SIP-informed metagenomics link novel genomic diversity to ecosystem function

Timothy J Rogers1,2, Paul Freire1,2, Bayleigh Benner3, Joe Vallino4, Gareth Trubl5, Julie Huber3, & Elaine Luo1,2

1 Department of Biological Sciences, University of North Carolina at Charlotte

2 Research Center to Predict Health & Environmental Risks, University of North Carolina at Charlotte

3 Woods Hole Oceanographic Institution

4 Marine Biological Laboratory, University of Chicago

5 Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore

Author note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Timothy J Rogers: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Paul Freire: Writing - Review & Editing; Bayleigh Benner: Writing - Review & Editing; Joe Vallino: Writing - Review & Editing; Gareth Trubl: Writing - Review & Editing; Julie Huber: Writing - Review & Editing; Elaine Luo: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Timothy J Rogers, Postal address. E-mail: [troger50@uncc.edu](mailto:troger50@uncc.edu)

Abstract

Viruses impact aquatic biogeochemistry via host cell lysis and rewiring host metabolic processes. However, their functional role in aquatic dark carbon cycling is mostly unexplored. Here, we established a method to identify active viruses targeting primary producers (chemoautotrophs), XXX, and XXX using 13C-DNA stable isotopic probing combined with metagenomics. Water samples were taken 10 meters below the surface of Siders Pond, a salt-stratified meromictic pond in Falmouth, Massachusetts. At 10 meters of depth the water is anoxic and sulfidic, making Siders Pond ideal for studying microbial interactions within an euxinic environment. Parallel incubations were conducted on the samples with equimolar levels of 12C- and 13C-DIC. Treatments were then filtered through a 0.2 m filter (cell-enriched retentate) followed by a 0.02 m filter (virus-enriched retentate) and extracted for metagenomic analysis. Calculating the difference in DNA density between the 12C- and 13C-fractions, we were able to identify the metagenomic assembled genomes of active chemoautotrophic community members as well as the virus sequences of virus populations targeting them. Most of the virus sequences belonged to the class Caudoviricetes with > 75% of sequences being novel at the Order and Family levels. Further, virus populations were found to contain auxiliary metabolic genes associated with cofactors/vitamins, amino acids, carbohydrates, and energy metabolisms.

*Keywords:* metagenome, bioinformatics, prophage, phage, prokaryotic host

*Word count:* X

SIP-informed metagenomics link novel genomic diversity to ecosystem function

# Introduction

## Primary production in aquatic environments

* (AND) Within aquatic systems dissolved inorganic carbon (DIC) is transformed into organic carbon and incorporated into biomass through the use of carbon fixation pathways by autotrophic organisms known as primary producers.
* (AND) Most primary production is carried out by photoautotrophs.
* (BUT) However, in light-limited ecosystems, primary production is primarily carried out by chemolithoautotrophs, known as dark carbon fixation (DCF).
* (AND) During DCF, organic carbon is produced through light-independent conversion of DIC using inorganic compounds as electron donors and acceptors in the absents of light.
* (AND) DCF contributes a large exent of primary production in aqutic systems.
* For example:
  + Within deep ocean regions, DCF can reach up to 30% of surface primary production in the marine pelagic redoxcline
  + Within deep sea hydrothermal vents, DCF production can be 2-4 orders of magnitude higher than photosynthetic biomass reaching this depth
  + Within terrestrial hot springs, where 35% of total MAG abundance was identified as chemolithoautotrophs(Rogers 2022), DCF accounts for up to 13% of carbon sequestered from released slab carbon with little evidence of photoautotrophic fixation (Fullerton 2021).
  + Within lake water columns, DCF contributes from 0.5 to 50% of the total carbon fixation within the whole lake.
* (BUT) Much less is known about other components of the microbial loop within aquatic systems, such as viruses.

