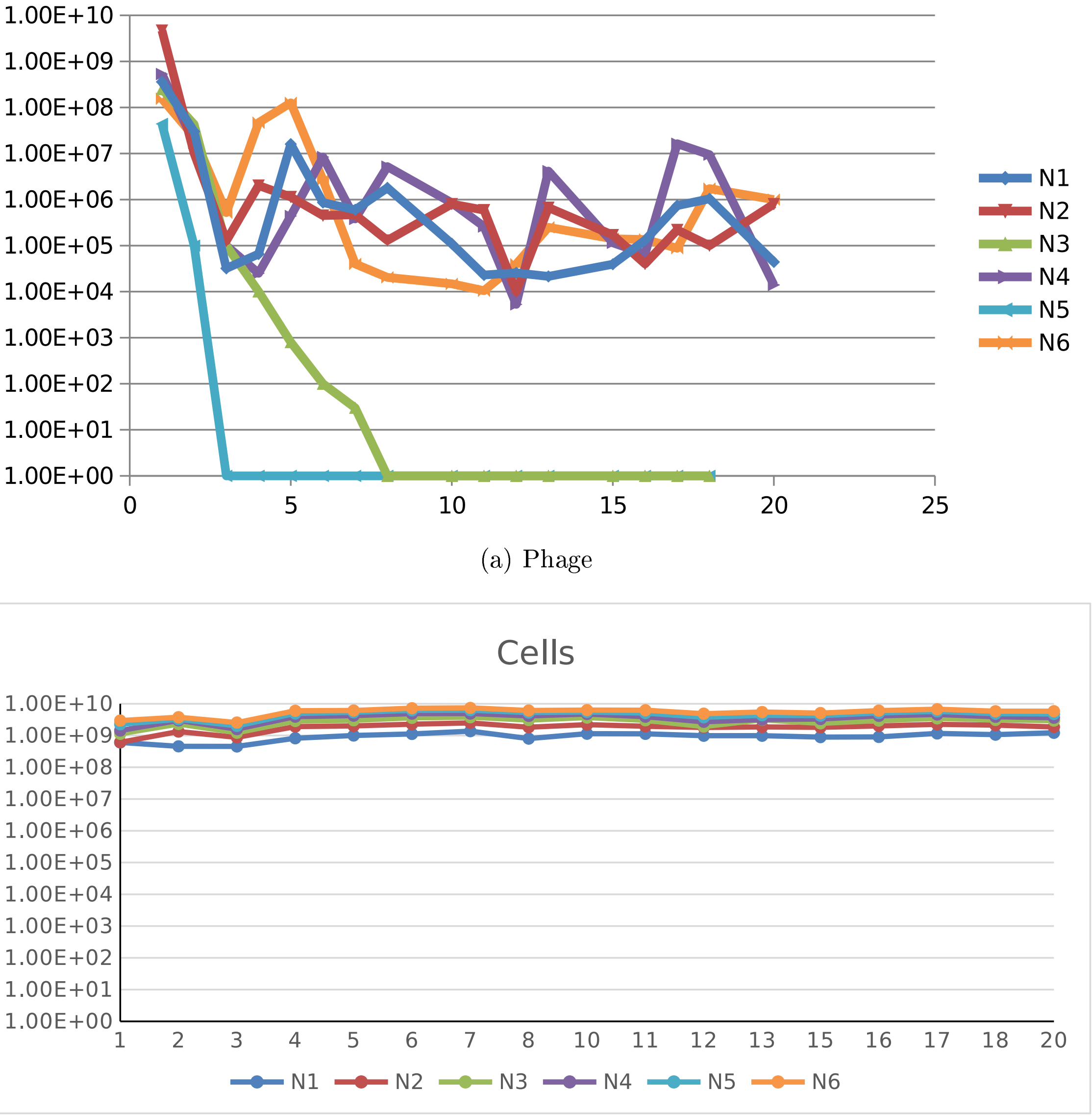
(a) $\phi_u = 1$

(b) $\phi_u = 0.5$

(c) $\phi_u = 0.1$

(d) $\phi_u = 0.05$

Figure C.10: Phase diagram of general model with innate immunity. Phase diagram of model behavior under variations in the rates of autoimmunity (α) and CRISPR system loss (μ) for different values of (ϕ_u). Orange indicates a stable equilibrium with all populations coexisting and defended host dominating phage populations, green indicates that all populations coexist but phages dominate, and blue indicates that defended bacteria have gone extinct but phages and undefended bacteria coexist.



(a) Phage

(b) Bacteria

Figure C.11: Replicate serial transfer experiments. Densities of (a) phage and (b) bacteria measured daily at serial transfer. All replicate experiments start with the same conditions and strains as in the main text.

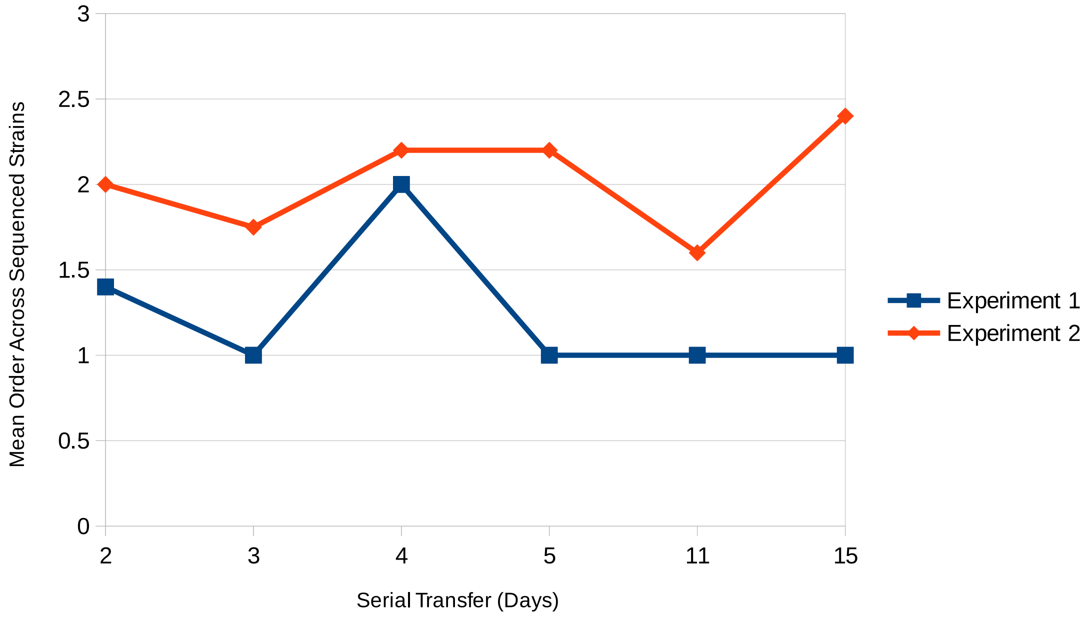
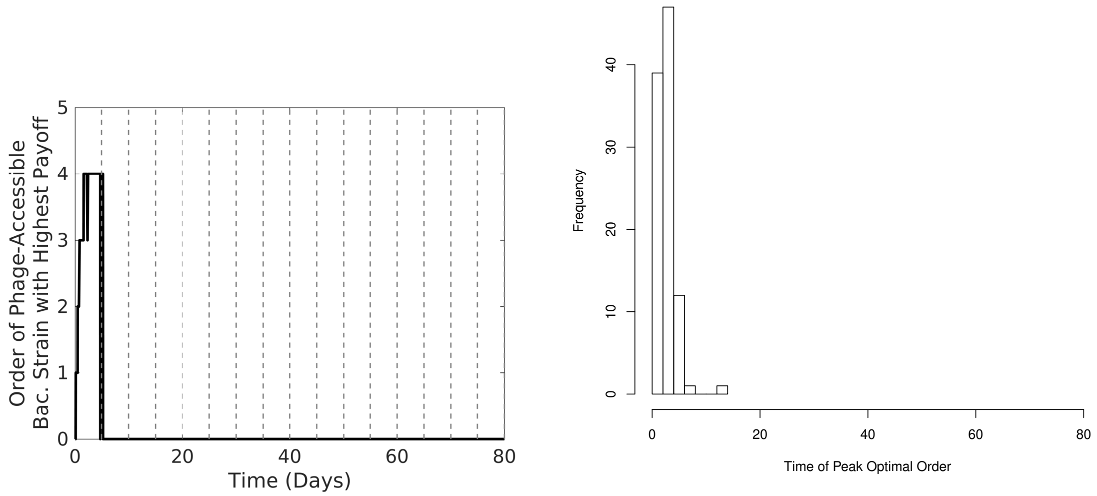


Figure C.12: Mean sequenced order of host over time in serial transfer experiments 1 and 2. Sum over CRISPR1 and CRISPR3 loci.



(a)

(b)

Figure C.13: Optimal host order for phages to infect over time. The optimal host strain that is either currently infected by a phage strain or one PAM mutation away from being infected. Optimality is defined in terms of population size times the burst size of the phage strain that does or could infect that strain, so that the balance between abundant host and mutation cost is taken into account. In (a) we track the order of the “best” available host strain at any given point in a single example simulation (Fig C.22), and in (b) look at the timing of the peak optimal order across 100 simulations (Fig 4.4, $\mu_q = 5 \times 10^{-9}$). Note that after the initial arms race dynamic the best available host strain is the CRISPR-lacking host in all simulations. The order of the best host strain peaks early on in all simulations and then drops to zero (CRISPR-lacking), signifying an early end to the arms race between host and phage.

