RIBONUCLEOTIDE REDUCTASE GENES INFLUENCE THE BIOLOGY AND ECOLOGY OF MARINE VIRUSES

by

Amelia O. Harrison

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Marine Studies

Fall 2019

© 2019 Amelia O. Harrison
All Rights Reserved
RIBONUCLEOTIDE REDUCTASE GENES INFLUENCE THE BIOLOGY AND ECOLOGY OF MARINE VIRUSES

by

Amelia O. Harrison

Approved: ________________________________________________________

K. Eric Wommack, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: ________________________________________________________

Mark A. Moline, Ph.D.
Director of the School of Marine Science and Policy

Approved: ________________________________________________________

Estella A. Atekwana, Ph.D.
Dean of the College of Earth, Ocean, and Environment

Approved: ________________________________________________________

Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education and Dean of the Graduate College
ACKNOWLEDGMENTS

First, I would like to thank my advisor, Eric Wommack, for taking a chance on me as a freshman undergrad and again as a masters student. It has been a privilege to work with someone so passionate and excited about science. I would also like to thank Shawn Polson, who has also supported me and helped foster my love for research over the past six years. I would like to thank the final member of my committee, Jennifer Biddle, for her insight and general awesomeness.

I am deeply thankful for all of my past and current labmates. Barb Ferrell who somehow always has time for everyone. Prasanna Joglekar, for always making me laugh. Rachel Marine, who took me on as an intern despite not having the time. Sean McAllister, though technically not a labmate, for always being up for a good debate. And, of course, Ryan Moore, without whom none of this work would have been possible.

I also have some people back home to acknowledge. Steve Jordan, who first taught me to pipette and how to run a PCR. Jeremy Lauver and Angela Gockley for helping to spark my interest in science. Sienna Fisher, Wesley Sigl, and Scott Leighow, who always encouraged me to be myself.

My deepest thanks also go, of course, to my family. My siblings, Maggie and Henry, who (usually) put up with my science babble. My mother, Brenda, for her endless support and for keeping me grounded. And my father, Dr. Harry Harrison, who stressed to us the value of hard work and education. I love you all.

I would also like to acknowledge funding from the National Science Foundation and the CEOE. Special thanks also to the CBCB core facility and those who keep Biomix up and running so we can continue to submit doomed slurm scripts.
TABLE OF CONTENTS

LIST OF TABLES .......................................................... vii
LIST OF FIGURES .......................................................... viii
ABSTRACT ................................................................. x

Chapter

1 INTRODUCTION ......................................................... 1

2 REANNOTATION OF THE RIBONUCLEOTIDE REDUCTASE IN A CYANOPHAGE REVEALS LIFE HISTORY STRATEGIES WITHIN THE VIRIOPLANKTON ......................... 6

2.1 Introduction ...................................................... 6
2.2 Methods ............................................................ 7

2.2.1 The Cyano SP Clade ............................................. 7
2.2.2 Putative $\alpha$ and $\beta$ subunit identification ................ 8
2.2.3 Phylogenetic analysis ........................................... 9

2.2.3.1 Phylogenetic reference sequence curation ............... 9
2.2.3.2 Reference sequence preparation for phylogeny construction ......................................................... 10
2.2.3.3 Phylogenetic tree construction ............................ 12

2.2.4 Sequence similarity network .................................. 13

2.3 Results ............................................................. 13

2.3.1 P-SSP7 Class I $\alpha$ subunit identification .................. 14
2.3.2 P-SSP7 Class I $\beta$ subunit identification ................... 14
2.3.3 Assignment of P-SSP7 RNR to a Class I subclass ............ 16
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Cyano SP clade reference sequences and their hosts.</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>RNR Class I subclass membership of RNRdb groups.</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>Metal-binding amino acid residues in each of the β subunit references and P-SSP7.</td>
<td>18</td>
</tr>
<tr>
<td>2.4</td>
<td>Metal-binding amino acid residues in each of the RNRdb groups and the Cyano SP clade.</td>
<td>19</td>
</tr>
<tr>
<td>3.1</td>
<td>Number of GOV 2.0 RNRs analyzed by type.</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Number of samples taken at each depth.</td>
<td>53</td>
</tr>
<tr>
<td>A.1</td>
<td>Clustering statistics for Class I α subunit sequences removed from RNRdb groups NrdE and NrdAz prior to phylogenetic analysis.</td>
<td>90</td>
</tr>
<tr>
<td>A.2</td>
<td>Annotated Class I α, Class I β, and Class II references used for alignment of putative α and candidate β subunits and curation of phylogenetic reference sequences.</td>
<td>91</td>
</tr>
<tr>
<td>A.3</td>
<td>Catalytic residues in Class I RNR α and β subunits and their positions in the putative α and β sequences from Prochlorococcus phage P-SSP7.</td>
<td>92</td>
</tr>
<tr>
<td>A.4</td>
<td>Accession numbers for genomes and RNR and Pol I protein sequences of the Cyano SP clade.</td>
<td>93</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

1.1 Summary of RNR class and Class I subclass divisions. 4

2.1 Cyano SP $\alpha$ subunits group with Class I RNRs. 15

2.2 Cyano SP $\beta$ subunits show no close relationship to known Class I clades. 20

2.3 Cyano SP RNRs cluster separately from known RNRs. 22

2.4 Cyano SP $\alpha$ subunits are most closely related to RNRdb clades NrdAk and NrdAi. 23

2.5 There is evolutionary pressure for cyanophage to carry Class I RNRs. 31

3.1 GOV 2.0 samples are from globally-distributed marine sites. 34

3.2 GOV 2.0 contains much of the known Class I and Class II RNR diversity. 46

3.3 GOV 2.0 ecological zones are preserved by the RNR-containing virus community, but $\beta$-diversity indicates new sample clusters. 47

3.4 Viral community structure changes with biogeographical regions and environmental variables. 49

3.5 RNR subclass abundances drive separations among samples. 51

3.6 RNR subclasses show different latitudinal patterns of abundance. 52

3.7 RNR subclasses display different depth profiles. 54

3.8 RNR subclasses have different biogeographical distributions. 55

A.1 Maximum-likelihood phylogenetic tree of Class I alpha subunit with Class II sequences from the RNRdb. 94
A.2 Maximum-likelihood phylogenetic trees of trimmed RNRdb Class I beta subunits. .......................... 95
A.3 Maximum-likelihood phylogenetic trees of trimmed RNRdb Class I alpha subunits. .......................... 96
A.4 Protein SSNs for Class I $\alpha$ with Class II sequences. ........................................ 97
B.1 Workflow of the Protein Active Site Validator (PASV). ................. 99
B.2 Construction of RNR Gold. ................................................. 101
C.1 Class I $\alpha$ and Class I $\beta$ RNR sequences from marine viruses make identical trees. .......................... 104
D.1 Clustering GOV 2.0 RNRS did not affect the overall shape of the data. 105
ABSTRACT

In marine systems, virioplankton are critical to biogeochemical cycling, microbial mortality, and horizontal gene transfer. However, relatively little is known about the ecology and biology of environmental phage populations, largely due to the lack of a universal marker. Viral marker genes such as ribonucleotide reductase (RNR) can provide insight into both phage diversity and phenotype. RNRs are ancient enzymes that reduce ribonucleotides to deoxyribonucleotides. This is the only known de novo pathway for dNTP synthesis and is the rate-limiting step of DNA synthesis. Thus, RNRs are common in the genomes of lytic dsDNA phage. RNRs are divided into three classes based on their reactivity with oxygen: Class I RNRs are O$_2$-dependent; Class II RNRs are O$_2$-independent, B$_{12}$-dependent; and Class III RNRs are O$_2$-sensitive. Class I is further divided into subclasses (Ia-If) based on metal requirements. These biochemical differences tie RNRs and their cellular carriers closely to their environments. We hypothesized that this connection would hold true for viruses as well as cells, making RNRs informative at both the individual virus and community levels. At the individual level, we corrected a misannotation of the RNR in a cyanophage, revealing an RNR with highly specific adaptations to the intracellular environment of its host. This cyanophage RNR also showed adaptations that may have a role in driving its observed highly lytic phenotype. At the community level, the RNR-containing virus community was reflective of the total viral community and that viruses with different RNR types had different ecological distributions in the oceans. RNR type distributions mirrored those of the trace metals required for RNR activity, except for subclass If. Subclass If is exclusive to cyanophage, so this subclass mirrored the distributions of marine Synechococcus and Prochlorococcus. Together, these two stories show RNRs as drivers of marine viral ecology and influencers of viral phenotype and life history strategies.
Chapter 1

INTRODUCTION

Viruses are the most abundant biological entities on the planet, with an estimated $10^{31}$ viral particles globally [1]. While viruses are known to infect cellular life from all three domains, viruses largely influence ecosystems through the infection of microbial hosts. In the oceans, $10^{23}$ viral infections are estimated to take place every second, resulting in the mortality of approximately 20% of marine microbial biomass each day [2]. Cell lysis resulting from viral infection influences ocean biogeochemical cycling by returning particulate and dissolved organic matter to the water column [1, 3], where it may be taken up by microbial populations to fuel new growth, or exported to the deep ocean [2, 4]. Viral predation can also influence biogeochemical cycles through the restructuring of microbial populations [5], metabolic reprogramming of host cells [6, 7], and horizontal gene transfer [8]. While the importance of viruses within marine microbial communities is now commonly accepted, the biological and ecological details of viral-host interactions that influence the transformations of nutrient elements in ecosystems are largely unknown. Bridging the gap between genetic observations and ecosystem-level effects requires an understanding of the connections between genes and phenotypes. Among viruses infecting marine microbes, genes involved in nucleotide metabolism and viral replication are highly predictive of viral phenotype and evolutionary history [9, 10, 11].

For example, a point mutation in motif B of the family A DNA polymerase gene (polA) is indicative of viral life style [12, 13]. Another useful viral marker gene is ribonucleotide reductase (RNR). RNRs catalyze the rate-limiting step of DNA synthesis (ribonucleotide reduction) [14, 15], and are therefore prominent in the genomes of lytic dsDNA phage [10, 16, 17]. Because RNRs have evolved into several types with diverse
biochemical mechanisms and nutrient requirements [18], the RNR used by a cell or virus can reflect the environmental conditions surrounding DNA replication [16, 19, 20].

All RNRS share a common catalytic mechanism in which a thiol radical in the active site removes a hydrogen atom from the 3’ hydroxyl group of the ribose sugar, thereby activating the substrate [21, 22, 23]. The mechanism by which the thiol radical is generated varies greatly among RNRS and provides the biochemical basis dividing the three major RNR classes [23]. Extant RNRS are also commonly divided by their reactivity with O₂ [20]: Class I RNRS are O₂-dependent; Class II RNRS are O₂-independent; and Class III RNRS are O₂-sensitive (Figure 1.1A).

Class III RNRS are the most dissimilar of the extant types, bearing no sequence similarity to Class I and II RNRS despite a common ancestry [23, 24]. They consist of two subunits that create radicals by cleaving S-adenosylmethionine molecules using iron-sulfur clusters [25]. Class III RNRS are inactivated by O₂ [26, 27], and are therefore found only in strict or facultative anaerobes and their viruses [28]. Class II RNRS are the only RNRS that do not require separate subunits for radical generation and catalysis [18]. Instead, Class II RNRS are encoded by a single gene, nrdJ. Class II RNRS require adenosylcobalamin (AdoCbl), a form of B₁2, to produce a radical [29, 30]. There are two subclasses of Class II RNR: monomeric and dimeric [18].

Class I RNRS are the most recent [23] and the most complex of the extant RNRS (Figure 1.1B). Radical generation takes place on a smaller subunit (β or R2) and is transferred to a larger catalytic subunit (α or R1) [31]. The α subunit is encoded by nrdA or nrdE and the β subunit is encoded by nrdB or nrdF. These genes form exclusive pairs: nrdA is found only with nrdB (nrdAB), and nrdE is found only with nrdF (nrdEF). Notably, the Class I α subunit is thought to have evolved directly from Class II RNRS, so they share several catalytic sites, though sequence similarity between the two classes remains low otherwise [23]. The radical initiation mechanism of the β subunit further divides Class I RNRS into five subclasses (a-e) [19, 32, 33, 34]. The subclasses are divided based on the identity of the metallocofactor (or absence thereof), the identity of the oxidant, and whether the β subunit contains (and utilizes)
the tyrosine radical site (Figure 1.1B). Class I RNRs are generally presumed to be subclass Ia enzymes unless they can be assigned to another subclass based on sequence homology to a close relative that has been biochemically characterized [35].

To further our understanding of how carrying RNR genes influences the biology and ecology of viruses in the environment, this thesis will focus on viral RNRs at both the individual and community levels. First, the RNR in a cyanophage reveals important life history strategies for survival in both the marine and intracellular environments. Then, RNRs from a metagenomic viral dataset show the utility of RNR as a marker gene and reveal details about the connections between RNR-carrying viruses and the environment. Through the analysis of viral RNRs, this thesis advances our understanding individual genes as drivers viral biology and ecology.
Figure 1.1: Summary of A) RNR class and B) Class I subclass divisions. Gray outlined boxes to the left indicate categories. In B, like colors indicate common traits and light gray filled boxes indicate missing traits.
PREFACE TO CHAPTER 2

The work presented in Chapter 2 of this thesis has previously been published with myself as the first author [36]. I performed the analysis and wrote the manuscript. Ryan M. Moore created the sequence similarity networks, assisted with the analysis, and edited the manuscript. K. Eric Wommack and Shawn W. Polson contributed to study design, data interpretation, and manuscript preparation.
Chapter 2

REANNOTATION OF THE RIBONUCLEOTIDE REDUCTASE IN A CYANOPHAGE REVEALS LIFE HISTORY STRATEGIES WITHIN THE VIRIOPLANKTON

2.1 Introduction

The addition rate of predicted protein sequences to reference databases has far surpassed the rate at which they can be experimentally characterized [37, 38]. Consequently, most functional annotation of proteins is done bioinformatically [37, 38, 39]. Both sequence and structure based methods of annotation require the use of similarity metrics to transfer existing annotations from a related protein to the query protein [37]. Such homology-based methods have faced scrutiny since the annotation of the first genome [40] because of the potential for misannotation.

Misannotations can arise for a number of reasons. They may be the result of confusion surrounding protein classification [41], errors in literature [42], annotating only by top BLAST hit [43], and non-orthologous gene displacement [43]. Numerous studies have estimated misannotation in public databases [39, 44, 45, 46], with most studies finding that computationally annotated entries were more likely to contain mis-annotations than manually annotated entries [45, 46, 47]. When annotating proteins, it is important for researchers to account for errors within subject sequence databases that may bias functional assignments of their query sequences [43].

While the diversity of RNR biochemistry makes this enzyme a potential marker for inferring aspects of viral biology and ecology [16], proper annotation of RNR genes

---

1 Adapted from Amelia O. Harrison, Ryan M. Moore, Shawn W. Polson, and K. Eric Wommack. Reannotation of the ribonucleotide reductase in a cyanophage reveals life history strategies within the virioplankton. Frontiers in Microbiology, 10(134), 2019.
is imperative for this purpose. Unfortunately, the diversity of RNRS has fostered high misannotation rates, with one study reporting that 77% of RNRS submitted to GenBank had misannotations [48]. Most of those misannotations (88%) were due to RNRS sequences being assigned to the wrong class. In response, a specialty database (RNRSdb) was created for maintaining a collection of correctly annotated RNRS [48]. Even with resources such as the RNRSdb, however, the complexity of RNRS annotation remains daunting for non-experts. Class I RNRS can be particularly difficult to identify, as their classification relies largely on the annotation of both an α and β subunit.

Our prior work examining the phylogenetic relationships among RNRS from marine virioplankton revealed two large clades of cyanophage RNRS, with the first made up of Class I RNRS and the second of Class II RNRS [16]. The hosts of these cyanophage, marine *Synechococcus* and *Prochlorococcus*, carry Class II RNRS. Thus, the presence of such a large cyanophage clade with Class I RNRS was intriguing, and in contradiction to earlier findings that phage tend to carry an RNRS gene similar to that of their host cell [17]. Now, the reanalysis of an RNRS from the Class II-carrying cyanophage has revealed that the RNRS in this second clade are, in fact, Class I RNRS that were misannotated as Class II. The reannotation of the RNRS from *Prochlorococcus* phage P-SSP7 from Class II to Class I implies that most known marine cyanophage carry RNRS that are not host-derived, nor dependent on B$_{12}$. Additionally, our analysis suggests that the P-SSP7 RNRS may represent a novel Class I RNRS subclass.

