6-2020

From Genes to Ecosystems: Resource Availability and DNA Methylation Drive the Diversity and Abundance of Restriction Modification Systems in Prokaryotes

Spiridon E. Papoulis
University of Tennessee, Knoxville, spapouli@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Spiridon E. Papoulis entitled "From Genes to Ecosystems: Resource Availability and DNA Methylation Drive the Diversity and Abundance of Restriction Modification Systems in Prokaryotes." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Microbiology.

Erik R. Zinser, Major Professor

We have read this dissertation and recommend its acceptance:

Steven W. Wilhelm, David Talmy, Scott J. Emrich, J. Jeffrey Morris

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

From Genes to Ecosystems:
Resource Availability and DNA Methylation Drive the Diversity and Abundance of Restriction Modification Systems in Prokaryotes

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Spiridon Evangelos Papoulis
August 2020

I dedicate this work to my Mother and Baba, who I owe everything

ACKNOWLEDGEMENTS

Before reading this dissertation, you should be aware of the many people that contributed and supported me though this doctoral work. First, this dissertation would have not been possible without the guidance of my advisor, Erik Zinser, and the input of the committee members who all helped shape the ideas and direction of this dissertation and include David Talmy, Steven Wilhelm, Scott Emrich, and Jeff Morris. Jeff Morris deserves special thanks for encouraging me to apply for a research experience in the summer of 2014 at the University of Tennessee and a fantastic undergraduate research experience that I carried into my graduate work. Gary LeClair should be recognized to his exceptional commitment to the summer research program at the University of Tennessee, as it ultimately convinced me to attend the university for my graduate studies. I would like to also thank former committee members Barry Bruce and especially Igor Jouline, whom taught the initial computational skills at beginning of this dissertation. I would also like to thank Dr. and Mrs. Penely for their financial support of graduate student studies.

It cannot be emphasized enough how critical my family was in contributing to this dissertation. My parents' love and support throughout my life have been pivotal, because without it, none of this would have been possible. My wife, Katherine Moccia, not only contributed intellectually by helping me organize and assess my ideas for every chapter presented in this dissertation but has been the best partner I could have even imagined. Her limitless compassion and profound love have rooted me in happiness, and I am eternally thankful for every minute I experience it.

ABSTRACT

Together, prokaryotic hosts and their viruses numerically dominate the planet and are engaged in an eternal struggle of hosts evading viral predation and viruses overcoming defensive mechanisms employed by their hosts. Prokaryotic hosts have been found to carry several viral defense systems in recent years with Restriction Modification systems (RMs) were the first discovered in the 1950s. While we have biochemically elucidated many of these systems in the last 70 years, we still struggle to understand what drives their gain and loss in prokaryotic genomes. In this work, we take a computational approach to understand the underlying evolutionary drivers of RMs by assessing 'big data' signals of RMs in prokaryotic genomes and incorporating molecular data in trait-based mathematical models. Focusing on the Cyanobacteria, we found a large discrepancy in the frequency of RMs per genome in different environmental contexts, where Cyanobacteria that live in oligotrophic nutrient conditions have few to no RMs and those in nutrient-rich conditions consistently have many RMs. While our models agree with the observation that increased nutrient inputs make the selective pressure of RMs more intense, they were unable to reconcile the high numbers of RMs per genome with their potent defensive properties- a situation of apparent overkill. By incorporating viral methylation, an unavoidable effect of RMs, we were able to explain how organisms could carry over 15 RMs. With this discovery, we then tried and reassess the distribution of methyltransferases, an essential component of RMs that can also have alternate physiological rolls in the cell. We expand on conventional wisdom, that

methyltransferases that are widely phylogenetically conserved are associated with global cellular regulation. However, we also find that organisms with high numbers of RMs also have a surprising amount of conservation in the methyltransferases that they carry. This data suggests caution should be used in associating phylogenic signals with functional rolls in methyltransferases as different functional rolls seem to overlap in their phylogenetic signal. Indeed, we suggest trait-based modeling may be the best tool in elucidating why organisms with a high selective pressure to maintain RMs appear to have conserved methyltransferase.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Prokaryotic Hosts and Their Viruses

ABSTRACT

The microbial arms race between prokaryotic hosts and their viruses have produced an exceptional number of molecular mechanisms. In this review chapter, I discuss a few extracellular and several intracellular mechanisms that prokaryotic hosts employ to overcome their viral predators. With emphasis on intracellular defense systems, we find that there are two primary consistencies between all of them despite their vastly different defensive strategies: each seem to have a vast diversity in configurations, and all are horizontally transferred. Importantly, for nearly all intracellular defense systems, we have found viral solutions to counter host defenses, suggesting the pressure for innovation is incredibly strong between both hosts and viruses. Unfortunately, the variance and horizontal transfer of theses defense systems in prokaryotes makes them difficult to study when comparing different organisms, therefore, their exact evolutionary drivers remain elusive, even in the case of defense systems that have been know for 70 years, such as the Restriction Modification system.

I. Introduction: Escalation and Innovation in the Microbial Arms Race

In Lewis Carroll's Through the Looking-Glass, the Red Queen shows Alice that you must always run in Wonderland to stay in the same spot, and if you want to go somewhere else, you must move twice as fast[1]. The idea of running in-place was a completely foreign concept to Alice, and likely, to all that have never visited wonderland. Yet, this strangely fantastical idea is accurate metaphor for the interactions between organisms. When Dr. Leigh Van Valen described the Red Queen hypothesis, he described a zero-sum game between organisms, where no species can ever win and new adversaries replace losers in the struggle for finite resources[2]. Much like running in place, avoiding extinction is a constant struggle of adaption to acquire more resources, and innovations by competitors ultimately negatively impacts all others in competition. These underpinnings drive the engine of diversity between hosts and their respective viruses, where developments of new traits in one drives the innovations in another[3–6].

In this review chapter, we will give a brief overview on the defense mechanisms discovered in prokaryotes and how phages overcome these defensive barriers. We generalize these defense mechanisms into two major categories: those that prevent the entry of phage DNA/RNA into the cytosol and intracellular defense systems that protect hosts once the viral DNA/RNA has entered the cytosol. As will become apparent, each mechanism exemplifies the microbial arms race between hosts and viruses, and we have discovered at least one viral mechanism to overcome most host defensives. Although some of these systems have been extensively studied and biochemically elucidated, we still do not understand what drives the gain, primarily though horizontal gene transfer, and loss of these systems in prokaryotic genomes.

II. Prokaryotic Extracellular Defenses

To start the infective process, bacteriophage (also referred to as phage, viruses) must successfully attach to a prokaryotic cell and inject their DNA or RNA genome into the cytosol[5,7]. The interaction between the phage and bacterial cell is specific - phage take advantage of cell surface proteins and other strucutres as gateways into the cytosol. Phage targets are extremely diverse, including flagella tail fibers, cell wall structural proteins, porins, receptors, antibiotic efflux pumps, various lipid and polysaccharide cell wall moieties, and transporters in involved iron, vitamin, and lipid transporters among others[8,9]. Viral predation, however, selects for hosts that can resist phage attachment. For example, OmpA is a outer membrane protein that maintains structural integrity *Escherichia coli*, but also serves as an entry point for several phage[10]. Non-synonymous mutations to OmpA result in either complete resistance or reduction of infection of T-like phage. It is important to note that each mutation had different impacts on the infectivity of the 14 test phage, highlighting the diversity of the phage and the different ways they interact with OmpA. Moreover, just as the host is selected to resist phage attachment, phage are selected to attach to hosts. Mutations in the phage receptor binding proteins can not only change host specificity, but also overcome mutations to receptors to once again become infectious[11].

Other mechanisms have been found to prevent phage attachment via physically blocking phage adsorption cites. For example, phage adsorb though a cell-wall-anchored virulence factor in *Staphylococcus aureus*, however, *S. aureus* produces immunoglobulin G-binding protein A that masks the phage receptor and reduces adsorption[12]. *Escherichia coli* uses lipoproteins in a similar manner to mask OmpA from phage[13]. Interestingly,

phage can use physical blockages to their advantage as well. For example, T5 phage induce lipoprotein synthesis to block their own receptor to prevent super infection and to prevent newly produced virions from attaching to receptors on lysed cells[14]. Biofilms, composed of extracellular polymers, can also act as a physical barrier between phage and their adsorption cite[15]. Indeed, virions have been found to have polysaccharase activity, suggesting these viral particles can physically remove exopolysaccharides to find their receptor site[16].

III. Prokaryotic Intracellular Defense

Over the years, researchers have discovered several intracellular antiviral defense systems in prokaryotes. Many of these defense systems are foundational tools used by molecular biologists in the manipulations of DNA. In the context of viral predation, these systems offer a last-ditch effort to save the cell after a virus has successfully attached to the cell and injected nucleic acids into the cytosol. As described below, many of these systems are effective in reducing infection and, for most, we have discovered viral mechanisms that overcome these defense systems.

Restriction Modification Systems. Restriction Modification Systems (RMs) are innate microbial immune systems, typically composed of endonuclease and methyltransferase activities, and are arguably the best studied antiviral system in Prokaryotes[17]. Endonucleases hydrolyze the phosphodiester bond of the DNA sugar backbone, resulting in a DNA cutting function[18]. Unlike exonucleases, which cleave nucleotides in a stepwise like manner from either 5´ or 3´ ends, endonucleases target internal phosphodiester bonds within DNA. The cutting activity is discriminatory, however, as endonucleases are only catalytically activate at specific recognition

sequences within DNA, also called a restriction site. The other functional requirement of RMs is methyltransferase activity, which targets the same recognition sequence as the endonuclease. Methyltransferases transfer the methyl group from S-adenosyl methionine to the C-5 carbon, or the N^4 amino group of cytosine or to the N^6 amino group of adenine, and inhibit endonuclease activity[19]. Without methyltransferase activity, the endonuclease would function unhindered in a cell and digest the host chromosome, leading to cell death. Thus, through methylation, RMs can discriminate between the host DNA from which they are expressed and foreign DNA, such as plasmids or bacteriophage that have successfully injected their DNA into the cytosol.

While being some of the best studied proteins, RMs are incredibly diverse in domain and genomic architecture. RMs can be broken down into four types, each with their own unique properties emerging as a result of their underlying structure. Type I RMs are composed of *hsdR, hsdM,* and *hsdS* genes which are responsible for endonuclease activity, methyltransferase activity, and DNA recognition, respectively, where these gene names are specific to type I nomenclature[20]. To form a functional methyltransferase, a hetero-oligomeric enzyme complex is formed between two HsdM methyltransferase subunits and one HsdS recognition sequence subunit. This same complex is used to form endonuclease activity after the addition of two HsdR subunits in the presence of ATP. Type II RMs are composed of two genes, one for endonuclease activity while the other for methyltransferase activity, and are functionally independent. Type III RMs are composed of two genes, referred to as *mod* and *res* genes responsible for the methyltransferase and endonuclease activities, respectively[21]. The Mod subunit can independently recognize and methylate DNA but becomes an endonuclease when

complexed together with the Res subunit and requires ATP for hydrolysis. Lacking the methylation activity altogether, type IV systems are only composed of a single endonuclease; however, this endonuclease is only active towards methylated recognition sequences[22]. Of course, there are many examples of RMs that defy conventional classification, such as type IIG that have both endonuclease and methyltransferase domains fused in a single polypeptide[23].

RMs are extremely potent antiviral defense systems. For example, in the development of efficient transposon mutagenesis in *Nostoc* PCC7120 (formerly *Anabaena*), Elhai et al. shows that transformation efficiency is a function of the number of unmodified restriction sites[24]. Impressively, two out of the three endonucleases tested decreased transformation efficiency by an order of magnitude per restriction site, where the other required two restriction sites to drop the efficiency by an order of magnitude. Consistently, a study of BsuMI restriction showed that three unmethylated sites within a plasmid dropped conjugal transfer efficiency by 3.5 orders of magnitude[25], and a study of EcoRI showed that 4 unmethylated sites dropped conjugal efficiency by nearly 5 orders of magnitude[26]. Extending this logic to phage as the foreign DNA, it is unsurprising how a single RMs can reduce viral infection rates by almost 7 orders of magnitude[27]. RMs, however, act as a powerful selective force in viruses, selecting for mechanisms to evade these defense systems.

There are several different strategies employed by phage to avoid restriction endonucleases. Viruses can elevate digestion via point mutations to the restriction site, rendering endonucleases useless and producing an active infection, killing the host[28]. A more interesting mechanism involves the inherent weakness of RMs: methylation.

Methyltransferases are used by viruses to avoid host endonucleases, both passively and actively. Although RMs are highly effective, some virions will escape restriction, and because methyltransferases indiscriminately methylate DNA, viral progeny will carry the host methylation patterns[29,30]. The passive methylation of lucky virions that avoid restriction leads to enhanced infectivity of viral progeny- they will carry the methylation patterns of the host and have increased infectivity to hosts that share that same RM system. The methylation pattern of the last host, however, is reset after infecting a new host- without methyltransferases to methylate newly synthesized daughter strands, the original methyl group is "diluted" as replication ensues. Viruses have also been found to actively methylate their own genome. For example, viruses have been found to carry methyltransferases, possibly as a bet hedging strategy in anticipation of future host RMs[31]. Another example of active methylation, phage have been found to express proteins that bias type I RMs to promote methylation over restriction, ensuring that viral progeny are fully methylated for the next unfortunate host[32].

Methylation is not the only base modification viruses use to evade host endonucleases. Phage have been found to modify their genomes by replacing thymine with 5-hydroxymethyluracil, or modify cytosine to 6-Hydroxymethylcystosine which can resist type IV endonucleases due to the glycosyl group[33]. Interestingly, it was discovered that a cryptic prophage in *E. coli* CT596 transcribed the genes *gmrS* and *gmrD*, a novel type IV endonuclease that targets glycosylated nucleic acids[34]. As a way to protect nucleic acids susceptible to different proteins, *Myoviridae* phage have been found to co-inject IP* protein, an inhibitor of GmrSD[35]. Other co-injected proteins, have also been

observed such as *darA* and *darB* from bacteriophage P1 that also inhibit type I RMs, although the mechanism of inhibition is unknown[36].

Another example of protein endonuclease protein inhibitors is Ocr, which is immediately transcribed by viral DNA after entering they cytosol[37,38]. Ocr is a DNA mimicry protein that inhibits type I RMs. The mimicry is accomplished by having similar charge distribution and bend to that of DNA, allowing the protein to block restriction and increase infection rates by nearly three orders of magnintute[37]. Because of the anti-restriction efficacy, researchers have utilized this protein in the lab as a way to increase transformation efficiency during electroporation protocols[39].

Phosphorthioation. DNA phosphonothioate defense systems, commonly referred to as DND systems, are another innate immune system. In this system, a non-bridging oxygen atom in the phosphodiester bond in the DNA backbone is replaced with a sulfur atom by the products of *dptBCDE*[40]. The remaining portion of the *dpt* gene cluster, *dptFGH*, is responsible for endonuclease activity and reduces the infection rate of foreign DNA by two orders of magnitude[41].

Indeed, the DND system may be widespread throughout the oceans. A DND gene cluster was identified in *Pelagibacter ubique* strain HTCC1002 of the SAR11 clade[42]. With a global estimated population of over 10^{28} cells, SAR11 bacteria account for almost 1 in 4 plankton[43]. While prevalence of the DND defense system in SAR11 is not known, metagenomic contigs from the Sargasso Sea showed evidence of DND clusters, despite this group largely lacking CRISPR or RMs[42,43].

DND systems are one of the few examples where we have not found an anti-restriction system employed by phage. While possible that one may not exist, it is more

likely that we have not discovered such a mechanism. Indeed, if these systems are largely distributed in SAR11 organisms, the phage that infect SAR11 may have the highest likelihood of carrying such a mechanism. Unfortunately, SAR11 are difficult to culture, thus extensive physiological studies of these organisms (and their viruses) are currently limited[44]. We can confirm, however, that current SAR11 viruses in culture are largely represented in sequence data from environmental samples, suggesting that viral predation is a common ecological reality for this numerically dominate group of heterotrophs[45].

Clustered Regularly Interspaced Short Palindromic Repeats and Associated Genes. In addition to innate immunity, Prokaryotes also have adaptive immune systems. Coined CRISPR-Cas in 2002, these adaptive immune systems are found throughout the prokaryotic world[46]. The CRIPSR component is composed of direct repeats, separated by stretches of variable sequences, called spacers, that contain captured viral and/or plasmid DNA. These spacers act as the memory bank for the adaptive immunity and are collectively referred to as a CRISPR array, storing sequences to be used in degradation invading DNAs/RNAs for cleavage. Spacers work in tandem with CRISPR-associate-genes (Cas) and can be highly varied, ranging from 4 to more than 20 Cas genes in different organisms[47]. Generally speaking, CRISPR RNAs complex together with Cas proteins to provide homology-based nuclease activity. Invading DNAs/RNAs are cleaved if they are similar enough to the original spacer sequence and contain a protospacer-adjacent (PAM) motif. Cas proteins require a PAM sites in the foreign DNA as a safety mechanism to ensure that the CRISPR-Cas defense system does not recognize and digest its own CRISPR array in the genome when active, thus, Cas proteins incorporate spacers lacking the PAM motif to differentiate self from non-self[48,49]. Indeed, incorporation of

new spacers can decrease infection rates by five orders of magnitude[50]. Phage, however, can escape digestion if mutations are introduced into the portion of the viral genome that is recognized by the corresponding CRISPR spacer, or by removing the PAM motif.

Abortive Infection. Ubiquitous in the bacterial world, abortive infection (Abi) can be key in controlling infections at a population level[51]. During infection from viruses, Abi genes are expressed and cause premature cell death to either stop or limit production of virions[52]. An example is AbiK, which caused a 14 to 11-fold reduction in viral burst sizes by interrupting the packaging process of viral genomes, greatly limiting the spread of the viruses to the rest of the population. However, phage mutants can emerge that avoid this reduction, as is true of all Abi systems[53,54]. Toxin-antitoxin systems have also been considered to cause abortive infection, such as *mazEF* and *hok-sok*[55,56].

Bacteriophage Exclusion. A more recent mechanism in the microbial arms race, the novel Bacteriophage Exclusion (BREX) defense system was discovered in *Bacillus subtilis*[57]. Goldfarb et al. show that the BREX system in *B. subtilis* contains six genes. Among these genes, authors find evidence of proteins that interact with other proteins, including a protease, a putative alkaline phosphatase, and serine/threonine kinase and DNA methyltransferase targeting TAGGAG. Although the mechanism is still unclear, this phage defense system does not protect from phage attachment or DNA entering the cytosol, nor is it an abortive infection system, but prevents phage DNA replication. This inhibition does not seem to be like that of RMs as the authors failed to detect degradation of phage DNA in infected BREX-containing cells. Currently, the authors hypothesize that BREX is a protein-protein interaction system that, either upon co-injection of viral DNA/proteins or synthesis of viral proteins, somehow deactivates proteins necessary for

viral replication. Currently, it is unclear how the DNA methyltransferase plays a role in this system; however, without it, cells lose their ability to defend against phage. Phylogenetic analysis and operon organizations suggest there are 6 different types of BREX systems found in ~10% of genomes and show sporadic distributions across many phyla.

IV. Intracellular Defense Systems and Horizontal Gene Transfer– A Paradox

Horizontal gene transfer is a well known phenomenon in the microbial world in which genomic material is shared between two cells without reproduction and takes place between closely related species and even domains of life[58]. There are several microbial mechanisms that mediate HGT; however, the three most recognized mechanism are transformation, conjugation and transduction.

Natural transformation is the process by which a recipient cell meditates exogenous (naked) DNA uptake. In this mechanism, a transformation pilus guides exogenous double-stranded DNA from outside the cell into the cytoplasm, during which one strand is degraded and the remaining strand becomes bound to the mediator protein DprA[59]. Here, DprA loads recombinase protein RecA and scans chromosomal DNA to initiate homologous recombination, allowing this mutant to propagate the newly acquired gene through replication. Conjugation is the transfer of genetic material from one cell to another through formation of a pilus. Conjugative plasmids are primarily composed of four different gene modules: replication, which aligns plasmid replication with a host cell's growth cycle and maintains a stable copy number to ensure that the cell is not overburdened; propagation, which houses the genes for creating the conjugation machinery; stability, which ensures proper dissemination of plasmids within the cell and

protective mechanism to avoid degradation or plasmid breaking homologous recombination; and adaption, which provides the host extra genes thereby giving it a selective advantage[60]. Lastly, transduction is the movement of genetic material from donor to recipient through phage. This happens when phage accidently incorporates host DNA during lysis, in generalized transduction, or specialized transduction when prophage incorrectly packages surrounding host DNA when excising itself out of the genome[58].

All intracellular defense systems described above are effective in defending against foreign DNA. Due to their potent defensive properties, one would hypothesize that organism that carry more defense systems are less likely to participate in HGT because they would degrade incoming DNAs once they entered the cytosol. What we observe, however, is the exact opposite trend in RMs[61], CRIPR-Cas[62–64], DND systems[42,65,66], abortive infection[67,68], and BREX[57]. Moreover, with the exception of CRISPR-Cas, defense systems are colocalized to defense islands[69–71]. Defense islands, analogous to pathogenicity islands, show signs of high genomic plasticity relative to the rest of the genome and are considered hotspots for HGT. We are able to associate HGT with these areas by observing deviation from genome wide GC content, abrupt changes in oligonucleotide frequencies, structural features (e.g. repeat regions), and mobile genetic elements located in these defense islands[72]. To gain further insight into possible selective pressures that may be driving the relationship between HGT and defense systems, we will focus on RMs as they are the best studied.

The linkage between HGT and RMs is well documented. RMs have been found on/with plasmids[73,74], prophages[75], transposons[76], integrative conjugative elements[77] and

integrons[78,79], suggesting they have been transferred via all major mechanisms of HGT. One hypothesis that may explain these phenomena is that RMs are selfish genomic elements[80,81]. Much of this hypothesis rests on observation of post-segregationally killing of the host after RM loss[82]. In post-segregationally killing, cells are killed by remnant protein products of genes they have lost, such as endonucleases as in the case of RMs. As cells divide, the remaining methyltransferases are unable to keep the modifications required to prevent endonucleases from degrading the host genome, killing the cell. Indeed, this 'addiction' to the RM system can encourage stabilization of other mobile elements, such as a plasmids, explains why they are repeatedly found with each other[83]. We note that the phenomena of post-segregation killing is also true for toxin-antitoxin systems as they can also stabilize plasmids[84]. The selfish behavior of RMs may describe their propensity to be coupled with mobile elements; however, it does not adequately explain why some organisms have more or less RMs than others[17].

V. Conclusion: A Search for Selective Forces without Context

Above, we described several well known and recently-discovered viral defense systems, with emphasis on intracellular defense systems. Intracellular defense systems have been foundational in driving the development of the molecular biology tools for working with nucleic acids, such as RMs and CRISPRs. What we lack, however, are mechanistic understandings of the selective forces that govern gene gain and loss of these defense systems. Moreover, it is unclear how these defense systems interact with one another; however, we know they are compatible and found together in genomes[70,85].

Despite being the best studied defense system, we still do not understand why there is such an immense diversity of RMs, nor a rationale for the presence of multiple

RMs in a single organism[17]. The close association of HGT and large strain-to-strain variance in RMs, as reasoned by others, suggest that ecological pressures may be responsible for driving the gain and loss of RMs in microbial genomes[70,86]. Overall, my strategy will aim to integrate the well characterized molecular mechanisms of RMs, their distribution across the microbial genomes, and the ecological systems in which the organisms inhabit.

In this dissertation, we aim to understand the selective forces that govern RMs in Prokaryotes. As a matter of practical necessity, we develop a codebase to increase the tractability of working with large amounts of publicly available sequencing data and is outlined in Chapter 2. In Chapter 3, We find the distribution of RMs across microbial genomes and incorporate observations from literature in trait-based mathematical models to explain the bioinformatic observation that extracellular nutrients covary with RMs per genome. Importantly, we find that the trait-based modeling approach was the only way to integrate the informatic observation of high RMs per genome with the reported defensive efficiencies of RMs. In Chapter 4, we explore phylogenetic signal of different methyltransferases showing that, surprisingly, methyltransferases appear to be conserved in some high RM carrying organisms, emulating that of methyltransferases with other physiological roles. In our conclusion chapter, we explore how trait-based modeling of RMs could be expanded and possibly apply to other defense systems. We also discuss how our models from chapter 3 may explain how some methyltransferases are being conserved in high RM organisms, as seen in chapter 4.

CHAPTER 2

Finditfasta: A Python Module for Managing and Accessing Sequencing Data at NCBI

ABSTRACT

Since next-generation sequencing hit the marketplace in 2007, we have seen an exponential growth in the amount of sequencing data available. The limitation with such rapid growth, however, is that our current toolsets are limited in not only accessing this abundance of data but managing it as well. Moreover, there remains a fair amount of confusion in the literature about what is the proper and most efficient way to reference these publicly available resources, sometimes making it difficult or impossible to know which sequence data was used in their analyses. To address these problems, we have developed Finditfasta, a python module aimed at working with the large quantity of sequencing data available at the National Center for Biotechnology Information (NCBI). As a use case, we show that Finditfasta, in tandem with alignment software, can rapidly integrate 3rd party databases back into NCBI resources without the need to search all microbial genomes, allowing researchers to associate metadata easily. Our goal in development is not only decrease the barrier to resources at NCBI by automating many of the mundane tasks required for working with this data, but foster clearer documentation in computational projects by directly guiding users to which sequence identifiers are best to reference.

Publication Note

This Chapter and chapter 4 are being combined into a manuscript for submission and subsequent publication.

Contributions

Spiridon Papoulis wrote the code base. Spiridon Papoulis and Katherine Moccia wrote the documentation and tested the module. Spiridon Papoulis perform the analysis.

Acknowledgements

We would like to thank Erik Gann and Joseph Jackson for coming up with the clever and catchy name Finditfasta.

Code Availability Statement

All source code is freely available at github.com/SEpapoulis/Finditfasta for python ≥3.7.

I. Introduction: The Rapid Growth of Computation in Biology

1977 marked a historic year in biological sciences with the publication of the first sequenced genome, the single-stranded DNA coliphage PhiX174 [87]. DNA sequencing, especially when coupled with the polymerase chain reaction[88] later developed in 1986, provided scientists with everything they needed to move biology into the information age. The potential to gain insights from sequencing nucleic acids was immediately recognized by the international community, and after further improvements in shotgun sequencing[89,90], the international collaboration of the human genome project in 1990 had begun[91].

Sequencing efforts were by no means exclusive to the human genome. Since the advent of sequencing and PCR, researchers had been generating sequence data from all domains of life, necessitating the creation of a new resource of data sharing and leading to the creation of the National Center for Biotechnology Information (NCBI) in 1988[92]. The following years marked the creation of familiar databases and resources, such as the introduction of the Basic Local Alignment Search Tool (BLAST) in 1990, and GenBank in 1992, although these resource were not available through the internet until 1994[92]. By 1998, the National Institutes of Health had 781 entries in their genome division, 17 of which were closed microbial genomes from model organisms, such as *Escherichia coli*, *Synechocystis PCC6803*, and *Saccharomyces cerevisiae*[93].

With the initial sequencing of the human genome project complete in 1999 and the growing number of highly researched microbial genomes, NCBI created RefSeq, a well-annotated set of reference genomic, transcript, and protein sequences[94]. Importantly, RefSeq introduced the non-redundant database, a collection of well-annotated sequences

where no two proteins are identical, and spans all organisms in Refseq. By mid-2004, Refseq had grown to include 2,467 species and over a million non-redundant proteins [95].

2008 marked a pivotal year in our ability to sequence nucleic acids with the development of the second generation of sequencing technology. This technology allowed for parallel sequencing of multiple nucleic acids in one reaction at a fraction of the cost[96,97]. Thus, a positive feedback loop emerged: as sequencing improved, costs decreased and enabled a wider range of scientists to pursue sequencing based projects, which incentivized innovation in sequencing technology. As the technology improved, sequencing costs continued to plummet orders of magnitude from a single human genome costing 100 million USD and 10 million USD in 2001 and 2007, respectively, to just 1000 USD in 2019[98]. Today, we are slowly seeing the emergence of a third-generation sequencing technologies that generates long reads along with other features, such as the possibility of generating methylome data with singe molecule real time sequencing from Pacific Biosciences or the ability to sequence on flash drive sized machines with MinION from Oxford nanopore Technologies[97].

II. The Prokaryotic Data 'problem'

The extreme reduction of sequencing costs over the last 20 years has provided a wealth of new microbial genomes available at NCBI. However, with this explosion of genomic data comes a new set of problems that create bottlenecks for analysis and interpretation. For example, due to the volume of data, bioinformatic programs must be both scalable and validated to ensure proper operation[99]. Scalability is important because without it, our tools will deteriorate in performance the larger the datasets become. Much like running controls during experiments, validation of a program can be accomplished

by checking the output of data that has already been analyzed, ensuring all algorithms are operating as intended. These issues notwithstanding, a more elusive problem remains unaddressed: how to efficiently find, access, document, and manage data of interest located within massive publicly available datasets. Web interfaces, such as those at NCBI[100], provide access to publicly available genomes for researchers but quickly become unwieldy when locating and accessing larger datasets of tens of thousands of genomes. A more convenient option for accessing large volumes of data are file transfer protocols (FTP) that streamline data access, however, these methods have a steep learning curve as users would need to know 1) the internal file structure of the FTP servers, 2) intermediate scripting skills for file retrieval, and 3) the identifier logic used to catalog data. Specific to NCBI, identifiers called accessions are required to find all nucleic acid, protein, or assembly data within NCBI's FTP servers.

Because large volumes of data are being generated from every scientific research field and multiple sectors of our economy[101], many open source projects have been under development to ease and streamline analysis workflows, such as Project Juypter[102]. In biological sciences, analysis solutions such as Qiime[103] have emerged for processing large amounts of sequencing data; however, a lack of software to streamline data access and organization in computational workflows makes it difficult to build custom reference datasets from high quality assemblies, such as those found in Refseq. To address this issue, I have developed finditfasta, a python module aimed at increasing accessibility of publicly available data at NCBI.

III. Finditfasta: A solution to access and management of publicly available sequence data

FinditFasta (FIF) is a lightweight python module designed to improve the access and integration of sequence data located at different places within NCBI, such as Genbank, Refseq, and Taxonomy databases. The resources at NCBI can facilitate computational research by giving access to high quality sequencing data, making it easier for creating positive controls and building custom databases. For example, integration of these resources can allow researchers to find the genomic locations and taxonomic distributions of proteins without needing to search genomes with alignment software, allowing researchers to easily document all DNAs or proteins used in each database to maximize clarity in computational projects. This can be especially useful in gathering references for transcriptomic or metagenomic studies as it builds off our current framework of cataloging sequence data. Moreover, finding the genomic source of a protein can be invaluable to phylogenomic investigations for understanding genomic context and regulatory mechanisms, or can provide clues for proper annotation[104].

At the core of the FIF module is the Catalog, a data structure that manages biological data with three primary features. Foremost, the FIF Catalog creates a low memory way of iterating through large amounts of data, thus maintaining scalability through database growth. Secondly, the FIF Catalog automatically integrates data that are stored at different locations within NCBI. Lastly, the FIF Catalog is designed to work with past, current, and future database releases by allowing some data structures to be dynamic.

To handle the volume of data, the FIF Catalog is compiled using sqlite3, a database application programming interface (API) module for SQLite databases[105].

SQLite is a relational database management system that efficiently retrieves and stores tabulated data from computer storage (i.e. hard drive or solid-state drive) using primary keys, also known as unique identifiers. Using storage instead of random-accesses memory (RAM) allows for working with large datasets- if datasets get to big, they can quickly overwhelm conventional personal computers with commonly only 4 gigabytes of unused RAM as compared to storage drives that commonly have between 1000-500 gigabytes of available space. While there is a speed trade-off between using storage over RAM to retrieving data from the database, primary keys can greatly speed up data calls because primary keys are unique identifiers organized to a binary-tree that decreases search times of a database of n size from searching all elements, to searching $\log n$ elements. This increase in speed is due the traversal of the binary-tree from root to leaf to find an entry instead of searching every entry and is foundational for efficient calls from any database. The primary keys of the FIF Catalog are protein accession numbers, assembly accession numbers, and taxonomic identification numbers (Figure 1). Relational databases like SQLite greatly increase tractability of large datasets because they “relate” pieces of information internally and are ideal when information such as organisms and their associated taxonomies is highly interconnected.

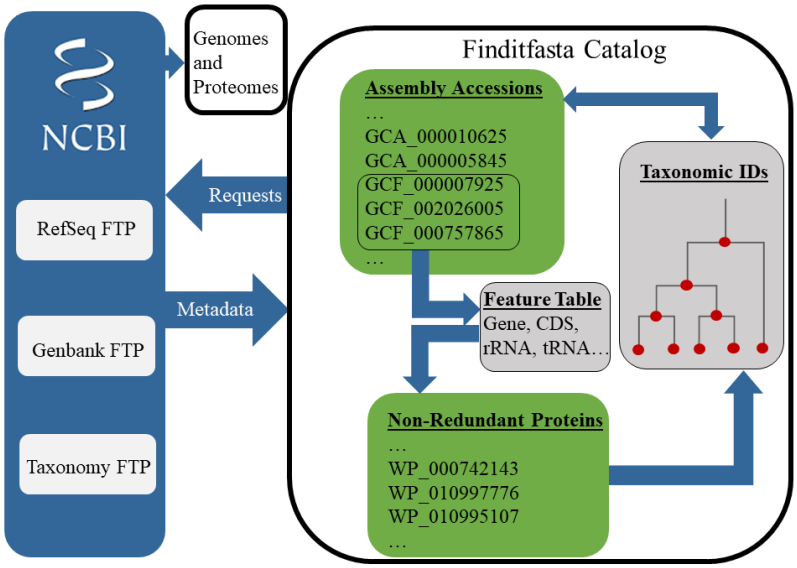

Figure 1. Finditfasta Database Structure. Finditfasta sources data from three FTP locations at NCBI: RefSeq, Genbank, and Taxonomy. Refseq and Genbank are used to link assembly accession numbers with the taxonomic identifiers that are organized in a tree structure and allows for taxonomic group calling of assemblies. Assemblies can be used to request genomes and proteomes associated with each assembly. Non-redundant proteins are mapped to their conserved taxonomic nodes, which can be subsequently used to call assemblies and feature tables to find their genomic locations. Assembly accessions and non-redundant proteins, in green, represent static primary keys, while taxonomic nodes and feature tables, in gray, represent dynamic data that may or may not change over time, pending review.

FIF improves data access by integrating several commonly used resources at NCBI. Upon initialization, NCBI FTP servers are automatically queued for download from three databases: Genbank, Refseq, and Taxonomy (Figure 1). Genbank and Refseq both contain organismal sequence metadata, where the primary keys are assembly accessions from Genbank (GCA) or Refseq (GCF). Because Refseq is a high quality, reannotated subset of Genbank assemblies, it is important to avoid using Genbank assemblies if their Refseq equivalents are already being used. FIF automatically associates Refseq assemblies with their sourced Genbank assemblies to allow users to intelligently reduce redundancies when building their own custom reference dataset.

Maintaining database integrity is a challenge, especially with dynamic data structures. This is especially true for data structures emulating taxonomy, as taxonomic identification of an organism is frequently redefined[106]. Accessions for both Refseq (GCA) and Genbank (GCF) assemblies are organized under different unique taxonomic identifications (taxids) according to NCBI taxonomy (Figure 1). In our implementation, taxids are pulled into RAM and built into a tree-like structure, emulating the tree of life, for easy iteration up and down the tree, allowing for batch calls of accession assemblies or non-redundant proteins based on any taxonomic criteria available (Figure 1). Because taxids often change, we use taxonomy merely as a hierarchical structure of convenience rather than documentation. This rationale follows because assembly accessions will always find the same sequence data, whereas taxids are fluid and may not always retrieve

the same information if assemblies become reclassified under a different taxid, or the association between taxids becomes altered.

FIF begins by building a local database. This initial step acts as a safety mechanism to protect against database changes that may alter assembly calls when using taxonomic criteria. While this does require a modest diskspace of ~500MB we find this to be more convenient than contacting NCBI server for data retrieval every time that catalog initialized. Additionally, we find compiling a local database the best way to maintain clarity between computational projects, so taxonomy does not change during the time from project conception to completion. Locally compiled databases also maximize reproducibility because database files can be conveniently shared.

