NEON USER GUIDE TO MICROBE GROUP ABUNDANCES (DP1.10109.001; DP1.20277.001; DP1.20278.001)

PREPARED BY	ORGANIZATION
Lee Stanish	FSU
Stephanie Parker	AOS
Hugh Cross	AOS/TOS

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	11/8/2017	Initial Release
B	08/24/2020	Included general statement about usage of neonUtilities R package and statement about possible location changes. Section 3: Added notification on discontinued soil microbe abundances data product and updated method references; Section 3.3: Added Sampling Design Changes section and included changes to sampling frequency for group abundances analysis; Section 3.7.1: Updated description of Associated Data Streams for bundled Soil Physical and Chemical Properties data product. Section 4.5: Updated details on data QAQC procedures.
C	03/16/2022	Added section 4.2.2 to add primers used in qPCR

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 Group Abundances data and associated metadata measured on aquatic and terrestrial samples by broad taxonomic (e.g. bacterial, archaeal, and fungal) groups. 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 Group Abundances (DP1.10109.001) (AD[05]), NEON Data Variables for Benthic Microbe Group Abundances (DP1.20277.001) (AD[06]), and NEON Data Variables for Surface Water Microbe Group Abundances (DP1.20278.001) (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; 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 Group Abundances (DP0.10109.001) (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 Data 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

Acronym	Definition
qPCR	Quantitative Polymerase Chain Reaction

3 DATA PRODUCT DESCRIPTION

The Microbe Group Abundances data products provide semi-quantitative estimates of the abundances of bacteria, archaea and fungi in soil and aquatic samples. Group abundances are quantified using Quantitative Polymerase Chain Reaction (qPCR), a method of measuring the abundance of a target DNA fragment in a sample, which is used to approximate the abundance of the organisms containing that DNA fragment within a sample. NEON measures the abundance of a fragment of the 16S subunit of the rRNA gene operon to quantify the abundances of bacteria and archaea: the abundances of fungi are quantified based on the abundance of the internal transcribe spacer (ITS) region of the rRNA gene operon (Ginzinger 2002). 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 (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 sample processing and qPCR analysis occurs (Figure 1). For data generated prior to Jan 1, 2016, 3 separate primer sets were used to quantify bacterial, archaeal, and fungal abundances. For data generated after Jan 1, two primer sets were used: 1 primer set that amplifies both bacteria and archaea (Takahashi et al., 2014), and 1 set that amplifies fungi (Walters et al., 2016). For specific methods and primer sets used, refer to the ***mga_labSummary*** data table, included in this download package.

NOTICE. The Soil Microbe Group Abundances data product was discontinued in 2020, with the last field samples having associated qPCR data in 10-2018. Users who are interested in generating this data from NEON samples may submit a request for samples or DNA extracts from the NEON Biorepository. Alternatively, the Soil Microbe Biomass data product (DP1.10104) may be utilized as a similar measurement of the quantities of microbes in soils.

