Pacific Northwest
NATIONAL LABORATORY

PNNL-39468

NW-BRaVE T3 Hydroplane Project Close

Project Close-out for T3 Hydroplane Analysis

Publish Date (June 2026)

Connah G Johnson
David D Pollock

U.S. DEPARTMENT of ENERGY

Prepared for the U.S. Department of Energy under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-39468

NW-BRaVE T3 Sampling Project Close

Project Close-out for T3 Hydroplane Analysis

Publish Date (June 2026)

Connah G Johnson
David D Pollock

Prepared for the U.S. Department of Energy under Contract DEAC0576RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Executive Summary

Thrust 3 of the Northwest Biopreparedness Research in a Virtual Environment was an expansive project including method development, sample collection and sequence analysis. The sampling occurred over a multi-year period to generate metagenomic datasets that inform cyanophage-picocyanobacterial interactions in the Salish Sea across a moderate timeframe and geographical range. Part of the thrust's aim was to validate how much experimental structural and multiomics work in a model organism (*Prochlorococcus Marinus*, str. MED4) from thrusts 1 and 2 would carry over into a broader range of related organisms in the natural world, to address a fundamental question in scientific preparation for epidemics: whether and how much experimental information from known and experimentally tractable species can translate to actionable biological information in unknown species. In other words, thrust 3 aimed to find out whether the model organism experiments matter in terms of how organisms interact. This report updates work described in Johnson and Pollock 2025 (1).

A key analytical aim was to examine evidence for adaptation related to interactions in the Salish metagenomic sequences by integrating with reference genome sequences and looking at signals of coevolution and convergence. However, it became clear as we proceeded that existing tools were insufficient to meet our needs. For reference, we aligned reference genes and genomes, created phylogenetic trees, assembled contigs, and were on the way to assembling genomes (MAGs) as well as phylogenetic placement of Salish sequences, but the assumptions made during these processes tend to annihilate critical information, or worse, bias towards specific outcomes. For these reasons Hydroplane was born as a means to implement a fast, one-shot Bayesian approach to metagenomic evolutionary analysis. In our first paper we validated the speed, sensitivity, and specificity of the approach in *Prochlorococcus* sequences. Our next aims were to integrate the compressed analytical database across multiple genomes and begin dissecting the structure of *Prochlorococcus* and then *Synechococcus* variation in our samples as well as other samples collected worldwide, setting the stage for a full description of origins, convergence, and coevolutionary trends. Here we present a few additional analytical results obtained before the project was terminated midstream.

Acknowledgments

This work was supported by the NW-BRaVE for Biopreparedness project funded by the U. S. Department of Energy (DOE), Office of Science, Office of Biological and Environmental Research, under FWP 81832. Pacific Northwest National Laboratory is a multi-program national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830. A portion of this research was performed on independent faculty support funds at the University of Colorado School of Medicine (Anschutz Medical Campus).

Acronyms and Abbreviations

NW-BRaVE	Biopreparedness Research in a Virtual Environment
PNNL	Pacific Northwest National Laboratory

