

# Collection of subsurface bacteria Nitrospirota and Nitrospinota genome data including IMG and NCBI accessions for sequence datasets in June 2021 (Slow Life in Crust project)

**Website:** <https://www.bco-dmo.org/dataset/933610>

**Data Type:** Synthesis

**Version:** 1

**Version Date:** 2024-12-29

## Project

» [Microbial activity in the crustal deep biosphere](#) (Slow Life in Crust)

Contributors	Affiliation	Role
<a href="#">Orcutt, Beth N.</a>	Bigelow Laboratory for Ocean Sciences	Principal Investigator
<a href="#">D'Angelo, Timothy</a>	Bigelow Laboratory for Ocean Sciences	Scientist
<a href="#">Merchant, Lynne M.</a>	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

## Abstract

The phyla Nitrospirota and Nitrospinota have received significant research attention due to their unique nitrogen metabolisms important to biogeochemical and industrial processes. These phyla are common inhabitants of marine and terrestrial subsurface environments and contain members capable of diverse physiologies in addition to nitrite oxidation and complete ammonia oxidation. We used phylogenomics and gene-based analysis with ancestral state reconstruction and gene-tree-species tree reconciliation methods to investigate the life histories of these two phyla. This dataset includes list of previously-published sequence datasets that were used for the analysis. The data and interpretations are published at DOI 10.1038/s41396-023-01397-x. Additional metadata such as NCBI accessions, assembly release dates, and NCBI taxon ids were added in December 2024.

## Table of Contents

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

## Coverage

**Location:** global

**Spatial Extent:** N:81.82 E:179.06 S:-77.01 W:-176.57

**Temporal Extent:** 2013-04-26 - 2023-01-12

## Methods & Sampling

*Genomic dataset collection, curation and quality control:* This study used publicly available genome assemblies. Existing publicly available genome assemblies were downloaded from the National Center for Biotechnology Investigation (NCBI) and the Integrated Microbial Genomes (IMG) database of the U.S. Department of Energy's Joint Genome Institute in June 2021. The Genome Taxonomy Database (GTDB) website (release 202) was used to access lists of NCBI assembly accession numbers for the following GTDB-assigned phyla: *Nitrospinota*, *Nitrospinota\_A* (now called *Tectomicrobia*), *Nitrospinota\_B*, *Nitrospirota*, *Nitrospirota\_A* (*Leptospirilla*). The IMG

assemblies were found using the same GTDB taxonomy classifier using the search function on the IMG website. IMG metagenome assemblies that were designated as “public” and “published” were also downloaded for these phyla. Duplicate entries between IMG and NCBI were manually removed.

## Data Processing Description

Quality control of the assemblies was performed using the CheckM qa workflow (v 1.07) to remove genomes with <50% genome completion and >10% sequence contamination, leaving genomes that fall within the MIMAG categories “medium” (>50% completion, <10% contamination) and “high” (>90% completion, <5% contamination). These resulting genomes were dereplicated with dRep, using default parameters, to remove nearly-identical assemblies. All genomes were then classified using the GTDB-tk classifier tool (v1.5.0, r202). Polyphyletic groups that were once considered a part of *Nitrospirota* and *Nitrospinota* (i.e., *Nitrospirota\_A* (*Leptospirilla*), *Nitrospinota\_A* (*Tectomicrobia*) and *Nitrospinota\_B*) were included only in the phylogenomic trees. These groups were not included in the gene cluster based functional analyses. All code to recreate these processes are available at [https://github.com/ts-dangelo/bioinformatic\\_scripts\\_python](https://github.com/ts-dangelo/bioinformatic_scripts_python).

## BCO-DMO Processing Description

Summary of processing on the submitted dataset by the dataset's BCO-DMO data manager.

### 1. Modified submitted accessions file

Fixed two shifted cell values in the submitted file "Supplemental Data 1.csv" and saved it as "Supplemental Data 1 fixed.csv" by using Excel.

The original file has a row number 333 with some values shifted into the wrong columns. Part of an "IsolationPlot" value was shifted to the right into the "Sample" column. And the "Sample" value was shifted to the right into the "Location" column. This was fixed by cutting the value in the "Sample" column and copying it into the "IsolationPlot" column for a final value of "saline water (ENVO:00002010) including plankton (ENVO:xxxxxxx)" and then cutting the value SAMN06450621 from the "Location" column and copying it into the "Sample" column.

