

# Single-cell transcriptomic data from ciliates isolated in New England waters between 2019 and 2023

**Website:** <https://www.bco-dmo.org/dataset/988253>

**Data Type:** Other Field Results

**Version:** 1

**Version Date:** 2025-11-04

## Project

» [Collaborative Research: Combining single-cell and community omics to test hypotheses about diversity and function of planktonic ciliates](#) (Ciliate Omics)

Contributors	Affiliation	Role
<a href="#">Katz, Laura A.</a>	Smith College	Principal Investigator
<a href="#">York, Amber D.</a>	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

## Abstract

This dataset contains sampling and genetic accession information for single-cell transcriptomic data from bottle samples collected with CTD profiles (BCO-DMO dataset doi:10.26008/1912/bco-dmo.879380.1) and also from shore-based tide pool collections. Bottle samples were collected with R/V Connecticut in the Long Island Sound on 14-15 June 2022. Individual ciliates were picked for sequencing. Single-cell transcriptomes were based on poly-A selected gene sequences from individual ciliates that had been washed after isolation. Raw sequences are available in the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) under BioProjects PRJNA1026950 and PRJNA1223830.

## Table of Contents

- [Coverage](#)
- [Dataset Description](#)
  - [Methods & Sampling](#)
  - [Data Processing Description](#)
  - [BCO-DMO Processing Description](#)
- [Related Publications](#)
- [Related Datasets](#)
- [Parameters](#)
- [Instruments](#)
- [Project Information](#)
- [Funding](#)

## Coverage

**Location:** Northern US coast (from Maine to Connecticut) and a dock located in Groton, Connecticut, USA

**Spatial Extent:** N:44.87 E:-68.311088 S:41.316018 W:-72.062171

**Temporal Extent:** 2019-05-22 - 2023-05-22

## Dataset Description

See the "Related Datasets" section for methods and data from additional datasets from this study (CTD profile data and marine metabarcoding data).

## Methods & Sampling

Data were collected near UConn Avery Point and on a research cruise. Data were deposited by George McManus.

Sample type:

Transcriptomes: single-cell amplification of polyA transcripts as described in Shazib et al. (2025, doi: 10.1016/j.ympev.2024.108239).

Single-cell transcriptomes were based on poly-A selected gene sequences from individual ciliates that had been washed after isolation.

Raw sequences are available in the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) under two BioProjects:

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1026950>

\* **BioProject Title:** "Phylogenomic pipeline for uncultivable microbial eukaryotes using single cell RNA sequencing data - A case study with planktonic ciliates (Protista, Ciliophora, Oligotrichea)."

\* **BioProject Description:** "We collected marine planktonic ciliates from a dock located in Groton, Connecticut, USA. The collection was made on multiple dates from July 2020 to January 2022. After isolating the single cell, we performed single-cell RNA sequencing. We created a phylogenomic pipeline relying on PhyloTol to construct single gene trees from the transcriptomic data. Our pipeline produced well-curated data, which we used to estimate species trees. Overall this project provides a guideline to infer species tree of uncultivable microbial eukaryotes using single cell RNA sequencing data."

<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1223830>

\* **BioProject Title:** Whole transcriptome and genome data from several marine ciliates

\* **BioProject Description:** Single-cell whole transcriptome amplification and whole genome amplification; and subsequent Illumina NovaSeq 6000 reads from three marine ciliates species (*Helicostemella* sp., *Stenosemella* sp., and *Tintinnopsis cylindrica*), sampled in the northern US coast (from Maine to Connecticut); including one conjugating *Tintinnopsis cylindrica*.

## Data Processing Description

Scripts developed under this study to analyze the NCBI data can be found at the GitHub (see Wiki <https://github.com/Katzlab/EukPhylo/wiki>), with versions of scripts developed through this project from 2020-2024. A full version of EukPhylo version 1 is published on Zenodo (doi: 10.5281/ZENODO.15866075). The parent DOI for all versions of the github repository, including past releases of PhyloTol is doi:10.5281/ZENODO.13323347.

## BCO-DMO Processing Description

\* Sheet 1 of submitted file "Katz\_transcriptomes.xlsx" was imported into the BCO-DMO data system for this dataset. Values "NA" imported as missing data values. Table will appear as Data File: 988253\_v1\_ciliate\_transcriptomes.csv (along with other download format options).

Missing Data Identifiers:

\* In the BCO-DMO data system missing data identifiers are displayed according to the format of data you access. For example, in csv files it will be blank (null) values. In Matlab .mat files it will be NaN values. When viewing data online at BCO-DMO, the missing value will be shown as blank (null) values.

\* Column names adjusted to conform to BCO-DMO naming conventions designed to support broad re-use by a variety of research tools and scripting languages. [Only numbers, letters, and underscores. Can not start with a number]

\* Additional data at NCBI (PRJNA1223830) was made public after Katz\_transcriptomes.xlsx was submitted to BCO-DMO so the new accession identifiers for the SRA, BioProject, and BioSample holdings were merged into this dataset using the SRA\_Run as the join key to data extracted from the NCBI Run Selector (run 2025-11-04).