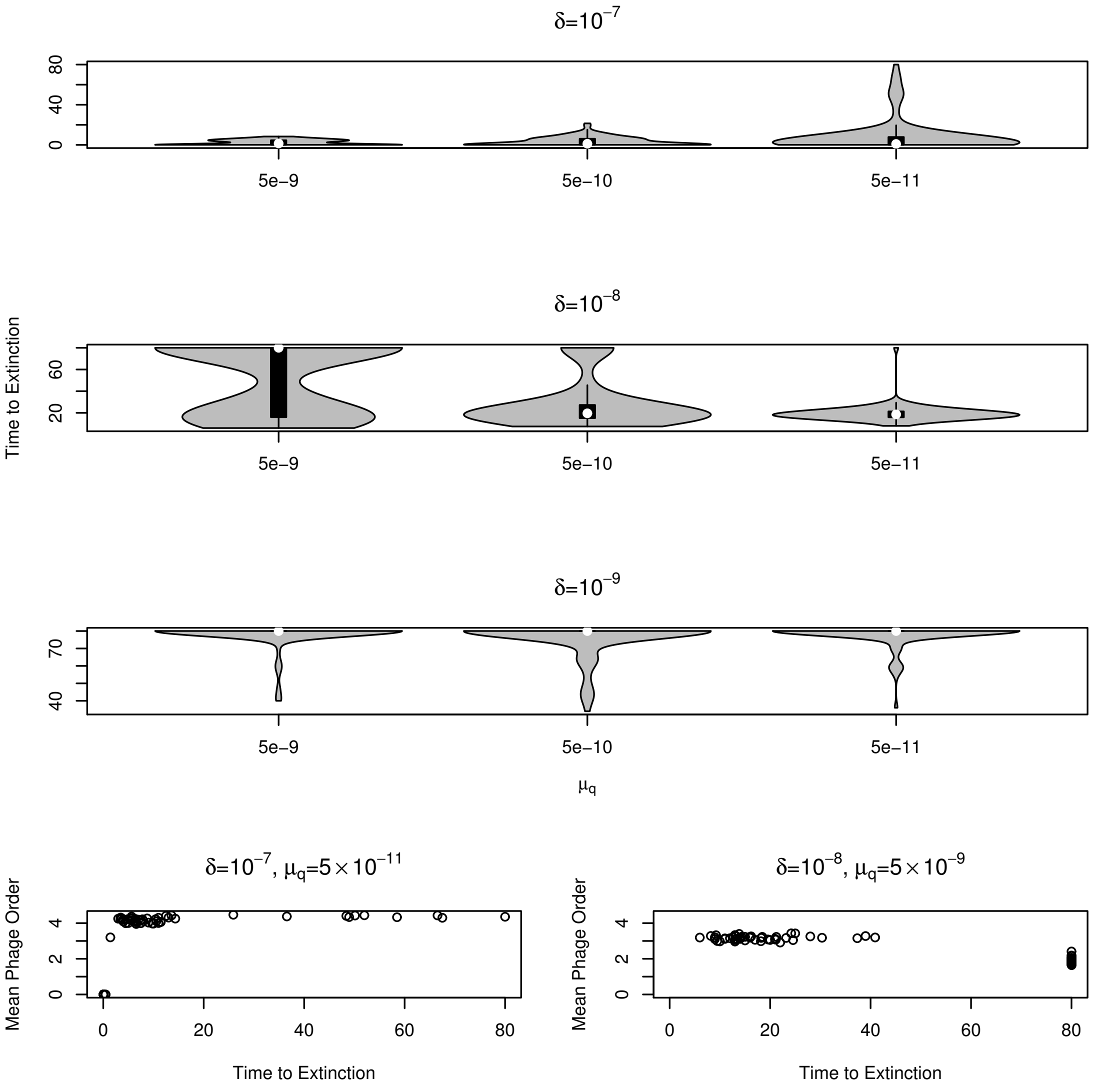


Figure C.14: Effect on simulations of varied phage adsorption rates. Adsorption rate has a profound effect on the outcome of host-phage interactions. At a high adsorption rate ($\delta = 10^{-7}$) either phages or bacteria tend to go extinct early on in the simulation, and phages are able to drive their host extinct approximately 50% of the time (49/100 simulations for all back-mutation rates). We see a reversed relationship of time to extinction with μ_q from our base adsorption rate ($\delta = 10^{-8}$, Fig 4.4), although in general at a very high adsorption rate few simulations demonstrate long-term coexistence as phage consume all host early on. This suggests that the costs associated with PAM mutations are required to keep phage growth rate low enough to prevent overconsumption of host, and indeed upon closer examination of individual simulations it is clear that back mutations to lower phage orders precipitate phage collapse. In the lowest panel we demonstrate that coexistence in the long term at high δ is associated with a high mean phage order over the course of a simulation, while the opposite is true of our typical intermediate δ. At a low adsorption rate ($\delta = 10^{-9}$) we see populations coexisting until the 80 day mark (max simulation length) in almost all simulations.

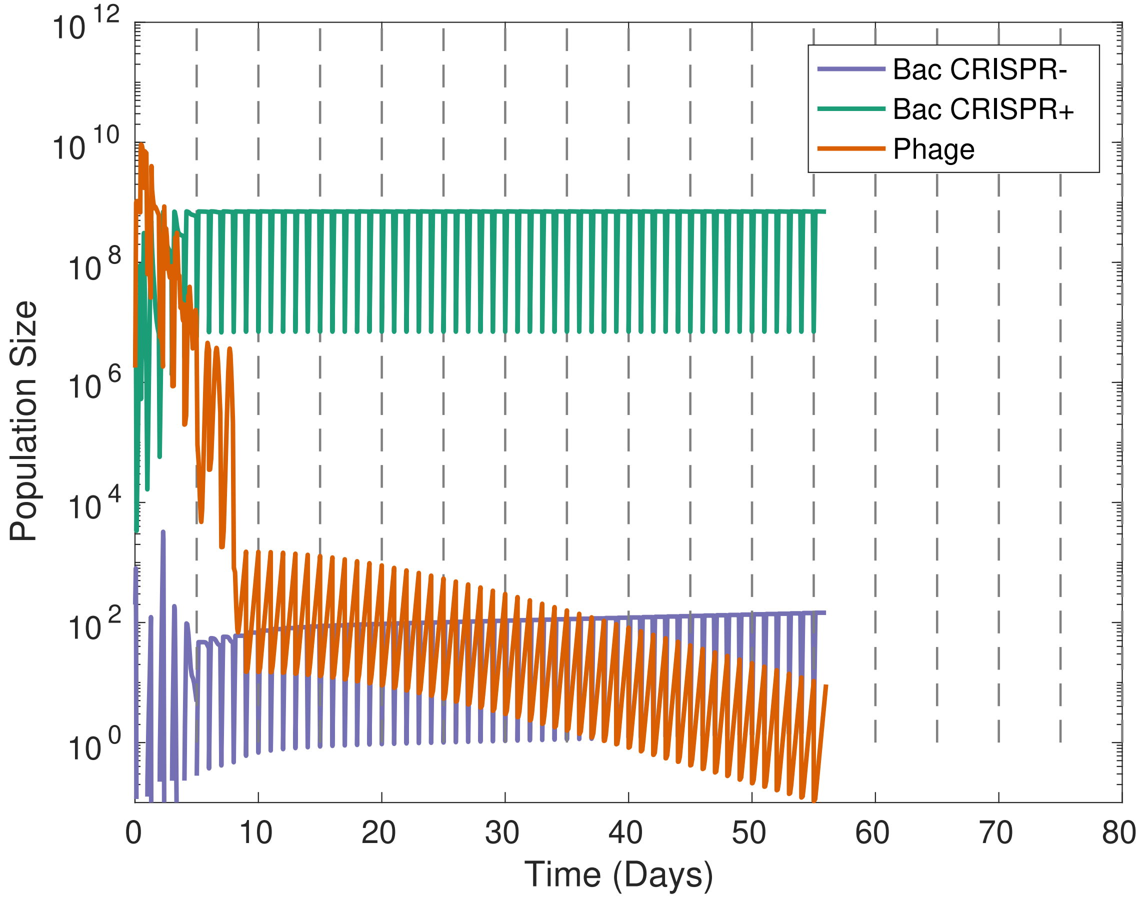


Figure C.15: Representative simulation with a floor on the susceptible host population and high autoimmunity. We let $B_s > 1 \, \forall t$ and $\alpha = 5 \times 10^{-4}$.

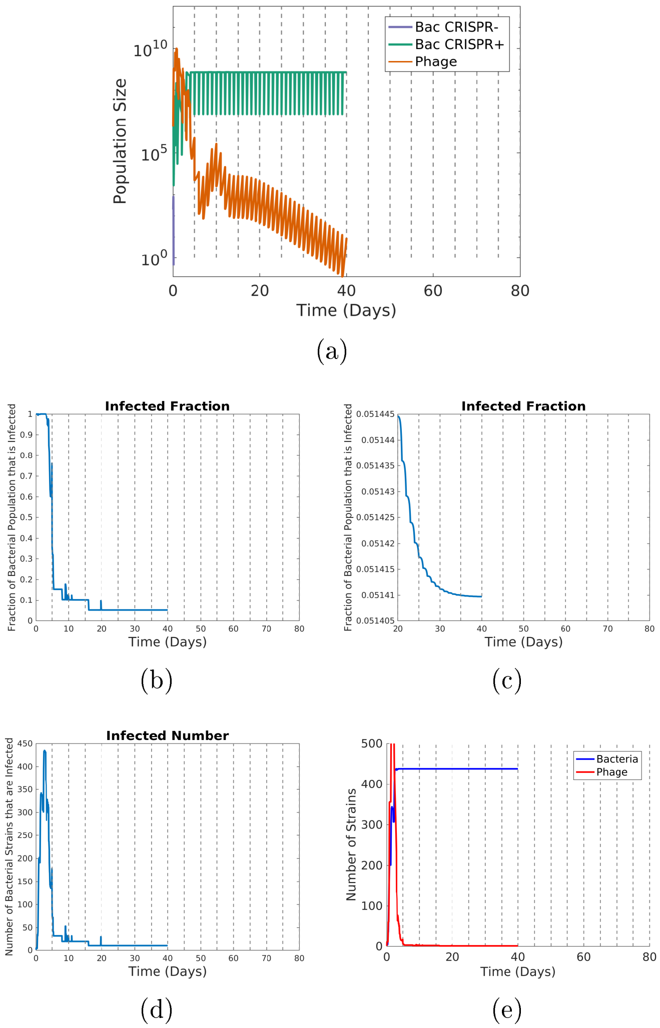


Figure C.16: Transient phage survival at low density Example of low-level phage persistence due to slow evolutionary dynamics. Here we see that (a) in the absence of the constant production of susceptible bacteria by CRISPR-enabled strains (i.e., $\mu = 0$) phages are still able to paradoxically persist despite a clear advantage to bacteria in the arms race and an absence of other sustaining mechanisms. In (b) a small fraction of the CRISPR enabled bacterial population is maintained that lacks spacers towards the infecting phages and in (c) we zoom in to show that this population is declining due to this infection, but extremely slowly, implying this coexistence is not stable in the long term. The number of bacterial strains that can be infected by phages over time is shown in (d), and (e) shows how the richness of phage and bacterial strains changes over time. Note that although the number of bacterial strains increases asymptotically, after an initial spike the number of strains that can be infected by phages drops dramatically to the single digits. This keeps the overall phage population growth constrained (balanced by adsorption to immune bacteria). In fact, over time phage act to suppress their own growth by negatively infecting the competitiveness of their host (although this effect is so small that phages can persist for an extended period of time seemingly stably). What looks flat is actually monotonically decreasing. All parameters as in Main Text Table 4.3 and $\alpha = 0$.