### 2. Updated this fixed accessions file with IMG and NCBI metadata

Python code in Jupyter notebooks was created by the data manager and run to add metadata to the submitted accessions file. The scripts were run using the fixed file "Supplemental Data 1 fixed.csv" as input.

Data manager GitHub code repo (last commit = 7e8e2462a on 12/30/2024 updating docs)

[https://github.com/BCODMO/DM\\_scripts\\_by\\_dataset/tree/master/Orcutt\\_acces...](https://github.com/BCODMO/DM_scripts_by_dataset/tree/master/Orcutt_acces...)

Code run on 2024-12-29 using the following two Jupyter notebooks

[https://github.com/BCODMO/DM\\_scripts\\_by\\_dataset/blob/master/Orcutt\\_acces...](https://github.com/BCODMO/DM_scripts_by_dataset/blob/master/Orcutt_acces...)

[https://github.com/BCODMO/DM\\_scripts\\_by\\_dataset/blob/master/Orcutt\\_acces...](https://github.com/BCODMO/DM_scripts_by_dataset/blob/master/Orcutt_acces...)

The code and processing procedure are documented in markdown files in the GitHub repository

[https://github.com/BCODMO/DM\\_scripts\\_by\\_dataset/blob/master/Orcutt\\_acces...](https://github.com/BCODMO/DM_scripts_by_dataset/blob/master/Orcutt_acces...)

[https://github.com/BCODMO/DM\\_scripts\\_by\\_dataset/blob/master/Orcutt\\_acces...](https://github.com/BCODMO/DM_scripts_by_dataset/blob/master/Orcutt_acces...)

Eight new columns were added using python Jupyter notebooks run on 12/29/2024 with the output saved to the file named "updated\_pi\_supplemental\_table.csv".

columns added to the fixed file "Supplemental Data 1 fixed.csv": IMG\_genome\_id, BioProject, Corrected\_BioSample, GenBank\_assembly, release\_date, last\_updated\_date, publication\_date, ncbi\_taxon\_id

## column definitions

IMG\_genome\_id: IMG genome assembly id (extracted from values in "ID" column)

BioProject: NCBI BioProject accession

Corrected\_BioSample: correct NCBI BioSample accession

GenBank\_assembly: NCBI GenBank genome assembly accession

release\_date: Date genome assembly was released to the public

last\_updated\_date: Last date BioSample was updated on NCBI

publication\_date: Date genome assembly submitted to either IMG or NCBI

ncbi\_taxon\_id: NCBI taxon id of the sampled organism

## Steps to retrieve extra metadata

Used a search on the IMG website and python Jupyter notebooks to add NCBI accessions, dates, and NCBI taxon id metadata columns to provide more context to the submitted file.

There are two sources for the new metadata. One source is a search query on the IMG website (Integrated Microbial Genomes and Microbiomes <https://img.jgi.doe.gov/> <https://img.jgi.doe.gov/>) and the other source is NCBI API calls retrieving JSON reports.

The goal was to add corresponding BioProject accessions for each row in the 'ID' column of a submitted accessions file.

In the process of gathering supporting metadata by the dataset's BCO-DMO data manager, 15 BioSample values retrieved with the NCBI API using NCBI genome assembly accessions were found to be different than those in the submitted file's "Sample" column, which contains BioSample values. The 15 new BioSample values in the column "Corrected\_BioSample" in the final BCO-DMO dataset

933610\_v1\_nitrospirota\_and\_nitrospirota\_genomes.csv were manually checked on their corresponding NCBI web pages. It was found that the BioSample metadata and the corresponding NCBI genome assembly value confirmed the BioSample values retrieved via NCBI API calls were correct. The original 15 BioSample values were manually checked on their corresponding NCBI web pages, and the BioSample metadata and NCBI genome assembly accession for that BioSample did not match those listed in the submitted file.