The non-redundant protein dataset provided by Refseq is an exceptional resource that is also utilized by FIF. In the last 20 years, the number of coding sequences in Refseq has increased by four orders of magnitude. As of February 6th 2020, there are 181,972 prokaryotic genome assemblies that constitute 695,280,251 coding sequences that reduce to 136,129,750 when applying a non-redundant criterion (Figure 2). At the current growth rates of assemblies, it is likely we could have ~10^6 assemblies by 2025 and may push the number of coding sequences over 10^9. Because the growth rate of the non-redundant protein dataset is slower, the divide between the non-redundant dataset and coding sequences will continue to grow. We note that Figure 2 was generated only with metadata available to FIF and demonstrates how linking assembly metadata, in this case release date, with the non-redundant proteins found per assembly, allows rebuild database growth since conception.

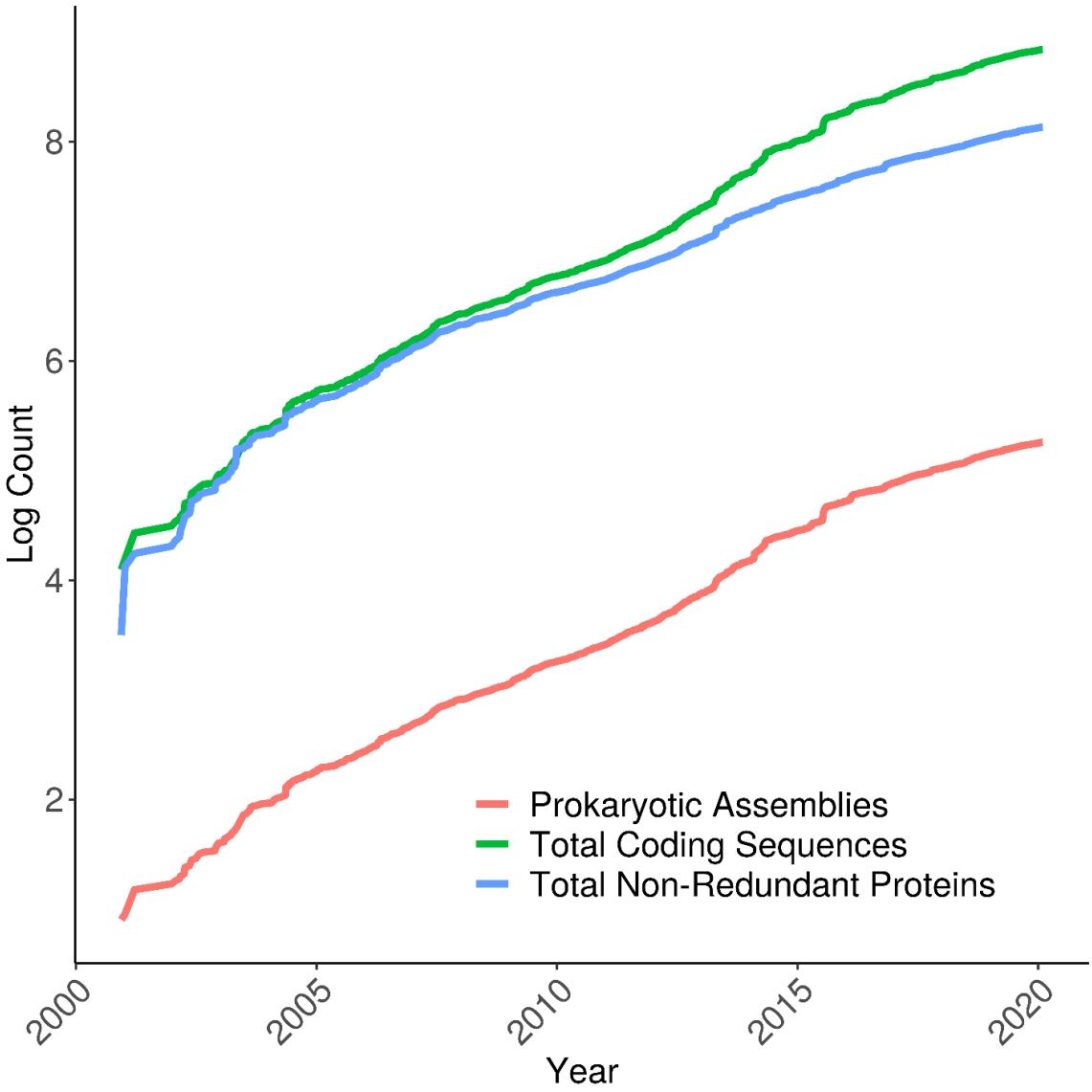

Figure 2. Growth of Prokaryotic Data in NCBI's RefSeq Database. Total prokaryotic assemblies, protein coding sequences, and number of non-redundant proteins over the years, as of January 2020. Each assembly corresponds to the genome of a single organism, while multiple coding sequences come from each organism. Non-redundant proteins, however, are only added if they are not observed in any other organisms when documented.

Importantly, we find applying the non-redundant criterion to all prokaryotic proteins reduces the protein search space for any alignment software by 5x. The trade-off, however is that extra steps are required to find the source organisms for each non-redundant protein. Refseq does provide mappings of non-redundant proteins to their genomic locations though a web interface for identical protein groups[107], however, this again becomes difficult to use with larger protein queries. With FIF, we take an alternate approach that makes calling the genomic location data for non-redundant proteins easier for users (Figure 1). This is accomplished using feature tables, condensed tab-delimited files reporting genomic location data for all annotated features per assembly, making integration of this information into the FIF's SQLite database simple. Each Refseq release provides non-redundant protein mappings to their conserved taxonomic species, therefore, FIF compiles feature tables by species taxid and requires just one database transaction per protein lookup. In contrast, if feature tables were organized by assemblies, looking up genomic location of a non-redundant protein would be comparatively slower because it would require an SQLite database transaction for each assembly documented under a species taxid. To maintain database performance, non-redundant proteins only found in one species are associated to species via primary keys, thus search times are log *n*. In contrast, querying multispecies proteins requires searching the entire multispecies dataset to find all species associated for a given multispecies protein. Therefore, the separation of multispecies and non-multispecies proteins maximizes search speeds.

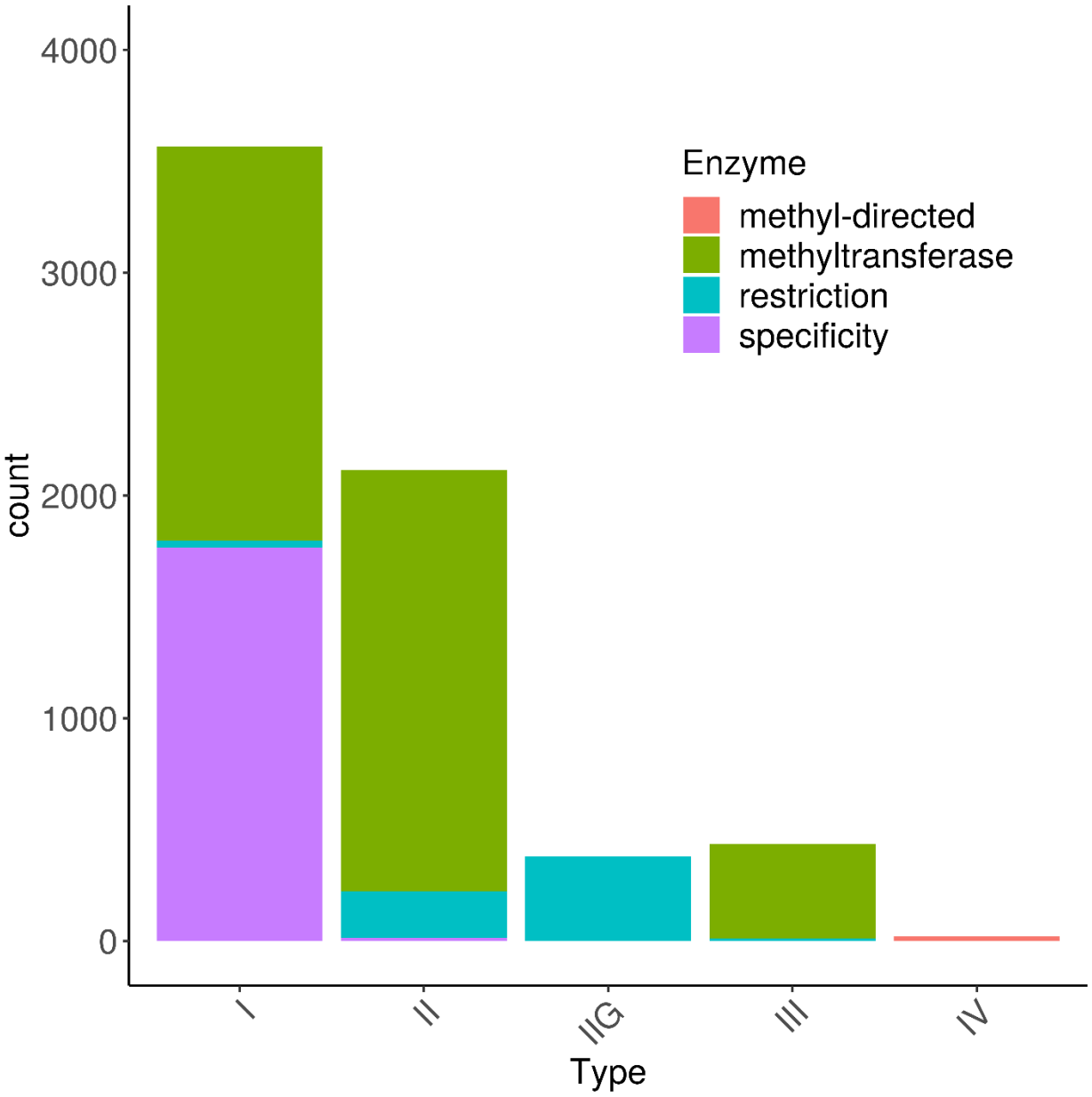

Figure 3. Distribution of Proteins belonging to the REBASE 'Gold-Standard' Protein Dataset. Using the avalible metadata at REBASE, we are only able to describe the proteins as part of different types, and what kind of enzyme they are.

By default, the FIF Catalog does not download non-redundant protein accessions, nor their taxonomic mappings. The justification for this design was to decrease the compile time when first initializing the database. When initializing the FIF Catalog for the first time, less than 200MB of compressed data are downloaded and compiled into a 500MB database for assembly operations. Subsequent to this initial download, FIF requires an additional download of 1.4GB of compressed data that compiles into a 31GB database file to perform index-based searches of 223,560,051 non-redundant proteins.

IV. Database Integration Case Study: New England Biolab's REBASE

Microbial databases have been extremely prolific as researchers attempt to organize and analyze the abundance of sequence data available[108]. However, it has become difficult to integrate curated information from several sources as database design and documentation varies greatly from author to author. Pertinent to this dissertation, New England Biolabs REBASE is a comprehensive database of endonucleases and methyltransferases that were originally sourced from sequenced genomes in Genbank. REBASE documents their "Gold Standard" proteins which have been experimentally validated for activity[109]. Moreover, REBASE delineates between the different types of endonucleases and methyltransferases by breaking them down into specific types with their own unique properties (see chapter 1), therefore, maintaining their associated metadata will be critical in assessing the precision of this database when being used as a reference dataset. Using annotations in REBASE, we find that this dataset is largely biased towards Type I and Type II methyltransferases, with majority of the

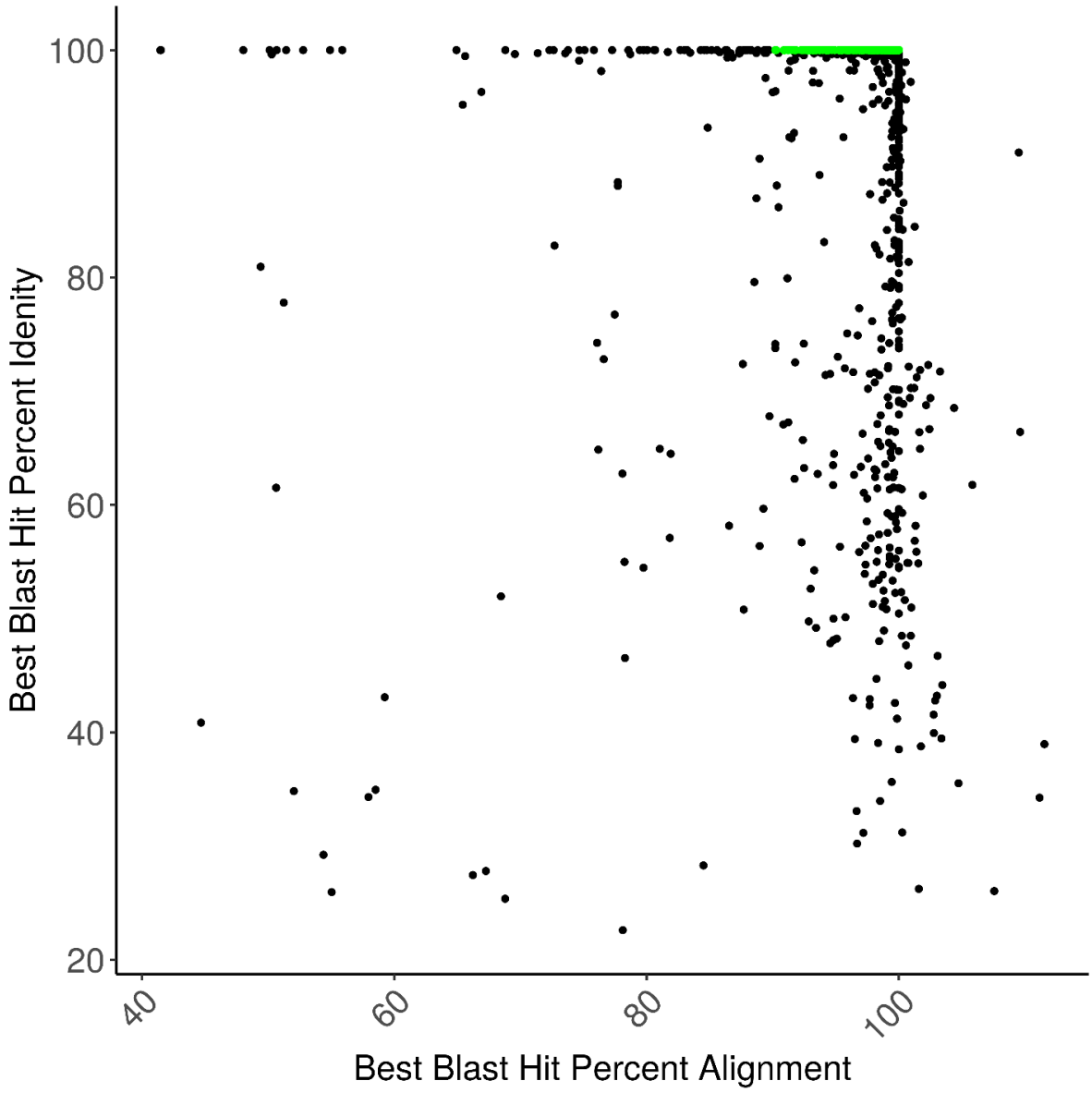

Figure 4.Best Blast Hits Mapping "Gold-Standard" REBASE Restriction Modification Proteins to Non-Redundant Proteins. REBASE queried proteins were considered a match (green data points) if there was 100% identity to a non-redundant protein and ≥90% query alignment length or were considers homologs (black data points) if the alignment failed to meet the criteria.

endonucleases belonging to Type II and Type IIG (Figure 3). Unfortunately, these characterized proteins are not associated with any assembly information and instead users would need rely on the organism name, which is ambiguous when considering strain levels are not documented for each entry in REBASE or ambiguities due to taxonomic reclassification. Thus, with the combination of BLAST to align to non-redundant proteins and FIF to use index-based searches to find their genomic locations, we will be able to locate the origin of these proteins and check their taxonomic distribution.

To find the non-redundant alias of each biochemically characterized restriction modification (RM) system, we query the "Gold-Standard" dataset against the non-redundant protein database to generate alignments using BLAST. Limiting results to 100 hits per query protein, we further refined our searches to proteins with 100 percent identity and ≥90% alignment length to the query protein. We justify our relaxed alignment length percentage relative to the query to account for any possible differences between annotation of the start site between Genbank assemblies used at REBASE and current Refseq assemblies downloaded from NCBI. In cases where there was 100% identity between two proteins, the longer alignment length was selected. Using these cutoffs, we found non-redundant accessions for 6,520 out of 7960 Gold-Standard proteins (Figure 4). We tracked down the taxonomic origin of the non-redundant proteins using the index-based searches of FIF and found most of these proteins were sourced from the phyla Proteobacteria and Firmicutes (Figure 5). Moreover, we found that nearly 20% of the Gold-Standard proteins are sourced from just 6 microbial species: *Escherichia coli,*

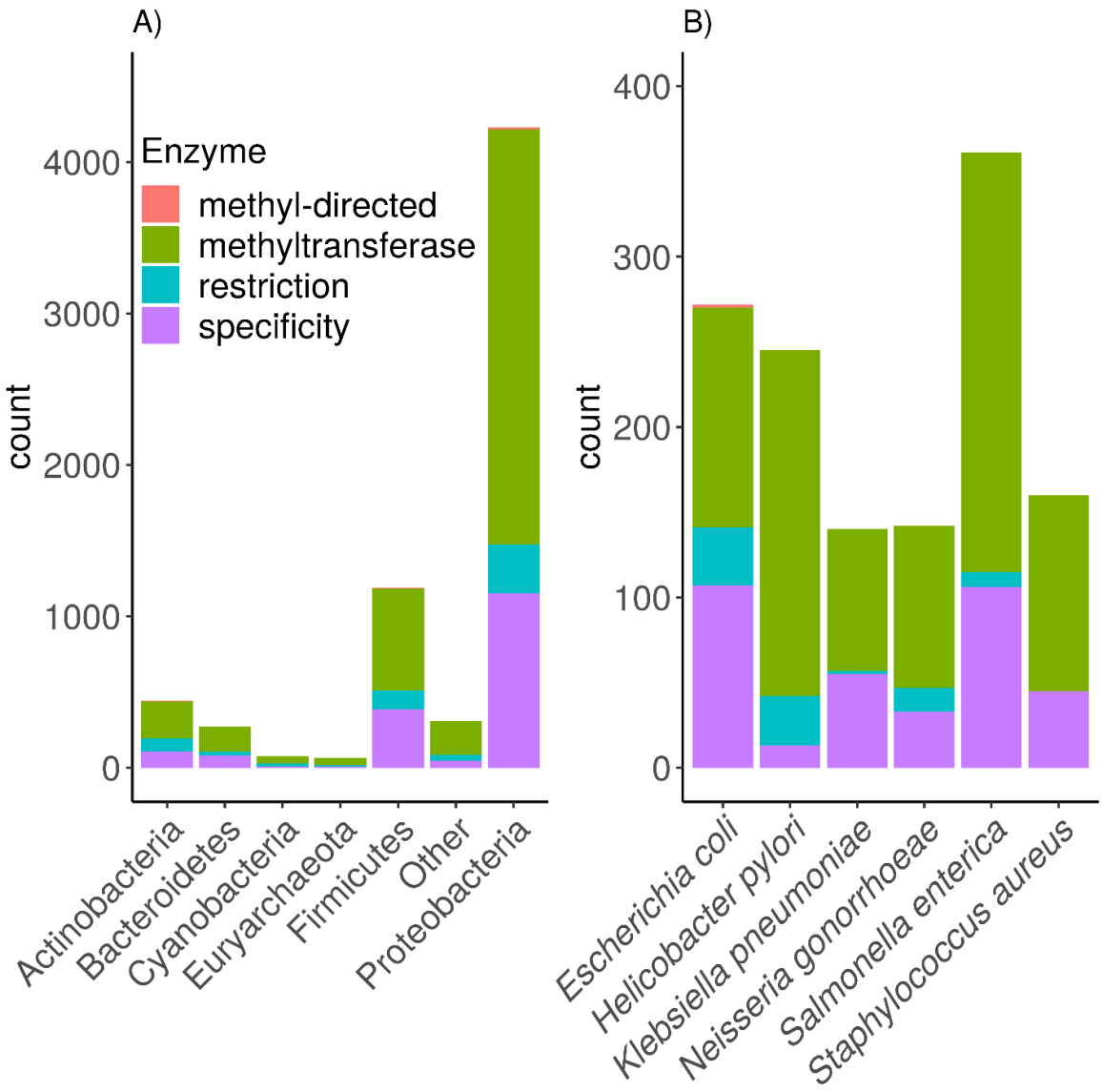

Figure 5. Taxonomic Representation among the "Gold-Standard" REBASE dataset. By mapping REABASE biochemically characterized proteins with non-redundant proteins, we are able to find the taxonomic distribitions. **A)** Distribution of proteins grouped at the taxonomic rank of Phylum. **B)** Top 6 source microbial species for Gold-Standard proteins, which constitute nearly 20% of the dataset.

Helicobacter pylori, Klebsiella pneumoniae, Neisseria gonorrhoeae, Salmonella enterica, and *Staphylococcus aureus*.

V. Discussion and Conclusions

The multidisciplinary nature of computational biology makes it difficult to generate any one program to address research needs. Moreover, developers must try and anticipate the knowledge of their user bases. This can be difficult as Computational Biology/Bioinformatics is a relatively young scientific field with institutions playing "catch-up" to develop robust curriculum to teach the necessary foundational knowledge[110]. The goal in developing FIF was not to create a monolithic program for all computational biology needs, but rather a lightweight tool that excels in a few operations. Hopefully, this can encourage a workflow mentality to computation, allowing for exploration of data at each processing step, and avoid data "pipelines" that have linear workflows.

As an example of a use case, we were able to integrate a biochemically characterized set of methyltransferases and endonucleases into the non-redundant protein dataset and leverage that dataset to find the taxonomic distribution using FIF in tandem with alignment software. We find that this strategy is ideal in a wide array of bioinformatic projects, especially those investigating the origin of unannotated sequences from environmental samples. For example, researchers could take a collection of unidentified proteins, find the non-redundant homologs with alignment software, and assess their taxonomic distribution without needing to independently search every single sequenced organism. This may be extremely useful in metagenomic/metatranscriptomic

studies that find genes/proteins of interest, where researchers could quickly assess the genomic context of several genomic elements at once, possibly elucidating their role the environment. Indeed, increasing database sizes will cause computational times to grow (Figure 2), therefore, maximizing search strategies will be critical in the coming years.

FIF streamlines the access to assembly data both at Genbank and Refseq though compilation of a local database. Compiling a local database provides two key benefits: reproducibility and workflow management. These benefits arise because there is no ambiguity in several important pieces of information, including FTP locations of sequence downloaded, mappings of non-redundant proteins to taxonomic nodes, nor assembly mappings to taxonomic nodes for any project. One potential weakness in the module, however, is the abandonment of assembly accessions and non-redundant protein accessions (the primary keys of database) at NCBI. While we do not expect NCBI to phase out assembly accessions nor non-redundant proteins, it would not be the first time since the discontinuation of assigning GI numbers as sequencing identifiers in Genbank[111]. Barring the abandonment of assembly and non-redundant accessions, we expect that the database design of FIF to avoid deprecation for the foreseeable future.

In our example application of FIF, we were able to successfully map ~82% of the biochemically characterized RM proteins at REBASE to non-redundant protein accessions and show the distribution of these proteins within the tree of life. More importantly, we are able to access the exact genomic locations of each mapped REBASE entry among Refseq genomes without needing to search proteomes individually, allowing for a massive reduction in computational time. The index-based searches of non-redundant proteins through FIF complements this strategy by compensating for a caveat

that we lose genomic location of proteins when forcing non-redundant criteria. For RM systems in particular, contextual analysis becomes critical in assessing if a Type II methyltransferase has a cognate Type II endonuclease form a full RM system, for example[61].

The emphasis of assembly and non-redundant accessions as primary keys are deliberate, both for practical relational database purposes, but also user documentation. Our hope is that by emphasizing these primary keys, documentation will improve the reproducibility of computational workflows and recapitulating precise analyses. FIF is freely available at github.com/SEpapoulis/Finditfasta for python ≥3.7 and will be available via pip.

CHAPTER 3
Resource availability and viral DNA methylation drive the diversity and abundance of Restriction Modification Systems

ABSTRACT

Restriction Modification systems (RMs) in prokaryotes serve as primitive immune systems that degrade foreign DNA. Here, genus-level analysis of 139,023 genomes revealed a wide variation in RM quantity per genome (0 to >15) across prokaryotic domains. Within the Cyanobacteria, genera that dominate nutrient-rich environments exhibited vastly higher RMs per genome than those adapted to nutrient-poor systems. Using models, we show how resource-driven increases in host and viral abundance select for acquisition of new RMs. Importantly, the methyltransferase activity of RMs that protects the host from DNA cleavage also confers partial protection to viruses, reconciling the apparent overkill of RMs' high efficiency with high per genome abundance in some genera. Furthermore, modeling reveals competing hosts with subsets of the same RMs often compete to exclusion, whereas hosts with unique RMs sets can coexist. Collectively, these models show how the diversity of RM abundances and specificities can be attributed to resource availability.

Publication Note

An alternate version of this chapter is currently under review at mBio.

Contributions: Spiridon Papoulis, David Talmy, Steven Wilhelm and Erik Zinser designed research; Spiridon Papoulis performed research; Spiridon Papoulis, David Talmy, Erik Zinser contributed analytic tools; Spiridon Papoulis, David Talmy, Erik Zinser analyzed data; Spiridon Papoulis, David Talmy, Steven Wilhelm and Erik Zinser wrote the paper

Code Availability

All bioinformatic and mathematical modeling work was conducted in executable Jupyter Notebooks which can be accessed along with all source code and data at https://github.com/SEpapoulis/EscalationAndDe-escalationOfRM

Acknowledgments

We thank Igor Jouline and Jeffery Morris for helpful discussions involving this study and Katherine Moccia for critical review of this manuscript. This work was supported by an NSF grant to ERZ and SWW (NSF: IOS-1451528). This work was also supported by funds from The *Great Lakes Center for Fresh Waters and Human Health* (NIH 1P01ES028939-01; NSF OCE1840715).

Data Availability (https://github.com/SEpapoulis/EscalationAndDe-escalationOfRM)

Dataset_S01. Refseq assemblies used in this study Refseq metadata, including ftp location of source materials, in csv format

Dataset_S02. RM counts aggregated at the genus levelCSV of organism genome and RM count data aggregated at the genus level, where all numeric columns are averages, except for those column names delimited with "_std", which are standard deviations. Fields include taxonomic information (genus, phyla), genome information (num_isolates, bp, NumContigs), and RM counts. RM counts are distinguished by codes, where r = restriction enzyme, m = methyltransferase, and the number indicates the type. Total RM counts are a summation of rmT1, rmT2, rmT3, T4, and T2G_posthoc, where total RM without putative type IIG RM systems (fig S2) are a summation of rmT1, rmT2, rmT3, T4 and T2G.

Dataset_S03. Rebase "Gold Standard" non-putative Methyltransferases and endonucleases proteins used to retrieve HMM profiles, in fasta format.

Dataset_S04. RM pfams found in Rebase "Gold Standard" proteins

A list of HMMs used to identify RM as a text file. This file can be used to retrieve HMMs out of pfam using hmmfetch.

Dataset_S05. Pfams that covary with false positives

A list of HMMs used to identify common false positives as a text file (see SI methods). This file can be used to retrieve HMMs out of pfam using hmmfetch.

Dataset_S06. Blast exception proteinsA subset of Dataset_S03 that do not have pfams detailed in Dataset_S04, in fasta format.

I. Introduction

Viruses (bacteriophages, or phages) are a powerful evolutionary driver and ubiquitous in the prokaryotic world[112–114]. The lysis of microbial cells contributes to biogeochemical recycling *via* a process known as the "viral shunt"[115], and selects for genotypes whose innovations decrease the rate of mortality from viral infection. Antiviral innovations fall into two general classes: those that prevent virus adsorption at the cell envelope, for instance through mutation of the virus receptor[116–118], and those that establish within the cytoplasm the ability to destroy the virus or kill the infected cell. Cytoplasmic defenses are widespread in prokaryotes[5,119], and include CRISPR[120], argonauts[121], toxin-antitoxin systems[122], abortive infection[123] and BREX[124]. While many of these cytoplasmic defenses have only recently been discovered, one has been known since pioneering research in the 1950s: the restriction modification (RM) system[125,126].

Restriction Modification systems (RMs, or RM system for singular) galvanized the molecular biology revolution through their ability to cleave double-stranded DNA (dsDNA) at sequence specific motifs. When expressed *in vivo*, the endonuclease (restriction enzyme) activity of the RM system can protect a cell from dsDNA viruses that contain the specific sequences. Individual RMs can reduce rates of infection by 2 to 6 orders of magnitude[127]. Because of this antiviral effect, RMs can be thought of as primitive innate immune systems whose targets are pre-determined by the specified recognition motifs of the endonucleases. This contrasts with the "adaptive immunity" conferred by CRISPR-Cas systems that use information gathered from prior infections to provide targets for DNA cleavage.

The motifs targeted by RMs usually recognize between 4-14 bases[128], and consequently also present in the host's genome. To protect the cell's genome from cleavage, most RMs provide DNA methyltransferase activity that methylates residues within the same target motif as the endonuclease. For Type I-III RMs (see below), the endonuclease activity is specific for unmethylated DNA. Thus, the role of the methyltransferase is to block the endonuclease from cleaving host DNA, while leaving it free to attack incoming, unmethylated viral DNA. One important drawback to this defense system is that any viral DNA that escapes endonuclease attack long enough will be "immunized" by the methyltransferase[125–127,129]. Consequently, methylated viral progeny released from the cell will be protected from endonuclease activity if infecting a new cell with the same RM defense.

RMs fall into one of several classes based on protein structure and DNA target. In Type II RMs, endonuclease and methyltransferase activities are in separate proteins that recognize DNA independently. Type I and III RMs involve separate proteins that complex together with or without a specificity unit, respectively. Type IIB, IIG, IIH (collectively referred to as Type IIG in this study), involve single polypeptides with both activities covalently linked. Finally, Type IV RMs are single endonuclease proteins that have the unusual property of recognizing and cleaving methylated, rather than unmethylated, DNA[130].

RMs are nearly ubiquitous among prokaryotes[61,131], suggesting that while they do not play essential roles in cellular growth, they do play important roles in prokaryote ecology. Prior studies have suggested the number of RMs increases with genome size[61,132,133]. Despite this apparent relationship with genome size, RM distributions lack a

clear and obvious phylogenetic signal in prokaryotic lineages. For example, inclusion of the genus *Helicobacter* within the Epsilonproteobacteria lineage skewed the RM per megabase from 1.5 to 5, due to the nearly 12 RMs per *Helicobacter* genome[61]. Oliveira et al. also showed in a pan-genomic analysis of 43 bacterial species, isolates share only ~4% of RM genes in the core genome, whereas the rest of the RM genes were in the flexible genome. Moreover, of the RM gene families in the flexible genome, 80% are only found in 1/3 of strains, suggesting they have been recently horizontally transferred. These findings are supported by subsequent studies showing conservation of RMs at very fine taxonomic resolution diminishes when larger taxonomic groups are considered[134,135].

The lack of a strong phylogenetic signal (*i.e.*, vertical transmission) of RMs within the prokaryotes suggests that horizonal transmission has played a significant role in the evolutionary history of RMs. This may seem counterintuitive, as the restriction endonucleases of RMs can limit horizontal gene flow between organisms[133,136]. Nonetheless, RM genes have been found within mobile elements associated with horizontal transfer as outlined by Oliveira et al. (2016): plasmids[73,74], prophages[75], transposons[76], integrative conjugative elements[77] and integrons [78,79]. Additionally, chromosomal positioning of RMs displays a non-random linkage to genome islands associated with horizontal transfer[132].

Collectively, variation generated by horizontal gene transfer and the expansive diversity of domain and genomic architectures in RMs has made it difficult to use evolutionary context to infer the drivers of retention, loss, and innovation of prokaryotic RMs. Because RMs can target different sequences, a newly-introduced RM should theoretically offer an additive (*i.e.*, non-redundant) effect on antiviral defense. Larger

genomes may thus afford extra space to add RMs, but this space argument does not fully account for the gain of RMs during genome expansion, nor the loss of RMs during genome reduction. For instance, the high number of RMs per genome (~12) in *Helicobacter* cannot be explained by a proportionally larger genome size (~2 Mbp)[61,137,138]. It is clear that genomes contain variable amounts of RMs, defense agents that move frequently between species, but the evolutionary rules that govern abundance and targeting breadth are not well understood.

We explored the distribution of known RMs amongst the 139,023 sequenced genomes in a reference sequence database. We observed that at the genus level, RM quantity per genome varied greatly throughout the bacterial and archaeal domains. Our statistical analyses revealed conspicuous patterns in the prevalence of RMs among bacteria adapted to contrasting resource environments. We thus used contrasting models to explore and explain how these patterns arise from shifts in selective pressure along resource supply gradients. We show how high *vs* low resource availability drives successive additions or subtractions of innate molecular defense systems. Critically, our models suggest that the ability of viruses to exploit the unique feature of RM defense - the immunity conferred by methylation - plays an essential role in driving the escalation or de-escalation of RM antiviral defenses along environmental resource supply gradients.

II. Results

Restriction Modification Distributions 139,023 genomes (Dataset_S01) were searched using our RM pipeline, resulting in a mean of 1.93 RMs per genome, and 95% of all genomes have ≤ 5 RMs. These statistics, however, are susceptible to bias due to

uneven sampling between taxonomic groups. To mitigate the disproportionate effect some taxa (*e.g.,* genomes from overrepresented genera such as *Shewanella* and *Escherichia*) we aggregated data at the genus level, providing mean values for 2,522 total genera of bacteria and archaea (Dataset_S02). At this level of resolution, a mean value of 2.387 RMs per genome was observed, with 5th and 95th quantiles at 0 and 6 RMs per genome, respectively. Some genera were represented by very few genomes, and consequently their calculated means may poorly represent their true means. Therefore, we restricted the data further to genera with 5 or more sequenced genomes and found the mean, median, and 5th percent and 95th percent quantiles of RMs to be 2.17, 1.91, 0.427 and 4.40 RMs, respectively (Figure 6A).

