NEON USER GUIDE TO MICROBE MARKER GENE SEQUENCES (NEON.DP1.10108; NEON.DP1.20280; NEON.DP1.20282)

PREPARED BY	ORGANIZATION	DATE
Lee Stanish	TOS	03/02/2018
Stephanie Parker	AOS	03/02/2018

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	11/22/2017	Initial 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 Microbe Marker Gene Sequences, and associated metadata, from input data on aquatic and terrestrial samples. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the files NEON Data Variables for Soil Microbe Marker Gene Sequences (NEON.DP1.10108) (AD[05]), NEON Data Variables for Benthic Microbe Marker Gene Sequences (NEON.DP1.20280) (AD[06]), and NEON Data Variables for Surface Water Microbe Marker Gene Sequences (NEON.DP1.20282) (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; with 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 Microbe Marker Gene Sequences (NEON.DP0.10108) (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 Level 1, Level 2 and Level 3 Data Products Catalog
AD[04]	NEON.DP0.10108.001 _dataValidation.csv	NEON Raw Data Validation for Microbe Marker Gene Sequences (NEON.DP0.10108)
AD[05]	NEON.DP1.10109.001 _variables.csv	NEON Data Variables for Soil Microbe Marker Gene Sequences (NEON.DP1.10108)
AD[06]	NEON.DP1.20277.001 _variables.csv	NEON Data Variables for Benthic Microbe Marker Gene Sequences (NEON.DP1.20280)
AD[07]	NEON.DP1.20278.001 _variables.csv	NEON Data Variables for Surface Water Microbe Marker Gene Sequences (NEON.DP1.20282)
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.003044	AOS Protocol and Procedure: Aquatic Microbial Sampling
AD[12]	NEON.DOC.000913	TOS Science Design for Spatial Sampling
AD[13]	NEON.DOC.000008	NEON Acronym List
AD[14]	NEON.DOC.000243	NEON Glossary of Terms
AD[15]	OS_Generic_Transitions .pdf	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[16]		NEON's Ingest Conversion Language (NICL) specifications

2.2 Acronyms

Acronym	**Definition**
16S	Small subunit of the ribosomal RNA gene
ITS	Intergenic spacer region of the ribosomal RNA cistron
qPCR	Quantitative Polymerase Chain Reaction

3 DATA PRODUCT DESCRIPTION

The Microbe Marker Gene Sequences data products provide DNA sequence data and metadata of bacteria, archaea, and fungi in soil and aquatic samples. These data are used for taxonomic identification of microbial taxa. NEON targets a region of the 16S ribosomal RNA gene to measure bacteria and archaea, and the internally-transcribed spacer (ITS) region of the ribosomal RNA gene 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). These data are used to generate taxon tables for the downstream data products for Microbial Community Composition (NEON.DP1.10081, NEON.DP1.20086, NEON.DP1.20141). The sample plan implements the guidelines and requirements in the Science Designs for TOS Terrestrial Microbial Diversity (AD[08]) and Aquatic Sampling (AD[09]). Information on sample collection methods such as frequencies per sample type can be found in the field user guides for each data product:

- Soils: NEON User Guide to Soil Physical Properties, Distributed Periodic (NEON.DP1.10086)
- Surface water: NEON User Guide for Surface Water Microbe Cell Count (NEON.DP1.20138)
- Benthic habitats: NEON User Guide for Aquatic Benthic Microbe Collection (NEON.DP0.20270)

Sample collection methods differ between aquatic and terrestrial samples, but 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. For most samples, including soil and epipsammon, native material is processed for analysis; however, certain aquatic sample types have additional processing steps (Figure 1). Samples are shipped to an analytical laboratory where sample processing, DNA extraction, sequencing library preparation and DNA sequencing occur.

