Scaffold-derived metaproteomic exclusive and total spectral counts associated with proteins from samples taken during R/V Atlantic Explorer cruise AE1913 from the Sargasso Sea to Northeast US shelf waters in June of 2019

Website: https://www.bco-dmo.org/dataset/934706
Data Type: Cruise Results
Version: 1
Version Date: 2024-08-01

Project
» Collaborative Research: Direct Characterization of Adaptive Nutrient Stress Responses in the Sargasso Sea using Protein Biomarkers and a Biogeochemical AUV (Nutrient Stress Responses and AUV Clio)

Contributors	Affiliation	Role
Saito, Mak A.	Woods Hole Oceanographic Institution (WHOI)	Principal Investigator
Cohen, Natalie	Woods Hole Oceanographic Institution (WHOI)	Scientist
York, Amber D.	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
These are the Scaffold-derived metaproteomic exclusive and total spectral counts associated with proteins. Samples were taken during R/V Atlantic Explorer cruise AE1913 in Subtropical North Atlantic, beginning at the Bermuda Atlantic Time-series Station (BATS) of the Sargasso Sea and ending in coastal Northeast US shelf waters in June of 2019.

Table of Contents

Coverage

Location: Subtropical North Atlantic, beginning at the Bermuda Atlantic Time-series Station (BATS) of the Sargasso Sea (31.666888 N64.166293 W) and ending in coastal Northeast US shelf waters (39.31658 N 71.123208 W)
Spatial Extent: **N**:38.527393 **E**:-64.165587 **S**:31.586387 **W**:-70.840703
Temporal Extent: 2019-06-16 - 2019-06-27

Dataset Description

Related data table and dataset descriptions:

The primary data table for this dataset is provided under the "Data Files" section and contains total protein spectral counts while the table under "Supplemental Files" provides the exclusive protein spectral counts.

Total spectral counts refer to the total number of spectra with peptide to spectrum matches (PSMs) that matches to each entry within the FASTA sequence database. This approach allows each peptide to map to multiple closely related sequences. In contrast, with exclusive spectral counts each peptide is only allowed to map to one sequence within the FASTA database, and when a peptide is found in multiple database sequences the one with the most peptides mapping (parsimony) to it is selected. There are pros and cons to each approach, where total spectral counts will double count peptides when two similar proteins are compared, and exclusive spectral counts will underrepresent less abundant proteins with shared peptides, favoring the most homolog with the most shared peptides. Considering protein groups with shared peptides or focusing on peptide-level analyses are alternative approaches that could be constructed from these results.

See "Related Datasets" section for:
* "AE1913 Peptide Spectral Counts" which includes the individual peptides associated with these proteins (includes total spectral counts for each peptide).
* "AE1913 Protein Identification FASTA"

CTD and other data from the same cruise are listed on deployment page AE1913: https://www.bco-dmo.org/deployment/916412

These data will become part of the Ocean Protein Portal (https://proteinportal.whoi.edu/; Saito et al., 2020).

The assembly, annotations, metatranscriptomic assembly products, the same exclusive protein spectral counts, and other useful information associated with this multi-omic analysis was published as a package at Zenodo (doi: 10.5281/zenodo.8287779).

Methods & Sampling

Methods are reported in Cohen et al. 2023 (biorxiv preprint doi: 10.1101/2023.11.20.567900) and are summarized below.
* This section describes how this and related datasets were generated (see "Related Datasets" section).

One half of the 142 mm filters (0.2-51 µm) collected by Clio were processed for metaproteomics. Proteins were extracted in an 1% SDS-based detergent in 50 mM HEPES at pH 8.5, reduced with dithiothreitol, alkylated with iodoacetamide, and purified using a polyacrylamide electrophoresis tube gel method. Protein quantification was performed using a BSA assay. Trypsin was added to the protein-bead mixture in a 1:20 trypsin:protein ratio. Peptides were purified using C18 tips and diluted to a concentration of 0.1 µg µL−1.

Approximately 2-5 µg of purified peptides were injected onto a Dionex UltiMate 3000 RSLCnano LC system with an additional RSLCnano pump, run in online 2D active modulation mode interfaced with a Thermo Fusion mass spectrometer. The mass spectrometer acquired MS1 scans from 380 to 1,580 m/z at 240K resolution in the Orbitrap. MS2 were collected in data dependent mode in the ion trap with a cycle time of 2 seconds between scans and acquisition of charge states 2 to 10. MS2 scans had 1.6 m/z isolation window, 50 ms maximum injection time and 5 s dynamic exclusion time.