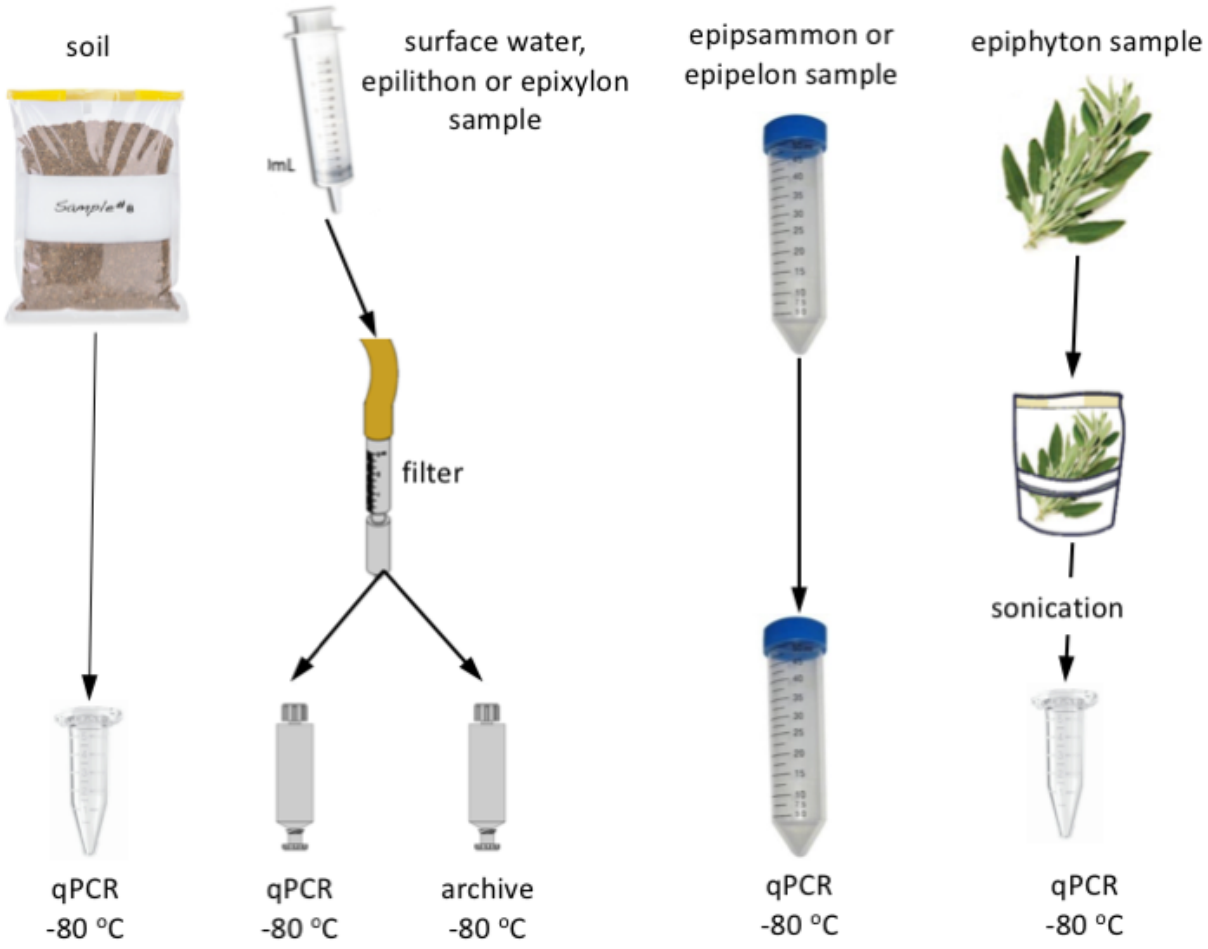

Figure 1: Overview of microbial field sample types, processing steps, and analyses.

3.1 Spatial Sampling Design

Sampling for microbial group abundance analysis is executed at all NEON sites and for all samples, data are reported at the resolution of a single sampling location.

For soils, this equates to a randomly-assigned X,Y coordinate (+/- 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).

