NEON USER GUIDE TO MICROBIAL COMMUNITY COMPOSITION (DP1.10081.001; DP1.20141; DP1.20086.001)

PREPARED BY	ORGANIZATION
Lee Stanish	TOS
Stephanie Parker	AOS

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	2/27/2018	Initial Release
B	12/16/2020	Included general statement about usage of neonUtilities R package and statement about possible location changes; Updated Figure 1; Section 3.3: Added Sampling Design Changes section and included changes to sampling frequency for microbial analyses; Section 3.7.1: Updated description of Associated Data Streams for bundled Soil physical and chemical properties data product; Section 3.9: Clarified data relationships and joining data across related data products.
C	04/25/2022	Updated section 4.3 Data Revision with latest information regarding data release.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

This document provides an overview of the data included in this NEON Level 1 data product, the quality controlled product generated from raw Level 0 data, and associated metadata. In the NEON data products framework, the raw data collected in the field - for example, soil temperature from a single collection event - are considered the lowest level (Level 0). Raw data that have been quality checked via the steps detailed herein, as well as simple metrics that emerge from the raw data are considered Level 1 data products.

The text herein provides a discussion of measurement theory and implementation, data product provenance, quality assurance and control methods used, and approximations and/or assumptions made during L1 data creation.

1.2 Scope

This document describes the steps needed to generate the L1 data products for Microbial Community Composition and associated metadata measured on aquatic and terrestrial samples from input data. Three separate data products are described herein:

1. Soil Microbe Community Composition (DP1.10081.001)
2. Surface Water Microbe Community Composition (DP1.20141.001)
3. Benthic Microbe Community Composition (DP1.20086.001)

This document also provides details relevant to the publication of the data products via the NEON data portal, with additional details in the files, NEON Data Variables for Soil Microbe Community Composition (DP1.10081.001) (AD[05]), NEON Data Variables for Surface Water Microbe Community Composition (DP1.20141.001) (AD[06]), and NEON Data Variables for Benthic Microbe Community Composition (DP1.20086.001) (AD[07]), provided in the download package for each of these three data products.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the laboratory data from samples generated by the following field sampling protocols: TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[10]) for upland soil samples; TOS Standard Operating Procedure: Wetland Soil Sampling (AD[11]) for wetland soil samples; or AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[12]) for aquatic samples. The raw data that are processed as described in this document are detailed in the file, NEON Raw Data Validation for Microbial Community Composition (DP0.10081.001) (AD[04]), provided in the download package for this data product. Please note that raw data products (denoted by 'DP0') may not always have the same numbers (e.g., '10033') as the corresponding L1 data product.

2 RELATED DOCUMENTS AND ACRONYMS

2.1 Associated Documents

AD[01]	NEON.DOC.000001	NEON Observatory Design (NOD) Requirements
AD[02]	NEON.DOC.000913	TOS Science Design for Spatial Sampling
AD[03]	NEON.DOC.002652	NEON Products Catalog
AD[04]	Available with data download	Validation csv
AD[05]	Available with data download	Variables csv
AD[06]	Available with data download	Variables csv
AD[07]	Available with data download	Variables csv
AD[08]	NEON.DOC.000908	TOS Science Design for Microbial Diversity
AD[09]	NEON.DOC.001152	NEON Aquatic Sample Strategy Document
AD[10]	NEON.DOC.014048	TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling
AD[11]	NEON.DOC.004130	TOS Standard Operating Procedure: Wetland Soil Sampling
AD[12]	NEON.DOC.003044	AOS Protocol and Procedure: Aquatic Microbial Sampling
AD[13]	NEON.DOC.000008	NEON Acronym List
AD[14]	NEON.DOC.000243	NEON Glossary of Terms
AD[15]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[16]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[17]	Available on NEON data portal	NEON Ingest Conversion Language
AD[18]	Available with data download	Categorical Codes csv

2.2 Acronyms

rRNA	Ribosomal ribonucleic acid
16S	Small subunit of the rRNA gene
ITS	Internal transcribed spacer region of the small subunit of the rRNA gene

3 DATA PRODUCT DESCRIPTION

The Microbial Community Composition data products provide taxonomic data for bacteria, archaea, and fungi for soil and aquatic samples. NEON targets a region of the 16S ribosomal RNA gene operon to measure bacteria and archaea, and the internal transcribed spacer (ITS) region of the ribosomal RNA gene operon to measure fungi. Data are generated using high-throughput technology that produces many thousands of sequence reads per sample (Armougom and Didier, 2009; Klindworth et al., 2013). The sequence data, which comprise the Microbe Marker Gene Sequences Data Products (DP1.10108.001; DP1.20280.001; DP1.20282.001), are used to generate taxon tables provided in this data product. The sampling plan implements the guidelines and requirements described in the Science Designs for TOS Terrestrial Microbial Diversity and Aquatic Sampling (AD[08] and AD[09]). Information on sample collection methods for each data product can be found in the field user guides:

- Soils: NEON User Guide to Soil Physical Properties, Distributed Periodic (DP1.10086.001)
- Surface water: NEON User Guide for Surface Water Microbe Cell Count (DP1.20138.001)
- Benthic habitats: NEON User Guide for Aquatic Benthic Microbe Collection (DP0.20270.001)

In general, samples are minimally processed in the field in order to reduce the introduction of microbial contaminants. After collection, samples are frozen in the field on dry ice and transported to ultra-low freezers at the NEON field laboratories. Samples are shipped to an analytical laboratory where DNA extraction, sample library preparation and DNA sequencing occur. The sequence data are quality-filtered and processed bioinformatically to generate taxon tables. An overview of the current workflow is shown in Figure 1. Note that, given the rapid pace of technological and bioinformatic development, the exact methods for sequence processing and taxon-calling will change over time.