2.2 Methods

2.2.1 The Cyano SP Clade

The RNRS from *Prochlorococcus* phage P-SSP7 is a member of the ‘Cyano II’ RNRS clade, as recognized by Sakowski et al. [16] in a study of virioplankton RNRS. Based on our analysis, and to avoid confusion with the nomenclature for RNRS classes, we have renamed the Cyano II clade to the Cyano SP clade, as RNRS in this clade are exclusively found within the cyanosipho- and cyanopodoviruses [16]. We have also renamed the Cyano I clade to the Cyano M clade, as RNRS in this clade are exclusively
Table 2.1: Cyano SP clade reference sequences and their hosts.

<table>
<thead>
<tr>
<th>Virus</th>
<th>Family</th>
<th>Host</th>
<th>Host RNR Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prochlorococcus phage P-SSP7</td>
<td>Podoviridae</td>
<td>Prochlorococcus marinus</td>
<td>II-monomeric</td>
</tr>
<tr>
<td></td>
<td></td>
<td>subsp. pastoris str. CCMP1986</td>
<td></td>
</tr>
<tr>
<td>Cyanophage P-SSP2</td>
<td>Podoviridae</td>
<td>P. marinus MIT9312</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Cyanophage 9515-10a</td>
<td>Podoviridae</td>
<td>P. marinus MIT9515</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Cyanophage NATL1A-7</td>
<td>Podoviridae</td>
<td>P. marinus NATL1A-7</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Cyanophage NATL2A-133</td>
<td>Podoviridae</td>
<td>P. marinus NATL2A-133</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Cyanophage SS120-1</td>
<td>Siphoviridae</td>
<td>P. marinus SS120</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Cyanophage Syn5</td>
<td>Podoviridae</td>
<td>Synechococcus str. WH8109</td>
<td>II-monomeric</td>
</tr>
<tr>
<td>Synecchococcus phage S-CBS4</td>
<td>Siphoviridae</td>
<td>Synechococcus CB0101</td>
<td>II-monomeric</td>
</tr>
</tbody>
</table>

seen in cyanomyoviruses. The aforementioned study included ten reference sequences from the (now) Cyano SP clade. Eight of those ten references were used in the current study (Table 2.1). Cyanophage KBS-S-1A was excluded because its genome has not been fully sequenced and Synechococcus phage S-CBP3 was excluded because its RNR was missing a conserved catalytic site. P-SSP7 was chosen as the clade representative because it is the most well-studied phage from this group, has a full genome available, and is the source of the original RNR misannotation.

2.2.2 Putative $\alpha$ and $\beta$ subunit identification

Putative $\alpha$ and $\beta$ subunit sequences were extracted from the genome of *Prochlorococcus* phage P-SSP7 (genome accession no. NC_006882.2). The putative Class I $\alpha$ subunit is the RNR currently identified in the P-SSP7 genome as ribonucleotide reductase class II (accession no. YP_214197.1) and was downloaded from NCBI in April
2018. As P-SSP7 has no annotated β subunit, candidate β sequences were identified based on length filtering of unannotated protein sequences. While Class I β subunits are typically between 350 and 400 amino acids [15], we expanded our search range to avoid excluding any potential Class I β subunits. Four candidate, unannotated proteins between 200 and 500 amino acids in length were downloaded for analysis in May 2018. Candidate proteins were searched against the Conserved Domain Database using batch CD-Search [49].

The P-SSP7 putative Class I RNR α subunit and four candidate β subunit proteins were imported into Geneious v10.2.4 (https://www.geneious.com) to analyze conserved residues. The putative α subunit peptide sequence was aligned with one representative of each of the known Class I subclasses (Table A.2) using the MAFFT v7.388 Geneious plug-in [50] on the FFT-NS-ix1000 (iterative refinement method with 1000 iterations) setting with the BLOSUM62 scoring matrix. If necessary, alignments were manually modified to ensure that annotated active sites in the subclass representatives were properly aligned. References have been biochemically characterized and have corresponding crystal structures, where possible. Active sites were annotated for each of the subclass representatives based on literature reports and crystal structures. Residues from the putative P-SSP7 Class I α subunit aligning with active sites in subclass representatives were recorded (Table A.3). Candidate Class I β subunit proteins were analyzed individually in the same manner, using the β subunits corresponding to the Class I α subclass representatives (Table A.2). P-SSP7 candidate β subunit proteins lacking key residues were removed from the analysis. This left a single candidate β subunit protein (accession no. YP_214198.1). Putative active sites identified in the putative β subunit are recorded in Table A.3.

2.2.3 Phylogenetic analysis

2.2.3.1 Phylogenetic reference sequence curation

Sequences from the RNRdb were used as phylogenetic references. The RNRdb pulls RNRSs from databases including RefSeq [51, 52] and GenBank [53], and includes
RNRs both from cultured and isolated organisms and viruses, as well as RNRs from environmental metagenomic samples. To create a reference sequence set for phylogenetic analyses, all available Class I α (NrdA and NrdE), Class I β (NrdB and NrdF), and Class II (NrdJ) sequences were downloaded from the RNRdb on August 20, 2018 [48]. Sequences were separated into three sets (Class I alpha, Class I beta, and Class II) before sequence curation. Exact and sub-string matches were removed from each set using CD-HIT v4.6 [54, 55]. Sequences were then divided into smaller groups of similar sequences identified by the RNRdb. RNRdb group assignment is based on phylogenetic clade membership [35, 56], so division increased sequence alignment quality. Group names and subclass membership are presented in Table 2.2. RNRdb sequences were aligned individually by group using the MAFFT v7.388 Geneious plug-in on AUTO setting with the BLOSUM62 scoring matrix. Sequence alignments were visualized and edited in Geneious v10.2.4. Inteins within RNRdb sequences were removed manually after the initial alignment step because they are evolutionarily mobile and confound phylogenetic analyses [57, 58]. After intein removal, sequences were realigned and those lacking essential catalytic residues were removed, as they are likely non-functional [16]. Other than the two tyrosine residues involved in Class I radical transport (Y730 and Y731, *E. coli*), the same conserved residues were used for Class I α and Class II sequences (Table A.3). Both intein removal and catalytic residue identification for all groups were done with guidance from the annotated Class I subclass and Class II representatives (Table A.2).

### 2.2.3.2 Reference sequence preparation for phylogeny construction

Broadly, three categories of phylogenies were constructed from protein sequences: (i) Class I α-only, (ii) Class I β-only, and (iii) Class I α with Class II. All phylogenies included Cyano SP clade members (Table 2.1). Class I α and Class II proteins share a common ancestor [23], but are phylogenetically unrelated to Class I β proteins, which belong to the ferritin-like superfamily [59]. Class I α and Class II proteins also share a common catalytic mechanism and several active sites, but are divergent enough that
Table 2.2: RNR Class I subclass membership of RNRdb groups.

<table>
<thead>
<tr>
<th>Class I subclass</th>
<th>RNRdb groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>NrdABe, NrdABg</td>
</tr>
<tr>
<td>Ia (presumed)*</td>
<td>NrdABh, NrdABk, NrdAm, NrdABn, NrdAq, some NrdABz (NrdABza)</td>
</tr>
<tr>
<td>Ib</td>
<td>some NrdEF (NrdEFb)</td>
</tr>
<tr>
<td>Ic</td>
<td>some NrdABz (NrdABzc)</td>
</tr>
<tr>
<td>Id</td>
<td>NrdABI</td>
</tr>
<tr>
<td>Ie</td>
<td>some NrdEF (NrdEFe)</td>
</tr>
</tbody>
</table>

*Ia (presumed) includes all groups without biochemically characterized members.

full-length protein sequences from both classes cannot be presented on the same phylogeny [30]. Thus, Class I \( \alpha \) and Class II protein sequences in this analysis were trimmed to a previously defined region of interest that excluded regions not shared between the two groups (N437-S625, *E. coli* CQR81730.1) [16]. The Class I alpha-only phylogeny allowed for greater resolution, as the phylogeny could be based on a longer protein sequence segment, being trimmed only before C225 in *E. coli* (CQR81730.1). Class I \( \beta \) sequences were trimmed to the region between W48 and Y356 (*E. coli*, KXG99827.1). For Class I alpha-only and Class I beta-only phylogenies, sequences were trimmed near the N-terminus to exclude evolutionarily mobile ATP cone domains [24]. Class I \( \beta \) sequences were also trimmed near the C-terminus to exclude any fused glutaredoxin domains [60]. In all cases, trimming was guided by annotated Class I (a-e) or Class II subclass (mono- or dimeric) representatives (Table A.2).

In addition to trimming, sequences were clustered prior to phylogenetic analysis, as each group contained a large number of sequences (after curation, Class I alpha: 15,894 sequences, Class I beta: 17,109 sequences, and Class II: 9,147 sequences). To avoid inter-group mixing within individual sequence clusters, sequences were clustered by RNRdb group (Table 2.2). Clustering of RNRdb sequences was performed at multiple identity thresholds (70%, 75%, and 80%) using CD-HIT v4.7 to ensure that the placement of the Cyano SP clade was not an artifact of the identity threshold, as
Cyano SP members have grouped with Class II sequences in the past [16]. Cyano SP sequences were not clustered before phylogenetic analysis. For Class I alpha-only and beta-only phylogenies, sequences were clustered over 80% of the alignment length. For the Class I α with Class II phylogeny, sequences were clustered over 100% of the alignment length due to the short length of the trimmed region.

Two RNRdb groups, NrdABz and NrdEF, contained member sequences belonging to two Class I subclasses (Table 2.2). In these cases, the Class I β sequences (NrdBz and NrdF) were assigned to subclasses based on active sites. For NrdBz, Class I β subunit enzymes were classified as subclass Ia (NrdBza) by the presence of a Tyr residue in the Tyr radical site (Tyr122 in E. coli R2), or as subclass Ic (NrdBzc) by the presence of a Phe, Leu, or Val mutation in the Tyr radical site [48]. For NrdF, Class I β subunit enzymes were classified as subclass Ib (NrdFb) or Ie (NrdFe) if carboxylate residues were conserved or missing, respectively, from the second, fourth, and fifth metal-binding sites in relation to the subclass Ib representative (Table A.2). Class I α sequences from clades NrdAz and NrdE were assigned to a subclass based on the assignment of their corresponding β subunits, as they contain more than one subclass and could not be assigned to subclasses based on primary sequence of the α subunit alone. Class I α subunit sequences that were not able to be paired with a β subunit, or that were paired with more than one β subunit, were excluded from further analysis. Pairing was done using string matching of sequence headers. Excluded Class I α subunit sequences included 1006 NrdAz and 2921 NrdE sequences, or 31% and 45% of total curated NrdAz and NrdE sequences, respectively. The excluded sequences comprised a small percentage of overall RNR diversity (Appendix Table A.1). Thus, their exclusion is not expected to have affected the phylogenetic analyses (Appendix Table A.1). All other RNRdb groups exclusively belonged to a single subclass.

2.2.3.3 Phylogenetic tree construction

For all phylogenetic analyses, cluster representatives were aligned with correspondingly trimmed α or β subunits from the Cyano SP clade. All alignments were
constructed in Geneious using the MAFFT v7.388 plug-in with setting FFT-NS-2 (fast, progressive method) and the BLOSUM62 scoring matrix. Trees were constructed using the FastTree v2.1.5 [61] Geneious plug-in with default settings. Trees were visualized and customized in Iroki [62]. Phylogenies inferred from sequences clustered at different identity thresholds can be found in Appendix A (Appendix Figures A.1, A.2, and A.3).

Finally, a phylogeny was constructed from trimmed Class I $\alpha$ subunit and Class II sequences from only cyanobacteria and cyanophage. No clustering was performed. The phylogeny was constructed as described above from an alignment done using the MAFFT v7.388 plug-in with setting FFT-NS-ix1000 (iterative refinement method with 1000 iterations).

### 2.2.4 Sequence similarity network

A protein sequence similarity network (SSN) was constructed with the same RNR Class I $\beta$ subunit sequences used for phylogenetic analysis. The SSN was generated with the Enzyme Similarity Tool (EFI-EST) [63] as in Rose et al. (E-value: 5, fraction: 1, minimum alignment score: 90) [34]. As the full network was too large to visualize in Cytoscape [64, 65], the 90% identity representative node network was used (i.e., each node in the network contained sequences that shared at least 90% amino acid identity).

### 2.3 Results

*Prochlorococcus* phage P-SSP7 is a cyanopodovirus that infects the marine cyanobacterium *Prochlorococcus marinus* subsp. pastoris str. CCMP1986 [66]. The RNR discovered in P-SSP7 was initially annotated as Class II based on the apparent lack of a Class I $\beta$ subunit in the phage genome. The RNR from P-SSP7 also lacks an ATP cone region, a domain that is common in Class I $\alpha$ subunits but rare in Class II enzymes [24, 67]. This was the first cyanophage RNR of its kind to be annotated, and consequently became the baseline annotation for closely related RNRs. Prior examination of RNRs in viral shotgun metagenomes (viromes) designated the phylogenetic
clade containing the RNR from P-SSP7 as the ‘Cyano II’ clade, recognizing that member RNRs (Table 2.1), exclusively from cyanophage, were annotated as Class II and seemed to fall on the Class II side of the tree [16]. This study also recognized a ‘Cyano I’ clade composed exclusively of cyanomyoviruses that carried Class I RNRs [16]. The Cyano II clade has been renamed to Cyano SP, as the clade is comprised solely of RNRs from cyanosiphoviruses and cyanopodoviruses. The Cyano I clade has been renamed to Cyano M, as it consists of RNRs strictly from cyanomyoviruses.

2.3.1 P-SSP7 Class I \(\alpha\) subunit identification

The first indication that the RNR from P-SSP7 was misannotated as a Class II RNR came from the observation of two consecutive tyrosine residues (Y730 and Y731 in \(E.\ coli\)) that are present in the C-terminus of Class I \(\alpha\) subunits and participate in long-range radical transport between the \(\alpha\) and \(\beta\) subunits of Class I RNRs [68, 69]. These tyrosines are not present in Class II RNRs but are present in the P-SSP7 RNR peptide (Table A.3). To confirm the classification of the P-SSP7 RNR as a Class I enzyme, a phylogenetic tree was constructed containing Class I \(\alpha\) subunits and Class II sequences from the RNRdb, together with the putative \(\alpha\) subunits from the Cyano SP clade (formerly Cyano II) reported in Sakowski et al. [16] (Figure 2.1). Trees were constructed at different clustering identities to ensure that the placement of Cyano SP sequences with a given RNR class was not an artifact of the clustering threshold (Figure A.1). The Cyano SP RNRs grouped with the Class I \(\alpha\) subunit sequences in the phylogenies constructed from sequences clustered at 75% and 80% identity, but clustered with Class II sequences in the tree made from sequences clustered at 70% identity.

2.3.2 P-SSP7 Class I \(\beta\) subunit identification

While the tyrosine residues within the P-SSP7 RNR are indicative of a Class I RNR, the initial annotation of the P-SSP7 RNR was made primarily because no \(\beta\) subunit gene could be identified within the P-SSP7 genome. Class I RNRs require a
Figure 2.1: Cyano SP α subunits group with Class I RNRs. Maximum-likelihood phylogenetic tree of Cyano SP clade α subunits with 80% clustered Class I α and Class II RNRdb sequences trimmed to a region of interest. Gray branches belong to Class II. Colored branches belong to one of the five Class I subclasses, or Cyano SP as indicated in the key. Light purple branches indicate RNRdb groups without characterized members, which are assumed to be subclass Ia enzymes. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.
β subunit for radical generation. Because the cyanobacterial host of P-SSP7 carries a Class II RNR, the phage would have to carry its own copy of the Class I β subunit gene in order for its α subunit to function. All unannotated proteins in the P-SSP7 genome approximately the length of a Class I β subunit in the P-SSP7 genome were considered RNR β subunit candidates. Four predicted proteins within the genome matched this length criteria. A batch CD-Search [49] of the candidate β subunit peptide sequences was unable to identify any conserved domains in any of the sequences. Thus, we aligned the candidate P-SSP7 β subunit sequences with the sequences of biochemically characterized β subunits from each of the known Class I subclasses (Table A.2). Only one of the candidate sequences, accession no. YP_214198.1, was found to contain residues experimentally shown to be required for β subunit function (Table A.3). The hypothetical protein also resided directly downstream of the α subunit, where the β subunit is typically found [17]. Thus, YP_214198.1 was identified as the missing P-SSP7 β subunit.