Note: This dataset contains two different missing data identifiers "NA" and "-". If there were partial matches to the functional annotation database, the missing ones were denoted with "-". If there were no matches at all, when the data frames were merged, the empty columns were denoted with "NA".

example lines in opp_TOTAL_spectralcounts.csv

"6","megahit_HN001_k141_101642.p1","-","-","-","SBP_bac_1,SBP_bac_8"...
vs
"4","megahit_HN001_k141_100671.p1",NA,NA,NA,NA,NA,NA,"X1_30_0.2"...

Data Processing Description

The metatranscriptomic ORFs were used as the protein database, and peptide-spectrum matches were performed using Sequest algorithm within IseNode Proteome Discoverer 2.2.0.388 with a parent ion tolerance of 10 ppm and fragment tolerance of 0.6 Da, and 0 max missed cleavage. Identification criteria consisted of a peptide threshold of 98% (false discovery rate [FDR] = 0.1%) and protein threshold of 99% (1 peptide

minimum, FDR = 1.5%) in Scaffold 5.1.2 (Proteome Software) resulting in 77,438 proteins and 3,155,061 exclusive spectral counts.

BCO-DMO Processing Description

BCO-DMO Data Manager Processing Notes:
* Data from source file opp_TOTAL_spectralcounts.csv was imported into the BCO-DMO data system as the primary table for this dataset and appears under the Data Files section. First column was an un-named row of sequential numbers so was given the name "row_id".
* Data from source file opp_spectralcounts.csv was attached as as supplemental file (contains exclusive counts). First column was an un-named row of sequential numbers so was given the name "row_id".

[table of contents | back to top]

Data Files

File
Protein total spectral counts filename: 934706_v1_ae1913-protein-total-spectral-counts.csv (Comma Separated Values (.csv), 1.06 GB) MD5:d6ab0593af68ea7207e254f34504bb1b Primary data file for dataset ID 934706, version 1. See supplemental files for exclusive spectral counts.

[table of contents | back to top]

Supplemental Files

File
Protein exclusive spectral counts filename: 934706_v1_ae1913-protein-exclusive-spectral-counts.csv (Comma Separated Values (.csv), 1.06 GB) MD5:743813f7eaee641bcbc94c995a025794 Exclusive spectral counts (see "Data Files" for the total spectral counts). This table is the same structure as the total spectral count table. See "Parameters" section for column information.

[table of contents | back to top]

Related Publications

Saito, M. A., Saunders, J. K., Chagnon, M., Gaylord, D. A., Shepherd, A., Held, N. A., Dupont, C., Symmonds, N., York, A., Charron, M., & Kinkade, D. B. (2020). Development of an Ocean Protein Portal for Interactive Discovery and Education. Journal of Proteome Research, 20(1), 326–336. https://doi.org/10.1021/acs.jproteome.0c00382
Related Research

[table of contents | back to top]

Related Datasets

IsRelatedTo

Cohen, N., Krinos, A., Alexander, H., & Saito, M. (2022). Protistan metabolism across the western North Atlantic Ocean revealed through autonomous underwater profiling (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.8287779 https://doi.org/10.5281/zenodo.8287779

Saito, M. A., Cohen, N. (2024) **Peptides associated with scaffold-derived metaproteomic proteins**

from samples taken during R/V Atlantic Explorer cruise AE1913 from the Sargasso Sea to Northeast US shelf waters in June of 2019. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2024-08-01 doi:10.26008/1912/bco-dmo.934718.1 [view at BCO-DMO]
Relationship Description: These datasets are from the same collection and study and will be included in the Ocean Protein Portal (https://proteinportal.whoi.edu).

Saito, M. A., Cohen, N. (2024) **Protein identification FASTA file (scaffold-derived metaproteomic proteins) from samples taken during R/V Atlantic Explorer cruise AE1913 from the Sargasso Sea to Northeast US shelf waters in June of 2019.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2024-08-01 doi:10.26008/1912/bco-dmo.934727.1 [view at BCO-DMO]
Relationship Description: These datasets are from the same collection and study and will be included in the Ocean Protein Portal (https://proteinportal.whoi.edu).