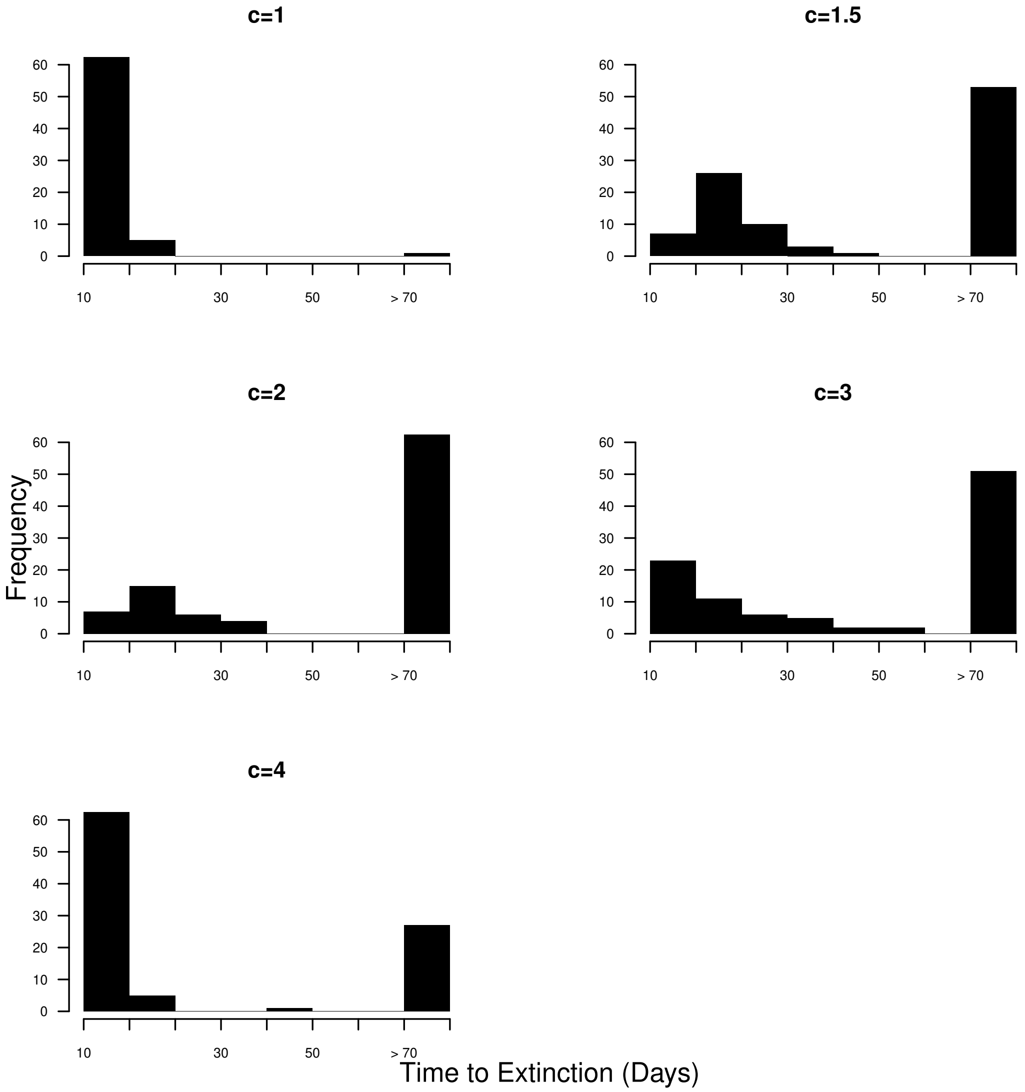
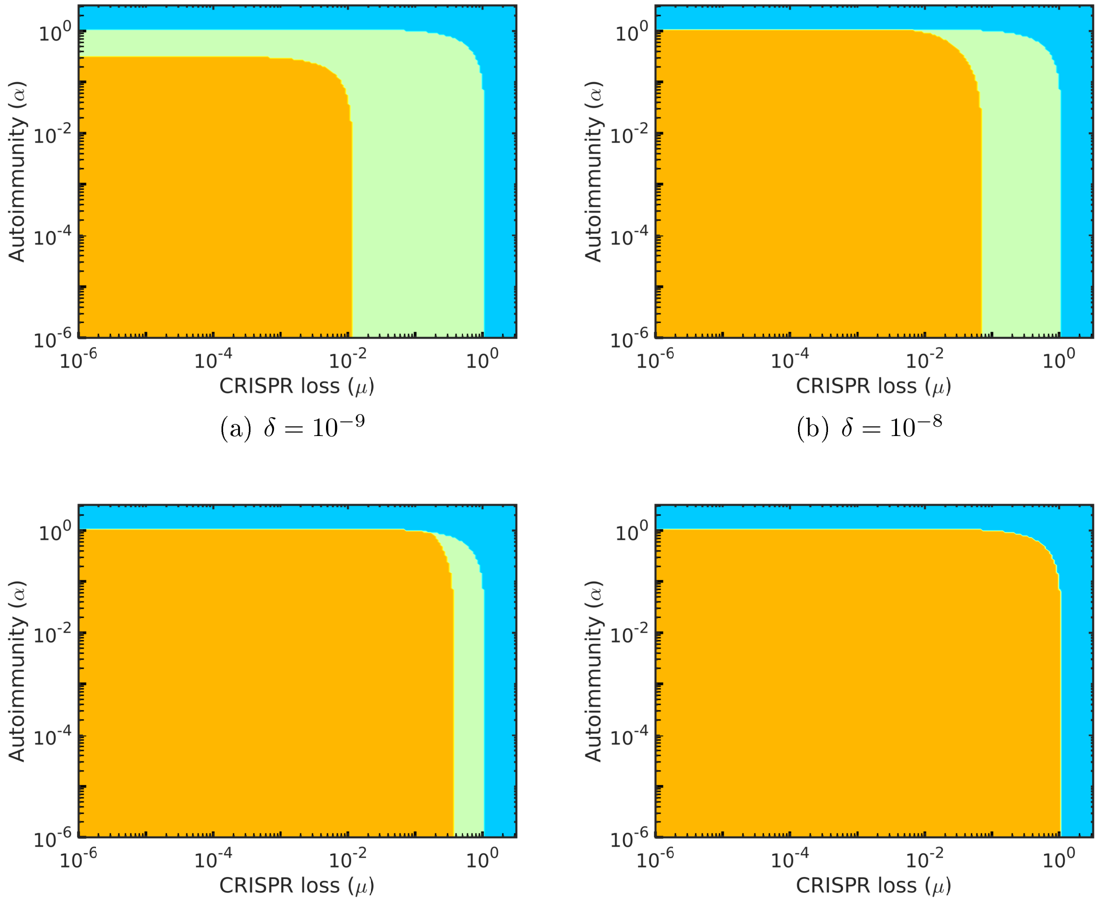


Figure C.17: Effect of changes in PAM mutation cost (c). Distribution of phage extinction times in bacterial-dominated cultures with different costs on PAM mutation in phage (c). The peak at 80 corresponds to stable coexistence (simulations ran for a maximum of 80 days). These results of for a locus-loss mechanism only ($\mu_L = 5 \times 10^{-4}$, $\alpha = 0$).



(a) $\delta = 10^{-9}$

(b) $\delta = 10^{-8}$

(c) $\delta = 10^{-7}$

(d) $\delta = 10^{-6}$

Figure C.18: Phase diagram of general model with phage coevolution. Phase diagrams of model behavior without coevolution or other forms of immunity ($\phi_d = 0$, $\phi_u = 1$) under variations in the rates of autoimmunity (α) and CRISPR system loss (μ) over various adsorption rates (δ). Orange indicates a stable equilibrium with all populations coexisting and defended host dominating phage populations, green indicates that all populations coexist but phages dominate, and blue indicates that defended bacteria have gone extinct but phages and undefended bacteria coexist. There is an apparent increase in the area of the coexistence region in which host dominate as adsorption rate increases.

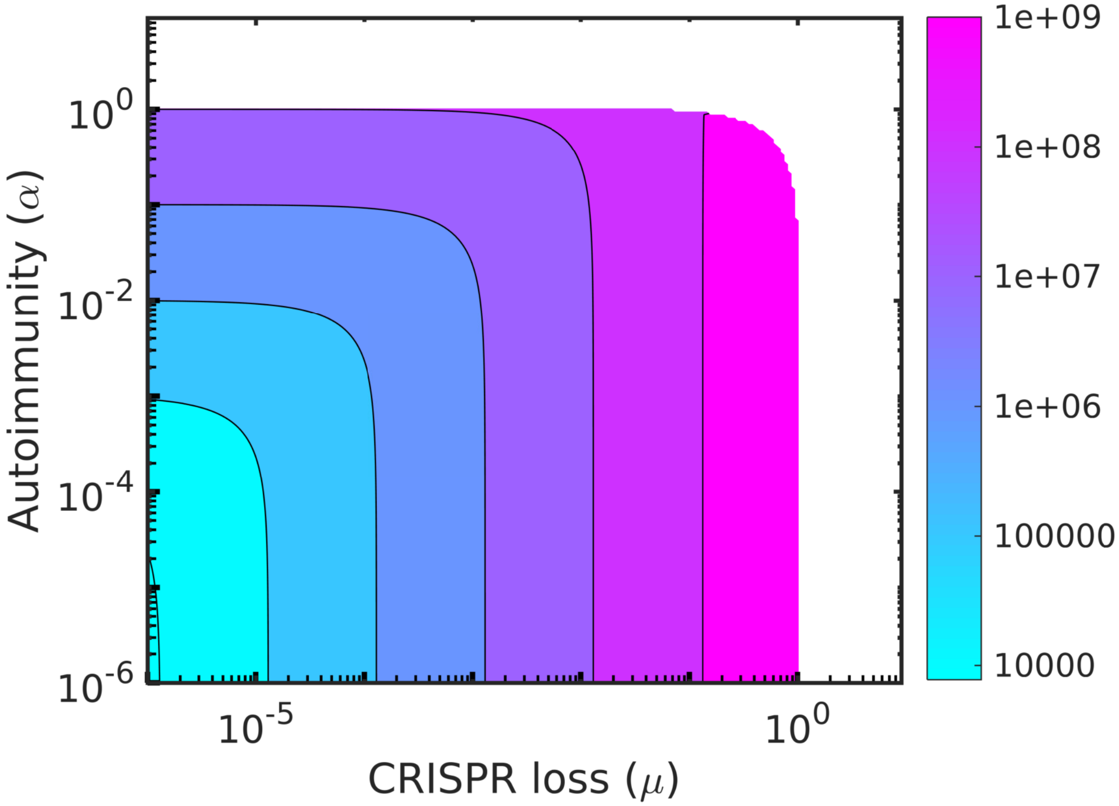


Figure C.19: Equilibrium phage population during coexistence. Equilibrium population of phages when there is full coexistence over a range of α and μ_L values for our general model without coevolution ($\phi_u = 1$, $\phi_d = 0$).

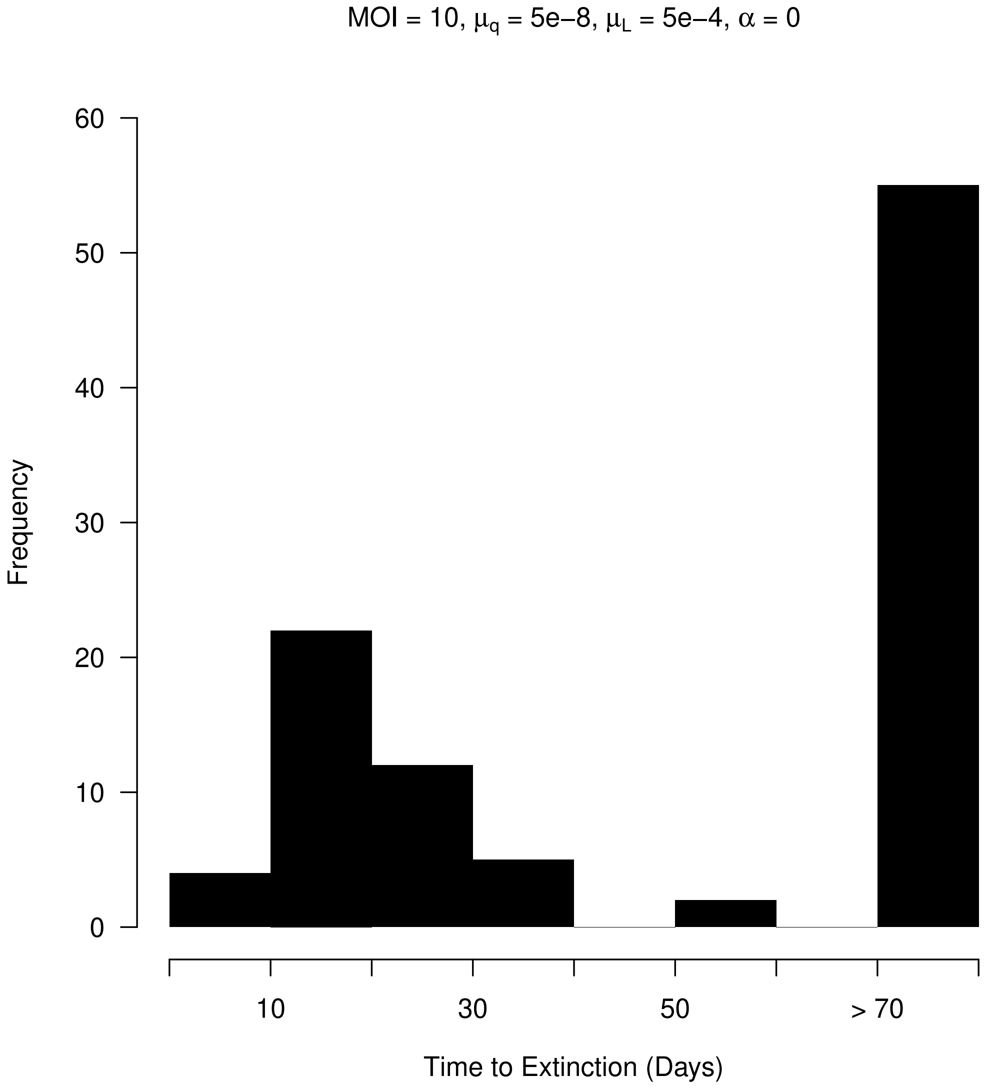


Figure C.20: Distribution of phage extinction times in bacterial-dominated cultures with an MOI of 10. The peak at 80 corresponds to what we call stable coexistence (simulations ran for a maximum of 80 days).