2.3.3 Assignment of P-SSP7 RNR to a Class I subclass

Class I subclass divisions are based on the mechanism of radical generation utilized by the β subunit. Alignment with representative Class I RNR β subunit sequences found that the P-SSP7 β subunit lacked the tyrosine residue (Y122 in E. coli R2) on which the stable protein radical is formed in subclasses Ia, Ib, and Ie (Figure 1.1B). The lack of the tyrosine residue seemed to indicate that the P-SSP7 β subunit belonged to subclass Ic, as Ic is the only described subclass that lacks this residue completely (the residue is conserved in Id but does not harbor a radical) [33, 34, 70]. Each subclass has a unique combination of metal-binding residues and uses a different metallocofactor (or does not bind metals at all, in the case of subclass Ie) [33]. The residues in the putative P-SSP7 β subunit aligning with the first sphere of metal-binding residues of the subclass representatives (Table 2.3) were consistent with Class I RNRs that require metallocofactors (subclasses Ia-Id) and exactly matched subclasses Ic and Id [33]. However, when considering second sphere binding residues,
the overall pattern of metal-binding residues in the P-SSP7 β subunit did not match that of any subclass representative (Table 2.3), nor of any existing RNRdb group (Table 2.4).

Known Class I subclasses are either monophyletic or contain members that are closely related [35, 56]. Thus, phylogenetic trees were constructed to confirm proper subclass assignment of the P-SSP7 RNR using Class I β subunit sequences from the RNRdb clustered at 70%, 75%, and 80% and β subunits from the Cyano SP clade members. In a phylogenetic analysis of the 70% identity cluster representative sequences, the P-SSP7 β subunit and Cyano SP homologs were phylogenetically distinct from known RNRs, and did not clearly join with RNRdb groups, instead branching directly off the backbone of the tree (Figure 2.2). In the phylogenetic reconstructions at 75% and 80% identity, the Cyano SP group remained distinct but branched closely with either the NrdBg group (75% identity, subclass Ia) or the NrdBh group (80% identity, subclass Ia presumed) (Appendix Figure A.2). Notably, the Cyano SP β subunits branched away from subclass Ic members (NrdBzc subgroup) in all phylogenies (Appendix Figure A.2), making it unlikely that the Cyano SP clade belongs to subclass Ic.

Because Class I subclass assignment was inconclusive based on the β subunit metal-binding residues and phylogenetic analysis, we constructed a protein sequence similarity network (SSN) using the Enzyme Similarity Tool (EFI-EST) [63] as per Rose et al. [34] with the same β subunit sequences used for phylogenetic tree construction (Figure 2.3). The SSN also provided an alignment-free method for viewing connections between RNR sequences [63]. Most sequences were members of large, distinct subgraphs with sequences exclusively from a single RNRdb group (e.g., NrdBk and NrdBg). However, some RNRdb groups were evenly spread across multiple subgraphs of similar size (e.g., NrdBh and NrdBi), likely indicating a higher level of sequence heterogeneity than other groups. The Cyano SP clade representatives formed exclusive subgraphs not connected to other RNRdb sequences and were divided into three singleton and one non-singleton cluster, indicating that the clade representatives are
### Table 2.3: Metal-binding amino acid residues in each of the β subunit references and P-SSP7.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Subclass</th>
<th>First Sphere</th>
<th>Second Sphere</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>Ia</td>
<td>1 2 3 4 5 6 7 8</td>
<td>S15 D238</td>
</tr>
<tr>
<td><em>S. typhimurium</em></td>
<td>Ib</td>
<td>1 2 3 4 5 6 7 8</td>
<td>M97 D191</td>
</tr>
<tr>
<td><em>C. trachomatis</em></td>
<td>Ic</td>
<td>1 2 3 4 5 6 7 8</td>
<td>E119 D226</td>
</tr>
<tr>
<td><em>F. johnsoniae</em></td>
<td>Id</td>
<td>1 2 3 4 5 6 7 8</td>
<td>C96 D194</td>
</tr>
<tr>
<td><em>A. urinae</em></td>
<td>Ie</td>
<td>1 2 3 4 5 6 7 8</td>
<td>M115 D209</td>
</tr>
<tr>
<td>P-SSP7</td>
<td>Cyano SP</td>
<td>1 2 3 4 5 6 7 8</td>
<td>D70 D146</td>
</tr>
</tbody>
</table>
Table 2.4: Metal-binding amino acid residues in each of the RNRdb groups and the Cyano SP clade. 
RNRdb groups are based on phylogenetic clades.

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Clade</th>
<th>First Sphere</th>
<th>Second Sphere</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 2 3 4 5 6</td>
<td>7 8</td>
</tr>
<tr>
<td>Ia</td>
<td>NrdBe</td>
<td>D E H E E E H</td>
<td>M/I/V D</td>
</tr>
<tr>
<td>Ia</td>
<td>NrdBg</td>
<td>D E H E E E H</td>
<td>S D</td>
</tr>
<tr>
<td>Ia (presumed)</td>
<td>NrdBh</td>
<td>D E H E E E H</td>
<td>E/Q D</td>
</tr>
<tr>
<td></td>
<td>NrdBk</td>
<td>D E H E E E H</td>
<td>M/R/I D/E</td>
</tr>
<tr>
<td></td>
<td>NrdBn</td>
<td>D E H E E E H</td>
<td>E D</td>
</tr>
<tr>
<td></td>
<td>NrdBza</td>
<td>D E H E E E H</td>
<td>E D</td>
</tr>
<tr>
<td>Ib</td>
<td>NrdFb</td>
<td>D E H E E H</td>
<td>M D</td>
</tr>
<tr>
<td>Ic</td>
<td>NrdBzc</td>
<td>E E H E E E H</td>
<td>E D</td>
</tr>
<tr>
<td>Id</td>
<td>NrdBi</td>
<td>E E H E E E H</td>
<td>C/S D/E</td>
</tr>
<tr>
<td>Ie</td>
<td>NrdFe</td>
<td>D Q/V H S/P K H</td>
<td>M D</td>
</tr>
<tr>
<td>Cyano SP</td>
<td>Cyano SP</td>
<td>E E H E E H</td>
<td>D D</td>
</tr>
</tbody>
</table>
Figure 2.2: Cyano SP $\beta$ subunits show no close relationship to known Class I clades. Cladogram of near full-length Cyano SP and 70% clustered RNRdb Class I $\beta$ subunit sequences. Branch colors indicate Class I subclass and leaf dot colors correspond to RNRdb group. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.
divergent even from each other.

Assignment of the Cyano SP RNRs to an existing Class I subclass could not be reliably made based on the analysis of $\beta$ subunit metal-binding residues, phylogenies, or the protein SSN. Instead, the missing tyrosine radical residue, unique pattern of metal-binding sites, and phylogenetic divergence of the Cyano SP $\beta$ subunits from RNRdb groups likely indicate that the Cyano SP clade represents a novel Class I subclass.

### 2.3.4 Origin of the P-SSP7 RNR

Class I $\alpha$ and $\beta$ subunits tend to evolve in units, producing highly similar phylogenies [17, 30]. Because placement of the Cyano SP $\beta$ subunits on phylogenetic trees changed with the percent amino acid identity used for clustering RNR sequences (Appendix Figure A.2), the Cyano SP $\alpha$ subunits were evaluated for clues to the origin of the RNR in P-SSP7. Class I $\alpha$-only phylogenies were built from sequences longer than those used for the combined Class I $\alpha$-Class II phylogenies, allowing greater phylogenetic resolution. Representative RNRdb Class I $\alpha$ subunit sequences from 70%, 75%, and 80% identity clusters were assessed. Regardless of the clustering identity, the Class I $\alpha$ subunit phylogenies showed consistent placement of the Cyano SP clade as an outgroup for the branch that contains RNRdb groups NrdAi (subclass Id) and NrdAk (subclass Ia presumed) (Figure 2.4 and Appendix Figure A.3). Like the Class I $\beta$ phylogenies, the Cyano SP $\alpha$ subunit clade was distinct and was not surrounded by any RNRdb group. The phylogenetic placement of the Cyano SP Class I $\alpha$ sequences among RNRdb groups (Figure 2.4 and Appendix Figure A.3) was different from that seen for the Cyano SP Class I $\beta$ sequences (Figure 2.2 and Appendix Figure A.2). Thus, a conclusive placement for the Cyano SP $\beta$ subunits among RNRdb groups was not possible.
Figure 2.3: Cyano SP RNRS cluster separately from known RNRS. Protein sequence similarity network of the Cyano SP clade and all RNRdb Class I β subunit sequences included in phylogenetic analysis. Nodes represent sequence clusters ≥ 90% similarity. Nodes are colored based on RNRdb group and match leaf dot colors on the cladogram in Figure 2.2. Edges connect nodes with minimum alignment score ≥ 90. Network was visualized and customized in Cytoscape.
Figure 2.4: Cyano SP α subunits are most closely related to RNRdb clades NrdAk and NrdAi. Cladogram of near full-length Cyano SP and RNRdb Class I α subunit sequences clustered at 80%. Branch colors indicate Class I subclass and leaf dot colors correspond to RNRdb group. Colors matching to clades in Figure 2.2 indicate α/β subunit pairs. Note there are α subunit clades that do not have corresponding, distinct β subunit clades, as the α subunits have diverged more than the β subunits. NrdAm α subunits pair with β subunit group NrdBh. NrdAq α subunits pair with β subunit subgroup NrdBza. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.
2.4 Discussion

2.4.1 The Cyano SP RNR has adapted to the intracellular environment

The perceived lack of a β subunit gene in the P-SSP7 genome may have led to the initial misannotation of the P-SSP7 RNR gene as a Class II RNR [66]. Additionally, it seems unusual for a virus to carry a different class of RNR than its host [17]. Given that cellular organisms carry RNRs that are adapted to their environmental niche [19, 20], viruses would also likely benefit from having the same RNR type as their host cell. The preference for a potentially iron-dependent Class I RNR enzyme among cyanophage seems puzzling considering that iron is often the primary limiting nutrient in the oceans, including in regions dominated by *Synechococcus* and *Prochlorococcus* [71, 72]. *Synechococcus* and *Prochlorococcus*, hosts infected by phage within the Cyano SP (cyanosipho- and cyanopodoviruses) (Table 2.1) and Cyano M (cyanomyoviruses) clades, are some of the few B$_{12}$ producers in the oceans [73, 74]. Therefore, B$_{12}$ availability would seem to be sufficient for viral replication with a B$_{12}$-dependent Class II RNR, while iron availability for phage-infected cells could be too low to support the highly lytic phenotype displayed by many of these phage.

However, carrying a Class I RNR would relieve marine cyanophage of their dependence on the host to produce sufficient levels of B$_{12}$ for deoxyribonucleotide synthesis by a Class II enzyme. Although it is less limiting in ocean waters, B$_{12}$ is likely to be more limiting than iron inside a cyanobacterial cell. In Cyanobacteria, B$_{12}$ is used as a cofactor for two enzymes, the Class II RNR (NrdJ) and methionine synthase (MetH) [73]. NrdJ is needed only while the cell is actively replicating, thus, transcription of this gene is closely tied with the cell cycle [75, 76]. Similarly, MetH expression is high during early growth of the B$_{12}$-producing cyanobacterium *Synechocystis* but decreases when cells enter the stationary growth phase [77]. Given that NrdJ and MetH are both tied to cellular growth, intracellular B$_{12}$ concentrations are likely highly variable.

In addition, cobalt, the metal at the center of B$_{12}$, is required almost exclusively for B$_{12}$ formation and is tightly controlled because of its toxicity to cells [78, 79]. In contrast, both iron and manganese are required for numerous proteins and molecules
within a cyanobacterial cell that are needed throughout the cell cycle [80, 81]. Cytoplasmic cyanobacterial iron and manganese quotas have been documented at $10^6$ atoms/cell [82, 83] and a study that aimed to identify and quantify metals in a cyanobacterium found that iron was present in high intracellular concentrations, while cobalt concentrations were below the detection limit [84]. Furthermore, some Prochlorococcus are able to maintain growth while up-taking just one atom of cobalt per cell per hour [85]. Therefore, upon infection, a cyanophage would encounter an intracellular pool of iron many fold larger than that of B$_{12}$.

The acquisition of B$_{12}$ from the surrounding environment also seems unlikely. B$_{12}$ is bulky and structurally complex, requiring special transporters which neither Prochlorococcus, Synechococcus, nor their phages are known to encode [86, 87, 88]. Furthermore, one study showed that while some organisms, such as eukaryotic microalgae, are able to import partial or finished forms of B$_{12}$, Synechococcus and likely Prochlorococcus are unable to do this [74]. Instead, Synechococcus is required to synthesize B$_{12}$ start to finish [74], likely because both Prochlorococcus and Synechococcus produce a form of B$_{12}$ that seems to be unique to Cyanobacteria [73].

Finally, B$_{12}$ is energetically expensive to synthesize. B$_{12}$ synthesis requires a long pathway made up of roughly twenty different enzymes [89]. By comparison, some Class I RNR metallocofactors are known to self-assemble [19]. At most, a metallocofactor may require a flavodoxin (NrdI) for assembly [33]. When considering that carrying a Class I enzyme relieves the phage of relying on a complex host-mediated pathway for a molecule that is not consistently produced throughout the cell cycle, the difference in RNR type between host and phage is not surprising.

The RNR from P-SSP7 also seems to have adapted to the environment inside the host cell in other ways. The P-SSP7 $\beta$ subunit lacks the tyrosine residue used for radical generation in most Class I RNR subclasses (Figure 1.1B). The tyrosine residue harbors a stable protein radical and is a target of nitric oxide [90, 91]. Tyrosine-radical scavenging nitric oxide is hypothesized to be present inside Synechococcus cells as an intermediate in nitrate reduction [92], which is widespread among freshwater and
marine *Synechococcus* species and is coupled to photosynthesis [93, 94, 95, 96]. Thus, the loss of the tyrosine radical site in the Class I β subunit genes of cyanophage, such as P-SSP7, would enable these phage to avoid RNR inactivation by nitric oxide.

### 2.4.2 Connections between RNR and cyanophage phenotype

Most marine cyanophage belong to the Cyano M and Cyano SP clades. The Cyano M clade consists of subclass Ia RNRs belonging solely to cyanomyoviruses. The Cyano SP clade consists of RNRs from the proposed novel subclass If belonging solely to cyanosipho- and cyanopodoviruses. This pattern also extends past the Cyano M and Cyano SP clades. Thus far, all sequenced Class II RNRs from marine cyanophage belong only to the P60 clade, which contains only cyanosipho- and cyanopodoviruses. Additionally, Cyanophage S-TIM5, a myovirus, carries a subclass Id RNR, a subclass not carried by cyanosipho- or cyanopodoviruses. In this way, RNR type is predictive of cyanophage morphology.

Most Class I RNR α subunits contain an ATP cone region. ATP cones are regulatory sites that essentially act as on/off switches for RNRs [24, 97]. When ATP is bound, the RNR holoenzyme enters a conformational state that allows for function [98]. Once dNTP levels rise high enough, dATP binds the ATP cone and the holoenzyme enters a non-functional conformation [98, 99]. Intriguingly, the Class I α subunits of the Cyano SP clade do not have ATP cones. This is unusual for Class I α subunits and likely represents an evolutionary loss, given that only two Class I α subunit clades (NrdAi/NrdAk and NrdEb/NrdEe) (Figure 2.4) lack ATP cones [24, 67]. In losing the ATP cone domain, the Cyano SP RNRs have lost this regulatory switch. As a consequence, the RNR of cyanopodo- and cyanosiphoviruses cannot be inactivated through dATP binding, which would be beneficial to a fast-replicating lytic phage [100].

The highly lytic nature of the Cyano SP clade is also reflected in the biochemistry of the family A DNA polymerase gene (*polA*) carried by some members of the clade (Appendix Table A.4). The amino acid residue at position 762 (*E. coli* numbering) plays a role in shaping the activity and fidelity of Pol I (*polA* peptide) and
is hypothesized to be reflective of phage lifestyle [13]. Prior work found that a mutation from phenylalanine to tyrosine at position 762 produced a 1,000-fold increase in processivity with a concomitant loss of fidelity [101]. Three of the member phages within the Cyano SP clade carry a Pol I with a tyrosine at position 762, indicating that Cyano SP members are capable of fast DNA replication. Other members carry polA genes that contain a frameshift mutation, preventing identification of the 762 position. Pairing an unregulated RNR, such as the Cyano SP RNR, with a highly processive DNA polymerase would be advantageous for a highly lytic phage. This phenotype is thought to be characteristic of most cyanopodoviruses [13, 102, 103]. Observations of gene associations such as Tyr762 Pol I and Cyano SP clade Class I RNR can thus inform predictions of the possible life history characteristics of unknown viruses.

