RNA taxonomy data from an iron incubation experiment during the PUPCYCLE II cruise (R/V Sally Ride) within the California Current System in May and June of 2023

Website: https://www.bco-dmo.org/dataset/1003525
Data Type: Cruise Results, experimental
Version: 1
Version Date: 2026-08-12

Project
» CAREER: An integrated molecular and physiological approach to examining the dynamics of upwelled phytoplankton in current and changing oceans (Upwelled Phytoplankton Dynamics)

Contributors	Affiliation	Role
Marchetti, Adrian	University of North Carolina at Chapel Hill (UNC-Chapel Hill)	Principal Investigator
Cohen, Natalie	University of Georgia (UGA)	Scientist
Schnetzer, Astrid	North Carolina State University (NCSU)	Scientist
Cook, Claire	University of Georgia (UGA)	Student
Jeong, Yubeen	University of North Carolina at Chapel Hill (UNC-Chapel Hill)	Student
Lim, Prisca	University of North Carolina at Chapel Hill (UNC-Chapel Hill)	Student
McClure, Will	North Carolina State University (NCSU)	Student
Speciale, Emily	University of North Carolina at Chapel Hill (UNC-Chapel Hill)	Student
Gerlach, Dana Stuart	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset includes metatranscriptomic (RNA-seq) taxonomy results from an iron incubation experiment using upwelled waters sampled in the California Current System (CCS) during the PUPCYCLE II cruise with Chief Scientist Adrian Marchetti. PUPCYCLE II (Phytoplankton response to the UPwelling CYCLE) took place onboard the R/V Sally Ride from May 29th to June 10th, 2023. Freshly upwelled seawater for the incubation experiment was collected within the northern CCS and placed into designated cubitainers. Three cubitainers were immediately harvested for the initial timepoint (T0). The remaining twenty-seven cubitainers were assigned treatments: nine were unamended (Ctrl), nine were amended with 5 nM FeCl2 (Fe), and nine were amended with 200 nM Desferrioxamine B, a strong iron chelator (DFB). Three cubitainers from each treatment were harvested for each of the three subsequent timepoints: 48 hours (T1), 168 hours (T2), and 264 hours (T3) after incubation. Approximately 4 liters of each sample was harvested for RNA-seq of the natural phytoplankton community, or metatranscriptomics. The dataset provides results of RNA-seq analyses for relative taxonomic compositions of each sample, focused on broad protist groups, mixotroph genera, and diatom genera. These results show the differences in transcriptional activity of phytoplankton taxa groups between different iron regimes, specifically the domination of diatoms in Ctrl/+Fe and the steady presence of mixotrophs in DFB.

Table of Contents

- Funding

Coverage

Location: California Current System, off the southern coast of Oregon at 43°02'42.7"N, 124°33'07.2"W
Spatial Extent: **Lat**:43.045194 **Lon**:-124.552
Temporal Extent: 2023-05-29 - 2023-06-10

Dataset Description

This dataset is one of many generated from an iron incubation experiment conducted as part of the PUPCYCLE II cruise in May and June 2023. The different analyses are listed here, with links to other datasets from this study in the Related Datasets section.

Analyses include:

- Physiology (dataset 994607)
  - Dissolved inorganic nutrients
  - Size-fractionated chlorophyll a
  - Size-fractionated isotope uptake rates
  - Flow cytometry
  - FlowCAM analysis
  - FLP experiments for mixotroph grazing
- **RNA taxonomy data (this dataset 1003525)**
- Differential expression: Pfam
  - Mixotroph Pfam (dataset 1004659)
  - Diatom Pfam (dataset 1004674)
- Differential expression: KEGG gene (dataset 1004689)
- Differential expression: KEGG pathways (dataset1004704)
- Dissolved iron concentrations (dataset 1004007)

See Related Datasets section below for links to above mentioned datasets.