The incorrect BioSample values in the "Sample" column are the following:

SAMN03222684, SAMN03222684, SAMN02862043, SAMN03222686, SAMN03222686, SAMEA3140934, SAMEA3140928, SAMN01922995, SAMN03203005, SAMN03203000, SAMN10411419, SAMN12518157, SAMN04315473, SAMN06659639, SAMN07631077

The correct BioSample values are found in the column "Corrected\_BioSample" in the primary dataset file 933610\_v1\_nitrospirota\_and\_nitrospirota\_genomes.csv.

The new columns were found with results from a query on the IMG website (Integrated Microbial Genomes and Microbiomes at <https://img.jgi.doe.gov/> (<https://img.jgi.doe.gov/>)) and python Jupyter notebooks using NCBI API calls and a series of joins with the fixed submitted accessions dataset.

To query the IMG website, a list of IMG genome ids was created using IMG id values in the "ID" column of the file "Supplemental Data 1 fixed.csv". The IMG genome ids in the "ID" column are prefixed with "IMG\_" in the "ID" column and some are suffixed with an underscore followed by a number.

IMG query URL: <https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=FindGenomes&page=geno...>

Downloaded the query results table from IMG (Integrated Microbial Genomes and Microbiomes <https://img.jgi.doe.gov/>) where the query used the list of IMG genome ids from the "ID" column. The results table contained NCBI accessions and other metadata associated with IMG genome ids.

Created a python Jupyter notebook to perform NCBI API calls using BioSample accessions, NCBI assembly accessions, and taxon names to retrieve data reports in JSON format. The reports JSON contained values for BioSample, BioProject, NCBI assembly accession, release date, last updated date, publication date, and NCBI taxon id.

The NCBI API calls used the NCBI version 2 REST API which is explained here

<https://www.ncbi.nlm.nih.gov/datasets/docs/v2/reference-docs/rest-api/>. The following three API endpoints were used:

/genome/accession/{accession}/dataset\_report

/biosample/accession/{accession}/biosample\_report  
/genome/taxon/{taxon\_name}/dataset\_report

The API endpoint /genome/accession/{accession}/dataset\_report was called with NCBI genome assembly accessions from the “ID” column of the accessions dataset table.

The API endpoint /biosample/accession/{accession}/biosample\_report was called with NCBI BioSample accessions from the “Sample” column of the accessions dataset table.

The API endpoint /genome/taxon/{taxon\_name}/dataset\_report was called using the phylum taxon names of Nitrospirota and Nitrospirota.

All the metadata results from the NCBI API endpoints were combined into one file with duplicates removed.

Created a python Jupyter notebook to run after attaining NCBI accessions and IMG genome ids and this retrieves and adds 8 extra metadata columns to the fixed submitted accessions dataset.

This notebook performed a series of joins on IMG genome ids, NCBI BioSamples and NCBI genome assembly accessions.

The first step was joining the IMG genome ids and its associated NCBI metadata on to the starting accessions dataset to create a new table.

The next steps were to perform a series of joins onto this new table.

Joined this new table with a table of NCBI accession values, which came from 3 NCBI API calls as discussed above, on the NCBI genome assembly values to create a new table.

A column containing NCBI BioSample values was created from this join. This column contains 15 BioSample values that are different than those in the original submitted accessions table.

A new column named ‘Fixed\_BioSample’ was created by merging NCBI BioSample values found with the IMG query and NCBI BioSample values found with NCBI API calls.

Using the BioSample values in the ‘Fixed\_BioSample’ column created by the Jupyter notebook, the NCBI API was called again and the new metadata values found for these corrected BioSample values was used in the final processed dataset. This retrieved metadata was used over any previously attained metadata. Any metadata values not retrieved using the ‘Fixed\_BioSample’ column values are filled in with the original metadata values.

Renamed columns in the table with new metadata, 'Fixed\_BioSample' was renamed 'Corrected\_BioSample' and this table was saved to a file named “updated\_pi\_supplemental\_table.csv”.

3. Processed the file “updated\_pi\_supplemental\_table.csv” with the BCO-DMO dataset processing tool named laminar

Renamed columns to conform with the BCO-DMO parameter naming convention of replacing spaces with underscores, and starting parameter names with an alphabetical character. Renamed the parameter ““# predicted genes” to “num\_predicted\_genes”.