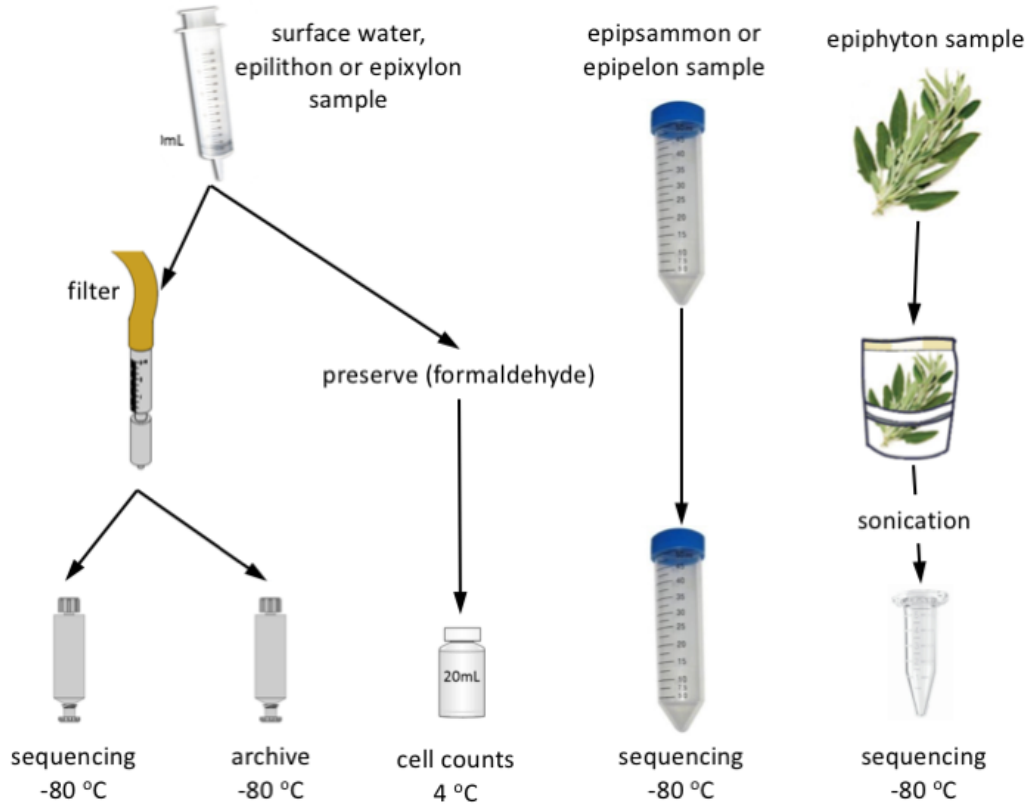

Figure 1: Overview of microbial field sample types, processing steps, and analyses. Note that samples destined for cell count analyses are part of a different data product, NEON.DP1.20138.

3.1 Spatial Sampling Design

Sampling for microbe marker gene sequence 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).