2.4.3 A novel Class I RNR in cyanophage

Reannotation of the P-SSP7 RNR from Class II to Class I is based primarily on the discovery of a Class I \(\beta\) subunit in the P-SSP7 genome. The P-SSP7 \(\beta\) subunit was identified using conserved residues, as no conserved domains could be identified in the previously hypothetical protein. Our discovery of the Class I \(\beta\) subunit via active sites and genome location demonstrates that some unknown viral proteins (i.e., the viral genetic dark matter) [104] could actually be well known proteins that are simply too divergent for annotation using homology searches or gene model approaches.

The reannotation is also supported by the presence of the consecutive tyrosine residues in the C-terminus of the newly annotated Class I \(\alpha\) subunit, which are essential for radical transfer between Class I \(\alpha\) and \(\beta\) subunits [68, 69] and are not found in Class II RNRs. Additionally, two trees constructed from Class I \(\alpha\) and Class II sequences showed the Cyano SP clade (represented by P-SSP7) on the Class I side of the tree (Figure 2.1 and Appendix Figure A.1B). While the 70% Class I \(\alpha\) with Class II tree showed the Cyano SP clade on the Class II side of the tree (Appendix Figure A.1A), we believe this is an artifact of the low identity threshold and short region of interest. Protein SSNs constructed from the same sequences used in the Class I \(\alpha\) with Class II
phylogeny showed the Cyano SP clade as being distinct from both Class I and Class II sequences (Appendix Figure A.4), with all Class I and II RNRs being connected and the Cyano SP clade remaining distinct (data not shown). This pattern also held when adding RNRs mined from the Earth Viromes dataset (data not shown) [105]. Thus, the high divergence of the Cyano SP clade as compared to Class I α and Class II sequences in the RNRdb are likely contributing to the Cyano SP clade grouping with Class II sequences on the 70% tree. A study of gene transcription in P-SSP7-infected Prochlorococcus cultures lends experimental support for the presence of a Class I RNR in P-SSP7. Both the P-SSP7 Class I RNR α subunit (identified as nrd-020) and the neighboring β subunit (identified as nrd-021) were co-expressed during the second stage of phage infection, during which DNA replication typically takes place [106].

Assignment of the P-SSP7 RNR to an existing Class I subclass was inconclusive as the radical-generating β subunit [19] could not be clearly assigned based on conserved residues. β subunits are used for subclassification because, unlike α subunits that all have a consistent mechanism, the mechanisms of β subunits are variable. While the P-SSP7 β subunit contains all of the conserved residues required for function (Table A.3), it lacks the tyrosine residue (Y122 in E. coli) that harbors the stable protein radical or is conserved in subclasses Ia, Ib, Id, and Ie [32, 33, 34, 107] (Figure 1.1B). Assignment also could not be made to subclass Ic, the only known subclass lacking the tyrosine residue [70], based on the outcome of phylogenetic (Figure 2.2 Appendix Figure A.2) and protein SSN analysis (Figure 2.3). We also examined the metal-binding sites in the P-SSP7 β subunit, as metallocofactor identity is used to discriminate between subclasses Ia-Id [32, 34]. The metal-binding residues for the P-SSP7 and other Cyano SP clade member β subunits formed a different pattern than is seen in any of the RNRdb groups (Table 2.4). The combination of the unique metal-binding residues, the lack of a tyrosine residue on which to generate a protein radical, and the phylogenetic distance between the Cyano SP clade and subclass Ic (NrdBzc) sequences, suggest that the P-SSP7 Class I β subunit may constitute a novel subclass of Class I RNRs, the proposed subclass If.
2.4.4 Origin of the P-SSP7 RNR

Because P-SSP7’s host, like most marine *Synechococcus* and *Prochlorococcus*, carries a Class II RNR, we were interested in the origin of the Class I RNR found in P-SSP7. The Class I β subunit phylogenies inconsistently placed the Cyano SP clade. Examination of Class I α subunit trees showed a consistent placement of the Cyano SP clade at the base of the branch harboring the RNRdb groups NrdAk (Ia presumed) and NrdAi (subclass Id) (Figure 2.4 and Appendix Figure A.3). This is perhaps to be expected as, like the NrdAk and NrdAi groups, the Cyano SP Class I α subunits do not contain ATP cone domains, a trait that is rare among Class I α subunits [67].

The observation that the Cyano SP clade does not have the same placement on the Class I β-only and Class I α-only trees is highly unusual. In viruses and cellular organisms, Class I α and β subunits are thought to evolve as units [17], producing trees with the same patterns [30]. However, viral genomes are known to be highly modular, consisting of genes from multiple sources [10, 108]. It seems possible that an ancestral phage of the Cyano SP clade incorporated the Class I α and β subunits separately. Given that Class I α and β subunits can only perform ribonucleotide reduction as a unit, i.e. both subunits are required for functionality, these acquisitions would have had to occur in quick succession to avoid loss by the phage. Perhaps in support of this hypothesis is that the Cyano SP β subunits sometimes cluster with the NrdB group (subclass Ia) which harbors the Cyano M clade, while the Cyano SP α subunits consistently cluster with the NrdAi group (subclass Id) that contains the *Synechococcus* phage S-TIM5. These phage groups (i.e. Cyano SP, S-TIM5, and Cyano M) all infect marine *Synechococcus* and *Prochlorococcus*, making the possibility more likely that the Cyano SP RNRs are a mosaic of these cyanomyoviral groups, with the α subunit having been acquired from a cyanophage related to S-TIM5 and the β subunit from a member of the Cyano M clade.

A phylogeny constructed using all Cyanobacteria and cyanophage present in the RNRdb with the Cyano SP clade demonstrates that the majority of known cyanophage carry Class I RNRs (Figure 2.5). The *Synechococcus* or *Prochlorococcus* hosts of phage
in the Cyano M, Cyano SP clades, *Synechococcus* phage S-TIM5, and the Cyanophage P60 clade all carry Class II RNRs [16, 109, 110]. Despite being a myovirus, S-TIM5 does not carry an RNR belonging to the Cyano M clade, likely because it is believed to represent a separate lineage of myoviruses [109]. Cyanosiphovirus- and cyanopodoviruses were found in two widely separated clades. Lytic cyanosiphovirus- and cyanopodoviruses within the Cyanophage P60 RNR clade contain a Class II RNR, which is the same type carried by their hosts, whereas cyanosiphovirus- and cyanopodoviruses in the Cyano SP clade contain a Class I RNR. The biological and ecological explanations behind this divergence are a mystery. Prior work has indicated that cyanopodoviruses can be broadly divided into two clusters, MPP-A and MPP-B, based on whole genome analyses [111], but no single gene or gene group clearly distinguishes the two clusters. Nevertheless, RNRs belonging to the Cyano SP clade seem to be more common among cyanosiphovirus- and cyanopodoviruses [16, 111]. Whether carrying a Class II RNR is the ancestral state of cyanosiphovirus- and cyanopodoviruses could not be determined from our phylogenies.

The use of marker genes such as RNR in studying viral ecology is important in connecting genomic information to phenotypic traits. However, correct annotation of these genes is essential if accurate information is to be gained. The reannotation also means that most marine cyanophage carry RNRs that did not come from their hosts (Figure 2.5), which has implications for our understanding about the acquisition of nucleotide metabolism genes by viruses. That Cyano SP clade members carry Class I RNRs and have lost the tyrosyl radical site in the \( \beta \) subunit is also a reminder that viruses have to adapt to the intracellular environment as well as the extracellular environment. Finally, the discovery of an overlooked \( \beta \) subunit implies that some unknown viral gene space may be composed of known genes that are too divergent for similarity-based annotation methods to detect but can still be identified by other means.
Figure 2.5: There is evolutionary pressure for cyanophage to carry Class I RNRs. A) Maximum-likelihood phylogenetic tree of Cyano SP clade α subunits with 80% clustered Class I α and Class II RNRdb sequences trimmed to a region of interest. B) Maximum-likelihood phylogenetic tree of a subset of Class I α subunit sequences limited to Cyanobacteria and cyanophage. In both trees, dark green branches indicate Cyanobacteria and light green branches indicate cyanophage. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bars represent amino acid changes per 100 positions.
Chapter 3

RIBONUCLEOTIDE REDUCTASE GENES DRIVE BIOGEOGRAPHICAL DISTRIBUTIONS OF MARINE VIRUSES

3.1 Introduction

The metabolic reprogramming of host cells by viruses is an important factor for biogeochemical cycling in the oceans. For example, the expression of the Photosystem-II encoding gene \textit{psbA} by viruses sometimes surpasses that of Cyanobacteria [112] and some cyanophage inhibit CO$_2$ fixation in hosts during infection [7], affecting the flow of carbon through the ecosystem. Marine viruses frequently encode cell-derived genes that participate in photosynthesis [6, 7, 66, 113], the pentose phosphate pathway [66, 114], phosphate acquisition [115, 116], and nucleotide metabolism [16, 66, 117]. Analysis of marine virome and cyanophage genome data has shown that viral genomes are enriched for genes related to nucleotide biosynthesis [114, 118, 119]. This has been proposed to be especially important for viruses in nutrient-limited environments where hosts may not have large dNTP pools [120, 121], making nucleotide synthesis limiting for viral replication [114, 122].

In this context, the role of RNR in viral genomes is invaluable, as it catalyzes the final and rate-limiting step of dNTP synthesis [14, 15]. Thus, to keep dNTP production running efficiently during infection, viral RNRs must be provided with their cofactors. All RNRs, excepting subclass Ie [33], require trace metals either as cofactors or as metal centers within their cofactors [18, 19, 34, 123]. In the oceans, trace metals are often limiting nutrients [72, 124, 125]. Most trace metals have different distributions in the ocean with regard to depth and geographical location because they have different input sources and undergo different chemical reactions [124, 125]. Because RNRs are of high importance to viral fitness and are reliant upon differentially distributed nutrients, we
hypothesized that marine viruses carrying different types of RNR would show different biogeographical patterns and that these patterns would mirror those of the trace metals required by RNRs.

The potential connection between RNR type and ecological distributions of viral populations make RNR an intriguing marker gene candidate. Currently, most marker genes target specific viral populations [126, 127, 128]. However, RNR is known to be present in the genomes of diverse viruses [10, 16], meaning they have the potential to be informative on the community rather than population level. RNRs are also indicative of phage biology, as they are strongly tied to lytic phage [16]. Finally, RNRs are of interest to many fields outside of viral ecology, including evolutionary biology [30, 129] and cancer research [18], meaning that they are abundant in sequencing databases and are well represented in the literature. Therefore, we believe that RNR is a valuable marker gene for environmental viral populations.

To test our hypothesis about RNR subclass driving ecological distributions of viruses and show the utility of RNR as a marker, we mined viral RNRs from the Global Ocean Viromes (GOV) 2.0 dataset [130]. The GOV 2.0 dataset contains 145 viromes from globally distributed marine sites (Figure 3.1). GOV 2.0 is accompanied by extensive metadata, making it useful for ecological study. Additionally, the total viral community from GOV 2.0 has previously been analyzed [130]; thus, we can compare results from the RNR-containing virus subset to the previous total community analysis.

3.2 Methods
3.2.1 GOV 2.0
3.2.1.1 Overview

GOV 2.0 is a dataset of 145 viromes from globally distributed marine sampling sites [130]. The viromes originate from three sampling expeditions: Malaspina 2010 Circumnavigation Expedition (Malaspina, [131]), Tara Oceans [132, 133], and Tara Oceans Polar Circle (TOPC, Gregory et al., 2019). GOV 2.0 expands upon the previous
Figure 3.1: GOV 2.0 samples are from globally-distributed marine sites. Point color indicates sampling expedition. Most sites were sampled at multiple depths.
GOV dataset [134] by including TOPC samples and additional mesopelagic samples from Tara Oceans [130].

3.2.1.2 Sample collection and sequencing

The Malaspina viromes [131] consist of fourteen viral samples collected between April and July 2011 from the tropical and subtropical Atlantic and Pacific oceans [134]. Thirteen samples were taken from the bathypelagic and one from the mesopelagic. All bathypelagic samples were collected at 4,000 m depth excepting sites 81 and 82, where samples were collected at 3,500 m and 2,150 m, respectively. The mesopelagic sample was collected at 294 m depth and targeted the oxygen minimum zone (OMZ) portion of the water column.

Tara Oceans and TOPC samples were collected from three depths: surface, deep chlorophyll maximum (DCM), and mesopelagic. Surface samples were all collected at 5 m depth. DCM samples were not taken at a consistent nominal depth (17-177 m). Instead, the DCM was identified at each sampling site using a Rosette Vertical Sampling System (RVSS) and then sampled [133]. When no DCM was observed, samples were collected at the bottom of the mixed layer (stations 123, 124, and 125; 100-150 m). Mesopelagic samples were collected between 250 m and 1,000 m depth. Additional samples were occasionally collected when environmental features of interest were encountered (i.e., OMZs) and are included in the mesopelagic group. The Tara Oceans expedition collected 90 viral samples from 45 sampling sites between October 2009 and December 2011. The TOPC expedition collected 41 viral samples from 20 sites between June and December 2013.

Sample collection and storage followed the same protocols, with few exceptions, for Malaspina [135], Tara Oceans [134, 136], and TOPC [130] viromes and a previous study found that variation between Malaspina and Tara Oceans samples was likely due to ecological differences rather than methodological artifact from sequencing differences [130]. Briefly, following previous filtration steps, 80 L (Malaspina) or 20 L (Tara Oceans and TOPC) of seawater were filtered through a 0.22 μm filter. Viruses were
concentrated from the filtrate using iron chloride flocculation. Viral concentrates were treated with DNase I to remove free DNA. DNA was extracted using phenol/chloroform [137] and sequenced on Illumina HiSeq 2000 using 150bp, paired end (Malaspina) or 101bp, paired end (Tara Oceans and TOPC) protocols.

3.2.1.3 Sequencing analysis

Details of the analysis of GOV 2.0 sequencing data can be found in Gregory et al. [130]. Briefly, sequences were quality controlled and assembled by sample. Contigs ≥ 1.5kb were entered into VirSorter [138] and VirFinder [139]. Contigs that mapped to the human, dog, or cat genome were removed. Contigs used for further analysis included those ≥ 5kb and linear, or ≥ 1.5 kb and circular that were verified as viral by both VirSorter and VirFinder and that had < 40% of their genome classified as cellular by CAT [140]. To create viral populations, contigs were clustered at 95% identity over 80% of the contig length. This resulted in 195,728 viral populations with representative contigs ≥ 10kb. Population abundances were determined for each sample by mapping the raw virome reads to the viral populations. Contigs with < 5kb coverage were removed. Average read depth was calculated for each population and used as a proxy for abundance.

3.2.2 RNR mining

ORFs were predicted and translated for viral population contigs ≥ 10kb using Prodigal v2.6.3 (default settings with option -meta) [141]. Class I α and Class II RNRs were mined from the translated ORFs using an in-house database, RNR Gold, as the query database. RNR Gold is composed of Class I α and Class II RNR sequences from the RNRdb [48], Viral RefSeq [52, 51], UniProt50 [47], Tara Oceans [136], MetaGenomes OnLine (MgOL, http://metagenomesonline.org/), and the Earth Viromes [105], and is used to identify RNRs from environmental datasets. Details on the construction of RNR Gold can be found in Appendix B. Only Class I and Class II RNRs were collected because Class III RNRs are rare in ocean waters due to their O₂
sensitivity. Class I $\beta$ subunits were not mined because they make virtually identical trees as compared to Class I $\alpha$ subunits (see Appendix C) [30, 36]. Class I $\beta$ subunits are also more difficult to identify because they are shorter and are part of the ferritin-like superfamily, which has many members [59]. Class I $\alpha$ subunits and Class II sequences, on the other hand, do not have any close protein relatives [23].

To maximize recovery of GOV 2.0 Class I $\alpha$ and Class II RNRs, iterative homology searches were performed using MMSeqs2 v6915b7875dfe819c3d6aaa8ff2304f60111dec42 easy-search with the highest sensitivity setting (-s 7) [142]. After each search, protein sequences were validated as Class I $\alpha$ or Class II RNRs using conserved active sites (Table A.3). Validated sequences were then added to the query database and another homology search was performed using the updated queries. Iterations were performed until no new sequences passed the filter (3 iterations). All homology searching steps were performed automatically using the MMSeqs2 Iterator v0.3.2 (https://github.com/mooreryan/mmseqs_iterator) with filtering enabled.