Renamed the parameter ncbi\_taxon\_id to NCBI\_organism\_taxid.

Converted UTC date parameters with the format %Y-%m-%dT%H:%M:%S.%f to the ISO UTC datetime format of %Y-%m-%dT%H:%M:%SZ.

The organism taxon id is an integer, so removed the trailing decimal 0 to convert it into an integer.

Reordered the parameters so that genome metadata is towards the end of the dataset table and accessions metadata is towards the start.

Saved this updated table as the final dataset named 933610\_v1\_nitrospirota\_and\_nitrospirota\_genomes.csv

## Problem Description

In the process of gathering supporting metadata by the dataset's BCO-DMO data manager, 15 BioSample values retrieved with the NCBI API using NCBI genome assembly accessions were found to be different than those in the submitted file's "Sample" column, which contains BioSample values. The 15 new BioSample values in the column "Corrected\_BioSample" in the final BCO-DMO dataset 933610\_v1\_nitrospirota\_and\_nitrospina\_genomes.csv were manually checked on their corresponding NCBI web pages. It was found that the BioSample metadata and the corresponding NCBI genome assembly value confirmed the BioSample values retrieved via NCBI API calls were correct. The original 15 BioSample values were manually checked on their corresponding NCBI web pages, and the BioSample metadata and NCBI genome assembly accession for that BioSample did not match those listed in the submitted file.

The incorrect BioSample values in the "Sample" column are the following:  
SAMN03222684, SAMN03222684, SAMN02862043, SAMN03222686, SAMN03222686, SAMEA3140934, SAMEA3140928, SAMN01922995, SAMN03203005, SAMN03203000, SAMN10411419, SAMN12518157, SAMN04315473, SAMN06659639, SAMN07631077

The correct BioSample values are found in the column "Corrected\_BioSample" in the primary dataset file 933610\_v1\_nitrospirota\_and\_nitrospina\_genomes.csv.

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

## Data Files

File
<b>933610_v1_nitrospirota_and_nitrospina_genomes.csv</b> (Comma Separated Values (.csv), 156.94 KB) MD5:236240cefcddfa9c0c298e17fd6cc62
Primary data file for dataset ID 933610, version 1

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

## Related Publications

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23), 5315-5316.

<https://doi.org/10.1093/bioinformatics/btac672>

*Software*

Dangelo, T.S. (2024). *bioinformatic\_scripts\_python*. GitHub. commit: d8394a9ed27. [https://github.com/ts-dangelo/bioinformatic\\_scripts\\_python](https://github.com/ts-dangelo/bioinformatic_scripts_python)

*Software*

D'Angelo, T., Goordial, J., Lindsay, M. R., McGonigle, J., Booker, A., Moser, D., Stepanauskus, R., & Orcutt, B. N. (2023). Replicated life-history patterns and subsurface origins of the bacterial sister phyla Nitrospirota and Nitrospina. *The ISME Journal*, 17(6), 891-902. <https://doi.org/10.1038/s41396-023-01397-x>

*Results*

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R. A., Otilar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D., & Dubchak, I. (2011). The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research*, 40(D1), D26-D32. <https://doi.org/10.1093/nar/gkr947>

*Methods*

National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2024 Dec 29]. Available from:

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

*Methods*

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, 11(12), 2864-2868. <https://doi.org/10.1038/ismej.2017.126>

*Software*

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2021). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and

complete genome-based taxonomy. Nucleic Acids Research, 50(D1), D785–D794.

<https://doi.org/10.1093/nar/gkab776>

#### Methods

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7), 1043–1055. <https://doi.org/10.1101/gr.186072.114>

#### Software

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

## Parameters

Parameter	Description	Units
ID	Name of sequence data	unitless
IMG_genome_id	IMG genome assembly ID (Integrated Microbial Genomes and Metagenomes)	unitless
GenBank_assembly	NCBI genome assembly accession (National Center for Biotechnology Information)	unitless
Sample	Submitter's original NCBI BioSample accession (National Center for Biotechnology Information)	unitless
Corrected_BioSample	NCBI BioSample accession retrieved with NCBI API call using NCBI Genome assembly accession (National Center for Biotechnology Information)	unitless
BioProject	NCBI BioProject accession (National Center for Biotechnology Information)	unitless
release_date	Release date for genome to be publicly accessible	unitless
last_updated_date	Last date BioSample metadata was updated	unitless
publication_date	Date that genome assembly was submitted to either IMG (Integrated Microbial Genomes and Metagenomes) or NCBI (National Center for Biotechnology Information)	unitless
Domain	Taxonomic phylogenetic domain of the sequence data	unitless
Phylum	Taxonomic phylogenetic phylum of the sequence data	unitless
Class	Taxonomic phylogenetic class of the sequence data	unitless