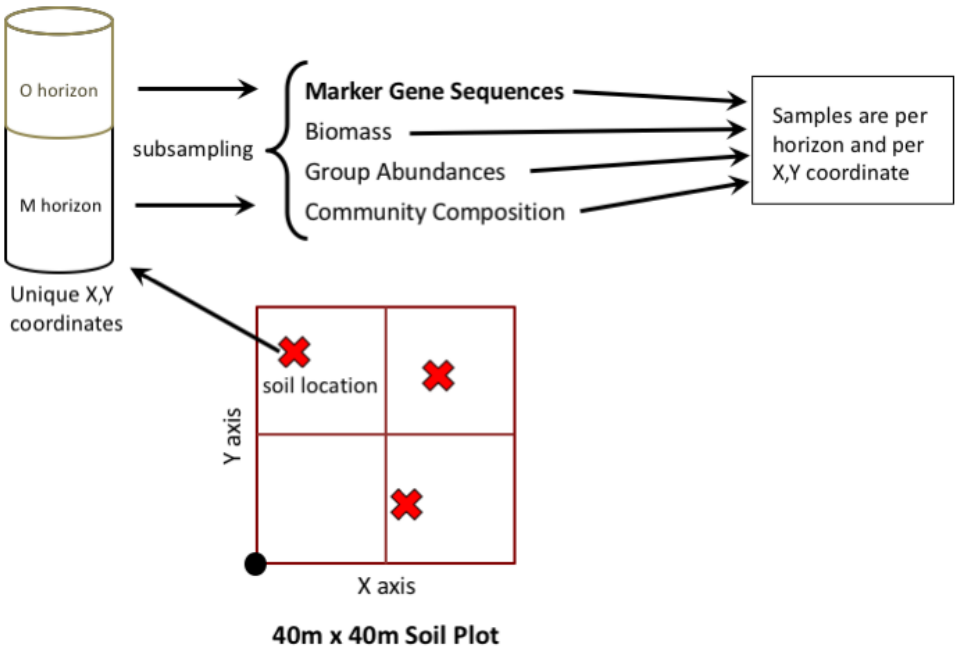

Figure 2: Overview of soil microbial field sampling, spatial design, and analysis workflow.

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 microbe marker genes data products are described in more detail in the Data Product User Guides for Soil Physical Properties (NEON.DP1.10086), Aquatic Surface Water Cell Counts (NEON.DP1.20138), and Aquatic Benthic Microbe Collection (NEON.DP0.20270). For a description of the methods used in terrestrial plot selection, refer to the TOS Science Design for Spatial Sampling (AD[02]).

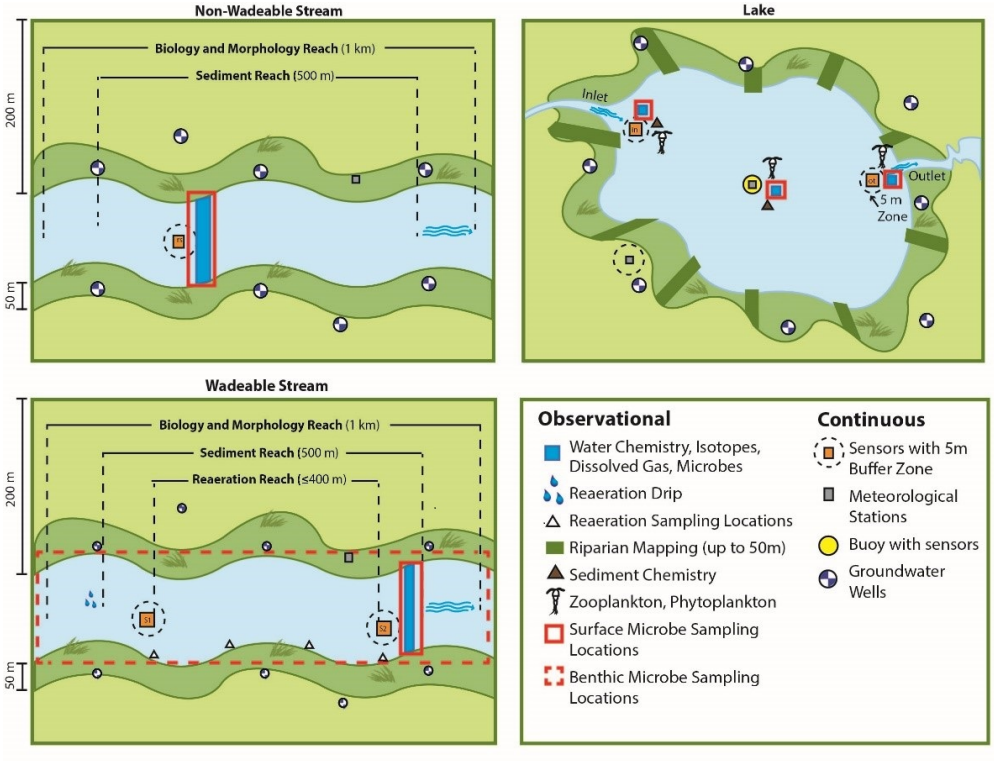

Figure 3: Generic NEON aquatic site layouts with microbial sampling locations highlighted in red boxes.