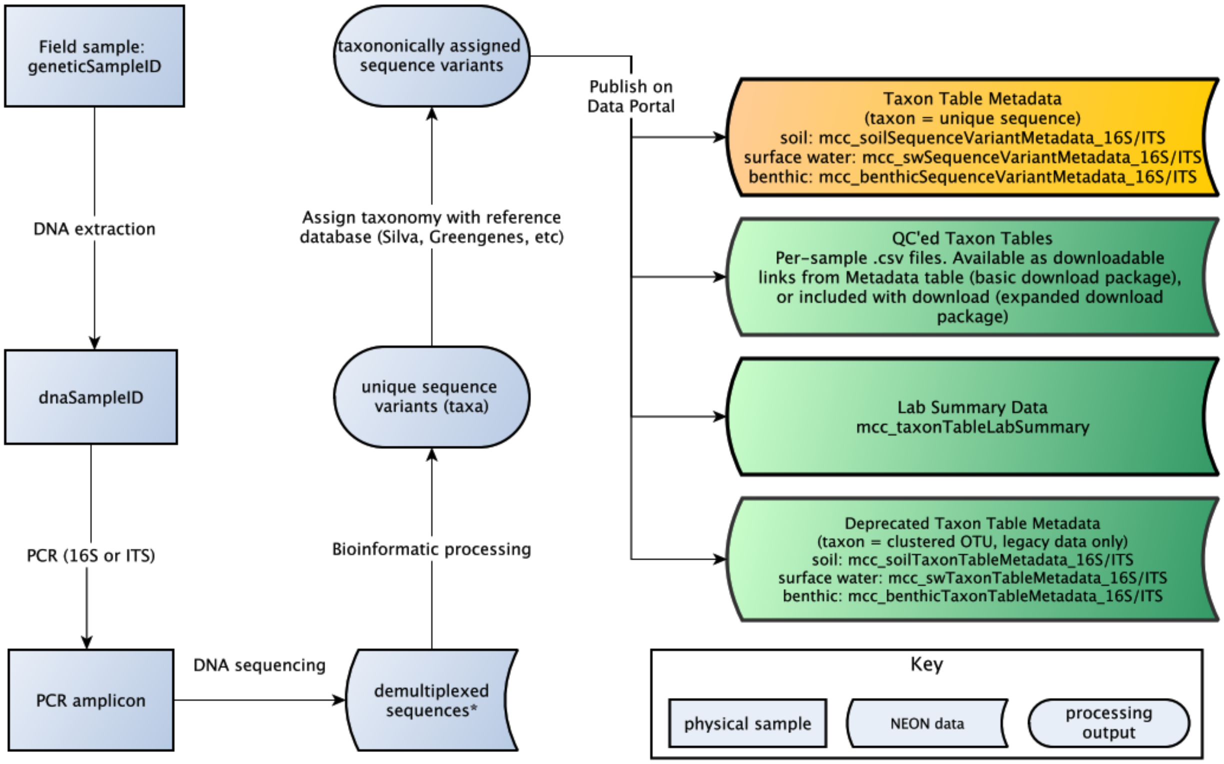
Field sample: geneticSampleID
DNA extraction
dnaSampleID
PCR (16S or ITS)
PCR amplicon
DNA sequencing
demultiplexed sequences*
Bioinformatic processing
unique sequence variants (taxa)
Assign taxonomy with reference database (Silva, Greengenes, etc)
taxononically assigned sequence variants
Publish on Data Portal
Taxon Table Metadata (taxon = unique sequence) soil: mcc_soilSequenceVariantMetadata_16S/ITS surface water: mcc_swSequenceVariantMetadata_16S/ITS benthic: mcc_benthicSequenceVariantMetadata_16S/ITS
QC'ed Taxon Tables Per-sample .csv files. Available as downloadable links from Metadata table (basic download package), or included with download (expanded download package)
Lab Summary Data mcc_taxonTableLabSummary
Deprecated Taxon Table Metadata (taxon = clustered OTU, legacy data only) soil: mcc_soilTaxonTableMetadata_16S/ITS surface water: mcc_swTaxonTableMetadata_16S/ITS benthic: mcc_benthicTaxonTableMetadata_16S/ITS
Key
physical sample
NEON data
processing output

Figure 1: The microbial community composition data product workflow from sample collection to publication on the NEON Data Portal. Published data tables in the gold box are part of the basic download package, while tables in the green boxes are part of the expanded download package. *Demultiplexed sequence data available in the microbe marker gene sequences data products, DP1.10108, DP1.20280, DP1.20282.

3.1 Spatial Sampling Design

Sampling for microbial community composition analysis is executed at all NEON sites, with data reported at the resolution of a single sampling location.

For soils, this equates to a randomly-assigned X,Y coordinate ($\pm$ 0.5 meters) within a NEON plot. Ten plots are sampled at 3 randomly selected locations within each plot (Figure 2). In general, only the surface horizon is sampled to a maximum depth of 30cm, and horizons are broadly defined as either organic (O) or mineral (M).

