Relative protein abundance from scaled and corrected exclusive peptide spectral counts from the ProteOMZ R/V Falkor expedition cruise FK160115 in the Pelagic central Pacific Ocean in 2016

Website: https://www.bco-dmo.org/dataset/868030
Data Type: Cruise Results
Version: 1
Version Date: 2022-01-13

Project
» The ProteOMZ Expedition: Investigating Life Without Oxygen in the Pacific Ocean (ProteOMZ (Proteomics in an Oxygen Minimum Zone))
» Marine Microbial Investigator Award: Investigator Mak Saito (MM Saito)

Program
» Marine Microbiology Initiative (MMI)

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

Abstract
Relative protein abundance from scaled and corrected exclusive peptide spectral counts from 20-1250 m in the water column (0.2-3 µm filter size fraction) from the ProteOMZ R/V Falkor expedition. There are a total of 107,579 unique peptide sequences from 56,543 protein groups (88,251 proteins). Exclusive spectral counts are provided per sample as are the full dataset scaled and normalized spectral counts. The protein distributions in this dataset highlight the microbial dynamics across biomes in the central Pacific Ocean. These data were submitted in Saunders et al. (2022).

Table of Contents

Coverage

Spatial Extent: **N**:10 **E**:-139.8 **S**:-10.6 **W**:-156
Temporal Extent: 2016-01-17 - 2016-02-05

Methods & Sampling

Methods & Sampling:

Samples were handled as described in (Saunders et al., submitted) and (McIlvin et al., 2021). There are a total of 107,579 unique peptide sequences from 56,543 protein groups (88,251 proteins). Proteins were extracted from biomass collected on a quarter section of a 142 mm 0.2 µm Supor filter (Pall Corporation) after pre-filtration through a 3.0 µm Supor filter. Proteins were extracted using a modified SP3 magnetic bead method (Hughes et al., 2014). Extracted proteins were quatified using the bicinchoninic acid method (Thermo Scientific Micro BCA protein assay kit) with an albumin protein reference standard. Extracted protein was purified and digested with trypsin. Purified and digested protein was then injected into an online nanoflow 2D active modulation liquid chromatography separation (McIlvin et al., 2021). Eluent flowed inline into the ion source of a Thermo Fusion quadrupole-Orbitrap mass spectrometer (Thermo Scientific). MS/MS spectra (and mass spec methods) are available at the PRIDE repository under accession PXD030684. Peptide to spectrum matching (PSM) on the MS/MS data was conducted with SEQUEST HT using Thermo Proteome Discoverer v 2.1 software against a database of predicted proteins from an assembled metagenome from the central Pacific Ocean. Additional protein inference and FDR calculations were conducted with Scaffold v 4.8.7. Identification was conducted with decoy false discovery rates (FDR) with a threshold of 95% minimum for peptides (FDR=0.1%) and a threshold of 99% (1 peptide minimum) for proteins (FDR=1.6%). Protein level inference for parsimony-based assignments of specific proteins was conducted using experiment-wide grouping with binary peptide-protein weights in Scaffold. All PSM and protein quantification in Scaffold was conducted using Exclusive Counts (not double-counted, PSM assignment to protein groups using parsimony).

Location: Pelagic central Pacific Ocean from 20 – 1250 m depth

Data Processing Description

Data Processing:
The raw mass spectra files were searched against SEQUEST within Proteome Discoverer v2.2 software. Raw mass spectra files are available in ProteomeXchange via the PRIDE database with accession # PXD030684 and 10.6019/PXD030684. Processed files were then loaded into Proteome Software and exclusive peptide reports were exported. The files were modified to scale and normalize the spectral counts at across the entire dataset. The peptide report was too large to work within Excel and was modified in Pandas/Python to produce a CSV file. Lowest Common Ancestor (LCA) analysis was conducted with METATRYP v 2 (Saunders et al 2020). Grouping of taxonomic and KO levels was conducted with pandas in python.

BCO-DMO data manager processing notes:
* Un-named first data column with row index removed (verified with submitter).

[table of contents | back to top]

Data Files

File
ProteOMZ Exclusive Peptide Level Spectral Counts filename: proteomz_peptide_spec_counts.csv (Comma Separated Values (.csv), 1.75 GB) MD5:81ec05e6f8abbd6e2ac31a3c26963053 ProteOMZ Exclusive Peptide Level Spectral Counts data table in csv format. See Parameters section for data table column names, descriptions, and units.

