The Advantages of Submitting 'Omic Data Links to BCO-DMO

Biological and Chemical oceanography are increasingly using 'omic tools (e.g. genomic, transcriptomic, proteomics, and metabolomic) to study ocean environments, and there is considerable value of these high density datasets due to their ability to document the environment status in time and space. We want to make data generators aware that there are many advantages to submitting omics data links to BCO-DMO. While there are dedicated repositories for omics data (Table 1), data discoverability and accessing metadata and environmental data are well known to be major obstacles to reusing omic datasets. This is generally because the repositories are designed for biomedical research and hence they often lack the ability to connect to environmental research and associated metadata. **Although BCO-DMO does not host raw sequence or mass spectrometry data, it can easily link to these repositories.** In doing so BCO-DMO can serve as a valuable means to allow researchers to discover your data and place it in its appropriate environmental context. Because BCO-DMO can organize datasets by their expeditions, environmental data associated with where the meta'omic data was discovered can be easily connected (Figure 1). Similarly BCO-DMO can organize data by grants and/or projects, allowing laboratory experimental data to be associated with omic data. BCO-DMO's site is optimized for data discovery using search engines, and data types can be searched for within BCO-DMO's holdings. **With a minor effort in submission, all the hard work you have put into collecting your dataset can be used to help other researchers throughout the world by making your data easier to find and interpret.** You, the data generator, will benefit from submission through increased citations and collaborations.

We recommend that all sequencing efforts be submitted to national biological archiving services such as NCBI or ENA. Data submitted to these archives can be seamlessly connected to their associated environmental data stored in BCO-DMO by unifying all sequence projects under a BioProject in NCBI or Study in ENA (Figure 2). Thus, the BioProject structure provides a single accession key which can be used to find all of the data associated with a single research expedition or experimental project. Within the larger NCBI BioProject (ENA Study), NCBI BioSamples (ENA Sample) should be assigned for each field sample or experimental treatment condition. BioProjects and BioSamples contain and associate both the raw and post-processed sequencing data. The raw sequencing data should be deposited into NCBIs corresponding Sequence Read Archive (SRA), organized into an SRA Study (which links to an overall NCBI BioProject) with individual SRA Samples (akin to the BioSample; Figure 3). All genetic sequencing data types can be organized into this broader organizational structure, whether they are metagenomic sequencing, cultured isolates, 16S amplicon sequencing, etc.. The BioProject ID and associated BioSample IDs can then be linked with data stored in BCO-DMO facilitating data discovery and access.

We also recognize that sequence assemblies are particularly valuable data resources that are often not accepted by raw sequence repositories as described above, and that these resources are either not being archived or are in impermanent locations (dropbox and lab websites etc) that are difficult to discover and whose links could change. If you have assemblies that are of potential community use, you can work with the BCO-DMO data managers to explore a means to archive it, such as using Zenodo or BCO-DMO's servers.

For mass spectra datasets (e.g. proteomics and metabolomics) links to raw data repositories (e.g. PRIDE and MassIVE) can be shared and smaller processed datasets (protein or metabolite concentrations) can be hosted at BCO-DMO. The mass spectrometry community has created a consortium (The PX consortium) to "provide a common framework and infrastructure for the cooperation of proteomics resources by defining and implementing consistent, harmonized, user-friendly data deposition and exchange procedures among the members". As a result, data deposited in one of the four participating repositories (PRIDE, PeptideAtlas, MassIVE, and jPost) are discoverable throughout the ProteomeXchange system using universal PX identifiers. We encourage submissions of raw data to this consortium following its standards. The identifier and weblink can then be hosted with valuable metadata and environmental data at BCO-DMO.

The 'omic landscape is an exciting and rapidly changing field with potential for contributing to discoveries through data re-analysis and synthesis. At BCO-DMO we aim to help make your data discoverable and to facilitate data reuse. Feedback is always welcome as well at info@bco-dmo.org.

Omic Data and BCO-DMO document. Version 6/10/2021

Table 1. Raw Data Repositories BCO-DMO has linked to

Repository	**Data types**
NCBI (SRA, Bioproject)	DNA, RNA sequence
EBI	DNA, RNA sequence
PRIDE	Mass Spectra (proteomics, metabolomics)
MassIVE	Mass Spectra (proteomics, metabolomics)
ProteomeXChange (PX consortium)	Mass Spectra (proteomics, metabolomics)
MetaboLights	Metabolomics database