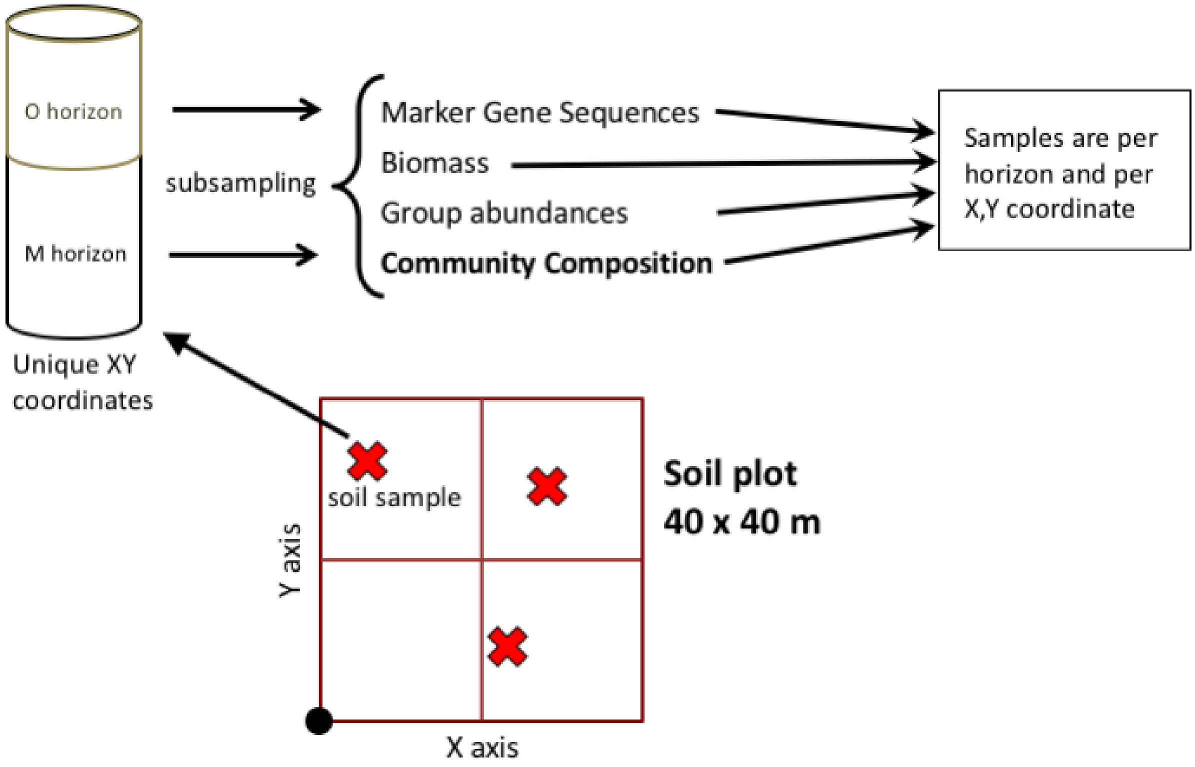

Figure 2: Overview of soil microbial field sampling, spatial design, and analysis workflow.

For aquatic surface water samples, this equates to the buoy sensor station and inlet/outlet locations within a lake, the buoy sensor station for large rivers, or the downstream sensor array for wadeable streams. For aquatic benthic samples, this equates to up to eight locations within a 1 km reach (Figure 3).

The spatial designs for the microbial community composition data products are described in more detail in the Data Product User Guides for Soil Physical Properties (DP1.10086.001), Aquatic Surface Water Cell Counts (DP1.20138.001), and Aquatic Benthic Microbe Collection (DP0.20270.001). For a description of the methods used in terrestrial plot selection, refer to the TOS Science Design for Spatial Sampling (AD[02]).

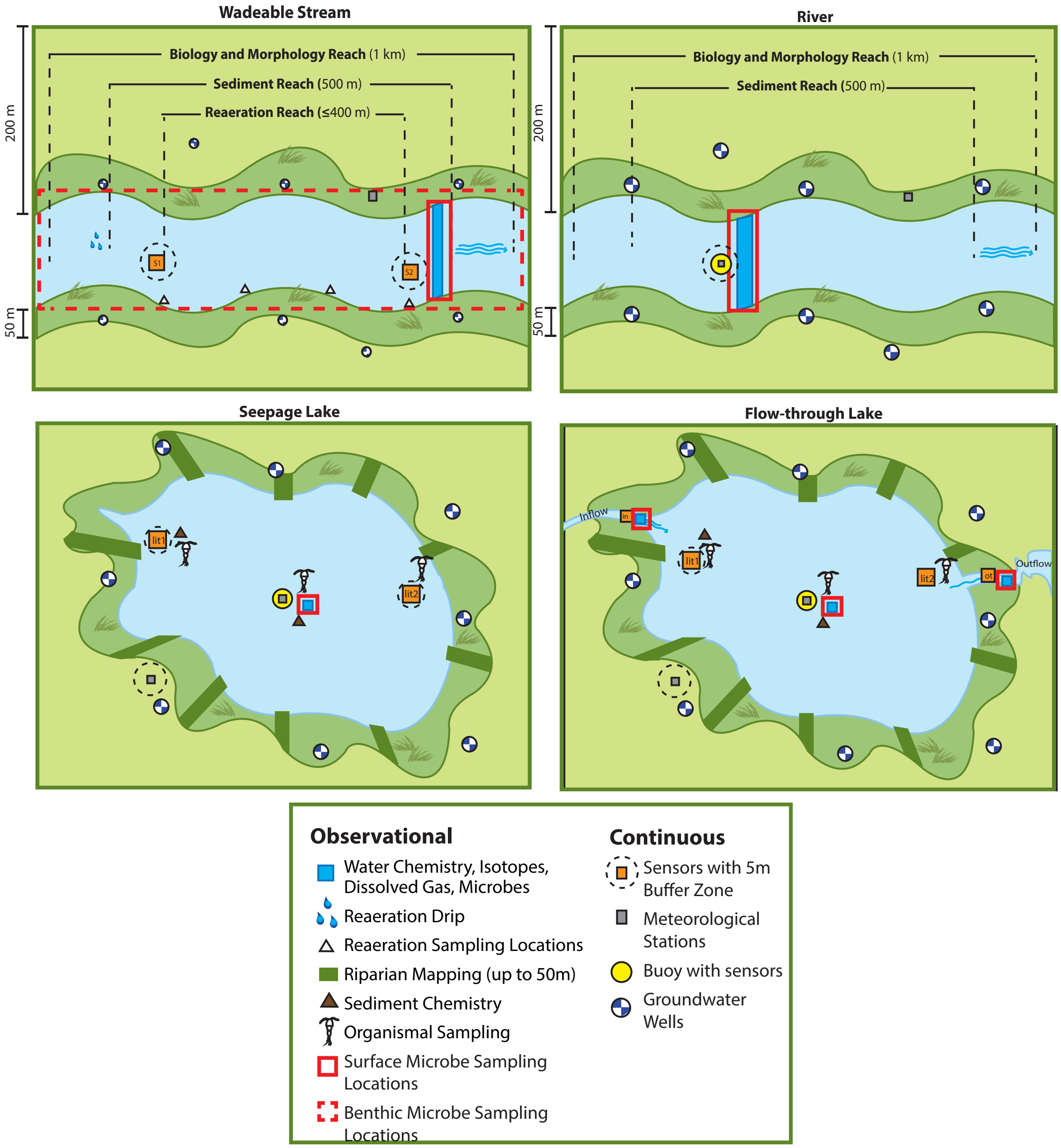
Wadeable Stream
Biology and Morphology Reach (1 km)
Sediment Reach (500 m)
Reaeration Reach (≤400 m)
200 m
50 m
S1
S2
River
Biology and Morphology Reach (1 km)
Sediment Reach (500 m)
200 m
50 m
Seepage Lake
lit1
lit2
Flow-through Lake
Inflow
in
lit1
lit2
ot
Outflow
Observational
Water Chemistry, Isotopes, Dissolved Gas, Microbes
Reaeration Drip
Reaeration Sampling Locations
Riparian Mapping (up to 50m)
Sediment Chemistry
Organismal Sampling
Surface Microbe Sampling Locations
Benthic Microbe Sampling Locations
Continuous
Sensors with 5m Buffer Zone
Meteorological Stations
Buoy with sensors
Groundwater Wells