[table of contents | back to top]

Related Publications

Hughes, C. S., Foehr, S., Garfield, D. A., Furlong, E. E., Steinmetz, L. M., & Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. Molecular Systems Biology, 10(10), 757. doi:10.15252/msb.20145625
Methods

McIlvin, M. R., & Saito, M. A. (2021). Online Nanoflow Two-Dimension Comprehensive Active Modulation Reversed Phase–Reversed Phase Liquid Chromatography High-Resolution Mass Spectrometry for Metaproteomics of Environmental and Microbiome Samples. Journal of Proteome Research, 20(9), 4589–4597.

doi:10.1021/acs.jproteome.1c00588
Methods

Saunders, J. K., Gaylord, D. A., Held, N. A., Symmonds, N., Dupont, C. L., Shepherd, A., … Saito, M. A. (2020). METATRYP v 2.0: Metaproteomic Least Common Ancestor Analysis for Taxonomic Inference Using Specialized Sequence Assemblies—Standalone Software and Web Servers for Marine Microorganisms and Coronaviruses. Journal of Proteome Research, 19(11), 4718–4729. doi:10.1021/acs.jproteome.0c00385
Methods

Saunders, J. K., McIlvin, M. R., Dupont, C. L., Kaul, D., Moran, D. M., Horner, T., Laperriere, S. M., Webb, E. A., Bosak, T., Santoro, A. E., & Saito, M. A. (2022). Microbial functional diversity across biogeochemical provinces in the central Pacific Ocean. Proceedings of the National Academy of Sciences, 119(37). https://doi.org/10.1073/pnas.2200014119
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Saito, M. A. (2024) **Total spectral count of proteins from R/V Falkor cruise 160115 for the ProteOMZ expedition in the Central Pacific in 2016.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 3) Version Date 2022-06-06 doi:10.26008/1912/bco-dmo.737620.3 [view at BCO-DMO]
Relationship Description: These datasets are part of the Ocean Protein Portal "ProteOMZ" dataset (https://proteinportal.whoi.edu/; Saito et al., 2019).

Saito, M. A., Santoro, A. E. (2018) **Hydrographic data from the CTD mounted on the trace metal rosette (TMR) aboard R/V Falkor cruise (160115) during the ProteOMZ expedition in the Central Pacific in 2016.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2018-05-01 http://lod.bco-dmo.org/id/dataset/734608 [view at BCO-DMO]
Relationship Description: This dataset was collected asynchronously using another instrument at the same stations during the expedition.

Saito, M. A., Santoro, A. E. (2018) **Macronutrient analysis and selected hydrographic data from the R/V Falkor ProteOMZ expedition (FK160115) in the Central Pacific in 2016.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2018-11-19 http://lod.bco-dmo.org/id/dataset/730912 [view at BCO-DMO]
Relationship Description: This dataset was collected asynchronously using another instrument at the same stations during the expedition.

Saito, M. A., Santoro, A. E. (2018) **SeaBird SBE19 underway CTD information for the R/V Falkor 160115 cruise in the Central Pacific for the ProteOMZ expedition in 2016.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2018-06-15 http://lod.bco-dmo.org/id/dataset/730925 [view at BCO-DMO]
Relationship Description: This dataset was collected asynchronously using another instrument at the same stations during the expedition.

Saito, M. A., Santoro, A. E. (2024) **R/V Falkor 160115 McLane pump log from the ProteOMZ expedition in the Central Pacific during 2016 (ProteOMZ project).** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 3) Version Date 2024-10-21 doi:10.26008/1912/bco-dmo.708495.3 [view at BCO-DMO]
Relationship Description: Dataset "RV Falkor 160115 McLane Pump Log" is the log for the sample collection via McLane pumps used for the protein sampling for dataset "ProteOMZ Exclusive Peptide Level Spectral Counts."

[table of contents | back to top]

Parameters

Parameter	Description	Units

MSMS_sample_name	RAW file name; crossreferences to proteomexchange or OPP repository	unitless
Stn	Station number	unitless
Depth	Depth of sampling	meters
Longitude	Longitude	decimal degrees
Latitude	Latitude	decimal degrees
DepthGroup	K-means cluster of depths across transect	unitless
Region	Hierarchical cluster of stations across transect	unitless
Peptide_sequence	Unique Peptide sequence; this is the most unique identifier	unitless
Protein_name	Metagenome ID of best inferred protein from Scaffold protein inference; needed to map to protein IDs.	unitless
Exclusive_Sum_PSM	Exclusive Sum of +2 +3 +4 data = total unnormalized spectral counts; Quantitative Value	count
Scaling_Factor	Scaling factor used to normalize samples across experiment to control for identification bias of PSMs	unitless
Calculated_Total_Protein	Concentration of protein per L of seawater passed through filter	ug/L
Scaled_Corrected_Exclusive_Sum	Scaled and normalized exclusive sum of spectral counts	sc/L
blast_accession	Accession number in NCBI of best blast hit for query of ORF from protein name identified	unitless
blast_best_hit_taxon_id	best NCBI taxon ID for best blast hit in NCBI for query of ORF from protein name identified	unitless
KO	accession number of KEGG Orthology (KO) identifier for ORF from protein name	unitless