Previous studies reported a correlation between the genome size and the number of RMs[61,132,133]. To revisit these analyses with an updated, larger set of completed genomes, we performed both linear and negative binomial regression on the mean RMs of the genera with 5 or more sequenced genomes. Both regressions give the same result: genome size is a poor predictor of the number of RMs in prokaryotes as it can explain no more than ~2% of the variation (linear: Estimate = 0.145, R^2 = 0.0217, p = 4.93e-05, negative binomial: Estimate= 0.06401, McFadden Pseudo R^2=7.02e-3, p = 1.07e-05). Moreover, while these data are statistically significant, the estimates from each regression suggest that there would need to be an extremely large increase in genome size for there

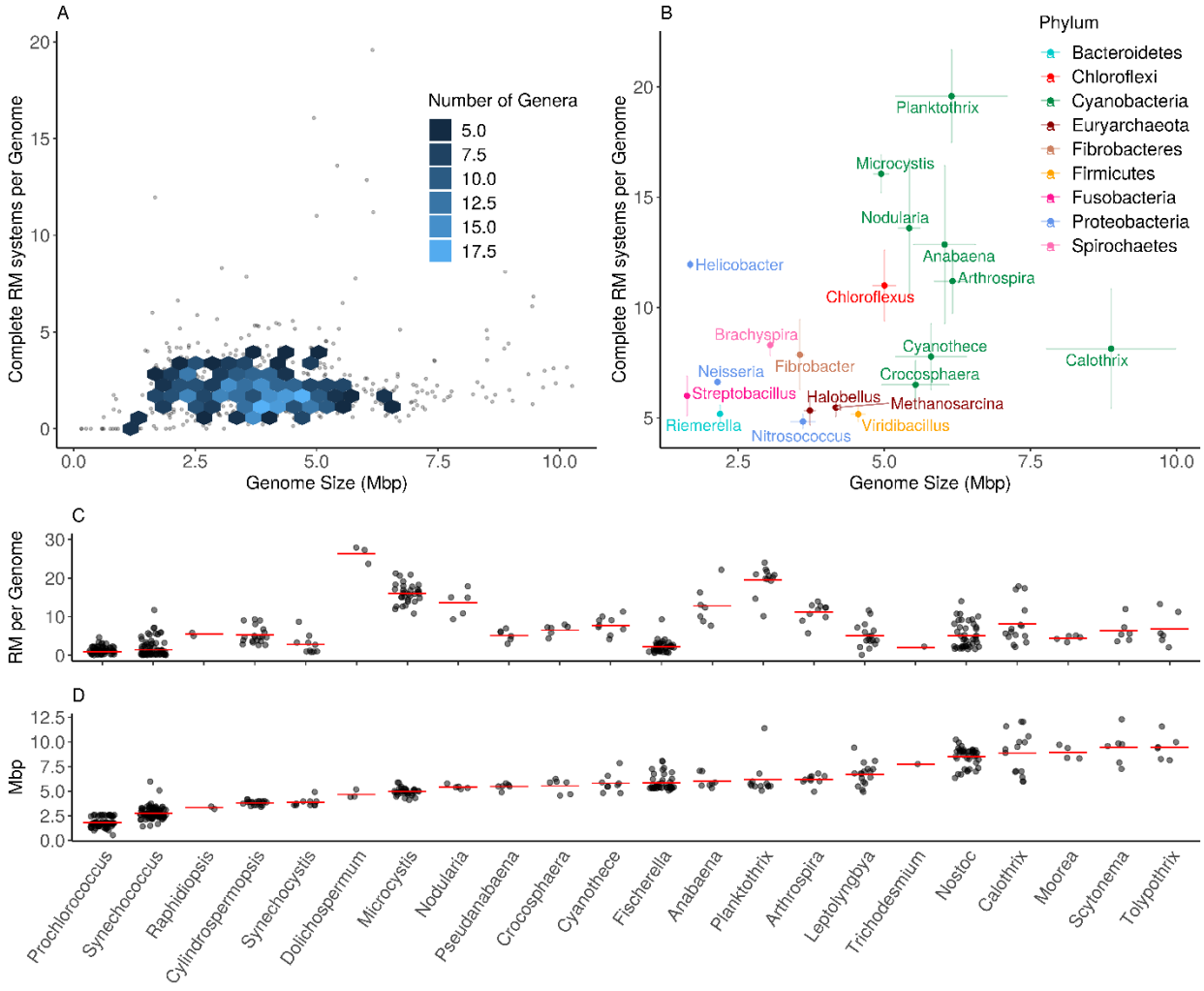

Figure 6. Distribution of Restriction Modification Systems in Prokaryotic Organisms. A) Mean number of complete RM systems per genome plotted against mean genome size in prokaryotic genera. Data points represent the mean of 5 or more isolates, hexagons are rendered when there are 5 or more data points. Mean RM and 95% confidence interval = 2.17+/- 0.119, Median RM = 1.91. **B)** 95% quantile (≥4.4 RM) of the RM distribution. Error bars represent 95% confidence intervals of the genus mean and data points represent 5 or more isolates. Genera were dropped if the 95% confidence interval feel below the 95th quantile. **C&D)** Subset of data from **(A)** which is restricted to only the Cyanobacteria Phyla, showing complete RMs **(C)** and genome size **(D)** per genome. Genera are in ascending rank order by genome size. Red bars indicate genus mean while datapoints are individual isolates. All genera have 5 or more isolates except for *Raphidiopsis* (n=2), *Trichodesmium* (n=1), *Dolichospermum* (n=3).

to be an impact in RM count, if genome size is the sole predictive indicator. For example, an organism with an initial genome size of 2 Mbp would need to expand its genome by an additional 6.90 Mbp or 6.50 Mbp according to the linear or negative binomial regression respectively, to gain one additional RM system. However, for small genomes within the range of 0.5 to 2.5 Mbp, we observed a more pronounced scaling of RM counts per genome as a function of genome size, a trend that is consistent with earlier studies[61,133].

Consistency Among Extremes of the RM Distribution. To investigate factors other than genome size that could drive RM gain or loss, we next examined the extreme cases of very few or very many RMs per genome, for genera with at least five genomes sequenced. The low-RM genera were defined as those with 95% confidence intervals below the 5th quantile line at 0.427 RMs per genome. This category included several genera that are exclusively intracellular or have a large intracellular component to their lifestyle. These include the obligate intracellular parasites *Wolbachia* and *Rickettsia*[139] (Figure 7). Given that a strict intracellular lifestyle should limit the contact rate with infectious virus and thus reduce the pressure to maintain viral defense, it was not surprising to find these genera in the low-RM category.

The high-RM genera were defined as those whose 95% confidence intervals were above the 95th quantile line at 4.40 RMs (Figure 6B). *Helicobacter* and *Neisseria*, noted previously for their high number of RMs[138,140] fell into this category, as expected. The cyanobacterium *Microcystis*, a photosynthetic freshwater microbe associated with harmful algal blooms (HABs), has been previously reported to contain an extensive

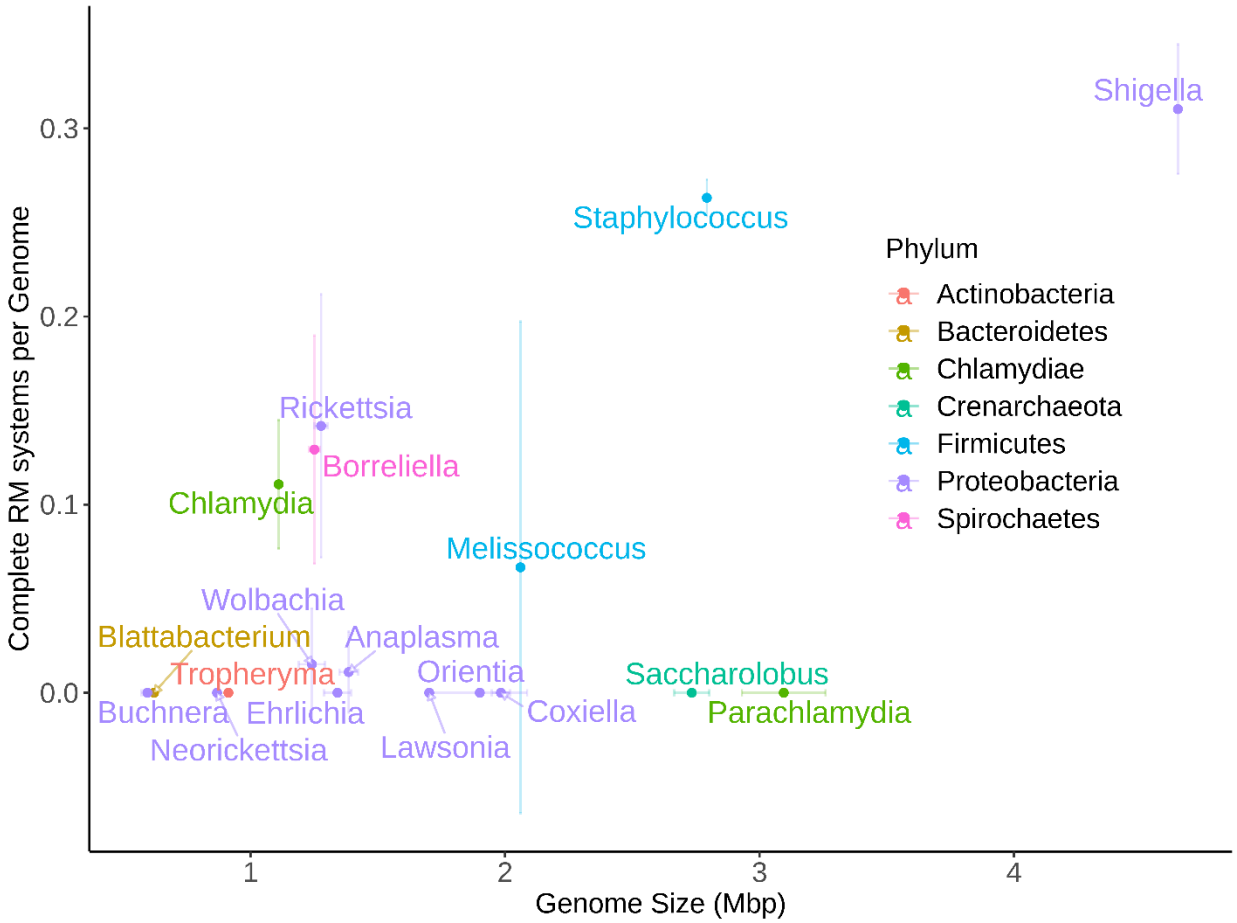

Figure 7. Genera Represent the 5% Quantile (≤0.427 RM) of the Prokaryotic RM Distribution. Mean number of complete RM systems per genome plotted against mean genome size in prokaryotic genera. Data points represent the mean of 5 or more isolates, hexagons are rendered when there are 5 or more data points. Genera were removed from this plot of the 95% confidence interval crossed the 5% quantile threshold.

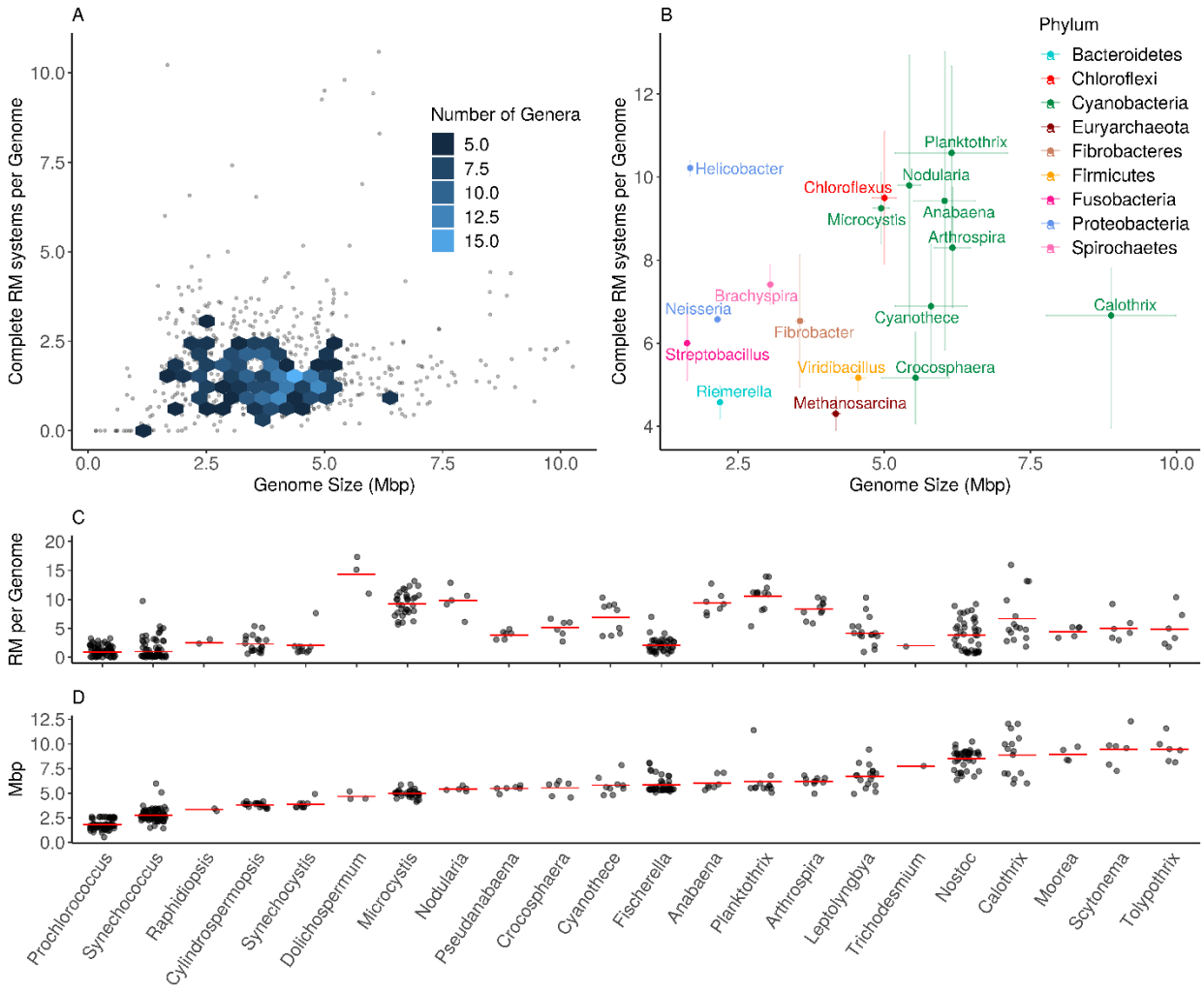

Figure 8. Distribution of Restriction Modification Systems in Prokaryotic Organisms without Type IIG RM found with HHblits. A) Mean number of complete RM systems per genome plotted against mean genome size in prokaryotic genera. Data points represent the mean of 5 or more isolates, hexagons are rendered when there are 5 or more data points. Mean RM and 95% confidence interval = 2.17+/- 0.119, Median RM = 1.91. **B)** 95% quantile (≥4.4 RM) of the RM distribution. Error bars represent 95% confidence intervals of the genus mean and data points represent 5 or more isolates. Genera were dropped if the 95% confidence interval feel below the 95th quantile. **C&D)** Subset of data from **(A)** which is restricted to only the Cyanobacteria Phyla, showing complete RMs **(C)** and genome size **(D)** per genome. Genera are in ascending rank order by genome size. Red bars indicate genus mean while datapoints are individual isolates. All genera have 5 or more isolates except for *Raphidiopsis* (n=2), *Trichodesmium* (n=1), *Dolichospermum* (n=3).

number of methyltransferases[141]. Our analysis confirmed these findings and indicated that many of these methyltransferases have associated endonuclease complements, making *Microcystis* one of the most RM-rich genera in this study (Figure 6B). In addition to *Microcystis*, we see a consistently strong signal from other bloom-forming cyanobacteria including *Planktothrix*, *Nodularia*, and *Anabaena* (Figure 6B)[142]. Moreover, we find that this signal is robust even with more stringent annotation calls (See methods, Figure 8), suggesting a strong association between bloom formation and RM abundance that necessitated further investigated (see below).

RM patterns in Cyanobacteria. Considering the planetary wide effects of phage, assessing the evolutionary drivers of RM gain and loss for all lineages of bacteria and archaea is a daunting and perhaps impossible task. We reasoned that a useful first approach to uncover such drivers is the analysis of related genera with well-characterized and distinct ecologies. To this end, we narrowed our investigation to the phylum Cyanobacteria.

As noted earlier, the high side of the Cyanobacteria RM distribution was dominated by the freshwater genera *Microcystis*, *Planktothrix*, *Nodularia*, *Dolichospermum* and *Anabaena,* where genome size is a poor indicator for RMs (Figure 6 C-D). *Dolichospermum*, a newly defined genus from isolates formerly aligned to the genus *Anabaena,* has the most RMs among these organisms and is characterized by the development of large blooms from eutrophication of water bodies[143]. Indeed, bloom formation is a phenotype that is consistent among these organisms as these genera all form dense blooms during their life history, and secondary metabolite (toxin)-producing

strains of these genera are known agents of HABs. These genera have more complete RMs in their genome than any other prokaryotic genera currently represented in our public databases.

The connection between very high copies of RMs and bloom formation in cyanobacteria is striking, though not universal. For example, *Trichodesmium* is well documented to form blooms in marine surface waters[144,145], but have a low number of RMs. Critically, blooms of organisms like *Microcystis* are much more severe in terms of biomass accumulation than most marine bloom formers, such as *Trichodesmium.* For example, satellite monitoring of *Trichodesmium* blooms across the planet show that blooms very rarely achieve chlorophyll *a* (a proxy for phytoplankton biomass) higher than 1 μg/L with the majority only showing 0.25 μg/L[146]. In contrast, bloom conditions of the western basin of Lake Erie are much more severe: *Microcystis* chlorophyll *a* levels in late July and early August averaged 14.8, 22.4, and 46.1 μg/L, in 2012, 2013 and 2014, respectively, where some stations peaked as high as 126.1 μg/L[147]. In lake Taihu, the highest recorded chlorophyll *a* concentrations from *Microcystis* blooms were ~105, ~115, and ~70 μg/L in 2012, 2013, and 2014, respectively[148]. Indeed, high density blooms seem to be a consistent ecological phenomenon of this genus, regardless of geological location.

The low extreme of the RM distribution is composed of the unicellular marine picocyanobacteria of the *Prochlorococcus* and *Synechococcus* genera (Figure 6C). Unlike the heterotrophic bacteria that populated the extreme low-RM category (Figure 7), these low-RM cyanobacteria are free-living. *Prochlorococcus* numerically dominates the low nutrient (oligotrophic) oceans and, while peaking only at about 10^5 cells ml^{-1}, is the most abundant photosynthetic organism on Earth[149]. *Synechococcus* also contributes

significantly to the oligotrophic phytoplankton community, and some genotypes can also be found at high abundance in nutrient-rich coastal environments or in freshwater systems[150].

Many *Prochlorococcus* genomes lack RMs altogether, and have a genus mean of 0.974 RM per genome. *Synechococcus* genomes on average contain more RMs (1.459 per genome), but together with *Prochlorococcus* are well below the mean for the Cyanobacterial phylum (7.444 RM per genome). Interestingly, genomes of *Synechococcus* strains more closely related to the oligotrophic specialist *Prochlorococcus*[151] showed statistically fewer RM ($p = 8.4E-05$, Wilcoxon rank-sum; Figure 9A). Moreover, when separated into freshwater and marine clusters, the marine *Synechococcus* had a statistically lower number of RM ($p = 0.0014$, Wilcoxon rank-sum; Figure 9B).

The low abundance or complete absence of RMs in these picocyanobacteria is a curious observation, because far from a virus-free existence, *Prochlorococcus* and *Synechococcus* are hosts to a diverse array of viruses, and these viruses are suspected to contribute significantly to mortality of their hosts *in situ*[152,153]. Thus, the low number of innate defense systems in these genera cannot be attributed to a lack of phages in their ecosystems.

Each extreme of RM abundance for the Cyanobacteria phylum was thus characterized by genera that numerically dominate their respective phytoplankton communities, but do so at vastly different population densities and under vastly different

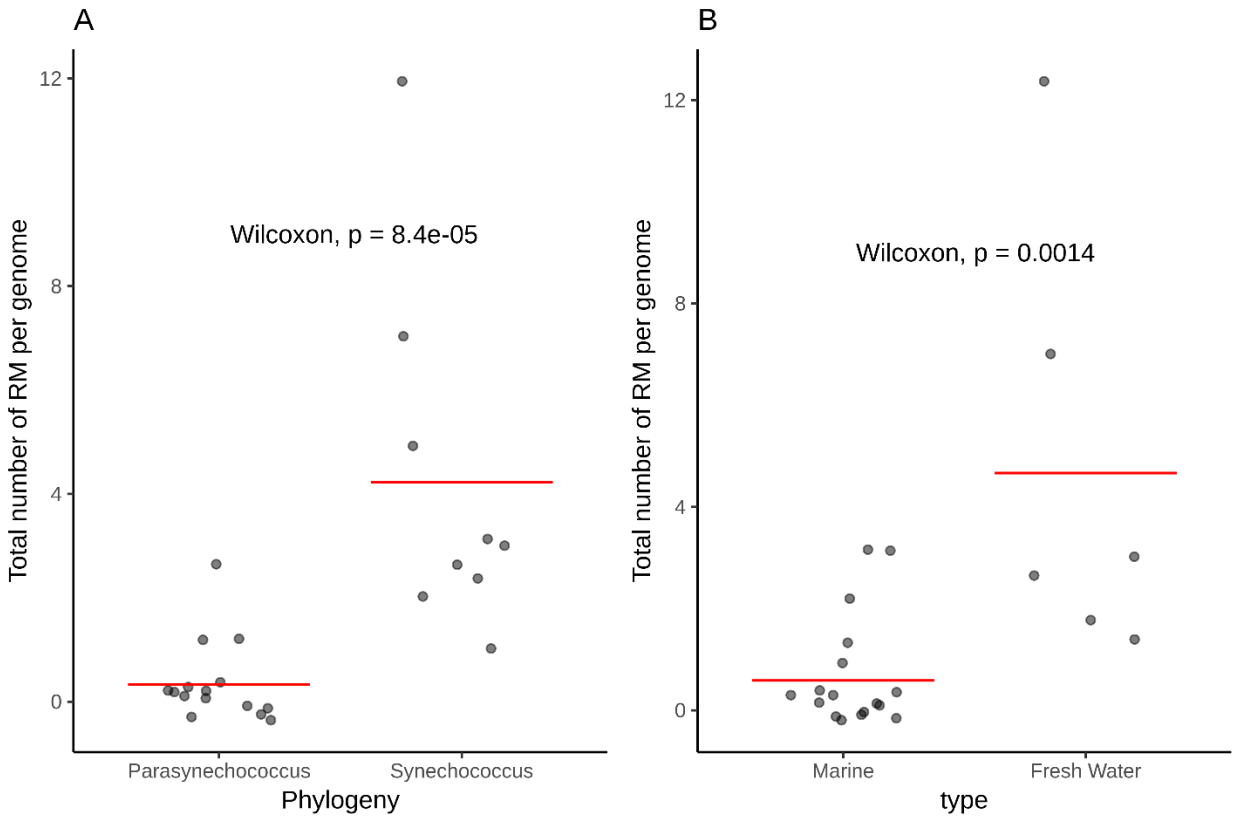

Figure 9. Comparing Total Number of Restriction Modification systems between different groups of Synechococcus. Individual points represent isolates while the red bar represents the category mean. **A)** RM counts in two phylogenetically distinct groups, isolates more closely related to the oligotrophic picocyanobacterium *Prochlorococcus*, referred to as *Parasynechococcus*, and the rest of the genus. **B)** *Synechococcus* isolates separated by their isolation from either marine or fresh water. See Coutinho *et al.* 2016 for further description of phylogenetic characterization.

resource availabilities. The high end of the RM distribution was dominated by HAB forming cyanobacteria, whose blooms are largely attributed to eutrophication of bodies of water from farm runoff carrying fertilizer, flooding the system with nitrogen and phosphate which promote life at high density[154]. Whereas the low end was dominated by oligotrophic picocyanobacteria, that are deprived of nutrients due to temperature stratification and large geographic distances from coastal inputs.

Given this pairing of extremes between resource and RM abundance, we hypothesized that high resource availability selects for the acquisition of RMs, to improve defenses at high cell density, whereas low resource availability selects for the loss of costly RMs to improve competitive fitness for scarce resources. To explore this hypothesis, we developed several models - involving various forms of viral-host interaction - that investigate how nutrient load affects the selective value of RM acquisition or loss in prokaryotes.

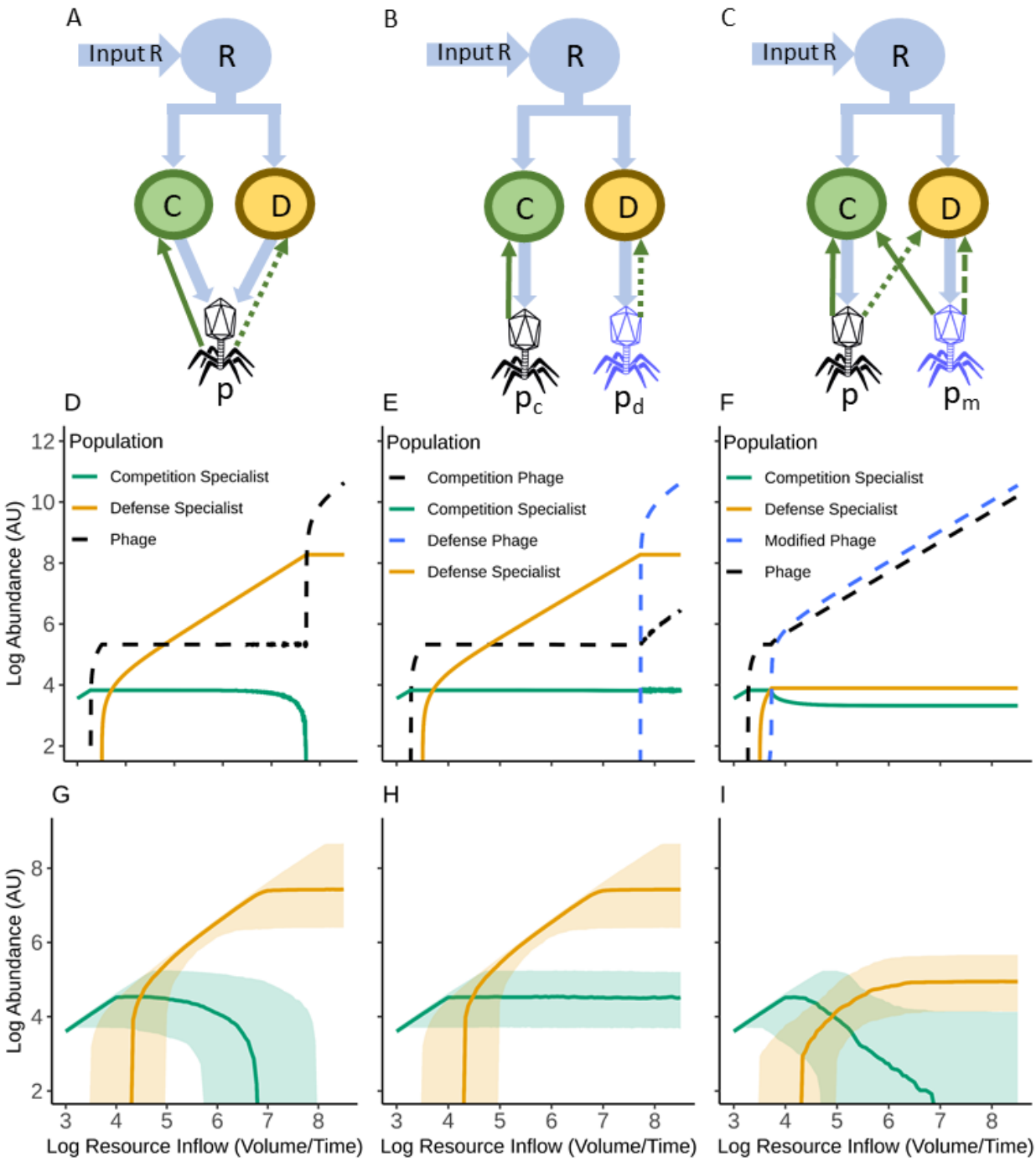

Figure 10. General, Parallel, and Memory Virus-Host Interaction Models. Figure columns correspond to general interaction **(A, D, G)**, parallel interaction **(B, E, H)** and memory interaction **(C, F, I)**. Model structures **(A-C)** show the mass transfer from resource, to two competing prokaryotic populations, and finally into phage. Green circles represent a population of competition specialist that are more competitive to resources **.**relative to the defense specialists, represented by the gold circles. Solid green arrows

Figure 10 (Continued)represent high infection rates, dashed lines represent intermediate infection rates, while dotted lines represent low infection rates. **(D-F)** shows the steady state abundance of each prokaryotic and viral populations across a wide range of resource input. Solid gold and green lines are the defense and competition specialist, respectively. Dashed black lines show the abundance of phage, while the dashed blue line shows the abundance of the competition phage in the parallel model, or the modified phage in the memory model. **(G-I)** Variation in steady state values for the competition and defense specialists from 990 simulations with parameters drawn from a LHS scheme. Solid lines are the median value of each population while the shaded regions show the 75^{th} and 25^{th} quantiles. For parameters, please see Table 1.

Competitive Exclusion at Nutrient Extremes. Three viral-host interaction models representing contrasting ecological contexts (Figure 10A-C, see methods) all produce similar results when competing prokaryotic populations in the presence of phage: increased resources always selects for populations with more RMs per genome (Figure 10D-F) and these results are robust to a large range of parameters (Fig 10G-I, see methods). However, only a 'memory model' (Fig 10C) - which implements the efficiency of memory for host-methylated virions - can reconcile realistic population sizes with observed counts of RM per genome (Fig 11). Moreover, we find that differentiation in the identity of RMs between competing populations promotes coexistence in the memory model and suggests intense pressure for RM innovation (Figure 12). We describe these findings and argue that viral methylation is a critical mechanism which links molecular efficiency of endonucleases and genomic distribution of RMs in prokaryotes.

With three viral-host interaction models, numerical simulations were performed over a variety of resource inflow rates to explore how resource supply affects co-cultures of a prokaryotic community composed of a competition specialist (n=0 RM) and defense specialist (n=1 RM) in the presence of phage. Outcomes for each model could be binned generally by resource inflow: low, mid-range, and high. In the 'general model' – where a single generalist phage can infect both hosts - low nutrient inflow established a steady state monoculture of the competition specialist, as both the defense specialist and the phage were eliminated from the system (Figure 10D, 10G). Mid-range resource inflow is characterized by co-existing steady-state populations of the competition and defense

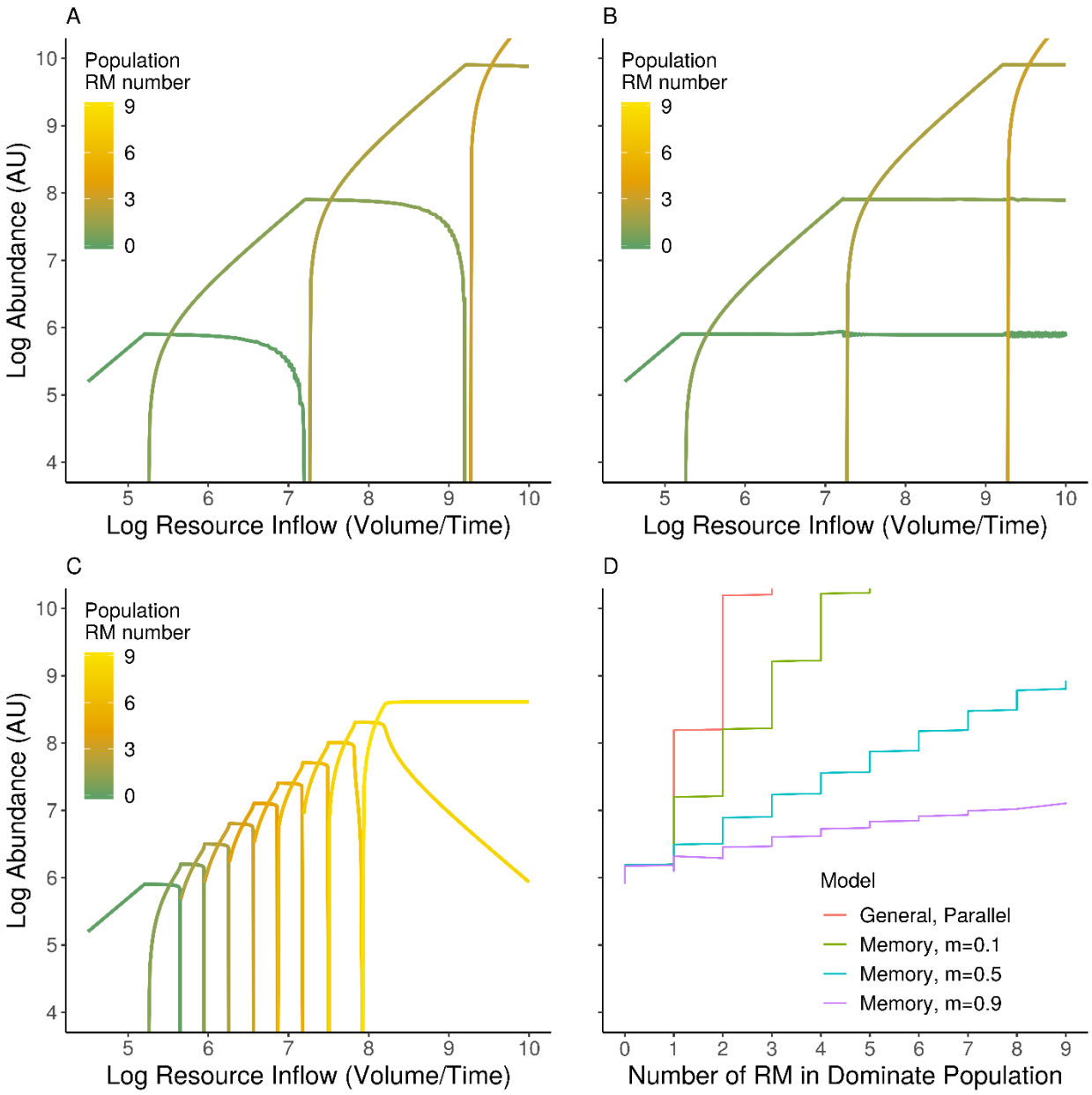

Figure 11. Abundance Scaling with Increasing Defense Types in General, Parallel and Memory Models. (A), **(B)** and **(C)** show the steady state abundances of populations carrying different numbers of RM systems. For simplicity, we assume the cost and resistance of each additional RM system are identical. **C)** Efficiency of memory (m) is set to 0.5 and RM systems are in a "subset" arrangement (see main text for description). **D)** Total abundance of each community plotted against the number of RM systems in the dominant subpopulation. General and Parallel models are identical, while the scaling of the memory model depends on the partial resistance conferred. For other parameters, Table 1.

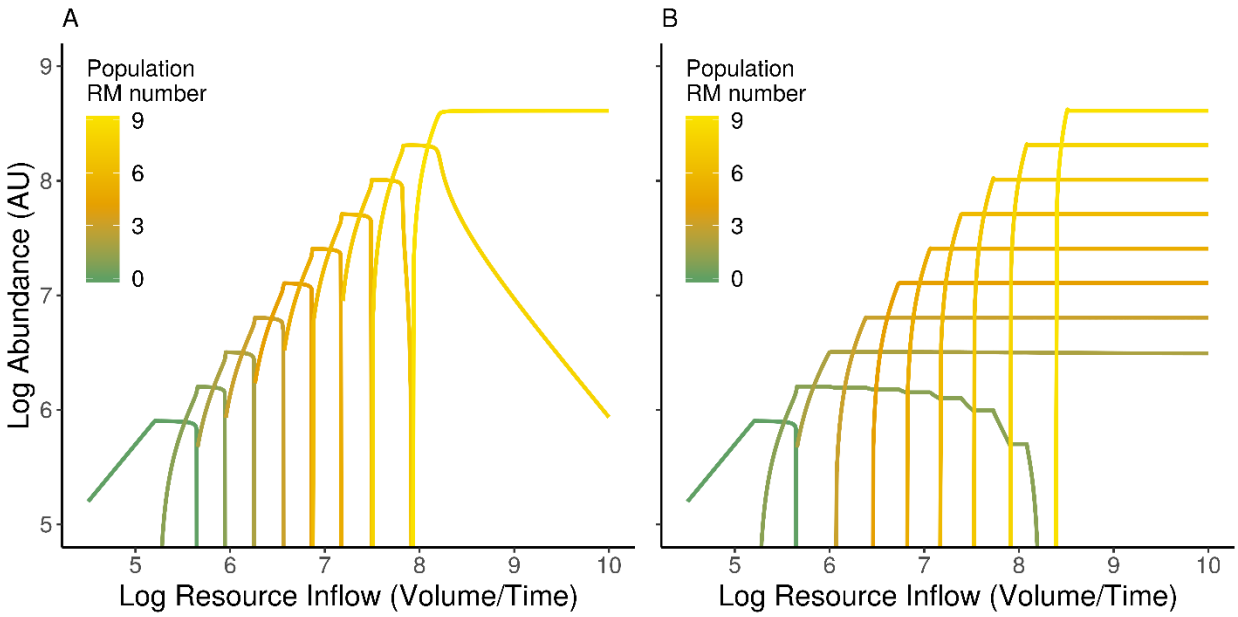

Figure 12. Identity of RM Systems Among Populations determine coexistence of Competitive and Defensive types in the Memory Model. (A) and **(B)** show "subset" and "unique set" RM communities, respectively, with theoretical methylated viral bursts (m = 0.5).

specialist. Within this range, the defense specialist cell density scales with resource flow rate, whereas the competition specialist density is held in check by the virus, facilitating the transition in numerical dominance from the competition to the defense specialist. At the highest nutrient inflow examined, the system enters a new state where the competition specialist is driven to extinction, and the defense specialist scales with resource until it its density is held in check by phage, the latter scaling with resource input.

In the 'parallel model' – where each host is infected by distinct phages - outcomes for the competition and defense specialists at low- to mid-range resource inflows are similar to the general model (Figure 10E, 10H). In contrast to the general model, however, the parallel model predicts stable co-existence of defense and competition specialists at high resource inflow.

Qualitatively, competitive outcomes in the memory model appear to resemble a mix of the general and the parallel model (Figure 10G-I). Low resource inflow selects monocultures of the competition specialist, and the defense specialist invades the system at roughly the same resource inflow rate as the general and parallel model predict (Figure 10F, 10I). At high density, fitness of competitive (n=0 RM) or defensive (n=1 RM) types was dependent upon parameter selection. In Fig 2F, the defense specialist is dominant at high resource inflow, resembling the general model. Yet, a small number of the replicates in the memory model resulted in the competition specialist being the dominant member (Figure 13). The spread of the green ribbon in Fig 10I reflect this heterogeneity of outcome in the memory model.