Figure 3: Generic NEON aquatic site layouts with microbial sampling locations highlighted in red boxes.

As much as possible, sampling occurs in the same locations over the lifetime of the Observatory. However, over time some sampling locations may become impossible to sample, due to disturbance or other local changes. When this occurs, the location and its location ID are retired. A location may also shift to slightly different coordinates. Refer to the locations endpoint of the NEON API for details about locations that have been moved or retired: https://data.neonscience.org/data-api/endpoints/locations/

3.2 Temporal Sampling Design

For all samples, the temporal resolution is that of a single collection date. For a more detailed description of the temporal frequency of sampling and general field methods, refer to TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[10]) or AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[12]) for soil and aquatic sampling protocols, respectively. Descriptions of the upstream field data products can be found in the Data Product User Guides for soil (DP1.10086.001), aquatic surface water (DP1.20138.001), and benthic (DP1.20270.001) field sampling.

3.3 Sampling Design Changes

Over the course of early operations, the design for microbial sampling has changed. Below is a list of previous sampling strategies that differ from the current design, with applicable years indicated.

3.3.1 Soils

- 2013 - 2018: Subsamples were collected for microbial community composition analysis during every soil sampling bout and at all sites.
- 2018 - current: Subsamples are collected for microbial community composition analysis during every soil sampling bout at one site per domain and during all ‘coordinated’ bouts at any site.

3.3.2 Aquatics

- 2014 - 2018: At seepage lake sites (lacking a true inlet and outlet), surface water samples were collected at the buoy sensor station and inlet/outlet locations.
- 2018 - current: Surface water samples are collected at only the buoy sensor station at seepage lake sites (lacking a true inlet and outlet).

3.4 Variables Reported

All variables reported from the analytical laboratory (L0 data) are listed in the file, NEON Raw Data Validation for Microbial Community Composition (DP0.10081.001) (AD[04]). All variables reported in the published data (L1 data) are also provided separately. For variables that are not present in the taxon tables, the variables are located in the following files:

- NEON Data Variables for Soil Microbe Community Composition (DP1.10081.001) (AD[05])
- NEON Data Variables for Surface Water Microbe Community Composition (DP1.20141.001) (AD[06])
- NEON Data Variables for Benthic Microbe Community Composition (DP1.20086.001) (AD[07])

Field names that are present in the taxon tables are defined in the Table below.

Table 1: Variables present in the taxon tables. Note that scannable barcodes refer to physical barcode labels used for sample tracking.

fieldName	**Description**	**dataType**	**units**
dnaSampleID	Identifier for DNA sample	string	NA
dnaSampleCode	Scannable barcode of a DNA sample	string	NA
completeTaxonomy	Full taxonomic hierarchy for identified organism with taxonomic levels separated by a semicolon	string	NA
domain	The scientific name of the domain in which the taxon is classified	string	NA
kingdom	The scientific name of the kingdom in which the taxon is classified	string	NA
phylum	The scientific name of the phylum or division in which the taxon is classified	string	NA
class	The scientific name of the class in which the taxon is classified	string	NA
order	The scientific name of the order in which the taxon is classified	string	NA
family	The scientific name of the family in which the taxon is classified	string	NA
genus	The scientific name of the genus in which the taxon is classified	string	NA
specificEpithet	The specific epithet (second part of the species name) of the scientific name applied to the taxon	string	NA
scientificName	Scientific name, or name of the lowest level taxonomic rank that can be determined	string	NA
individualCount	Number of individuals of the same type	unsigned integer	number

Field names have been standardized with Darwin Core terms (http://rs.tdwg.org/dwc/; accessed 16 February 2014), the Global Biodiversity Information Facility vocabularies (http://rs.gbif.org/vocabulary/gbif/; accessed 16 February 2014), and the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014).

To the extent possible, metadata names and terms are standardized according to the Genomics Standards Consortium, http://gensc.org/ (Kottmann et al., 2008; Yilmaz et al., 2011; Field et al., 2011). Efforts are also made to conform with the ENVO ontology (http://www.obofoundry.org/ontology/envo.html).

NEON TOS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and GEOID09 for its reference gravitational ellipsoid. NEON Aquatic spatial data uses the Earth

Graviational Model 96 (EGM96) for its reference graviational ellipsoid. Latitudes and longitudes are denoted in decimal notation to six decimal places, with longitudes indicated as negative west of the Greenwich meridian.

Some variables described in this document may be for NEON internal use only and will not appear in downloaded data.

3.5 Spatial Resolution and Extent

The finest resolution at which spatial data are reported is a single sampling location. For soils, this corresponds to a single X,Y coordinate location within a plot. For aquatics, this corresponds to a single station or habitat unit within a site.

3.5.1 Soils

sampleID (unique ID given to the individual soil sampling location and horizon) → **plotID** (ID of plot within site) → **siteID** (ID of NEON site) → **domainID** (ID of a NEON domain).

The spatial data are located in the data product Soil Physical Properties, distributed periodic (DP1.10086), in the table ***sls_soilCoreCollection***. The spatial data are measured at the plot *centroid*, and have an accuracy of ± 20 meters. A more precise measurement may be determined by calculating the offset from the plot centroid using the variables **coreCoordinateX** and **coreCoordinateY**. Refer to the User Guide for Soil Physical Properties, distributed periodic, for more information and instructions.