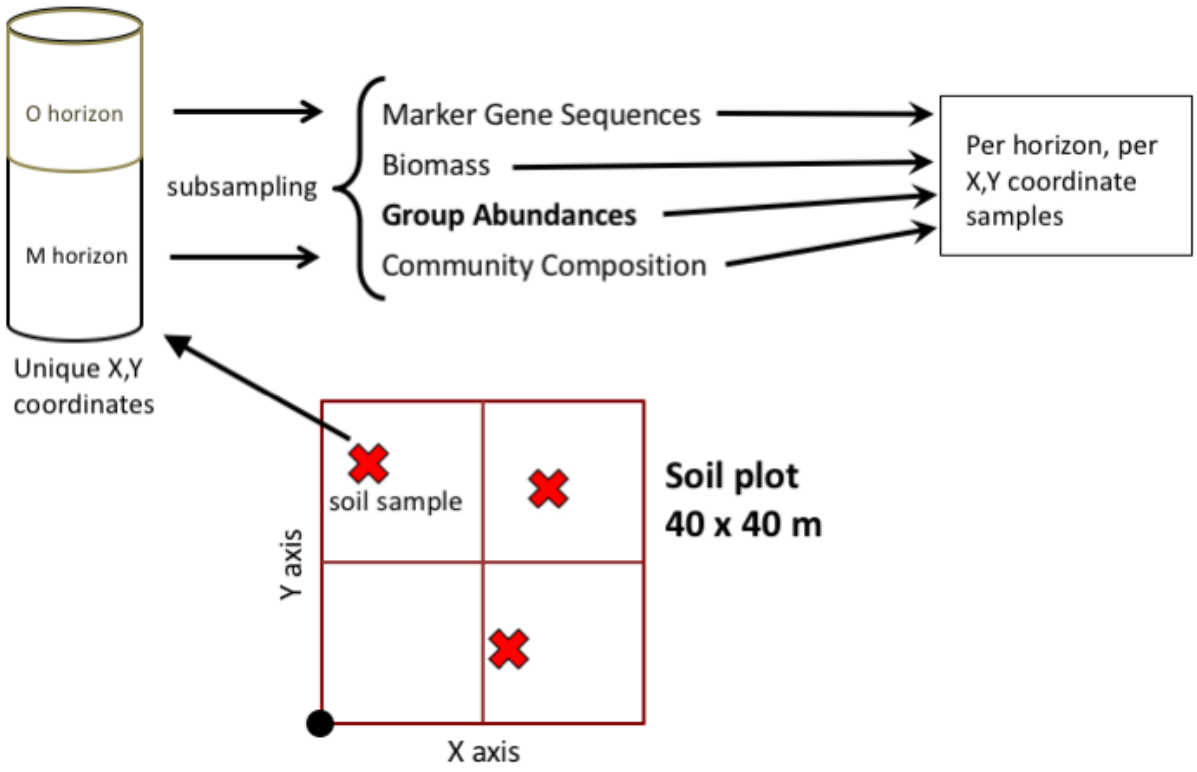

Figure 2: Overview of soil microbial field sampling and analysis workflow.

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 group abundances 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 Field Sampling (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]).