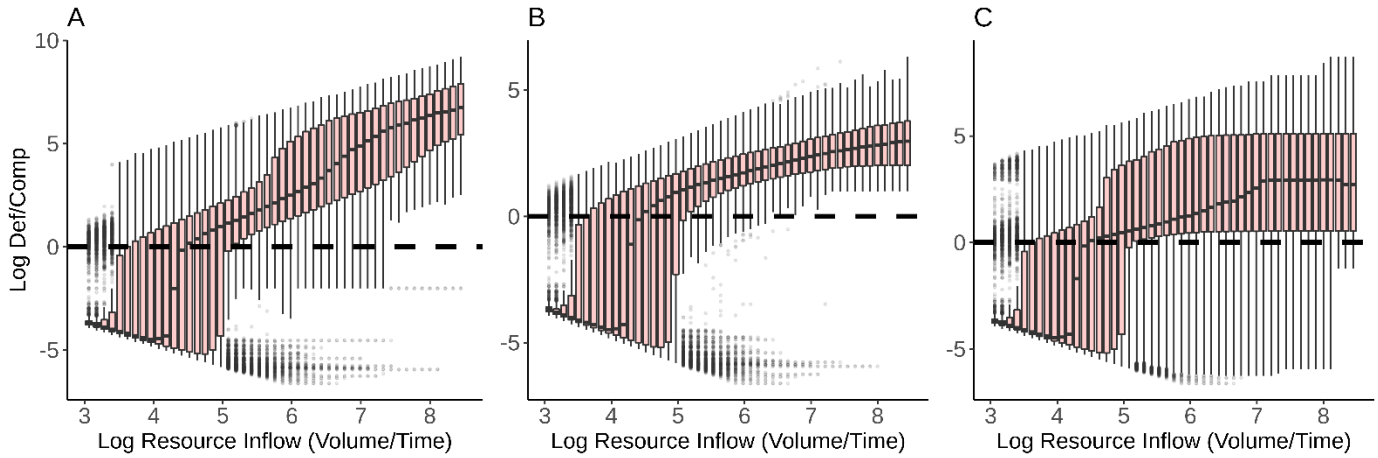

Figure 13. Ratios of Defense and Competition Specialists from 990 LHS replicates. Box and whiskers plot showing the log ratios of the steady state abundance between the competition and defense specialist. For plotting purposes, the abundance of either population was forced to 1 if it was less than 1. Dashed black line shows the point at which the abundance of each population is equal.

The parallel and general model may be thought of as endpoints of the memory model. When efficiency of memory is low, viral progeny from the defense specialist are hypomethylated and are susceptible to the defense specialist's endonuclease. Thus, the host has high resistance to both unmethylated and hypomethylated virions. As such, the range of coexistence with low efficiency of memory diminishes along supply gradients as in the general model (Figure 14, 15). Conversely, when efficiency of memory is high, viruses of the defense specialist are hypermethylated and can resist the defense specialist's endonuclease. With hypermethylation, the effectiveness of the defense specialists RMs is greatly diminished against methylated virions which effectively establishes parallel infections and facilitates coexistence, even at high resource inflow (Figure 14, 15).

In all three models, defense specialists with more RM are always selected at high resource inflow, unless the cost of RM's was so high that defensive groups were unable to compete, regardless of resource inflow (Figure 14, 15).

Modeling RM escalation and de-escalation.

Carrying one RM system at high resource inflow appears to confer selective advantages at high cell densities over hosts with no RM. When present in the same cell, RMs targeting different DNA sequences confer additive effects on viral defense[129,155]. One might therefore expect multiple RMs to confer additive protective benefit. If each RM system confers a moderate infection reduction of 10^2 and the protection is additive, two systems have a reduction of 10^4. The protective effect of additional RMs would quickly outstrip the number required to protect an entire population. Eight RM would protect populations up to 10^{16}. It appears puzzling, therefore, that the genera *Microcystis*

which has maximal cell densities ~10^8 ml^{-1} [148], carry in excess of 15 RMs (Figure 6B) and makes reconciling the efficiency of endonucleases difficult with their copy number per genome.

Interestingly, this simple arithmetic is evident when hosts with n= 1, 2, 3 RMs compete for resources in both the general and the parallel model (Figure 11A,B). Even when moderate resistance is assumed ($r = 10^{-2}$), just three RMs is enough to fully protect against viruses over a realistic range of cell densities. Hosts with greater than three RM are not selected, even for modest assumed costs of carrying an RM (Figure 11A,B). The memory model drastically differs from the predictions of the other two models (Figure 11C). Viral methylation weakens the protective effect of each additional RM, leading to far more modest gains in cell abundance per RM along the resource gradient (Figure 11C).

The weakening of RM-mediated protection is strongly dependent on the efficiency of viral memory. Deviations from 100% efficiency may derive from imperfect viral memory of prior host's RMs due to methylase limitation[156] and imperfect efficiency of foreign DNA invasion due to unmethylated host restriction sites[155]. A high efficiency of memory leads to a very gradual increase in cell abundance as a function of resource (Fig 11D, purple line, $m = 0.9$). Declines in memory efficiency lead to commensurate increases of cell density and numbers of host RMs ($m = 0.5, 0.1$; Fig 11D), approaching those for the general and parallel models (where $m = 0$; Fig 11D).

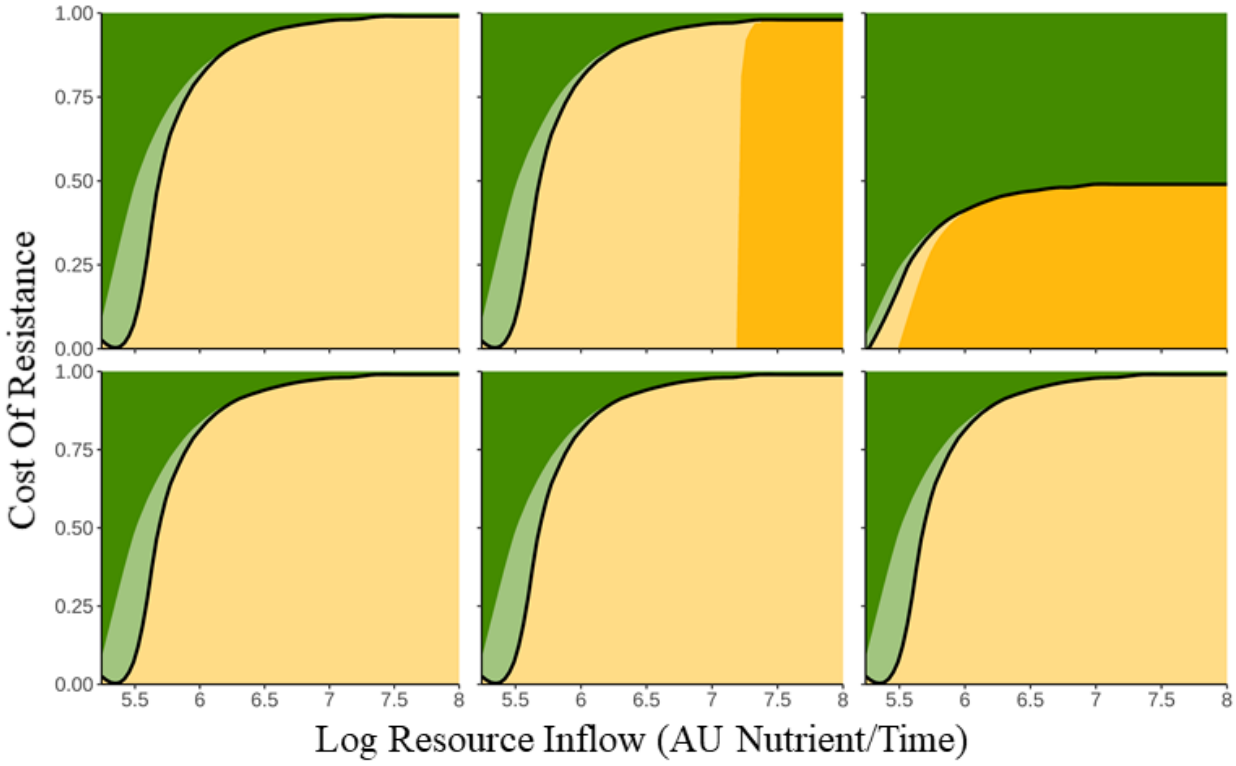

Figure 14. Cost of Resistance of General and Parallel Models. Steady states of numerical simulations at a variety of costs, resistances, and nutrient inputs for the general (top row) and parallel models (bottom row). The columns, from left to right, correspond to a resistance of 0.001, 0.01, and 0.5, respectively. Dark green = competition specialist only; Pale green = competition specialist is the dominate member; Pale gold = defense specialist is the dominate member; Dark gold = defense specialist only. Black line represents where the competition and defense specialists are in equal abundance.

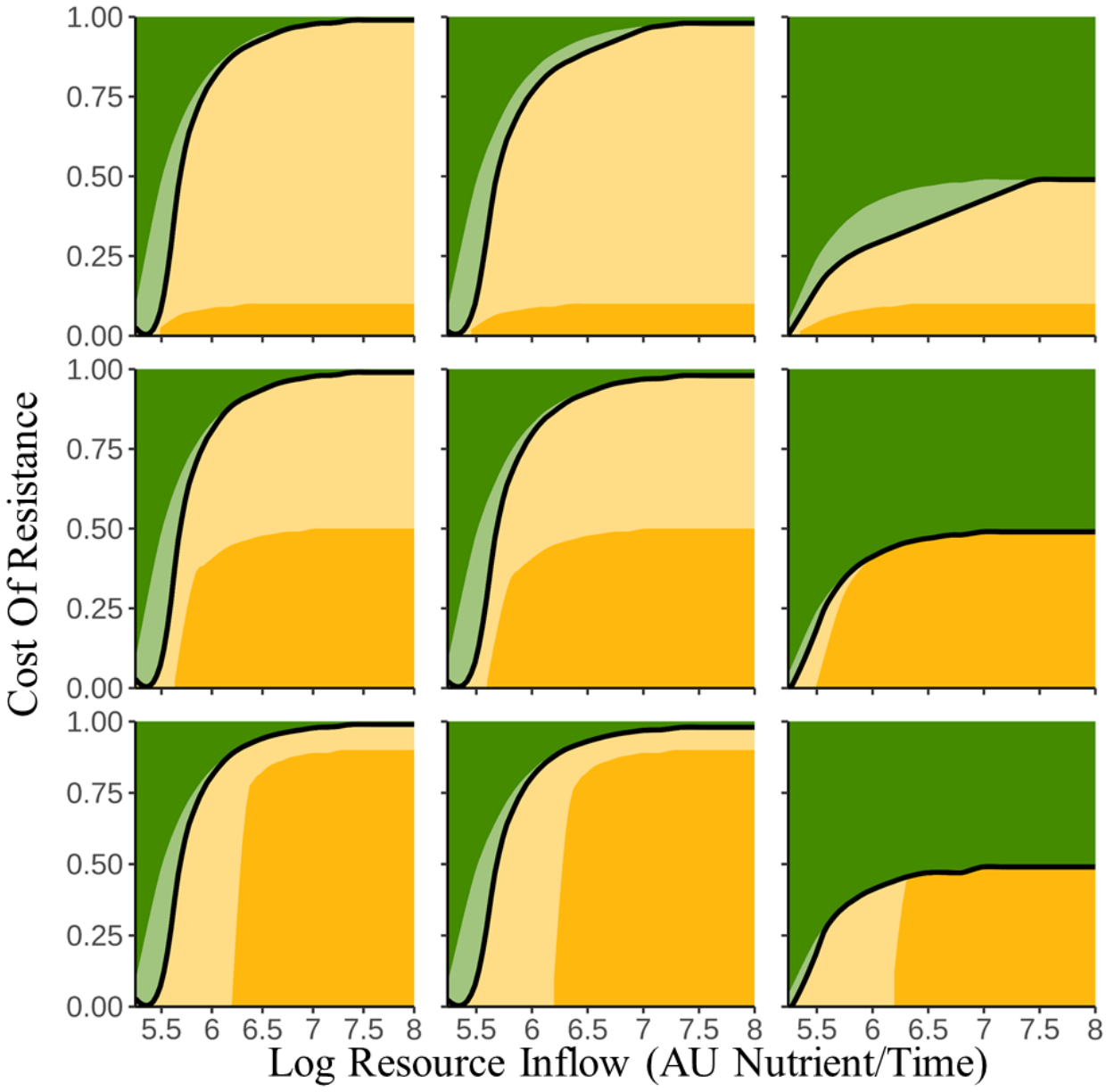

Figure 15. Cost of Resistance of Memory Model. Steady states of numerical simulations at a variety of costs, resistances, partial resistances, and nutrient inputs for the memory model. The columns correspond to a resistance of 0.001, 0.01, and 0.5, respectively. The rows, from top to bottom, correspond to a efficiency of memory of 0.9, 0.5, and 0.1, respectively. Dark green = competition specialist only; Pale green = competition specialist is the dominate member; Pale gold = defense specialist is the dominate member; Dark gold = defense specialist only. Black line represents where the competition an defense specialists are in equal abundance.

RM identity impacts coexistence between competitive and defensive populations

So far, we have assumed that host acquisition of a new RM system augments existing resistance. This presumes that organisms are only able to increase their RM suite though gene gain in a linear fashion. We relax this assumption by allowing RM diversification to emerge through multiple rounds of gene gain and loss. We hypothesized that innovation of novel RM would promote coexistence among diverse RM number. To test this scenario, we altered the original distribution in RM identity by making all populations have "unique sets" of RMs. In other words no subsets in RMs are shared between competitive and defensive types (empty intersection of RMs). For both the "subset" and "unique set" scenario, increases in nutrient inflow rate leads to numerical dominance by genotypes with progressively higher number of RMs (Figure 12). Communities that share RMs in a subset structure promote competitive exclusion of cells with fewer RM (Figure 12A), while a unique set structure in RMs promotes coexistence between nearly all populations (Figure 12B) at sufficiently high resource loads. In other words, strict sub-setting of RMs drives to extinction all but the competitor with the optimal number of RMs, whereas types with varying amounts of RMs can coexist, so long as the RMs are not exclusively of the same subset.

III. Discussion

We explored patterns in the distribution of RMs (*i.e.*, RMs per genome) in almost 140,000 bacterial and archaeal genomes, and used mathematical modeling to demonstrate that evolutionary gain or loss of RMs can be driven by resource availability. Bioinformatic data indicated that RM distribution varies extensively at the genus level,

ranging from zero to over a dozen per genome, and these distributions lack robust patterns across bacterial and archaeal domains. At the phylum level, however, we observed a clear link between environmental resource availability and RM abundance in Cyanobacteria. Consistent with this observation, general, parallel, and memory models all predicted cells gain RMs at high resource availability, and lose RMs at low resource availability. Of the three, only the memory model, which incorporates the unique "virus immunization" feature of RMs, could account for the extensive escalation and de-escalation of RM defenses that was evident in the cyanobacterial genomes.

Prior reports indicated that the number of RMs per genome correlates with genome size[61,132,133]. While this correlation held true in our study, especially for small genomes, we found that genome size was overall a poor predictor of RM content across all bacteria and archaea, and could not account for the extreme cases of RM accumulation we observed. For instance, within the Cyanobacteria, the seven genera with the largest genome sizes had only average RM abundance, whereas the genera with the most RMs had average to below-average genome sizes (Fig 6C). This diminished role of genome size may be due to our much larger sample size, binning by genus, and/or restricting our analysis to genera with at least five genomes sequenced. Regardless, it was clear that we still lacked an evolutionary understanding for why some genera have so many RM, while others have so few.

Reasoning as others have[86,132] that the quantity of RMs per genome was intrinsic to ecological strategy, we narrowed our investigation to a bacterial phylum with genera of divergent and well-characterized ecologies, the Cyanobacteria. We reasoned that the evolutionary drivers of RM gain and loss would be most apparent in genera occupying

the extremes of RM abundance. Critically, we found that these RM extremes were occupied by genera that numerically dominate aquatic systems at the two extremes of resource availability. The oligotrophic ocean is dominated by *Prochlorococcus* and *Synechococcus*[157], and these genera had very few RMs. In contrast, eutrophic systems promote dense blooms of genera[142,154,158] such as *Microcystis*, *Planktothrix*, *Anabaena*, and *Dolichospermum*, and these genera had the highest number of RMs per genome in the Cyanobacteria (Fig 6C).

From this phylum-level survey, we hypothesized that growth in high resource environments selects for acquisition of new RMs, evolving a specialization for defense, whereas growth in low resource environments selects for the loss of RMs during the evolution of competition specialization. General, parallel, and memory population models of virus-host interactions all predict that total number of RMs is modified by environmental resource supply. Clearly evident in the contrasting predictions of the general, parallel, and memory models are the effects of virus methylation and predator-prey 'memory' on host-virus dynamics. Critically, only the memory model can reconcile the magnitude of RMs per genome observed bioinformatically with experimentally measured levels of protection, and also explain the pressures that lead to rapid genomic turnover of RMs[138,159–162]. A link between methylation-based viral memory and population dynamics is novel with respect to other theoretical[86,163–166] and experimental[167–169] investigations linking resource availability and anti-predator defense.

The memory model may help to explain why RM composition varies over small phylogenetic distances. RMs are positively associated with homologous recombination and horizontal gene transfer between microbial species[136], and we suspect it is not a

coincidence that RM are rapidly turned over in genomes[138,159–162]. We note that the evolutionary history of the genus *Microcystis* is largely plagued with indicators of extreme genomic plasticity and include high proportions of genes deviating from genome average GC content, high numbers of repeat sequences, high numbers of insertion sequences, and poor genomic synteny between isolates[170,171] which could be the result of extreme pressure to innovate viral defense systems.

In our theoretical community where RMs were subsets, defensive hosts exclude competitive hosts with increasing nutrients, effectively purging the community of all genotypes but the one with the optimal number of RMs (Figure 12A, Figure 16 top). In contrast, for communities with unique sets, competition specialists with different, albeit richer, suites of RMs could invade and/or coexist (at lower abundances) rather than suffer from competitive exclusion under high resource conditions (Figure 12B, 16 bottom). This outcome for unique sets could help explain why RMs are so tightly associated with horizontal transfer - the protective value of a newly acquired RM system is not only maximal when modified (i.e. protected) viruses are rare, but divergence in methylomes can promote the coexistence between n RM and n+1 RM populations (Figure 12, 16).

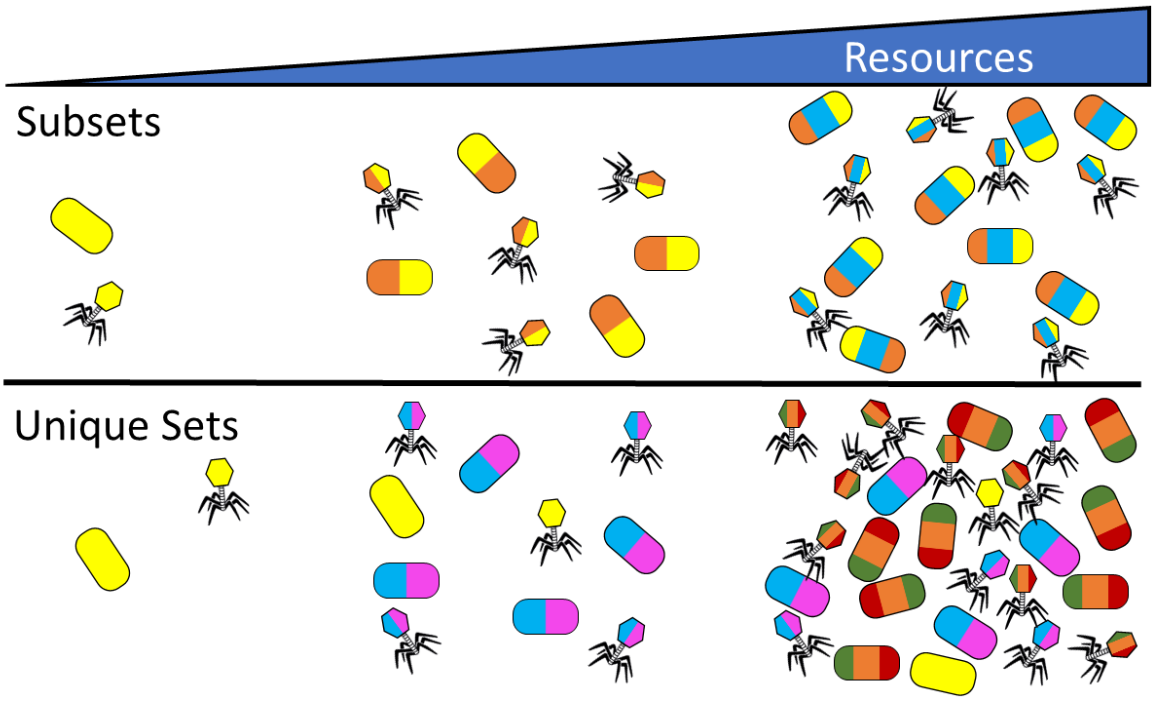

Figure 16. The Impact of RM Identity on Community Succession. Prokaryotes are in competition for resources in the presence of phage. Each color rendered on the bacilli represents a different complete RM system, while colors rendered on phage represent the adopted methylome of the host. As resources increase, the optimal number of RM per genome increases to gain resistance to phage, however, the identity of RM distributed among the different subpopulations can promote competitive exclusion (subsets of RM) or coexistence (unique sets of RM).

The subset arrangement of RMs (Figure 16 top) fails to acknowledge the importance of gene loss as an evolutionary driver of diversification. And, because it leads to extinction of competitors with sub-optimal RM abundances, would be also expected to perform poorly under high variance in selective pressures, such as high temporal variation in nutrient loads. In contrast, with a unique RM arrangement (Figure 16 bottom), competition specialists with low RM abundances are protected from purging during periods of nutrients increase, but can also dominate when nutrients become limited. We believe this variation between increased and decreased selective pressures, coupled with gene loss and gain, ultimately leads to the diversification of RMs we observe in bioinformatic data.

We suspect that both subsets and unique sets exist in nature, at least temporarily for the former. Sub-setting will drive the optimization of RM abundances within a population that adds or loses RMs sequentially, but unique sets will drive the diversification of RM abundances within a community of populations, because they promote co-existence of genotypes with unrelated RMs. Importantly, in both set arrangements, the dominant member of the community is the one with the optimal number of RMs per given resource availability; the distinction is what happens to the suboptimal members of the community.

Virus methylation and RM mediated defense are novel mechanisms linking host-virus populations with molecular control, but many other mechanisms drive realized dynamics. Prokaryotes can develop resistance to phage through a variety of mechanisms including alteration of phage receptors, production of extracellular matrix, and the

production of inhibitors[5,119]. Given the disparity in the quantity of RM per genome between picocyanobacterial and HAB cyanobacteria, we expect other costly defensive mechanisms (e.g. CRISPR, toxin-antitoxin) to also be enriched in eutrophic environments and dispensed in oligotrophic environments over evolutionary time[172], and the observation that prokaryotes with CRISPR-Cas systems have statistically higher numbers of RMs per genome may reflect this reality[61]. Additionally, Forde et al. (2008) demonstrated how cell surface phenotypes, lipopolysaccharide lengths and their interaction with outer membrane proteins, can generate different infection mechanisms and is qualitatively identical to our general and parallel model structure and their outcomes[173]. Notably, cultures of *Prochlorococcus*, cyanobacteria of low RM content, are readily taken over by cell envelope mutants upon exposure to phage[116]. While nutrients were not varied in that study, it is tempting to speculate how resource availability could impact fitness of those resistant genotypes. While our current models focus exclusively on the benefits of RM loss and gain, a future area of investigation will be to explore the interplay between RM and CRISPR, receptor modification, and other defense innovations as a function of resource availability.

Of all models investigated, the model invoking imperfect viral memory led to the closest qualitative match with genomic and environmental observations. The extent to which released viruses become immunized by methylation has been understudied and perhaps unappreciated thus far, leaving us with several unanswered questions. What is the range in efficiency of memory of RM-mediated viral methylation amongst the diverse classes of Type I, II and III RMs? How do Type IV RMs, which target methylated rather than un-methylated DNA, impact the relationship between memory and fitness under

varying resource conditions? How does hemi-methylation effect both the cost of an RM system, as well as the possibility for partial methylation of viral progeny? Addressing these questions is critical to link mechanisms of molecular defense with population dynamics and to understand how RMs function in natural prokaryotic communities.

IV. Methods

Bioinformatic Search Strategy

Because of the diversity in both genomic and domain architecture of RMs, we chose a strategy that uses both BLAST 2.7.1+[174] and HMMER 3.1b2[175] to generate alignments to our reference database and then refined our results by using genomic context. Protein profiles are built from Hidden Markov Models (HMMs) and allow us to identify putative methyltransferases or endonucleases by searching for the specific functional motifs in proteins. By using profiles, we could explicitly detect functional motifs within proteins regardless of the domain architecture, a problem local alignment algorithms like BLAST cannot resolve unless there is a protein with an identical architecture capable of generating full alignments. To ensure we were not aligning multiple profiles to the same residues in each protein, we 'competed' profiles that align to 75% of the same residues in a protein and select the profile with the lower e-value. We used hmmscan with gathering cutoffs to collect all Pfam (release 31) HMMs[176] that represent experimentally characterized 'Gold Standard' methyltransferases and endonucleases found in New England Biolabs' REBASE (Dataset_S03)[128]. We finalized our reference HMMs after manual curation (Figure 17-20, Dataset_S04). In curation, we found ResIII (PF04851) domains were repeatedly observed in various helicases and

transcriptional regulators. Since ResIII was common in type I, type IIG, and type III RMs, we retained this domain in our reference HMMs, but added HMMs that would covary with ResIII when a protein was a not part of a RM system to flag false positives (Dataset_S05). Endonucleases or methyltransferases that could not be detected with our reference HMMs were used in secondary BLAST searches to generate alignments to prokaryotic proteomes (Dataset_S06). Alignments were considered a match if the total alignment length was 75% of the query length, and the evalue ≤ 1E-5. Once we generated alignments, we used genomic context to count the number of full RMs. An RM system was considered complete if there was an endonuclease ≤ 4000 base pairs away from a methyltransferase, or if both motifs were detected in one peptide.

Large proteins (>750 AA) that contained a methyltransferase domain but did not show any additional motifs to indicate endonuclease activity were subjected to a more sensitive search algorithms part of the HHsearch suite[177] to evaluate if they were type IIG RMs as protein size alone can discriminate between type IIG RM and other methyltransferases (Figure 21). We first pre-clustered these putative type IIG RMs using psi-cd-hit[178,179] with a clustering threshold of 35% sequence identity and an alignment that covers at least 85% of each protein (parameters: -c .35 -aL .85 -aS.85 -g 1). Once the clusters were defined, representatives from each cluster were used to build profiles for HHblits. Clusters were considered type IIG proteins if the representative sequence aligned to 3S1S[180], 4PXG[181], 4XQK[182], 4ZCF[183], 5FFJ[184], or 5HR4[185]. Three iterations were used to build multiple sequence alignments with mact=0.35. Parameters for hhsearch are as follows: p=20, Z=250, loc, z=1, b=1, B=250, ssm=2, sc=1, seq=1, norealign, maxres=32000, contxt = context_data.crf.

This more sensitive analysis revealed that 27,147 of the original 33,633 flagged proteins could be aligned to verified type IIG RMs in the protein data bank. While we believe there is a high likelihood these are RMs, we wanted to determine if our initial findings (Figure 6) depend on the veracity of our type IIG calls. When we reanalyzed our RM collection without these putative type IIG RMs, the RM distribution was qualitatively the same: *Planktothrix, Microcystis, Nodularia* and *Anabaena* still dominated the tail end of the distribution, but quantitatively closer to with *Chloroflexus* and *Helicobacter* in RM per genome (Figure 8B).

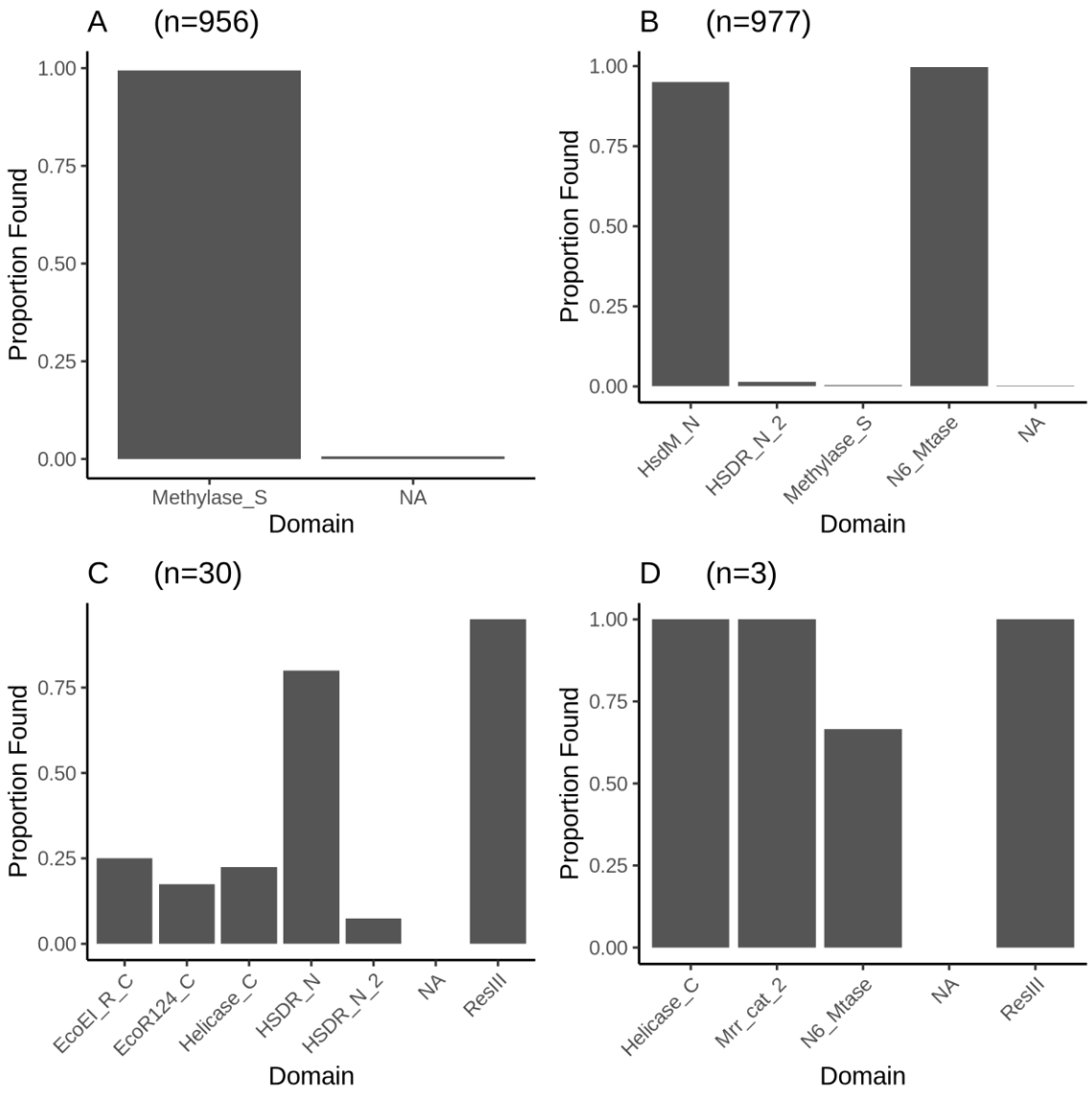

Figure 17. Domains in biochemically characterized Type I RM systems from New England Biolab's REBASE. Individual columns represent the frequency of a protein profile within n number of proteins. The NA column shows the proportion of proteins that did not have a detectable protein profile. **A)** Frequency of domains found is the specificity subunit of Type I RM systems. **B)** Frequency of domains found in Type I methyltransferases. **C)** Frequency of domains found in type I endonucleases. **D)** Frequency of domains found in Type I RM systems that had subunits concatenated together.

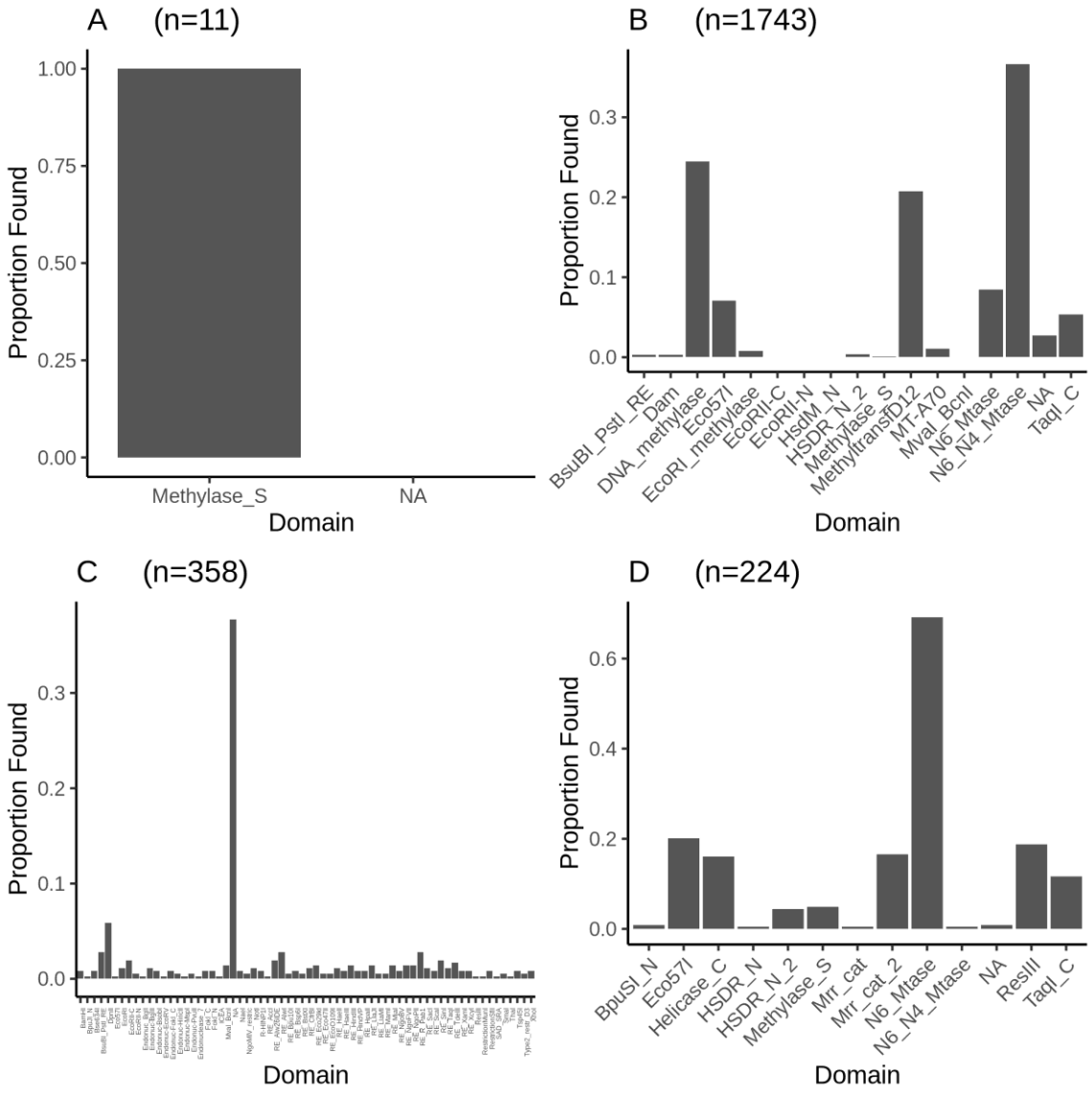

Figure 18. Domains in biochemically characterized Type II RM systems from New England Biolab's REBASE. Individual columns represent the frequency of a protein profile within n number of proteins. The NA column shows the proportion of proteins that did not have a detectable protein profile. **A)** Frequency of domains found is the specificity subunit of type II RM systems. **B)** Frequency of domains found in type II methyltransferases. **C)** Frequency of domains found in type II endonucleases. Note that the largest column is NA, indicating that most of these proteins require alignments via BLAST **D)** Frequency of domains found in type IIG RM systems.

