RNA-seq gene expression data for Synechococcus sp. WH8102 under continuous high, mid, and low phosphorus treatments in laboratory EFB experiments in 2025

Website: https://www.bco-dmo.org/dataset/998248
Data Type: experimental
Version: 1
Version Date: 2026-05-08

Project
» Collaborative Research: Assessing the role of polyphosphate production and cycling in marine ecosystem functioning (Polyphosphate production and cycling)

Contributors	Affiliation	Role
Filella, Alba	University of Arizona (UA)	Scientist
Mickle, Audrey	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset contains normalized RNA-seq gene expression counts from Synechococcus sp. WH8102 grown under controlled laboratory phosphorus treatments (High-P, Mid-P, and Low-P) in 2025. Gene expression was measured across three biological replicates per treatment and includes 2,718 genes with locus_tag identifiers, along with gene names and functional product annotations where available. Normalized counts were generated using DESeq2 (median-of-ratios size factor method), enabling comparison across samples and are provided in the supplemental table 998248_v1_Syn_WH8102_normalized_counts_annotated.csv. Many genes lack assigned names, reflecting the prevalence of hypothetical or predicted proteins typical of cyanobacterial genomes. In addition to normalized counts, the dataset includes a differential expression table reporting pairwise comparisons between treatments as log2 fold changes (primary dataset: 998248_v1_log2fc_gene_expression_wh8102.csv), along with adjusted p-values and directionality of expression changes. Mean expression values derived from variance-stabilized (VST) data are also provided to facilitate interpretation of expression trends. Together, these data tables provide a comprehensive resource for examining phosphorus-responsive gene expression and patterns of differential regulation in Synechococcus sp. WH8102.

Table of Contents

Coverage

Location: Laboratory experiment
Temporal Extent: 2025 - 2025

Methods & Sampling

RNA samples were collected from *Synechococcus* sp. WH8102 cultures grown under continuous laboratory phosphorus treatments (High-P, Mid-P, and Low-P), with three biological replicates per treatment. On the final

day of the experiment, ~700 mL of each culture was filtered through 47 mm, 0.4 µm GFF filters (previously combusted at 450°C) using a peristaltic pump and stored at -80°C until analysis.

Total RNA was extracted from cells collected on filters using a modified TRIzol-based extraction protocol. Filters were transferred aseptically into screw-cap bead tubes containing 0.1 mm disruptor beads and kept on ice after removal from −80 °C storage. Cells were enzymatically lysed in lysis buffer (30 mM Tris, 10 mM EDTA, and 10 mg mL^{-1} lysozyme) and incubated at 37 °C for 30 min with slow rotation. Samples were subsequently bead-beaten by vortexing at maximum speed for 5 min to ensure mechanical disruption. Following lysis, RNA was extracted using TRIzol Reagent (Thermo Fisher Scientific, Waltham, MA, USA). TRIzol and chloroform were added to each sample, followed by vigorous mixing and incubation at room temperature prior to centrifugation to separate phases. The aqueous phase containing RNA was transferred to a new tube and RNA was precipitated with saline solution and ice-cold isopropanol. Samples were incubated at −20 °C and centrifuged to pellet RNA. The RNA pellet was washed with 70% ethanol, briefly air-dried, and resuspended in nuclease-free water. To remove contaminating DNA, RNA samples were treated with TURBO DNase (Thermo Fisher Scientific) according to the manufacturer's instructions. A subsequent phenol–chloroform–isoamyl alcohol purification step was performed, followed by chloroform extraction and ethanol precipitation to further purify RNA. The final RNA pellet was washed with 70% ethanol, air-dried, and resuspended in nuclease-free water. RNA concentration and purity were assessed using NanoDrop Spectrophotometer (Thermo Fisher Scientific) and Qubit Fluorometer (Thermo Fisher Scientific).