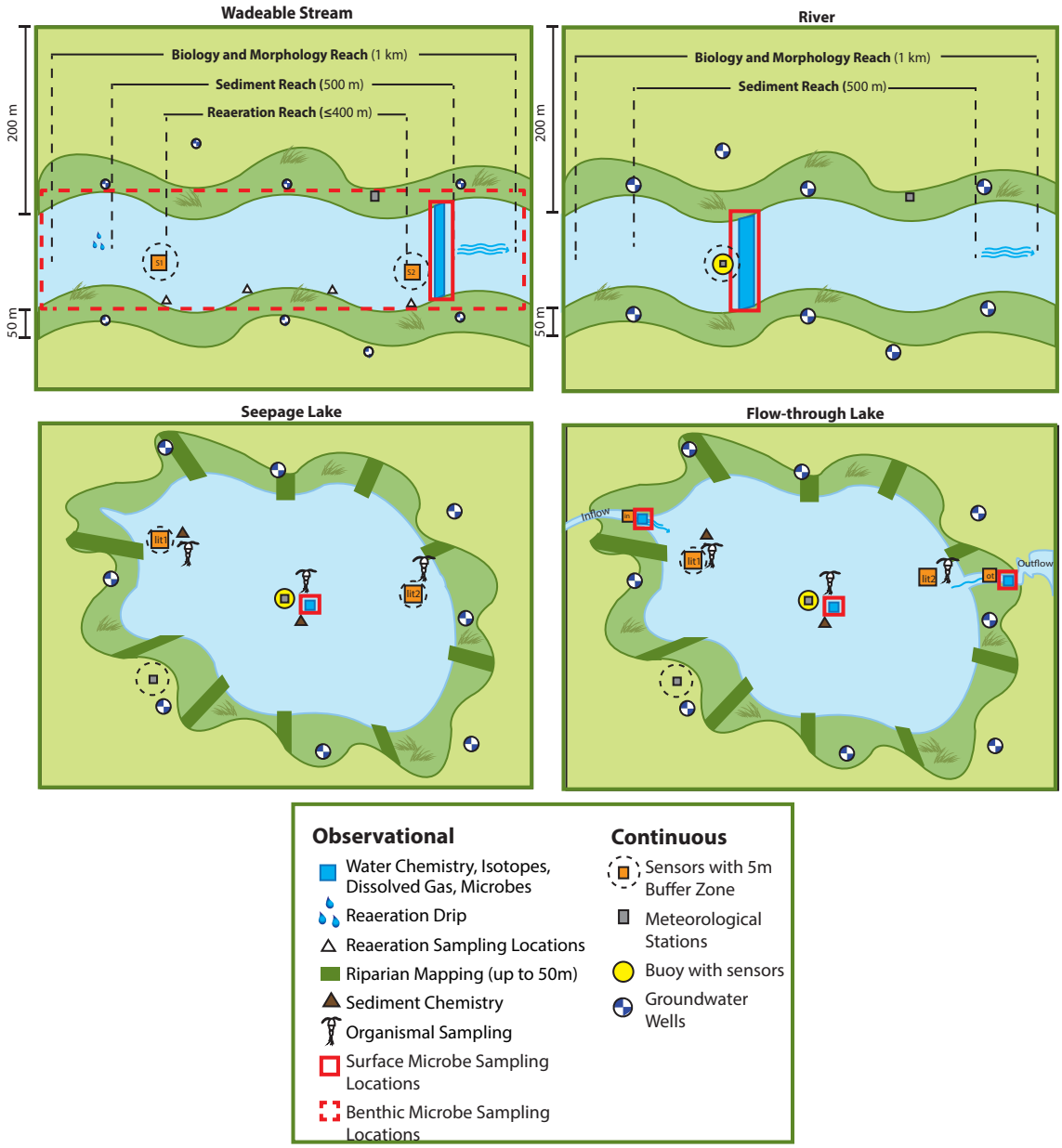

Figure 3: Generic NEON aquatic site layouts with microbial sampling locations highlighted in red boxes.

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 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 products for soil (DP1.10086), and aquatic surface water (DP1.20138) can be found in those respective Data Product User Guides.

3.2.1 Soils

The soil group abundance sample analysis was discontinued. Refer to the Sampling Design Changes Section for details on the previous temporal sampling designs.

3.2.2 Aquatics

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.

3.3 Sampling Design Changes

Over the course of early operations, the design for soil periodic 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 - Dec 2017: Soil samples were collected for microbial group abundances analysis during every bout and for all sites. Sampling occurred 3 times per year in conjunction with the soil physical and chemical properties data product (DP1.10086). Two sampling bouts occurred during periods of seasonal transitions (e.g. winter-spring or wet-dry), and once during the period of peak greenness (as measured by remote sensing data). At sites with short growing seasons (e.g. tundra and taiga), sampling occurred once annually during peak greenness. Once every five years, a 'coordinated' bout occurs in which additional biogeochemical and isotopic measurements are made, along with measurements of nitrogen transformation rates and microbe biomass (see bundled soil data product DP1.10086 and microbial biomass data product DP1.10104). During a 'coordinated' bout, up to 2 soil horizons were sampled for microbial group abundance analyses to a maximum depth of 30 cm.
- Dec 2017 - Oct 2018: Soil samples were collected for microbial group abundances analysis during every soil sampling bout at one site per domain and during all 'coordinated' bouts at any site.
- Oct 2018 - current: Soil microbe group abundances analysis discontinued at all sites.

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 field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Microbe Group Abundances (DP0.10109.001) (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 Group Abundances (DP1.10109.001) (AD[05]).
- NEON Data Variables for Benthic Microbe Group Abundances (DP1.20277.001) (AD[06]).
- NEON Data Variables for Surface Water Microbe Group Abundances (DP1.20278.001) (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. 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*, however, a more precise measurement may be desired 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 (DP1.20138.001), in the table **amc_fieldSuperParent**.
- Benthic samples: Benthic microbe marker gene sequences (DP1.20086.001), in the field data 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 stack 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 group abundances 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 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 group abundances, and vice-versa. Data from each table can be joined to the linking table, ***sls_soilCoreCollection***, by the **sampleID**. These data can then be joined to the table ***mmg_soilDnaExtraction***, which is part of the group abundances data product, by the **geneticSampleID**.
- Soil microbe community composition (DP1.10081.001) - Microbial community composition data derived from marker gene sequencing. The **dnaSampleID** variable in the tables ***mcc_soilTaxonTable_16S*** and ***mcc_soilTaxonTable_ITS*** may be used to link data in this product to soil microbe group abundances data.
- Soil microbe marker gene sequences (DP1.10108.001): Microbial 16S and ITS sequence data. The **dnaSampleID** variable in the tables ***mmg_soilDnaExtraction***, ***mmg_soilPcrAmplification*** and ***mmg_soilMarkerGeneSequencing*** can be used to link data in this product to the soil microbe group abundances data.