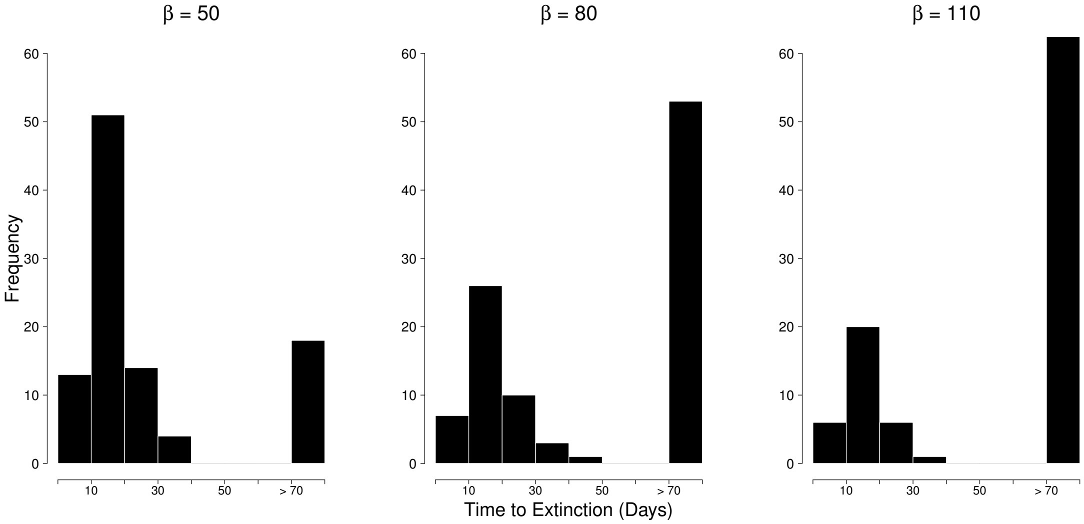
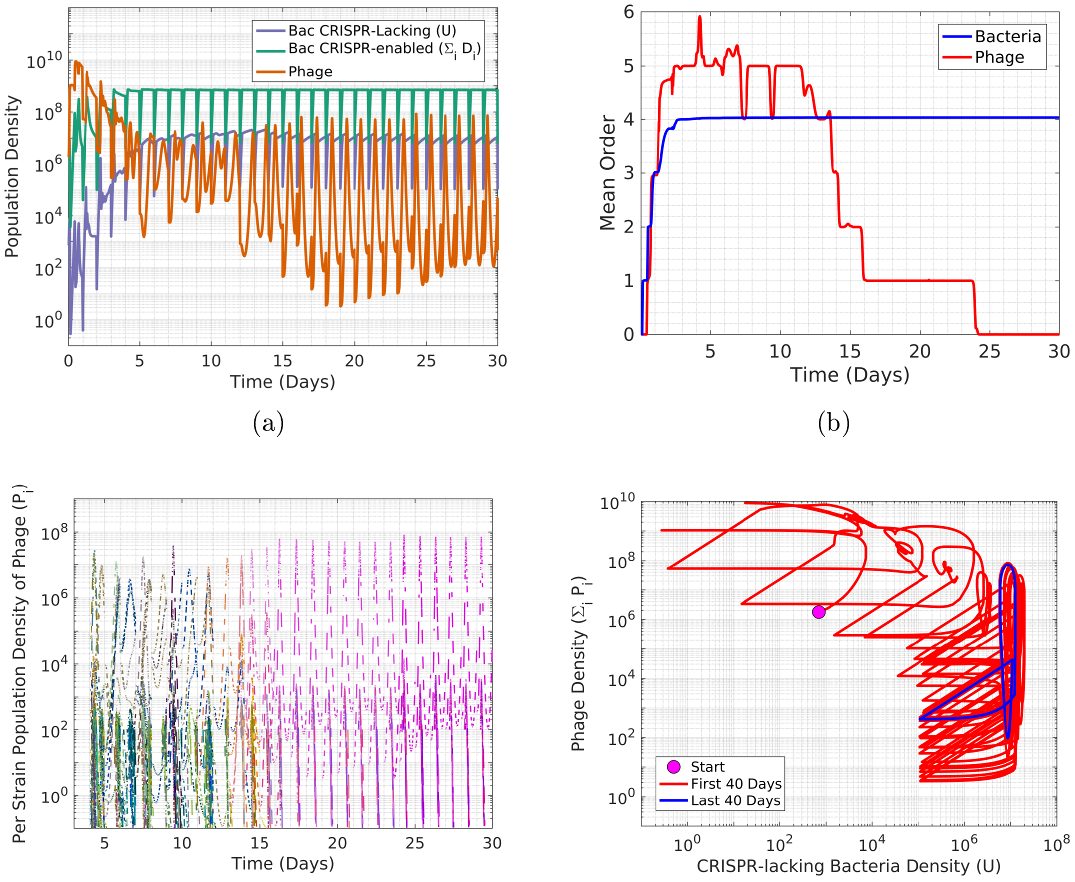


Figure C.21: Distributions of phage extinction times in bacterial-dominated cultures with various burst sizes. The peak at 80 corresponds to what we call stable coexistence (simulations ran for a maximum of 80 days).



(a)

(b)

(c)

(d)

Figure C.22: Representative example of a simulation demonstrating stable coexistence under a loss mechanism ($\mu_L = 5 \times 10^{-4}$, $\alpha = 0$, $\mu_q = 5 \times 10^{-9}$). In (a) we show the archetypal shift from phage-dominance during an initial arms race to unstable host-dominated coexistence where fluctuating selection dynamics are observed to stable host-dominated predator-prey cycling of phages and CRISPR-lacking hosts as seen in (d) where evolution ceases to occur. In (b) we see that a drop in mean phage order leads to stable cycling and in (c) that this corresponds to a single phage strain becoming dominant after previous cycling of strains. This corresponds to a shift away from fluctuating selection dynamics. In (c) the colors specify different phage strains.

References

[1] Jake L. Weissman, Hao H. Yiu, and Philip L.F. Johnson. What bacteria do when they get sick. *Fronteirs Young Minds*, 7(102), 2019.

[2] John H Paul. Microbial gene transfer: an ecological perspective. *Journal of Molecular Microbiology and biotechnology*, 1(1):45–50, 1999.

[3] Curtis A Suttle. Marine viruses-major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801, 2007.

[4] Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, Léo Berline, Jennifer R Brum, et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 532(7600):465, 2016.

[5] T Frede Thingstad. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, 45(6):1320–1328, 2000.

[6] Sarit Avrani, Omri Wurtzel, Itai Sharon, Rotem Sorek, and Debbie Lindell. Genomic island variability facilitates *Prochlorococcus*–virus coexistence. *Nature*, 474(7353):604, 2011.

[7] Marcia F Marston, Francis J Pierciey, Alicia Shepard, Gary Gearin, Ji Qi, Chandri Yandava, Stephan C Schuster, Matthew R Henn, and Jennifer BH Martiny. Rapid diversification of coevolving marine synechococcus and a virus. *Proceedings of the National Academy of Sciences*, 109(12):4544–4549, 2012.

[8] Sergey N. Rodin and Vadim A. Ratner. Some theoretical aspects of protein coevolution in the ecosystem "phage-bacteria" I. The problem. *Journal of Theoretical Biology*, 100(2):185–195, January 1983.

[9] Sergey N. Rodin and Vadim A. Ratner. Some theoretical aspects of protein coevolution in the ecosystem "phage-bacteria" II. The deterministic model of microevolution. *Journal of Theoretical Biology*, 100(2):197–210, January 1983.

[10] Eugene V Koonin, Kira S Makarova, and Yuri I Wolf. Evolutionary genomics of defense systems in archaea and bacteria. *Annual Review of Microbiology*, 71:233–261, 2017.

[11] Stineke van Houte, Angus Buckling, and Edze R. Westra. Evolutionary ecology of prokaryotic immune mechanisms. *Microbiology and Molecular Biology Reviews*, 80(3):745–763, September 2016.

[12] Kira S. Makarova, Yuri I. Wolf, Sagi Snir, and Eugene V. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of Bacteriology*, 193(21):6039–6056, November 2011.

[13] Shany Doron, Sarah Melamed, Gal Ofir, Azita Leavitt, Anna Lopatina, Mai Keren, Gil Amitai, and Rotem Sorek. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, page eaar4120, January 2018.

[14] Kira S. Makarova, Vivek Anantharaman, L. Aravind, and Eugene V. Koonin. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biology Direct*, 7:40, 2012.

[15] Jake L Weissman, William F Fagan, and Philip LF Johnson. Selective maintenance of multiple CRISPR arrays across prokaryotes. *The CRISPR Journal*, 1(6):405–413, 2018.

[16] Pere Puigbò, Kira S. Makarova, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin. Reconstruction of the evolution of microbial defense systems. *BMC Evolutionary Biology*, 17:94, April 2017.

[17] Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research*, 41(8):4360–4377, April 2013.

[18] Francisco J. M. Mojica, Cesar Díez-Villaseñor, Jesús García-Martínez, and Elena Soria. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution*, 60(2):174–182, 2005.

[19] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A. Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, March 2007.

[20] Ruud Jansen, Jan DA van Embden, Wim Gaastra, and Leo M Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology*, 43(6):1565–1575, 2002.

[21] Reidun K. Lillestøl, Shiraz A. Shah, Kim Brügger, Peter Redder, Hien Phan, Jan Christiansen, and Roger A. Garrett. CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Molecular Microbiology*, 72(1):259–272, April 2009.

[22] Ümit Pul, Reinhild Wurm, Zihni Arslan, René Geißen, Nina Hofmann, and Rolf Wagner. Identification and characterization of *E. coli* CRISPR-Cas promoters and their silencing by H-NS. *Molecular Microbiology*, 75(6):1495–1512, 2010.

[23] Yunzhou Wei, Megan T. Chesne, Rebecca M. Terns, and Michael P. Terns. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Research*, 43(3):1749–1758, February 2015.

[24] Anders F. Andersson and Jillian F. Banfield. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science (New York, N.Y.)*, 320(5879):1047–1050, May 2008.

[25] Adi Stern, Leeat Keren, Omri Wurtzel, Gil Amitai, and Rotem Sorek. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends in Genetics*, 26(8):335–340, August 2010.

[26] David Paez-Espino, Wesley Morovic, Christine L. Sun, Brian C. Thomas, Ken-ichi Ueda, Buffy Stahl, Rodolphe Barrangou, and Jillian F. Banfield. Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nature Communications*, 4:1430, February 2013.

[27] Yunzhou Wei, Rebecca M. Terns, and Michael P. Terns. Cas9 function and host genome sampling in Type II-A CRISPR–Cas adaptation. *Genes & Development*, 29(4):356–361, February 2015.

[28] Eugene V. Koonin and Yuri I. Wolf. Is evolution Darwinian or/and Lamarckian? *Biology Direct*, 4:42, November 2009.

[29] Eugene V. Koonin and Yuri I. Wolf. Just how Lamarckian is CRISPR-Cas immunity: the continuum of evolvability mechanisms. *Biology Direct*, 11:9, 2016.

[30] Kira S. Makarova, Yuri I. Wolf, Omer S. Alkhnbashi, Fabrizio Costa, Shiraz A. Shah, Sita J. Saunders, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Daniel H. Haft, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Rebecca M. Terns, Michael P. Terns, Malcolm F. White, Alexander F. Yakunin, Roger A. Garrett, John van der Oost, Rolf Backofen, and Eugene V. Koonin. An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*, 13(11):722–736, November 2015.

[31] Ariel D. Weinberger, Christine L. Sun, Mateusz M. Pluciński, Vincent J. Denef, Brian C. Thomas, Philippe Horvath, Rodolphe Barrangou, Michael S. Gilmore, Wayne M. Getz, and Jillian F. Banfield. Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Computational Biol*, 8(4):e1002475, April 2012.

[32] Sukrit Silas, Georg Mohr, David J Sidote, Laura M Markham, Antonio Sanchez-Amat, Devaki Bhaya, Alan M Lambowitz, and Andrew Z Fire. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science*, 351(6276):aad4234, 2016.

[33] Caryn Hale, Kyle Kleppe, Rebecca M Terns, and Michael P Terns. Prokaryotic silencing (psi) RNAs in *Pyrococcus furiosus*. *RNA*, 14(12):2572–2579, 2008.

[34] Caryn R Hale, Peng Zhao, Sara Olson, Michael O Duff, Brenton R Graveley, Lance Wells, Rebecca M Terns, and Michael P Terns. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, 139(5):945–956, 2009.

[35] Stan JJ Brouns, Matthijs M Jore, Magnus Lundgren, Edze R Westra, Rik JH Slijkhuis, Ambrosius PL Snijders, Mark J Dickman, Kira S Makarova, Eugene V Koonin, and John Van Der Oost. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 321(5891):960–964, 2008.