3.2 Temporal Sampling Design

At most terrestrial sites, soil sampling for marker gene sequence analysis occurs 3 times per year in conjunction with the soil physical properties data product (DP1.10086). Two sampling bouts occur during periods of seasonal transitions (e.g. winter-spring or wet-dry), and one during the period of peak greenness (as measured by remote sensing data). At sites with short growing seasons (e.g. tundra and taiga), sampling occurs once annually during peak greenness.

Once every five years, a 'coordinated' bout occurs in which additional biogeochemical and isotopic measurements are made (DP1.10078), along with measurements of microbe biomass (DP1.10104) and nitrogen transformation rates (DP1.10080). During a coordinated bout, up to 2 soil horizons (organic and mineral) are sampled for microbial analyses to a maximum depth of 30 cm.

Surface water samples are collected monthly in wadeable streams, and every other month in lakes and rivers in conjunction with surface water chemistry sampling. Benthic microbe samples are collected three times per year, roughly spring, summer, and autumn at the same time as algal periphyton samples.

For all samples, the temporal resolution is that of a single collection date. For a comprehensive description of 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 for soil (NEON.DP1.10086), aquatic surface water (NEON.DP1.20138) and

benthic (NEON.DP0.20270) sampling can be found in the Data Product User Guides for those respective Data Products.

3.3 Variables Reported

All variables reported from the field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Microbe Marker Gene Sequences (NEON.DP0.10108) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the following files:

- NEON Data Variables for Soil Microbe Marker Gene Sequences (NEON.DP1.10108) (AD[05]).
- NEON Data Variables for Benthic Microbe Marker Gene Sequences (NEON.DP1.20280) (AD[06]).
- NEON Data Variables for Surface Water Microbe Marker Gene Sequences (NEON.DP1.20282) (AD[07]).

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), the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014), where applicable.

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 Gravitational Model 96 (EGM96) for its reference gravitational 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.

3.4 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.

3.4.1 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 accuracury of $\pm$ 20 m. However, 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.

3.4.2 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 respective marker gene sequences data product download in the following tables:

- Surface water samples: Field data for the parent sample of surface water microbes, table ***amc_fieldSuperParent*** and ***amc_fieldGenetic***.
- Benthic samples: Aquatic benthic microbes field data, table ***amb_fieldParent***.

3.5 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 stack files across months is available here: https://github.com/NEONScience/NEON-utilities

3.6 Associated Data Streams

This section describes the data products that are directly linked or closely related to the microbe marker gene sequences data products.

3.6.1 Soils

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

- Soil Physical Properties, distributed periodic (DP1.10086) - includes all field data associated with a soil sample. These data are linked to the marker genes data by the **geneticSampleID** in the table ***sls_soilCoreCollection***.
- Soil microbe community composition (NEON.DP1.10081) - Microbial community composition data derived from marker gene sequencing. The **dnaSampleID** variable may be used to link data in this product to soil microbe marker genes data.
- Soil microbe group abundances (NEON.DP1.10109): Bacterial/archaeal and fungal abundances as measured by qPCR. The **dnaSampleID** variable in the table ***mga_soilGroupAbundances*** can be used to link data in this product to the soil microbe marker gene sequences data.

- Soil microbe biomass (NEON.DP1.10104) - Microbial biomass as measured by PLFA. Use information in the Soil Physical Properties data product (NEON.DP1.10086, table ***sls_soilCoreCollection***) to obtain the **biomassID** corresponding to the **sampleID**. The **sampleID** will map to a corresponding **geneticSampleID**, which can then be used to link data in the two data products.
- Soil inorganic nitrogen pools and transformations (NEON.DP1.10080) - Measurements derived by field incubations of soil cores or buried bags. As described for soil microbe biomass, use the **sampleID** from table ***sls_soilCoreCollection*** to link these data products.
- Soil chemical properties (Distributed periodic) (NEON.DP1.10078) - Measurements of soil carbon and nitrogen. As with soil microbe biomass, the corresponding **sampleID** can be used to link data.
- Soil stable isotopes (Distributed periodic) (NEON.DP1.10100) - Measurements of soil carbon and nitrogen stable isotopes. As with soil microbe biomass, the corresponding **sampleID** can be used to link data.