- Soil microbe biomass (DP1.10104.001) - Microbial biomass as measured by PLFA. Use information in the Soil physical and chemical properties, periodic data product (DP1.10086.001, table ***sls_soilCoreCollection***) to obtain the **biomassID**. The **biomassID** will map to a corresponding **geneticSampleID**, which can then be used to link data in the two data products.

3.7.2 Aquatics

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

- Surface water microbes field data are found in the Aquatic Cell Counts data product (DP1.20138.001). The field **geneticSampleID** within the table ***amc_fieldCellCounts*** can be used to link these data products.
- Benthic microbes field data are part of the download package for the Benthic microbe marker gene sequences data product (DP1.20280.001), and can be linked by the **geneticSampleID**.
- Chemical properties of surface water (DP1.20093.001) - Measurements of chemical constituents in water. The field **parentSampleID** in the table ***swc_fieldSuperParent*** can be used to link these data products.
- Periphyton, seston and phytoplankton collection (DP1.20166.001) - 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 (DP1.20163.001): 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**.
- Benthic (DP1.20086.001) and surface water (DP1.20141.001) microbe community composition: Taxonomic data derived from 16S and ITS marker gene sequencing. The field **dnaSampleID** in the tables ***mcc_benthicTaxonTable_16S***, ***mcc_benthicTaxonTable_ITS***, ***mcc_swTaxonTable_16S*** and ***mcc_swTaxonTable_ITS*** can be used to link these data to this data product.
- Benthic (DP1.20280.001) microbe marker gene sequence data. The field **geneticSampleID** in the tables ***amb_fieldParent*** and ***mmg_benthicDnaExtraction*** can be used to link these data to this data product.
- Surface water (DP1.20282.001) microbe marker gene sequences data. The field **geneticSampleID** in the tables ***mmg_swDnaExtraction*** can be used to link these data to this data product.

3.8 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 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 30-120 product instances generated per site per year.

Aquatic samples are collected at all aquatic NEON sites. For surface water sampling, a maximum of 4 sample locations will be sampled at every site 6-12 times per year, for a maximum of 24 product instances collected per site per year in a lake, and 12 product instances per site per year in a wadeable stream or

river. Benthic microbial sampling occurs only at wadeable stream sites, where up to 8 samples are collected three times per year, for a maximum of 24 product instances per site per year.

3.9 Data Relationships

Each **geneticSampleID** is a subsample of the parent sampleID in the relevant field collection data table (soil, benthic, surface water), and is sent for DNA extraction. The DNA extraction laboratory data appear in the table *mmg_(soil)(benthic)(sw)DnaExtraction*, and are linked by the **geneticSampleID**. One **dnaSampleID** is expected per geneticSampleID, although more may exist depending on the number of DNA extractions that occur on a sample. The combination of **dnaSampleID** and **targetTaxonGroup** represents an independent record in the **mga_groupAbundances** data and only one record should exist.

3.9.1 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 (DP1.10086.001). Every bout type that includes microbes (e.g. the variable **boutType** includes the string 'microbe') should sample for group abundance analysis. A record from *sls_soilCoreCollection* may have zero or one child records in table *mga_soilGroupAbundances* of this data product.

Soil Physical Properties (DP1.10086.001)

sls_soilCoreCollection.csv -> One record expected per **sampleID**. Generates samples used in Soil microbe group abundances (DP1.10109.001), Soil microbe marker gene sequences (DP1.10108.001), Soil microbe community composition (DP1.10081.001), and Soil microbe biomass (DP1.10104.001). Additionally, subsamples generated from soil sampleIDs are used in Soil inorganic nitrogen pools and transformations (DP1.10080.001). Each **geneticSampleID** is a subsample of the parent **sampleID** and is sent for DNA extraction.