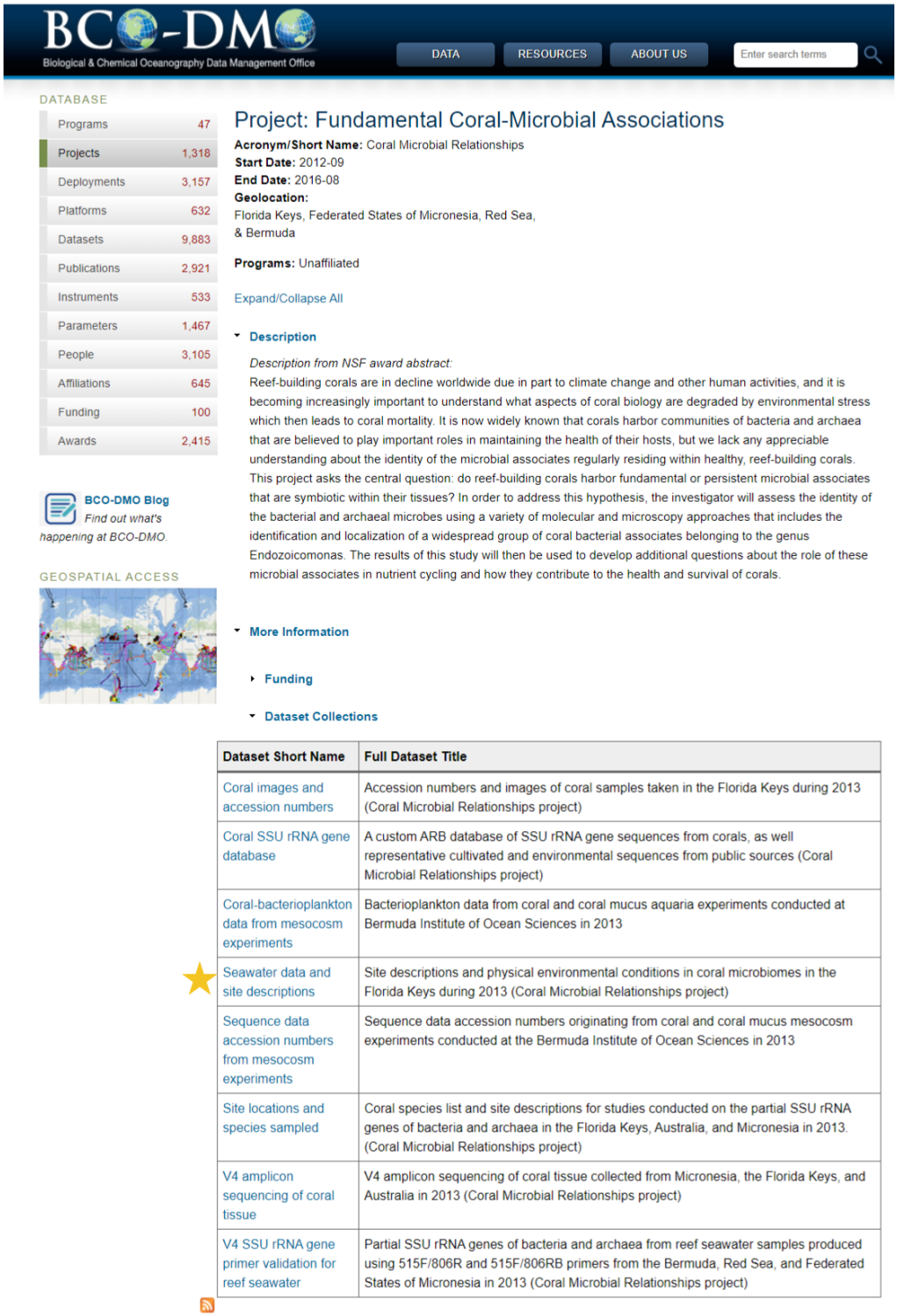

Figure 1. Example of a BCO-DMO project page that connects 'omic data with environmental datasets and sample collection metadata. The star symbol indicates a dataset within this larger project that is illustrated in Figures 2 & 3.