Total RNA samples were submitted to SeqCoast for library preparation and sequencing. Ribosomal RNA (rRNA) was depleted and sequencing libraries were prepared using the Illumina Stranded Total RNA Prep Ligation Kit with Ribo-Zero Plus Microbiome (Illumina, #20072063) and Illumina Unique Dual Indexes, following the manufacturer's protocol. Sequencing was performed on an Illumina NextSeq 2000 platform using a 300-cycle XLEAP-SBS flow cell to generate 2 × 150 bp paired-end reads. A 1–2% PhiX spike-in control was included to support optimal base calling. Base calling, demultiplexing, adapter trimming, and initial run quality control were performed using DRAGEN (v4.2.7; Illumina) onboard the NextSeq 2000 system. Sequencing quality was assessed at both the run level and per-sample level, including inspection of FastQC reports. FastQ files corresponding to paired-end reads were generated for downstream analysis.

The reads were mapped to the *Synechococcus* sp. WH8102 genome obtained from the JGI Genome Portal using Bowtie2 v.2.4.1 (Langmead et al., 2012) with default parameters in paired-end mode. Counts were generated with the featureCounts function in the Rsubread v.2.12.3 R Bioconductor package (Liao et al., 2019).

Differential gene expression analysis was performed using DESeq2 (v1.42.1). Raw read counts were imported into a DESeqDataSet object with experimental conditions (Low-P, Mid-P, High-P) specified as the design factor. Size-factor normalization was applied using the median-of-ratios method to account for differences in sequencing depth, and gene-wise dispersion estimates were fitted under a negative binomial generalized linear model framework. Two complementary DESeq2 approaches were used to assess differential expression. First, a likelihood ratio test (LRT) was applied to compare a full model including the condition effect against a reduced model excluding it, in order to identify genes exhibiting any significant expression variation across treatments. Second, pairwise contrasts between conditions (Low-P vs High-P, Mid-P vs High-P, and Low-P vs Mid-P) were performed using Wald tests to estimate $\log_2$ fold changes (log2FC). These effect sizes represent shrinkage-adjusted estimates from the DESeq2 model, where positive values indicate higher expression in the first condition of each comparison. P-values from Wald tests were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) procedure, and genes with adjusted p-values (padj) < 0.05 were considered statistically significant. For visualization and exploratory analyses, variance-stabilizing transformation (VST) was applied to normalized counts using DESeq2. Mean expression values per condition were computed from VST-transformed data across biological replicates. These values were used exclusively for visualization and interpretation of expression patterns and were not used for statistical inference.

Notes on experimental design:

- Treatments: High-P, Mid-P, Low-P
- Replicates: 3 biological replicates per treatment
- RNA extractions using TRIzol reagent

Data Processing Description

The submitted dataset consists of normalized RNA-seq gene expression counts derived from raw sequencing reads. **Raw sequencing data** have been deposited in the Gene Expression Omnibus (GEO) at National Center

for Biotechnology Information (NCBI) under accession number **GSE327000**. Data processing included the following steps:

1. **Quality control and trimming:**
   - Demultiplexing, adapter trimming, and quality control were performed using **BCL Convert v3.9.3**.
   - Raw read quality was assessed with **FastQC v0.73** via Galaxy (galaxy0).
2. **Alignment:**
   - Reads were mapped to the *Synechococcus* sp. WH8102 reference genome using **Bowtie2 v2.4.1** in paired-end mode with default parameters.
3. **BAM file processing:**
   - SAM files were converted to BAM sorted and indexed.
4. **Count generation:**
   - Gene-, CDS-, and exon-level counts were generated using the **featureCounts** function from **Rsubread v2.12.3**.
5. **Counts normalization:**
   - Raw counts were normalized using **DESeq2 v1.40.2** (median-of-ratios size-factor method).
   - Normalized counts were exported as a wide table (norm_counts_annotated) containing locus_tag, gene_name (when available), product annotation, and counts for each biological replicate and treatment.
6. **Software and computing environment:**
   - Analyses were performed in **R 4.3.3 (GUI 1.80, Big Sur ARM build 8340)** using **RStudio Version 2023.12.1+402**.
   - R packages: **DESeq2 v1.40.2**, **Rsubread v2.12.3**, and standard R packages for data manipulation and plotting.