[36] Hong Li. Structural principles of CRISPR RNA processing. *Structure*, 23(1):13–20, 2015.

[37] Jason Carte, Ruiying Wang, Hong Li, Rebecca M Terns, and Michael P Terns. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & development*, 22(24):3489–3496, 2008.

[38] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S. Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561, 2005.

[39] Giedrius Gasiunas, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*, 109(39):E2579–E2586, 2012.

[40] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, 2012.

[41] Pedro H. Oliveira, Marie Touchon, and Eduardo P. C. Rocha. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research*, 42(16):10618–10631, September 2014.

[42] Tamara Goldfarb, Hila Sberro, Eyal Weinstock, Ofir Cohen, Shany Doron, Yoav Charpak-Amikam, Shaked Afik, Gal Ofir, and Rotem Sorek. BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal*, 34(2):169–183, January 2015.

[43] Gal Ofir, Sarah Melamed, Hila Sberro, Zohar Mukamel, Shahar Silverman, Gilad Yaakov, Shany Doron, and Rotem Sorek. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nature Microbiology*, 3(1):90, 2018.

[44] Francisco J. M. Mojica, Cesar Díez-Villaseñor, Elena Soria, and Guadalupe Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology*, 36(1):244–246, January 2002.

[45] Kira S. Makarova, Nick V. Grishin, Svetlana A. Shabalina, Yuri I. Wolf, and Eugene V. Koonin. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct*, 1:7, March 2006.

[46] Rika E. Anderson, William J. Brazelton, and John A. Baross. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiology Ecology*, 77(1):120–133, July 2011.

[47] Ariel D. Weinberger, Yuri I. Wolf, Alexander E. Lobkovsky, Michael S. Gilmore, and Eugene V. Koonin. Viral diversity threshold for adaptive immunity in prokaryotes. *mBio*, 3(6):e00456–12, December 2012.

[48] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Evolutionary dynamics of the prokaryotic adaptive immunity system CRISPR-Cas in an explicit ecological context. *Journal of Bacteriology*, 195(17):3834–3844, September 2013.

[49] Jake L Weissman, Rohan MR Laljani, William F Fagan, and Philip LF Johnson. Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *The ISME journal*, 2019.

[50] Shiraz A. Shah and Roger A. Garrett. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Research in Microbiology*, 162(1):27–38, January 2011.

[51] Wenyan Jiang, Inbal Maniv, Fawaz Arain, Yaying Wang, Bruce R. Levin, and Luciano A. Marraffini. Dealing with the evolutionary downside of crispr immunity: Bacteria and beneficial plasmids. *PLOS Genetics*, 9(9):e1003844, September 2013.

[52] David Burstein, Christine L. Sun, Christopher T. Brown, Itai Sharon, Karthik Anantharaman, Alexander J. Probst, Brian C. Thomas, and Jillian F. Banfield. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications*, 7:10613, February 2016.

[53] David Burstein, Lucas B. Harrington, Steven C. Strutt, Alexander J. Probst, Karthik Anantharaman, Brian C. Thomas, Jennifer A. Doudna, and Jillian F. Banfield. New CRISPR–Cas systems from uncultivated microbes. *Nature*, 542(7640):237–241, February 2017.

[54] Lin-Xing Chen, Basem Al-Shayeb, Raphaël Méheust, Wen-Jun Li, Jennifer A Doudna, and Jillian F Banfield. Candidate Phyla Radiation Roizmanbacteria from hot springs have novel and unexpectedly abundant CRISPR-Cas systems. *Frontiers in Microbiology*, 10:928, 2019.

[55] Aude Bernheim, David Bikard, Marie Touchon, and Eduardo PC Rocha. Co-occurrence of multiple CRISPRs and *cas* clusters suggests epistatic interactions. *bioRxiv*, page 592600, 2019.

[56] Philippe Horvath, Anne-Claire Coûté-Monvoisin, Dennis A. Romero, Patrick Boyaval, Christophe Fremaux, and Rodolphe Barrangou. Comparative analysis of CRISPR loci in lactic acid bacteria genomes. *International Journal of Food Microbiology*, 131(1):62–70, April 2009.

[57] C. Diez-Villasenor, C. Almendros, J. Garcia-Martinez, and F. J. M. Mojica. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology*, 156(5):1351–1361, May 2010.

[58] Fei Cai, Seth D. Axen, and Cheryl A. Kerfeld. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. *RNA Biology*, 10(5):687–693, May 2013.

[59] Daniel J Nasko, Barbra D Ferrell, Ryan M Moore, Jaysheel D Bhavsar, Shawn W Polson, and K Eric Wommack. CRISPR spacers indicate preferential matching of specific virioplankton genes. *mBio*, 10(2):e02651–18, 2019.

[60] Joe Bondy-Denomy, April Pawluk, Karen L. Maxwell, and Alan R. Davidson. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, 493(7432):429–432, January 2013.

[61] Ido Yosef, Moran G. Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research*, page gks216, March 2012.

[62] Pedro F. Vale, Guillaume Lafforgue, Francois Gatchitch, Rozenn Gardan, Sylvain Moineau, and Sylvain Gandon. Costs of CRISPR-Cas-mediated resistance in *Streptococcus thermophilus*. *Proceedings. Biological Sciences / The Royal Society*, 282(1812):20151270, August 2015.

[63] Edze R. Westra, Stineke van Houte, Sam Oyesiku-Blakemore, Ben Makin, Jenny M. Broniewski, Alex Best, Joseph Bondy-Denomy, Alan Davidson, Mike Boots, and Angus Buckling. Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Current Biology*, 25(8):1043–1049, April 2015.

[64] Ellinor O Alseth, Elizabeth Pursey, Adela M Luján, Isobel McLeod, Clare Rollie, and Edze R Westra. Bacterial biodiversity drives the evolution of CRISPR-based phage resistance in *Pseudomonas aeruginosa*. *bioRxiv*, page 586115, 2019.

[65] Yong Joon Chung, Christel Krueger, David Metzgar, and Milton H. Saier. Size comparisons among integral membrane transport protein homologues in bacteria, archaea, and eucarya. *Journal of Bacteriology*, 183(3):1012–1021, February 2001.

[66] Luciano Brocchieri and Samuel Karlin. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research*, 33(10):3390–3400, 2005.

[67] Heidi Ledford. Five big mysteries about CRISPR's origins. *Nature News*, 541(7637):280, January 2017.

[68] Kira S. Makarova, Daniel H. Haft, Rodolphe Barrangou, Stan J. J. Brouns, Emmanuelle Charpentier, Philippe Horvath, Sylvain Moineau, Francisco J. M. Mojica, Yuri I. Wolf, Alexander F. Yakunin, John van der Oost, and Eugene V. Koonin. Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, 9(6):467–477, June 2011.

[69] Jake L. Weissman, Rayshawn Holmes, Rodolphe Barrangou, Sylvain Moineau, William F. Fagan, Bruce Levin, and Philip L. F. Johnson. Immune loss as a driver of coexistence during host-phage coevolution. *The ISME Journal*, 12(2):585–597, February 2018.

[70] Jacob H. Munson-McGee, Shengyun Peng, Samantha Dewerff, Ramunas Stepanauskas, Rachel J. Whitaker, Joshua S. Weitz, and Mark J. Young. A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *The ISME Journal*, page 1, February 2018.

[71] Simon J. Labrie, Julie E. Samson, and Sylvain Moineau. Bacteriophage resistance mechanisms. *Nature Reviews. Microbiology*, 8(5):317–327, May 2010.

[72] Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. The basic building blocks and evolution of CRISPR–Cas systems. *Biochemical Society Transactions*, 41(6):1392–1400, December 2013.

[73] David Bikard, Asma Hatoum-Aslan, Daniel Mucida, and Luciano A Marraffini. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell host & microbe*, 12(2):177–186, 2012.

[74] Aude Bernheim, Alicia Calvo-Villamañán, Clovis Basier, Lun Cui, Eduardo P. C. Rocha, Marie Touchon, and David Bikard. Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. *Nature Communications*, 8(1):2094, December 2017.

[75] Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *Nucleic Acids Research*, 44(21):10074–10090, December 2016.

[76] Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, January 2012.

[77] Ambarish Biswas, Raymond H.J. Staals, Sergio E. Morales, Peter C. Fineran, and Chris M. Brown. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, 17:356, 2016.

[78] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, December 2009.

[79] Richard J. Roberts, Tamas Vincze, Janos Posfai, and Dana Macelis. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research*, 38(suppl_1):D234–D236, January 2010.

[80] S. R. Eddy. Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[81] Doherty Aidan J., Jackson Stephen P., and Weller Geoffrey R. Identification of bacterial homologues of the Ku DNA repair proteins. *FEBS Letters*, 500(3):186–188, July 2001.

[82] L. Aravind and Eugene V. Koonin. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Research*, 11(8):1365–1374, August 2001.

[83] Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14):6614–6624, August 2016.

[84] Jenna Morgan Lang, Aaron E. Darling, and Jonathan A. Eisen. Phylogeny of bacterial and archaeal genomes using conserved genes: Supertrees and supermatrices. *PLOS ONE*, 8(4):e62510, April 2013.

[85] Aaron E. Darling, Guillaume Jospin, Eric Lowe, Frederick A. Matsen Iv, Holly M. Bik, and Jonathan A. Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243, January 2014.

[86] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):e9490, March 2010.

[87] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[88] Jesse H. Krijthe. *Rtsne: t-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*, 2015. R package version 0.15.

[89] Roberts David R., Bahn Volker, Ciuti Simone, Boyce Mark S., Elith Jane, Guillera-Arroita Gurutzeta, Hauenstein Severin, Lahoz-Monfort José J., Schröder Boris, Thuiller Wilfried, Warton David I., Wintle Brendan A., Hartig Florian, and Dormann Carsten F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, August 2017.

[90] A. P. Reynolds, G. Richards, B. de la Iglesia, and V. J. Rayward-Smith. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504, December 2006.

[91] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2018. R package version 2.0.7-1.

[92] Anthony R. Ives and Theodore Garland. Phylogenetic logistic regression for binary dependent variables. *Systematic Biology*, 59(1):9–26, January 2010.

[93] Lam si Tung Ho and Cécile Ané. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology*, 63(3):397–408, May 2014.

[94] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, 78(1-2):103–112, 2005.

[95] Olga Morozova, Olga Levina, Anneli Uusküla, and Robert Heimer. Comparison of subset selection methods in linear regression in the context of health-related quality of life and substance abuse in Russia. *BMC medical research methodology*, 15(1):71, 2015.

[96] Donald E. Farrar and Robert R. Glauber. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics*, 49(1):92–107, 1967.

[97] Muhammad Imdadullah, Muhammad Aslam, and Saima Altaf. mctest: An R package for detection of collinearity among regressors. *The R Journal*, 8(2), December 2016.

[98] Kim-Anh Lê Cao, Simon Boitard, and Philippe Besse. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253, June 2011.

[99] Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS Computational Biology*, 13(11):e1005752, November 2017.

[100] Florian Rohart, Aida Eslami, Nicholas Matigian, Stéphanie Bougeard, and Kim-Anh Lê Cao. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics*, 18:128, February 2017.

[101] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001.

[102] Andy Liaw, Matthew Wiener, and others. Classification and regression by randomForest. *R news*, 2(3), 2002.

[103] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960.

[104] Ruth E. Ley, Catherine A. Lozupone, Micah Hamady, Rob Knight, and Jeffrey I. Gordon. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, 6(10):776–788, October 2008.