3.5.2 Aquatics

namedLocation (unique ID given to the location within a site) → **siteID** (ID of NEON site) → **domainID** (ID of a NEON domain).

The spatial data can be found in the following Data Products:

- Surface water samples: Surface water microbe cell count (DP1.20138), in the table ***amc_fieldSuperParent*** and ***amc_fieldGenetic***.
- Benthic samples: Benthic microbe marker gene sequences (DP1.20280), in the table ***amb_fieldParent***.

3.6 Temporal Resolution and Extent

The finest resolution at which temporal data are reported is the **collectDate**, the date and time of day when the sample was collected in the field.

The NEON Data Portal provides data in monthly files for query and download efficiency. Queries including any part of a month will return data from the entire month. Code to combine files across months is available here: https://github.com/NEONScience/NEON-utilities.

3.7 Associated Data Streams

This section describes the data products that are directly linked or closely related to the microbe community composition data products.

3.7.1 Soils

Soil data are derived from subsamples collected during soil biogeochemical and microbial sampling and include numerous related data products:

- Soil physical and chemical properties, periodic (DP1.10086.001) - This data product bundle includes field data, soil moisture and pH, laboratory measurements of soil carbon and nitrogen concentrations (DP1.10078.001) and stable isotopes (DP1.10100.001), and inorganic nitrogen measurements derived by field incubations of soil (DP1.10080.001). Note that not all measurements are made on every corresponding sample measured for community composition, and vice-versa. Data from each soil phyiscal and chemical properties table can be joined to the linking table, ***sls_soilCoreCollection***, by the **sampleID**. These data can then be joined to the community composition data by the **geneticSampleID**.

- Soil microbe marker gene sequences (DP1.10108.001): Microbial 16S and ITS sequence data. The **dnaSampleID** variable can be used to link the data in the ***mmg_soilDnaExtraction***, ***mmg_soilPcrAmplification***, ***mmg_soilMarkerGeneSequencing***, and ***mmg_soilRawDataFiles*** tables to the soil community composition data.

- Soil microbe metagenome sequences (DP1.10107.001): Shotgun metagenomics sequences, typically derived from plot-level composited soil samples. The **genomicsPooledIDList** from the table ***sls_metagenomicsPooling*** contains the parent **sampleID**'s that link to the table ***sls_soilCoreCollection***. These data can be joined to the community composition data by the **geneticSampleID**.

- Soil microbe group abundances (DP1.10109.001): Bacteria/archaeal and fungal abundances as measured by quantitative PCR (qPCR). The **dnaSampleID** variable in the table ***mga_soilGroupAbundances*** may be used to link data in this product to the community composition data.

- Soil microbe biomass (DP1.10104.001): Microbial biomass as measured by phospholipid fatty acid analysis. The **biomassID** links to the Soil physical and chemical properties data product (DP1.10086.001, table ***sls_soilCoreCollection***), which links to the community composition data product by the **geneticSampleID**.

3.7.2 Aquatics

Aquatic microbial community composition data are derived from samples collected in conjunction with other physical, chemical, and biological measurements. These include:

- Surface water microbe field data: Found in the Aquatic Cell Counts data product (DP1.20138.001). The variable **geneticSampleID** in the table ***amc_fieldCellCounts*** can be used to link these data to this data product.

- Benthic microbe field data: Found in the Benthic microbe marker gene sequences data product (DP1.20280.001) and the metagenomics sequences (DP1.20279.001) data products. The variable **geneticSampleID** in the table ***amb_fieldParent*** can be used to link these data.

- Benthic microbe 16S and ITS marker gene sequences (DP1.20280.001). The **dnaSampleID** variable in the tables ***mmg_benthicDnaExtraction***, ***mmg_benthicPcrAmplification***, ***mmg_benthicMarkerGeneSequencing***, and ***mmg_benthicRawDataFiles*** can be used to link data in this product to community composition data.

- Surface water microbe 16S and ITS marker gene sequences (DP1.20282.001). The **dnaSampleID** variable in the tables ***mmg_swDnaExtraction***, ***mmg_swPcrAmplification***, ***mmg_swMarkerGeneSequencing***, and ***mmg_swRawDataFiles*** can be used to link data in this product to community composition data.

- Chemical properties of surface water (DP1.20093.001): Measurements of chemical constituents in water. Link to the surface water community composition data using the **parentSampleID** in the table ***swc_fieldSuperParent***.

- Periphyton, seston and phytoplankton collection (DP1.20166.001): Field data associated with sample collection. These data are linked to the benthic community composition data by the field data table ***amb_fieldParent***, which is part of the Benthic microbe marker gene sequences data product (DP1.20280.001). The **sampleID** in this table links to the **parentSampleID** in the table ***alg_domainLabChemistry***.

- Periphyton, seston and phytoplankton chemical properties (DP1.20163.001): Measurements of chemical constituents in algal samples. The field **parentSampleID** in the table ***alg_domainLabChemistry*** links to the **sampleID** in the table ***amb_fieldParent***, which can then be linked to the benthic community composition data product by the **geneticSampleID**.

- Benthic microbe group abundances (DP1.20277.001): Bacteria/archaeal and fungal abundances as measured by qPCR. The **geneticSampleID** in the table ***mga_benthicGroupAbundances*** can be used to link to the community composition data.

- Surface water microbe group abundances (DP1.20278.001): Bacteria/archaeal and fungal abundances as measured by qPCR. The **geneticSampleID** in the table ***mga_swGroupAbundances*** can be used to link to the community composition data.

3.8 Product Instances

For soil samples, up to 10 plots will be sampled at a subset of NEON terrestrial sites one to three times per year. During most years, the surface soil horizon (organic or mineral) will be collected, while once every 5 years during a coordinated microbes/biogeochemistry bout, up to 2 soil horizons will be collected as separate samples. For each soil horizon sampled, 3 unique locations are collected at each plot, for up to 6 samples per plot. Thus, there will be 0 records at sites that are not sampled and 30-120 unique records generated per site per year at sampled sites.

