SAGs from the Pacific Ocean OMZ

Website: https://www.bco-dmo.org/dataset/998887
Version: 1
Version Date: 2026-05-19

Project
» Collaborative Research: Microdiversity drives ecosystem function: SAR11 bacteria as models for oceanic nitrogen loss (SAR11 in OMZs)

Contributors	Affiliation	Role
Konstantinidis, Kostas	Georgia Institute of Technology (GA Tech)	Principal Investigator
Rauch, Shannon	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset describes assembled genomes of single-cell amplified genomes (SAGs) used in Zhao et al., ISME 2025.The abstract of this manuscript follows: Surveys of microbial communities (metagenomics) or isolate genomes have revealed sequence-discrete species. That is, members of the same species show >95% average nucleotide identity (ANI) of shared genes among themselves vs.<83% ANI to members of other species while genome pairs showing between 83% and 95% ANI are comparatively rare. In these surveys, aquatic bacteria of the ubiquitous SAR11 clade (Class Alphaproteobacteria) are an outlier and often do not exhibit discrete species boundaries, suggesting the potential for alternate modes of genetic differentiation. To explore evolution in SAR11, we analyzed high-quality, single-cell amplified genomes, and companion metagenomes from an oxygen minimum zone in the Eastern Tropical Pacific Ocean, where the SAR11 make up ~20% of the total microbial community. Our results show that SAR11 do form several sequence-discrete species, but their ANI range of discreteness is shifted to lower identities between 86% and 91%, with intra-species ANI ranging between 91% and 100%. Measuring recent gene exchange among these genomes based on a recently developed methodology revealed higher frequency of homologous recombination within compared to between species that affects sequence evolution at least twice as much as diversifying point mutation across the genome. Recombination in SAR11 appears to be more promiscuous compared to other prokaryotic species, likely due to the deletion of universal genes involved in the mismatch repair, and has facilitated the spread of adaptive mutations within the species (gene sweeps), further promoting the high intraspecies diversity observed. Collectively, these results implicate rampant, genome-wide homologous recombination as the mechanism of cohesion for distinct SAR11 species. The single-cell amplified genomes are available in the National Center for Biotechnology Information (NCBI) under BioProject number PRJNA1124867.

Table of Contents

Coverage

Location: Eastern Tropical Pacific Ocean
Spatial Extent: **N**:18.8833 **E**:-104.9 **S**:18.1667 **W**:-106.2833

Methods & Sampling

Water samples were collected using Niskin bottles on a rosette containing a conductivity–temperature–depth profiler (Sea-Bird SBE 911plus). The water was prepared by cryopreservation according to the protocol recommended by the Bigelow Single Cell Genomics Center (SCGC). Sorting was performed on 4 April 2023 (within <2 months from the date the first sample was collected) and SAGs were generated with the modified genomic DNA amplification technique, WGA-Y, which enables a substantially improved average genome recovery from single cells (service S-202). In total, 105 SAGs with Cp values <3 h were randomly selected for sequencing.

Genome assembly and draft annotation were performed by SCGC as described in the center's webpage https://scgc.bigelow.org/capabilities/service-description/.

Data Processing Description

Raw reads of SAGs were processed as described in Zhao et al., ISME 2025. They are available in NCBI under BioProject number PRJNA1124867.

BCO-DMO Processing Description

- Imported original file "Table S1. Sample information for the SAGs from the ETNP OMZ.xlsx" (sheet 1) twice: once with header row 106 as a lookup table named "station_list", and once with header row 2 as the main data table.
- Filtered the main data file to keep only rows where ROW_NUMBER < 103.
- Joined "station_list" to the main data file in half-outer mode, matching on the SAG ID number, bringing in columns Depth, Latitude, Longitude, SAG identifier, and Station.
- Renamed columns to comply with BCO-DMO naming conventions.
- Converted Latitude from degrees-decimal_minutes format (North) to decimal degrees.
- Converted Longitude from degrees-decimal_minutes format (West) to decimal degrees
- Rounded Latitude and Longitude to maximum 4 decimal places.
- Saved the final file as "998887_v1_sags_pacific_omz.csv".

[table of contents | back to top]

Related Publications

Zhao, J., Pachiadaki, M., Conrad, R. E., Hatt, J. K., Bristow, L. A., Rodriguez-R, L. M., Rossello-Mora, R., Stewart, F. J., & Konstantinidis, K. T. (2025). Promiscuous and genome-wide recombination underlies the sequence-discrete species of the SAR11 lineage in the deep ocean. The ISME Journal, 19(1). https://doi.org/10.1093/ismejo/wraf072
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Georgia Institute of Technology. Candidatus Pelagibacterales Raw sequence reads. 2024/06. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; 2011-. Available from: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA1124867. NCBI:BioProject: PRJNA1124867.