This workflow yields a comprehensive dataset combining normalized expression values, statistical significance, pairwise comparisons, and gene functional annotations.

BCO-DMO Processing Description

- Loaded data from log2_FC_final_table.xlsx (Sheet1), using row 1 as header, with empty string and "nd" values interpreted as missing
- Renamed three columns to remove hyphens for BCO-DMO parameter-naming compliance: mean_Low-P to mean_Low_P, mean_Mid-P to mean_Mid_P, mean_High-P to mean_High_P
- Output table as 998248_v1_log2fc_gene_expression_wh8102.csv
- Syn_WH8102_normalized_counts_annotated.xlsx was converted to CSV format and parameter descriptions were added to the description. The resulting file, 998248_v1_Syn_WH8102_normalized_counts_annotated.csv, is listed as a supplemental file.
- Some descriptive context was added to the description of the README – Synechococcus WH8102 RNA-seq Data (2025 P Experiment).txt file and it was listed as a supplemental file.

[table of contents | back to top]

Related Publications

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Software

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. doi:10.1038/nmeth.1923
Software

Liao, Y., Smyth, G. K., & Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research, 47(8), e47–e47. https://doi.org/10.1093/nar/gkz114
Software

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). doi:10.1186/s13059-014-0550-8
Software

R Core Team. (2024). R: A language and environment for statistical computing (Version 4.3.3) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/
Software

[table of contents | back to top]

Related Datasets

IsRelatedTo

Filella, A., Duhamel, S., Diaz, J. (2026) **Cell abundance, nutrient concentrations, and intracellular phosphorus pools in Synechococcus sp. WH8102 continuous (EFB) culture experiments.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-13 http://lod.bco-dmo.org/id/dataset/1005084 [view at BCO-DMO]

Filella, A., Duhamel, S., Segura-Noguera, M., Diaz, J. (2026) **Single-cell elemental quotas and stoichiometry of Synechococcus sp. WH8102 measured by X-ray microanalysis (XRMA) under varying phosphorus conditions.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-14 http://lod.bco-dmo.org/id/dataset/1005186 [view at BCO-DMO]

Software

RStudio Team (2023) RStudio: Integrated Development for R. Version 2023.12.1+402. RStudio, Inc., Boston, MA. http://www.rstudio.com/

References

Filella A, Duhamel S, Diaz J, Segura-Noguera M, Gomez-Buckley A, Carini P. RNA-seq gene expression data for Synechococcus sp. WH8102 under continuous high, mid, and low phosphorus treatments in laboratory EFB experiments in 2025. [Dataset]. NCBI Gene Expression Omnibus; 2026. GEO Series GSE327000. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE327000

[table of contents | back to top]

Parameters

Parameter	Description	Units
locus_tag	Unique gene identifier in Synechococcus sp. WH8102	unitless
gene_name	Common gene name (if available)	unitless
product	Functional description of gene product	unitless
go_process	Gene Ontology (GO) biological process annotation	unitless
go_function	Gene Ontology (GO) molecular function annotation	unitless
go_component	Gene Ontology (GO) cellular component annotation	unitless