Aquatic samples are collected at all aquatic NEON sites. For surface water sampling, wadeable stream sites produce one sample up to 12 times per year, for a maximum of 12 product instances per site per year. Rivers produce up to 2 samples 6 times per year, for a maximum of 12 product instances per site per year. Flow-through lakes produce up to 5 samples 6 times per year, for a maximum of 30 product instances per site per year. Seepage lakes produce up to 2 samples 6 times per year, for a maximum of 12 product instances per site per year. Benthic microbial sampling occurs only at wadeable stream sites, where up to 8 samples are collected 3 times per year, for a maximum of 24 product instances per site per year.

3.9 Data Relationships

Data downloaded from the NEON Data Portal are provided in separate data files for each site and month requested. The neonUtilities R package contains functions to merge these files across sites and months into a single file for each table described above. The neonUtilities package is available from the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/web/packages/neonUtilities/index.html) and can be installed using the install.packages() function in R. For instructions on using neonUtilities to merge NEON data files, see the Download and Explore NEON Data tutorial on the NEON website: https://www.neonscience.org/download-explore-neon-data

Duplicates and/or missing data may exist where protocol and/or data entry abberations have occurred; *users should check data carefully for anomalies before joining tables*.

3.9.1 Soils

TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling dictates that each X,Y soil sampling location yields a unique sampleID per horizon per collectDate (day of year, local time) in the table ***sls_soilCoreCollection*** for the data product Soil physical and chemical properties, periodic (DP1.10086.001). A subset of bouts will sample for microbe community composition analysis: records in these bouts will have a **geneticSampleID**. There is a 1:1 relationship between the **geneticSampleID** and the **sampleID**, which is used throughout the suite of soil physical and chemical properties data products (DP1.10086). A record from ***sls_soilCoreCollection*** may have zero or one child records in tables ***mcc_soilSequenceVariantMetadata_16S*** and ***mcc_soilSequenceVariantMetadata_ITS*** of this data product. Additionally, each sample will be associated with 0-2 records in the table ***mcc_taxonTableLabSummary***, which describes the methods used to generate the taxon tables.

Each **geneticSampleID** is sent for DNA extraction. The DNA extraction laboratory data appear in the Soil Microbe Marker Gene Sequences (DP1.10108.001) data product, in table ***mmg_soilDnaExtraction***, and are linked by the **geneticSampleID/geneticSampleCode**. There are one or more **dnaSampleIDs/dnaSampleCode**s expected per **geneticSampleID**. Duplicate records for an individual **dnaSampleID/dnaSampleCode** should not exist, however multiple DNA extractions may occur. These DNA extracts will appear as separate records, and individual DNA extracts are tracked throughout the entire sequencing and analysis workflow by the **dnaSampleID** and/or **dnaSampleCode**.

The tables ***mcc_soilSequenceVariantMetadata_16S*** and ***mcc_soilSequenceVariantMetadata_ITS*** are linked by the **dnaSampleID/dnaSampleCode**. These tables include the per-sample metadata for each taxon table and target gene (16S for bacteria and archaea, ITS for fungi). Per-sample taxon tables are provided as downloadable .csv files using the link in the metadata table or are included in the expanded download package.

Soil Physical and Chemical Properties, periodic (NEON DP1.10086)

sls_soilCoreCollection.csv - > One record expected per **sampleID**. Generates samples used in Soil microbe marker gene sequences (DP1.10108.001), Soil microbe community composition (DP1.10081.001), Soil microbe group abundances (DP1.10109.001, discontinued), and Soil microbe biomass (DP1.10104.001). Additionally, subsamples generated from soil sampleIDs are used in Soil Microbe Metagenome Sequences (DP1.10107.001), Soil inorganic nitrogen pools and transformations (DP1.10080), Soil chemical properties (DP1.10078.001) and Soil stable isotopes (DP1.10100.001).

Soil Microbe Marker Gene Sequences (DP1.10108.001)

mmg_soilDnaExtraction.csv -> One record expected per **dnaSampleID**. A geneticSampleID will represent one sample per plot/horizon/X,Y-coordinate combination and per collectDate (day of year, local time). Generally there will be only one DNA extraction per **geneticSampleID** but in some cases multiple extractions will be necessary.

Important Note: The DNA extraction table is generic: samples that may not be relevant to this data product may appear in the data table. To limit the DNA extraction dataset to those that are relevant to soil community composition, filter the records in the ***mmg_soilDnaExtraction*** table to include only those with a **dnaSampleID** that is also contained in the target taxon metadata table (mcc_soilSequenceVariantMetadata_16S, mcc_soilSequenceVariantMetadata_ITS).

Soil Microbe Community Composition (DP1.10081.001)

mcc_soilSequenceVariantMetadata_16S.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

mcc_soilSequenceVariantMetadata_ITS.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

mcc_taxonTableLabSummary.csv -> This table describes the laboratory methods used for tabulating the taxon tables, with each unique set of methods (**testMethod**) corresponding to a new record. Each record in this table will correspond with many records in *mcc_soilSequenceVariantMetadata_16S.csv* and *mcc_soilSequenceVariantMetadata_ITS.csv*, and can be linked by the **testMethod**.

mcc_soilTaxonTableMetadata_16S.csv -> Deprecated table. Taxa represent unique operational taxonomic units (OTUs) based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

mcc_soilTaxonTableMetadata_ITS.csv -> Deprecated table. Taxa represent unique OTUs based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

3.9.2 Aquatics

3.9.2.1 Surface water The protocol dictates that each namedLocation sampled yields a unique **parentSampleID**, one sample per collectDate (day of year, local time) in the table ***amc_fieldSuperParent*** for the data product Surface Water Microbe Marker Gene Sequences (DP1.20282.001). Each **parentSampleID** may be subsampled into one **geneticSampleID** that is used for microbial analyses, and an archive sample, described in the table ***amc_fieldGenetic***. The **geneticSampleID**s are sent for DNA extraction and correspond to the the **geneticSampleID** in table ***mmg_swDnaExtraction***. One **dnaSampleID** is expected per **geneticSampleID**, although more than one may exist if multiple DNA extractions occur on a sample. Duplicate records for an individual **dnaSampleID** should not exist.