Group	NCBI taxonomic group for ORF from protein name	unitless
Domain	NCBI taxonomic domain for ORF from protein name	unitless
Phylum	NCBI taxonomic phylum for ORF from protein name	unitless
Class	NCBI taxonomic class for ORF from protein name	unitless
Order	NCBI taxonomic order for ORF from protein name	unitless
Family	NCBI taxonomic family for ORF from protein name	unitless
Genus	NCBI taxonomic genus for ORF from protein name	unitless
Species	NCBI taxonomic species for ORF from protein name	unitless
LCA_Group	Lowest Common Ancestor for groups of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Domain	Lowest Common Ancestor for domains of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Phylum	Lowest Common Ancestor for phyla of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Class	Lowest Common Ancestor for classes of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Order	Lowest Common Ancestor for orders of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Family	Lowest Common Ancestor for families of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_Genus	Lowest Common Ancestor for genuses of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless

LCA_Level	lowest taxonomic level of the Lowest Common Ancestor of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
LCA_taxon	lowest taxonomic identification of the Lowest Common Ancestor of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Group	All groups of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Domain	All domains of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Phylum	All phyla of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Class	All classes of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Order	All orders of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Family	All families of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Genus	All genuses of all NCBI taxons from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_KO	All KEGG Orthology (KO) identifiers from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_KO_Orthology	All KEGG Orthology (KO) orthologies from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_KO_Class	All KEGG Orthology (KO) classes from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_KO_Path	All KEGG Orthology (KO) paths from ORFs containing peptide sequence in the entire metagenome	unitless
Combo_Gene_Name	All KEGG Orthology (KO) gene names form ORFs containing peptide sequence in the entire metagenome	unitless

Combo_Gene_Description	All KEGG Orthology (KO) gene descriptions containing peptide sequence in the entire metagenome	unitless
Combo_EC	All Enzyme Commission (E.C.) numbers from ORFs from ORFs containing peptide sequence in the entire metagenome	unitless
Best_Peptide_identification_probability	Metric of peptide quality	unitless
Best_Sequest_XCorr_Only_deltaCn	Metric of peptide quality	unitless
Best_Sequest_XCorr_Only_XCorr	Metric of peptide quality	unitless
Number_of_identified_plus2H_spectra	Count of +2H spectra	count
Number_of_identified_plus3H_spectra	Count of +3H spectra	count
Number_of_identified_plus4H_spectra	Count of +4H spectra	count
Median_Retention_Time	Median retention time	minutes
Total_TIC	Total Ion Current	unitless
SOM_Label	Taxon and KO grouping label for SOM neural net	unitless
SOM_Label_Flag	Flag for SOM neural net that indicates whether more than one group listed condensed into SOM Label	unitless
All_Other_Proteins	All other ORFs in metagenome that contain the peptide sequences. Used to calculate LCA and Combos for taxonomy and KO annotations.	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	McLane in situ filtration pump
Generic Instrument Name	McLane Pump
Generic Instrument Description	McLane pumps sample large volumes of seawater at depth. They are attached to a wire and lowered to different depths in the ocean. As the water is pumped through the filter, particles suspended in the ocean are collected on the filters. The pumps are then retrieved and the contents of the filters are analyzed in a lab.

[table of contents | back to top]

Deployments

FK160115

Website	https://www.bco-dmo.org/deployment/708387
Platform	R/V Falkor
Report	https://service.rvdata.us/data/cruise/FK160115/doc/FK160115_OfficialCruiseReport_Saito_v3.pdf
Start Date	2016-01-16
End Date	2016-02-11
Description	Project: Using Proteomics to Understand Oxygen Minimum Zones (ProteOMZ) More information is available from the ship operator at https://schmidtocean.org/cruise/investigating-life-without-oxygen-in-the... Additional cruise information is available from the Rolling Deck to Repository (R2R): https://www.rvdata.us/search/cruise/FK160115

[table of contents | back to top]

Project Information

The ProteOMZ Expedition: Investigating Life Without Oxygen in the Pacific Ocean (ProteOMZ (Proteomics in an Oxygen Minimum Zone))

Website: https://schmidtocean.org/cruise/investigating-life-without-oxygen-in-the-tropical-pacific/#team

Coverage: Central Pacific Ocean (Hawaii to Tahiti)

From Schmidt Ocean Institute's ProteOMZ Project page:

Rising temperatures, ocean acidification, and overfishing have now gained widespread notoriety as human-caused phenomena that are changing our seas. In recent years, scientists have increasingly recognized that there is yet another ingredient in that deleterious mix: a process called deoxygenation that results in less oxygen available in our seas.