Protein validation based on active site recognition was performed using the Protein Active Site Validator (PASV) v1.1.11 (https://github.com/mooreryan/pasv). Details on the function of PASV can be found in Appendix B. Briefly, PASV individually aligns a query against a set of user-provided alignment references and records the amino acid residue at each user-specified key position (site of conserved residues in the first reference protein). For this study, 23 alignment references were used, which included a mix of Class I $\alpha$ and Class II RNR sequences, with *Escherichia coli* Class Ia RNR as the primary alignment reference. References were chosen by randomly selecting a protein sequence from each Class I $\alpha$ and Class II RNRdb clade [48] and the Cyano SP clade [36]. All PASV references were trimmed to the ROI except for the primary alignment reference (*E. coli* subclass Ia) which was left as full-length. The reference sequences were tested by running PASV with manually validated RNRdb sequences [36] as queries. All manually validated RNRdb sequences passed the PASV filter. When using the PASV setting in the MMSeqs2 Iterator, the user also provides the sequence of residues expected for a functional protein. The optional PASV setting
requiring a query to span a region of interest (ROI) was also enabled in this case. Query protein sequences that both spanned the region of interest and included all required active site residues were added to the query database for the next homology search iteration. The expected residue sequence was Asp, Cys, Glu, Cys, Pro (NCECP) using positions 437, 439, 441, 462, and 621 and the ROI spanned positions 437-625 in *E. coli*.

### 3.2.3 RNR curation, typing, and phylogenies

Putative GOV 2.0 Class I α and Class II RNR protein sequences were imported into Geneious v10.2.6 (https://www.geneious.com) for manual curation. Sequences were aligned with annotated references, separate from those used for PASV, consisting of one sequence from each Class I and Class II subclass (Table A.2). using the MAFFT v7.388 Geneious plug-in [50] on the AUTO setting with the BLOSUM62 scoring matrix. Key active site residues (the same used in PASV for validation) were visually inspected, as occasional false positive proteins, unrelated to RNR, pass the PASV filter due to random alignment. Sequences without the proper residues were discarded. If present, inteins were manually removed from sequences. Residues aligning at position 225 in *E. coli* were also inspected and sequences without a Cys residue in this position were discarded. This residue was not included in the PASV profile because it does not align well when Class I α and Class II RNRs are combined. This process ensured that any viral RNRs analyzed were likely to be functional during viral infection. Any viral population contigs with more than one RNR were removed from analysis after manual validation of active sites. Such contigs were identified using string matching of sequence headers.

Sequences passing the first manual round of validation were then trimmed to a region of interest (ROI). The ROI extends from residue 225 to 759 in *E. coli* and includes all RNR active sites and catalytic residues [36]. This ROI allows for precise placement on a phylogenetic tree while also not requiring the protein sequence to be absolutely complete, which means that more sequences can be kept for analysis. Additionally, this ROI also excludes domains such as ATP cones [24, 67] or glutaredoxins.
that may be fused to the N-terminus or C-terminus of RNR protein sequences and are neither evolutionarily informative nor consistently present.

Trimmed sequences were divided by class (Class I α or Class II) using the residue corresponding to position 438 in E. coli before a final validation of key functional residues. Sequences with a Leu, Ile, Val, or Met residue in this position were sorted as Class I. Sequences with a Pro or Ala residue were sorted as Class II. Subsequent phylogenetic analyses confirmed these divisions. For Class I α sequences, residues aligning at E. coli positions 730 and 731 were checked for the presence of Tyr residues. The Tyr residues are involved in radical transport with the β subunit and are therefore not present in Class II RNRs [69]. The Class I α and Class II sequences were checked for C-terminal Cys residues aligning with E. coli RNR positions 754 and 759 or L. leichmanii RNR residues 731 and 736, respectively. In both classes, C-terminal cysteines interact with reducing agents, and proteins lacking these residues are incapable of enzymatic turnover [143, 144, 145]. The C-terminal regions of Class I α and Class II RNRs do not align, so PASV could not be used for these residues. The mined, trimmed, and validated Class I α and Class II GOV 2.0 sequences were then aligned with trimmed reference Class I α or Class II sequences from the RNRdb [48] using the MAFFT v7.388 Geneious plug-in on the FFT-NS-ix2 (iterative refinement method with 2 iterations) with the BLOSUM62 scoring matrix. Details on curation, trimming, and clustering of RNRdb sequences can be found in the methods section of Chapter 2 (Section 2.2). Briefly, Class I α and Class II sequences were manually reviewed for active sites and then trimmed to the same ROI described above. Then, Class I α and Class II RNRdb sequences were clustered at 75% and 80% identity, respectively. Phylogenies of GOV 2.0 RNR sequences and RNRdb cluster representatives were inferred using the FastTree v2.1.11 [61] Geneious plug-in with default settings. Two trees were constructed, one each for Class I α and Class II sequences. GOV 2.0 RNR sequences were then assigned to RNR types based on phylogenetic clade. Class I α sequences were typed by RNRdb clade (Table 2.2). Class II sequences were assigned to dimeric and monomeric subclasses. The number of sequences sorted into each type can be
found in Table 3.1. One clade of GOV 2.0 Class I $\alpha$ sequences formed a clade without any references and were typed as unknown NrdA.

All typed GOV 2.0 sequences were then clustered to reduce the computing power necessary for downstream analysis. Clustering was performed at 95% identity using the MMSeqs2 v9-d36de easy-cluster program [142]. At this identity threshold, RNRs from different clades or subclasses form separate clusters. After clustering, representative sequences were trimmed to a smaller ROI that extends from residues 437 to 625 in *E. coli* [16]. This ROI allows for Class I $\alpha$ and Class II RNR sequences to be placed on the same phylogeny, as this region has been evolutionarily conserved in both classes. The trimmed representative sequences were aligned using the MAFFT v7.388 Geneious plug-in on the FFT-NS-ix1000 (iterative refinement method with 1,000 iterations) with the BLOSUM62 scoring matrix. A phylogeny was constructed using the v2.1.11 FastTree Geneious plug-in with default settings [61]. The resulting tree was colored and visualized in Iroki [62].

### 3.2.4 Compositional data analysis

Sequencing datasets are compositional, meaning they only contain relative information (i.e., relative abundance) about their components [146, 147]. This is largely due to restrictions imposed by sequencing instruments, which are capable of sequencing only an arbitrary number of molecules, thereby subjecting sequencing data to a constant sum constraint [146, 147, 148, 149]. Therefore, this study applies compositional data analysis (CoDA) methods to the GOV 2.0 dataset. All following data analysis was performed in RStudio v1.2.1335 [150] with R v3.6.0 [151] unless otherwise specified. All data visualizations, except phylogenetic trees, were created using ggplot2 v3.2.1 [152].

#### 3.2.4.1 Data preparation

Population abundances of viral population contigs $\geq$ 10kb with validated RNR sequences were collected for each sample. The “raw” (not normalized for number of
Table 3.1: Number of GOV 2.0 RNRs analyzed by type.

<table>
<thead>
<tr>
<th>Subclass</th>
<th>RNRdb clade</th>
<th>No. of sequences</th>
<th>% of total sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>NrdAe</td>
<td>159</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>NrdAg</td>
<td>810</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>NrdAm</td>
<td>5</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>NrdAn</td>
<td>3</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>NrdAza*</td>
<td>1</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>NrdA unknown</td>
<td>242</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Ib</td>
<td>NrdEb*</td>
<td>6</td>
<td>0.05</td>
</tr>
<tr>
<td>Id</td>
<td>NrdAi</td>
<td>1715</td>
<td>12.9</td>
</tr>
<tr>
<td>Ia (presumed)</td>
<td>NrdAk</td>
<td>1381</td>
<td>10.4</td>
</tr>
<tr>
<td>If (proposed)</td>
<td>CyanoSP</td>
<td>242</td>
<td>1.8</td>
</tr>
<tr>
<td>II dimeric</td>
<td>NrdJa+b, NrdJd</td>
<td>2335</td>
<td>17.5</td>
</tr>
<tr>
<td>II monomeric</td>
<td>NrdJm</td>
<td>6427</td>
<td>48.2</td>
</tr>
</tbody>
</table>

*Subsets of RNRdb clades, defined in Table 2.2*

reads) average read depth was used as a proxy for population abundance. Unnormalized values were used because metagenomic data do not meet the assumptions of read count normalization [148, 149]. Sample 155_SUR was excluded from any further analysis because only one RNR-containing viral population had any reads mapping from this sample. Gregory et al. also excluded this sample from much of their analysis [130].

A community abundance matrix was generated with samples as rows and viral population abundances (calculated as the sum of the abundance of all members of a 95% RNR cluster) as columns. Clustering was performed to reduce the computational burden of analysis and did not greatly affect the shape of the data in exploratory analysis (Appendix D). Next, an RNR subclass abundance matrix was generated with samples as rows and RNR subclass or clade abundances (calculated as the sum of the abundance of each viral population containing an RNR from the same subclass or clade) as columns. The Class I α clade NrdAk was kept separate from subclass Ia because it does not have a characterized representative [123] and was large enough to be analyzed separately (Table 3.1).
3.2.4.2 Zero replacement and CLR transformation

The 95% clustered population and subclass abundance matrices underwent the same pre-analysis processing. Zero values in the abundance matrices were replaced using the square root Bayesian-multiplicative method in the R package zCompositions v1.3.2.1 [153]. Then, zero-replaced abundance matrices were transformed using the centered log-ratio transformation [148, 154]. The CLR-transformed abundances were used as input for the remaining data analysis.

There are two major reasons for replacing zeros in metagenomic abundance data. First, these zeros are not considered to be true zeros. Instead, they are count zeros, meaning that the feature of interest (i.e. an OTU) was present in the sample at low abundance, but counting (sequencing depth) was not sufficient to observe it [149]. Second, compositional data have one fewer dimension than regular sampling data [154]. Compositional data, therefore, exist in simplex space rather than real space [147, 153, 154], but can be transformed to exist in real space, where common statistics (e.g., regression, t-tests) can be performed [154, 155]. Transformation is done using log-ratios [154, 156]. Because the log of zero cannot be computed, zeros in the data must be replaced in order to perform the data transformation [153].

CLR transformations center the data around the geometric mean of the sample, such that all observations in each sample sum to zero. Observations with positive values are more abundant than the geometric mean, such that a CLR value of one indicates that an observation has an abundance one log greater than the geometric mean. An observation with a negative value of one, then, is one log less abundant than the geometric mean.

3.2.4.3 Viral community composition

Aitchison distance [156] was used as a $\beta$-diversity metric in place of the count-based Bray-Curtis dissimilarity [148]. Aitchison distance was calculated by taking the pairwise Euclidean distances between all samples based on the CLR-transformed abundances of the community abundance matrix. This returns the same sample distances as
would performing Aitchison distance on zero-replaced values before CLR transforma-
tion. To visualize distances between samples based on community composition, PCA
was performed on the Aitchison sample distance matrix using the R package biplotR
v0.0.9 (https://github.com/mooreryan/biplotr). Additionally, samples were hier-
archically clustered using function hclust to produce a sample dendrogram based on
the same Aitchison distance matrix. PCA was also performed on the original commu-
nity abundance matrix of CLR-transformed values to create a compositional biplot.
Samples were also assigned to sample clusters based on cluster membership on the β-
diversity dendrogram. Eight sample clusters were identified based on the dendrogram:
Epipelagic 1, 2, and 3, Mesopelagic, Bathypelagic, Arctic, Antarctic/Upwelling, and
Open Ocean.

perMANOVA tests were used to determine if different GOV 2.0 ecological groups
represented different viral RNR communities. perMANOVA tests were performed in R
using the Adonis function in package vegan v2.5.5 [157]. Mantel tests with Spearman
correlation were used to test for significant relationships between community composi-
tion and continuous environmental variables including latitude, temperature, oxygen,
salinity, nitrate, ammonium, chlorophyll a, iron, phosphate, silicate, and nitrite. Iron
measurements were only taken at the surface (5 m depth), so only surface sample
communities were assessed.

3.2.4.4 RNR subclass distribution

The biplotR package was used to construct a sample PCA plot based on the
CLR-transformed abundance of viral populations containing each RNR subclass (or
clad). Kruskal-Wallis tests were used to test for differences in the abundance of each
RNR subclass among GOV 2.0 sample depths. In the case of a significant result,
a post-hoc Dunn test was performed to determine which pairwise comparisons were
significantly different based on RNR subclass abundance. Kruskal-Wallis and Dunn
tests were also performed to test for significant differences in RNR subclass abundance
among sample clusters based on the β-diversity dendrogram. Dunn tests were performed with the R package dunn.test v1.3.5 [158] using the Hochberg adjustment for multiple comparisons.

3.3 Results

To explore the diversity of RNRs and ecology of RNR-containing viruses in the world’s oceans, RNRs were mined from the GOV 2.0 dataset, which contains 145 viromes from Tara Oceans, Tara Oceans Polar Circle (TOPC), and the Malaspina expedition. These viromes are globally distributed and from four major sample depths: surface (5 m depth), deep chlorophyll maximum (DCM), mesopelagic, and bathypelagic.

3.3.1 RNR mining and curation

RNRs were mined from GOV 2.0 viral populations with contigs ≥ 10kb using an iterative homology searching procedure with PASV validation enabled (Appendix B.1). 14,883 viral protein sequences were identified as Class I α or Class II RNR by PASV. Of these, 14,590 (98%) properly spanned the PASV region of interest (ROI) and had all required conserved residues upon visual inspection. Two sequences contained inteins, which were manually trimmed from the sequences. Nine viral population contigs contained more than one RNR and were removed from analysis. 13,326 sequences, representing 6.8% of the GOV 2.0 ≥ 10kb viral populations, spanned the longer ROI used for phylogenetic classification (225-759 in E. coli) and contained all of the catalytic sites required for RNR functionality (Table A.3). These were considered bona fide RNRs and used for analysis. Bona fide RNR sequences were clustered at 95%, resulting in 7,247 clusters.

3.3.2 Phylogenetic diversity of marine viral RNRs

Marine viruses contain most of the described RNR Class I and Class II subclasses (Figure 3.2). Class I subclasses Ic and Ie were not identified and only six subclass Ib sequences were found (Table 3.1). All other Class I subclasses are well-represented in
the phylogeny, though some Class I RNRdb clades were absent (NrdAh, NrdAq) or found in very low numbers (NrdAm, NrdAn, NrdAz). The unknown GOV 2.0 RNR clade, which formed a clade lacking any RNRdb reference sequences, was grouped with subclass Ia for later analyses. The unknown clade was closest to the NrdAz RNRdb clade, which contains sequences belonging to both subclass Ia and subclass Ic. While subclass Ia and Ic RNRs cannot be distinguished based on the primary sequence of the $\alpha$ subunit, they can be distinguished phylogenetically [36]. The unknown GOV 2.0 clade is more closely related to the subclass Ia sequences of the NrdAz clade, and so is likely to belong to subclass Ia. Additionally, a survey of Class I $\beta$ subunits in the original GOV dataset [134] found no subclass Ic $\beta$ subunit sequences (Appendix C).

### 3.3.3 Community composition of RNR-containing viruses

RNR-carrying viral communities in global marine samples were compared based on sample abundances of GOV 2.0 viral populations (95% clusters of viral contigs). Principle components analyses (PCA) performed on the community abundance matrix (CLR-transformed abundances based on sequencing reads mapped to contigs, with zero replacement) and on the sample distance matrix (pairwise Aitchison distance between samples) demonstrate how samples differ based on the structure of the RNR-containing virus community (Figure 3.3).

Eight clusters were observed in a sample dendrogram based on hierarchical clustering of the distance matrix (Figure 3.3) and identified based on GOV 2.0 sample membership: Arctic, Antarctic/Upwelling, Bathypelagic, Epipelagic groups 1, 2, and 3, Mesopelagic, and Open Ocean. The Arctic cluster contained most of the Arctic samples, including mesopelagic samples from the Arctic. This was in contrast to the samples from non-polar regions, where epipelagic and mesopelagic samples formed separate clusters. The Antarctic/Upwelling cluster was so named because it contains all of the Antarctic samples (except Station85_MES) and samples from major upwelling zones, such as the Peruvian and Benguelan coasts. This cluster also contained some Arctic samples. Bathypelagic and Mesopelagic clusters consisted only of samples from those
Figure 3.2: GOV 2.0 contains much of the known Class I and Class II RNR diversity. Maximum-likelihood phylogenetic tree of 95% cluster representatives of GOV 2.0 RNRs trimmed to a region of interest. Branch color represents RNR subclass and the gray dotted line indicates the Class separation. GOV 2.0 sequences were typed using reference sequences prior to placement on this phylogeny. The tree was constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.
Figure 3.3: GOV 2.0 ecological zones are preserved by the RNR-containing virus community, but $\beta$-diversity indicates new sample clusters. A) PCA of GOV 2.0 samples based on CLR-transformed community abundance matrix colored by the ecological zones set forth in the original analysis of the GOV 2.0 dataset [130]. B) and C) PCA of GOV 2.0 samples based on Aitchison distance between samples. Aitchison distance was calculated based on the CLR-transformed community abundance matrix. Samples are colored by B) GOV 2.0 ecological zone and C) Aitchison distance dendrogram cluster. D) Hierarchically clustered sample dendrogram based on the Aitchison distance matrix. Branch color represents observed dendrogram clusters. Leaf dot color represents GOV 2.0 ecological zone. Leaf label color corresponds to the Longhurst province in which a sample was taken.
depths. When mesopelagic samples clustered away from other mesopelagic samples, it was generally to join samples from the same sampling station (e.g., Station72_MES clusters with Station72_SRF and Station72_DCM, rather than with other mesopelagic samples). Epipelagic clusters all were composed mainly of surface and DCM samples from temperate and tropical (non-polar) regions. The Open Ocean cluster consists of DCM and mixed layer samples (mixed layer sampled because no DCM was detected) from Stations 122-125, and all samples from Station 72. These stations are from the equatorial Pacific and Atlantic, respectively, and are the samples furthest away from a continent. GOV 2.0 samples can also be described by environmental metadata such as Longhurst provinces or sample depth. In addition, there are five ecological zones identified by Gregory et al. as producing different viral assemblages: temperate and tropical epipelagic, temperate and tropical mesopelagic, bathypelagic, Arctic, and Antarctic [130]. Three samples were also considered statistical outliers and were not assigned to an ecological zone.