3.6.2 Aquatics

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

- Surface water microbes field data: included in the download package for this data product (Surface water microbe marker gene sequences). The field **geneticSampleID** within the table ***amc_fieldGenetic*** can be used to link these data products.
- Benthic microbes field data: included in the download package for this data product (Benthic microbe marker gene sequences), and can be linked by the **geneticSampleID**.
- Benthic (NEON.DP1.20086) and surface water (NEON.DP1.20141) microbe community composition: Taxonomic data derived from the Microbe marker gene sequencing data products described in this User Guide. The field **dnaSampleID** can be used to link these data to this data product.
- Surface water microbe cell count (NEON.DP1.20138) - Measurements of the abundances of microbiota in preserved surface water samples. The field **cellCountSampleID** in the table ***amc_cellCounts*** = **cellCountSampleID** in ***amb_fieldParent*** and can be used to link these data products.
- Chemical properties of surface water (NEON.DP1.20093) - Measurements of chemical constituents in water. The field **parentSampleID** in the table ***swc_fieldSuperParent*** = **parentSampleID** in tables ***amc_fieldSuperParent*** and ***amc_fieldGenetic*** and can be used to link these data products.
- Periphyton, seston and phytoplankton collection (NEON.DP1.20166) - Field data associated with sample collection. The field **parentSampleID** in the table ***alg_fieldData*** links to the **sampleID** in the table ***amb_fieldParent***, which can then be linked to this data product by the **geneticSampleID**.
- Periphyton, seston and phytoplankton chemical properties (NEON.DP1.20163): Measurements of chemical constituents of 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 this data product by the **geneticSampleID**.

3.7 Product Instances

For soil samples, a maximum of 10 plots will be sampled at every site one to three times per year. Most years, the surface soil horizon (organic or mineral) will be collected, while once every 5 years during a coordinated mi-

crobes/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 30-120 product instances generated per site per year.

Aquatic samples are collected at all aquatic NEON sites. For surface water sampling, wadeable streams 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. Lakes produce up to 4 samples 6 times per year, for a maximum of 24 product instances collected 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.

3.8 Data Relationships

3.8.1 Soils

The protocol dictates that each X,Y location sampled yields a unique **sampleID** per horizon per collectDate (day of year, local time) in the table ***sls_soilCoreCollection*** for the data product Soil Physical Properties, distributed periodic (NEON.DP1.10086). Every bout type that includes microbes (e.g. the variable **boutType** includes the string 'microbe') should sample for marker gene sequence analysis. A record from ***sls_soilCoreCollection*** may have zero or one child records in tables ***mmg_soilMarkerGeneSequencing_16S*** and ***mmg_soilMarkerGeneSequencing_ITS*** of this data product.

Each **geneticSampleID** is a subsample of the parent **sampleID** in the table ***sls_soilCoreCollection***, and is sent for DNA extraction. The DNA extraction laboratory data appear in the table ***mmg_soilDnaExtraction***, and are linked by the **geneticSampleID**. There are one or more **dnaSampleID**s expected per **geneticSampleID**, depending on the number of DNA extractions that occur on a sample. Duplicate records for an individual **dnaSampleID** should not exist.

One record in tables ***mmg_soilPcrAmplification_16S*** and ***mmg_soilPcrAmplification_ITS*** is expected per **dnaSampleID**. This table includes the PCR amplification processing metadata for each sample.

Note that only metadata are available on the NEON data portal. Actual sequence data are available on external public sequence repositories (see Special Considerations section below on how to access).

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*.

Soil Physical Properties (NEON DP1.10086)

sls_soilCoreCollection.csv - > One record expected per **sampleID**. Generates samples used in Soil microbe marker gene sequences (NEON.DP1.10108), Soil microbe community composition (NEON.DP1.10081), Soil microbe group abundances (NEON.DP1.10109), and Soil microbe biomass (NEON.DP1.10104). Additionally, subsamples generated from soil sampleIDs are used in Soil inorganic nitrogen pools and transformations (NEON.DP1.10080).