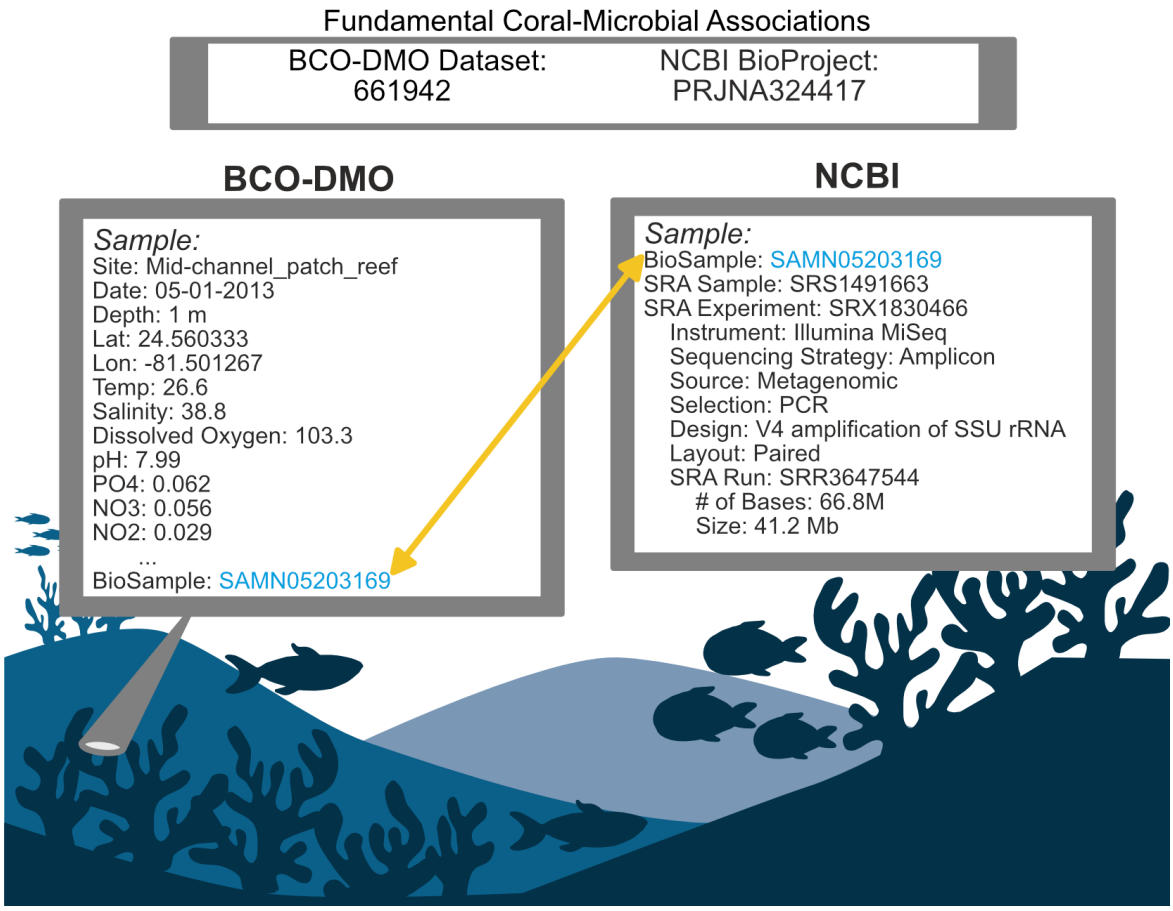

Figure 2. An example of a physical marine sample found within the Fundamental Coral-Microbiome Project in BCO-DMO. An individual physical sample has associated environmental data within BCO-DMO. Additionally, an accession for a BioSample associated with this physical sample is found within BCO-DMO which facilitates linkage to the nucleotide sequencing data for this sample found within the NCBI repository. Additional sequencing information can be obtained for this genomic data within the NCBI repository based upon this accession link via the BioSample.

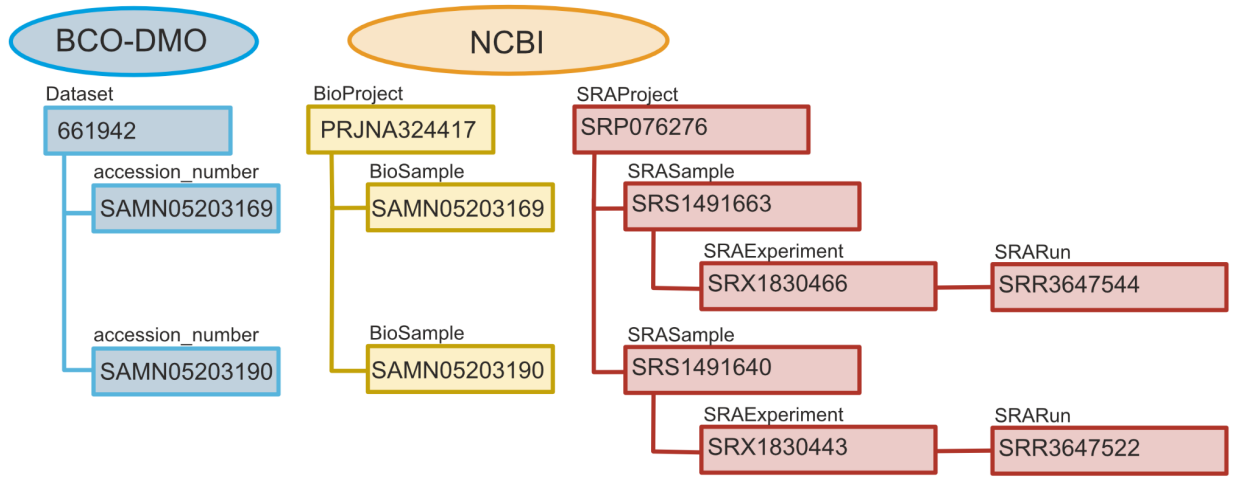

Figure 3. Example of data structures from the Coral-Microbial Association Dataset #661942 found within BCO-DMO and the genomic data associated with the samples found within the NCBI repository where the sequencing data is maintained. Within NCBI, the sequencing data is organized under an overarching BioProject, with raw sequencing data maintained within the SRA database in the depicted structure. Note, only 2 biosamples are depicted in this figure whereas the dataset actually has 10 separate BioSamples.