Soil Microbe Marker Gene Sequences (DP1.10108.001)

mmg_soilDnaExtraction.csv -> This table contains the DNA extraction laboratory data. Data 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 provided to the lab. Duplicate records for an individual **dnaSampleID** should not exist. *Important Note*: This DNA extraction table is generic: samples that may not be relevant to the soil data product may appear in the data table. To limit the DNA extraction dataset to those that are relevant to the group abundances samples, filter the records in the *mmg_soilDnaExtraction* table to include only those with a **dnaSampleID** that is also contained in the *mga_soilGroupAbundances* table.

Soil Microbe Group Abundances (DP1.10109.001)

mga_soilGroupAbundances.csv -> This table includes the gene copy number data for each sample. One record is expected per **dnaSampleID** per **targetTaxonGroup**.

mga_batchResults.csv -> This table describes the batch-level data associated with a qPCR run. One record is expected per batch of samples analyzed (**batchID**), and links to the table *mga_soilGroupAbundances* by

the **batchID**. *Important Note*: The batch results table is generic for all soil and aquatic data: samples that may not be relevant to this data product may appear in the data table. To limit the dataset to those that are relevant to the soil group abundances data, filter the records to only those with **batchID**'s matching the **batchID**'s in the *mga_soilGroupAbundances* table.

mga_labSummary.csv -> This table describes the laboratory methods used to analyze samples, with **labSpecificStartDate** and **labSpecificEndDate** indicating the date range over which the methods apply. The summary table is generic for all soil and aquatic data. One record is expected per unique set of methods. The start and end dates can be used to filter the data in ***mga_soilGroupAbundances*** using the fields **laboratoryName**, **processedDate**, and **targetTaxonGroup**.

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*.

3.9.2 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 Surface water microbe cell count (DP1.20138), 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_fieldCellCounts*** within the Surface water microbe cell count product. These **geneticSampleID**s are sent for DNA extraction such that the **geneticSampleID** from ***amc_fieldCellCounts*** = **geneticSampleID** in ***mmg_swDnaExtraction***.

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 Cell Count (DP1.20138.001)

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

amc_fieldCellCounts.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 Surface water microbe metagenome sequences (DP1.10107.001) variable **geneticSampleID** in the table ***mmg_swDnaExtraction***.

Surface Water Microbe Marker Gene Sequences (DP1.20282.001)

mmg_swDnaExtraction.csv -> This table contains the DNA extraction laboratory data. Data 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 provided to the lab. Duplicate records for an individual **dnaSampleID** should not exist.

Surface Water Microbe Group Abundances (DP1.20278.001)

mga_swGroupAbundances.csv -> This table includes the gene copy number data for each sample. One record expected per **dnaSampleID** per **targetTaxonGroup**.

mga_batchResults.csv -> This table describes the batch-level data associated with a qPCR run. One record is expected per batch of samples analyzed (**batchID**), and links to the table ***mga_swGroupAbundances*** by the **batchID**. *Important Note*: The batch results table is generic for all soil and aquatic data: samples that may not be relevant to this data product may appear in the data table. To limit the dataset to those that are relevant to the soil group abundances data, filter the records to only those with **batchID**'s matching the **batchID**'s in the **mga_swGroupAbundances_** table.

mga_labSummary.csv -> This table describes the laboratory methods used to analyze samples, with **labSpecificStartDate** and **labSpecificEndDate** indicating the date range over which the methods apply. The summary table is generic for all soil and aquatic data. One record is expected per unique set of methods. The start and end dates can be used to filter the data in ***mga_swGroupAbundances*** using the fields **laboratoryName**, **processedDate**, and **targetTaxonGroup**.

3.9.2.2 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 (DP1.20280.001)

amb_fieldParent.csv -> One record expected per namedLocation and collectDate (day of year, local time), and generates a unique **sampleID**. Record represents a subsample (**geneticSampleID**) of the field-collected sample.

mmg_benthicDnaExtraction.csv -> This table contains the DNA extraction laboratory data. Data 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 provided to the lab. Duplicate records for an individual **dnaSampleID** should not exist.