Soil Microbe Marker Gene Sequences (NEON.DP1.10108)

mmg_soilDnaExtraction.csv - > One record expected per **dnaSampleID**. A geneticSampleID will represent one sample per plot/horizon/X,Ycoordinate 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 the marker genes samples, it may be helpful to filter the records in the ***mmg_soilDnaExtraction*** table to include only those with a value of 'marker gene' or 'marker gene and metagenomics' in the variable **sequenceAnalysisType**.

mmg_soilPcrAmplification_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

mmg_soilPcrAmplification_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_soilDnaExtraction*** table.

mmg_soilMarkerGeneSequencing_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables ***mmg_soilPcrAmplification_16S*** and ***mmg_soilDnaExtraction***.

mmg_soilMarkerGeneSequencing_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables ***mmg_soilPcrAmplification_ITS*** and ***mmg_soilDnaExtraction***.

3.8.2 Aquatics

3.8.2.1 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***. Each **parentSampleID** may be subsampled into one **geneticSampleID** that is used for microbial analyses, and an archive sample, described in the table ***amc_fieldGenetic***. These **geneticSampleID**s are sent for DNA extraction such that the **geneticSampleID** from ***amc_fieldGenetic*** = **geneticSampleID** in ***mmg_swDnaExtraction***.

Note that only metadata are available on the NEON data portal. Actual sequence data are available on external public sequence repositories (see Special Considerations section below on how to access).

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*.

Surface Water Microbe Marker Gene Sequences (NEON.DP1.20282)

amc_fieldSuperParent.csv -> One record expected per namedLocation sampled and collectDate (day of year, local time), generates a unique **parentSampleID**.

amc_fieldGenetic.csv -> One record expected per namedLocation per collectDate (day of year, local time). 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 variable **geneticSampleID** in the table ***mmg_swDnaExtraction***.

mmg_swDnaExtraction.csv - > One record expected per **dnaSampleID**. A geneticSampleID will represent one sample 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.

mmg_swPcrAmplification_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

mmg_swPcrAmplification_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_swDnaExtraction*** table.

mmg_swMarkerGeneSequencing_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables ***mmg_swPcrAmplification_16S*** and ***mmg_swDnaExtraction***.

mmg_swMarkerGeneSequencing_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables ***mmg_swPcrAmplification_ITS*** and ***mmg_swDnaExtraction***.

3.8.2.2 Benthic

The protocol dictates that each namedLocation sampled yields a unique **sampleID**, one sample per collectDate (day of year, local time) in Benthic microbe marker gene sequences (DP1.20280), in the table ***amb_fieldParent***. Each **sampleID** may be subsampled into one **geneticSampleID** that is used for microbial analyses, and an archive sample, described in the same table. These **geneticSampleID**s are sent for DNA extraction such that the **geneticSampleID** from ***amb_fieldParent*** = **geneticSampleID** in ***mmg_benthicDnaExtraction***.

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*.

Benthic Microbe Marker Gene Sequences (NEON.DP1.20280) amb_fieldParent.csv -> One record expected per namedLocation sampled and collectDate (day of year, local time), generates a unique **sampleID**. Record represents a subsample (**geneticSampleID**) of the field-collected sample.

mmg_benthicDnaExtraction.csv - > One record expected per **dnaSampleID**. A geneticSampleID will represent one sample 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.

mmg_benthicPcrAmplification_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

mmg_benthicPcrAmplification_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the ***mmg_benthicDnaExtraction*** table.

mmg_benthicMarkerGeneSequencing_16S.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables

mmg_benthicPcrAmplification_16S and ***mmg_benthicDnaExtraction***.

mmg_benthicMarkerGeneSequencing_ITS.csv - > One record is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the upstream tables ***mmg_benthicPcrAmplification_ITS*** and ***mmg_benthicDnaExtraction***.