mean_Low_P	Mean variance-stabilizing transformation (VST)-normalized expression value, averaged across three biological replicates for the Low-P condition. Used for visualization/interpretation only; not used for statistical inference	VST units
mean_Mid_P	Mean variance-stabilizing transformation (VST)-normalized expression value, averaged across three biological replicates for the Mid-P condition. Used for visualization/interpretation only; not used for statistical inference.	VST units
mean_High_P	Mean variance-stabilizing transformation (VST)-normalized expression value, averaged across three biological replicates for the High-P condition. Used for visualization/interpretation only; not used for statistical inference.	VST units
log2FC_Low_vs_High	Log2 fold-change between Low-P and High-P	log2 fold-change
padj_Low_vs_High	Adjusted p-value (Benjamini–Hochberg correction) for Low-P vs High-P comparison	unitless
dir_Low_vs_High	Direction of change for Low-P vs High-P; direction of differential expression (e.g., “Up in Low-P”, “Up in High-P”)	unitless
sig_Low_vs_High	Significance label for Low-P vs High-P comparison; significance status for each comparison (e.g., significant or not significant [ns])	unitless
log2FC_Mid_vs_High	Log2 fold-change between Mid-P and High-P	log2 fold-change
padj_Mid_vs_High	Adjusted p-value (Benjamini–Hochberg correction) for Mid-P vs High-P comparison	unitless
dir_Mid_vs_High	Direction of change for Mid-P vs High-P; direction of differential expression (e.g., “Up in Low-P”, “Up in High-P”)	unitless
sig_Mid_vs_High	Significance label for Mid-P vs High-P comparison; significance status for each comparison (e.g., significant or not significant [ns])	unitless
log2FC_Low_vs_Mid	Log2 fold-change between Low-P and Mid-P	log2 fold-change
padj_Low_vs_Mid	Adjusted p-value (Benjamini–Hochberg correction) for Low-P vs Mid-P comparison	unitless
dir_Low_vs_Mid	Direction of change for Low-P vs Mid-P; direction of differential expression (e.g., “Up in Low-P”, “Up in High-P”)	unitless

sig_Low_vs_Mid	Significance label for Low-P vs Mid-P comparison; significance status for each comparison (e.g., significant or not significant [ns])	unitless
pattern	Overall expression pattern classification across conditions (e.g., “High > Mid > Low”, “Low > Mid > High”, or “complex” for non-monotonic patterns)	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina NextSeq 2000
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Sequencing was performed on an Illumina NextSeq 2000 platform using a 300-cycle XLEAP-SBS flow cell to generate 2 × 150 bp paired-end reads. A 1–2% PhiX spike-in control was included to support optimal base calling.
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

Dataset-specific Instrument Name	peristaltic pump
Generic Instrument Name	Pump
Dataset-specific Description	On the final day of the experiment, ~700 mL of each culture was filtered through 47 mm, 0.4 µm GFF filters (previously combusted at 450°C) using a peristaltic pump and stored at -80°C until analysis.
Generic Instrument Description	A pump is a device that moves fluids (liquids or gases), or sometimes slurries, by mechanical action. Pumps can be classified into three major groups according to the method they use to move the fluid: direct lift, displacement, and gravity pumps

Dataset-specific Instrument Name	Qubit Fluorometer (Thermo Fisher Scientific)
Generic Instrument Name	Qubit fluorometer
Dataset-specific Description	RNA concentration and purity were assessed using NanoDrop Spectrophotometer (Thermo Fisher Scientific) and Qubit Fluorometer (Thermo Fisher Scientific).
Generic Instrument Description	Benchtop fluorometer. The Invitrogen Qubit Fluorometer accurately and quickly measures the concentration of DNA, RNA, or protein in a single sample. It can also be used to assess RNA integrity and quality. Manufactured by Invitrogen, Carlsbad, CA, USA (Invitrogen is one of several brands under the Thermo Fisher Scientific corporation.)