## Viral effecton DCF

* (AND) In the photic zone of oceans, viral infections are observed in ~5-30% of the microbial cells.
* (AND) Cell lysis transforms up to 45% of cellular organic carbon into DOC and particulate organic carbon (POC).
* (AND) In fact, viral lysis in the surface ocean is estimated to transform ~150 gigatonnes of carbon annually.
* (SOMETHING HERE ABOUT LAKE SYSTEMS)
* (AND) This organic carbon can then flow into other trophic levels.
* (AND) These processes have implications for carbon sequestration, biochemical cycling, and the overall balance of carbon dioxide in the atmosphere.

## Meromictic aquatic systems

* (AND) Within meromictic bodies of water, the lack of intermixing between water layers leads to strong vertical chemical gradients leading to a stark chemocline between the mixolimnion and monimolimnion layers.
* (AND) Siders pond is a glacial kettle hole that contains a salt-stratified meromictic pond in Falmouth, Massachusetts.
* (AND) The upper fresh water portion has a salinity range between 2.5 to 5.5 ppt whereas the deep salt water has a salt content of ~13 ppt.
* (AND) The stratified nature of this pond causes the bottom water to be permanently anoxic and extremely reducing.
* (AND) This setting offers a convenient/inexpensive access to dark carbon cycelers.

## Microbial Dark carbon fixation

## Viruses as a key player in carbon cycling

* (AND) Chemosynthetic communities within aquatic systems have important viral-host relationships that affect the metabolic interactions between hosts and bio-geo-chemical cycles.
* (AND) Viruses can effect these interactions through viral lysis of host cells and/or through auxiliary metabolic genes (AMGs) they carry in their genomes and use to modulate and/or augment host metabolic processes.
* (But) However, the diversity, identity, and functional roles of viruses in dark carbon fixation remains mostly unexplored.
* Using DNA-SIP in combination with metagenomics (SIP-metagenomics), we have established a method to identify active viruses targeting primary producers (chemoautotrophs).

## Viral interactions with Cheomolithoautorophy in aquatic environments

## Motivation and proposed methods

## Roll of viruses

Viruses, likely the most numerous biological agents (CITATION), play essential diverse rows within every ecosystem (CITATION). Within prokaryotic systems, viruses modulate population sizes, alter host metabolic function, and drive evolutionary processes through horizontal gene transfer events. For example, within aquatic systems viruses kill X%-Y% of their host, some stat about biochemical cycles, and another state on the rate of genetic exchange per day (CITATION, CITATION, CITATION).

## Current methods of VH-pairing

## Limitation of these methods

## Advantages of DNA-SIP and metagenomics

Stable isotopic probing (SIP) has been used to characterize

# Methods

## Field site description and sample processing

Water samples for microbial analysis were collected from Siders Pond, a salt-stratified meromictic pond in Falmouth, Massachusetts (pond center latitude: 41.549006, longitude: -70.622039, altitude: 0 m). Stratification of this pond is due to inputs of both seawater and groundwater. These conditions lead to the water at 10 m of depth becoming anoxic and sulfidic, ideal conditions for studying microbial interactions within an euxinic environment. To understand the functional row of viruses on aquatic biogeochemical cycling and dark carbon fixation, 3 N2 pre-purged 1L glass bottles were filled with water from 10 m below the surface of Siders Pond. Two of the vials were supplemented with either 12C- or 13C-bicarbonate at an initial concentration of \*\*\*. The third bottle was sacrificed for DNA extraction as an environmental control. No duplicate samples were taken.

## DNA extraction and density-gradient centrifugation

The uninoculated sample was collected for DNA extraction at the start of the SIP incubation (time 0), as well as both the 12C- and 13C-bicarbonate inoculated bottles after 7 days of incubation. After 7 days of incubation both the 12C- and 13C-bicarbonate inoculated bottles were sacrificed for DNA extraction. Inoculated bottles were filtered through 0.2 m sterivex filters to collect the cell enriched portion of the samples. The 0.2 m filtrate was then further filtered through 0.02 m disc filter to collect the virus enriched portion of the samples. (Need to discribe qPCR and DNA extraction process).