[105] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, Anupriya Tripathi, Sean M. Gibbons, Gail Ackermann, Jose A. Navas-Molina, Stefan Janssen, Evguenia Kopylova, Yoshiki Vázquez-Baeza, Antonio González, James T. Morton, Siavash Mirarab, Zhenjiang Zech Xu, Lingjing Jiang, Mohamed F. Haroon, Jad Kanbar, Qiyun Zhu, Se Jin Song, Tomasz Kosciolek, Nicholas A. Bokulich, Joshua Lefler, Colin J. Brislawn, Gregory Humphrey, Sarah M. Owens, Jarrad Hampton-Marcell, Donna Berg-Lyons, Valerie McKenzie, Noah Fierer, Jed A. Fuhrman, Aaron Clauset, Rick L. Stevens, Ashley Shade, Katherine S. Pollard, Kelly D. Goodwin, Janet K. Jansson, Jack A. Gilbert, Rob Knight, The Earth Microbiome Project Consortium, Jose L. Agosto Rivera, Lisa Al-Moosawi, John Alverdy, Katherine R. Amato, Jason Andras, Largus T. Angenent, Dionysios A. Antonopoulos, Amy Apprill, David Armitage, Kate Ballantine, Jir?í Bárta, Julia K. Baum, Allison Berry, Ashish Bhatnagar, Monica Bhatnagar, Jennifer F. Biddle, Lucie Bittner, Bazartseren Boldgiv, Eric Bottos, Donal M. Boyer, Josephine Braun, William Brazelton, Francis Q. Brearley, Alexandra H. Campbell, J. Gregory Caporaso, Cesar Cardona, JoLynn Carroll, S. Craig Cary, Brenda B. Casper, Trevor C. Charles, Haiyan Chu, Danielle C. Claar, Robert G. Clark, Jonathan B. Clayton, Jose C. Clemente, Alyssa Cochran, Maureen L. Coleman, Gavin Collins, Rita R. Colwell, Mónica Contreras, Benjamin B. Crary, Simon Creer, Daniel A. Cristol, Byron C. Crump, Duoying Cui, Sarah E. Daly, Liliana Davalos, Russell D. Dawson, Jennifer Defazio, Frédéric Delsuc, Hebe M. Dionisi, Maria Gloria Dominguez-Bello, Robin Dowell, Eric A. Dubinsky, Peter O. Dunn, Danilo Ercolini, Robert E. Espinoza, Vanessa Ezenwa, Nathalie Fenner, Helen S.

Findlay, Irma D. Fleming, Vincenzo Fogliano, Anna Forsman, Chris Freeman, Elliot S. Friedman, Giancarlo Galindo, Liza Garcia, Maria Alexandra Garcia-Amado, David Garshelis, Robin B. Gasser, Gunnar Gerdts, Molly K. Gibson, Isaac Gifford, Ryan T. Gill, Tugrul Giray, Antje Gittel, Peter Golyshin, Donglai Gong, Hans-Peter Grossart, Kristina Guyton, Sarah-Jane Haig, Vanessa Hale, Ross Stephen Hall, Steven J. Hallam, Kim M. Handley, Nur A. Hasan, Shane R. Haydon, Jonathan E. Hickman, Glida Hidalgo, Kirsten S. Hofmockel, Jeff Hooker, Stefan Hulth, Jenni Hultman, Embriette Hyde, Juan Diego Ibáñez-Álamo, Julie D. Jastrow, Aaron R. Jex, L. Scott Johnson, Eric R. Johnston, Stephen Joseph, Stephanie D. Jurburg, Diogo Jurelevicius, Anders Karlsson, Roger Karlsson, Seth Kauppinen, Colleen T. E. Kellogg, Suzanne J. Kennedy, Lee J. Kerkhof, Gary M. King, George W. Kling, Anson V. Koehler, Monika Krezalek, Jordan Kueneman, Regina Lamendella, Emily M. Landon, Kelly Lane-deGraaf, Julie LaRoche, Peter Larsen, Bonnie Laverock, Simon Lax, Miguel Lentino, Iris I. Levin, Pierre Liancourt, Wenju Liang, Alexandra M. Linz, David A. Lipson, Yongqin Liu, Manuel E. Lladser, Mariana Lozada, Catherine M. Spirito, Walter P. MacCormack, Aurora MacRae-Crerar, Magda Magris, Antonio M. Martín-Platero, Manuel Martín-Vivaldi, L. Margarita Martínez, Manuel Martínez-Bueno, Ezequiel M. Marzinelli, Olivia U. Mason, Gregory D. Mayer, Jamie M. McDevitt-Irwin, James E. McDonald, Krista L. McGuire, Katherine D. McMahon, Ryan McMinds, Mónica Medina, Joseph R. Mendelson, Jessica L. Metcalf, Folker Meyer, Fabian Michelangeli, Kim Miller, David A. Mills, Jeremiah Minich, Stefano Mocali, Lucas Moitinho-Silva, Anni Moore, Rachael M. Morgan-Kiss, Paul Munroe, David Myrold, Josh D. Neufeld, Yingying Ni, Graeme W. Nicol, Shaun Nielsen, Jozef I. Nissimov, Kefeng Niu, Matthew J. Nolan, Karen Noyce, Sarah L. O'Brien, Noriko Okamoto, Ludovic Orlando, Yadira Ortiz Castellano, Olayinka Osuolale, Wyatt Oswald, Jacob Parnell, Juan M. Peralta-Sánchez, Peter Petraitis, Catherine Pfister, Elizabeth Pilon-Smits, Paola Piombino, Stephen B. Pointing, F. Joseph Pollock, Caitlin Potter, Bharath Prithiviraj, Christopher Quince, Asha Rani, Ravi Ranjan, Subramanya Rao, Andrew P. Rees, Miles Richardson, Ulf Riebesell, Carol Robinson, Karl J. Rockne, Selena Marie Rodriguezl, Forest Rohwer, Wayne Roundstone, Rebecca J. Safran, Naseer Sangwan, Virginia Sanz, Matthew Schrenk, Mark D. Schrenzel, Nicole M. Scott, Rita L. Seger, Andaine Seguin-Orlando, Lucy Seldin, Lauren M. Seyler, Baddr Shakhsheer, Gabriela M. Sheets, Congcong Shen, Yu Shi, Hakdong Shin, Benjamin D. Shogan, Dave Shutler, Jeffrey Siegel, Steve Simmons, Sara Sjöling, Daniel P. Smith, Juan J. Soler, Martin Sperling, Peter D. Steinberg, Brent Stephens, Melita A. Stevens, Safiyh Taghavi, Vera Tai, Karen Tait, Chia L. Tan, Neslihan Tas¸, D. Lee Taylor, Torsten Thomas, Ina Timling, Benjamin L. Turner, Tim Urich, Luke K. Ursell, Daniel van der Lelie, William Van Treuren, Lukas van Zwieten, Daniela Vargas-Robles, Rebecca Vega Thurber, Paola Vitaglione, Donald A. Walker, William A. Walters, Shi Wang, Tao Wang, Tom Weaver, Nicole S. Webster, Beck Wehrle, Pamela Weisenhorn, Sophie Weiss, Jeffrey J. Werner, Kristin

West, Andrew Whitehead, Susan R. Whitehead, Linda A. Whittingham, Eske Willerslev, Allison E. Williams, Stephen A. Wood, Douglas C. Woodhams, Yeqin Yang, Jesse Zaneveld, Iratxe Zarraonaindia, Qikun Zhang, and Hongxia Zhao. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, November 2017.

[106] Adrian G. Patterson, Simon A. Jackson, Corinda Taylor, Gary B. Evans, George P. C. Salmond, Rita Przybilski, Raymond H. J. Staals, and Peter C. Fineran. Quorum sensing controls adaptive immunity through the regulation of multiple CRISPR-Cas systems. *Molecular Cell*, 64(6):1102–1108, December 2016.

[107] C. Condon, D. Liveris, C. Squires, I. Schwartz, and C. L. Squires. rRNA operon multiplicity in *Escherichia coli* and the physiological implications of *rrn* inactivation. *Journal of Bacteriology*, 177(14):4152–4156, July 1995.

[108] Sara Vieira-Silva and Eduardo P. C. Rocha. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLOS Genetics*, 6(1):e1000808, January 2010.

[109] Benjamin R. K. Roller, Steven F. Stoddard, and Thomas M. Schmidt. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nature Microbiology*, 1(11):16160, November 2016.

[110] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R Mende, Adriana Alberti, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015.

[111] Zarir E. Karanjawala, Niamh Murphy, David R. Hinton, Chih-Lin Hsieh, and Michael R. Lieber. Oxygen metabolism causes chromosome breaks and is associated with the neuronal apoptosis observed in DNA double-strand break repair mutants. *Current Biology*, 12(5):397–402, March 2002.

[112] Robert S. Pitcher, Nigel C. Brissett, and Aidan J. Doherty. Nonhomologous end-joining in bacteria: a microbial perspective. *Annual Review of Microbiology*, 61:259–282, 2007.

[113] E Jończyk, M Kłak, R Międzybrodzki, and A Górski. The influence of external factors on bacteriophages. *Folia microbiologica*, 56(3):191–200, 2011.

[114] Guilhem Faure, Kira S Makarova, and Eugene V Koonin. CRISPR-Cas: Complex functional networks and multiple roles beyond adaptive immunity. *Journal of Molecular Biology*, 2018.

[115] Grigory L Dianov, Tatyana V Timehenko, Olga I Sinitsina, Andrew V Kuzminov, Oleg A Medvedev, and Rudolf I Salganik. Repair of uracil residues closely spaced on the opposite strands of plasmid DNA results in double-strand break

and deletion formation. *Molecular and General Genetics MGG*, 225(3):448–452, 1991.

[116] Stanislav G Kozmin, Yuliya Sedletska, Anne Reynaud-Angelin, Didier Gasparutto, and Evelyne Sage. The formation of double-strand breaks at multiply damaged sites is driven by the kinetics of excision/incision at base damage in eukaryotic cells. *Nucleic Acids Research*, 37(6):1767–1777, 2009.

[117] Yuzhi Hong, Liping Li, Gan Luan, Karl Drlica, and Xilin Zhao. Contribution of reactive oxygen species to thymineless death in *Escherichia coli*. *Nature Microbiology*, 2(12):1667, 2017.

[118] Sarah S Henrikus, Camille Henry, John P McDonald, Yvonne Hellmich, Elizabeth A Wood, Roger Woodgate, Michael M Cox, Antoine M van Oijen, Harshad Ghodke, and Andrew Robinson. DNA double-strand breaks induced by reactive oxygen species promote DNA polymerase iv activity in *Escherichia coli*. *bioRxiv*, page 533422, 2019.

[119] Tulip Mahaseth and Andrei Kuzminov. Prompt repair of hydrogen peroxide-induced DNA lesions prevents catastrophic chromosomal fragmentation. *DNA repair*, 41:42–53, 2016.

[120] Thomas Bonura, Christopher D Town, Kendric C Smith, and Henry S Kaplan. The influence of oxygen on the yield of DNA double-strand breaks in x-irradiated *Escherichia coli K-12*. *Radiation research*, 63(3):567–577, 1975.

[121] Michael J Tilby and Pamela S Loverock. Measurements of DNA double-strand break yields in *E. coli* after rapid irradiation and cell inactivation: the effects of inactivation technique and anoxic radiosensitizers. *Radiation research*, 96(2):309–321, 1983.

[122] GP Van der Schans and Joh Blok. The influence of oxygen and sulphhydryl compounds on the production of breaks in bacteriophage DNA by gamma-rays. *International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine*, 17(1):25–38, 1970.

[123] Robert S Pitcher, Andrew J Green, Anna Brzostek, Malgorzata Korycka-Machala, Jaroslaw Dziadek, and Aidan J Doherty. NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation. *DNA repair*, 6(9):1271–1276, 2007.

[124] Pierre Dupuy, Benjamin Gourion, Laurent Sauviac, and Claude Bruand. DNA double-strand break repair is involved in desiccation resistance of *Sinorhizobium meliloti*, but is not essential for its symbiotic interaction with *Medicago truncatula*. *Microbiology*, 163(3):333–342, 2017.

[125] Carsten T Charlesworth, Priyanka S Deshpande, Daniel P Dever, Joab Camarena, Viktor T Lemgart, M Kyle Cromer, Christopher A Vakulskas,

Michael A Collingwood, Liyang Zhang, Nicole M Bode, et al. Identification of preexisting adaptive immunity to Cas9 proteins in humans. *Nature medicine*, 25(2):249, 2019.

[126] Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. A new view of the tree of life. *Nature Microbiology*, 1:16048, April 2016.

[127] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25, 2000.