Contents

Figures

1. Introduction

Microbes and viruses interact strongly with each other over evolutionary time, driving adaptation and innovation in each other to competitively manage cellular resources in a battle to survive. This management is done through controlling a complex web of interacting genes and regulatory effects, and in response to diverse biotic and abiotic environments. To interpret these interactions requires detailed modeling of the evolutionary dynamics of individual genes, particularly to detect convergence and coevolution. Briefly, the analytical aim of the T3 thrust was to examine evidence for adaptation related to interactions in the Salish metagenomic sequences by integrating with reference genome sequences and looking at signals of coevolution and convergence. However, it became clear as we proceeded that existing tools were insufficient. For reference, we aligned reference genes and genomes, created phylogenetic trees, assembled contigs, and were on the way to assembling genomes (MAGs) as well as phylogenetic placement of Salish sequences, but the assumptions made during these processes tend to annihilate critical information, or worse, bias towards specific outcomes. For these reasons Hydroplane was born as a means to implement a fast, one-shot Bayesian approach to metagenomic evolutionary analysis, and feed into our planned probabilistic analyses of origins, coevolution, and convergence across genomes and metagenomes. This analysis is designed to rapidly detect and organize homologous sections of bacterial and virus genomes, while doing so in a “rigorous model-based approach to evolutionary metagenomics that avoids treating inferences as empirical data”. In our first paper we validated the speed, sensitivity, and specificity of the approach in *Prochlorococcus* sequences. Our next aims were to integrate the compressed analytical database across multiple genomes and begin dissecting the structure of *Prochlorococcus* and then *Synechococcus* variation in our samples as well as other samples collected worldwide, setting the stage for a full description of origins, convergence, and coevolutionary trends. Below, we present some additional analytical results obtained prior to project ramp down.

2. Materials and Methods

The methods used here were as given in Johnson and Pollock 2025 (1), but in brief, Hydroplane is written in the Go programming language. Code, examples, and instructions are available at github.com/NWBRaVE/Hydroplane and github.com/PollockLaboratory/hydroplane and was derived from previously published code base github.com/PollockLaboratory/AnCovMulti. No new code base is presented in this report. The details of the statistical justification and validation for our approach are given in Johnson and Pollock 2025 (1). The basic concept is to build up what we call *pclouds* of kmers probably related by divergence in a sample, which are organized into short clusters in a genome region we call *storms*, which are then further organized into observed long chains of storms in individual genomes or contigs which we call *tempests*. This approach was validated to be quite fast. We analyzed the same 109 complete *Prochlorococcus* genomes described in Johnson and Pollock 2025 (1), and where metagenomic sequences are considered, they come from Boise et al, 2026 (2).

3. Results and Discussion

In our earlier paper (1), we demonstrated the speed and sensitivity of analysis in pairwise comparisons among *Prochlorococcus* genomes. While we focused on the approach's statistical robustness (empirical sensitivity and specificity allowing clean separation of signal from noise depending on parameter choice), the next obvious step is to focus on building up the *pclouds* at probably homologous sites across multiple genomes. In a test run to demonstrate practical feasibility with the existing Hydroplane software, we built clouds on a focal core of 11 genomes, which were then tested on a further set of 109 genomes (the same 109 complete *Prochlorococcus* genomes described in Johnson and Pollock 2025 (1)). In figure 1 we show the results for a storm of size 20 non-overlapping kmers, with the divergent kmers at predicted homologous positions stacked to the right as they were found (Figure 1, top). In this example storm, there were 3 to 7 different kmers identified, and visual inspection reveals that the kmers on any line are quite similar to one another, as expected if they are homologous kmers. Across different storms and across different positions within each storm, it was typical that the number of different kmers identified varied considerably from 1 to, rarely, the maximum 11. We note that the kmers identified were not identified by similarity, but by position within the storm, bounded by at least three identical kmers within the same storm, and gaps between identified kmers limited to +/- three nucleotides. This approach allows that some identified kmers at homologous positions may be offset from the founding or focal kmer for a given *pcloud*, with the offset among kmers deferred until post-analysis to avoid unnecessary (and compute-intensive) repetition during the identification process.