Benthic Microbe Group Abundances (DP1.20277.001)

mga_benthicGroupAbundances.csv -> This table includes the gene copy number data for each sample. One record expected per **dnaSampleID** per **targetTaxonGroup**.

mga_batchResults.csv -> This table describes the batch-level data associated with a qPCR run. One record is expected per batch of samples analyzed (**batchID**), and links to the table ***mga_benthicGroupAbundances*** by the **batchID**. *Important Note*: The batch results table is generic for

all soil and aquatic data: samples that may not be relevant to this data product may appear in the data table. To limit the dataset to those that are relevant to the soil group abundances data, filter the records to only those with **batchID**'s matching the **batchID**'s in the **mga_benthicGroupAbundances_** table.

mga_labSummary.csv -> This table describes the laboratory methods used to analyze samples, with **labSpecificStartDate** and **labSpecificEndDate** indicating the date range over which the methods apply. The summary table is generic for all soil and aquatic data. One record is expected per unique set of methods. The start and end dates can be used to filter the data in ***mga_swGroupAbundances*** using the fields **laboratoryName**, **processedDate**, and **targetTaxonGroup**.

3.10 Special Considerations

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

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 Microbe Group Abundances (DP0.10109.001), 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]).

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 Analyzing Real-Time PCR Data

To the extent possible, microbe group abundances data follow the *M*inimum *I*nformation of *Q*uantitative Real-Time PCR *E*xperiments (MIQE) metadata and QA/QC reporting guidelines outlined in Bustin et al. (2009). The MIQE standards define the essential and desired metadata to be reported for a qPCR reaction and include parameters related to experimental design, target gene information, oligonucleotides, the SOP/protocol, qPCR data validation and data analysis.

For each data product (soil, surface water, and benthic). the data table ***mga_groupAbundances*** presents the abundance data for each sample as a unique record for each targetTaxonGroup. For example, samples that are analyzed for **targetTaxonGroup**s 'bacteria', 'archaea' and 'fungi' separately should contain 3 records, while samples that are analyzed for the **targetTaxonGroup**s 'bacteria and archaea' and 'fungi' should contain 2 records.

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]).

4.2.2 Primer Information for Gene Regions Used

As these data are derived from PCR reactions, a specific region of the target organism is amplified, which for bacteria and archaea is the 16S ribosomal region, and for fungi is the internal transcribed spacer (ITS) region that lies between the small and large subunits of ribosomal RNA. As a guide for end-users, below are the primer sequences for each of the PCR targets. Primers are summarized in Table 1.

Table 1: Primers used for qPCR

Target Gene	Primer Name	Sequence
16S	341F	CCTACGGGNBGCASCAG
	805R	GGACTACNVGGGTATCTAATCC
ITS	ITS1f	CTTGGTCATTTAGAGGAAGTAA
	ITS2r	GCTGCGTTCTTCATCGATGC

Bacterial/Archaea 16S Primers

The primers used to measure abundance of bacteria and archaea produce an amplicon of 464 base pairs, and they span the **V3** and **V4** hypervariable regions of the 16S ribosomal region (Figure 4).

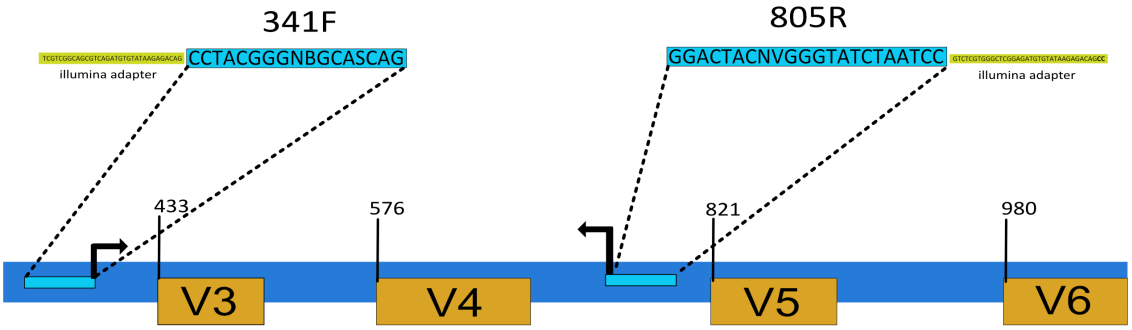

Figure 4: Map of bacterial/archaea primers in the 16S ribosomal region. Hypervariable regions are in Orange. Scale is approximate.