[128] Marcus C Chibucos, Adrienne E Zweifel, Jonathan C Herrera, William Meza, Shabnam Eslamfam, Peter Uetz, Deborah A Siegele, James C Hu, and Michelle G Giglio. An ontology for microbial phenotypes. *BMC Microbiology*, 14(1):294, 2014.

[129] V H TierrafrÃŋa, C MejÃŋa-Almonte, J M Camacho-Zaragoza, H Salgado, K Alquicira, S Gama-Castro, C Ishida, and J Collado-Vides. Mco: towards an ontology and unified vocabulary for a framework-based annotation of microbial growth conditions. *Bioinformatics*, page bty689, 2018.

[130] Jaime Iranzo, Alexander E. Lobkovsky, Yuri I. Wolf, and Eugene V. Koonin. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. *BMC Evolutionary Biology*, 15:43, 2015.

[131] Moran Goren, Ido Yosef, Rotem Edgar, and Udi Qimron. The bacterial CRISPR/Cas system as analog of the mammalian adaptive immune system. *RNA Biology*, 9(5):549–554, May 2012.

[132] Luciano A. Marraffini and Erik J. Sontheimer. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science (New York, N.Y.)*, 322(5909):1843–1845, December 2008.

[133] Luciano A. Marraffini. CRISPR-Cas immunity in prokaryotes. *Nature*, 526(7571):55–61, October 2015.

[134] Rotem Sorek, C. Martin Lawrence, and Blake Wiedenheft. CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annual Review of Biochemistry*, 82(1):237–266, 2013.

[135] Pierre Boudry, Ekaterina Semenova, Marc Monot, Kirill A. Datsenko, Anna Lopatina, Ognjen Sekulovic, Maicol Ospina-Bedoya, Louis-Charles Fortier,

Konstantin Severinov, Bruno Dupuy, and Olga Soutourina. Function of the CRISPR-Cas system of the human pathogen *Clostridium difficile*. *mBio*, 6(5):e01112–15, October 2015.

[136] Joakim M. Andersen, Madelyn Shoup, Cathy Robinson, Robert Britton, Katharina E. P. Olsen, and Rodolphe Barrangou. CRISPR diversity and microevolution in *Clostridium difficile*. *Genome Biology and Evolution*, 8(9):2841–2855, September 2016.

[137] A. Mira, H. Ochman, and N. A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in genetics: TIG*, 17(10):589–596, October 2001.

[138] Chih-Horng Kuo and Howard Ochman. Deletional bias across the three domains of life. *Genome Biology and Evolution*, 1:145–152, January 2009.

[139] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(Database issue):D733–D745, January 2016.

[140] Cristóbal Almendros, Francisco JM Mojica, César Díez-Villaseñor, Noemí M Guzmán, and Jesús García-Martínez. CRISPR-Cas functional module exchange in *Escherichia coli*. *MBio*, 5(1):e00767–13, 2014.

[141] Kirill A. Datsenko, Ksenia Pougach, Anton Tikhonov, Barry L. Wanner, Konstantin Severinov, and Ekaterina Semenova. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications*, 3:945, July 2012.

[142] Daan C. Swarts, Cas Mosterd, Mark W. J. van Passel, and Stan J. J. Brouns. CRISPR interference directs strand specific spacer acquisition. *PLoS ONE*, 7(4):e35888, April 2012.

[143] Roger A. Garrett, Shiraz A. Shah, Gisle Vestergaard, Ling Deng, Soley Gudbergsdottir, Chandra S. Kenchappa, Susanne Erdmann, and Qunxin She. CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochemical Society Transactions*, 39(1):51–57, February 2011.

[144] Soley Gudbergsdottir, Ling Deng, Zhengjun Chen, Jaide V. K. Jensen, Linda R. Jensen, Qunxin She, and Roger A. Garrett. Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Molecular Microbiology*, 79(1):35–49, January 2011.

[145] David L. Bernick, Courtney L. Cox, Patrick P. Dennis, and Todd M. Lowe. Comparative genomic and transcriptional analyses of CRISPR systems across the genus *Pyrobaculum*. *Frontiers in Microbiology*, 3, July 2012.

[146] Caryn R. Hale, Sonali Majumdar, Joshua Elmore, Neil Pfister, Mark Compton, Sara Olson, Alissa M. Resch, Claiborne V. C. Glover, Brenton R. Graveley, Rebecca M. Terns, and Michael P. Terns. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Molecular Cell*, 45(3):292–302, February 2012.

[147] Hagen Richter, Judith Zoephel, Jeanette Schermuly, Daniel Maticzka, Rolf Backofen, and Lennart Randau. Characterization of CRISPR RNA processing in *Clostridium thermocellum* and *Methanococcus maripaludis*. *Nucleic Acids Research*, 40(19):9887–9896, October 2012.

[148] Bridget NJ Watson, Raymond HJ Staals, and Peter C Fineran. CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio*, 9(1):e02406–17, 2018.

[149] Jeff Nivala, Seth L Shipman, and George M Church. Spontaneous CRISPR loci generation in vivo by non-canonical spacer integration. *Nature Microbiology*, page 1, 2018.

[150] David Paez-Espino, Itai Sharon, Wesley Morovic, Buffy Stahl, Brian C. Thomas, Rodolphe Barrangou, and Jillian F. Banfield. CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio*, 6(2):e00262–15, May 2015.

[151] Cheryl-Emiliane T Chow and Jed A Fuhrman. Seasonality and monthly dynamics of marine myovirus communities. *Environmental microbiology*, 14(8):2171–2183, 2012.

[152] M. Senthil Kumar, Joshua B. Plotkin, and Sridhar Hannenhalli. Regulated CRISPR modules exploit a dual defense strategy of restriction and abortive infection in a model of prokaryote-phage coevolution. *PLoS Computational Biology*, 11(11):e1004603, November 2015.

[153] Asaf Levy, Moran G. Goren, Ido Yosef, Oren Auster, Miriam Manor, Gil Amitai, Rotem Edgar, Udi Qimron, and Rotem Sorek. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature*, 520(7548):505–510, April 2015.

[154] Alexander P. Hynes, Manuela Villion, and Sylvain Moineau. Adaptation in bacterial CRISPR-Cas immunity can be driven by defective phages. *Nature Communications*, 5:4399, July 2014.

[155] Marie Touchon and Eduardo P. C. Rocha. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE*, 5(6):e11126, June 2010.

[156] Marie Touchon, Sophie Charpentier, Olivier Clermont, Eduardo P. C. Rocha, Erick Denamur, and Catherine Branger. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *Journal of Bacteriology*, 193(10):2460–2467, May 2011.

[157] Rongpeng Li, Lizhu Fang, Shirui Tan, Min Yu, Xuefeng Li, Sisi He, Yuquan Wei, Guoping Li, Jianxin Jiang, and Min Wu. Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Research*, 26(12):1273–1287, December 2016.

[158] Alexander Martynov, Konstantin Severinov, and Iaroslav Ispolatov. Optimal number of spacers in CRISPR arrays. *PLoS Computational Biology*, 13(12):e1005891, 2017.

[159] Ibtissem Grissa, Gilles Vergnaud, and Christine Pourcel. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC bioinformatics*, 8:172, May 2007.

[160] Nora C Pyenson, Kaitlyn Gayvert, Andrew Varble, Olivier Elemento, and Luciano A Marraffini. Broad targeting specificity during bacterial type III CRISPR-Cas immunity constrains viral escape. *Cell host & microbe*, 22(3):343–353, 2017.

[161] Aris-Edda Stachler and Anita Marchfelder. Gene repression in Haloarchaea using the CRISPR (clustered regularly interspaced short palindromic repeats) - Cas I-B system. *Journal of Biological Chemistry*, page jbc.M116.724062, May 2016.

[162] Aris-Edda Stachler, Israela Turgeman-Grott, Ella Shtifman-Segal, Thorsten Allers, Anita Marchfelder, and Uri Gophna. High tolerance to self-targeting of the genome by the endogenous CRISPR-Cas system in an archaeon. *Nucleic Acids Research*, March 2017.

[163] April Pawluk, Joseph Bondy-Denomy, Vivian HW Cheung, Karen L Maxwell, and Alan R Davidson. A new group of phage anti-crispr genes inhibits the type ie CRISPR-Cas system of *Pseudomonas aeruginosa*. *MBio*, 5(2):e00896–14, 2014.

[164] Sukrit Silas, Patricia Lucas-Elio, Simon A Jackson, Alejandra Aroca-Crevillén, Loren L Hansen, Peter C Fineran, Andrew Z Fire, and Antonio Sánchez-Amat.

Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. *eLife*, 6, 2017.

[165] Anne Kupczok, Giddy Landan, and Tal Dagan. The contribution of genetic recombination to CRISPR array evolution. *Genome Biology and Evolution*, 7(7):1925–1939, July 2015.

[166] Raymond H. J. Staals, Simon A. Jackson, Ambarish Biswas, Stan J. J. Brouns, Chris M. Brown, and Peter C. Fineran. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nature Communications*, 7:12853, October 2016.

[167] Haiyan Zeng, Jumei Zhang, Chensi Li, Tengfei Xie, Na Ling, Qingping Wu, and Yingwang Ye. The driving force of prophages and CRISPR-Cas system in the evolution of *Cronobacter sakazakii*. *Scientific Reports*, 7:40206, January 2017.

[168] William B. Whitman, David C. Coleman, and William J. Wiebe. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, 95(12):6578–6583, June 1998.

[169] Patrick D. Schloss, Rene A. Girard, Thomas Martin, Joshua Edwards, and J. Cameron Thrash. Status of the archaeal and bacterial census: an update. *mBio*, 7(3):e00201–16, July 2016.

[170] Steven W. Wilhelm and Curtis A. Suttle. Viruses and nutrient cycles in the sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, 49(10):781, Oct 1999.

[171] K E Wommack and R R Colwell. Virioplankton: viruses in aquatic ecosystems. *Microbiology and Molecular Biology Reviews : MMBR*, 64:69–114, March 2000.

[172] Curtis A. Suttle. Viruses in the sea. *Nature*, 437(7057):356âĂŞ361, Sep 2005.

[173] Joshua S. Weitz and Steven W. Wilhelm. Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biology Reports*, 4, September 2012.

[174] Charles H Wigington, Derek Sonderegger, Corina P D Brussaard, Alison Buchan, Jan F Finke, Jed A Fuhrman, Jay T Lennon, Mathias Middelboe, Curtis A Suttle, Charles Stock, William H Wilson, K Eric Wommack, Steven W Wilhelm, and Joshua S Weitz. Re-examination of the relationship between marine virus and microbial cell abundances. *Nature Microbiology*, 1:15024, January 2016.

[175] Øivind Bergh, Knut Yngve BØrsheim, Gunnar Bratbak, and Mikal Heldal. High abundance of viruses found in aquatic environments. *Nature*, 340(6233):467–468, August 1989.

[176] John McN. Sieburth, Paul W. Johnson, and Paul E. Hargraves. Ultrastructure and ecology of *Aureococcus anophageferens* gen. et sp. nov. (chrysophyceae): The dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, summer 1985. *Journal of Phycology*, 24(3):416–425, September 1988.

[177] Lita M. Proctor and Jed A. Fuhrman. Viral mortality of marine bacteria and cyanobacteria. *Nature*, 343(6253):60–62, January 1990.

[178] Gunnar Bratbak, Mikal Heldal, Svein Norland, and T. Frede Thingstad. Viruses as Partners in Spring Bloom Microbial Trophodynamics. *Applied and Environmental Microbiology*, 56(5):1400–1405, May 1990.

[179] Gunnar Bratbak, T. Frede Thingstad, and Mikal Heldal. Viruses and the microbial loop. *Microbial Ecology*, 28(2):209–221, 1994.

[180] Markus G Weinbauer and Fereidoun Rassoulzadegan. Are viruses driving microbial diversification and diversity? *Environmental microbiology*, 6:1–11, January 2004.

[181] S. J. Schrag and J. E. Mittler. Host-parasite coexistence: The role of spatial refuges in stabilizing bacteria-phage interactions. *The American Naturalist*, 148(2):348–377, 1996.