Methods & Sampling

Data collection took place on the R/V Sally Ride from May 29th to June 10th, 2023. To simulate upwelling conditions under different iron treatments, an onboard incubation experiment was conducted. Seawater for the incubation experiment was collected within the northern CCS, off the southern coast of Oregon at 43°02'42.7"N, 124°33'07.2"W. The collected seawater was deemed as freshly upwelled water and taken from a depth of 55 m – which corresponds to a depth slightly below the euphotic zone receiving less than 1% irradiance – using trace-metal clean techniques on May 30th, 2023, 13:30 GMT. The seawater was pumped and homogenized using trace metal clean techniques, then transferred into a total of thirty 20 L low-density polyethylene cubitainers. Three cubitainers were immediately harvested for the initial timepoint (T0). The remaining twenty-seven cubitainers were assigned treatments: nine were unamended (Ctrl), nine were amended with 5 nM $FeCl_2$ (+Fe), and nine were amended with 200 nM desferrioxamine B, a strong iron chelator that inhibits dissolved iron uptake (DFB). The cubitainers were placed in an on-deck incubator covered with two layers of neutral density screening to simulate 26% of incident irradiance supplied with flow-through surface seawater to maintain ambient surface temperature. Three cubitainers from each treatment were harvested for each of the three subsequent timepoints: 48 hours (T1), 168 hours (T2), and 264 hours (T3) after incubation.

RNA Collection and Metatranscriptomics. Cubitainers were harvested for RNA by collecting and filtering approximately 2.5 to 4 L of each sample onto 0.8 µm Pall Supor filters (142 mm) using a peristaltic pump; samples were then immediately flash frozen in liquid nitrogen and stored at -80°C. RNA was extracted using the RNAqueous-4PCR kit, per instructions from the manufacturer with few modifications – filters were cut up due to their large size, 200 µL of glass beads were added, and 3 mL of lysis buffer was added. RNA samples were sent to GENEWIZ for library preparation and sequencing with PolyA tail selection. Sequencing was performed on an Illumina HiSeq 4000 with a 2x150 bp configuration. GENEWIZ provided raw paired-end read sequences for each sample.

Raw reads were trimmed using Trim Galore v0.6.10 (Krueger, 2015) and quality control was assessed through FastQC (Andrews, 2010). A *de novo* metatranscriptome assembly was conducted using rnaSPAdes v3.15.5 for individual assemblies (Bushmanova et al., 2019) and CD-HIT v4.8.1 for a grand assembly (Li & Godzick, 2006). TransDecoder v5.7.1 was used to determine the most likely open reading frame (ORF) for each contig (Haas, 2023). ORFs were annotated using the Marine Functional Eukaryotic Reference Taxa (MarFERReT) database v1.1.1 (e-value < $1e^{-06}$) (Groussman et al., 2023) for NCBI taxonomies (Federhen, 2012), PR^2 taxonomies (Guillou et al., 2012) and Pfam 34.0 functions (Mistry et al., 2021). To remove any potential bacterial contamination, ORFs were also annotated against the PhyloDB v1076 database using EUKulele (e-value < $1e^{-06}$) (Krinos et al., 2020). Any ORF with a higher bitscore for a bacteria or virus compared to its MarFERReT annotation was removed (resulting in < 3% of ORFs removed). To maximize functional annotation coverage, ORFs were annotated using eggNOG-mapper v2.1.12 (Cantalapiedra et al., 2021) to obtain annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa, 2000). Trimmed samples were then aligned to the grand assembly using Salmon v1.10.3 (Patro et al., 2017) and exported into a comprehensive counts table using tximport (Soneson et al., 2016). Reads not mapping to a protist group as defined by MarFERReT annotations were removed before downstream analysis.

For taxonomic composition, all protist reads were normalized using DESeq2's median of ratios methods (Love et al., 2014). The mean and standard deviation for the number of reads mapping to each protist group and the proportion of reads mapping to each protist group were calculated for each timepoint/treatment. Mean/standard deviation for number and proportion of protist reads specifically mapping to mixotroph genera and diatom genera were also calculated. Mixotrophs were identified based on constitutive mixotroph genera within the Mixoplankton Database (MDB) (Mitra et al., 2023) and diatom genera were identified based on mapping to the lineage Bacillariophyta.