Significant differences in RNR-containing viral communities between sample environmental groups defined by dendrogram cluster, Longhurst region, sample depth, or GOV 2.0 ecological zone were determined with perMANOVA tests. All environmental descriptors had a significant perMANOVA result (p < 0.05), meaning that at least one of the sample groupings was characterized by a significantly different community (Figure 3.4A). Longhurst region explained the most variation in community structure with an R² value of 0.396, followed by dendrogram cluster (R² = 0.271) and ecological zone (R² = 0.219). Depth explained the least variation (R² = 0.079) but was still a significant factor (p = 0.001). Mantel tests were performed to determine which environmental factors may be driving changes in community structure. Latitude, temperature, oxygen, salinity, nitrate, silicate, phosphate, and ammonium were all shown to have a significant (p < 0.05) relationship with changes in community structure (Figure 3.4B).
**Figure 3.4: Viral community structure changes with biogeographical regions and environmental variables.** Column charts showing the results of A) Mantel and B) perMANOVA tests. In both charts, the color of the column indicates a significant ($p \leq 0.05$) or insignificant ($p > 0.05$) test statistic. In A) column height indicates the value of the Mantel R statistic, with higher values indicating greater correlation between the two distance matrices. Positive values indicate positive correlations while negative values indicate negative correlations. Values approaching zero indicate no correlation between distance matrices. In B) column height indicates the value of the perMANOVA $R^2$ statistic, with higher values indicating a higher portion of explained variance.
3.3.4 RNR subclass ecology

The abundance of RNRs by subclass was used to explore the potential role of RNR in shaping the biology and ecology of marine viruses. RNR subclass abundance was determined by summing the abundance of viral populations carrying the same RNR subclass, followed by zero replacement and CLR-transformation. Subclass Ia includes RNRs from RNRdb clades (Table 2.2) NrdAe, NrdAg, NrdAm, NrdAn, NrdAz, and the unknown clade of 95% clustered GOV 2.0 RNR sequences. RNRdb clade NrdAk was analyzed individually and not grouped into any subclass because the clade has no biochemically characterized representatives [123], groups away from subclass Ia clades on a phylogenetic tree (Figure 3.2), and had enough viral population members (n = 1381) to be analyzed separately. Class I subclasses Id and If (proposed novel subclass containing the Cyano SP RNRs, 2.4.3), and Class II dimeric (IId) and monomeric (IIm), were also included in the analysis.

Principal components analysis was performed on the CLR-transformed subclass abundance matrix, with samples as points and subclasses as loadings (Figure 3.5). Together, PC1 and PC2 explained 92.3% of variation in the data based on subclass abundances. Subclass If appeared to be driving the separation between non-polar and polar samples. The Class II subclasses (IId and IIm) and the Clade NrdAk seem to be driving bathypelagic and some mesopelagic samples. Subclass Id abundance appears to be a major influence on polar samples, and subclass Ia shows no clear patterns. In the case of RNR subclasses, GOV 2.0 ecological zones seem to map better onto the PCA than the dendrogram clusters defined by the community structure (Figures 3.5 3.3).

The global distribution of RNR subclasses was explored using loess lines of CLR-transformed subclass abundance plotted against latitude (Figure 3.6). Subclass If had the greatest change in abundance with latitude, with an observed > 6 log reduction in abundance at the poles compared to the equator. Subclass Ia was least relatively abundant near 30°S, with an increase in relative abundance near 30°N. All other subclasses became more relatively abundant at the poles. At all latitudes, Class
Figure 3.5: RNR subclass abundances drive separations among samples. Points are samples and loadings are subclasses. Samples are colored by A) dendrogram cluster and B) GOV 2.0 ecological zone.

II monomeric RNRS displayed the highest CLR abundance.

Significant differences in RNRdb subclass abundance among sample dendrogram clusters and sample depths were determined with Kruskal-Wallis tests, with post-hoc Dunn tests to identify significantly different pairs. Longhurst provinces were not tested because many of the provinces contain few sequences (Figure 3.3). Sample depth is an important aspect of the GOV 2.0 dataset as it reflects important marine ecological features such as the DCM or bathypelagic. Samples taken from mixed layer depths were not included in this analysis due to a lack of sampling at this depth (n = 6) (Table 3.2). Sample depth was associated with changes in abundance (Kruskal-Wallis p < 0.05) for all subclasses, though RNRdb subclasses had different depth profiles (Figure 3.7). The Class I uncharacterized clade NrdAk and both Class II subclasses showed increased relative abundance with depth. All other Class I subclasses showed decreased abundance with depth. Similar to latitude, Class II monomeric RNRS also had the highest CLR abundance at all sample depths. The difference in abundance between surface and DCM samples was never significant, while the differences in abundance between bathypelagic and surface or DCM were always significant, no matter the subclass. The difference in abundance between mesopelagic and all other depths
Figure 3.6: RNR subclasses show different latitudinal patterns of abundance. CLR abundances of viral populations containing the same RNR subclass were pooled within each sample. Sample abundances of each subclass pool were plotted against the latitude at which the sample was taken for all meso- and epipelagic samples. Loess (local regression) lines were plotted for each subclass. Shadows represent the 95% confidence intervals. Points, lines, and shadows are colored by subclass. Subclass colors correspond to branch colors in Figure 3.2.
Table 3.2: Number of samples taken at each depth.

<table>
<thead>
<tr>
<th>Sample depth</th>
<th>No. of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface (SRF)</td>
<td>60</td>
</tr>
<tr>
<td>Deep chlorophyll maximum (DCM)</td>
<td>39</td>
</tr>
<tr>
<td>Mixed layer (MXL)</td>
<td>6</td>
</tr>
<tr>
<td>Mesopelagic (MES)</td>
<td>26</td>
</tr>
<tr>
<td>Bathypelagic (BTH)</td>
<td>13</td>
</tr>
</tbody>
</table>

was generally significant, though this varied by subclass.

Subclasses were then tested to see if their abundances were significantly different among the viral community Aitchison distance dendrogram clusters (Figure 3.8). It is important to note here that the Mesopelagic dendrogram cluster does not contain all mesopelagic samples, in contrast to the mesopelagic grouping discussed in the above sample depth section. Kruskal-Wallis tests returned significant p-values ($p < 0.05$) between clusters for all subclasses. Post-hoc Dunn tests showed that, for all subclasses, a significant difference was never observed between Arctic and Antarctic/Upwelling or Mesopelagic and Bathypelagic pairs. Significant differences were also never observed between any pair combinations of Epipelagic and Open Ocean clusters. The abundance of an RNR subclass was generally different in the Bathypelagic cluster compared to other clusters. The Bathypelagic cluster was less distinct for subclasses Id and If. For subclass Id, all clusters displayed a significantly lower CLR abundance than the Arctic and Antarctic/Upwelling clusters. The only additional significantly different pair was Bathypelagic and Epipelagic 1. For subclass If, the Bathypelagic cluster was only significantly different from the Open Ocean and Epipelagic clusters, while being statistically indistinguishable from the Arctic and Antarctic/Upwelling clusters. For all samples, the Mesopelagic cluster was less distinct than the Bathypelagic cluster but showed the same general patterns.
Figure 3.7: RNR subclasses display different depth profiles. CLR abundances of viral populations containing the same RNR subclass were pooled by sample. For each subclass pool, samples were divided based on sample depth and subclass abundances plotted with sample depths as groups. Boxplot fill colors indicate sample depth. For subclass If, only non-polar samples ($60^\circ S < \text{latitude} < 60^\circ N$) were included because the CLR abundance of subclass If is very low in all polar samples, regardless of sample depth. Test statistics for Kruskal-Wallis tests comparing subclass abundances among sample depths are shown beneath the corresponding plot. Based on Dunn tests, all pairwise comparisons were significant except the following: (all subclasses) SRF/DCM; (Ia) DCM/MES; (Id) SRF/MES, DCM/MES, (If) MES/BTH, (IId) MES/BTH.
Figure 3.8: RNR subclasses have different biogeographical distributions. CLR abundances of each RNR subclass were tested for differences in abundance among sample dendrogram clusters. Boxplot fill colors indicate dendrogram clusters. For each dendrogram cluster, significantly different clusters are indicated by the abbreviations in parentheses above the corresponding boxes.
3.4 Discussion

The viruses analyzed in this study represented 6.8% of the GOV 2.0 viral populations, though the actual percentage of viruses carrying RNR genes is likely to be much higher. Our strict validation process for RNR protein sequences excluded many RNRs that appeared functional but were too short to be considered. A past study of RNRs in marine viromes estimated that up to 93% of marine viruses carry RNR genes [16], though this is likely an overestimate. Another study found that RNRs were present in 17.4% of viruses surveyed [159]. However, this study included few marine viruses and is not representative of the GOV 2.0 dataset.

3.4.1 Phylogenetic diversity of marine viral RNRs

The viral RNRs found in the GOV 2.0 dataset represented most of the known Class I and Class II RNR subclasses. Class II monomeric sequences accounted for the largest percentage of recovered sequences (48.6%), consistent with a previous study of RNR in viromes which found that Class II monomeric RNRs made up roughly 50% of RNRs [16]. However, it is unknown if the exact count of subclass members is meaningful in this study because RNRs were mined from GOV 2.0 viral populations, which had previously been clustered [130]. Class II monomeric sequences were also the most phylogenetically diverse subclass, an unexpected finding as the subclass constitutes one of the smaller clades in the RNRdb. Additionally, Class II monomeric sequences are also relegated to a single clade in the RNRdb, whereas Class I $\alpha$ and Class II dimeric sequences are divided among several clades [48]. In contrast, Class II monomeric sequences from GOV 2.0 formed multiple distinct clades (Figure 3.2). This is a strong indication that the RNR diversity in environmental datasets is not currently captured in sequencing databases.

Viruses carrying Class II monomeric RNRs were the most relatively abundant group at all sample depths, at all latitudes, and in all sample dendrogram clusters, consistent with a previous survey of RNRs in viromes [16]. Viruses carrying Class II
dimeric RNRs were found to have generally high CLR abundances as well. The dominance of Class II RNR-carrying viruses in the ocean may be related to the availability of their enzymatic cofactor, $\text{B}_{12}$. $\text{B}_{12}$ is generally biologically synthesized and has a cobalt atom at its center [89]. Therefore, cobalt and $\text{B}_{12}$ concentrations are directly related to one another [160]. While cobalt is the scarcest trace metal utilized by biological entities in the oceans [161], it is generally not limiting to primary production [72] because cobalt and $\text{B}_{12}$ are required for relatively few enzymes as compared to other metals, such as iron, manganese, or zinc [73, 162]. When cobalt is limiting, it is also rarely the primary limiting nutrient [72, 71]. Thus, $\text{B}_{12}$ should be readily available in most areas of the oceans.

RNRs belonging to the Class II dimeric subclass and Class I subclasses Ia, Id, If, and clade NrdAk were also found in high numbers. Subclasses Ic and Ie were entirely absent in the GOV 2.0 dataset, and very few subclass Ib sequences were recovered. The iterative homology searching method used in this study was designed specifically to recover all RNR sequences in the dataset, so it is unlikely that these sequences would have been missed. The Ic, Ie, and Ib subclasses are commonly found in host-associated, pathogenic bacteria [33, 70, 163], which should be absent from marine systems. An earlier study of viral RNRs in metagenomic datasets found subclass Ib to be absent from epipelagic marine samples, though they were abundant in a hydrothermal vent system [17]. Subclass Ic was not examined in that study, nor was subclass Ie, as it had not yet been described.

3.4.2 RNR-containing viruses reflect the total viral community

Our results confirmed that RNR is a useful marker for environmental phage populations. First, RNR genes appeared to be single copy in viral genomes. Just 9 GOV 2.0 viral populations were found to contain more than one RNR and no populations contained more than two RNR genes. In all cases, the two RNRs were located far apart (thousands of bases) on the viral population contig and were virtually identical,
perhaps indicating that these contigs originated from the fusion of two highly related viral genomes, rather than a single virus with multiple RNRs.

Second, and perhaps most importantly, the RNR-containing virus subset reflected the viral community as a whole. Despite the low percentage of viral populations analyzed, we were able to recreate the same basic data structure using only the RNR-containing viral community. The GOV 2.0 ecological zones characterized by different total viral communities were also associated with changes in structure of the RNR-containing viral community, as the zones mapped well onto PCA plots (Figures 3.3A-C and 3.5) and a dendrogram (Figure 3.3D) constructed from Aitchison sample distance. Statistical testing also showed that GOV 2.0 ecological zone was a significant factor in the differentiation of RNR-containing virus community structure. Additionally, the eight sample clusters identified on the dendrogram largely overlapped with the ecological zones defined by Gregory et al., with polar and non-polar samples being distinct and sample depth driving separation in low and mid-latitude regions [130]. Altogether, this indicates that the RNR-containing community is similarly structured to the total viral community.

While this may seem remarkable given the relatively small subset analyzed, this result is perhaps unsurprising when considering what is known about the diversity of RNR-containing viruses and the GOV 2.0 dataset. Viruses that infect bacteria and archaea are overwhelmingly dsDNA viruses [10], with the viral order Caudovirales considered to be the most diverse and abundant group [164, 10, 165]. RNRs are known to be present in many lytic dsDNA viruses, including all three families of Caudovirales [16, 166]. The sample preparation used by all expeditions included in GOV 2.0 are known to be biased toward dsDNA viruses smaller than 0.22 μm [130, 167], meaning that the GOV 2.0 dataset may be enriched towards RNR-carrying viruses. Additionally, some viruses carrying RNR are considered to be ubiquitous in the oceans [166] and many infect hosts known to be highly abundant in marine ecosystems, such as members of the SAR11 [16] and SAR116 clades [168], and Cyanobacteria [16, 66].

58
3.4.3 RNR subclasses have different ecological distributions

Finally, RNR subclasses showed different biogeographical distributions that largely tracked with distributions of their cofactors, indicating that RNR is ecologically informative and connects viruses to their external environments. Depth and latitude were the most important factors shaping subclass abundance. Sample clusters from the dendrogram based on community structure seemed to be too granular for the detection of changes in subclass abundance, as pairwise comparisons of subclass abundance between the polar clusters, any combination of Open Ocean and Epipelagic clusters, or the Mesopelagic and Bathypelagic clusters were not significant.

Viruses carrying subclass Ia RNRs, which utilize a diiron cofactor (Figure 1.1B) seemed to be tied to geographical patterns of iron concentration in the surface ocean, but not to depth increases in iron concentration. Subclass Ia-carrying viruses had an asymmetrical change in abundance with latitude, increasing in abundance between roughly 20°N and 40°N with steady abundance into the high northern latitudes. Additionally, no significant difference in abundance was found between any of the polar and epipelagic sample clusters. Samples from 20-40°N originate from the Mediterranean and Red Seas and the Northern Indian Ocean. Along with much of the coastal Arctic, these geographic regions have some of the highest surface iron concentrations in the global ocean [169, 170]. In contrast, samples taken from the southerly counterparts of these latitudes are largely from upwelling zones and open ocean sites, where iron is scarcer [169, 170], perhaps resulting in lower subclass Ia RNR abundance. However, viruses carrying subclass Ia enzymes became less relatively abundant with depth, in opposition with the classical nutrient-like depth profile of iron, in which concentrations are depleted at the surface and increase with depth [124, 125, 169]. It is possible that the hosts of viruses carrying subclass Ia RNRs are confined to surface waters, tying their viruses to the same environment. This seems a likely explanation as most cyanomyoviruses are known to carry subclass Ia RNRs (Cyano M clade, [16, 36]), and their hosts, marine Synechococcus and Prochlorococcus are limited to the sunlit ocean.