[table of contents | back to top]

Parameters

Parameter	Description	Units
row_id	sequential row identifier	unitless
protein_id	Protein identifier. Uniquely identifies a protein within the dataset and FASTA file	unitless
kegg_id	Kegg identifier	unitless
enzyme_comm_id	Enzyme Commission identifer	unitless
protein_name	Protein descriptive name	unitless
pfams_id	Protein family ID number	unitless
supergroup	Supergroup	unitless
classification	Classification	unitless
sample_id	Identifies the sample associated with this annotation	unitless
spectral_count	Spectral count	unitless
cruise_id	Cruise identifier	unitless
station_id	Station identifier where sample was taken	unitless
depth_m	The depth in meters at which the sample as taken	meters
minimum_filter_size_microns	Minimum size of the collection filter	microns (um)
maximum_filter_size_microns	Maximum size of the collection filter	microns (um)
date_y_m_d	The date of sample collection	unitless
latitude_dd	The latitude at the station in decimal degrees (-90 to 90)	decimal degrees
longitude_dd	The longitude at the station in decimal degrees (-180 to 180)	decimal degrees

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	
Generic Instrument Name	AUV Clio
Generic Instrument Description	Clio is an autonomous underwater vehicle (AUV) created to accomplish the dual goals of global ocean mapping and biochemistry sampling. The ability to sample dissolved and particulate seawater biochemistry across ocean basins while capturing fine-scale biogeochemical processes sets it apart from other AUVs. Clio is designed to efficiently and precisely move vertically through the ocean, drift laterally to observe water masses, and integrate with research vessel operations to map large horizontal scales up to a depth of 6,000 meters. More information is available at https://www2.whoi.edu/site/deepsubmergencelab/clio/

Dataset-specific Instrument Name	Thermo Fusion mass spectrometer
Generic Instrument Name	Mass Spectrometer
Generic Instrument Description	General term for instruments used to measure the mass-to-charge ratio of ions; generally used to find the composition of a sample by generating a mass spectrum representing the masses of sample components.

Dataset-specific Instrument Name	Dionex UltiMate 3000 RSLCnano LC system
Generic Instrument Name	Ultra high-performance liquid chromatography
Generic Instrument Description	Ultra high-performance liquid chromatography: Column chromatography where the mobile phase is a liquid, the stationary phase consists of very small (< 2 microm) particles and the inlet pressure is relatively high.

[table of contents | back to top]

Deployments

AE1913

Website	https://www.bco-dmo.org/deployment/916412
Platform	R/V Atlantic Explorer
Start Date	2019-06-16
End Date	2019-06-28
Description	coordinated deployments: McLane pumps, AUV Clio, CTD, trace metal rosette

[table of contents | back to top]

Project Information

Collaborative Research: Direct Characterization of Adaptive Nutrient Stress Responses in the Sargasso Sea using Protein Biomarkers and a Biogeochemical AUV (Nutrient Stress Responses and AUV Clio)

Coverage: Bermuda Atlantic Time Series

NSF Award Abstract:
Microscopic communities in the ocean can be surprisingly diverse. This diversity makes it difficult to study the individual organisms and reactions that control specific reactions controlling nutrient cycles. Past studies confirm that iron and nitrogen are vital elements for biological growth. There is increasing evidence, however, that other chemicals such as silica, zinc, cobalt, and vitamin B12 may be just as important. This project will provide an unprecedented view of community distributions using new molecular methods to isolate and link active proteins to specific chemical cycles during the very first research deployment of a brand-new autonomous underwater vehicle (AUV). The AUV will collect samples in programed patterns by pumping water directly into its filtering mechanism and then return the samples to the ship for analysis. The Bermuda Atlantic Time-series Study (BATS) station, which provides abundant supporting data, is the site for this innovative investigation into the microbial ecology and chemistry of the open oceans. Additionally, data will be widely distributed to other scientists through the Ocean Protein Portal website being developed by the Woods Hole Oceanographic Institute (WHOI) and the Biological and Chemical Oceanography Data Management Office. Data will also contribute a new teaching module in the Marine Bioinorganic Chemistry course at WHOI.

This first scientific deployment of the newly engineered and constructed biogeochemical AUV, Clio, will generate a novel dataset to examine marine microbial biogeochemical cycles in the Northwestern Atlantic oligotrophic ocean in unprecedented detail and at high vertical resolution. First the project proposes to understand if the microbial community reflects the varying chemical composition and cyanobacterial species through nutrient response adaptations. Additionally, the research will determine if iron stress in the low light Prochlorococcus ecotype found in the deep chlorophyll maximum is a persistent feature influenced by seasonal dust fluxes. The highly resolved vertical data from the in situ pumping capabilities of Clio are fundamental to a rigorous examination of these biogeochemical questions. This highly transformative dataset will greatly advance understanding of the nutrient and trace element cycling of this region and will be the first field validation of the potentially revolutionary capability these new approaches represent for the study of marine microbial biogeochemistry.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1658030

[table of contents | back to top]