\* NCBI Run Selector data were also extracted from the other BioProject PRJNA1026950 since the submitter indicated the Organism column had been updated there. The "Organism" column from the NCBI metadata were added to this dataset as column Organism\_NCBI.

\* The "Organism" column from the provided file Katz\_transcriptomes.xlsx was also kept but renamed "Organism\_label" since it included other information (example "Tintinnopsis sp\_LIS-Tintinnina-11").

\* SRA\_sample for BioProject PRJNA1223830 was not included in the SRA run selector metadata so this column was not included here. The BioSample is included and sample information can be found from that.

\* The names in the "Organism\_NCBI" column were matched to ncbi\_txid identifiers using the Global Names Verifier. NCBI advises consulting authoritative sources for taxonomic name information. To this end, the World Register of Marine Species taxon match tool was used to match the names on 2025-11-04 and the Life Science Identifiers (LSIDs) were added along with the accepted spelling of the names.

\*\* Note: "Stenosemella steini" (ncbi\_txid:1594485) was a phonetic (not exact) match to the currently accepted name "Stenosemella steinii" LSID (urn:lsid:marinespecies.org:taxname:178827).

\* Column lat\_lon separated into Lat,Lon columns in decimal degrees. Some lacked a negative for lon to be W so that was corrected.

\* Instrument column in the dataset updated. The model number was sequentially increasing each row in the original excel file. Corrected so all models for Illumina HiSeq say "Illumina HiSeq 2500"

[ [table of contents](#) | [back to top](#) ]

---

## Related Publications

Auden Cote-L'Heureux, Godwin N. Ani, Katzlab, Adri K. Grow, MCLeleu, Elinor Sterner, GiuliaRibeiro, & rebeccagawron. (2025). *Katzlab/EukPhylo: EukPhylo version 1.0* (Version v1.0) [Computer software]. Zenodo. <https://doi.org/10.5281/ZENODO.15866075> <https://doi.org/10.5281/zenodo.15866075>  
*Software*

Grow, A., Sleith, R., Sehein, T., Labare, M., & Katz, L. (2023). Exploring the diversity of microeukaryotic communities in New England tide pools. *Aquatic Microbial Ecology*, 89, 143–155.  
<https://doi.org/10.3354/ame02003>  
*Results*

Katzlab. (2025). EukPhylo version 1.0 wiki. GitHub. <https://github.com/Katzlab/EukPhylo/wiki>  
*Methods*

Shazib, S. U. A., Ahsan, R., Leleu, M., McManus, G. B., Katz, L. A., & Santoferrara, L. F. (2025). Phylogenomic workflow for uncultivable microbial eukaryotes using single-cell RNA sequencing – A case study with planktonic ciliates (Ciliophora, Oligotrichea). *Molecular Phylogenetics and Evolution*, 204, 108239.  
<https://doi.org/10.1016/j.ympev.2024.108239>  
*Results*

[ [table of contents](#) | [back to top](#) ]

---

## Related Datasets

### IsRelatedTo

Katz, L. A. (2025) **Metabarcoding data from samples collected at shore-based tide pools and ocean samples in New England waters in 2019**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2025-11-04 <http://lod.bco-dmo.org/id/dataset/988266> [[view at BCO-DMO](#)]

*Relationship Description: The Single-cell transcriptomics dataset (988253) and metabarcoding dataset (988266) are the omics data associated with the collections from the CTD dataset (879380). Both omics datasets used the same bottle samples collected with the CTD profiles.*

McManus, G., Santoferrara, L., Katz, L. A. (2022) **CTD profiles collected with RV Connecticut on the Continental shelf and slope south of Montauk, NY on 14-15 June 2022**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2022-09-16  
[doi:10.26008/1912/bco-dmo.879380.1](https://doi.org/10.26008/1912/bco-dmo.879380.1) [[view at BCO-DMO](#)]

*Relationship Description: The Single-cell transcriptomics dataset (988253) and metabarcoding dataset (988266) are the omics data associated with the collections from the CTD dataset (879380). Both omics datasets used the same bottle samples collected with the CTD profiles.*

Smith College (2023). Phylogenomic pipeline for uncultivable microbial eukaryotes using single cell RNA

sequencing data - A case study with planktonic ciliates (Protista, Ciliophora, Oligotrichea). 2023/10. In: NCBI:BioProject: PRJNA1026950 [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA1026950>.