## Raw read processing, metagenomic assembly, and viral operational taxonomic unit identification

Raw reads were trimmed using two passes through BBDuk (v39.01; CITATION) to remove Illumina adapters and phiX with the following parameters: ktrim=r k=21 mink=11 hdist=2 tbo tpe for adapter removal during the first pass and k=27 hdist=1 qtrim=rl trimq=17 cardinality=t mingc=0.05 maxgc=0.95 for phiX removal during the second pass. Cell enriched reads from the environmental control and the inoculated samples were then assembled de novo into 3 assemblies (12C-cell enriched, 13C-cell enriched, and cell enriched environmental control) in megahit (CITATION). Virsorter2 (CITATION) was used to identify potential viral contigs (vContgs) within the cell enriched incubation and control assemblies. Cell enriched reads from the individual quality controlled fastq files were then mapped to potential vContigs from their corresponding assemblies to identify virus specific reads within the cell enriched reads. Virus specific reads from the cell enriched incubation and the virus enriched incubation as well as Virus specific reads from the cell enriched control and viral enriched controls were co-assembled in metaSPAdes with a minimum contig length of 1.5 kb to form 2 more assemblies (viral enriched control and viral enriched incubation). Virsorter2 was then used to identify viral operational taxonomic units (vOTUs) within the virus enriched assembly. The “anvi-script-reformat-fasta” program of anvi’o v7.1 (CITATION) was used to generate simplified definition lines (deflines) for each contig within the assembly sets. QUAST (vXXX; CITATION) was then used to assess quality of all 4 assembly sets. (Add in here about dereplication of vContigs before mapping. Original count 27,054. Dereplacated count 27,023)

## Metagenomic binning and taxonomic identification.

The 3 cell enriched assembles were then used to construct 158 medium to high quality metagenome assembled genomes (MAGs/bins) with ≥50% completeness and <10% contamination (Supplementary Table SX). Briefly, the Binning and Bin\_refinement modules of MetaWRAP (CITATION) were used to construct the metagenomic assembled genomes (MAGs/bins). Within the Binning module MaxBin2 (CITATION), MetaBAT2 (CITATION), and CONCOCT (CITATION) were used for curating the initial bin sets. Next, the Bin\_refinement module was used to refine the bin sets into a consensus bin set for each assembly which were 50% completeness and 10% contamination (Table SX) for a total of 227 bins. Finally, the consensus bin sets were then dereplicated using dRep (CITATION) into one consensus bin set (our final MAGs) of 158 MAGs ( 34 from 12C treament, 30 from 13C treatment, and 94 from the environmental control). dRep was used for this step as it clusters MAGs based on simularity and selects the highest quality MAG from each cluster as the representative for that cluster. This allows for dereplication and comparison of MAGs across multiple assemblies. CheckM v1.1.3 was used for both the Bin\_refinement module and dRep to evaluate MAG completeness and contamination based on prokaryotic lineage-specific marker genes.

## Annotation and taxonomic assignment

DRAM (CITATION) and Metacerubus (CITATION) were used for MAG gene annotation. Metacerubus uses the KEGG (CITATION), COG (CITATION), VOG (CITATION), PHROG (CITATION), and PFAM (CITATION) databases. (Add something in here about how AMGs were identified). KRAKEN2 (CITATION) was used for assigning taxonomy to contigs within all four assembly sets, while the Genome Taxonomy Database Toolkit (GTDB-TK (CITATION); version X.X.X) was used for taxonomic identiﬁcation of the MAGs. For consistency, GTDB nomenclature has been used for MAGs throughout this manuscript. We refer to other commonly used names when clarity is necessary.