A record from the field data may have zero or one child records in tables ***mcc_swSequenceVariantMetadata_16S*** and ***mcc_swSequenceVariantMetadata_ITS*** of this data product. Additionally, each sample will be associated with 0-2 records in the table ***mcc_taxonTableLabSummary***, which describes the methods used to construct the taxon tables.

The tables ***mcc_swSequenceVariantMetadata_16S*** and ***mcc_swSequenceVariantMetadata_ITS*** are linked by the **dnaSampleID**. These tables include the per-sample metadata for each taxon table and target gene (16S for bacteria and archaea, ITS for fungi). Per-sample taxon tables are provided as downloadable .csv files using the link in the metadata table or are included in the expanded download package.

Surface Water Microbe Marker Gene Sequences (DP1.20282.001)

amc_fieldSuperParent.csv -> Field data associated with a surface water genetic sample. One record expected per namedLocation sampled and per collectDate, generates a unique **parentSampleID**.

amc_fieldGenetic.csv -> One record expected per namedLocation and per collectDate. Record represents a subsample (geneticSampleID) of the field-collected samples (parentSampleID). Depending on the time of year, each record generates zero or one **geneticSampleID**s, corresponding to the **geneticSampleID** in the table ***mmg_swDnaExtraction***.

mmg_swDnaExtraction.csv -> One record expected per **dnaSampleID**. Generally there will be only one DNA extraction per **geneticSampleID** but in some cases multiple extractions will be necessary. Duplicate records for an individual **dnaSampleID** should not exist.

Surface Water Microbe Community Composition (DP1.20141.001)

mcc_swSequenceVariantMetadata_16S.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

mcc_swSequenceVariantMetadata_ITS.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

mcc_taxonTableLabSummary.csv -> This table describes the laboratory methods used for tabulating the taxon tables, with each unique set of methods (**testMethod**) corresponding to a new record. Each record in this table will correspond with many records in *mcc_swSequenceVariantMetadata_16S.csv* and *mcc_swSequenceVariantMetadata_ITS.csv*, and can be linked by the **codeVersion**.

mcc_swTaxonTableMetadata_16S.csv -> Deprecated table. Taxa represent unique OTUs based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

mcc_swTaxonTableMetadata_ITS.csv -> Deprecated table. Taxa represent unique OTUs based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

3.9.2.2 Benthic habitats The AOS Protocol and Procedure for Aquatic Microbial Sampling dictates that each namedLocation sampled yields a unique **sampleID**, one sample per collectDate (day of year, local time) in the table ***amb_fieldParent*** in the data product Benthic Microbe Marker Gene Sequences (DP1.20280.001). Each **sampleID** may be subsampled into one **geneticSampleID** that is used for microbial analyses, and an archive sample, described in the same table. The **geneticSampleID**s are sent for DNA extraction and correspond to the **geneticSampleID** in ***mmg_benthicDnaExtraction***. One **dnaSampleID** is expected per **geneticSampleID**, although more than one may exist if multiple DNA extractions occur on a sample. Duplicate records for an individual **dnaSampleID** should not exist.

A record from the field data may have zero or one child records in tables ***mcc_benthicSequenceVariantMetadata_16S*** and ***mcc_benthicSequenceVariantMetadata_ITS*** of this data product. Additionally, each sample will be associated with 0-2 records in the table ***mcc_taxonTableLabSummary***, which describes the methods used to construct the taxon tables.

The tables ***mcc_benthicSequenceVariantMetadata_16S*** and ***mcc_benthicSequenceVariantMetadata_ITS*** are linked by the **dnaSampleID**. These tables include the per-sample metadata for each taxon table and target gene (16S for bacteria and archaea, ITS for fungi). Per-sample taxon tables are provided as downloadable .csv files using the link in the metadata table or are included in the expanded download package.

Benthic Microbe Marker Gene Sequences (DP1.20280.001)

amb_fieldParent.csv -> Field data associated with a benthic genetic sample. One record expected per namedLocation sampled and per collectDate generates a unique **sampleID**. Each record is a subsample (**geneticSampleID**) of the field-collected sample.

mmg_benthicDnaExtraction.csv -> DNA extraction data associated with a **geneticSampleID**. One record expected per namedLocation and per collectDate. Each unique DNA extraction from a **geneticSampleID** is given a **dnaSampleID**. Generally there will be only one DNA extraction per geneticSampleID but in some cases multiple extractions will be necessary.

Benthic Microbe Community Composition (DP1.20086.001)

mcc_benthicSequenceVariantMetadata_16S.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

mcc_benthicSequenceVariantMetadata_ITS.csv -> One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

mcc_taxonTableLabSummary.csv -> This table describes the laboratory methods used for tabulating the taxon tables, which each unique set of methods (**testMethod**) corresponding to a new record. Each record in this table will correspond with many records in *mcc_benthicSequenceVariantMetadata_16S.csv* and *mcc_benthicSequenceVariantMetadata_ITS.csv*, and can be linked by the **codeVersion**.

mcc_benthicTaxonTableMetadata_16S.csv -> Deprecated table. Taxa represent unique OTUs based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

mcc_benthicTaxonTableMetadata_ITS.csv -> Deprecated table. Taxa represent unique OTUs based on a representative sequence from a cluster of sequences that share at least 97% sequence similarity. One record is expected per **dnaSampleID**. Each record contains a **dnaSampleID** that corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

3.10 Special Considerations

3.10.1 Downloading Taxon Tables

Due to their large size, microbial taxon tables can be downloaded in two ways depending on whether the basic or expanded download package is selected.

- Basic Download: Taxon tables can be accessed from the web using the static URL links in the field **downloadFileUrl**, associated with each record in mcc_*taxonTableMetadata_16S|ITS.csv (* = soil, sw, or benthic). Pasting a URL into a web browser will automatically download the taxon table associated with a **dnaSampleID**.