Fungal ITS Primers

For the fungal ITS amplicon products, the primers target the ITS-1 region, which lies between the 18S large rRNA subunit and the small 5.8S subunit. The primers are ITS1f and ITS2r, which is not to be confused with the *regions* ITS1 and ITS2. The primers are located within the ribosomal subunits, as these regions are much more conserved than the spacer (ITS1f is at the 3 prime end of the 18S subunit, and ITS2r is at the 5 prime end of the 5.8S unit). See Figure 5.

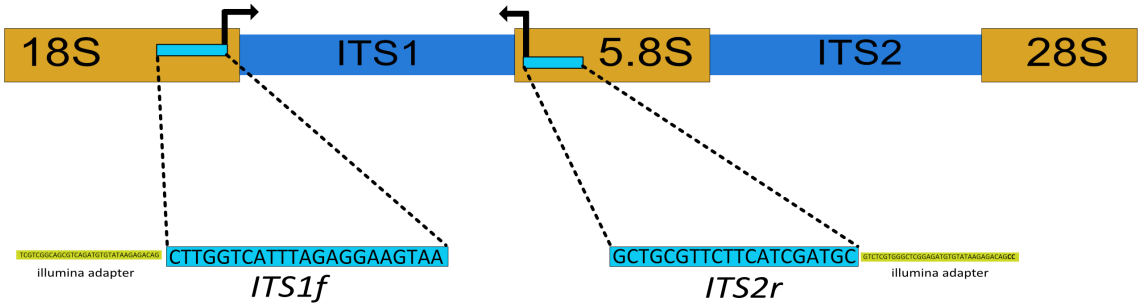

Figure 5: Map of fungal primers in the ITS ribosomal region. rRNA regions are in Orange. Not to scale.

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. Please see the table below for an explanation of **dataQF** 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

4.5 Analytical Facility Data Quality

At a minimum, the analytical laboratory conforms to the following rules to ensure data quality:

- For each batch of samples, the standard curve must achieve an $R^2 \geq 0.95$ based on at least 3 concentrations from a dilution series, with each concentration run in triplicate (batch-level results are reported in the data table **mga_batchResults**);
- Limit of detection for a batch is defined as the highest concentration obtained from a negative (no-template) control sample in that batch: samples at or below this threshold are considered below detection and will be reported in the data table **mga_groupAbundances**, field **qaqcStatus** as “Fail”;
- Test samples are run in triplicate and no more than 1 replicate can fail to produce usable data.
- Any records with a **qaqcStatus** = “Fail” should also be accompanied by free-form notes in the “remarks” variable

More detailed descriptions of laboratory methods and QAQC procedures can be found in the laboratory SOP, which appears in the field **testProtocolVersion** in the mga_groupAbundances data tables. Refer also to the data table ***mga_labSummary*** - available in the expanded package for each data product download - which provides additional details on the general analytical approach and documentation on the methods used during the specified period of time.

5 REFERENCES

1. Bustin, S. A., V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, et al. 2009. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. Clinical Chemistry 55:611-622. https://doi.org/10.1373/clinchem.2008.112797.
2. Ginzinger, D. G. 2002. Gene quantification using real-time quantitative PCR: An emerging technology hits the mainstream. Experimental Hematology 30:503-512.

3. Takahashi, S., J. Tomita, K. Nishioka, T. Hisada & M. Nishijima. 2014. Development of a prokaryotic universal primer for simultaneous analysis of bacteria and archaea using next-generation sequencing. PloS One 9:e105592.
4. Walters, W., E.R. Hyde, D. Berg-Lyons, G. Ackermann, G. Humphrey, A. Parada, J.A. Gilbert, J.K. Jansson, J.G. Caporaso, & J.A. Fuhrman. 2016. Improved Bacterial 16S RRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. Msystems 1 (1): e00009–15.