Viruses carrying RNRs from the uncharacterized NrdAk clade showed similar
latitudinal abundance patterns as subclass Ia-carrying viruses, but also increased in relative abundance with depth. Viruses with clade NrdAk RNRs, then, seemed to be tracking iron patterns with respect to both geographical area and sample depth, indicating that RNRs in the clade NrdAk may use iron as a cofactor for radical generation. It is possible that clade NrdAk and subclass Ia RNRs use iron atoms in different oxidation states, allowing clade NrdAk to track iron concentrations generally, but restricting subclass Ia RNRs to surface waters. Viruses with NrdAk RNRs may also have a more diverse set of hosts, allowing them to be more connected to the environment, rather than a few specific microbial populations.

Subclass Id-carrying viruses showed abundance patterns similar to those of manganese concentrations in the oceans, reaching their highest relative abundances at the poles, and their lowest in bathypelagic samples. This was expected as subclass Id RNRs use a dimanganese cofactor for radical generation [34, 123]. Manganese concentrations are higher in surface waters than deep waters [125, 171] because of photochemical reactions that only proceed due to light availability [124, 171]. Manganese is also highly abundant in the Arctic due to sediment flux [171], where viruses with subclass Id RNRs reach their highest latitudinal abundance.

Viruses carrying RNRs of the proposed subclass If clearly shadowed the distribution of their host organisms, marine *Synechococcus* and *Prochlorococcus* [16, 66, 36], showing dramatic changes with both latitude and depth. Marine *Synechococcus* and *Prochlorococcus* are most abundant at the equator and decrease in abundance with increasing latitude [172]. These organisms also require light to perform photosynthesis, and so cannot live in the dark ocean. A few mesopelagic samples did show elevated abundance of subclass If-carrying viruses (Figure 3.5), which could be due to the presence of low-light *Prochlorococcus* ecotypes that can be abundant below the thermocline [173] or due to mixing events that moved cyanobacterial cells below the mixed layer. This result provides evidence for the hypothesis presented in Harrison et al. that cyanophage RNRs are unconnected to the extracellular environment, instead relying on intracellular pools of iron or manganese within their cyanobacterial hosts.
Viruses containing either of the Class II RNR subclasses showed similar ecological distributions, which was expected as they utilize the same cofactor (adenosylcobalamin, a form of B$_{12}$). Class II RNR-containing virus abundance followed patterns similar to the distribution of cobalt, the metal at the center of B$_{12}$ [89], with CLR abundances increasing with depth and latitude. Cobalt is typically depleted in non-polar surface waters due to biological uptake [124] and increases in concentration with depth [174]. In contrast, surface seawater in the Arctic contains some of the highest concentrations of cobalt in the global ocean [174], which likely accounts for the increased CLR abundance of Class II RNRs at all depths in the high northern latitudes.

The fact that RNR subclass abundances largely shadow the distributions of their cofactors makes it somewhat puzzling why increases in abundance are observed in the Antarctic for most subclasses. Unlike the Arctic Ocean, the Southern Ocean is known to have low concentrations of trace metals, including cobalt [174], iron [170], and manganese [171]. In fact, all three metals have been shown to be limiting nutrients for primary production in the Southern Ocean [72, 171]. There are several possible explanations for this discrepancy. First, the Antarctic results are relatively unstable as there are many fewer Antarctic (n = 5) than Arctic (n = 40) samples. Second, the increased abundances may be an artifact of the data itself. Because metagenomic sequencing data is compositional [148], it is not possible to extract true counts from the data, only relative abundances. Therefore, it is possible that RNR-containing viruses are less abundant in Antarctic waters than other regions as a whole, but some subclasses are more relatively abundant in the Antarctic as compared to non-polar regions. This would be consistent with a study that found that temperate viruses are more abundant in the Antarctic than in lower-latitude regions [164], as RNRs are rarely found in temperate phage. A quantitative method such as qPCR would be required to test this hypothesis. Third, elevated nitrogen concentrations in the Southern Ocean could be driving an increase in abundance of RNR-containing viruses. Nitrate is the primary limiting nutrient in most marine surface waters [72, 71], but is greatly enriched
in the Southern Ocean [175]. One of the key biological uses for nitrogen is for the synthesis of nucleotides, which has been established as important for viral replication [114, 118]. A recent study showed that during infection, a cyanophage caused its host cell to uptake nitrogen from the environment, which was incorporated into proteins and presumably into dNTPs as well [122]. Thus, it is possible that the increased availability of nitrogen in the Antarctic could have caused the RNR-containing virus population to be larger, as dNTP production could have been less limited. The explanation could also be different for different subclasses. For subclass Id, a recent model has shown that the area where Antarctic samples were collected is relatively enriched for manganese compared with surrounding waters [171], which could perhaps contribute to the more moderate relative abundance increase in Antarctic samples. For Class II RNRs, the difficulty in measuring B$_{12}$ could have led to underestimates of its concentration in the Antarctic [176]. Class II RNR abundance could also be greater despite apparent B$_{12}$ limitation because RNRs are in some ways flexible about the type of B$_{12}$ required. B$_{12}$ forms can be divided into two main groups, cobalamins and pseudocobalamins, which differ in the identity of the lower-axial ligand. Cyanobacteria and eukaryotic phytoplankton use different forms of B$_{12}$, with Cyanobacteria producing and using pseudocobalamin, while eukaryotic phytoplankton prefer cobalamin [73, 74]. Class II RNRs, on the other hand, have no preference for either group. Instead, Class II RNRs require an adenosyl group be attached as the upper-axial ligand. Therefore, while vitamin B$_{12}$, a cobalamin, has been shown to be limiting for the growth of eukaryotic algae [71], it may not be limiting for a viral Class II RNR, which can also use pseudocobalamin as a cofactor. This could also help explain why Class II RNRs are still found at a high relative abundance in surface waters, because pseudocobalamins produced by Cyanobacteria are found in high concentrations above the mixed layer [73].
3.4.4 Differences between this study and the original GOV 2.0 analysis

The placement of individual samples into groups sometimes differed between this study and the original GOV 2.0 analysis that included the total viral community [130]. In the sample dendrogram presented here (Figure 3.3), the three GOV 2.0 outlier (singleton) samples clustered with other samples. Sample 85_MES grouped with other mesopelagic samples. Samples 102_MES and 72_MES grouped with samples from the same station, but different sample depths. Some group rearrangements were observed. For example, MSP114, the single mesopelagic sample from the Malaspina expedition grouped with other Malaspina samples in this study, but with other mesopelagic samples in the GOV 2.0 study. Sample 63_MES grouped near the Malaspina samples in this study, but with other Arctic samples in the GOV 2.0 analysis. Finally, this study's Antarctic/Upwellings cluster was a novel grouping. In the full GOV 2.0 study, Antarctic samples grouped separately from other samples, though they were most closely associated with Arctic samples. In the present study, Antarctic samples grouped with a handful of Arctic samples, and, surprisingly, with some samples from non-polar regions. These non-polar samples were all from upwelling regions, specifically from the Benguelan (Station 67), Peruvian (Station 102), and Patagonian continental shelf (Station 82) upwelling zones. All of these locations were experiencing upwelling events during the sampling period [177, 178, 179]. A recent study also revealed that the areas of the Antarctic sampled during Tara Oceans are also upwelling zones [180]. The Antarctic and non-polar upwelling samples likely grouped together because upwellings bring deep water that has often originated from polar regions up to the surface, producing similar conditions. The Arctic samples included the Antarctic/Upwelling cluster are all along the Transpolar current. After crossing the Arctic, this current continues down the eastern coast of Greenland into the North Atlantic. This is one of the areas where ocean water sinks and becomes part of the Great Ocean Conveyor Belt. This water would eventually travel to the Antarctic and the upwelling zones, creating a possible connection between these specific Arctic and Antarctic samples.

Longhurst province [181] was shown by the Aitchison distance dendrogram and
by a perMANOVA to be a significant factor and explain the most variation in viral community structure. This was surprising because the original GOV 2.0 study concluded that viral assemblages were not beholden to traditional biogeographical ocean biomes such as Longhurst province. The grouping that explained the second most variation community structure was the dendrogram clusters, which was also the second most granular grouping of samples after Longhurst province. This indicates that perhaps the RNR-containing virus community is more connected to shifts in environmental conditions, and therefore is more likely to correspond with biogeographical regions defined by nutrient characteristics, such as Longhurst provinces [181].

It is unclear why these differences exist between the sample groups in this study and those defined by Gregory et al [130]. One possible source is the data analysis itself. Aitchison distance was used as a $\beta$-diversity metric in this study, while the previous Gregory et al. used Bray-Curtis dissimilarity. In addition to being compatible with CoDA methods, Aitchison distance has also been shown to be invariant to many biases inherent in sequencing studies [182]. This makes Aitchison distance a useful choice in comparing samples from three separate survey expeditions conducted over multiple years. However, it is just as possible that these differences in sample relationships are the result of analyzing only a subset of the data (6.8% of viral populations $\geq 10$kb). To verify the true source of variation, the analysis completed in this study would need to be replicated using all GOV 2.0 viral populations.

3.4.5 Future work

Future work still needs to be done in determining the driving factors of RNR-containing viral communities. This study has so far been unable to leverage the wealth of metadata collected during the Tara Oceans and TOPC expeditions. This is because ecological surveys such as this suffer from spatial autocorrelation, in which sites that are nearer to one another also tend to have similar metadata and biological communities [183, 184]. In the GOV 2.0 dataset, both geographical and temporal spatial autocorrelation are present, because sites that were sampled closer together in space
were also sampled closer together in time [130, 133, 134]. This must be controlled for when testing relationships between environmental data and sequencing results. Otherwise, autocorrelation can mask significant relationships or create spurious correlations between variables [183, 184].

Additional future work should include the study of viral RNRs from other habitats. In addition to the global ocean, viral RNRs have also been found in viruses from hot springs [185], near hydrothermal vents [17], in desert hypoliths [186], and in fermented foods such as kimchi [17]. However, the factors driving the distributions of different RNRs in such systems have yet to be studied. Studying RNRs in diverse habitats and along environmental gradients would be helpful in determining how carrying RNR genes influences the biology and ecology of environmental viruses.
Chapter 4
CONCLUSIONS

From the work presented in this thesis, it is clear that RNR genes play a role in shaping the ecology and biology of the viruses that carry them. Additional research will be needed to assess the extent of the influence of viral RNRs over their carriers. For instance, I hypothesized that losing the ATP cone region would contribute to the highly lytic phenotype of *Prochlorococcus* phage P-SSP7. Benchtop experiments would be required to test this hypothesis, and it would still be unclear if an increase in burst size observed in the laboratory would be observed in the natural environment. An alternative hypothesis may be simply that the domain was lost because of evolutionary pressure to shorten the gene. However, this work shows great promise for RNRs and their nutrients as specific factors that may be driving individual viral populations, and more strongly identifies RNR as a contributor to observed viral phenotypes.

The biogeographical survey presented in Chapter 3 showed that when viral populations were grouped by the carried RNR subclass, viral abundances corresponded to the distributions of the biochemical cofactors required by those subclasses. This makes RNR a powerful marker, as it is able to not only reflect the total viral community, but is also informative about the ecology and biology of the viruses, which is important for making genome-to-phenome connections. Most viral marker genes currently in use are structural proteins [112, 127, 128, 187, 188]. While this ensures that amplified products are viral, structural genes do not connect viruses to the environment. They also generally provide information about single virus lineages (e.g., *gp23* only amplifies myoviral sequences) and would not reflect the viral community as a whole. Other viral marker genes include functional proteins such as *psbA* or *polB*, but still target only specific viral populations (T4-like cyanophage) [128, 189, 126]. RNR, in contrast, is
present in diverse viruses [10, 16] and connects its carrier to the environment [19, 20]. A limitation of RNR, however, is that viral RNRs are not necessarily distinguishable from cellular RNRs [17, 30], meaning that a sample would still have to undergo filtration, concentration, and DNase steps prior to PCR amplification to ensure that any amplified RNR sequences are viral. Additionally, because RNRs are highly divergent [23, 30], a suite of primers would need to be designed to capture the full RNR diversity.

Reannotation of the Cyano SP clade, which constitutes the proposed subclass If, was essential for interpreting the results of the biogeographical survey. The hypothesis proposed in Chapter 2 that carrying Class I RNRs allows cyanophage to take advantage of large intracellular pool of iron helped explain why viruses with subclass If RNRs did not follow known patterns of manganese or iron concentrations. More importantly, had the Cyano SP RNRs still been annotated as Class II RNRs, their inclusion in the analysis of Class II RNR-containing viruses could have masked patterns and altered the conclusions. This would have been the likely result considering that subclass If and Class II RNR-containing viruses showed opposite patterns with respect to both latitude 3.6 and depth 3.7.

Taken together, the studies in this thesis show the utility of RNR as a viral marker gene, and the power of RNR genes to influence the biology and ecology of the viruses that carry them.
BIBLIOGRAPHY


Laura M. K. Dassama, Amie K. Boal, Carsten Krebs, Amy C. Rosenzweig, J. Martin Bollinger, and Jr. Evidence that the β subunit of *Chlamydia trachomatis* ribonucleotide reductase is active with the manganese ion of its manganese(IV)/iron(III) cofactor in site 1. *Journal of the American Chemical Society*, 134(5):2520–2523, feb 2012.


A.1 Supplemental Methods

Protein sequence similarity networks (SSN) were constructed with the same trimmed RNR sequences used for the Class I α with Class II phylogeny and for the Class I α-only phylogeny. Cyano SP α subunit sequences trimmed to the corresponding region were also included. The SSNs were generated with the Enzyme Similarity Tool (EFI-EST) [63] as in Rose et al. (E-value: 5, fraction: 1) [34], except that multiple alignment scores (25 and 40 for the Class I α with Class II sequences, 110 and 155 for Class I α-only sequences) were chosen to generate a more complete picture of protein connectivity. As the full networks were too large to visualize in Cytoscape [65, 64] representative node networks were presented instead (for Class I α with Class II sequences: 55% clustering for alignment score 25, 60% clustering for alignment score 40; for Class I α-only sequences: 40% clustering for alignment score 110, 55% for alignment score 155).

A.2 Supplemental Tables
### Table A.1: Clustering statistics for Class I $\alpha$ subunit sequences removed from RNRdb groups NrdE and NrdAz prior to phylogenetic analysis.

<table>
<thead>
<tr>
<th>Cluster identity</th>
<th>Only included sequences</th>
<th>Mixed</th>
<th>Only excluded sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. clusters</td>
<td>% total seqs.</td>
<td>No. clusters</td>
</tr>
<tr>
<td>70%</td>
<td>27</td>
<td>1.7</td>
<td>40</td>
</tr>
<tr>
<td>75%</td>
<td>60</td>
<td>2.0</td>
<td>64</td>
</tr>
<tr>
<td>80%</td>
<td>114</td>
<td>4.2</td>
<td>103</td>
</tr>
</tbody>
</table>
Table A.2: Annotated Class I α, Class I β, and Class II references used for alignment of putative α and candidate β subunits and curation of phylogenetic reference sequences.