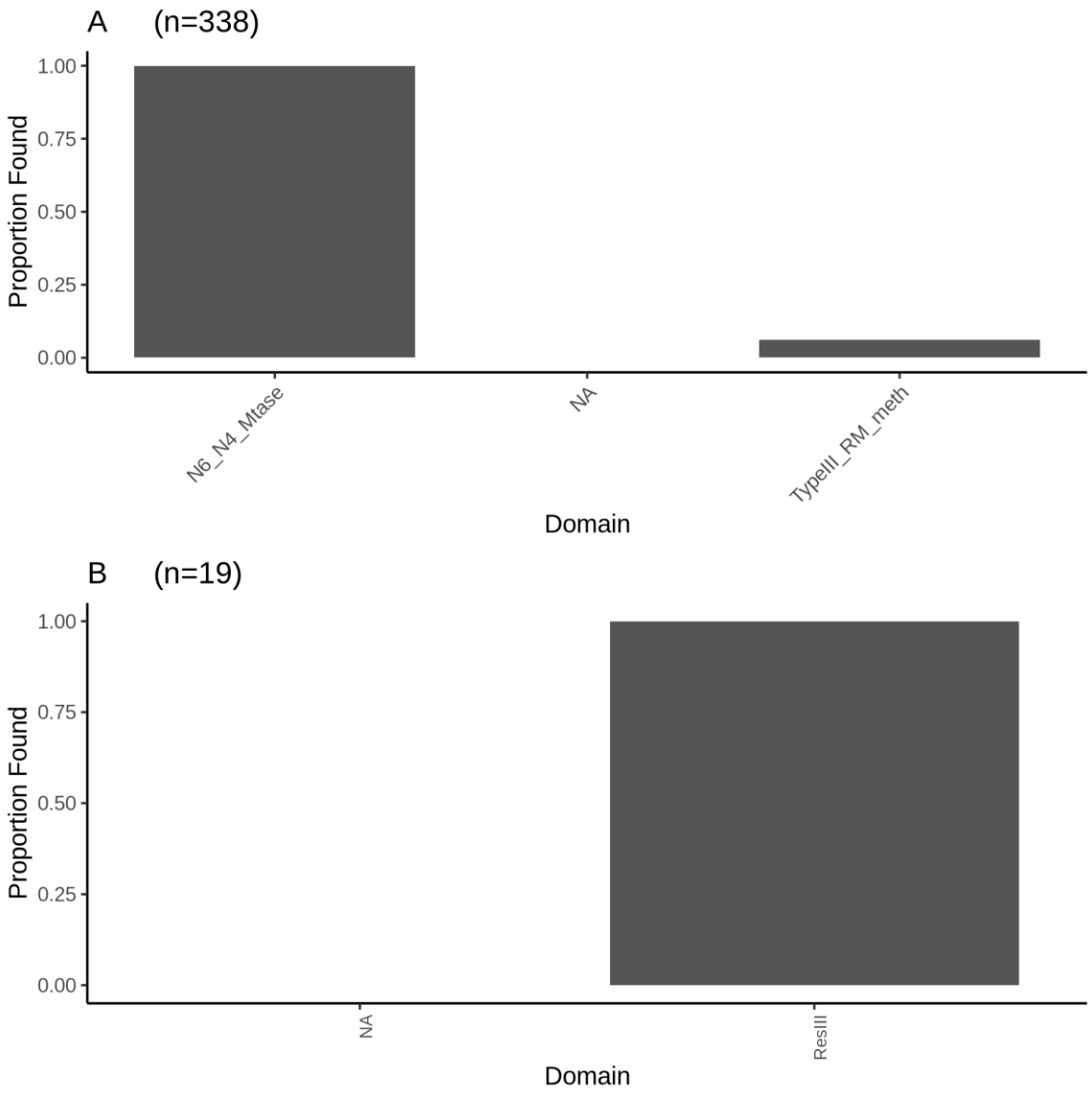

Figure 19. Domains in biochemically characterized Type III RM systems from New England Biolab's REBASE. Individual columns represent the frequency of a protein profile within n number of proteins. The NA column shows the proportion of proteins that did not have a detectable protein profile. **A)** Frequency of domains found in type III methyltransferases. **B)** Frequency of domains found in type III endonucleases.

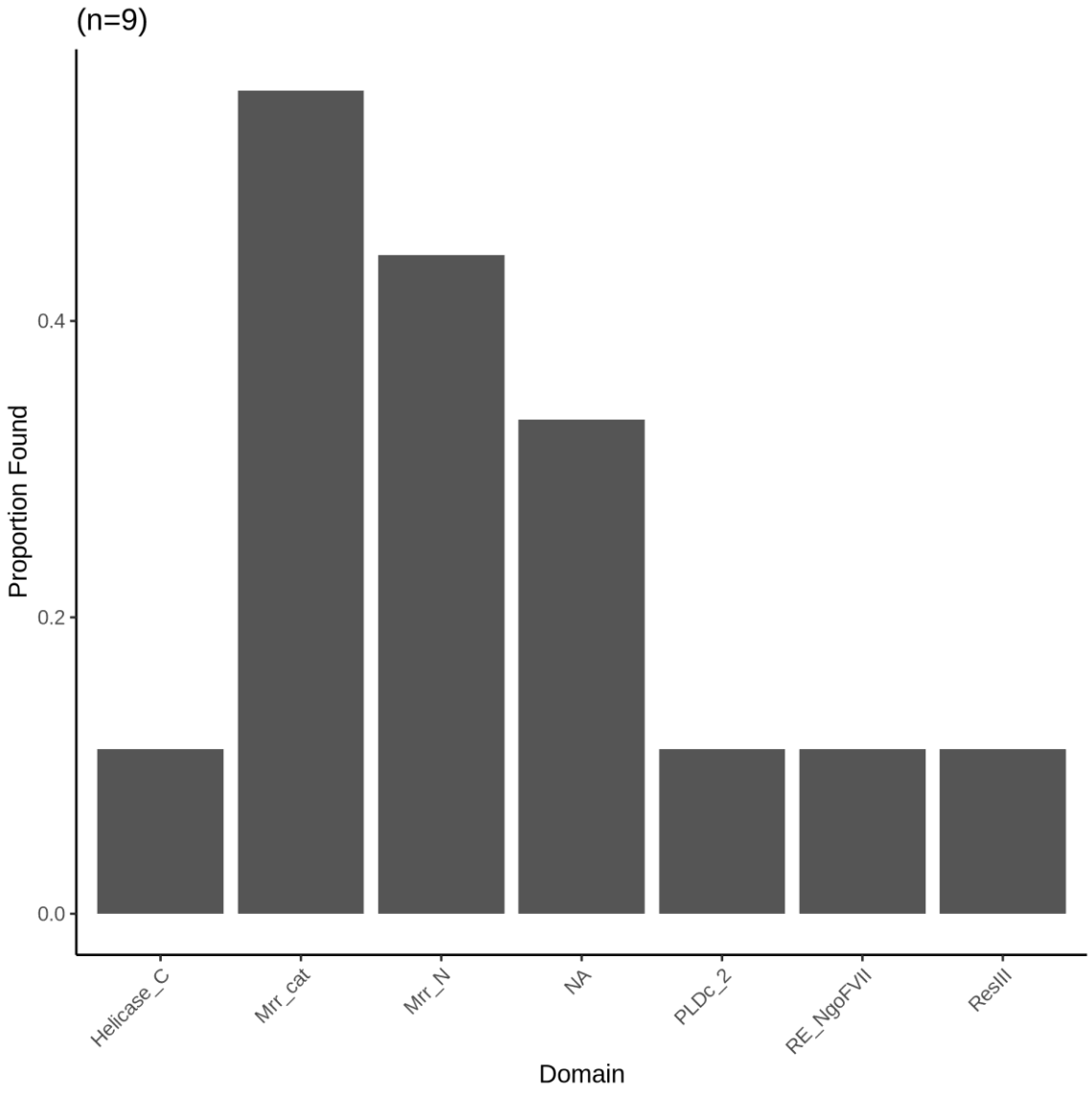

Figure 20. Domains in biochemically characterized Type IV RM systems from New England Biolab's REBASE. Individual columns represent the frequency of a protein profile within n number of proteins. The NA column shows the proportion of proteins that did not have a detectable protein profile.

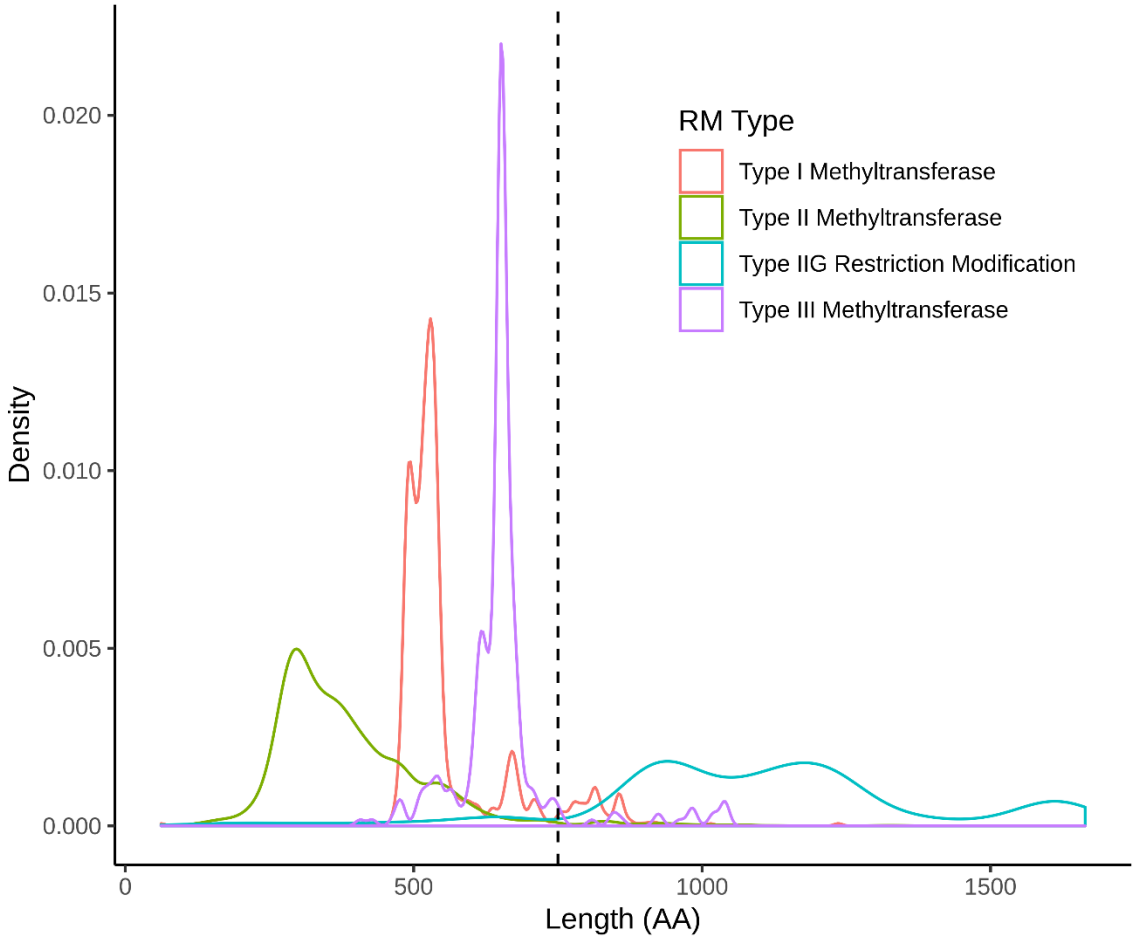

Figure 21. Distribution of Protein Lengths from Gold Standard Methyltransferases. All biochemically characterized proteins from New England Biolab's REBASE were plotted to evaluate the size distributions. We see that the minimum size for most of the type IIG RM systems is 750 AA at the dashed black line.

Viral-Host Interaction Models

To model the effects of resource on the selection for defense, we competed two theoretical prokaryotic populations for a single resource in the presence of phage. Modeled prokaryotic populations differ only in the number of defense systems (*i.e.,* RMs) they carry, where defensive types have more RM, and thus greater resistance to phage, relative to competitive types that have fewer RMs. We further assumed a trade-off: investment in RM increases the resistance of the defense specialist to phage, but this comes at a cost to resource utilization[186].

We explored the influence of this trade-off on competitive outcomes within three hypothetical system structures with contrasting representations of virus-host interaction (Figure 10A-C). All three systems involved competition between two host prokaryotic strains. In the general interaction model (Figure 10A), the competition and defense specialists are infected by the same phage. In the parallel model (Figure 10B), the competition and defense specialists are infected by phage that do not cross-infect the other host. In both of these first two models, the viral defense is generic, and RM could be substituted with phage receptor modification, CRISPR, etc. In contrast, the memory model (Figure 10C) is a variation of the general model that accounts for the unique feature of RM defenses: the "memory" bestowed upon surviving phage by the methyltransferase component of the RM defense system in the defense specialist, which renders the phage resistant to the restriction endonuclease[125–127,129].

All population model simulations were computed using SciPy's integrate module[187]. All simulations launched with different fixed nutrient inflow were allowed to

reach an equilibrium steady state. The final abundances of each population were plotted against the simulation resource inflow to show system state changes in R using ggplot2. A Latin Hypercube Sampling (LHS) scheme[188] was used to randomly pick model parameters over uniform distributions in specified ranges, except for resistance (r) and baseline infection rate (ϕ), which were drawn from log-uniform distributions. LHS is favorable over brute-force random sampling because previous samples are used to make intelligent draws for the next sample, ensuring that random draws are representative of parameter variation in multidimensional space. LHS was done with the pyDOE module. Sampling ranges for cost and resistance were restricted to meet the assumption costs do not outweigh resistance, or in our model, $r > 1\text{-}c$. Data was managed using Pandas[189].

Table 1. LHS Parameter Values. Values used for simulations in Figures 10, 11, 12, and 13. Please see methods for a detailed description of equations.

Description	Parameter	Figure 2D-F (LHS replicate 230)	Figure 2G-I LHS Sampling Ranges	Figure 3,4 (Unless noted)
Growth Rate	α	1.1791	Uniform: 0.75 - 1.25	1.0
Burst Size	β	38	Uniform: 5 - 49	25
Baseline defense	ϕ	9.1648E-7	Log Uniform: 1E-8 - 1E-6	1E-8
Cost	c	0.5	Uniform: 5E-6 - 0.999	0.1
Resistance	r	0.4114	Log Uniform: 1E-4 - 0.0991	0.01
Bacterial loss	δ_b	0.2783	Uniform: 0.15 – 0.3	0.2
Phage loss	δ_p	0.2345	Uniform: 0.15 – 0.3	0.2
Efficiency of memory	m	0.8529	Uniform: 4.08E-4 – 0.999	0.5

Model Structure

To explore the selective pressures on bacteria (and other prokaryotes) to increase their defense, we model the competition of *i* bacteria in the presence of *j* phage. The model explicitly defines the rate of change of resource in the system as

$$\frac{\mathrm{dR}}{\mathrm{dt}} = \mathrm{S} - \sum_{i=1}^{n} \alpha_i \, (1 - c_i) b_i R \qquad (1)$$

Where R and b_i are the abundances of resource and i^{th} bacterial population at time t, respectively. S is the nutrient flowing into the system (mol N time^{-1}), while α is the resource utilization rate (cell/mol N) and c is the cost of resistance where 0 < c < 1 and is a function of the total number of RMs in bacterial population i. We assume that RMs' costs are identical, and that cost is linear, thus, if one RM system causes c_i to be 0.05, two RMs will cause c_i to be 0.10 for a respective bacterial population. Growth of each bacterial population is defined as

$$\frac{\mathrm{db}_i}{\mathrm{dt}} = \alpha_i (1 - c_i) b_i R - \sum_{j=1}^{n} \phi r_{ij} b_i p_j - \delta_b \mathrm{b}_i \qquad (2)$$

Where ϕ is the base line infection rate of a bacterial population, δ_b is bacterial loss, and p_j is the density of a phage population j at time t. The resistance of bacterial population i to phage population j, stored in the matrix r_{ij}, is a function of the number of RMs of the bacterial population i, or

$$r_{ij} = r^{|RM_i|} \qquad (3)$$

Where r is the resistance conferred by each RMs and RM_i is the total number RMs carried by bacterial population i, thus total resistance to phage j is the multiplicative protection of all RMs. We treat RMs in an organism as a mathematical set which imposes all RMs are unique. Vertical bars denote the cardinality of the set, where cardinality is the total number of elements in RM_i. Biologically, we can think of this as each RM system targeting different recognition sequence in DNA. The assumption that RM are multiplicative is not unfounded as Arber and Wauters-Willems (1970) were able to demonstrate that the defensive value of multiple RMs greatly reduce the efficiency of plating (e.o.p) of phage on host *E. coli* strains[129]. For example, in these experiments, one RM system decreased e.o.p by $1x10^{-2}$, another RM decreased e.o.p by $3x10^{-5}$, and together decreased e.o.p of phage to $6x10^{-7}$. For simplicity, we will assume that all RMs in our models have the same protective value. Finally, phage replication is determined by

$$\frac{\mathrm{dp}_j}{\mathrm{dt}} = \beta_j \sum_{i=1}^{n} \phi r_{ij} b_i p_j - \delta_p \mathrm{p}_j \qquad (4)$$

Where β_j is burst size of phage j and δ_p is phage loss. Altogether, equations 1-4 represent a diamond food-web ecosystem with predation being "keystone" to maintaining diversity within the ecosystem[163–165], where our general model reduces to a simple diamond food web with a single phage as the predator of competing bacterial populations. Equations 1-4 could be extended in an infinite number of ways, for example to include separate compartments for different infection pathways[190] or non-linear interactions[191]. Equations 1-4 represent the most transparent and parsimonious model to explore how, to first-order,

resource environment selects number of RMs. Explicit representation of non-linear interactions and additional infection pathways would introduce additional unconstrained model parameters, while also limiting the clarity with which the mechanisms driving RM selection can be presented. Nevertheless, we also considered two additional ecologically pertinent model configurations.

The parallel interaction model differs from the general interaction model in that for each host population, there is a single viral population can cause infection, which is accomplished by forcing off-diagonal values of matrix r_{ij} to be zero. In this scenario, the phage are restricted to a smaller host range relative to those in the general interaction model, and collectively these two models encapsulate the diverse, often nested viral-host interaction networks among microbial communities, where there are some generalist phage that infect most members and specialist phage that infect few members[192].

The 'memory interaction model' addresses the biological consequences of DNA methylation by RM defenses. Pioneering research in the 1950s revealed that the efficiency of infection depended on which host strain the virus was replicated[125,126]. This dependency was later shown to result from modification of viral progeny[193], specifically, the methylation of viral DNA. Host methyltransferases methylate all DNA indiscriminately, including any replicating viral DNA that evades host defenses long enough to be modified. In this manner, viral progeny adopt the host's methylation pattern, which confers immunity to the virus when infecting another bacterial cell with the same RM system(s). This adoption of host methylation patterns can be thought of as viral 'memory' of its most recent prey. We can incorporate this detail by altering the summation in equation (4):

$$\frac{\mathrm{dp}_j}{\mathrm{dt}} = \beta_j \sum_{j=1}^{n} \phi r_{ij} b_i p_j - \delta_p \mathrm{p}_j \quad (5)$$

With this alteration, viral populations can only be replenished from a specific bacterial population and implies that the number of bacterial subpopulations is equal to the number of viral subpopulations. To incorporate the differential infection generated from the modification of viral DNA from host RMs, we alter how our r_{ij} matrix is generated:

$$r_{ij} = r^{|RM_i - RM_j|}\, m^{|RM_i \cap RM_j|} \quad (6)$$

RM_i and RM_j denote the RM carried by the host and methylation state of the virus, respectively, and are treated as mathematical sets. Similar to equation 3, RMs contribute to host resistance, however, the cardinality in the difference between RM_i and RM_j (RM in *i* and not in *j*) contribute to resistance. The cardinality in the intersection between RM_i and RM_j (RM in *i* and in *j*) can still contribute to resistance of bacterial *i* to phage *j*, however, this depends on the efficiency of memory, *m*. Thus, methylation shared between hosts and viruses cause partial resistance. The assumption that difference between viral and host RM sets determines resistance has empirical support. For example, Arber and Wauters-Willems (1970) showed viruses methylated by one out of two RMs resulted in an e.o.p. as if the host had only a single RM system. Due to e.o.p. being a relative metric for viral success in infection, it cannot quantify efficiency of memory, which to the best of our knowledge currently has no direct empirical constraint. We consider instead sensitivity of our main predictions allowing inefficient / incomplete methylation due to methylase limitation[156], and unmethylated restriction sites[155]. These two observations

necessitate that, as long as the methylation of viral progeny is not perfect during replication, viral progeny are susceptible to RM albeit at a much lower frequency, or $m > r$. For simplicity, we assume the efficiency of memory conferred from all RMs is equal.

We consider the general model to be a special case of the memory model. When $m = r$, we can simplify the exponents in equation 6. This simplification leads to equation 6 becoming identical to equation 3 because $\left|RM_i - RM_j\right| + \ |RM_i \cap RM_j\ |$ is equal to a cardinality of $|RM_i|$. Intuitively, we can think of this as a complete lack of viral methylation which leads to the predicted outcome of the general model.

CHAPTER 4
Phylogenetic Signals and Functional Roles of Methyltransferases

Contributions

Spiridon Papoulis and Erik Zinser designed research; Spiridon Papoulis performed research; Spiridon Papoulis and Erik Zinser analyzed data.

Acknowledgments

We would like thank Scott Emrich for useful conversations in protein classification and Steven Wilhelm for the many conservations about methyltransferases.

ABSTRACT

Methyltransferases are essential parts of the Restriction Modifications systems in Prokaryotic genomes, but also play functional roles in other parts of prokaryotic physiology, such as DNA repair or global regulation. In terms of phylogenetic distribution among prokaryotic genomes, methyltransferases empirically shown to be important in global regulation, such as Dam and CcrM methyltransferases, are rooted in the phylogeny of the organisms in which they are found, showing patterns of conservation. In contrast, methyltransferases used as part of restriction modifications systems are typically sporadically distributed, a likely result of horizontal gene transfer. In this work, we expand on how well Dam and CcrM methyltransferases are conserved in a large dataset of Proteobacterial genomes, confirming that these methyltransferases are vertical transferred though evolution. We find this signal of conservation may indicate a functional role, however, we find interesting patterns of methyltransferase conservation in organisms with high amounts of Restriction Modification systems. These data suggest that phylogenetic distribution should be used with caution when trying to infer the functional role of a methyltransferase, especially when in the genomic context of many RMs, as they produce similar phylogenetic signals.

I. Introduction

Restriction Modification systems (RMs) are ubiquitous in prokaryotes and are exceptionally robust phage defense systems due to the efficiency of the DNA cutting activity of endonucleases. Methyltransferases are essential components of RMs because they protect host genomic DNA from degradation (see chapter 1). In the context of viral host interactions, we have proposed that methylation incentivizes the constant innovation of RMs in organisms, either though changing the DNA target recognition sequence or through loss and gain of new RMs.

DNA methylation as a biological function, however, can go far beyond that of viral defense. One role some methyltransferases play is in the epigenetic modification of DNA. Methylation of nucleic acids causes physical alterations in the curvature of the double helix[194]. This, in turn, can affect the binding affinity of proteins to nucleic acids[195], such as transcriptional regulators, and is the underlying mechanism in epigenetic gene regulation. Colloquially, epigenetics has been described as non-genetic heritable changes in gene expression, generating further variation in organismal phenotypes[196]. In bacteria, the effects of epigenetic DNA modification have been heavily studied through experimentation of well characterized methyltransferases, namely *dam*, and *ccrM* methyltransferases, both of which lack restriction endonuclease partners[197,198].

The *dam* methyltransferase targets GATC and has been implicated in several cellular functions in *Escherichia coli*[197], including chromosome replication initiation and methyl-directed mismatch repair (see below). Microarray data shows that *dam*- mutants have large changes in transcriptional profiles, where genes responsible for respiration and bacterial motility were significantly different than that of the wildtype[199]. These

transcriptional changes covaried with GATC motifs in the promoter regions and suggest, at least in *E. coli*, that *dam* methylation is important for proper gene regulation. Indeed, single-molecule real-time (SMRT) sequencing was able to show that most sites along the wildtype *E. coli* genome were indeed methylated and those methylation patterns changed over the cell cycle[200]. Importantly, Westphal et al. demonstrated that *dam*- mutants had a significant negative impact to fitness in long term stationary phase and which suggests that the loss of epigenetic modification can be selected against in certain genetic backgrounds[200].

Independent of regulation, *dam* methylation is also used by DNA mismatch repair (MMR) to differentiate between the methylated template and the newly synthesized daughter strand that remains unmethylated for a short time after replication[197]. MMR in *E. coli* is composed of *mut* proteins functioning in tandem with dam methylation. Initially, *MutS* identifies mismatched base pairs, which then recruits *MutL* in an ATP-dependent fashion[201,202]. *MutL* then recruits *MutH*, which is responsible for digesting the erroneous, unmethylated daughter strand[203]. After digestion with *MutH*, exonucleases are recruited to remove the mismatched site and then polymerase III refills the removed error in the daughter strand[204]. The operation is complete after the repaired section is covalently linked with DNA ligase.

CcrM (cell-cycle-regulated methyltransferase) is similar to the *dam* methyltransferase in that it has large effects on global cellular regulation. CcrM, targeting GANTC, was first observed in *Caulobacter crescentus*[205]. *C. crescentus* asymmetrically divides into two morphologically distinct cells. The chromosome replicates in predivisional cells, and the two copies remain in a hemi-methylated state until just prior

to cell division, when ccrM is expressed and the daughter strands become methylated[205,206]. Interestingly, if *ccrM* is ectopically expressed throughout the cell cycle, cells become morphologically abnormal. Transcriptional studies showed that *ccrM* methylation, when absent or overexpressed, caused hundreds of genes to be misregulated, including 3 global transcriptional regulators[198]. The influence of CcrM methylation on *Caulobacter crescentus* regulatory network is thus profound, even if the molecular nature of this influence is not well understood: the lack of CcrM methylation on the binding sites of global regulators suggests that the effects of the methylation are more cryptic and may be affecting regulation on the peripherals of the regulatory network.

Since methyltransferases such as Dam and CcrM have alternative cellular functions beyond restriction signals that care must be taken when trying to infer the selective forces that govern the retention of any single methyltransferase. Our own previous research has helped us understand the selective forces that govern RMs: increased environmental nutrient loads increases host (and viral) abundance. Because of this, the selective pressure for viral defense increases to compensate for the increased contact rate between host and viruses. Moreover, our own mathematical models show that diversity of target recognition sequences are necessary to combat the viral memory associated with RMs. In contrast, methyltransferases, such as *dam* and *ccrM* seem to be essential for the proper function of cellular regulatory networks in their respective hosts. Indeed, previous research shows that *dam* methylation is found throughout *Gammaproteobacteria*[207] and *ccrM* found throughout *Alphaproteobacteria*[198], suggesting that the functional role of these methyltransferases are deeply rooted in their phylogeny. In this study, we attempt to get an up-to-date contextualized phylogenetic view of

methyltransferases. We hypothesize that methyltransferases important for global regulation, like *dam* and *ccrM,* will show a non-sporadic taxonomic distribution and a monophyletic evolutionary signal, while methyltransferases of RM systems will show sporadic distributions and display a polyphyletic signal. In addition, we investigate if phylogenomic context is a reasonable criterion for differentiating between regulatory methyltransferases from those involved in phage defense. Differentiating between these methyltransferases will be key in guiding our interpretation of why organisms might retain methyltransferases and guide future experimentation.

II. Results

Methyltransferase distributions in Proteobacteria Our analysis of methyltransferase phylogeny began with the Proteobacteria because many of their methyltransferases have been tested empirically and their recognition sequences validated. To ensure our data were from high quality genomic material, our analysis was restricted to complete genomes, where each genome belonged to a Proteobacterial genus that has 5 or more isolates. Our final dataset of 8,683 genomes, distributed among 142 genera and 19,850 unique methyltransferases, resulted in 3,673 arbitrarily named clusters

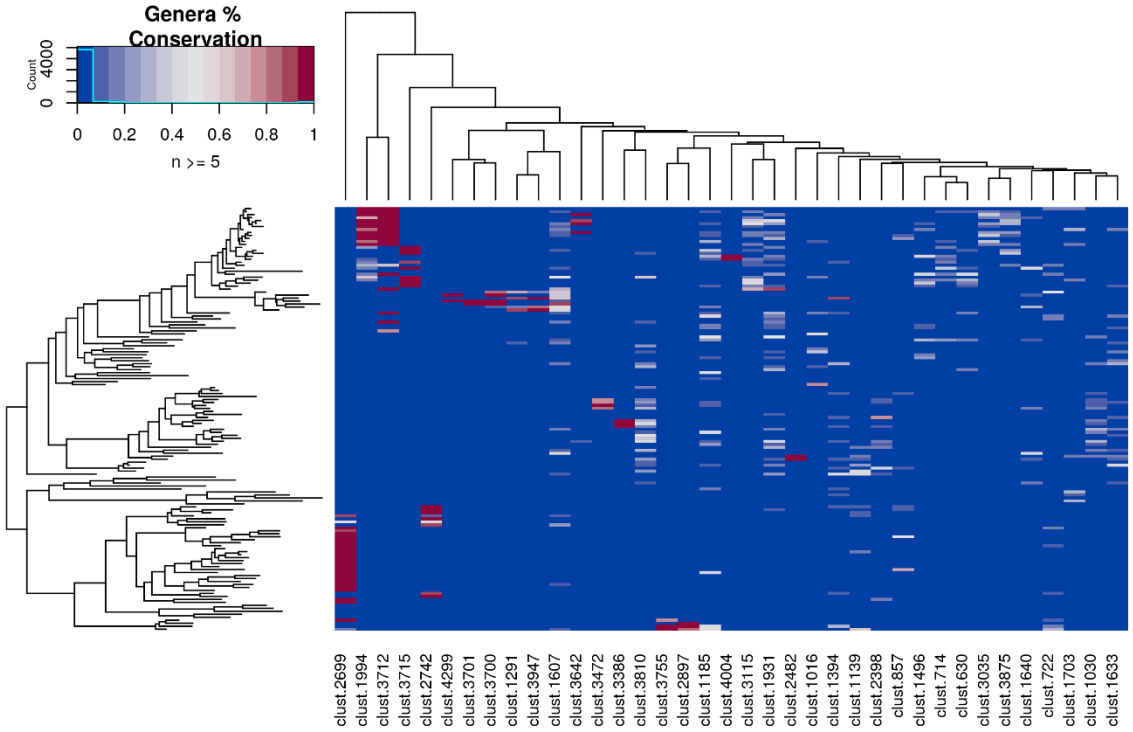

Figure 22. Phylogenetic Distribution of Most Common Methyltransferase Clusters. A heatmap showing the presence/absence of the most widely-distributed methyltransferase clusters (columns) in different Proteobacterial genera (rows). Columns were arranged by complete-linkage hierarchical clustering, whereas rows were arranged by phylogenetic distance using the r16S gene of each genus. Colors in the heatmap show percent distribution of each methyltransferase cluster in each genus, with red indicating all isolates from that genus have a methyltransferase homolog from the respective cluster. The blue line in the "Genera % Conservation" show the count distribution for the values in the heatmap.

at 60% identity cutoff (see methods). Using our custom classification scheme, we found that most (66.2%) of these protein clusters were type II methyltransferases, while other clusters were identified as type IIG (15.7%), type III (11.1%), or type I (10.9%) methyltransferases. We note that a very small percentage of clusters were labeled as ambiguous (0.1%) or concatenated versions of type I RMs (0.1%).

We found that most methyltransferase clusters are present in only single genera in Proteobacteria (not shown), however, a few methyltransferase clusters were consistently observed in multiple Proteobacterial genera (Figure 22). Figure 22 shows the distributions of these conserved methyltransferase clusters: 20 of which are type II methyltransferases, 15 are type II, and only 2 are type IIG. Clusters 2699 and 3712 were highly conserved within the genera they are found, and spanned higher taxonomic ranks above the genus level. These clusters contain experimentally-validated *ccrM* or *dam* methyltransferase homologs, respectively, whose activity indicates methyltransferases in these clusters are involved in other cellular functions beyond that of RMs. Moreover, we found that the methyltransferase tree topologies of these two clusters are reflective of 16S rRNA trees, suggesting the selective pressure has been consistently maintained throughout these organisms' evolutionary history (Figure 23-24). Specifically, we see methyltransferases from the same genus or very closely related genera, such as *Shigella* and *Escherichia*, forming distinct clades.

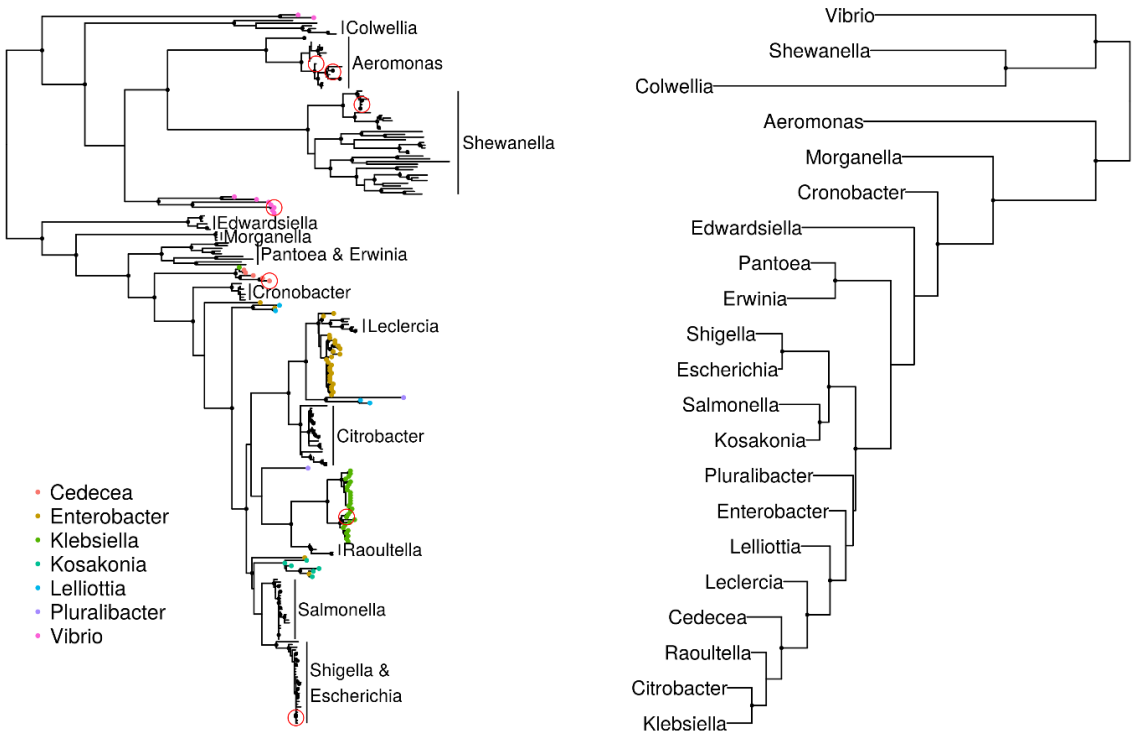

Figure 23. Phylogeny of Dam Methyltransferases vs ribosomal 16S rRNA gene. At the left, the phylogeny of cluster 3,712 from Figure 1 was annotated with the source genera and built using protein sequences. The right shows the 16S rRNA gene phylogeny from the genera from where methyltransferases were found. The internal nodes of each tree were annotated with black points if the support values were ≥ 0.80. Methyltransferase leaf nodes were annotated with colors if they were not considered monophyletic, while monophyletic nodes were annotated with their genus source. Nodes were considered monophyletic if all methyltransferase from a single genus formed a clade or if two closely related genera, such as *Shigella* and *Escherichia* or *Pantoea* and

(Figure 23 continued) *Erwinia*, formed a clade. Red circles indicate proteins with empirical support for the recognition sequence GATC.