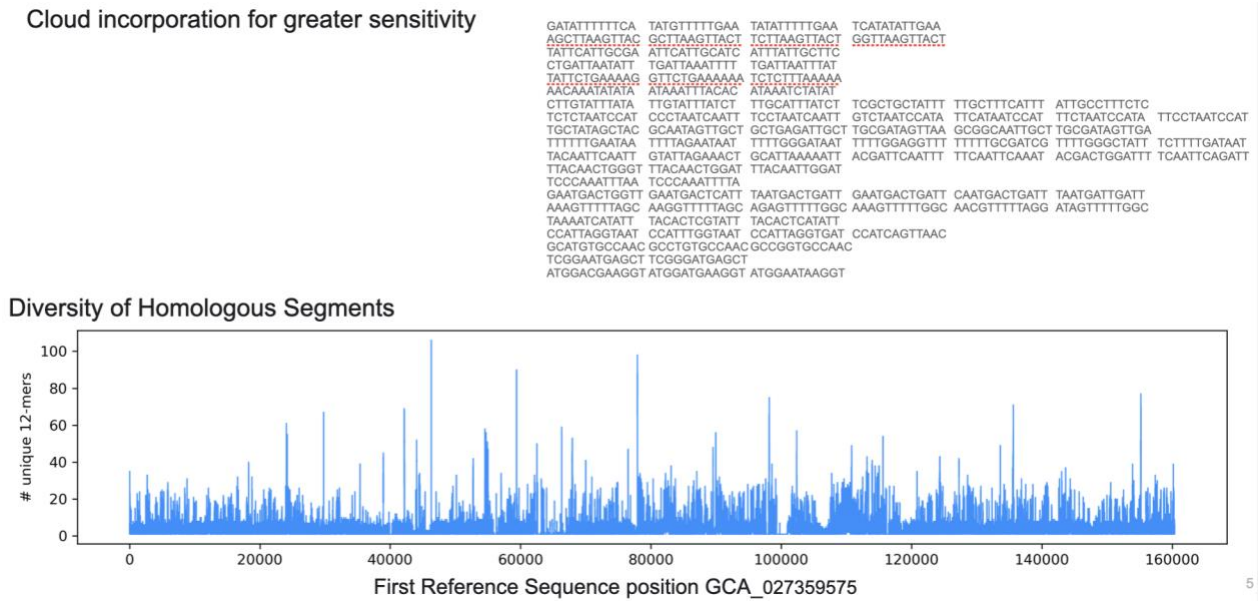

Figure 1. Cloud incorporation and homologous segment diversity.

As further demonstration, the clouds were then compared against the 109-genome dataset to visualize the propensity to identify sites and the diversity among an expanded set of genomes (Figure 1, lower). We can see here that a huge number of sites were identified, a large majority of sites had 10 or fewer unique 12mers identified, and nearly all had fewer than 20-40 unique 12mers. At the same time, the scan of each of the 109 genomes took only slightly longer than the scan of each genome from storms without built-up *pclouds*, demonstrating the reduction of the overall speed of analysis from an O(N^2) problem (in this case, 109x109 ~ $1.2*10^4$) to an O(N) problem (here, 109). Furthermore, we demonstrate that an effective analysis can be obtained by building up clouds on only 11 out of 109 genomes. In principle that means that *storms* of *pclouds* can be built to be highly sensitive yet specific with a minimal amount of computation. We also note that the *stormclouds* here were built up in sequential fashion through the 11 genomes, whereas it is likely that a subsequent pass through the 11 or the pass through the 109 could identify frequencies of kmers, which could be used to whittle down to the subset most likely to identify homologs accurately but not so big that it dramatically increased noise levels.

While a preliminary test run on some metagenomic sequences from Boise et al, 2026 (2) indicated that these *stormclouds* can easily be applied, as expected, to identify *Prochlorococcus*-derived sequence reads in metagenomic samples, there are a number of simple steps that need to be complete before this is sensible. First, there should be simple repetition of parsing the genomes to create clouds so that there is less genome order dependence in creating the clouds. Second, new storms need to be created for regions of subsequent genomes that are not matched to existing storms. Additionally, it would be prudent to make sure that no *stormclouds* start to collect kmers in runaway fashion; we have not yet observed this, but it seems possible, and there are rare cases in the current run with kmer counts over 60 among the 109 genomes. In any case, as we start to build very large *stormclouds*, it is necessary to evaluate whether the stringency of detection should be increased, or again if *pcloud* counts should be limited by frequency to create well balanced, sensitive yet specific detectors. These steps should not be too difficult to implement if future funding becomes available.