Smith College (2025). Whole transcriptome and genome data from several marine ciliates. 2025/02. In: NCBI:BioProject: PRJNA1223830. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA1223830>

[ [table of contents](#) | [back to top](#) ]

## Parameters

Parameter	Description	Units
BioProject	NCBI BioProject accession	unitless
SRA_Experiment	NCBI Sequence Read Archive (SRA) Study accession	unitless
SRA_study	NCBI Sequence Read Archive (SRA) Study accession	unitless
SRA_Run	NCBI Sequence Read Archive (SRA) Run accession	unitless
BioSample	NCBI BioSample accession	unitless
sample_name	Sample name. Uses in house tracking system sample identifier referred to as LKH Number""	unitless
library_ID	library identifier. Uses in house tracking system sample identifier referred to as LKH Number""	unitless
title	title of accession	unitless
Isolate	Isolate (Not applicable)	unitless
Isolation_source	Isolation source (Not applicable)	unitless
Tissue	Tissue (not applicable)	unitless
Collection_Date	Collection date in format mm/dd/YY	unitless
geo_loc_name	geolocation name	unitless
Lat	Latitude	decimal degrees

Lon	Longitude	decimal degrees
description	Not used	unitless
Organism_label	Organism taxonomic name (full or abbreviated). May include additional descriptors in the label.	unitless
Organism_NCBI	Organism scientific name. This is the organism included in the NCBI metadata. See supplemental species list for more information and identifiers for this name.	unitless
library_strategy	Library strategy	unitless
library_source	Library source	unitless
library_selection	Library selection	unitless
library_layout	Library layout	unitless
platform	platform (ILLUMINA)	unitless
instrument_model	Instrument model	unitless
design_description	Design description	unitless

[ [table of contents](#) | [back to top](#) ]

---

## Instruments

<b>Dataset-specific Instrument Name</b>	Illumina HiSeq 2500
<b>Generic Instrument Name</b>	Automated DNA Sequencer
<b>Generic Instrument Description</b>	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

<b>Dataset-specific Instrument Name</b>	Illumina NovaSeq 6000
<b>Generic Instrument Name</b>	Automated DNA Sequencer
<b>Generic Instrument Description</b>	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

[ [table of contents](#) | [back to top](#) ]

---

## Project Information

**Collaborative Research: Combining single-cell and community 'omics' to test hypotheses about diversity and function of planktonic ciliates (Ciliate Omics)**

**Website:** <http://microzooplankton.uconn.edu>

**Coverage:** New England continental shelf

### *NSF Award Abstract:*

Planktonic ciliates are key members of marine food webs where they serve diverse roles, including as food chain links between smaller microbes and larger plankton. Due to their small size and difficulties in identifying and cultivating them, we know less about ciliate diversity and distributions in the ocean than we do about larger organisms such as fish and invertebrates. Previous work from this team measured ciliate diversity in coastal waters and found that distinct genetic variants were separated in time and space in a way that could be related to factors such as ocean temperature, salinity, and depth gradients. Many questions remained unanswered, and it is important to understand the environmental factors that control the diversity and distribution of plankton such as ciliates to predict how these organisms may respond to a changing environment in the coming decades. This project focuses on: 1) how ciliate species are delineated using single-cell genomics and transcriptomics; 2) DNA-based studies of all ciliates and other planktonic members of the SAR clade (Stramenopila, Alveolata, Rhizaria), which will provide ecological context; 3) in situ gene expression by single-cell and meta-transcriptomics; and 4) laboratory studies of gene expression in cultivated ciliate species. This project involves training of postdoctoral scholars, graduate students, and undergraduates. The researchers are committed to creating diverse and inclusive research labs; recruitment of participants will be done through partnership with appropriate groups on our campuses. The project integrates with summer Research Experiences for Undergraduates (REU) activities at both Smith College and UCONN (including the UCONN/Mystic Aquarium joint REU), which are especially focused on underrepresented students. This project also enhances efforts to broaden understanding of biodiversity in partnership with the UCONN Noyce Scholars Program, which facilitates career-changing STEM professionals to become teachers in underserved secondary schools.

This project will assess distributions of reproductively-isolated species, determined using a new method to characterize regions of the ciliate germline genome. Furthermore, it will use phylogenomic methods to identify clade-specific transcripts (e.g. those of spirotrich ciliates) within metatranscriptomes from the shelf environment and to expand knowledge of ciliate function with single-cell transcriptomics of field-collected cells. These approaches will be a substantial improvement over the culture-based methods that are potentially biased towards "weedy" species in the ocean. The combination of definitive species identification with assessment of function via single-cell and meta-transcriptomics promises to provide significant advances in marine plankton ecology. The research focuses on two broad questions: 1) does the observed high diversity in phylogenetically-informative genes reflect reproductive isolation and functional differentiation in planktonic ciliates? and 2) do different co-occurring species of planktonic ciliates show substantial functional differences that correspond to different niches in the ocean? The project assesses species boundaries (i.e. reproductive isolation) through analyses of patterns in the germline micronuclei of planktonic ciliate morphospecies; characterizes transitions of closely-related ciliates across ecological gradients in the ocean; and examines functional differences within and between species, and in communities, through analyses of transcriptomics.

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

<b>Funding Source</b>	<b>Award</b>
<a href="#">NSF Division of Ocean Sciences (NSF OCE)</a>	<a href="#">OCE-1924570</a>
NSF Division of Environmental Biology (NSF DEB)	<a href="#">DEB-1651908</a>
NSF Division of Environmental Biology (NSF DEB)	<a href="#">DEB-1541511</a>

[ [table of contents](#) | [back to top](#) ]