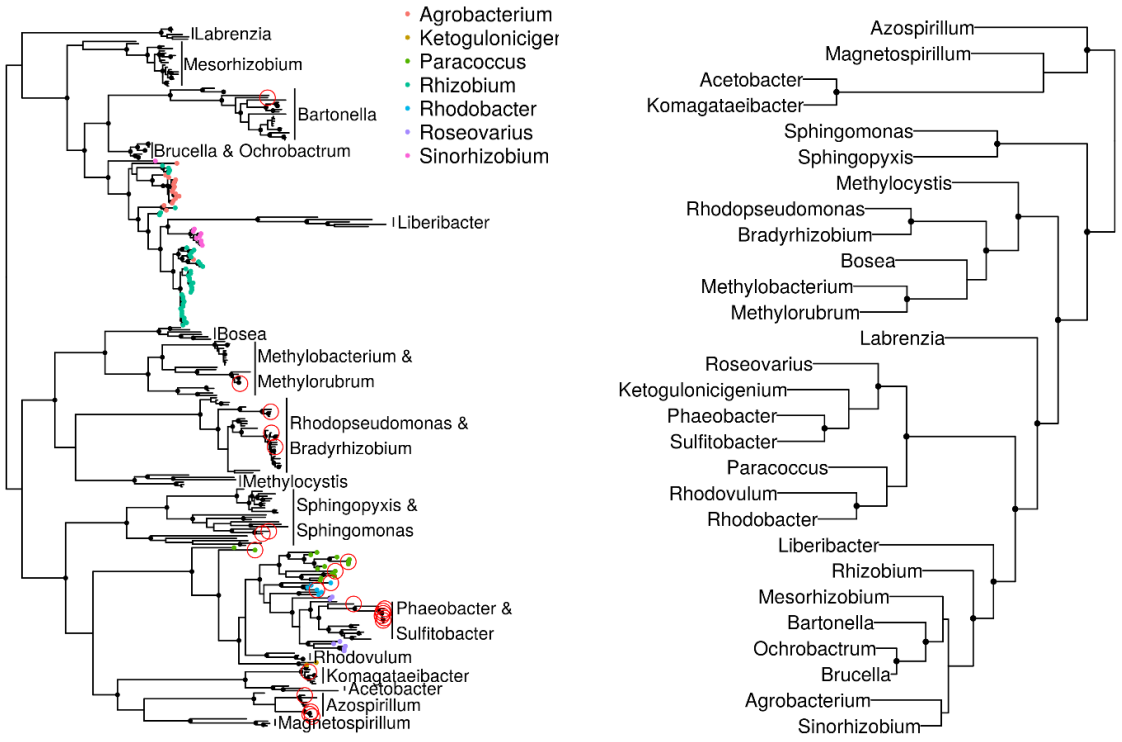

Figure 24. Phylogeny of Methyltransferase Cluster 2699 vs ribosomal 16S. At the left, the phylogeny of cluster 2699 from Figure 1 was annotated with the source genera. The right shows the r16S phylogeny from the genera from where methyltransferases were found. The internal nodes of each tree were annotated with red points if the support values were ≥ 0.80. Methyltransferase leaf nodes were annotated with colors if they were not considered monophyletic while monophyletic nodes were annotated with their genus source. Nodes were considered monophyletic if all methyltransferase from a single genus formed a clade or if two closely related genera, such as *Methylobacterium* and

(Figure 24 continued) *Methylorubrum* and *Brucella* or *Ochrobactrum*, formed a clade. Red circles indicate methyltransferases with empirical support for the recognition sequence GANTC.

Cluster 3712 is one of the most frequent clusters represented, specifically in *Gammaproteobacteria*, containing the characterized *dam* methyltransferase, targeting GATC, from *E. coli.* Throughout the proteobacterial tree of cluster 3712 we found additional examples of orthologs experimentally validated to target GATC (Figure 23), suggesting that this entire cluster shares the same recognition sequences and is indeed functionally identical. Moreover, this cluster follows 16S rRNA phylogeny to an impressive degree. For example, all methyltransferases from *Colwellia*, *Aeromonas*, and *Shewanella* form distinct clades for each genus respectively and cluster together similarly when compared to the corresponding 16S rRNA tree (Figure 23). Moreover, this tree topology is consistent among the family of *Enterobacteriaceae*, with *Shigella*, *Escherichia*, *Samonella* and *Kosakonia* separating from *Citrobacter*, *Enterobacter*, *Lecercia*, *Lelliottia*, *Klebsiella*, and *Raoultella*. We also found some methyltransferases mixing between closely related genera, such as *Pantoea* and *Erwinia* or *Shigella* and *Escherichia*, showing a limitation in the phylogenetic resolution of our protein trees.

Similar to cluster 3712 but restricted to *Alphaproteobacteria*, cluster 2699 has a large taxonomic distribution and, when present in a genus, was consistently found among all isolates of that genus (Figure 22). Throughout the phylogeny, we find empirically characterized proteins that target the recognition sequence GANTC, suggesting all proteins in this cluster likely target GANTC and are CcrM-like methyltransferases (Figure 24). Moreover, methyltransferase tree topology is like that of the 16S rRNA gene phylogeny, for both large and small trends. We also find larger taxonomic clades emerging belonging to the orders Rhizobiales, Rhodospirillales, and Rhodobacterales. Rhodobacterales, represented by the genera *Roseovarius*, *Kentogulonicigenium*,

Phaeobacter, *Sulfitobacter*, *Paracoccus*, *Rhodovulum* and *Rhodobacter* forms a distinct methyltransferase clade. Rhodospirillales, represented by *Azospirillium*, *Magnetospirillum*, *Komagataeibacter*, and Acetobacter form a distinct clade of methyltransferases, reflecting their 16S rRNA gene phylogeny. Two groups of Rhizobiales emerge in each clade, with strong support values for both methyltransferase and 16S rRNA gene trees. Distinct clades of *Methylobacterium* and *Methylorubrum* or *Rhodospedomonas* and *Bradyrhizobium* cluster together along with *Bosea* and *Methylocystis*, reflecting the tree topology of the 16S rRNA gene phylogeny. The other group of Rhizobiales, constituted of *Liberibacter*, *Mesorhizobium*, *Bartonella*, *Ochrobactrum*, *Brucella*, *Agrobacterium*, and *Sinorhizobium*, also maintains tree topology between the methyltransferase and 16S rRNA trees. We do note that the ability to differentiate between the methyltransferases of some closely related genera, such as *Brucella* and *Ochrobactrum* or *Sphingopyxis* and *Sphingomonas* is limited (Figure 24).

Methyltransferase cluster 1994 contains *dcm* methyltransferase, which targets CCWGG. We found several other empirically characterized methyltransferases in cluster 1994 that also target CCWGG, thus, other proteins within this cluster likely target the same recognition sequence (Figure 25). While highly conserved in some *Gammaproteobacterial* genera, isolates from others seem to have patterns of loss (Figure 22). Unlike *dam* or *ccrM*, this methyltransferase cluster does not from distinct clades of methyltransferases when considering larger taxonomic groups. However, when we subset

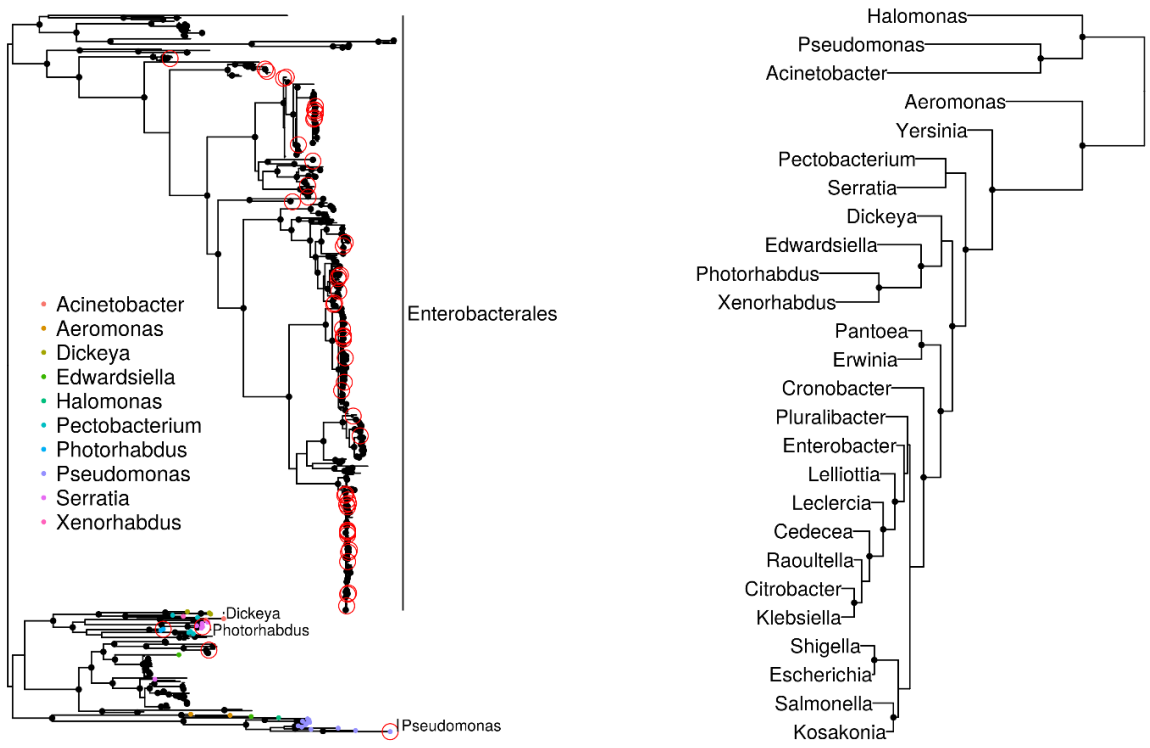

Figure 25. Methyltransferase Cluster 1994. Methyltransferases from Cluster 1,994 (right) compared to the 16S rRNA gene tree (left). Methyltransferase tree leaf nodes were annotated only if they were not sourced from an Enterobacterales genome. The internal nodes of each tree were annotated with black points if the support values were ≥ 0.80. Red circles indicate methyltransferases with empirical support for the recognition sequence CCWGG.

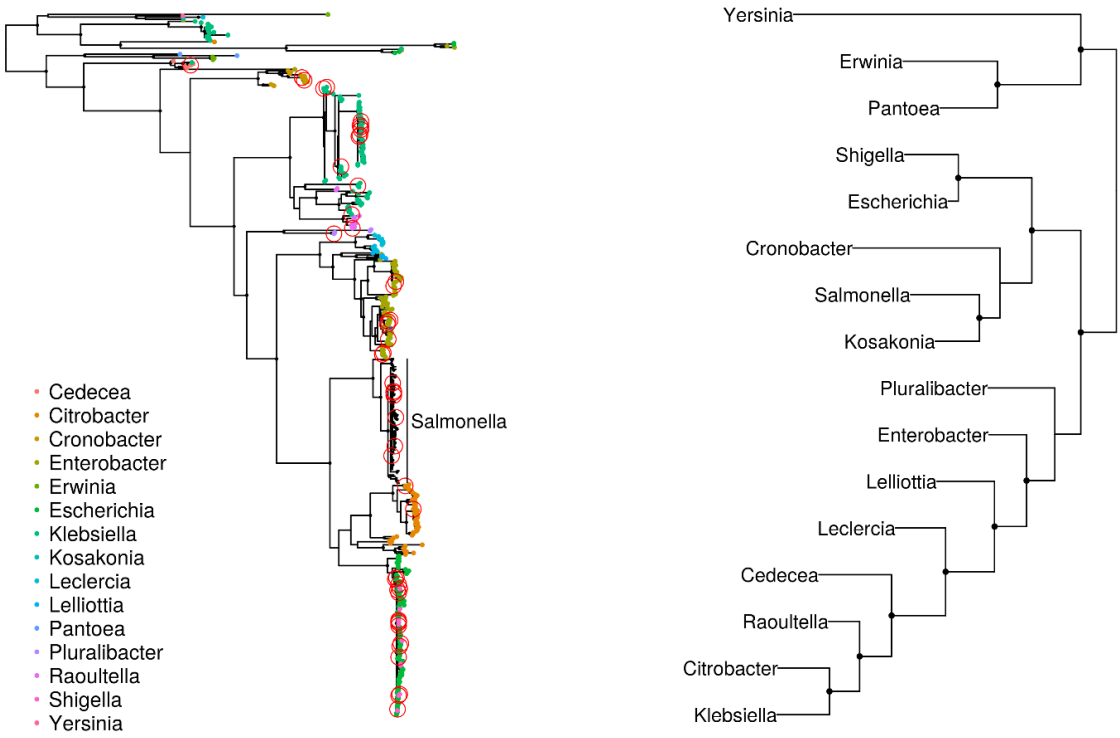

Figure 26. Enterobacterales containing Dcm Methyltransferases. At the left, the phylogeny, built with protein sequences, of a subset of cluster 1,994 from Figure 1 was annotated with the source genera. The right shows the r16S phylogeny from the genera from where methyltransferases were found. The internal nodes of each tree were annotated with black points if the support values were ≥ 0.80. Methyltransferase leaf nodes were annotated with colors if they were not considered monophyletic while monophyletic nodes were annotated with their genus source. Nodes were considered monophyletic if all methyltransferase from a single genus formed a clade or if two closely related genera.

(Figure 26 continued) Red circles indicate methyltransferases with empirical support for the recognition sequence CCWGG. See Figure S1 for the full cluster.

the methyltransferases to only include those from *Enterobacterales*, we can recapitulate 16S rRNA gene tree topologies (Figure 26). For example, closely related genera such as *Shigella* and *Escherichia* form a distinct clade of methyltransferases as seen for clusters 3712 or 2699.

Interestingly, cluster 3715 inversely correlated with the Dam cluster 3712, and after cross-referencing with empirically characterized enzymes, we found 3 members of 3715 target GATC (Figure 22). This trend of inverse correlation between clusters of the same recognition sequences was also true for cluster 2699, the *ccrM* cluster and cluster 2742 which contains 9 proteins with the confirmed recognition sequence of GANTC.

The remaining methyltransferase clusters detected in the Proteobacteria are sporadically distributed, inconsistently found in genera (Figure 22), and often include methyltransferases with known differences in recognition sequence. For example, cluster 1607, a type I methyltransferase cluster, has over 57 empirically tested methyltransferases all with different recognition sequences (Figure 22). However, we note a few interesting clusters. Clusters 3701 and 4299 are both type IIG clusters highly conserved in the family *Pasteurellaceae*.

A phylum-level scope of methyltransferases shows that they are sporadically distributed, with the exception of *dam* or *ccrM* like methyltransferases (Figure 22). We were curious if finer taxonomic resolution would reveal patterns of methyltransferase co-occurrence among isolates. To this end, we evaluated the methyltransferase distributions for genera belonging to the family *Enterobacteriaceae* and order *Campylobacterales*.

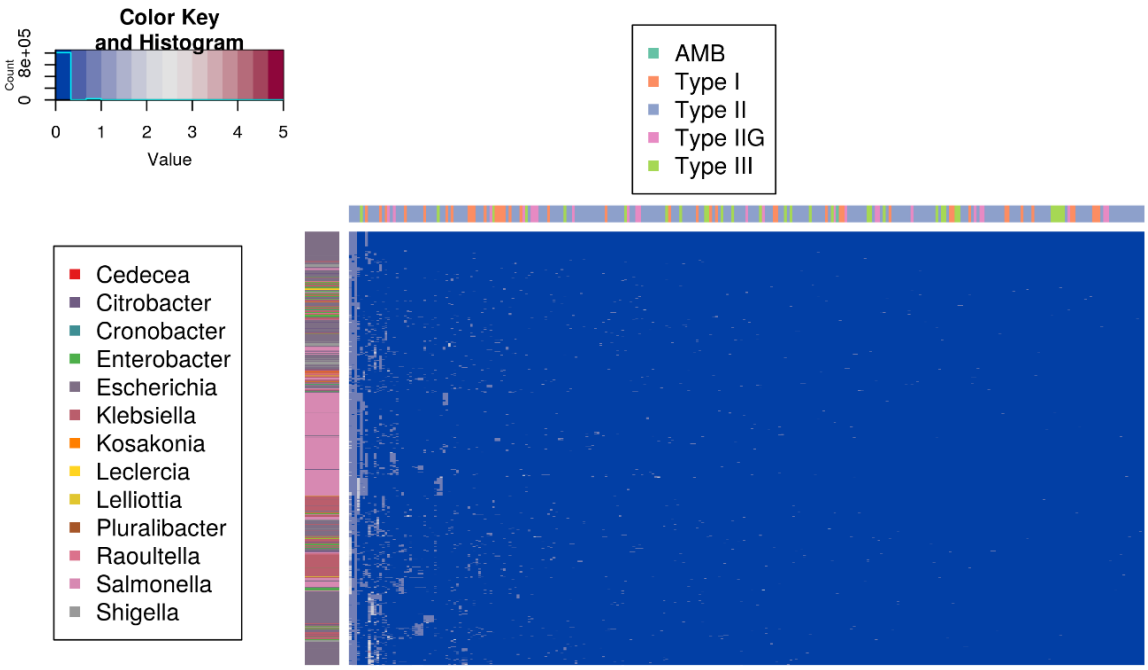

Figure 27. The distribution of Methyltransferase Clusters from the family Enterobacteriaceae. Each row is a single genome assembly with genus being identified at the left of the heatmap. 290 Columns represent methyltransferase clusters and are annotated with their type at the top of the heatmap. Methyltransferase clusters that are present in only one isolate (n=123) have been removed. Both rows and columns were arranged via hierarchical clustering. The blue line in the “Color Key and Histogram” shows the count distribution for the values in the heatmap.

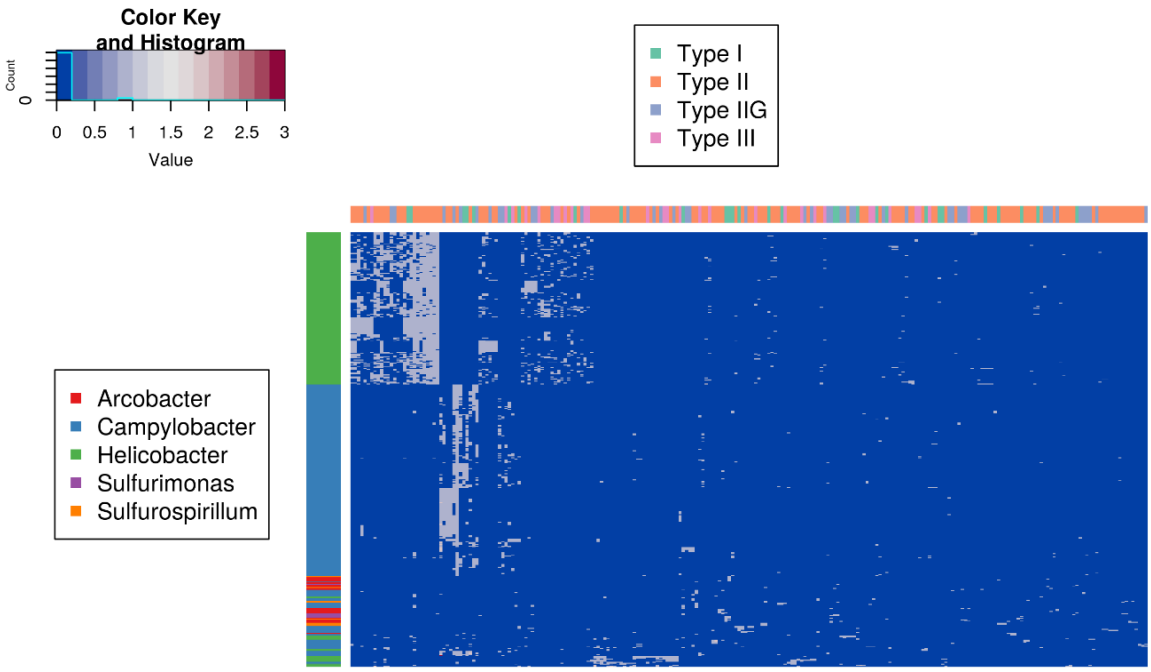

Figure 28. The distribution of Methyltransferase Clusters from the order Campylobacterales. Each row is a single assembly genome with genus being identified at the left of the heatmap. 244 columns represent methyltransferase clusters and are annotated with their type at the top of the heatmap. Methyltransferase clusters that are present in only one isolate (n=240) have been removed. The blue line in the "Color Key and Histogram" shows the count distribution for the values in the heatmap.

Campylobacterales was specifically chosen because it contains the genus *Helicobacter*, a Proteobacterium notorious for their high numbers of complete RMs. Within the *Enterobacteriaceae*, methyltransferases were sporadically distributed, with the exceptions of *dam* and *dcm* methyltransferase clusters (Figure 27). Moreover, we found the distribution pattern of methyltransferases to be a poor criterion for grouping organisms at their genus level. However, we found something very different in *Campylobacterales*: while no single methyltransferase cluster was present in all genus members, *Helicobacter* and *Campylobacter* appeared to have genus-specific patterns of methyltransferase cluster presence/absence that was shared by almost all genus members (Figure 28). Unfortunately, it is difficult to know if the clustering of *Helicobacter* genomes via methyltransferase clusters is due to the lack of closely related organisms with high numbers of RMs.

Methyltransferase distributions in Cyanobacteria. Our analyses in Proteobacteria have provided us with context for methyltransferases- those that are widely distributed throughout the phylum tend to be connected epigenetic phenomena/alternative functions, and organisms with high numbers of RMs can be grouped by the methyltransferase they contain alone. We expand our analysis to Cyanobacteria with interest in harmful bloom forming (HAB) cyanobacteria due to the prevalence of RMs found among these organisms.

Similar to Proteobacteria, while a few clusters have larger taxonomic scopes, the majority of the methyltransferase clusters are dispersed sporadically throughout the Cyanobacterial phylum (Figure 29). Closer phylogenetic investigations of

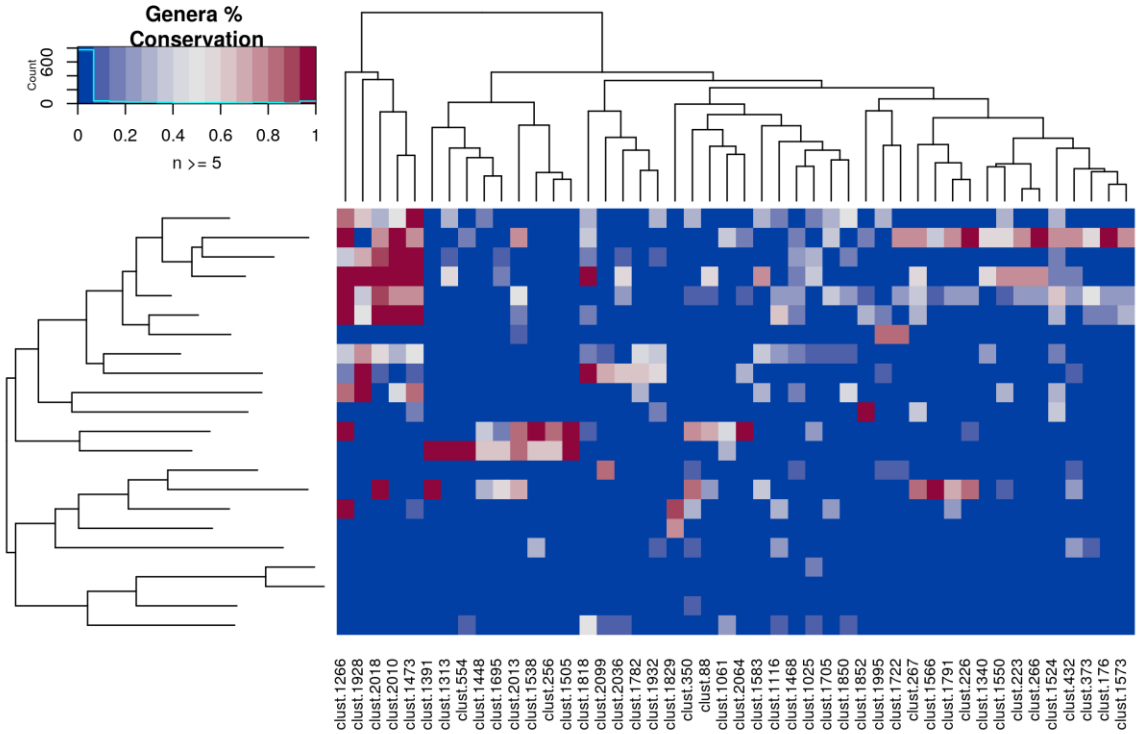

Figure 29. Phylogenetic Distribution of the Most Common Methyltransferase Clusters. A heatmap showing the presence/absence of methyltransferase clusters (columns) in different Cyanobacterial genera (rows). Columns were arranged by complete-linkage hierarchical clustering of presence/absence among genera, whereas rows were arranged by phylogenetic distance using the 16S rRNA gene of each genus. Colors in the heatmap show percent distribution of each methyltransferase cluster in each genus, with red indicating all isolates from that genus have a methyltransferase from the respective cluster. The blue line in the "Genera % Conservation" shows the count distribution for the values in the heatmap.

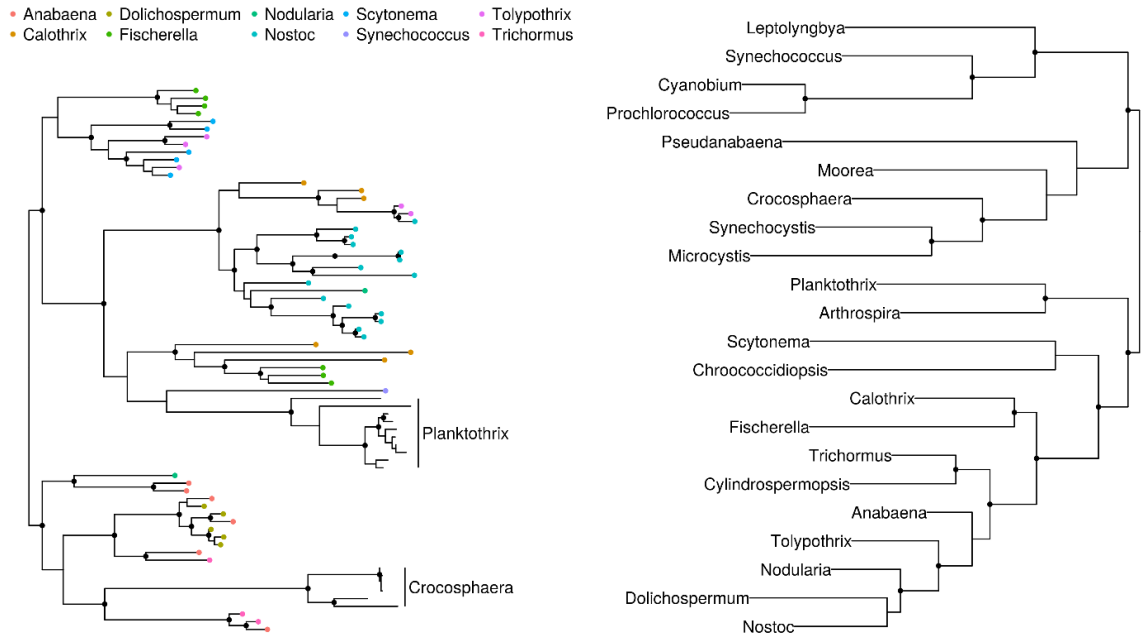

Figure 30. Cyanobacterial Methyltransferase Cluster 1266. At the left, the phylogeny of a subset of cluster 1266 from Figure 5 was annotated with the source genus. The right shows the r16S phylogeny from the genera from the Cyanobacteria phyla. The internal nodes of each tree were annotated with black points if the support values were ≥ 0.80. Methyltransferase leaf nodes were annotated with colors if they were not considered monophyletic while monophyletic nodes were annotated with their genus source. Nodes were considered monophyletic if all methyltransferase from a single genus formed a clade or if two closely related genera.

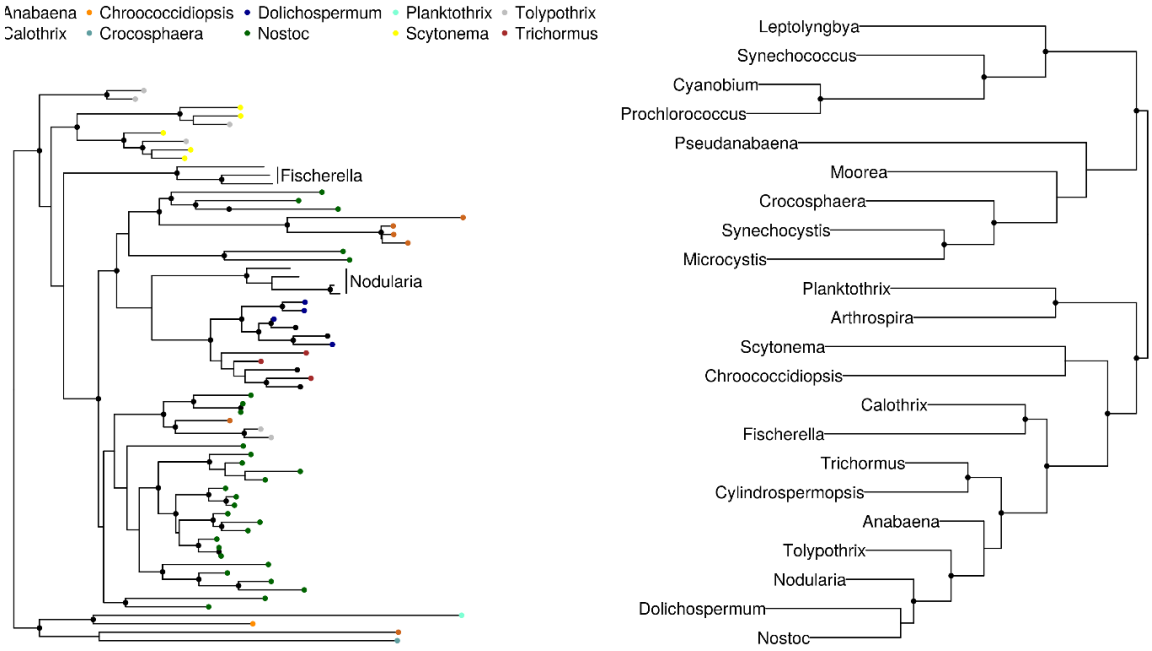

Figure 31. Cyanobacterial Methyltransferase Cluster 1473. At the left, the protein phylogeny of a subset of cluster 1,266 from Figure 5 was annotated with the source genus. The right shows the 16S rRNA gene phylogeny from the genera from the Cyanobacteria phyla. The internal nodes of each tree were annotated with black points if the support values were ≥ 0.80. Methyltransferase leaf nodes were annotated with colors if they were not considered monophyletic while monophyletic nodes were annotated with their genus source. Nodes were considered monophyletic if all methyltransferase from a single genus formed a clade or if two closely related genera.

methyltransferase clusters show very few instances of grouping by genus, suggesting these proteins are extremely polyphyletic. For example, neither cluster 1266 nor cluster 1473 show consistency between methyltransferase and 16S rRNA trees, with the noted exceptions of all *Planktothrix* and *Crocosphaera* homologs (Figure 30 and 31).

A closer look the isolate breakdown of the methyltransferase cluster distribution in Cyanobacteria show patterns similar to those observed in *Helicobacter*. Interestingly, we find the distribution of methyltransferase clusters among order *Oscillatoriophycidaea* is sufficient to cluster assemblies from the same genera, with the exception of *Planktothrix*, which forms two separate groups (Figure 32). In genera from order *Nostocales*, we find that the distribution of methyltransferases is sufficient to group assemblies from the genera *Nodularia*, *Cylindrospermopsis*, and *Fischerella* by hierarchical clustering. However, hierarchical clustering of methyltransferase distributions serves a poor criterion for other Cyanobacterial genera in this order (Figure 33).

The high degree to which methyltransferase distributions were able to distinguish genome assemblies of *Microcystis* from other *Oscillatoriophycidaea* was not expected, nor was the identification of 8 methyltransferase clusters found in each isolate of *Microcystis* (Figure 32). We investigated if there was a difference between the distribution of orphan methyltransferases and those that are components of RMs. We found that the likelihood of being part of RMs was the same for well-conserved and poorly-conserved methyltransferases in *Microcystis*: all were biased to being orphan methyltransferases (Figure 34). Moreover, no RMs were present in all isolates.

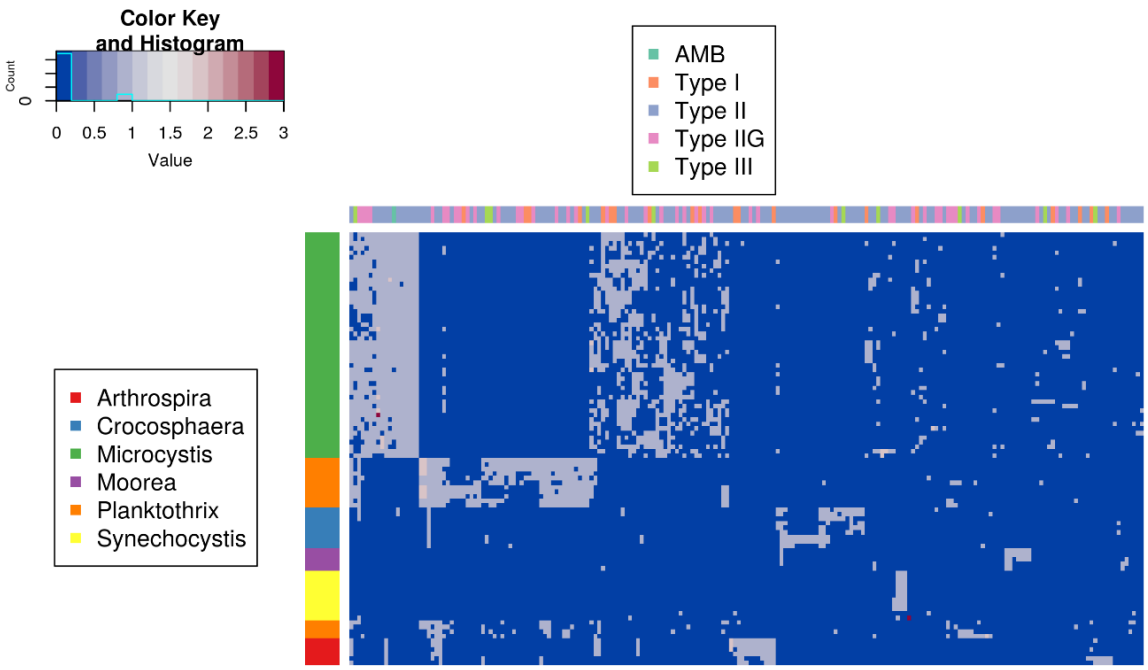

Figure 32. Methyltransferase cluster distributions in order Oscillatoriophycideae. Each row is a single assembly with genus being identified at the left of the heatmap. Columns represent methyltransferase clusters and are annotated with their type at the top of the heatmap. We note that *Synechocystis* was added to this analysis as its 16S rRNA gene alignments suggest it is a close relative of *Microcystis*. Methyltransferase clusters that are present in only one isolate (n=103) have been removed. The blue line in the "Color Key and Histogram" shows the count distribution for the values in the heatmap.

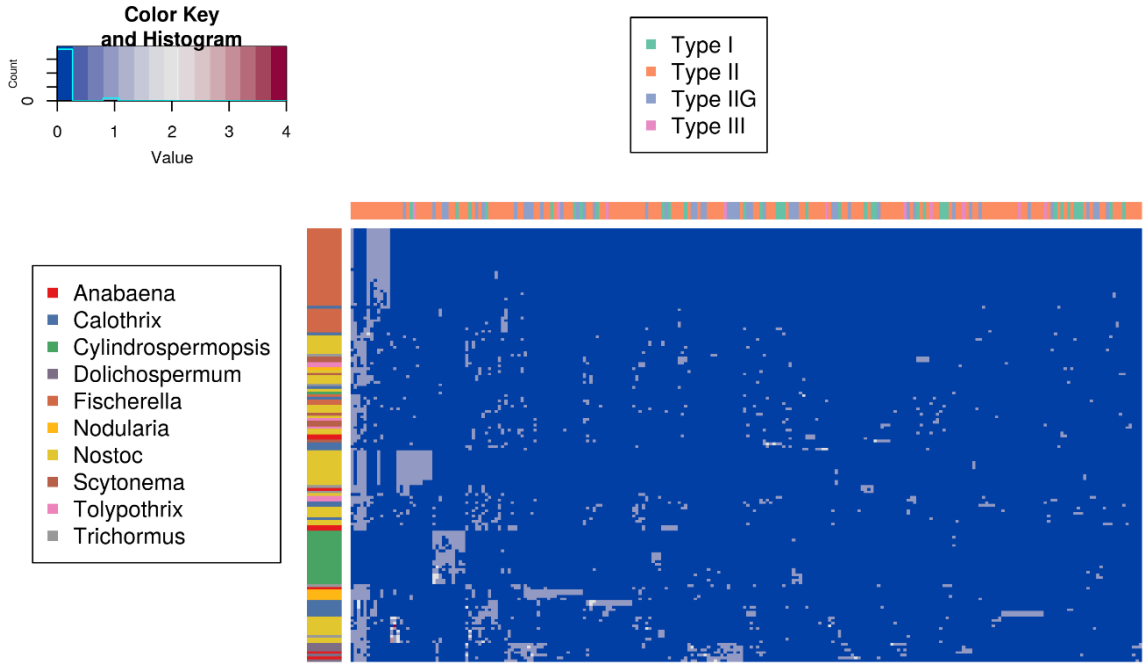

Figure 33. Methyltransferase cluster distributions in order Nostocales. Each row is a single assembly with genus being identified at the left of the heatmap. Columns represent methyltransferase clusters and are annotated with their type at the top of the heatmap. Methyltransferase clusters that are present in only one isolate (n=191) have been removed. The blue line in the "Color Key and Histogram" shows the count distribution for the values in the heatmap.