Data Processing Description

The metatranscriptomic pipeline code used for this project is publicly available at https://github.com/emilyspeciale/Speciale-Metatranscriptomics.

- CD-HIT v4.8.1
- DESeq2
- eggNOG-mapper v2.1.12
- FastQC
- Marine Functional Eukaryotic Reference Taxa (MarFERReT) database v1.1.1
- PhyloDB v1076 database using EUKulele
- rnaSPAdes v3.15.5
- Salmon v1.10.3
- TransDecoder v5.7.1
- Trim Galore v0.6.10
- tximport

BCO-DMO Processing Description

This section documents curation actions performed prior to publication review with the submitter, and additional information relevant to understanding and reusing this dataset. It distinguishes changes made to the submitted (meta)data from unresolved issues and/or enhancements that improve future reuse and interoperability.

CURATION ACTIONS PERFORMED ON DATA
- Imported source file "BCO_DMO_PUPCYCLEII_RNATaxonomy.csv" with NA as missing data identifier replaced with empty cells.
- Added columns for latitude and longitude of sampling station
- Set data types as string or numeric
- Exported the final file as "1003525_v1_rna_taxonomy.csv"

CURATION ACTIONS PERFORMED ON METADATA
- added summary for broader study showing connections between related datasets

ISSUES POTENTIALLY IMPACTING REUSE

N/A

[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

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience, 8(9). https://doi.org/10.1093/gigascience/giz100
Software

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution, 38(12), 5825–5829. https://doi.org/10.1093/molbev/msab293
Software

Cook, C. C. Z., Ewton, E. M., Marchetti, A., Menden-Deuer, S., Millette, N. C., Slomka, S., Speciale, E. V., Wilken, S., & Cohen, N. R. (2025). Evaluating acidotropic dyes for detecting mixotrophy in protists: Insights from cultures and field communities. https://doi.org/10.1101/2025.09.29.679303
Methods

Federhen, S. (2011). The NCBI Taxonomy database. Nucleic Acids Research, 40(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178
Related Research

Groussman, R. D., Blaskowski, S., Coesel, S. N., & Armbrust, E. V. (2023). MarFERReT, an open-source, version-controlled reference library of marine microbial eukaryote functional genes. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02842-4
Software

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... & Christen, R. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic acids research, 41(D1), D597-D604. https://doi.org/10.1093/nar/gks1160
General

Haas, B. J. (2023). TransDecoder (Version 5.7.1) [Computer software]. GitHub. https://github.com/TransDecoder/TransDecoder/releases/tag/TransDecoder-v5.7.1
Software

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. doi:10.1093/nar/28.1.27
Methods

Krinos, A. I., Hu, S. K., Cohen, N. R., & Alexander, H. (2020). EUKulele: Taxonomic annotation of the unsung eukaryotic microbes (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2011.00089
Software

Krueger, F. (2015). Trim Galore!: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. Babraham bioinformatics - trim galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Software

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158
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). https://doi.org/10.1186/s13059-014-0550-8
Software

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), 10. doi:10.14806/ej.17.1.200
Software

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E.,

Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. Nucleic Acids Research, 49(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913
Methods

Mitra, A., Caron, D. A., Faure, E., Flynn, K. J., Leles, S. G., Hansen, P. J., McManus, G. B., Not, F., do Rosario Gomes, H., Santoferrara, L. F., Stoecker, D. K., & Tillmann, U. (2023). The Mixoplankton Database (MDB): Diversity of photo-phago-trophic plankton in form, function, and distribution across the global ocean. Journal of Eukaryotic Microbiology, 70(4). Portico. https://doi.org/10.1111/jeu.12972
Methods
,
Methods

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods, 14(4), 417–419. doi:10.1038/nmeth.4197
Software

Pierce, E., Torano, O., Lin, Y., Schnetzer, A., & Marchetti, A. (2023). Comparison of advanced methodologies for diatom identification within dynamic coastal communities. Limnology and Oceanography: Methods, 21(11), 687–702. Portico. https://doi.org/10.1002/lom3.10575
Methods

Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research, 4, 1521. doi:10.12688/f1000research.7563.2
Methods

Speciale, E. V. (2026). Speciale-Metatranscriptomics [Computer software]. GitHub. https://github.com/emilyspeciale/Speciale-Metatranscriptomics
Methods
,
Software

Speciale, Emily. (2025). THE MOLECULAR PHYSIOLOGY OF MIXOTROPHIC PHYTOPLANKTON UNDER IRON-LIMITED UPWELLING CONDITIONS. The University of North Carolina at Chapel Hill University Libraries. https://doi.org/10.17615/C0G7-RP52 https://doi.org/10.17615/c0g7-rp52
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Speciale, E., Marchetti, A., Lim, P., Jeong, Y., Cook, C., Cohen, N., McClure, W., Schnetzer, A. (2026) **KEGG gene differential expression data for mixotrophs and diatoms.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-13 http://lod.bco-dmo.org/id/dataset/1004689 [view at BCO-DMO]

Speciale, E., Marchetti, A., Lim, P., Jeong, Y., Cook, C., Cohen, N., McClure, W., Schnetzer, A. (2026) **KEGG pathway differential expression data for mixotrophs and diatoms.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-14 http://lod.bco-dmo.org/id/dataset/1004704 [view at BCO-DMO]

Speciale, E., Marchetti, A., Lim, P., Jeong, Y., Cook, C., Cohen, N., McClure, W., Schnetzer, A. (2026) **Pfam differential expression data for diatoms from an iron incubation experiment during the PUPCYCLE II R/V Sally Ride cruise within the California Current System in May and June of 2023.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-13 http://lod.bco-dmo.org/id/dataset/1004674 [view at BCO-DMO]

Speciale, E., Marchetti, A., Lim, P., Jeong, Y., Cook, C., Cohen, N., McClure, W., Schnetzer, A. (2026) **Pfam differential expression data for mixotrophs from an iron incubation experiment during the PUPCYCLE II R/V Sally Ride cruise within the California Current System in May and June of 2023.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-12 http://lod.bco-dmo.org/id/dataset/1004659 [view at BCO-DMO]

Speciale, E., Marchetti, A., Lim, P., Jeong, Y., Cook, C., Cohen, N., McClure, W., Schnetzer, A. (2026) **Physiological measurements from iron incubation experiment using upwelled waters in the**

California Current System during the PUPCYCLE II cruise in May and June 2023. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-11 http://lod.bco-dmo.org/id/dataset/994607 [view at BCO-DMO]

[table of contents | back to top]

Parameters

Parameter	Description	Units
Timepoint	Timepoint identification (T0, T1, T2, or T3)	unitless
Treatment	Treatment identification (Ctrl, +Fe, or DFB)	unitless
Protist_Group	Classified general protist groups based on MarFERReT. Resulted in nine protist groups, any non-protist reads were excluded.	unitless
Mixotroph_Genera	Classified constitutive mixotroph genera based on Mixoplankton Database. Only includes 8 named genera, the rest are grouped into Other_mixotroph, or if not mapping to a mixotroph, Nonmixotroph.	unitless
Diatom_Genera	Classified diatom genera based on MarFERReT. Only includes 8 named genera, the rest are grouped into Other_diatom, or if not mapping to a diatom, NonDiatom.	unitless
Mean_Num_Reads	Mean number of normalized reads mapping to the identified protist group, mixotroph genera, or diatom genera for triplicates within each treatment/timepoint.	number of normalized reads
SD_Num_Reads	Standard deviation of number of normalized reads mapping to the identified protist group, mixotroph genera, or diatom genera for triplicates within each treatment/timepoint.	number of normalized reads
Mean_Percent_Reads	Mean proportion of normalized reads mapping to the identified protist group, mixotroph genera, or diatom genera for triplicates within each treatment/timepoint.	percentage
SD_Percent_Reads	Standard deviation of proportion of normalized reads mapping to the identified protist group, mixotroph genera, or diatom genera for triplicates within each treatment/timepoint.	percentage
Latitude	Latitude of sampling station	decimal degrees
Longitude	Longitude of sampling station	decimal degrees