[182] Richard E. Lenski. Coevolution of bacteria and phage: Are there endless cycles of bacterial defenses and phage counterdefenses? *Journal of Theoretical Biology*, 108(3):319–325, June 1984.

[183] Richard E. Lenski and Bruce R. Levin. Constraints on the coevolution of bacteria and virulent phage: A model, some experiments, and predictions for natural communities. *The American Naturalist*, 125(4):585–602, 1985.

[184] Alex R. Hall, Pauline D. Scanlan, Andrew D. Morgan, and Angus Buckling. Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecology Letters*, 14(7):635–642, July 2011.

[185] Stineke van Houte, Alice K. E. Ekroth, Jenny M. Broniewski, Hélène Chabas, Ben Ashby, Joseph Bondy-Denomy, Sylvain Gandon, Mike Boots, Steve Paterson, Angus Buckling, and Edze R. Westra. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature*, 532(7599):385–388, April 2016.

[186] John B. Waterbury and Frederica W. Valois. Resistance to co-occurring phages enables marine synechococcus communities to coexist with cyanophages abundant in seawater. *Applied and Environmental Microbiology*, 59(10):3393–3399, October 1993.

[187] Pedro Gómez and Angus Buckling. Bacteria-phage antagonistic coevolution in soil. *Science*, 332(6025):106–109, April 2011.

[188] M. T. Horne. Coevolution of *Escherichia coli* and bacteriophages in chemostat culture. *Science*, 168(3934):992–993, 1970.

[189] Simon A. Levin and J. Daniel Udovic. A mathematical model of coevolving populations. *The American Naturalist*, 111(980):657–675, 1977.

[190] Lin Chao, Bruce R. Levin, and Frank M. Stewart. A complex community in a simple habitat: An experimental study with bacteria and phage. *Ecology*, 58(2):369–378, March 1977.

[191] Brendan J. M. Bohannan, Richard E. Lenski, and Associate Editor: Robert D. Holt. Effect of prey heterogeneity on the response of a model food chain to resource enrichment. *The American Naturalist*, 153(1):73–82, 1999.

[192] Yan Wei, Amy Kirby, Bruce R. Levin, Associate Editor: Pejman Rohani, and Editor: Judith L. Bronstein. The population and evolutionary dynamics of *Vibrio cholerae* and its bacteriophage: Conditions for maintaining phage-limited communities. *The American Naturalist*, 178(6):715–725, 2011.

[193] L. Van Valen. A new evolutionary law. *Evolutionary Theory*, 1:1–30, 1973.

[194] Leigh Van Valen. Molecular evolution as predicted by natural selection. *Journal of Molecular Evolution*, 3(2):89–101, June 1974.

[195] Aneil Agrawal and Curtis M. Lively. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evolutionary Ecology Research*, 4(1):91–107, 2002.

[196] S. Gandon, A. Buckling, E. Decaestecker, and T. Day. Host–parasite coevolution and patterns of adaptation across time and space. *Journal of Evolutionary Biology*, 21(6):1861–1866, November 2008.

[197] A. Buckling and P. B. Rainey. Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1494):931–936, May 2002.

[198] Bruce R. Levin, Frank M. Stewart, and Lin Chao. Resource-limited growth, competition, and predation: A model and experimental studies with bacteria and bacteriophage. *The American Naturalist*, 111(977):3–24, 1977.

[199] Luis F. Jover, Michael H. Cortez, and Joshua S. Weitz. Mechanisms of multi-strain coexistence in host–phage systems with nested infection networks. *Journal of Theoretical Biology*, 332:65–77, September 2013.

[200] Justin R. Meyer, Devin T. Dobias, Sarah J. Medina, Lisa Servilio, Animesh Gupta, and Richard E. Lenski. Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science*, 354(6317):1301–1304, December 2016.

[201] Takehito Yoshida, Stephen P. Ellner, Laura E. Jones, Brendan J. M. Bohannan, Richard E. Lenski, and Nelson G. Hairston Jr. Cryptic population dynamics: Rapid evolution masks trophic interactions. *PLoS Biology*, 5(9):e235, September 2007.

[202] Laura E. Jones and Stephen P. Ellner. Effects of rapid prey evolution on predator–prey cycles. *Journal of Mathematical Biology*, 55(4):541–573, May 2007.

[203] M. Delbrück. Bacterial viruses or bacteriophages. *Biological Reviews*, 21(1):30–40, January 1946.

[204] Richard E. Lenski. Dynamics of interactions between bacteria and virulent bacteriophage. In K. C. Marshall, editor, *Advances in Microbial Ecology*, number 10 in Advances in Microbial Ecology, pages 1–44. Springer US, 1988.

[205] Justin R. Meyer, Devin T. Dobias, Joshua S. Weitz, Jeffrey E. Barrick, Ryan T. Quick, and Richard E. Lenski. Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, 335(6067):428–432, January 2012.

[206] Joshua S. Weitz. *Quantitative Viral Ecology: Dynamics of Viruses and Their Microbial Hosts*. Princeton University Press, January 2016. Google-Books-ID: 0zNJCgAAQBAJ.

[207] Reuben B. Vercoe, James T. Chang, Ron L. Dy, Corinda Taylor, Tamzin Gristwood, James S. Clulow, Corinna Richter, Rita Przybilski, Andrew R. Pitman, and Peter C. Fineran. Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLOS Genetics*, 9(4):e1003454, April 2013.

[208] Ron L. Dy, Andrew R. Pitman, and Peter C. Fineran. Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. *Mobile Genetic Elements*, 3(5):e26831, September 2013.

[209] Josiane E. Garneau, Marie-Ève Dupuis, Manuela Villion, Dennis A. Romero, Rodolphe Barrangou, Patrick Boyaval, Christophe Fremaux, Philippe Horvath, Alfonso H. Magadán, and Sylvain Moineau. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 468(7320):67–71, November 2010.

[210] Hélène Deveau, Rodolphe Barrangou, Josiane E. Garneau, Jessica Labonté, Christophe Fremaux, Patrick Boyaval, Dennis A. Romero, Philippe Horvath, and Sylvain Moineau. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology*, 190(4):1390–1400, February 2008.

[211] Philippe Horvath, Dennis A. Romero, Anne-Claire Coûté-Monvoisin, Melissa Richards, Hélène Deveau, Sylvain Moineau, Patrick Boyaval, Christophe Fremaux, and Rodolphe Barrangou. Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *Journal of Bacteriology*, 190(4):1401–1412, February 2008.

[212] Philip J. Gerrish and Richard E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103(0):127, 1998.

[213] Michael M. Desai, Daniel S. Fisher, and Andrew W. Murray. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17(5):385–394, March 2007.

[214] Lauren M. Childs, Nicole L. Held, Mark J. Young, Rachel J. Whitaker, and Joshua S. Weitz. Multiscale model of CRISPR-induced coevolutionary dynamics: Diversification at the interface of Lamarck and Darwin. *Evolution*, 66(7):2015–2029, July 2012.

[215] Serena Bradde, Marija Vucelja, Tiberiu Teşileanu, and Vijay Balasubramanian. Dynamics of adaptive immunity against phage in bacterial populations. *PLoS Computational Biology*, 13:e1005486, April 2017.

[216] Hélène Chabas, Stineke van Houte, Nina Molin Høyland-Kroghsbo, Angus Buckling, and Edze R. Westra. Immigration of susceptible hosts triggers the evolution of alternative parasite defence strategies. *Proceedings. Biological Sciences / The Royal Society*, 283(1837), August 2016.

[217] Lauren M. Childs, Whitney E. England, Mark J. Young, Joshua S. Weitz, and Rachel J. Whitaker. CRISPR-induced distributed immunity in microbial populations. *PLoS ONE*, 9(7):e101710, July 2014.

[218] Kelli L. Palmer and Michael S. Gilmore. Multidrug-resistant Enterococci lack CRISPR-Cas. *mBio*, 1(4):e00227–10, October 2010.

[219] Ekaterina Semenova, Matthijs M. Jore, Kirill A. Datsenko, Anna Semenova, Edze R. Westra, Barry Wanner, John van der Oost, Stan J. J. Brouns, and Konstantin Severinov. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences*, 108(25):10098–10103, June 2011.

[220] Bruno Martel and Sylvain Moineau. CRISPR-Cas: an efficient tool for genome engineering of virulent bacteriophages. *Nucleic Acids Research*, 42(14):9504–9513, 2014.

[221] Martin T. Ferris, Paul Joyce, and Christina L. Burch. High frequency of mutations that expand the host range of an RNA virus. *Genetics*, 176(2):1013–1022, June 2007.

[222] Lin Chao. Fitness of RNA virus decreased by Muller's ratchet. *Nature*, 348(6300):454–455, November 1990.

[223] Ekaterina Semenova, Ekaterina Savitskaya, Olga Musharova, Alexandra Strotskaya, Daria Vorontsova, Kirill A. Datsenko, Maria D. Logacheva, and Konstantin Severinov. Highly efficient primed spacer acquisition from targets destroyed by the *Escherichia coli* type I-E CRISPR-Cas interfering complex. *Proceedings of the National Academy of Sciences*, 113(27):7626–7631, July 2016.

[224] James S. Godde and Amanda Bickerton. The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*, 62(6):718–729, June 2006.

[225] Sajib Chakraborty, Ambrosius P. Snijders, Rajib Chakravorty, Musaddeque Ahmed, Ashek Md. Tarek, and M. Anwar Hossain. Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. *Molecular Phylogenetics and Evolution*, 56(3):878–887, September 2010.

[226] Uri Gophna, David M Kristensen, Yuri I Wolf, Ovidiu Popa, Christine Drevet, and Eugene V Koonin. No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. *The ISME Journal*, 9(9):2021–2027, September 2015.

[227] Edze R. Westra, Andrea J. Dowling, Jenny M. Broniewski, and Stineke van Houte. Evolution and ecology of CRISPR. *Annual Review of Ecology, Evolution, and Systematics*, 47(1):307–331, 2016.

[228] Pablo Yarza, Michael Richter, Jörg Peplies, Jean Euzeby, Rudolf Amann, Karl-Heinz Schleifer, Wolfgang Ludwig, Frank Oliver Glöckner, and Ramon Rosselló-Móra. The All-Species Living Tree project: A 16s rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31(4):241–250, September 2008.

[229] Simon P Blomberg, Theodore Garland Jr, and Anthony R Ives. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745, 2003.

[230] Liam J Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223, 2012.

[231] Ekaterina Savitskaya, Anna Lopatina, Sofia Medvedeva, Mikhail Kapustin, Sergey Shmakov, Alexey Tikhonov, Irena I Artamonova, Maria Logacheva, and Konstantin Severinov. Dynamics of *Escherichia coli* type I-E CRISPR spacers over 42 000 years. *Molecular Ecology*, 26(7):2019–2026, 2017.

[232] Ibtissem Grissa, Gilles Vergnaud, and Christine Pourcel. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, 35(Web Server issue):W52–57, July 2007.

[233] Céline Lévesque, Martin Duplessis, Jessica Labonté, Steve Labrie, Christophe Fremaux, Denise Tremblay, and Sylvain Moineau. Genomic organization and molecular analysis of virulent bacteriophage 2972 infecting an exopolysaccharide-producing *Streptococcus thermophilus* strain. *Applied and Environmental Microbiology*, 71(7):4057–4068, July 2005.

[234] Chaoyou Xue, Arun S. Seetharam, Olga Musharova, Konstantin Severinov, Stan J. J. Brouns, Andrew J. Severin, and Dipali G. Sashital. CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Research*, 43(22):10831–10847, December 2015.

[235] Mason J. Van Orden, Peter Klein, Kesavan Babu, Fares Z. Najar, and Rakhi Rajan. Conserved DNA motifs in the type II-A CRISPR leader region. *PeerJ*, 5:e3161, 2017.

[236] Bruce R. Levin, Sylvain Moineau, Mary Bushman, and Rodolphe Barrangou. The population and evolutionary dynamics of phage and bacteria with CRISPR–mediated immunity. *PLoS Genet*, 9(3):e1003312, March 2013.

[237] Brendan J. M. Bohannan and Richard E. Lenski. Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. *Ecology*, 78(8):2303–2315, December 1997.