Interestingly, 8 methyltransferases clusters were conserved in all isolates of *Microcystis.* With one key exception, every conserved methyltransferase cluster included at least one isolate whose methyltransferase was suspected to be a component of an RM. The only methyltransferase cluster that is widely distributed and didn't not have any evidence of being part of a full RM system was predicted to target GATC. Previous SMRT sequencing of 5 *Microcystis* isolates found large variations in the sites methylated, with the exception of 5 methylated sites in each isolate: GAATTC, GATATC, GATC, GGCC, and RGATCY[208]. These genomes also showed greater numbers of orphan methyltransferases than those part of full restriction modification systems and authors suggest that these orphan methyltransferases are likely of some alternative importance to the cell[208]. We expand on these results, showing that clusters that putatively target GATATC, GGCC, RGATCY, and GATC are in all *Microcystis*. Overall, this data suggests that organisms with large numbers of RMs per genome exhibit robust patterns of methyltransferase conservation at the genus level.

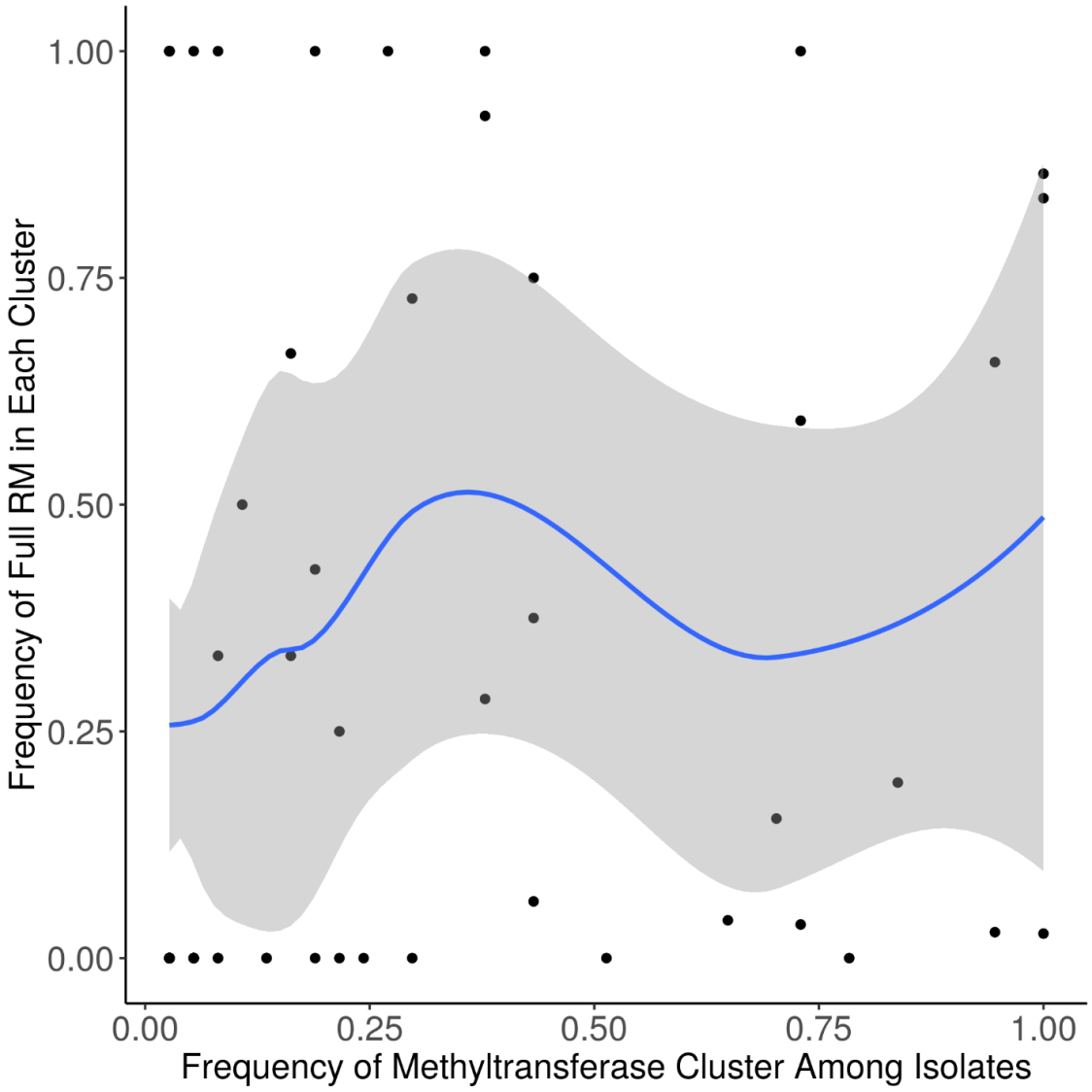

Figure 34. Frequency of full RMs as a function of phylogenetic breadth in *Microcystis*. A local polynomial regression (also known as loess) with span=0.75 was used to generate local fittings. The solid blue line represents the local mean, while the gray area is the local confidence interval. Datapoints represent individual clusters. n=37

III. Discussion

With the development of SMRT sequencing, we are able to measure methylation across microbial genomes[209]. However, we still understand little about the effects of this methylation beyond those that have been heavily empirically tested, such as *dam* or *ccrM*. In this investigation, we aimed to understand the distribution of methyltransferase clusters to discriminate those in genomic flux from being part of RMs and others being actively maintained though selection via alternative mechanisms.

Measuring methylation within a genome has led to a flurry of hypotheses in the regulatory implications of modification on host regulatory pathways. SMRT sequencing has been used in larger surveys of prokaryotes to show that methylation is extremely pervasive throughout the bacterial domain, with 92% having methylation[209]. Interestingly, that survey identified that 57% of orphan methyltransferases (i.e. methyltransferases without a cognate endonuclease) are conserved at the genus level, while only 9% of methyltransferases with a cognate endonuclease are conserved at the genus level. In another study looking at only type II RMs, nearly 24% of orphan methyltransferases analyzed across 559 genomes had evidence of degraded endonucleases[210]. These data suggest that orphan methyltransferases may have alternate functions and are being maintained via selection. With the advantage of larger databases, our study suggests that such interpretations should be used cautiously when evaluating methyltransferases in organisms with high numbers of complete RMs.

The functional requirements of type II RMs are separated between two proteins- the endonuclease and the methyltransferase. Loss of function mutations in this system will always be biased towards the endonucleases since active endonucleases, without

their cognate methyltransferase function, results in the host chromosome being susceptible to digestion, leading to cell death[211]. Therefore, unless the full RM system is lost together, orphan methyltransferases will always emerge under genomic loss. We believe that this is reflected in the methyltransferase distributions of organisms under selection to maintain large number of RMs, namely *Helicobacter* and *Microcystis* (Figures 6,8). Within each genus, we find patterns of methyltransferase loss to be sporadic, yet sequestered to their respective genera, allowing for hierarchical clustering to group these isolates together appropriately. Some of these methyltransferases even appear to be conserved at the genus level, as in the case of *Microcystis*. With the limitations of this study, it is not certain if the conservation of these methyltransferases is due to working in tandem with endonucleases, performing an alternative function (e.g. epigenetics) or, as reported for some RMs, doing both[212].

The confidence to assign functional role(s) for a methyltransferase depends on the lineages sampled. For example, *Gammaproteobacteria* are some of the most studied and best understood organisms in the bacterial domain of life, thus our ability to use phylogenetic context is maximal compared to other groups of organisms. Indeed, cyanobacteria are far less well sampled[213], therefore, identifying methyltransferases like *dam* or *ccrM* though phylogenomic methods may be impossible without more empirical evidence for methyltransferase functional roles in a cell or more sampling of closely related organisms. Fortunately, efforts have been made to improve the coverage of the phyla[214], however, the sampling still lags far behind that of other phyla.

Another potential weakness in our analysis is our combination of using genera annotations from NCBI. While we do not think this impacted our interpretations of larger

phylum-level analyses, the isolate-level analysis is susceptible to database errors. Unfortunately, this is especially true for Cyanobacteria as morphological characteristics were the historical criteria for naming conventions in Botanical Code[215]. These naming conventions add additional uncertainty in our hierarchical clustering when relying on proper genus level annotation from NCBI. For example, it is unclear if the separation of *Planktothrix* methyltransferases into two clusters (Figure 8) is due to true differences among methyltransferase distributions, or if the smaller of these two clusters belongs to *Arthrospira*. Another concern is *Synechocystis*, which was not annotated as *Oscillatoriophycidaea* even though the 16S rRNA trees suggest is very closely related to *Microcystis* (Figure S2). Indeed, strains named *Synechocystis* seem to have different phylogenies in higher resolution taxonomic studies[214] and suggests validating annotation is required. Therefore, our assembly level analysis will require expansion using phylogenetic markers for all isolates.

One interesting observation that we found in our analysis was the inverse correlation of different methyltransferase clusters with empirical evidence for the recognition sequence GATC. This suggests that the suite of methyltransferases encoded by a genome are not redundant in the recognition sequences they target, and it is possible to have methyltransferases of separate evolutionary origin performing the same functional role (i.e. convergent evolution). This, however, this suggests another problem when using phylogenetic context: phylogenetic signal of presence/absence may not coincide with presence/absence of function. This disconnect undermines our ability to assess the phylogenetic breadth of *dam* functional analogs, as they will be distributed between different clusters. We do, however, find that an identity threshold of 60%, when

applying alignment length criteria covering most of the query and subject sequences, is adequate in forming clusters with consistent predicted recognition sequences (Figures 2,3). The data, however, only supports this cutoff for *ccrM* and *dam* methyltransferase clusters, therefore caution should be used when only using protein alignments. As a matter of investigating functional redundancy, it would be interesting to evaluate if all methyltransferases targeting GATC are all anti-corollary (i.e. mutually-exclusive) with one another, and if the evolution of GATC methylation is a matter of convergent evolution.

Methylation is widespread throughout the prokaryotes, suggesting that it is a fundamental part of microbial life. Our results in Proteobacteria confirm previous notions of methylation: it is widespread, and those methyltransferases empirically shown to have essential roles in host global regulation are well maintained in a larger taxonomic ranks. The robustness of this criterion as a functional indicator, however, may not apply to organisms under selection for high numbers of RMs. Indeed, methyltransferase distribution in high RM organisms are very similar, suggesting that it may not be possible to make this differentiation without additional data.

IV. Methods

To avoid any bias associated with investigating the evolutionary history of methyltransferases, we took a clustering approach to a large dataset that contains well-characterized methyltransferases, but which only revealed clusters whose members had strong empirical support once phylogenetic trees were built (e.g. *dam* and *ccrM*). Our initial analysis was restricted to the proteomes of well characterized genera (5 or more isolates with complete genomes) from the phylum Proteobacteria, which were

downloaded with finditfasta (Chapter 1). To increase sampling for Cyanobacteria, we relaxed our complete genome criteria to include more genera; however, we still restricted the analysis to genera with 5 or more isolates.

Proteomes were searched for methyltransferase motifs using HMMER[216] with trusted cutoffs of the following pfams: DNA_methylase, Eco57I, EcoRI_methylase, HsdM_N, MethyltransfD12, MT-A70, N6_Mtase, N6_N4_Mtase, TypeIII_RM_meth, and Dam[217]. Because this analysis includes orphan methyltransferases, we took a machine learning approach to our methyltransferase type classification as opposed to our previous pipeline (Chapter 2). This ensured classification was consistent between methyltransferases part of full RMs and those that are orphans, therefore, endonucleases could not be used to assist in classification. Methyltransferases were classified using a random forest classifier available in the python module sklearn[218] that was trained on methyltransferases with empirical support from NEB's REBASE[109]. A combination of presence/absence of pfams and protein length was used to classify methyltransferase type with 95 ± 5 % accuracy from cross validation. Classification was ambiguous for a small subset (~0.1%) of these methyltransferases; however, when we applied a secondary round of classification using a decision tree classifier that was only trained on protein length, we are able to correctly classify over half (5/9) of the ambiguously labeled methyltransferases from the first round of classification, with others remaining ambiguous.

Protein clusters were formed in a neighbor joining fashion using cdhit [219]. Protein clusters were formed at 90% clustering identity and required at least 75% alignment length between the query and subject. The representative methyltransferases from each

cluster were then used for subsequent clustering at a lower identity threshold with the same alignment length requirements. This process was repeated for 80%, 70%, and 60% identities. A cluster was annotated as a type I, type II, or type III methyltransferase by checking the most frequent classifications of the proteins that constitute each cluster, while clusters composed of mostly ambiguously classified proteins were labeled as AMB.

The 16S rRNA gene from a representative genome, typically the highest quality, from each genus was used to build phylogenetic trees, while all methyltransferases belonging to a cluster at 60% identity were used for building phylogenetic trees. Clustal omega was used to generate multiple sequence alignments and columns were masked if over 30% of sequences contained gaps[220]. Trees were built using fasttree[221], newick files were annotated with ete3 in python[222], and then visualized in R using ggtree[223].

CHAPTER 5
A Trait Based Approach to Phylogenomics

ABSTRACT

Computational biology has become a rapidly emerging field within biology. As is hopefully evident from the works shown in previous chapters, we were able to tackle scientific questions that have remained elusive for many years, such as the selective forces that govern Restriction Modification systems in nature. In this conclusion chapter, I argue that trait-based modeling is an extremely effective tool in elucidating the selective forces of prokaryotic viral defense systems when other phylogenomic methods fail due to extreme variance generated by horizontal gene transfer. I discuss how our own trait-based models could be expanded upon to try and investigate the extreme variance in Restriction Modification genomic and domain architectures. Moreover, I try and apply the concepts learned from our memory model (Chapter 3) to the unexpected results of methyltransferase conservation in organisms with high numbers of complete Restriction Modification systems (Chapter 4). The framework of the memory model provides an alternative hypothesis to conservation in which orphan methyltransferases may help strains lacking the cognate endonuclease mitigate costs and still shed methylated viruses, effectively devaluing the endonucleases of their competitor. Lastly, due to computational biology being a relatively new field, I discuss some things I wish I knew in hindsight to have made my work during this dissertation easier. I emphasize that laboratory notebook standards should be applied in computational notebook and encourage all to use these resources as they are intended.

I. Introduction

The genomic revolution in biology has had profound effects on our ability to infer evolutionary relationships, and thus has greatly contributed to growing the tree of life. Through the work of pioneering molecular analysis by Carl Wose and George Fox in 1977, the use of the 16S ribosomal RNA (rRNA) gene sequence was used for the first quantitative phylogenetic study that showed that Archaea are a distinct domain in the tree of life[224]. Since then, we witnessed an unprecedented increase in the availability of molecular data (Chapter 2). As microbiologists, we have been able to leverage the wealth of this data to reconstruct the phylogenies of sequenced organisms in the context of genome-wide scale annotations, now commonly referred to as phylogenomics. Phylogenomics can be summarized in the following: to discover mechanisms of molecular evolution via phylogenetic context and the use of genomic data from multiple species to infer putative functions of DNAs and protein sequences[225]. Although genomes have been available for several years, the selective forces that govern the gain, loss, and diversity of Restriction Modification systems (RMs) had largely escaped our understanding[17]. We found that trait-based modeling was effective in relating molecular data of RMs with larger bioinformatic results and recapitulated pressures of Horizontal Gene Transfer (HGT) when incorporating viral methylation (Chapter 3). Moreover, we found that organisms under a high pressure to maintain RMs produce similar signals of methyltransferase distribution, suggesting the phylogenetic inference may be misleading when trying to infer the importance of conserved functional groups without properly controlling for alternative functions (Chapter 4). Unfortunately, integrating data from literature and databases with protein sequences is not trivial, as we have found in our own

investigations (Chapter 2). Through the development of code for data management (Chapter 2) and investigating RMs (Chapter 3 and 4), I offer some lessons learned to hopefully assist others in their own computational projects.

II. A Trait-based framework for RMs

A trait-based approach to mathematical modeling allowed us to integrate molecular observations with our own 'big data' bioinformatic signals (Chapter 3). Our models were simplistic yet effective in explaining the discrepancy between low counts of RMs in oligotrophic-dwelling Cyanobacteria and high counts of RMs in eutrophic-dwelling Cyanobacteria. It was not until I began writing the manuscript and ran additional numerical simulations, however, that I recognized the important the fact that having more than 3 RMs per genome, as many cyanobacteria do, would produce absurdly unrealistic resistances. Admittedly, if not for the modeling component, this detail would have largely gone unnoticed as we were prepared to move onto other investigations.

Our models forced us to revisit literature and develop alternative hypotheses to account for the escalation of RM defenses. Ironically, the key mechanistic detail absent from our model involved the observation that led to the discovery of RMs: that virions have a non-genetic change to the infection of hosts[226,227]. This explained not only how an organism could carry over 15 RMs in their genome, but also provided the mechanistic rationale that organisms are incentivized to diversify their recognition sequence as to not be driven to extinction by a n + 1 RM competitor (Chapter 3). We believe that the key to understanding the genomic and architectural diversity of RMs lies in the details of how cost and resistance is expressed in each RM type, and the trade-offs involved in each of their configurations. In the following sections, I discuss the sources of cost and

resistance, and how they can be used in future frameworks, as well as the structural trade-off that may make some RMs better at innovation than others.

Characterizing Costs and Resistance of RMs. Cells will suffer costs both in the DNA needed to maintain the genetic information from each RM as well as the RNA and amino acids needed for synthesis[228]. These costs, however, are not unique to a particular type of RM, nor any protein[228]. Therefore, assessing the total cost of carrying a single RM system will require understanding both endonuclease and methyltransferase activities and their impacts on the cell. The sources of cost, however, are likely varied due to different genomic and domain architecture of each RM system type. This can be plainly seen in their energy requirements for activity. For example, type I and type III RMs require ATP for hydrolysis, type IV require GTP, while type II and IIG are catalytically active without additional energy inputs[229]. In a modeling framework, this may suggest that the resistances of type I, III, and IV RMs may be a function of intracellular energy reserves, where starved cells are more susceptible to viral infection, whereas type II and IIG could operate independently of ATP/GTP reserves.

An important contributor to cost that is likely varied between RM types is the probability of autoimmunity, the accidental digestion of the host chromosome. Using a YPF reporter for the SOS-response, Pleska et al. were able to measure when cells suffered accidental chromosomal digestion due to RMs[211]. The SOS response is initialized when cells suffer increased DNA damage and can be costly to initiate[230,231], thus, the YFP reporter provided a real-time signal of accidental DNA digestion, and importantly, evidence that cells can survive this event. It was shown that for type II RMs, stochastic events can disrupt the stoichiometric balance between the methyltransferase

and endonuclease, resulting in unmethylated restriction site on the host chromosome and autoimmunity[211,232]. This likely explains why type II RMs are under tight regulatory control: to ensure that methyltransferase levels are always able to maintain host methylation. Most of these regulatory mechanism either inhibit endonuclease expression until methyltransferases have reached a threshold, or methyltransferase expression is reduced at steady state, and some systems even have additional controller proteins to ensure proper expression[233]. Although not known, it would be interesting to investigate if there is a difference in the probability of autoimmunity between different regulatory mechanisms of RMs and if there are trade-offs in different regulatory schemes, such as increased protection.

Importantly, autoimmunity resulting from stochiometric imbalance is either unlikely or not possible for some RM classes. For example, endonucleases are required to complex together with methyltransferase subunits to form a heterotetramer in type I RMs and a heterodimer in type III RMs to become active; thus, overproduction of the endonuclease subunit may not be as detrimental as overproduction of a type II endonuclease that does not require assembly of subunits to form an active enzyme. Moreover, type IIG RMs have both methyltransferase and endonuclease covalently linked in the same peptide, thus stochiometric imbalance is of no concern[234]. Indeed, hosts avoid type II restriction sites more often in their genomes than type IIG or type I, suggesting that the cost due to autoimmunity varies between types[211,235]. Evidence also suggests that the type of cut generated by the endonuclease may impact the cost of repair. Cells deficient in DNA repair mechanisms (*ΔrecA*) showed decrease fitness when carrying EcoRI, however EcoRV showed no additional fitness cost in the *ΔrecA*

mutant[211]. Authors hypothesized this difference is explained by how each endonuclease cuts DNA. EcoRV generates sticky end restriction sites and allows for complementary base pairing. In this situation, DNA ligase can repair the phosphodiester bond of the DNA backbone. EcoRV, on the other hand, produces blunt end cuts, therefore, the SOS response involving RecA is necessary to repair the damaged DNA. Therefore, sticky and blunt cutting could be another endonuclease trait in a modeling framework- sticky cutting enzymes may be less costly to cells during accidental restriction.

One complicating factor in autoimmunity is how nutrient quality affects the DNA repair. *E. coli* is more resistant to DNA damage caused by X rays, ultraviolet radiation, and methyl methanesulfonate when grown on yeast extract-nutrient broths than when grown on minimal media, suggesting the efficiency of RecA is nutrient dependent[236]. Additionally, the fitness cost of RMs increases in minimal media compared to rich media[211]. This may suggest that DNA repair is limited for organisms living in oligotrophic waters and may have enhanced costs for RMs. Indeed, many *Prochlorococcus* have lost DNA repair mechanisms altogether[237], thus the cost associated with RMs may be relatively higher for this genus as is similar to *ΔrecA E. coli*[211].

The costs associated with methyltransferases are largely unknown, but likely depend on the combination of genetic background and methylation site. As we discussed in Chapter 4, select methyltransferases are needed for global regulation. It is a poor assumption, however, that a methyltransferase cannot cause dysregulation. For example, HinfI is a type II system that was characterized from the Gammaproteobacterium *Haemophilus influenzae* and targets the recognition sequence GANTC[109]. We would hypothesize that this type II system would be incompatible with *C. crescentus* because

ccrM, the regulatory methyltransferase which targets GANTC, is temporally restricted in expression[205]. Therefore for this organism, GANTC would be "off-limits" to be used in RMs because the cost to global regulation would be too high- if inappropriately methylated, cells morphologically abnormal[205]. Likely, there are similar motifs in organisms that would have detrimental effects to fitness if methylated, and others likely have more intermediary effects on the cell.

Trade-offs in RMs. Typically associated with defense systems is the trade-off between cost and resistance[238], and for RMs, trade-offs exist between risk of autoimmunity and defensive efficacy. For type II RMs, those that are more efficient in defending host from infecting viruses also have a high rate of autoimmunity[211]. One possible source for this variation is the intrinsic stoichiometry in the expression between different type II RMs, where some may express slightly more endonuclease than methyltransferase, risking autoimmunity for increased protection. As described earlier, such a trade-off through altering stoichiometry is not possible for other types of RMs; however, our modeling results suggest that the rate of methylation may play a significant role in population level protection (Chapter 3) and may pose an unexplored trade-off.

Our modeling suggests that the efficiency of memory is a critically unmeasured parameter for RMs. In cases of high efficiency of memory, virions have a high likelihood of becoming modified from host methyltransferases, increasing infectiousness of the viral progeny via endonuclease evasion. In contrast, with low efficiency of memory, virions have a low likelihood of becoming modified and progeny virions are no more infectious than the parental virus. Because methylation is limited by the concentration of active methyltransferases in the cytosol[156], we propose that there is a trade-off between

efficiency of memory and autoimmunity. If there are too few methyltransferases expressed, the cell risks hypomethylation. However, increasing the number of methyltransferases to avoid the risk of hypomethylation may allow for more methylation to occur under active infection, increasing the efficiency of memory. Structural differences between RMs may pose different trade-offs with efficiency of memory.

Type IIG RMs have an interesting feature where they only hemi-methylate DNA, due to their non-palindromic recognition sequence, and require two specific unmodified sites to initiate endonuclease activity[234]. Requiring two unmodified sites is likely a mechanism to avoid autoimmunity because, if DNA is hemi-methylated, one round of replication will produce an unmodified restriction site. A compelling hypothesis is that hemi-methylation by type IIG establishes a significant trade-off between autoimmunity and efficiency of memory: hemi-methylation might lead to more autoimmunity, but because endonuclease and methyltransferase activities are colocalized to the same peptide, viral DNA is unlikely to be accidently methylated. This logic would extend to any RMs that recognizes non-palindromic DNA. Indeed, the unknown trade-offs between efficiency of memory, resistance, and cost may explain the diversity of RMs.

Recognition Sequence Variance of RMs. With high efficiency of memory, our models suggest that the identity of the recognition sequence is important in protecting hosts at a population level from viruses (Chapter 3). We found that increased diversity in recognition sequences fostered coexistence between multiple subpopulations, suggesting that viral memory is a driver in RMs recognition sequence variance. This highlights the advantage of not having methyltransferase and endonuclease activities separated. To successfully change the recognition sequence of a type II RM, two independent,

compatible mutations would need to occur in both the coding sequence of the endonuclease and methyltransferase, a very improbable scenario. In type IIG RMs, only a single mutation is required to alter the recognition sequence for both endonuclease and methyltransferase activity[234], allowing a population carrying type IIG RMs more potential to deviate their restriction site. Type I RMs may be even more likely to generate variation in their recognition sequence. In *Mycoplasma pulmonis*, researchers found repeat regions around multiple *hsdS* genes, responsible for DNA recognition in type I RMs, that facilitated homologous recombination, thus producing different recognition sequences[239]. Single-Molecule-Real-Time sequencing was able to precisely show how the recombination of the target recognition domains between different *hsdS* genes can generate new recognition sequences in very short timeframes, and is facilitated by the repeat regions in the *hsdS* gene[240]. Type III RMs also showed similar trends of mobility of the recognition domain between non-orthologous methyltransferase genes[241]. In summary, the architectures of the different RMs have large implications of recognition sequence variance, where type II RMs are poor at changing their target DNA, while others are able to deviate recognition sequences though point mutations. Moreover, it seems that type I and type III RMs have an accelerated rates of change as is evident in their propensity to engage in homologous recombination between target recognition domains.

Trait-Based Modeling when other Phylogenomic Approaches Fail. Hopefully, we have convinced readers that trait-based modeling is a useful tool in furthering our understanding of RMs. Phylogenomic inference alone had largely failed to explain the trends seen in RMs. This is likely due to two reasons. Foremost, the combination of RMs

being ubiquitous in Prokaryotes and their seeming random distributions failed to generate any meaningful inference.

As an example of useful inference, by comparing multiple strains of *Prochlorococcus*, we find there is strong genomic synteny between isolates from the same clades, however, there are a few locations in the genome that serve as rearrangement hotspots[242]. Moreover, these hotspots typically held genes associated with ecological adaptations, suggesting these genes were recently innovations in the evolutionary trajectory of the *Prochlorococcus* strains they were found in. In this example, genomic synteny to identified regions of large variance and, when considering phylogenetic context of different *Prochlorococcus* strains, show that genes in these areas were of recent evolutionary importance. This type of inference has been impossible for RMs - they are distributed across all Prokaryotes, and their variances within closely related taxa were generally vast (Chapter 3), thus, it was unclear what was selecting for increased numbers of RMs in certain strains.

In a more protein-centric example, phylogenomic inference revealed intermediate signaling systems that are the evolutionary link between two-component and chemotaxis systems, suggesting a larger sensory array in Prokaryotes and new targets for empirical testing[243]. Again, this type of investigation was difficult to apply to RMs - many RMs are extremely diverse. For example, it was impossible to investigate the evolutionary origins of endonucleases, and what ecological or physiological factors might drive their evolution, because many shared no homology to each other (Chapter 2 & 3).

Mathematical modeling of the traits of RMs, however, proved to me far more insightful (Chapter 3). There are several key observations that make defense from RMs a

viable trait: the functional mechanisms are well defined, we can quantify their frequency in microbial genomes, we can standardize their protective value via biochemical assays, and they are species independent though horizontal gene transfer[244]. Our success in using trait-based modeling for understanding the selective pressures of RMs may severe as a template to be generalized to other defense systems that are horizontally transferred (Chapter 1).

III. A Cautionary Tale of Data Interpretation

The previous chapter in this dissertation aimed to explore the distributions of methyltransferases and investigate if phylogenetic distribution alone could be a useful indicator in determining its functional role. We found the unsurprising results that *dam* and *ccrM* methyltransferases have larger phylogenic distributions and largely follow 16S rRNA phylogeny, indicative of vertical transfer (Chapter 4). This result confirmed and extended previous reports[197,198] that methyltransferases needed for global regulation were deeply tied to the phylogeny of the organism (Chapter 4). In contrast, the methyltransferases of RMs have been characterized as extremely varied in phylogenetic distribution[17]. While this trend largely held true, we found that increased selection for RMs defied this conventional wisdom.

To our surprise, genera with a high mean of RMs per genome, such as *Helicobacter* and *Microcystis*, showed remarkably consistent distributions of methyltransferases within the genus. Some have suggested that these methyltransferases may be indicative of regulatory importance in *Microcystis*[208]. In light of the reproducible methyltransferase distributions we find in other high RM carrying organisms, we argue

that this signal of phylogenetic breadth may be an artifact of preferential gene loss in type II, and possibly viral memory.

As discussed previously, type II RMs are largely biased towards loss of the endonuclease (Chapter 4). Therefore, the presence of a methyltransferases without an endonuclease may be intermediates in a streamlining path of RM decay. Indeed, these 'conserved' methyltransferases were part of a full RM system in at least one isolate in *Microcystis* (Chapter 4). One might suspect that the orphaned methyltransferases are subsequently lost by drift or selection. However, given the population level effects of viral methylation (Chapter 3), there is a possibility that hosts are selected to retain the methyltransferase of degraded RMs because it decreases the protective value of their competitors' RMs. For example, let us imagine the scenario of two competing host subpopulations in the presence of bacteriophage. In this scenario, one population has a single RM system, while the other only carries the methyltransferase component, thus all phage are modified. To demonstrate the predicted outcomes from our model, recall how resistance is calculated:

$$r_{ij} = r^{|RM_i - RM_j|}\, m^{|RM_i \cap RM_j|}$$

Where *r* is the resistance conferred by the endonuclease and *m* is the efficiency of memory, where we assume *r*<*m* and lower values lead to better protection from viruses. Also, recall RM_i and RM_j denote the RM carried the host and the methylation state of the virus. In our scenario, the population with only the methyltransferase has a resistance of 1, or in other words, no resistance. However, because this subpopulation still carries the methyltransferase, all phage retain the modified state. Therefore, the subpopulation with the endonuclease will always have a resistance equal to the efficiency of memory, or *m*.

In contrast, if a population has a full RM system and the other has lost the methyltransferase, there will be the production of unmodified phage. In this situation, the population carrying the full RM system will have a resistance of *r* to the unmodified phage and a resistance of *m* to the modified phage, increasing the overall protection of this population and their biomass. Thus, if a methyltransferase is not too costly, it is better to retain it if the competitor has the full RM version, because it devalues the resistance of that RM system by shedding methylated virions.

If true, this hypothesis does not suggest there is a lack of epigenetic gene regulation in high RM organisms; these hypotheses are not mutually exclusive. Rather, this provides reasons for caution in the interpretation of why methyltransferases may appear to be conserved in some organisms. As with most phylogenomic investigations, including a large taxonomic breadth is generally recommended as it improves the quality of the anlaysis[225]. In the case of Cyanobacteria, however, we do not see conserved methyltransferases shared with closely related genera that have few RMs (Chapter 4, Figure 8). Indeed, methyltransferases can serve many functional roles in Prokaryotes including RMs, epigenetic gene regulation, and even BREX[17,57,197,198,209], making their role in the cell far more difficult to predict. Indeed, some methyltransferases have been observed to have multiple physiological roles, making the selective forces even more difficult to elucidate[212].

IV. Final Thoughts: Hindsight and Personal Lessons Learned in Computational Biology

One of the primary challenges of this dissertation was learning the best practices of how to do computational biology. It has only been 13 years since the emergence of

next generation sequencing and the landscape of computational resources has rapidly been changing over that time (chapter 1). However, this rapidly changing landscape can result in confusion about the best practice to tackling a computational project. In hindsight, there are a few changes that would have greatly accelerated the progress of this dissertation from day one. Here, I will discuss these hard-learned lessons during the development of this dissertation and argue some changes early on in day to day work will greatly accelerate not only productivity, but possibly build a framework to easily learn more computational skills.

After 5 years of playing with different methods of documenting the doctoral work within this dissertation, I cannot understate the importance of computational notebooks[245]. Notebook files, along with their source code, are analogous to laboratory scientist's notebook and should have a large amount of time to their curation. Composed of individual executable cells, they can store code, images, text and hyperlinks in one cohesive document recording exactly how data was generated[245]. Intuitively, these files can be structured as an experimentalist's notebook would be. A table of contents can be listed at the beginning of each notebook file, with hyperlinks for easy navigation to a different section. Each section should have some sort of summary of the purpose, ideally including a hypothesis of what is being tested. The code imbedded in the notebook should be like the ideal protocol in a laboratory notebook: easy to read, well detailed, and explains different steps. Finally, each project should have some sort of conclusions sections, discussing the data and possible next steps. Admittedly, the initial beginnings of this dissertation were far more unorganized: a folder labeled 'bioinformatic_scripts', a folder for data, and a text file with steps of the sequential order of scripts to run and what

they do. This makes the projects unmanageable as results are accumulated and will make it extremely difficult for colleagues to follow.

It is extremely common to come across pipelines in bioinformatic work[246]. Pipelines are a series of functions/binaries that automatically process data, and while pipelines can certainly be helpful, they often make data exploration at intermediary steps difficult. Notebooks foster a more modular approach as the execution of code in cells makes it easier to break processes apart into logical steps. This will help you notice when you are applying the same logical steps repeatedly, making it easier to identify when you should invest your time into developing a generalized function for future use. While this may be slow at first, over time you will begin to build a portfolio of well-maintained, well-documented, and tested functions that will increase tractability in data analysis of future projects. When only using pipelines, it is easy to become detached from the underlying methodologies/logic, and if problems require a novel workflow, it will be harder to solve. As an analogy, the use of pipelines is like using kits at a laboratory bench, which can greatly accelerate routine work. Much like the implementations of algorithms outsourced to developers of computational packages, laboratory kits outsource the chemistry to the company scientists that design the kits for a specific laboratory procedure, such as DNA extraction. While laboratory kits undoubtedly save time in the lab, they detach researchers from the important details that makes the chemistry possible. This may never become a problem for some, but disassociation of the protocol from the underlying science will make developing new protocols, or perhaps more importantly, troubleshooting kits when they do not work extremely challenging. This can be extended to computation as dissociation from the underlying algorithms can make future analysis

of novel problems far more difficult. Insidiously, detecting problems in computation can require extra effort- barring errors, computers always will provide an answer, but it is the responsibility of the user to investigate if the validity of results.

The final, and possibly most important lesson is that of statistical vs biological significance. In our work, we were able to show a statistical significance between genome size and the number of RMs per genome as other have (Chapter 3)[17]. While we go on to show that this relationship does little to explain the variance, it does demonstrate a critical weakness when working with big datasets: when applying standard analytical tools, statistically significant signals can appear when they may not be biologically meaningful[247]. Moreover, is it always appropriate to have a statistical null of zero for the slope? Are there any genomic elements that do not scale with genome size? If not, what constitutes a relationship outside the norm? The analyses required to answer these questions are examples of possible controls and are critical when investigating any hypothesis. For anyone starting in computational biology, it is important to realize that the computer will always provide an answer, but it is up to the researcher to thoroughly test the validity of the results and provide the necessary controls to convince the reader that the signals contribute to the expansion of knowledge in biology.