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina HiSeq 4000
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Sequencing was performed on an Illumina HiSeq 4000 with a 2x150 bp configuration.
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

Dataset-specific Instrument Name	Sea Bird SBE 9 plus CTD
Generic Instrument Name	Sea-Bird SBE 9plus CTD
Generic Instrument Description	High precision and accuracy CTD comprising an SBE 9plus underwater unit (SBE 3plus temperature, SBE 4C conductivity, and Paroscientific Digiquartz pressure sensors, and an SBE 5T submersible pump). Can be used for either real-time data acquisition or for autonomous operations at a sampling speed of up to 24 Hz. The instrument package also includes a TC duct, to reduce salinity spiking caused by ship heave for improved resolution of water column features, and to ensure that temperature and conductivity measurements are made on the same parcel of water. Supplied with both an aluminium and titanium main housing, allowing for use up to 6800 and 10,500 metre depths respectively. Also capable of measuring from eight auxiliary sensors.

[table of contents | back to top]

Deployments

SR2311

Website	https://www.bco-dmo.org/deployment/994610
Platform	R/V Sally Ride
Start Date	2023-05-29
End Date	2023-06-10
Description	California Current System, off the southern coast of Oregon at 43°02'42.7"N, 124°33'07.2"W, depth 55m

[table of contents | back to top]

Project Information

CAREER: An integrated molecular and physiological approach to examining the dynamics of upwelled phytoplankton in current and changing oceans (Upwelled Phytoplankton Dynamics)

Coverage: California Upwelling Zone

NSF Award Abstract:
Upwelling zones are hotspots of photosynthesis that are very dynamic in space and time. Microsocopic algae,

known as phytoplankton, bloom when deep, nutrient-rich waters are upwelled into sunlit surface layers of the ocean, providing nourishment that supports productive food webs and draws down carbon dioxide (CO2) from the atmosphere to the deep ocean. Photosynthetic microbes in these regions must constantly adapt to changes in their chemical and physical environments. For example, subsurface populations respond to changes in light as they approach the surface. When upwelled waters move offshore, cells sink out of the illuminated zone, establishing seed populations that remain inactive until the next upwelling event. This process is called the upwelling conveyor belt cycle (UCBC). How phytoplankton respond to these changes in environmental conditions and how they may influence their nutrient requirements remains unknown. With future ocean changes predicted to alter seawater chemistry, including ocean acidification and decreased iron availability, some phytoplankton groups may be more vulnerable than others. Accompanying educational activities provide learning experiences to enhance understanding and awareness of marine microbes. The development of a research hub at UNC aims to provide infrastructure and support for scientists and students conducting research on environmental genomics. A laboratory component for an upper-level undergraduate course focused on marine phytoplankton is being developed. Educational outreach activities to broader communities include creation of a lesson plan on phytoplankton in upwelling zones and a virtual research cruise experience for middle-school students, as well as a hands-on lab activity for a local museum focused on marine phytoplankton and the important roles they play in shaping our planet.

The project examines how phytoplankton respond at the molecular and physiological level to the different UCBC stages, which seed populations (i.e., surface versus subsurface) contribute most to phytoplankton blooms during upwelling events of varying intensity, how phytoplankton elemental compositions are altered throughout UCBC stages, and how future predicted ocean conditions will affect the phytoplankton responses to UCBC conditions. This project contains both laboratory and fieldwork. In the laboratory, phytoplankton isolates recently obtained from upwelling regions are exposed to simulated UCBC conditions to examine changes in gene expression, growth and photosynthetic characteristics and elemental composition. Cultures are subjected to both current and future ocean conditions, including reduced iron availability and higher CO2. In the field, research cruises within upwelling regions study the dynamics of natural phytoplankton communities (both surface and subsurface) experiencing upwelling and relaxation and within simulated upwelling incubation experiments. Knowledge of how phytoplankton are affected by UCBC conditions at an integrated molecular, physiological and elemental level under both current and future scenarios is imperative for the proper conservation and management of these critically important ecosystems.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1751805

[table of contents | back to top]