[table of contents | back to top]

Parameters

Parameter	Description	Units
SAG_identifier	SAG identifier	unitless
Station	Station number	unitless
Latitude	Latitude of sample collection	decimal degrees
Longitude	Longitude of sample collection	decimal degrees
Depth	Depth of sample collection	meters (m)
Genome	ID of the SAG sequenced	unitless
Clade	Clade designation according to the concatenated gene tree shown in Figure 1 of Zhao et al., ISME J	unitless
Completeness	Genome quality was assessed using CheckM v1.5.3	ranging from 0 to 100%
Contamination	Genome quality was assessed using CheckM v1.5.3	ranging from 0 to 100%
Strain_heterogeneity	Genome quality was assessed using CheckM v1.5.3	ranging from 0 to 100%
Experiment_Accession	NCBI experiment accession number	unitless
Study_Accession	NCBI study accession number	unitless
Sample_Accession	NCBI sample accession number	unitless
SAG_ID	ID of the SAG sequenced (same as Genome above but with hyphens included)	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina NextSeq 500
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	SAG paired-end libraries were created with Nextera XT kits (Illumina), sequenced with a NextSeq 500 (Illumina) available at the Bigelow Center.
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 911plus
Generic Instrument Name	CTD Sea-Bird SBE 911plus
Dataset-specific Description	Water samples were collected using Niskin bottles on a rosette containing a conductivity–temperature–depth profiler (Sea-Bird SBE 911plus).
Generic Instrument Description	The Sea-Bird SBE 911 plus is a type of CTD instrument package for continuous measurement of conductivity, temperature and pressure. The SBE 911 plus includes the SBE 9plus Underwater Unit and the SBE 11plus Deck Unit (for real-time readout using conductive wire) for deployment from a vessel. The combination of the SBE 9 plus and SBE 11 plus is called a SBE 911 plus. The SBE 9 plus uses Sea-Bird's standard modular temperature and conductivity sensors (SBE 3 plus and SBE 4). The SBE 9 plus CTD can be configured with up to eight auxiliary sensors to measure other parameters including dissolved oxygen, pH, turbidity, fluorescence, light (PAR), light transmission, etc.). more information from Sea-Bird Electronics

Dataset-specific Instrument Name	Niskin bottles
Generic Instrument Name	Niskin bottle
Dataset-specific Description	Niskin bottles on a rosette were used to collect the water samples.
Generic Instrument Description	A Niskin bottle (a next generation water sampler based on the Nansen bottle) is a cylindrical, non-metallic water collection device with stoppers at both ends. The bottles can be attached individually on a hydrowire or deployed in 12, 24, or 36 bottle Rosette systems mounted on a frame and combined with a CTD. Niskin bottles are used to collect discrete water samples for a range of measurements including pigments, nutrients, plankton, etc.

[table of contents | back to top]

Project Information

Collaborative Research: Microdiversity drives ecosystem function: SAR11 bacteria as models for oceanic nitrogen loss (SAR11 in OMZs)

Coverage: Eastern Tropical North Pacific, off Colima, Mexico

NSF Award Abstract:

This project studies how low oxygen availability influences the biodiversity and ecological role of SAR11 bacteria, one of the most abundant microbial groups in the ocean. The work involves oceanographic sampling across a range of oxygen and nutrient levels in the Eastern Tropical North Pacific Ocean. Using a combination of genomic, microbiological, and biogeochemical methods, the study identifies the mechanisms by which SAR11 strains diversify into separate niches and species and contribute biochemically to the ecosystem, likely through removing nitrogen from seawater. The project equips the next generation of researchers and educators, notably those from underrepresented minority groups, to use oceanographic, genomic, and microbiological concepts to meet contemporary scientific challenges. This goal is met through a combination of bioinformatic workshops that target undergraduate students from the University System of Puerto Rico, middle school teacher-training workshops, and middle or high school teacher internships in the investigator's labs. This multifaceted research and educational agenda fills a gap in our understanding of marine biological diversity, identifies the contribution of SAR11 bacteria to nutrient and carbon cycles in low oxygen oceans, and provides lessons and analytical tools to study microbial processes in other ecosystems.

This project has two aims. Aim 1 employs comparative metagenomic and single-cell genomic analyses to identify metabolic properties that distinguish SAR11 clades from low oxygen regions and processes of selection or gene flow operating across the clades. Aim 2 combines microbial transcriptomics, incubation experiments with isotope tracers, and culturing to delimit the oxygen and nutrient conditions that define the niche space of each SAR11 clade and to correlate SAR11 gene transcription with community biochemical outcomes, including nitrogen loss through denitrification. The results of these aims and the informatic methods used to probe microbial microdiversity are disseminated through genomics-focused undergraduate workshops, and new teacher-training educational modules, including lab-based modules focused on the importance of microorganisms under environmental change in the oceans. Data, manuscripts, and informatics workflows from this project are made publicly available. The results are critical for resolving the processes that create and sustain microbial diversity in the oceans and informing biogeochemical models that predict how diversity influences ecosystem processes.

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-2129823

[table of contents | back to top]