Order	Taxonomic phylogenetic order of the sequence data	unitless
Family	Taxonomic phylogenetic family of the sequence data	unitless
Genus	Taxonomic phylogenetic genus of the sequence data. Blank if could not be identified.	unitless
Species	Taxonomic phylogenetic species of the sequence data. Blank if could not be identified.	unitless
NCBI_organism_taxid	NCBI taxon id of organism	unitless
Isolation_Source	descriptive category of environment where data originated from	unitless
IsolationPlot	Grouping of Isolation Source into predefined categories for plotting	unitless
Location	Descriptive location where sequence data originated	unitless
Coordinates	Downloaded latitude and longitude data from NCBI	degrees
Latitude	latitude of sample, south is negative	decimal degrees
Longitude	longitude of sample, west is negative	decimal degrees
Completeness	Estimated genome completeness	unitless
Contamination	Estimated contamination of genome data	unitless
Genome_Size_bp	Length of genome in basepairs	basepairs
Estimated_Genome_Length	basepairs length of contig	unitless
num_predicted_genes	The number of predicted genes in the contig	unitless
GC	percentage of GC basepairs out of all basepairs	unitless
Coding_density	percent of contig with genomic information in a known gene	unitless

## Project Information

### Microbial activity in the crustal deep biosphere (Slow Life in Crust)

**Coverage:** Juan de Fuca Ridge flank CORs, 47N/127W

#### *NSF Award Abstract:*

The marine deep biosphere is the habitat for life existing under the sea floor. The zone has remarkably low energy sources creating a paradox of how life can persist there. Resolving this energy paradox is a grand challenge in deep biosphere research. The Juan de Fuca Ridge flank off the coast of Washington, USA, is an accessible, low energy environment making it an attractive location for addressing this challenge. A series of experiments will be conducted on the seafloor at the Juan de Fuca Ridge flank, using established subseafloor observatories that access the crustal deep biosphere, to provide the first direct in situ measurement of microbial activity in the crustal subsurface. This project will provide essential information about the ability of life to survive under conditions that we are not able to replicate in the laboratory, but that are increasingly important for understanding microbial community interaction in the environment. This information can then be used in models of global microbial activity for estimating the impact of this biosphere on elemental cycling, transforming our understanding of microbial processes within this vast subseafloor habitat. To communicate these discoveries to the public, the project will include a ship-to-shore outreach program during the cruise. In addition public lectures will be presented, and an interactive display of deep-sea video footage will be set up for the annual public Open House at the Bigelow Laboratory for Ocean Sciences in Maine. Diverse undergraduate students and a postdoctoral researcher will be recruited to participate in the research and public outreach activities.

This project proposes to leverage existing subsurface infrastructure on the eastern flank of the Juan de Fuca Ridge with advances in single-cell based molecular and geochemical approaches to make fundamental new discoveries about the activity of life in the deep crustal biosphere. During a two-week research cruise, the research team will incubate crustal fluids in situ and in the laboratory with labeled substrates for tracking single-cell activity, coupled with radioisotope tracer activity and potentiostat measurements, with the objective of determining in situ and potential rates of activity and cellular physiology. The research will also identify which metabolisms active microorganisms utilize under in situ and laboratory conditions, the rates of these processes, and the microorganisms involved. The results are expected to provide explicit hypothesis testing of microbial activity and in situ microbial growth rates from the crustal deep biosphere to transform understanding of microbial activity in the crustal deep biosphere and generate critical information about the ability of life to survive under low energy conditions.

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

---

## Funding

Funding Source	Award
<a href="#">NSF Division of Ocean Sciences (NSF OCE)</a>	<a href="#">OCE-1737017</a>

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