References

1. Carroll, L., 1832-1898. *Through the looking-glass and what Alice found there*. Chicago : W.B. Conkey Co. (1900).
2. Valen, L. V. A New Evoluitonary Law. *New Evoluitonary Law* (1973).
3. Liow, L. H., Van Valen, L. & Stenseth, N. C. Red Queen: from populations to taxa and communities. *Trends Ecol. Evol.* **26**, 349–358 (2011).
4. Strotz, L. C. *et al.* Getting somewhere with the Red Queen: chasing a biologically modern definition of the hypothesis. *Biol. Lett.* **14**, 20170734 (2018).
5. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).
6. Wilhelm, S. W. & Matteson, A. R. Freshwater and marine virioplankton: a brief overview of commonalities and differences. *Freshw. Biol.* **53**, 1076–1089 (2008).
7. Braun, V. & Hantke, K. Bacterial Receptors for Phages and Colicins as Constituents of Specific Transport Systems. in *Microbial Interactions* (ed. Reissig, J. L.) 99–137 (Springer US, 1977). doi:10.1007/978-1-4615-9698-1_3.
8. Rakhuba, D. V., Kolomiets, E. I., Dey, E. S. & Novik, G. I. Bacteriophage Receptors, Mechanisms of Phage Adsorption and Penetration into Host Cell. *Pol. J. Microbiol.* **59**, 145–155 (2010).
9. Bertozzi Silva, J., Storms, Z. & Sauvageau, D. Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**, (2016).
10. Cole, S. T., Chen-Schmeisser, U., Hindennach, I. & Henning, U. Apparent bacteriophage-binding region of an *Escherichia coli* K-12 outer membrane protein. *J. Bacteriol.* **153**, 581–587 (1983).

11. Le, S. *et al.* Mapping the Tail Fiber as the Receptor Binding Protein Responsible for Differential Host Specificity of *Pseudomonas aeruginosa* Bacteriophages PaP1 and JG004. *PLOS ONE* **8**, e68562 (2013).
12. Nordström, K. & Forsgren, A. Effect of Protein A on Adsorption of Bacteriophages to *Staphylococcus aureus*. *J. Virol.* **14**, 198–202 (1974).
13. Riede, I. & Eschbach, M. L. Evidence that TraT interacts with OmpA of *Escherichia coli*. *FEBS Lett.* **205**, 241–245 (1986).
14. Pedruzzi, I., Rosenbusch, J. P. & Locher, K. P. Inactivation in vitro of the *Escherichia coli* outer membrane protein FhuA by a phage T5-encoded lipoprotein. *FEMS Microbiol. Lett.* **168**, 119–125 (1998).
15. Sutherland, I. W., Hughes, K. A., Skillman, L. C. & Tait, K. The interaction of phage and biofilms. *FEMS Microbiol. Lett.* **232**, 1–6 (2004).
16. Sutherland, I. W. Polysaccharases for microbial exopolysaccharides. *Carbohydr. Polym.* **38**, 319–328 (1999).
17. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol. Mol. Biol. Rev. MMBR* **77**, 53–72 (2013).
18. Pingoud, A., Fuxreiter, M., Pingoud, V. & Wende, W. Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.* **62**, 685–707 (2005).
19. Cheng, X. & Roberts, R. J. AdoMet-dependent methylation, DNA methyltransferases and base flipping. *Nucleic Acids Res.* **29**, 3784–3795 (2001).
20. Murray, N. E. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev. MMBR* **64**, 412–434 (2000).

21. Dryden, D. T. F., Murray, N. E. & Rao, D. N. Nucleoside triphosphate-dependent restriction enzymes. *Nucleic Acids Res.* **29**, 3728–3741 (2001).

22. Characterization and expression of the Escherichia coli Mrr restriction system. https://www-ncbi-nlm-nih-gov.proxy.lib.utk.edu/pmc/articles/PMC208215/.

23. Roberts, R. J. *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805–1812 (2003).

24. Elhai, J., Vepritskiy, A., Muro-Pastor, A. M., Flores, E. & Wolk, C. P. Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena sp*. strain PCC 7120. *J. Bacteriol.* **179**, 1998–2005 (1997).

25. Trieu-Cuot, P., Carlier, C., Poyart-Salmeron, C. & Courvalin, P. Shuttle vectors containing a multiple cloning site and a lacZ alpha gene for conjugal transfer of DNA from *Escherichia coli* to gram-positive bacteria. *Gene* **102**, 99–104 (1991).

26. Guiney, D. G. Promiscuous Transfer of Drug Resistance in Gram-Negative Bacteria. *J. Infect. Dis.* **149**, 320–329 (1984).

27. Korona, R., Korona, B. & Levin, B. R. Sensitivity of naturally occurring coliphages to type I and type II restriction and modification. *J. Gen. Microbiol.* **139 Pt 6**, 1283–1290 (1993).

28. Krüger, D. H., Barcak, G. J. & Smith, H. O. Abolition of DNA recognition site resistance to the restriction endonuclease EcoRII. *Biomed. Biochim. Acta* **47**, K1-5 (1988).

29. Krüger, D. H., Hansen, S. & Schroeder, C. Host-dependent modification of bacterial virus T3 affecting its adsorption ability. *Virology* **102**, 444–446 (1980).

30. Krüger, D. H., Hansen, S. & Schroeder, C. Virus adaptation to host cells: The non-classical modification of phage T3. *Z. Für Allg. Mikrobiol.* **20**, 495–502 (1980).

31. McGrath, S., Seegers, J. F. M. L., Fitzgerald, G. F. & van Sinderen, D. Molecular Characterization of a Phage-Encoded Resistance System in *Lactococcus lactis*. *Appl. Environ. Microbiol.* **65**, 1891–1899 (1999).

32. Toothman, P. Restriction alleviation by bacteriophages lambda and lambda reverse. *J. Virol.* **38**, 621–631 (1981).

33. Tock, M. R. & Dryden, D. T. The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* **8**, 466–472 (2005).

34. Bair, C. L. & Black, L. W. A Type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. *J. Mol. Biol.* **366**, 768–778 (2007).

35. Bair, C. & Black, L. W. Exclusion of Glucosyl-Hydroxymethylcytosine DNA Containing Bacteriophages. *J. Mol. Biol.* **366**, 779–789 (2007).

36. Iida, S., Streiff, M. B., Bickle, T. A. & Arber, W. Two DNA antirestriction systems of bacteriophage P1, darA, and darB: characterization of darA− phages. *Virology* **157**, 156–166 (1987).

37. Walkinshaw, M. D. *et al.* Structure of Ocr from Bacteriophage T7, a Protein that Mimics B-Form DNA. *Mol. Cell* **9**, 187–194 (2002).

38. Studier, F. W. Analysis of bacteriophage T7 early RNAs and proteins on slab gels. *J. Mol. Biol.* **79**, 237–248 (1973).

39. Patrick, S., Houston, S., Thacker, Z. & Blakely, G. W. Mutational analysis of genes implicated in LPS and capsular polysaccharide biosynthesis in the opportunistic pathogen *Bacteroides fragilis*. *Microbiol. Read. Engl.* **155**, 1039–1049 (2009).

40. Wang, L. *et al.* Phosphorothioation of DNA in bacteria by dnd genes. *Nat. Chem. Biol.* **3**, 709–710 (2007).

41. Xu, T., Yao, F., Zhou, X., Deng, Z. & You, D. A novel host-specific restriction system associated with DNA backbone S-modification in *Salmonella*. *Nucleic Acids Res.* **38**, 7133–7141 (2010).

42. Wang, L. *et al.* DNA phosphorothioation is widespread and quantized in bacterial genomes. *Proc. Natl. Acad. Sci.* **108**, 2963–2968 (2011).

43. Giovannoni, S. J. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annu. Rev. Mar. Sci.* **9**, 231–255 (2017).

44. Rappé, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).

45. Zhao, Y. *et al.* Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360 (2013).

46. Jansen, R., Embden, J. D. A. van, Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.* **43**, 1565–1575 (2002).

47. Deveau, H., Garneau, J. E. & Moineau, S. CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. *Annu. Rev. Microbiol.* **64**, 475–493 (2010).

48. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

49. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology,* **155**, 733–740 (2009).

50. Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709–1712 (2007).

51. Chopin, M.-C., Chopin, A. & Bidnenko, E. Phage abortive infection in *lactococci*: variations on a theme. *Curr. Opin. Microbiol.* **8**, 473–479 (2005).

52. Kjos, M., Snipen, L., Salehian, Z., Nes, I. F. & Diep, D. B. The Abi Proteins and Their Involvement in Bacteriocin Self-Immunity. *J. Bacteriol.* **192**, 2068–2076 (2010).

53. Boucher, I., Émond, É., Dion, É., Montpetit, D. & Moineau, S. Microbiological and molecular impacts of AbiK on the lytic cycle of *Lactococcus lactis* phages of the 936 and P335 species. *Microbiol. Read. Engl.* **146 (Pt 2)**, 445–453 (2000).

54. Bouchard, J. D. & Moineau, S. *Lactococcal* Phage Genes Involved in Sensitivity to AbiK and Their Relation to Single-Strand Annealing Proteins. *J. Bacteriol.* **186**, 3649–3652 (2004).

55. Exclusion of T4 phage by the hok/sok killer locus from plasmid R1. https://www-ncbi-nlm-nih-gov.proxy.lib.utk.edu/pmc/articles/PMC177903/.

56. Hazan, R. & Engelberg-Kulka, H. *Escherichia coli* mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1. *Mol. Genet. Genomics* **272**, 227–234 (2004).

57. Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).

58. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).

59. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J.-P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).

60. Norman, A., Hansen, L. H. & Sørensen, S. J. Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 2275–2289 (2009).

61. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).

62. McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D. & Boyd, E. F. CRISPR-Cas systems are present predominantly on mobile genetic elements in Vibrio species. *BMC Genomics* **20**, 105 (2019).

63. Varble, A., Meaden, S., Barrangou, R., Westra, E. R. & Marraffini, L. A. Recombination between phages and CRISPR-cas loci facilitates horizontal gene transfer in *staphylococci*. *Nat. Microbiol.* **4**, 956–963 (2019).

64. Gophna, U. *et al.* No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. *ISME J.* **9**, 2021–2027 (2015).

65. Ho, W. S., Ou, H.-Y., Yeo, C. C. & Thong, K. L. The dnd operon for DNA phosphorothioation modification system in *Escherichia coli* is located in diverse genomic islands. *BMC Genomics* **16**, 199 (2015).

66. He, X. *et al.* Analysis of a genomic island housing genes for DNA S-modification system in *Streptomyces lividans* 66 and its counterparts in other distantly related bacteria. *Mol. Microbiol.* **65**, 1034–1048 (2007).

67. Viral Evasion of a Bacterial Suicide System by RNA–Based Molecular Mimicry Enables Infectious Altruism. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003023.

68. Blower, T. R., Short, F. L., Fineran, P. C. & Salmond, G. P. C. Viral molecular mimicry circumvents abortive infection and suppresses bacterial suicide to make hosts permissive for replication. *Bacteriophage* **2**, 234–238 (2012).

69. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).

70. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **193**, 6039–6056 (2011).

71. Juhas, M. *et al.* Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol. Rev.* **33**, 376–393 (2009).

72. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring Horizontal Gene Transfer. *PLOS Comput. Biol.* **11**, e1004095 (2015).

73. Betlach, M. *et al.* A restriction endonuclease analysis of the bacterial plasmid controlling the ecoRI restriction and modification of DNA. *Fed. Proc.* **35**, 2037–2043 (1976).

74. Kulakauskas, S., Lubys, A. & Ehrlich, S. D. DNA restriction-modification systems mediate plasmid maintenance. *J. Bacteriol.* **177**, 3451–3454 (1995).

75. Kita, K., Kawakami, H. & Tanaka, H. Evidence for horizontal transfer of the EcoT38I restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in *Escherichia coli* TH38 strains. *J. Bacteriol.* **185**, 2296–2305 (2003).

76. Takahashi, N., Ohashi, S., Sadykov, M. R., Mizutani-Ui, Y. & Kobayashi, I. IS-linked movement of a restriction-modification system. *PloS One* **6**, e16554–e16554 (2011).

77. Burrus, V., Bontemps, C., Decaris, B. & Guédon, G. Characterization of a novel type II restriction-modification system, Sth368I, encoded by the integrative element ICESt1 of *Streptococcus thermophilus* CNRZ368. *Appl. Environ. Microbiol.* **67**, 1522–1528 (2001).

78. Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N. & Uchiyama, I. Shaping the genome – restriction–modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.* **9**, 649–656 (1999).

79. Rowe-Magnus, D. A. *et al.* The evolutionary history of chromosomal super-integrons provides an ancestry for multiresistant integrons. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 652–657 (2001).

80. Kobayashi, I. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29**, 3742–3756 (2001).

81. Werren, J. H. Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl. Acad. Sci.* **108**, 10863–10870 (2011).

82. Naito, T., Kusano, K. & Kobayashi, I. Selfish behavior of restriction-modification systems. *Science* **267**, 897–899 (1995).

83. Kulakauskas, S., Lubys, A. & Ehrlich, S. D. DNA restriction-modification systems mediate plasmid maintenance. *J. Bacteriol.* **177**, 3451–3454 (1995).

84. Tsang, J. Bacterial plasmid addiction systems and their implications for antibiotic drug development. *Postdoc J. J. Postdr. Res. Postdr. Aff.* **5**, 3–9 (2017).

85. Dupuis, M.-È., Villion, M., Magadán, A. H. & Moineau, S. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. *Nat. Commun.* **4**, (2013).

86. Ruess, J., Pleška, M., Guet, C. C. & Tkačik, G. Molecular noise of innate immunity shapes bacteria-phage ecologies. *PLOS Comput. Biol.* **15**, e1007168 (2019).

87. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687–695 (1977).

88. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1**, 263–273 (1986).

89. Milner, R. J. & Sutcliffe, J. G. Gene expression in rat brain. *Nucleic Acids Res.* **11**, 5497–5520 (1983).

90. Putney, S. D., Herlihy, W. C. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**, 718–721 (1983).

91. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

92. Smith, K. *A Brief History of NCBI's Formation and Growth*. (National Center for Biotechnology Information (US), 2013).

93. Tatusova, T. A., Karsch-Mizrachi, I. & Ostell, J. A. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* **15**, 536–543 (1999).

94. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.* **28**, 126–128 (2000).

95. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).

96. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).

97. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

98. Våge, Adding a cost of resistance description extends the ability of virus–host model to explain observed patterns in structure and function of pelagic microbial communities, Environmental Microbiology (2013)

99. Yang, A., Troup, M. & Ho, J. W. K. Scalability and Validation of Big Data Bioinformatics Software. *Comput. Struct. Biotechnol. J.* **15**, 379–386 (2017).

100. Information, N. C. for B., Pike, U. S. N. L. of M. 8600 R., MD, B. & Usa, 20894. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/.

101. The world's most valuable resource is no longer oil, but data. *The Economist*.

102. Project Jupyter. https://www.jupyter.org.

103. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

104. Overmars, L., Kerkhoven, R., Siezen, R. J. & Francke, C. MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics* **14**, 209–209 (2013).

105. sqlite3 — DB-API 2.0 interface for SQLite databases — Python 3.8.3 documentation. https://docs.python.org/3/library/sqlite3.html.

106. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

107. Home - Identical Protein Groups - NCBI. https://www.ncbi.nlm.nih.gov/ipg.

108. Zhulin, I. B. Databases for Microbiologists. *J. Bacteriol.* **197**, 2458–2467 (2015).

109. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2015).

110. Welch, L. *et al.* Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Comput. Biol.* **10**, (2014).

111. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).

112. Azam, F. *et al.* The Ecological Role of Water-Column Microbes in the Sea. *Mar. Ecol. Prog. Ser.* **10**, 257–263 (1983).

113. Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol.* **12**, 686–698 (2014).

114. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).

115. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience* **49**, 781–788 (1999).

116. Avrani, S. & Lindell, D. Convergent evolution toward an improved growth rate and a reduced resistance range in *Prochlorococcus* strains resistant to phage. *Proc. Natl. Acad. Sci.* **112**, E2191 (2015).

117. Bohannan, B. J. M., Kerr, B., Jessup, C. M. & Hughes, J. B. Trade-offs and coexistence in microbial microcosms. 9.

118. Luria, S. E. & Delbrück, M. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**, 491–511 (1943).

119. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu. Rev. Microbiol.* **71**, 233–261 (2017).

120. Westra, E. R., Dowling, A. J., Broniewski, J. M. & van Houte, S. Evolution and Ecology of CRISPR. *Annu. Rev. Ecol. Evol. Syst.* **47**, 307–331 (2016).

121. Olovnikov, I., Chan, K., Sachidanandam, R., Newman, D. K. & Aravin, A. A. Bacterial argonaute samples the transcriptome to identify foreign DNA. *Mol. Cell* **51**, 594–605 (2013).

122. Yamaguchi, Y., Park, J.-H. & Inouye, M. Toxin-Antitoxin Systems in Bacteria and Archaea. *Annu. Rev. Genet.* **45**, 61–79 (2011).

123. Cohen, D. *et al.* Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature* **574**, 691–695 (2019).

124. Chaudhary, K. BacteRiophage EXclusion (BREX): A novel anti-phage mechanism in the arsenal of bacterial defense system. *J. Cell. Physiol.* **233**, 771–773 (2018).

125. Bertani, G. & Weigle, J. J. Host Controlled Variation in Bacterial Viruses. *J. Bacteriol.* **65**, 113–121 (1953).

126. Luria, S. E. & Human, M. L. A Nonhereditary, Host-Induced Variation of Bacterial Viruses. *J. Bacteriol.* **64**, 557–569 (1952).

127. Korona, R., Korona, B. & Levin, B. R. Sensitivity of naturally occurring coliphages to type I and type II restriction and modification. *J. Gen. Microbiol.* **139 Pt 6**, 1283–1290 (1993).

128. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2015).

129. Arber, W. & Wauters-Willems, D. Host specificity of DNA produced by *Escherichia coli*. *Mol. Gen. Genet. MGG* **108**, 203–217 (1970).

130. Roberts, R. J. *et al.* A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805–1812 (2003).

131. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**, 4360–4377 (2013).

132. Makarova, K. S., Wolf, Y. I., Snir, S. & Koonin, E. V. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* **193**, 6039–6056 (2011).

133. Vasu, K. & Nagaraja, V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. *Microbiol. Mol. Biol. Rev. MMBR* **77**, 53–72 (2013).

134. Fullmer, M. S., Ouellette, M., Louyakis, A. S., Papke, R. T. & Gogarten, J. P. The Patchy Distribution of Restriction Modification System Genes and the Conservation of Orphan Methyltransferases in Halobacteria. *Genes* **10**, 233 (2019).

135. Roer, L. *et al.* Is the Evolution of *Salmonella enterica subsp. enterica* Linked to Restriction-Modification Systems? *mSystems* **1**, e00009-16 (2016).

136. Oliveira, P. H., Touchon, M. & Rocha, E. P. C. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5658–5663 (2016).

137. Kong, H. *et al.* Functional analysis of putative restriction–modification system genes in the *Helicobacter pylori* J99 genome. *Nucleic Acids Res.* **28**, 3216–3223 (2000).

138. Nobusato, A., Uchiyama, I. & Kobayashi, I. Diversity of restriction–modification gene homologues in *Helicobacter pylori*. *Gene* **259**, 89–98 (2000).

139. Casadevall, A. Evolution of Intracellular Pathogens. *Annu. Rev. Microbiol.* **62**, 19–33 (2008).

140. Budroni, S. *et al. Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4494–4499 (2011).

141. Zhao, L. *et al.* The highly heterogeneous methylated genomes and diverse restriction-modification systems of bloom-forming *Microcystis*. *Harmful Algae* **75**, 87–93 (2018).

142. Huisman, J. *et al.* Cyanobacterial blooms. *Nat. Rev. Microbiol.* **16**, 471–483 (2018).

143. Li, X., Dreher, T. W. & Li, R. An overview of diversity, occurrence, genetics and toxin production of bloom-forming *Dolichospermum* (*Anabaena*) species. *Harmful Algae* **54**, 54–68 (2016).

144. Bergman, B., Sandh, G., Lin, S., Larsson, J. & Carpenter, E. J. *Trichodesmium* – a widespread marine cyanobacterium with unusual nitrogen fixation properties. *Fems Microbiol. Rev.* **37**, 286–302 (2013).

145. Darwin, C. *Narrative of the surveying voyages of His Majesty's Ships Adventure and Beagle between the years 1826 and 1836, describing their examination of the southern shores of South America, and the Beagle's circumnavigation of the globe. Journal and remarks. 1832-1836. pg 14-17*. vol. III (London: Henry Colburn, 1839).

146. Westberry, T. K. & Siegel, D. A. Spatial and temporal distribution of *Trichodesmium* blooms in the world's oceans. *Glob. Biogeochem. Cycles* **20**, (2006).

147. Steffen, M. M. *et al.* Ecophysiological Examination of the Lake Erie *Microcystis* Bloom in 2014: Linkages between Biology and the Water Supply Shutdown of Toledo, OH. *Environ. Sci. Technol.* **51**, 6745–6755 (2017).

148. Hu, L. *et al.* Multi-Year Assessment of Toxic Genotypes and Microcystin Concentration in Northern Lake Taihu, China. *Toxins* **8**, (2016).

149. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13 (2014).

150. Olson, R. J., Chisholm, S. W., Zettler, E. R. & Armbrust, E. V. Pigments, size, and distributions of *Synechococcus* in the North Atlantic and Pacific Oceans. *Limnol. Oceanogr.* **35**, 45–58 (1990).

151. Coutinho, F., Tschoeke, D. A., Thompson, F. & Thompson, C. Comparative genomics of *Synechococcus* and proposal of the new genus *Parasynechococcus*. *PeerJ* **4**, e1522–e1522 (2016).

152. Mann, N. H. Phages of the marine cyanobacterial picophytoplankton. *FEMS Microbiol. Rev.* **27**, 17–34 (2003).

153. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. *Nature* **424**, 1047 (2003).

154. Bullerjahn, G. S. *et al.* Global solutions to regional problems: Collecting global expertise to address the problem of harmful cyanobacterial blooms. A Lake Erie case study. *Harmful Algae* **54**, 223–238 (2016).

155. Elhai, J., Vepritskiy, A., Muro-Pastor, A. M., Flores, E. & Wolk, C. P. Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena sp*. strain PCC 7120. *J. Bacteriol.* **179**, 1998–2005 (1997).

156. Szyf, M. *et al.* DNA methylation pattern is determined by the intracellular level of the methylase. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3278–3282 (1984).

157. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci.* **110**, 9824 (2013).

158. Heisler, J. *et al.* Eutrophication and Harmful Algal Blooms: A Scientific Consensus. *Harmful Algae* **8**, 3–13 (2008).

159. Jeltsch, A. & Pingoud, A. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.* **42**, 91–96 (1996).

160. Kobayashi, I. Behavior of restriction–modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29**, 3742–3756 (2001).

161. Furuta, Y. *et al.* Birth and death of genes linked to chromosomal inversion. *Proc. Natl. Acad. Sci.* **108**, 1501–1506 (2011).

162. Furuta, Y., Abe, K. & Kobayashi, I. Genome comparison and context analysis reveals putative mobile forms of restriction–modification systems and related rearrangements. *Nucleic Acids Res.* **38**, 2428–2443 (2010).

163. Holt, R. D., Grover, J. & Tilman, D. Simple Rules for Interspecific Dominance in Systems with Exploitative and Apparent Competition. *Am. Nat.* **144**, 741–771 (1994).

164. Leibold, M. A. A Graphical Model of Keystone Predators in Food Webs: Trophic Regulation of Abundance, Incidence, and Diversity Patterns in Communities. *Am. Nat.* **147**, 784–812 (1996).

165. McPeek, M. A. Trade-Offs, Food Web Structure, and the Coexistence of Habitat Specialists and Generalists. *Am. Nat.* **148**, S124–S138 (1996).

166. Våge, S., Storesund, J. E., Giske, J. & Thingstad, T. F. Optimal Defense Strategies in an Idealized Microbial Food Web under Trade-Off between Competition and Defense. *PLOS ONE* **9**, e101415 (2014).

167. Bohannan, B. J. M. & Lenski, R. E. The Relative Importance of Competition and Predation Varies with Productivity in a Model Community. *Am. Nat.* **156**, 329–340 (2000).

168. Gómez, P., Bennie, J., Gaston, K. J. & Buckling, A. The impact of resource availability on bacterial resistance to phages in soil. *PloS One* **10**, e0123752–e0123752 (2015).

169. Levin, B. R., Stewart, F. M. & Chao, L. Resource-Limited Growth, Competition, and Predation: A Model and Experimental Studies with Bacteria and Bacteriophage. *Am. Nat.* **111**, 3–24 (1977).

170. Lin, S. *et al.* Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. *Gene* **473**, 139–149 (2011).

171. Steffen, M. M. *et al.* Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *ISME J.* **8**, 2080–2092 (2014).

172. Iranzo, J., Cuesta, J. A., Manrubia, S., Katsnelson, M. I. & Koonin, E. V. Disentangling the effects of selection and loss bias on gene dynamics. *Proc. Natl. Acad. Sci.* **114**, E5616 (2017).

173. Forde, S. E. *et al.* Understanding the limits to generalizability of experimental evolutionary models. *Nature* **455**, 220–223 (2008).

174. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

175. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).

176. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

177. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).

178. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28**, 3150–3152 (2012).

179. Godzik, A. & Li, W. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).

180. Shen, B. W. *et al.* Characterization and crystal structure of the type IIG restriction endonuclease RM.BpuSI. *Nucleic Acids Res.* **39**, 8223–8236 (2011).

181. Xu, C. Y., Yu, F., Hu, X. J. & He, J. H. The 2.45 A Crystal Structure of the Restriction Endonuclease Sau3AI Suggests a Self-Inhibition Mechanism. *BE Publ.* doi:10.2210/pdb4pxg/pdb.

182. Chand, M. K. *et al.* Translocation-coupled DNA cleavage by the Type ISP restriction-modification enzymes. *Nat. Chem. Biol.* **11**, 870–877 (2015).

183. Gupta, Y. K., Chan, S.-H., Xu, S. & Aggarwal, A. K. Structural basis of asymmetric DNA methylation and ATP-triggered long-range diffusion by EcoP15I. *Nat. Commun.* **6**, (2015).

184. Kulkarni, M., Nirwan, N., van Aelst, K., Szczelkun, M. D. & Saikrishnan, K. Structural insights into DNA sequence recognition by Type ISP restriction-modification enzymes. *Nucleic Acids Res.* **44**, 4396–4408 (2016).

185. Callahan, S. J. *et al.* Structure of Type IIL Restriction-Modification Enzyme MmeI in Complex with DNA Has Implications for Engineering New Specificities. *PLOS Biol.* **14**, e1002442 (2016).

186. Våge, S., Storesund, J. E. & Thingstad, T. F. Adding a cost of resistance description extends the ability of virus-host model to explain observed patterns in structure and function of pelagic microbial communities: Structuring of microbial communities by viruses. *Environ. Microbiol.* **15**, 1842–1852 (2013).

187. Millman, K. J. & Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **13**, 9–12 (2011).

188. Mckay, M. D., Beckman, R. J. & Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics* **42**, 55–61 (2000).

189. McKinney, W. Data Structures for Statistical Computing in Python. in 51–56 (2010).

190. Thamatrakoln, K. *et al.* Light regulation of coccolithophore host–virus interactions. *New Phytol.* **221**, 1289–1302 (2019).

191. Record, N. R., Talmy, D. & Våge, S. Quantifying Tradeoffs for Marine Viruses. *Front. Mar. Sci.* **3**, (2016).

192. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. *Proc. Natl. Acad. Sci.* **108**, E288–E297 (2011).

193. Arber, W. Host specificity of DNA produced by *Escherichia coli*: V. The role of methionine in the production of host specificity. 10 (1965).

194. Diekmann, S. DNA methylation can enhance or induce DNA curvature. *EMBO J.* **6**, 4213–4217 (1987).

195. Polaczek, P., Kwan, K. & Campbell, J. L. GATC motifs may alter the conformation of DNA depending on sequence context and N6-adenine methylation status: possible implications for DNA-protein recognition. *Mol. Gen. Genet. MGG* **258**, 488–493 (1998).

196. Casadesús, J. & Low, D. Epigenetic Gene Regulation in the Bacterial World. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).

197. Barras, F. & Marinus, M. G. The great GATC: DNA methylation in E. coli. *Trends Genet. TIG* **5**, 139–143 (1989).

198. Gonzalez, D., Kozdon, J. B., McAdams, H. H., Shapiro, L. & Collier, J. The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. *Nucleic Acids Res.* **42**, 3720–3735 (2014).

199. Oshima, T. *et al.* Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli*. *Mol. Microbiol.* **45**, 673–695 (2002).

200. Westphal, L. L., Sauvey, P., Champion, M. M., Ehrenreich, I. M. & Finkel, S. E. Genomewide Dam Methylation in *Escherichia coli* during Long-Term Stationary Phase. *mSystems* **1**, (2016).

201. Su, S. S., Lahue, R. S., Au, K. G. & Modrich, P. Mispair specificity of methyl-directed DNA mismatch correction in vitro. *J. Biol. Chem.* **263**, 6829–6835 (1988).

202. Grilley, M., Welsh, K. M., Su, S. S. & Modrich, P. Isolation and characterization of the *Escherichia coli* mutL gene product. *J. Biol. Chem.* **264**, 1000–1004 (1989).

203. Welsh, K. M., Lu, A. L., Clark, S. & Modrich, P. Isolation and characterization of the *Escherichia coli* mutH gene product. *J. Biol. Chem.* **262**, 15624–15629 (1987).

204. Grilley, M., Griffith, J. & Modrich, P. Bidirectional excision in methyl-directed mismatch repair. *J. Biol. Chem.* **268**, 11830–11837 (1993).

205. Zweiger, G., Marczynski, G. & Shapiro, L. A *Caulobacter* DNA methyltransferase that functions only in the predivisional cell. *J. Mol. Biol.* **235**, 472–485 (1994).

206. Marczynski, G. T. Chromosome Methylation and Measurement of Faithful, Once and Only Once per Cell Cycle Chromosome Replication in *Caulobacter crescentus*. *J. Bacteriol.* **181**, 1984–1993 (1999).

207. Løbner-Olesen, A., Skovgaard, O. & Marinus, M. G. Dam methylation: coordinating cellular processes. *Curr. Opin. Microbiol.* **8**, 154–160 (2005).

208. Zhao, L. *et al.* The highly heterogeneous methylated genomes and diverse restriction-modification systems of bloom-forming *Microcystis*. *Harmful Algae* **75**, 87–93 (2018).

209. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLoS Genet.* **12**, e1005854–e1005854 (2016).

210. Seshasayee, A. S. N., Singh, P. & Krishna, S. Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.* **40**, 7066–7073 (2012).

211. Pleška, M. *et al.* Bacterial Autoimmunity Due to a Restriction-Modification System. *Curr. Biol.* **26**, 404–409 (2016).

212. Anjum, A. *et al.* Phase variation of a Type IIG restriction-modification enzyme alters site-specific methylation patterns and gene expression in *Campylobacter jejuni* strain NCTC11168. *Nucleic Acids Res.* **44**, 4581–4594 (2016).

213. Dvořák, P. *et al.* Species concepts and speciation factors in cyanobacteria, with connection to the problems of diversity and classification. *Biodivers. Conserv.* **24**, 739–757 (2015).

214. Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci.* **110**, 1053–1058 (2013).

215. Oren, A. & Ventura, S. The current status of cyanobacterial nomenclature under the "prokaryotic" and the "botanical" code. *Antonie Van Leeuwenhoek* **110**, 1257–1269 (2017).

216. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).

217. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

218. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

219. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinforma. Oxf. Engl.* **28**, 3150–3152 (2012).

220. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

221. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

222. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

223. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

224. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74**, 5088–5090 (1977).

225. Young, A. D. & Gillung, J. P. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Syst. Entomol.* **45**, 225–247 (2020).

226. LURIA, S. E. & HUMAN, M. L. A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.* **64**, 557–569 (1952).

227. Bertani, G. & Weigle, J. J. Host Controlled Variation in Bacterial Viruses. *J. Bacteriol.* **65**, 113–121 (1953).

228. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci.* **112**, 15690–15695 (2015).

229. Bickle, T. A. & Krüger, D. H. Biology of DNA restriction. *Microbiol. Rev.* **57**, 434–450 (1993).

230. Baharoglu, Z. & Mazel, D. SOS, the formidable strategy of bacteria against aggressions. *FEMS Microbiol. Rev.* **38**, 1126–1145 (2014).

231. Pennington, J. M. & Rosenberg, S. M. Spontaneous DNA breakage in single living *Escherichia coli* cells. *Nat. Genet.* **39**, 797–802 (2007).

232. Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science* **297**, 1183–1186 (2002).

233. Klimuk, E. *et al.* Controller protein of restriction–modification system Kpn2I affects transcription of its gene by acting as a transcription elongation roadblock. *Nucleic Acids Res.* **46**, 10810–10826 (2018).

234. Morgan, R. D., Dwinell, E. A., Bhatia, T. K., Lang, E. M. & Luyten, Y. A. The MmeI family: type II restriction–modification enzymes that employ single-strand modification for host protection. *Nucleic Acids Res.* **37**, 5208–5221 (2009).

235. Rusinov, I. S., Ershova, A. S., Karyagina, A. S., Spirin, S. A. & Alexeevski, A. V. Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses. *BMC Genomics* **19**, 885 (2018).

236. Sargentini, N. J., Diver, W. P. & Smith, K. C. The Effect of Growth Conditions on Inducible, recA-Dependent Resistance to X Rays in *Escherichia coli*. *Radiat. Res.* **93**, 364–380 (1983).

237. Kettler, G. C. *et al.* Patterns and Implications of Gene Gain and Loss in the Evolution of *Prochlorococcus*. *PLOS Genet.* **3**, e231 (2007).

238. Våge, S., Storesund, J. E. & Thingstad, T. F. Adding a cost of resistance description extends the ability of virus-host model to explain observed patterns in structure and function of pelagic microbial communities: Structuring of microbial communities by viruses. *Environ. Microbiol.* **15**, 1842–1852 (2013).

239. Dybvig, K., Sitaraman, R. & French, C. T. A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 13923–13928 (1998).

240. Furuta, Y. *et al.* Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLOS Genet.* **10**, e1004272 (2014).

241. Furuta, Y. & Kobayashi, I. Movement of DNA sequence recognition domains between non-orthologous proteins. *Nucleic Acids Res.* **40**, 9218–9232 (2012).

242. Yan, W. *et al.* Genome Rearrangement Shapes *Prochlorococcus* Ecological Adaptation. *Appl. Environ. Microbiol.* **84**, (2018).

243. Wuichet, K. & Zhulin, I. B. Origins and Diversification of a Complex Signal Transduction System in Prokaryotes. *Sci. Signal.* **3**, ra50 (2010).

244. Lavorel, S. *et al.* Plant Functional Types: Are We Getting Any Closer to the Holy Grail? in *Terrestrial Ecosystems in a Changing World* (eds. Canadell, J. G., Pataki, D. E. & Pitelka, L. F.) 149–164 (Springer, 2007). doi:10.1007/978-3-540-32730-1_13.

245. Thomas, K. *et al.* Jupyter Notebooks- a publishing format for reproducible computational workflows. *Stand Alone* 87–90 (2016) doi:10.3233/978-1-61499-649-1-87.

246. Leipzig, J. A review of bioinformatic pipeline frameworks. *Brief. Bioinform.* **18**, 530–536 (2017).

247. Wang, C., Chen, M.-H., Schifano, E., Wu, J. & Yan, J. Statistical methods and computing for big data. *Stat. Interface* **9**, 399–414 (2016).

VITA

Spiridon (Spiro) Evangelos Papoulis was born September 16th, 1992 in Royal Oak, Michigan and is the son of Laurie and Evangelos Papoulis. He soon found himself part of a cabal composed of his two younger brothers, Yannis and Laki, and his cousin, Alex, causing most of his childhood to be in fine company. His fondest memories of childhood were of the many family adventures to national parks. After graduating Northville High School in 2011, he attended Michigan State University to earn his degree in Biochemistry and his minor in Computer Science. During his undergraduate years, he found he had taken the home cooking of his mother and yiayia for granted, so while initially born from necessity, he started his love for cooking.

He was given the opportunity to attend an REU program here at the University of Tennessee in 2014, which he later attended for graduate school in 2015. Not only was he introduced his PhD advisor, he also met and fell madly in love with his wife, Katherine. They spend the first two years of their relationship long distant, often taking a car, to a bus, to a train, to a taxi to see one another. After many miles traveled over the years, they started their life together in Knoxville and eventually married on July 4th in 2019. Being a tinkerer, Spiro enjoys building his own home network with old electronics, and he is fairly confident that his network is not self-aware.

Yet.