## Quantifying taxon-specific isotope incorporation

The “anvi-gen-contigs-database” program of anvi’o was used to construct contig databases. Bowtie2 (CITATION) and samtools (CITATION) were used in conjunction with aniv’o’s “anvi-profile” program to calculate the average depth of coverage within the 2nd and 3rd nucleotide quantiles across each contig (Q2Q3) for all assembly sets. To caculate MAG read coverage, the Q2Q3 coverage of all contigs within each MAG were summed together. Coverage dataframes were then called into Rstudio (CITATION) using the R (CITATION) data.table package vX.X.X. (CITATION). The function ‘decostand’ function of the vegan package vX.X (CITATION) was used to convert the Q2Q3 coverage into relative abundance. The excess atom fraction (EAF) of 13C within each contig was calculated using the methods laid out by Hungate et al. (2015) with the following modifications. The total genomic copy per L of contig within density fraction was calculated by multiplying the total number of genome copies (; determined by either qPCR for the cell enriched fractions or DNA yeild ratio for the viral enriched fraction) of density fraction () by the relative abundance () of contig within density fraction . Additionally, instead of calculating the GC content using the regression formula, we quantified the GC content using the program seqkit (CITATION) with the following parameters: ‘seqkit fx2tab –name –only-id –gc > results.txt’. Due to our experiment not being replicated, we were unable to use the means and standard errors approach to calculating confidence intervals for EAF incorporation that have been used previously. However, we were able to use a method laid out by Starr et al. 2021. In short, this method involves plotting the EAF against the ranked EAF and running a segmented linear regression against the resulting spline function to identify a break point at which 13C incorporation can be inferred (Fig. SX). The function segmented from the R package segmented (CITATION) was used identify the breakpoint. This breakpoint was identified at rank 9067.20 with a standard error of 3.28. Three times the standard error was added to this breakpoint to identify the EAF threshold of XXX% (Fig. X). Values above this threshold were interpenetrated as having detectable incorporated the 13C.

# Result

* (AND) To determine if viral carbon incorporation rates match that of their hosts, we constructed metagenomic libraries of the DNA-SIP incubations derived from samples take at 10m below the surface of Siders pond.

## General overview of the MAGs

* (AND) We recovered a total of 158 medium to high quailty non-redundant MAGs (CITATION:MIMAGs).
* (AND) 154 bacteria MAGs and 4 archeae MAGs are 50% complete with 5% multi-genomic contamination.
* (AND) Of these, 46 MAGs are 90% complete with 10% contamination.
* (BUT) However, only XXX of these are high quality based on MIMAGs criteria (CITATION (Supplementary Table SX)).
* (AND) The most abundant pylum was XYZ, with X, Y, and Z in much lower abundances
* (AND) The majority of these MAGs are lineages that have no cultured representative at the family level or higher (X% of archaea and X% of bacteria).

### Quantifying isotopic incorporation and predicting Metabolic capabilitys in MAGs.

* (AND) We found that shifts in DNA density reflected the carbon cycling metabolic predictions (FIG. X).
* (AND) MAGs that were predicted to have a carbon fixation pathway had significantly larger shifts in DNA density than did those that did not have a predicted carbon fixation pathway (FIG. X).
* For Thiomicrorhabdus (bin.XX) contains the complete Calvin Benson cylce and has a shift of XXX, while no carbon fixation pathway was predicted for Anaerolineales which had a shift of XX.
* (AND) By plotting the EAF of 13C against the AF of 13C, we found that binned contigs for clusters with contig members of the same MAG and that those with predicted carbon fixation pathways have an EAF that is well above the threshold of XX%.

## General overview of the vOTUs

* (AND) From both the cell- and viral- enrichment, we were able to identify XX non-redundnet low quality to complete viral populations (vOTUs) with a sequence lenth of 5 kbp.
* (AND) Of these, X were low quality, X were medium quality, X were high quality, and X were complete.
* (AND) X novel polulations are not represented in publicly availible viral data bases.
* (AND) Across all fractions, vOTUs requited X% of reads from the virus-enrichment and X% from the cell-enrichment
* (AND) The most abundant taxonomy was X, Y, and Z.

### VC ratio (maybe change to EI ratio for extracellular and intracellular abundance ratio).

### Viral AMGs.

-AMGs

### Quantifying isotopic incorporation within the vOTUs.

-^13^C incorporation

### Virus-host pairing.

-iPHoP  
-^13^C incorporation overlay

# Discussion

#

# References

#::: {#refs custom-style=“Bibliography”} #:::