<table>
<thead>
<tr>
<th>Class I α subunit representatives</th>
<th>Class</th>
<th>Subclass</th>
<th>Species name</th>
<th>Crystal structure</th>
<th>Reference</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>a</td>
<td><em>Escherichia coli</em></td>
<td>1RLR</td>
<td>[69]</td>
<td>CQR81730.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>b</td>
<td><em>Salmonella typhimurium</em></td>
<td>1PEU</td>
<td>[190]</td>
<td>WP_000246626.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>c</td>
<td><em>Chlamydia trachomatis</em></td>
<td>N/A</td>
<td>[70]</td>
<td>WP_009872213.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>d</td>
<td><em>Flavobacterium johnsoniae</em></td>
<td>N/A</td>
<td>[34]</td>
<td>WP_012026039.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>e</td>
<td><em>Aerococcus urinae</em></td>
<td>N/A</td>
<td>[33]</td>
<td>WP_060778521.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class I β subunit representatives</th>
<th>Class</th>
<th>Subclass</th>
<th>Species name</th>
<th>Crystal structure</th>
<th>Reference</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>a</td>
<td><em>Escherichia coli</em></td>
<td>1RIB</td>
<td>[107]</td>
<td>KXG99827.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>b</td>
<td><em>Salmonella typhimurium</em></td>
<td>1R2F</td>
<td>[98]</td>
<td>WP_000777903.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>c</td>
<td><em>Chlamydia trachomatis</em></td>
<td>1SYY;4D8F</td>
<td>[70, 191]</td>
<td>WP_009872214.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>d</td>
<td><em>Flavobacterium johnsoniae</em></td>
<td>6CWO-P</td>
<td>[34]</td>
<td>WP_012026040.1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>e</td>
<td><em>Aerococcus urinae</em></td>
<td>6EBO</td>
<td>[33]</td>
<td>WP_013669290.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class II representatives</th>
<th>Class</th>
<th>Subclass</th>
<th>Species name</th>
<th>Crystal structure</th>
<th>Reference</th>
<th>Accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>II</td>
<td>monomeric, RTPR</td>
<td><em>Lactobacillus leichmanii</em></td>
<td>1L1L</td>
<td>[192]</td>
<td>AAA03078.1</td>
</tr>
<tr>
<td></td>
<td>II</td>
<td>dimeric</td>
<td><em>Thermotoga maritima</em></td>
<td>3O0N</td>
<td>[193]</td>
<td>WP_004082700.1</td>
</tr>
</tbody>
</table>

*PDB = Protein Data Bank

91
**Table A.3:** Catalytic residues in Class I RNR α and β subunits and their positions in the putative α and β sequences from *Prochlorococcus* phage P-SSP7.

<table>
<thead>
<tr>
<th>Class I RNR α subunit</th>
<th>Residue</th>
<th>Position in P-SSP7</th>
<th>Position in <em>E. coli</em></th>
<th>Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>32</td>
<td>225</td>
<td>active site</td>
<td>disulfide bridge</td>
<td>[143, 145]</td>
</tr>
<tr>
<td>Q</td>
<td>105R</td>
<td>294</td>
<td>substrate specificity</td>
<td></td>
<td>[14]</td>
</tr>
<tr>
<td>R</td>
<td>110</td>
<td>298</td>
<td>substrate specificity</td>
<td></td>
<td>[14]</td>
</tr>
<tr>
<td>N*</td>
<td>187</td>
<td>437</td>
<td>hydrogen bonds</td>
<td></td>
<td>[194]</td>
</tr>
<tr>
<td>C*</td>
<td>189</td>
<td>439</td>
<td>thyl radical - removes H</td>
<td></td>
<td>[195, 196]</td>
</tr>
<tr>
<td>E*</td>
<td>191</td>
<td>441</td>
<td>hydrogen bonds</td>
<td></td>
<td>[197]</td>
</tr>
<tr>
<td>C*</td>
<td>200</td>
<td>762</td>
<td>active site</td>
<td>disulfide bridge</td>
<td>[195]</td>
</tr>
<tr>
<td>P*</td>
<td>319</td>
<td>621</td>
<td>substrate binding</td>
<td></td>
<td>[144]</td>
</tr>
<tr>
<td>Y</td>
<td>423</td>
<td>730</td>
<td>radical transfer</td>
<td></td>
<td>[69]</td>
</tr>
<tr>
<td>Y</td>
<td>424</td>
<td>731</td>
<td>radical transfer</td>
<td></td>
<td>[69]</td>
</tr>
<tr>
<td>C</td>
<td>464</td>
<td>734</td>
<td>interacts with reducing agent</td>
<td></td>
<td>[143, 145]</td>
</tr>
<tr>
<td>C</td>
<td>466</td>
<td>759</td>
<td>interacts with reducing agent</td>
<td></td>
<td>[143, 145]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class I RNR β subunit</th>
<th>Residue</th>
<th>Position in P-SSP7</th>
<th>Position in <em>E. coli</em></th>
<th>Function</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>W*</td>
<td>14</td>
<td>48</td>
<td>reduces dioxygen</td>
<td></td>
<td>[198, 199]</td>
</tr>
<tr>
<td>Y*</td>
<td>78L</td>
<td>122</td>
<td>protein radical</td>
<td></td>
<td>[200]</td>
</tr>
<tr>
<td>F</td>
<td>122</td>
<td>208</td>
<td>protein radical stability</td>
<td></td>
<td>[201]</td>
</tr>
<tr>
<td>F</td>
<td>127</td>
<td>212</td>
<td>protein radical stability</td>
<td></td>
<td>[201]</td>
</tr>
<tr>
<td>R</td>
<td>145</td>
<td>236</td>
<td>radical transport</td>
<td></td>
<td>[144, 107]</td>
</tr>
<tr>
<td>Y</td>
<td>236</td>
<td>356</td>
<td>radical transport</td>
<td></td>
<td>[202, 203]</td>
</tr>
</tbody>
</table>

Residues in **bold** were used in reference curation.

*Residues were used for PASV profile.
### Table A.4: Accession numbers for genomes and RNR and Pol I protein sequences of the Cyano SP clade.

<table>
<thead>
<tr>
<th>Virus</th>
<th>Genome accession no.</th>
<th>RNR α accession no.</th>
<th>RNR β accession no.</th>
<th>Pol I accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Prochlorococcus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phage P-SSP7</td>
<td>NC-000882.2</td>
<td>YP-214197.1</td>
<td>YP-214198.1</td>
<td>YP-002755.438.1</td>
</tr>
<tr>
<td><em>Cyanophage</em> P-SSP2</td>
<td>NC-016656.1</td>
<td>YP-005087372.1</td>
<td>YP-005087373.1</td>
<td>frameshift</td>
</tr>
<tr>
<td><em>Cyanophage</em> 9515-10a</td>
<td>NC-016657.1</td>
<td>YP-005087431</td>
<td>YP-005087442.1</td>
<td>frameshift</td>
</tr>
<tr>
<td><em>Cyanophage</em> NATL1A-7</td>
<td>NC-016658.1</td>
<td>YP-005087466.1</td>
<td>YP-005087466.1</td>
<td>frameshift</td>
</tr>
<tr>
<td><em>Cyanophage</em> NATL2A-133</td>
<td>NC-016659.1</td>
<td>YP-005087552.1</td>
<td>YP-005087552.1</td>
<td>frameshift</td>
</tr>
<tr>
<td><em>Cyanophage</em> SS120-1</td>
<td>NC-016660.1</td>
<td>YP-005087553.1</td>
<td>YP-005087553.1</td>
<td>frameshift</td>
</tr>
<tr>
<td><em>Cyanophage</em> Syn5</td>
<td>NC-016661.1</td>
<td>YP-005087589.1</td>
<td>YP-005087589.1</td>
<td>frameshift</td>
</tr>
</tbody>
</table>

| *Synechococcus*       |                      |                     |                     |                     |
| phage SCBS4           | NC-016662.1          | YP-005087300.1      | YP-005087300.1      | frameshift          |

93
Figure A.1: Maximum-likelihood phylogenetic tree of Class I alpha subunit with Class II sequences from the RNRdb trimmed to a region of interest and clustered at A) 70% and B) 75% with trimmed Cyano SP clade alpha subunits. Gray branches belong to Class II. Colored branches belong to one of the five Class I subclasses, or Cyano SP as indicated in the key. Light purple branches indicate RNRdb clades without characterized members, which are assumed to be subclass Ia enzymes. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bars represent amino acid substitutions per site.
Figure A.2: Maximum-likelihood phylogenetic trees of trimmed RNRdb Class I beta subunits clustered at A) 70%, B) 75%, and C) 80% with trimmed Cyano SP beta subunits. Branches are colored by Class I subclass. Leaf dots are colored by RNRdb clade/subgroup. Both branches and leaf dots indicating the Cyano SP clade are colored light green.
Figure A.3: Maximum-likelihood phylogenetic trees of trimmed RNRdb Class I alpha subunits clustered at A) 70%, B) 75%, and C) 80% with trimmed Cyano SP alpha subunits. Branches are colored by Class I subclass. Leaf dots are colored by RNRdb clade/subgroup. Both branches and leaf dots indicating the Cyano SP clade are colored light green. Colors correspond to figures in the main body of the paper.
Figure A.4: Protein SSNs for Class I α with Class II sequences (A and B) and for Class I α-only sequences (C and D) used for phylogenetic analyses. Nodes are colored based on RNRdb group and match leaf dot colors on the trees in Figure A.3. In A and B, nodes representing Class II sequences are colored dark grey. Nodes labeled NrdJm are monomeric Class II and nodes labeled NrdJ are dimeric Class II. In all networks, light grey nodes contain sequences from more than one RNRdb group. Larger nodes represent larger clusters of sequences. In A) nodes represent sequence clusters ≥ 55% similarity and edges connect nodes with alignment scores ≥ 25. In B) nodes represent sequence clusters ≥ 60% similarity and edges connect nodes with alignment scores ≥ 40. In C) nodes represent sequence clusters ≥ 40% similarity and edges connect nodes with alignment scores ≥ 110. In D) nodes represent sequence clusters ≥ 55% similarity and edges connect nodes with alignment scores ≥ 155.
Appendix B

BIOINFORMATIC TOOLS FOR THE ANALYSIS OF RNRS WITHIN VIROMES

B.1 PASV

The Protein Active Site Validator (PASV) is an alignment-based method for the automatic identification of functional peptides that leverages the domain knowledge of its user (Figure B.1). PASV is available at https://github.com/mooreryan/pasv and was inspired by the PolA 762 caller (https://github.com/dnasko/dna_pola_762_caller). The user provides a fasta file of alignment reference sequences and a set of key residues that can be used for protein identification. The key residues are indicated as the numbered positions of the amino acids in the first protein in the alignment reference file. The final input is a fasta file of query sequences. PASV takes each query sequence and aligns it individually against the reference sequences. The amino acids in the query sequence aligning with key residue sites are documented and the query is partitioned based on those amino acids. Queries can also be partitioned based on whether they span a user-specified region of interest, which is provided as start and stop positions along the first sequence in the reference file.

PASV can be used for any protein that has a distinctive set of conserved residues. RNRs are particularly good candidates because they contain several immutable sites essential to enzyme function that are well documented and have been characterized in several organisms and viruses [18, 22, 194, 195, 196, 204].

B.2 RNR Gold

RNR Gold, an in-house RNR database, was created in response to database biases potentially negatively affecting the identification of RNRs within environmental
**Figure B.1:** Workflow of the Protein Active Site Validator (PASV). Briefly, putative RNIs gathered via homology or similarity search (query sequences) are aligned individually against a set of reference sequences chosen by the user. Each sequence is then partitioned based on the amino acids aligning with user provided key residues and based on whether they span a user specified region of interest.
datasets. Specifically, viral sequences are rare in the RNRdb [48], accounting for just 2.3% of the available Class I α, Class II, and Class III large subunit sequences. The RNRdb also experiences RNR type bias, with Class I α sequences alone making up 53% of the database. Because sequences in the RNRdb are collected from large sequence databases (i.e., GenBank [53] and RefSeq [52, 51]), these biases could not be alleviated by collecting RNRs from sequence databases ourselves. Therefore, I created RNR Gold, which consists of PASV-verified Class I α and Class II RNRs mined from a combination of sequence databases and environmental datasets (Figure B.2).

Briefly, RNR peptide sequences from the RNRdb [48] were used to create the initial query database and were searched against RefSeq [52], UniProt50 [47], the Tara Oceans viromes [136], MgOL (http://metagenomesonline.org), and the Earth Viromes [105] using MMSeqs2 [142]. Putative RNRs were then validated using PASV. Sequences that were confirmed as RNRs by PASV were combined with the initial RNRdb query sequences and searched against the same subject database. Putative RNRs from this second search that were validated as functional RNRs were joined with the previous search database to create the RNR Gold database.
Figure B.2: Construction of RNR Gold. Briefly, an initial profile-based homology search was performed using the RNRdb as the query and UniProt50, RefSeq, and additional viromes as the subject database. Potential RNRs were validated using PASV, and confirmed RNRs were combined with RNRdb RNRs to form the query database of a second homology search against the original subject database. Confirmed RNRs from the second homology search became RNR Gold.
Appendix C

CLASS I $\alpha$ AND $\beta$ SUBUNITS FROM GOV

Class I $\alpha$ and $\beta$ subunits are thought to evolve as units [17] and phylogenies of the two genes are virtually identical [30, 36]. However, these phylogenies are constructed mainly of cellular protein sequences, as few viral RNR sequences are currently present in the RNRdb [48, 30]. Because viral genomes are known to be mosaic [10, 108], it is possible that viral Class I $\alpha$ and $\beta$ genes may have different evolutionary histories. If so, analyzing only one of the genes would bias any conclusions regarding phylogenetic diversity of viral RNRs. To determine if Class I $\alpha$ and $\beta$ subunits have identical or divergent evolutionary histories in marine viruses, we mined both Class I $\alpha$ and $\beta$ subunits from the Global Ocean Viromes (GOV) [134] and compared the resulting phylogenies.

C.1 Methods

The GOV dataset includes the same samples as GOV 2.0 [130] with the exception of TOPC and some mesopelagic Tara Oceans samples. Details on sequencing data analysis can be found in Roux et al. [134].

Class I subunits were mined from translated ORFs available from http://datacommons.cyverse.org/browse/iplant/home/shared/iVirus/GOV/Contigs_set. Class I $\alpha$ subunits were collected using the same methods described in Section 3.2.2. Class I $\beta$ subunits were mined using the same process (MMSeqs2 Iterator with PASV enabled), but with a different query database and PASV options. Curated Class I $\beta$ subunits from the RNRdb (see Section 2.2.3) were used as queries for homology searching. Class I $\beta$ subunit alignment references were chosen similarly to the Class I $\alpha$ subunit references (reference something). The *Escherichia coli* subclass Ia $\beta$ subunit
was used as the primary alignment reference. Positions 48 and 122 in *E. coli* were used as the β subunit key positions. The expected pattern for a functional Class I β subunit was WY, WL, WF, or WV as the residue at position 122 (relative to *E. coli*) is not totally conserved among RNR subtypes [70, 36]. The ROI used for the Class I β subunits spanned from residues 48-356 in *E. coli*, which accounts for the majority of the β subunit and was used for phylogenetic analysis in Chapter 2 (see Section 2.2.3.2). The PASV profile for Class I β subunits was tested using previously curated β subunits from the RNRdb (see Section 2.2.3) and other members of the ferritin-like superfamily (methane monooxygenase, R2lox, fatty-acid desaturase [35, 59]) downloaded from InterPro [205]. All curated β subunits passed the filter, while all other members of the superfamily were excluded (data not shown).

Following homology searching, putative Class I α and β subunits were imported into Geneious 10.2.6 (https://www.geneious.com). Class I α subunits were checked for the same residues and trimmed to the same ROI as described in Section 2.2.3.2. Class I β subunits were checked for residues in bold in Table A.3. The residue in position 122 relative to *E. coli* was permitted to be Tyr, Leu, or Phe. Class I β subunits were trimmed to an ROI spanning from residues 48 to 356 in *E. coli*.

After manual curation and trimming, remaining Class I α and Class I β sequences from GOV were aligned separately using MAFFT [50]. Phylogenies for each subunit were inferred using FastTree [61]. Phylogenies were visualized and customized in Iroki [62].

### C.2 Results and Conclusions

The survey returned α and β subunit sequences from subclasses Ia, Ib, Id, and If. More Class I β sequences (n = 1119) were recovered from GOV as compared to Class I α sequences (n = 981). This is likely because Class I α sequences are nearly twice as long as β sequences on average [15]. Despite the greater number of β sequences, the GOV Class I α and Class I β phylogenies were virtually identical (Figure C.1). No subclass Ic or Ie β subunits were found. This indicates that, even in viruses, Class I α
Figure C.1: Class I $\alpha$ and Class I $\beta$ RNR sequences from marine viruses make identical trees. Maximum-likelihood phylogenetic trees of trimmed A) Class I $\alpha$ and B) Class I $\beta$ subunit sequences mined from the GOV dataset.

and $\beta$ subunits have identical evolutionary histories. Thus, the phylogenetic diversity of Class I RNRs in marine viruses can be understood with the analysis of only one of the subunits.
Appendix D

PCA OF RNR-CONTAINING VIRUSES BEFORE RNR CLUSTERING

Figure D.1: Clustering GOV 2.0 RNRs did not affect the overall shape of the data.  
A) PCA of GOV 2.0 samples based on CLR-transformed abundance of all RNR-containing viruses.  
B) PCA of GOV 2.0 samples based on Aitchison distance between samples.  
Aitchison distance was calculated based on the CLR-transformed abundance of all RNR-containing viruses.  
Both plots are colored by the ecological zones set forth in the original analysis of the GOV 2.0 dataset [130].  
For comparison, PCAs based on 95% clusters can be found in Figure 3.3.