3.9 Special Considerations

For ease of integration with external data sets, marker gene sequence data are published on public sequence repositories. The primary data repository is MG-RAST (http://metagenomics.anl.gov, Meyer et al., 2008), which directly synchronizes its data with the European Bioinformatics Institute (EMBL-EBI) database and, through EMBL-EBI, synchronizes with the National Center for Biotechnology Information's Sequence Read Archive (SRA). A suite of metadata, compliant with minimum metadata standards defined by the Genomics Standards Consortium (e.g. MIxS, MIMARKS), accompanies the sequence data. While efforts are made to publish comprehensive sequencing metadata with the sequence data stored at public sequence repositories, potentially important data will only be available through the NEON Data Portal. These data include:

- Methods and SOPs
- QA data
- Sample identifiers to enable joining marker gene sequencing data with other related Data Products, such as biogeochemistry data
- Data for other related Data Products

3.9.1 Retrieving Marker Gene Sequence Data

There are a number of ways to search and retrieve minimally processed marker gene sequence data.

- From the NEON data portal:
    1. Links beginning with "MG-RAST Project: NEON Soil Microbe Marker Gene Sequences" and ending with the upload date will redirect the user to a single MG-RAST project containing the data for the queried samples. This is a dynamic link and will automatically update based on the user query.
    2. Links beginning with "MG-RAST Project: NEON Freshwater Benthic Microbe Marker Genes" will take the user to the MG-RAST project page for the queried data. This is a dynamic link and will automatically update based on the user query.
    3. Links beginning with "MG-RAST Project: NEON Surface Water Microbe Marker Genes" will take the user to the MG-RAST project page for the queried data. This is a dynamic link and will automatically update based on the user query.
    4. The link "MG-RAST Sample Search" takes the user to the MG-RAST page for searching individual records, pre-populated with NEON records based on the user query.
- From MG-RAST directly: Users who are interested in using the MG-RAST data analysis pipeline may want to combine NEON datasets with other datasets. This may be more easily achieved by querying the MG-RAST database directly. Users can analyze samples from a variety of NEON and non-NEON projects. A free user account may be required.

- From SRA directly: Data and metadata are available for download from the SRA using the SRA toolkit. Documentation on how to install and use the toolkit for downloading sequence data is available on the SRA website.
- From EMBL-EBI: MG-RAST also synchronizes data sets with the European Bioinformatics Initiative Repository (EMBL-EBI, https://www.ebi.ac.uk/), which has a web and API interface for downloading data. The NEON soil marker gene sequence data can be found by querying the NCBI Project ID PRJNA393362.

Note: There may be lags between publication of metadata on the NEON data portal and availability of sequence data on the public sequence repository.

4 DATA QUALITY

4.1 Data Entry Constraint and Validation

Constraints 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 Microbe Marker Gene Sequences (NEON.DP0.10108), 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

4.3 Sequencing Data

Marker gene sequencing data are generated in batches of multiple samples. After sequencing, the multiplexed sequence data are parsed into separate files on a per sample basis. For each sample, minimum quality criteria must be met in order to accept the data for the sample. The general criteria include meeting a minimum sequencing depth (e.g. number of sequences per sample), a maximum number of ambiguous base calls, and a minimum quality score. The actual criteria may change over time as technology evolves and standards change. The per sample QA results are published as part of the expanded download package.

Following laboratory 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]).

4.4 Data Revision

All data are provisional until a numbered version is released; the first release of a static version of NEON data, annotated with a globally unique identifier, is planned to take place in 2020. 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 Change Log section of the data product readme, provided with every data download, contains a history of major known errors and revisions.

4.5 Quality Flagging

The **dataQF** field in each data record is a quality flag for known errors applying to the record. Please see the table below for an explanation of **dataQF** codes specific to this product.

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

4.6 Analytical Facility Data Quality

Data analyses conducted on marker gene sequencing data conform to the current data quality standards used by practitioners. Each metadata table includes a variable, called **qaqcStatus**, in which the laboratory can indicate sample processing issues. Any records with a qaqcStatus = “Fail” should also be accompanied by free-form notes in the “remarks” variable.

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. 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.
3. 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.
4. 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.

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.