Large-scale ocean circulation naturally results in low-oxygen areas of the ocean called oxygen deficient zones (ODZs). The cycling of carbon and nutrients – the foundation of marine life, called biogeochemistry – is fundamentally different in ODZs than in oxygen-rich areas. Because researchers think deoxygenation will greatly expand the total area of ODZs over the next 100 years, studying how these areas function now is important in predicting and understanding the oceans of the future. This first expedition of 2016 led by Dr. Mak Saito from the Woods Hole Oceanographic Institution (WHOI) along with scientists from University of Maryland Center for Environmental Science, University of California Santa Cruz, and University of Washington

aimed to do just that, investigate ODZs.

During the 28 day voyage named “ProteOMZ,” researchers aboard R/V *Falkor* traveled from Honolulu, Hawaii to Tahiti to describe the biogeochemical processes that occur within this particular swath of the ocean’s ODZs. By doing so, they contributed to our greater understanding of ODZs, gathered a database of baseline measurements to which future measurements can be compared, and established a new methodology that could be used in future research on these expanding ODZs.

Marine Microbial Investigator Award: Investigator Mak Saito (MM Saito)

In support of obtaining deeper knowledge of major biogeochemically relevant proteins to inform a mechanistic understanding of global marine biogeochemical cycles.

[table of contents | back to top]

Program Information

Marine Microbiology Initiative (MMI)

Website: https://www.moore.org/initiative-strategy-detail?initiativeId=marine-microbiology-initiative

A Gordon and Betty Moore Foundation Program.

Forging a new paradigm in marine microbial ecology:

Microbes in the ocean produce half of the oxygen on the planet and remove vast amounts of carbon dioxide, a greenhouse gas, from the atmosphere. Yet, we have known surprisingly little about these microscopic organisms. As we discover answers to some long-standing puzzles about the roles that marine microorganisms play in supporting the ocean’s food webs and driving global elemental cycles, we realized that we still need to learn much more about what these organisms do and how they do it—including how they evolved and contribute to our ocean's health and productivity.

The Marine Microbiology Initiative seeks to gain a comprehensive understanding of marine microbial communities, including their diversity, functions and behaviors; their ecological roles; and their origins and evolution. Our focus has been to enable researchers to uncover the principles that govern the interactions among microbes and that govern microbially mediated nutrient flow in the sea. To address these opportunities, we support leaders in the field through investigator awards, multidisciplinary team research projects, and efforts to create resources of broad use to the research community. We also support development of new instrumentation, tools, technologies and genetic approaches.

Through the efforts of many scientists from around the world, the initiative has been catalyzing new science through advances in methods and technology, and to reduce interdisciplinary barriers slowing progress. With our support, researchers are quantifying nutrient pools in the ocean, deciphering the genetic and biochemical bases of microbial metabolism, and understanding how microbes interact with one another. The initiative has five grant portfolios:

Individual investigator awards for current and emerging leaders in the field.
Multidisciplinary projects that support collaboration across disciplines.
New instrumentation, tools and technology that enable scientists to ask new questions in ways previously not possible.
Community resource efforts that fund the creation and sharing of data and the development of tools, methods and infrastructure of widespread utility.
Projects that advance genetic tools to enable development of experimental model systems in marine microbial ecology.
We also bring together scientists to discuss timely subjects and to facilitate scientific exchange.

Our path to marine microbial ecology was a confluence of new technology that could accelerate science and an opportunity to support a field that was not well funded relative to potential impact. Around the time we began this work in 2004, the life sciences were entering a new era of DNA sequencing and genomics, expanding possibilities for scientific research – including the nascent field of marine microbial ecology. Through conversations with pioneers inside and outside the field, an opportunity was identified: to apply these new sequencing tools to advance knowledge of marine microbial communities and reveal how they support and influence ocean systems.

After many years of success, we will wind down this effort and close the initiative in 2021. We will have invested more than $250 million over 17 years to deepen understanding of the diversity, ecological activities and evolution of marine microbial communities. Thanks to the work of hundreds of scientists and others involved with the initiative, the goals have been achieved and the field has been profoundly enriched; it is now positioned to address new scientific questions using innovative technologies and methods.

[table of contents | back to top]

Funding

Funding Source	Award
Gordon and Betty Moore Foundation: Marine Microbiology Initiative (MMI)	GBMF3782
Schmidt Ocean Institute (SOI)	R/V Falkor 160115 SOI ProteOMZ Expedition

[table of contents | back to top]