- Expanded Download: Per-sample taxon tables for all records in the desired spatial and temporal range are downloaded as a .zip file along with the data publication tables. Please note that:

  a. Packaged files may be large and require longer to download.
  b. To facilitate parsing the downloaded files, taxon tables downloaded in the expanded package are re-named to the value specified in mcc_taxonTableMetata_16S|ITS **downloadFileName**.
  c. The **downloadFileUrl** field can be ignored.

For any download option, per-sample files should be merged prior to analysis.

4 DATA QUALITY

4.1 Data Entry Constraint and Validation

Constraints and data validation are implemented during the process of ingest into the NEON database. The product-specific data constraint and validation requirements built into data entry applications and database ingest are described in the document NEON Raw Data Validation for Microbial Community Composition (DP0.10081.001), provided with every download of this data product. Contained within this file is a field named ‘entryValidationRulesParser’, which describes syntactically the validation rules for each field built into the data ingest validation. Data entry constraints are described in Nicl syntax in the validation file provided with every data download, and the Nicl language is described in NEON’s Ingest Conversion Language (NICL) specifications (AD[16]).

Note: Data collected prior to 2017 were processed using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow.

4.2 Automated Data Processing Steps

4.2.1 Sequencing Data

Community composition data are derived from marker gene sequencing data, which must pass basic QA/QC checks, including meeting a minimum sequencing depth (e.g. number of sequences per sample), a maximum number of ambiguous base calls, and a minimum quality score. *Note:* The actual criteria may change over time as technology evolves and standards change. See the User Guides for marker gene sequencing data products (DP1.10108.001 for soil, DP1.20282.001 for surface water, DP1.20280.001 for benthic) for more details and documentation.

4.2.2 Community composition

After initial data QA/QC, sequence data are passed through a bioinformatics pipeline, where identical sequences are dereplicated and low-quality reads are filtered out. For 16S sequence data, the forward and reverse reads are merged for analysis, while for ITS sequence data only the forward read is used. Primer sequences are removed, and the sequence data that pass initial QA are denoised using the DADA2 pipeline (Callahan et al., 2016). This method denoises the sequence data using run-level sequence quality data, and produces a table containing all unique sequence reads, which are then taxonomically identified using the SILVA (for 16S data) or UNITE (for ITS data) reference database.

Following bioinformatics analysis, samples that do not meet the QC threshold for number of filtered reads receive a quality flag in the field **sequenceCountQF**, and samples with a sampleFilteredReadNumber=1 are removed.

Note: The specific method for taxonomic identificaion will change over time and is defined in the download package metadata as well as the supporting documentation (e.g. SOP’s and methods).

Following submission of metadata into the NEON automated data ingest process, the steps used to process the data through to publication on the NEON Data Portal are detailed in the NEON Algorithm Theoretical Basis Document: OS Generic Transitions (AD[15]).

4.3 Data Revision

All data are provisional until a numbered version is released. Annually, NEON releases a static version of all or almost all data products, annotated with digital object identifiers (DOIs). The first data Release was made in 2021. During the provisional period, QA/QC is an active process, as opposed to a discrete activity performed once, and records are updated on a rolling basis as a result of scheduled tests or feedback from data users. The Issue Log section of the data product landing page contains a history of major known errors and revisions.

4.4 Quality Flagging

The **dataQF** field in each data record is a quality flag for known errors applying to the record. The **sequenceCountQF** field is used to flag samples that do not have sufficient data for ecological analysis based on the NEON-defined minimum sampleFilteredReadDepth (currently 4000, although this could change over time). Please see the *Special Considerations* section of this document for a list of known errors that may be present in the data, and below for an explanation of quality flagging codes specific to this product.

Table 2: Descriptions of the dataQF codes for quality flagging

fieldName	value	definition
dataQF	legacyData	Data recorded using a paper-based workflow that did not implement the full suite of quality control features associated with the interactive digital workflow
sequenceCountQF	OK	Sample meets minimum requirement for number of sequences used for analysis
sequenceCountQF	Below threshold	Sample does not meet minimum requirement for number of sequences used for analysis

Records of land management activities, disturbances, and other incidents of ecological note that may have a potential impact are found in the Site Management and Event Reporting data product (DP1.10111.001)

4.5 Analytical Facility Data Quality

Bioinformatics analyses conducted on sequencing data conform to the current data quality standards used by practitioners. The data table ***mcc_taxonTableLabSummary*** (available in the expanded package) contains the bioinformatics methods and links to analysis code used to generate data, with each record describing the long-term methods and parameters used during the specified period of time. This table also provides the file names for the analysis code used by the lab to run the bioinformatics methods, which is not standardized for a particular computer language.

5 REFERENCES

1. Armougom F., and R. Didier. 2009. Exploring microbial diversity using 16S rRNA high-throughput methods. Journal of Computer Science and Systems Biology 2:74–92. https://doi.org/10.4172/jcsb.1000019.

2. Callahan, B.J., P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, and S.P. Holmes. 2016. DADA2: High Resolution Sample Inference from Illumina Amplicon Data. Nature Methods 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

3. Field, D., L. Amaral-Zettler, G. Cochrane, J.R. Cole, P. Dawyndt, G.M. Garrity, et al. 2011. The Genomic Standards Consortium: Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. PLoS Biol 9:e1001088.

4. Klindworth A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, and F. O. Glöckner. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research 41:e1–e1.

5. Kottmann, R., T. Gray, S. Murphy, L. Kagan, S. Kravitz, T. Lombardot, et al. 2008. A standard MIGS/MIMS compliant XML schema: Toward the development of the Genomic Contextual Data Markup Language (GCDML). OMICS: A Journal of Integrative Biology 12: 115–21.

6. Meyer F., D. Paarmann, M. D'Souza, R. Olson, E.M. Glass, M. Kubal, T. Paczian, et al. 2008. The Metagenomics RAST Server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

7. Yilmaz, P., R. Kottmann, D. Field, R. Knight, J.R. Cole, L. Amaral-Zettler, et al. 2011. Minimum information about a marker gene sequence (MIMARKS)and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol 29:415-420.