Dataset-specific Instrument Name	NanoDrop Spectrophotometer (Thermo Fisher Scientific)
Generic Instrument Name	Thermo Scientific NanoDrop spectrophotometer
Dataset-specific Description	RNA concentration and purity were assessed using NanoDrop Spectrophotometer (Thermo Fisher Scientific) and Qubit Fluorometer (Thermo Fisher Scientific).
Generic Instrument Description	Thermo Scientific NanoDrop spectrophotometers provide microvolume quantification and purity assessments of DNA, RNA, and protein samples. NanoDrop spectrophotometers work on the principle of ultraviolet-visible spectrum (UV-Vis) absorbance. The range consists of the NanoDrop One/OneC UV-Vis Spectrophotometers, NanoDrop Eight UV-Vis Spectrophotometer and NanoDrop Lite Plus UV Spectrophotometer.

[table of contents | back to top]

Project Information

Collaborative Research: Assessing the role of polyphosphate production and cycling in marine ecosystem functioning (Polyphosphate production and cycling)

Coverage: Mediterranean Sea and California Current Ecosystem

NSF Award Abstract:
Phosphorus (P) is a vital nutrient required by all forms of life. In the ocean, which sustains half of global photosynthesis and oxygen supply, P can be scarce enough to constrain biological productivity, carbon dioxide uptake, and therefore climate. Human activity is accelerating the delivery of nutrients like P to the ocean, but rates of nitrogen inputs are even greater. This imbalance may increase levels of P stress in marine ecosystems, placing ocean productivity more and more under the control of P supply. Given the critical role of P in sustaining ocean health and ecosystem services now and into the future, a comprehensive understanding of its utilization and fate in the marine environment is necessary. In this project, the research team investigates marine polyphosphate (polyP), a ubiquitous yet poorly understood form of P made by all living organisms. To close major knowledge gaps on marine polyP, the investigators are overcoming major technical barriers to produce the first quantitative measurements of polyP in marine microorganisms. These measurements are being conducted on laboratory microbial cultures, as well as field samples from environments with high P supply, such as the California Current Ecosystem, or very low P supply, such as the Mediterranean Sea. In addition, the research team is resolving the cellular function of marine polyP across these different organisms and environments in order to clarify its role in biological P nutrition. This work helps advance polyP research across disciplines, including terrestrial science and even cancer research, and has broad application to P bioremediation. This project supports a postdoctoral researcher, a graduate student, and several undergraduate students in the labs of two female scientists. New educational tools to teach the public about marine polyP are being produced and disseminated through this project. A K12 teacher is participating in the work and communicating the findings to their classrooms and to broad audiences online.

PolyP is ubiquitous in marine systems, where it plays critical roles in microbial P nutrition and P mineral formation and sequestration. In these ways, polyP has the potential to shape long- and short-term marine P cycling, primary productivity, microbial ecology, and global climate. However, major knowledge gaps still exist. Due to technical limitations, the scientific community currently lacks a quantitative understanding of marine polyP pools, their chain lengths, and biological origins. Furthermore, given the view of polyP as a P storage molecule, recent observations pointing to the preferential retention of particulate polyP in low phosphate (Pi) environments raise new questions about its ecophysiological functions. To close these knowledge gaps, two research questions are addressed: Q1: What is the total content and chain length distribution of polyP across different microbial groups and environmental conditions? Q2: How do the production of particulate polyP contribute to microbial P demand and stoichiometry across a broad range in Pi availability? The following hypotheses are tested: H1 (Q1): Functionally and environmentally diverse plankton produce a broad range of

polyP chain lengths and concentrations. H2 (Q2): The preferential retention of polyP in low Pi environments can be reconciled with its role as a P storage molecule by a combination of taxonomic and physiological factors. These hypotheses are tested in the laboratory using representative cultures of marine plankton and in the field using observational approaches along natural Pi gradients in the Pacific Ocean and Mediterranean Sea. Applying a new P-targeted method using mass spectrometry, the team is resolving an unprecedented level of detail in marine polyP content and speciation. By combining cell sorting with elemental, biochemical, and radiotracer analyses, the team is gaining a mechanistic understanding of polyP physiology and its cycling in the ocean.

This project is supported by the Biological Oceanography and Chemical Oceanography Programs. This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-2245249

[table of contents | back to top]