Should Hydroplane development resume, immediate goals would be to build *stormclouds* for a larger set of reference *Prochlorococcus* genomes, map the recombination of segments and rates of evolution of all these genomes, and then apply these to all available ocean metagenomes. Followed by similar analyses of the related *Synechococcus* genomes and cyanophage known to target either or both, this would set the stage for a probabilistic analysis of convergent events and origins of all gene recombination. The speed and efficiency of analysis using Hydroplane, combined with the care taken to evaluate the statistical outcomes of hydroplane analyses, is expected to make such analyses straightforward and speedy, even with numerous metagenomes. We have not overlooked the fact that it will also be possible to perform *de novo* analyses of metagenome samples using hydroplane, something we had considered doing: first in a sampling approach code-named *non idem*, which would theoretically greatly improve diversity estimators such as *non pareil* (3)*;* and second in a more thorough approach to build *stormclouds* that could then be linked and reverse mapped to known genomes. Further, an *amino acid aware* version of Hydroplane could link more deeply diverged but homologous sequences, either through translation to amino acids prior to *pcloud* construction, or more simply through using longer kmers that allow for synonymous differences.

4. Closing Remarks

The work described here summarizes a small piece of the analytics for an ambitious project integrated with the original vision of DOE's Bio-preparedness Research in a Virtual Environment (NW-BRaVE). The closing remarks for a major portion of the project, the sampling, sequencing, and preliminary analysis, are provided in a companion paper (4), and we won't repeat them here, but it is worth crediting the DOE's earlier "foresight to see the valuable and irreplaceable aspects of environmental sequence information as a means to evaluate molecular interactions within and between molecules and organisms in their natural habitats". Here is but the first phase of a system we call Hydroplane designed to rapidly identify convergence in metagenomic samples with associated reference genome sequences without incorporating and ossifying incorrect alignment and phylogenetic inference (1), a capability that is not currently available. We acknowledge that readers may find the results presented here a bit incomplete and dissatisfying; this report is meant to ensure that although incomplete, they are documented. It would have been exciting to integrate this work with our work on ecotypic divergence (5), our work on integrated harmonization of biotic and abiotic information with sequencing results (6), and were prepared to integrate multi-omics and structural annotation data for model organisms from other BRaVE thrusts. This work is also notably prepared to integrate with scientific AI-related goals, particularly in the areas of integrating multiple kinds of convergence and more subtle convergence of kind of molecular features, which undoubtedly occur but may not be easily recognizable by human eyes, as well as in integrating notoriously disparate and variably certain structural and omics data that may or may not be extensible to distant organisms or ecological situations. We hope that we will be able to remedy this in the future.

5. References

1. C. G. M. Johnson, D. D. Pollock, Hydroplane I: one-shot probabilistic evolutionary analysis for scalable organizational identification. *bioRxiv* 10.1101/2025.10.14.682387, 2025.2010.2014.682387 (2025).
2. N. R. Boise *et al.*, Sampling microbial dynamics in the Salish Sea estuary: evaluating methods to capture cyanobacteria and cyanophage. *Frontiers in Marine Science* **Volume 13 - 2026** (2026).
3. R. L. Rodriguez, K. T. Konstantinidis, Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**, 629–635 (2014).
4. N. R. Boise, Leiser, Owen P., Johnson, Connah G., Brooks, Kyle L., Edmundson, Scott J., Pollock, David D. (2026) NW-BraVE T3 Sampling Project Close.
5. E. Brenner, C. Vang, C. Johnson, J. Ravi, Genotype-phenotype modeling of light ecotypes in Prochlorococcus reveals genomic signatures of ecotypic divergence. *bioRxiv* 10.1101/2025.10.13.678797, 2025.2010.2013.678797 (2026).
6. A. George *et al.*, EcoKMER: One-stop shop for spatio-temporal metagenomic exploration using DataFed. *bioRxiv* 10.64898/2025.12.15.694490, 2025.2012.2015.694490 (2025).

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov