NEON USER GUIDE TO MICROBE MARKER GENE SEQUENCES (DP1.10108.001; DP1.20280.001; DP1.20282.001)

PREPARED BY	ORGANIZATION
Lee Stanish	TOS
Stephanie Parker	AOS

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	11/22/2017	Initial Release
B	06/05/2019	Revised Section 3.2.1: Temporal resolution of soil analyses; Included details on use of raw sequence data files in Sections 3.7-3.9
C	10/01/2020	Included general statement about usage of neonUtilities R package and statement about possible location changes; Section 3.3: Added Sampling Design Changes section and included changes to sampling frequency for microbial analyses; Section 3.6.1: Updated description of Associated Data Streams for bundled Soil Physical and Chemical Properties data product; Section 3.10.1: Updated download format of raw sequence data from per-batch to per-sample resolution; Section 3.10.2: Updated details on accessibility and availability of data on external sequence data repositories.

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 (DP1.10108.001) (AD[05]), NEON Data Variables for Benthic Microbe Marker Gene Sequences (DP1.20280.001) (AD[06]), and NEON Data Variables for Surface Water Microbe Marker Gene Sequences (DP1.20282.001) (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 Marker Gene Sequences (DP0.10108.001) (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.000913	TOS Science Design for Spatial Sampling
AD[14]	NEON.DOC.000008	NEON Acronym List
AD[15]	NEON.DOC.000243	NEON Glossary of Terms
AD[16]	NEON.DOC.004825	NEON Algorithm Theoretical Basis Document: OS Generic Transitions
AD[17]	Available on NEON data portal	NEON Ingest Conversion Language Function Library
AD[18]	Available on NEON data portal	NEON Ingest Conversion Language
AD[19]	Available with data download	Categorical Codes csv

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 (DP1.10081.001, DP1.20086.001, DP1.20141.001). 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)

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.

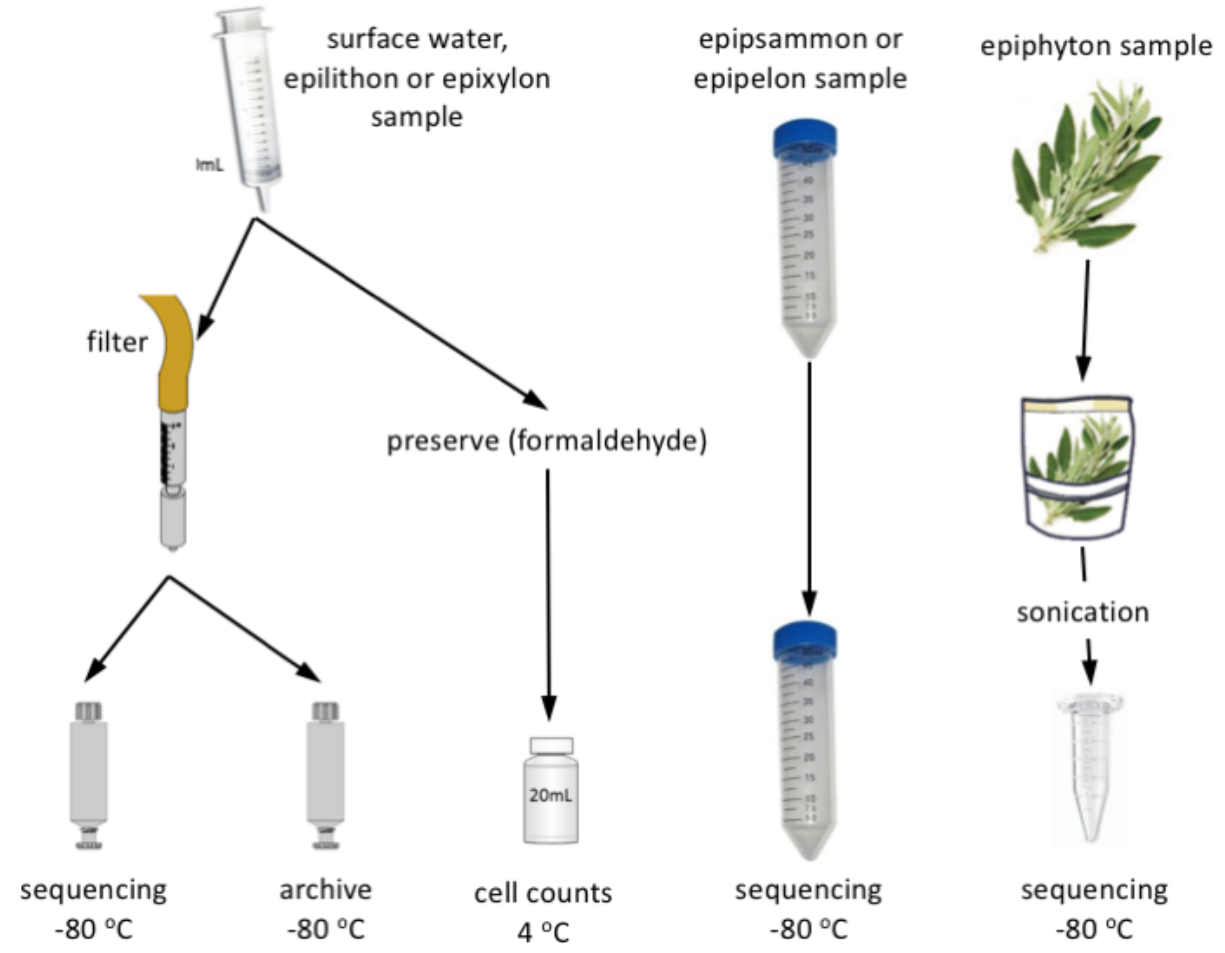

Figure 1: Overview of microbial field sample types, processing steps, and analyses. Note that samples destined for cell count analyses are part of a different data product, DP1.20138.001.

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).

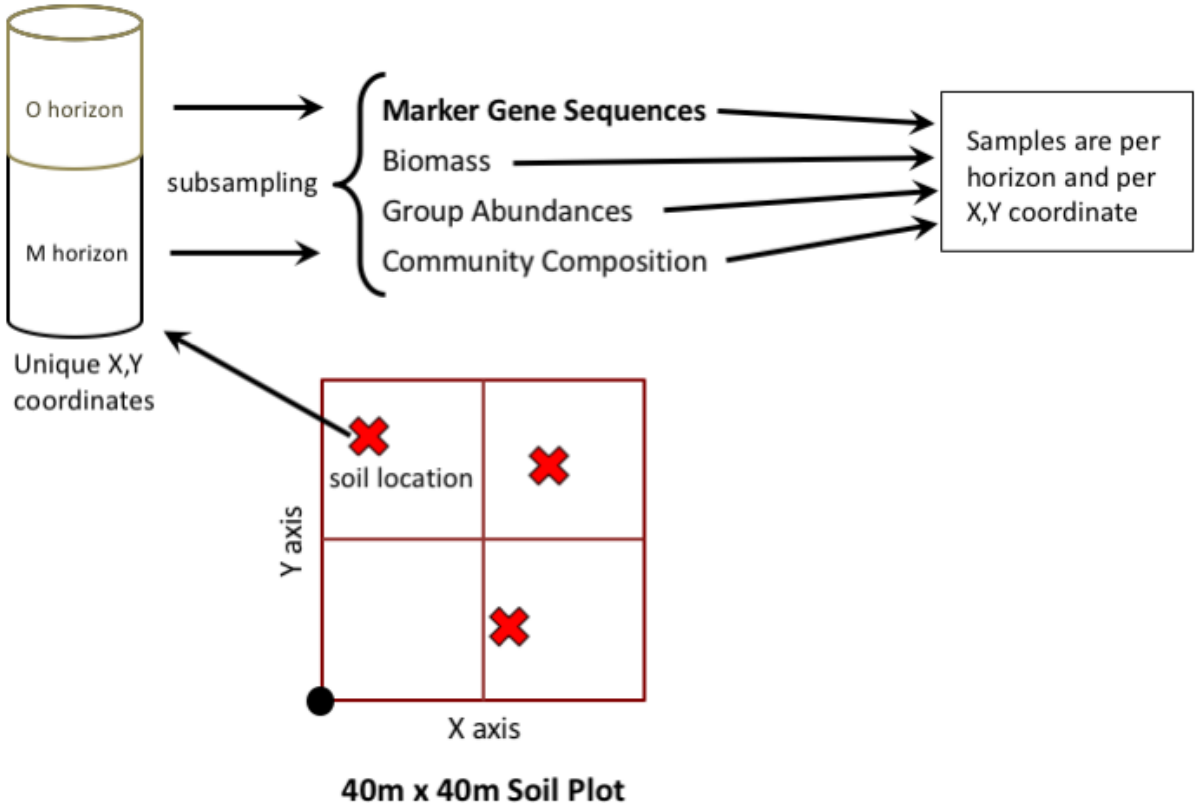

Figure 2: Overview of soil microbial field sampling, spatial design, and analysis workflow.

At aquatic sites, microbial surface water samples are collected in conjunction with water chemistry sampling (Figure 3). In lakes, up to 3 locations are sampled: the lake inlet, lake outlet, and profiling buoy. In seepage lakes (no true inlet and outlet), microbe samples are collected only at the buoy for samples collected in 2018 or later. In flow-through lakes (with a true inlet and outlet), samples are collected at all 3 lake locations. At large, non-wadeable streams (rivers), the sampling location is near the buoy sensor array. At both lakes and river buoy locations, either 1 or 2 samples are collected depending on whether the lake/river is stratified. In stratified systems, one sample is collected from the surface of the epilimnion, and one sample from the midpoint of the hypolimnion. In non-stratified sites, one surface sample is collected. In wadeable streams, one surface water sample is collected near the downstream sensor array.

Aquatic benthic microbial samples are collected in wadeable streams at up to 8 locations throughout the 1 km sampling 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 (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]).

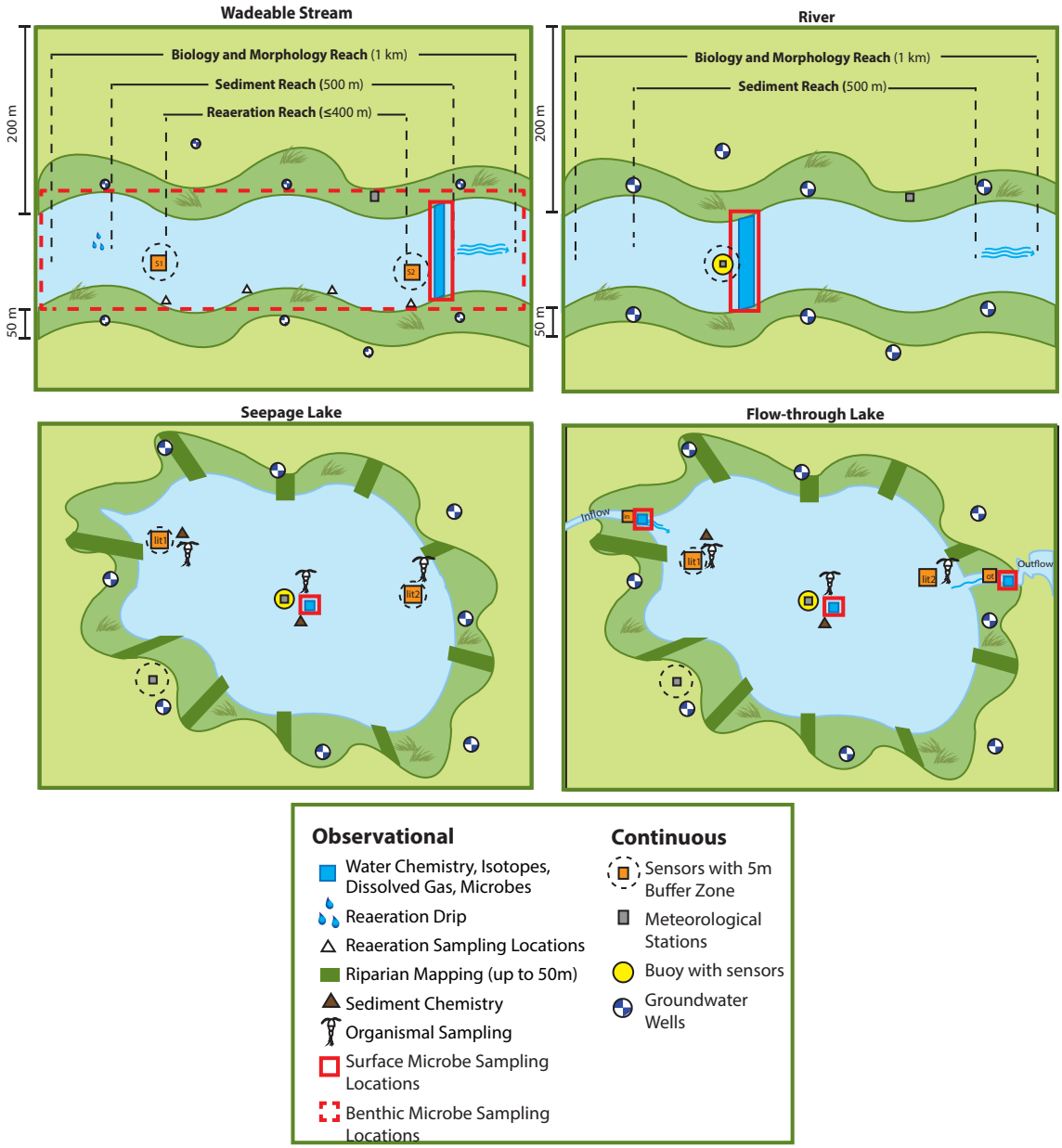

Figure 3: Generic NEON aquatic site layouts with microbial sampling locations highlighted in red boxes.

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 for soil (DP1.10086.001), aquatic surface water (DP1.20138.001) and benthic (DP0.20270.001) sampling can be found in the Data Product User Guides for those respective Data Products.

3.2.1 Soils

Soil sampling for marker gene sequence analysis occurs at a minimum of one site per domain and up to 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). Sites with short growing seasons (e.g. tundra and taiga) are sampled 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.

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 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 marker gene sequencing analyses (16S and ITS sequencing) during every soil sampling bout and at all sites.
- 2018 - current: Subsamples are collected for microbial marker gene sequencing analyses (16S and ITS sequencing) 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 field or laboratory technician (L0 data) are listed in the file, NEON Raw Data Validation for Microbe Marker Gene Sequences (DP0.10108.001) (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 Marker Gene Sequences (DP1.10108.001) (AD[05]).
- NEON Data Variables for Benthic Microbe Marker Gene Sequences (DP1.20280.001) (AD[06]).
- NEON Data Variables for Surface Water Microbe Marker Gene Sequences (DP1.20282.001) (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.

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.

3.5 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 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**. For more information and instructions, refer to the NEON User Guide to Soil physical and chemical properties, periodic (DP1.10086.001).

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 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***.

3.6 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 marker gene sequences 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 marker genes 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 may be used to link data in this product to soil microbe marker genes data.
- Soil microbe group abundances (DP1.10109.001): 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 (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** for a data record. 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: 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 (DP1.20086) and surface water (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 (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 (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 (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 (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**.

3.8 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 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 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 unique records per site per year.

Depending on the data delivery format from the analytical laboratory, raw sequence data may include multiple files, and each unique file that is associated with a sample generates a new record. Thus, a user may expect 1-4 x the number of unique records for raw sequence data, for a range of 24-480 unique records per site x year combination.

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

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 and chemical properties, periodic (DP1.10086). In general, every record that contains a **geneticSampleID** should be used for marker gene sequence analysis. A record from ***sls_soilCoreCollection*** may have zero or one child records in all tables 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. In general, each **dnaSampleID** represents an independent record. Sometimes, the lab may also report an **internalLabID**. In these instances, an independent record would be **dnaSampleID** + **internalLabID**. Duplicate records for an independent record (either **dnaSampleID** or **dnaSampleID** + **internalLabID**) should not exist. Lab replicates from the same DNA extraction will have the same **dnaSampleID** but different **internalLabID**'s.

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.

Both metadata and per sample raw/minimally processed sequence data are available on the NEON data portal. In addition, a subset of quality-filtered 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 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), Soil microbe community composition (DP1.10081), Soil microbe group abundances (DP1.10109), and Soil microbe biomass (DP1.10104). Additionally, subsamples generated from soil sampleIDs are used in Soil inorganic nitrogen pools and transformations (DP1.10080) and Soil carbon and nitrogen concentrations and stable isotopes (DP1.10100.001).

Soil Microbe Marker Gene Sequences (DP1.10108)

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 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***.

mmg_soilRawDataFiles.csv - > One or more records is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the tables ***mmg_soilMarkerGeneSequencing_16S/ITS***, ***mmg_soilPcrAmplification_16S/ITS*** and ***mmg_soilDnaExtraction***. One record per combination of **dnaSampleID**, **targetGene**, and **rawDataFileName** is expected.

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***.

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***. In general, each **dnaSampleID** represents an independent record. Sometimes, the lab may also report an **internalLabID**. In these instances, an independent record would be **dnaSampleID** + **internalLabID**. Duplicate records for an independent record (either **dnaSampleID** or **dnaSampleID** + **internalLabID**) should not exist. Lab replicates from the same DNA extraction will have the same **dnaSampleID** but different **internalLabID**'s.

Both metadata and per-sample raw/minimally processed sequence data are available on the NEON data portal. In addition, a subset of quality-filtered 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 (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***.

mmg_swRawDataFiles.csv - > One or more records is expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, which corresponds to the **dnaSampleID** in the tables ***mmg_swMarkerGeneSequencing_16S/ITS***, ***mmg_swPcrAmplification_16S/ITS*** and

mmg_swDnaExtraction. One record per combination of **dnaSampleID**, **targetGene** and **rawDataFileName** is expected.

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***. In general, each **dnaSampleID** represents an independent record. Sometimes, the lab may also report an **internalLabID**. In these instances, an independent record would be **dnaSampleID** + **internalLabID**. Duplicate records for an independent record (either **dnaSampleID** or **dnaSampleID** + **internalLabID**) should not exist. Lab replicates from the same DNA extraction will have the same **dnaSampleID** but different **internalLabID**'s.

Both metadata and per-sample raw/minimally processed sequence data are available on the NEON data portal. In addition, a subset of quality-filtered 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*.

Benthic Microbe Marker Gene Sequences (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***.

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

mmg_benthicMarkerGeneSequencing_16S/ITS, ***mmg_benthicPcrAmplification_16S/ITS*** and ***mmg_benthicDnaExtraction***. One record per combination of **dnaSampleID**, **targetGene**, and **rawDataFileName** is expected.

3.10 Special Considerations

There are multiple venues for retrieving NEON sequence data: raw data directly from the NEON data portal and quality processed data from external sequence data repositories.

3.10.1 From the NEON Data Portal

Raw sequence data can be downloaded from the URL listed in the NEON field **rawDataFilePath**, located in the rawDataFiles publication table. Clicking on the URL will initiate download of the sequence file. Files can also be automatically downloaded and un-zipped in the R software environment using the neonUtilities package (v1.2.2 or later), available at https://cran.r-project.org/web/packages/neonUtilities/index.html.

When downloading raw sequence data files directly from the NEON data portal, the following should be considered:

a) The raw data files are typically megabytes (MB) to gigabytes (GB) in size. Ensure you have sufficient space prior to downloading many files.
b) Downloaded files are in a compressed (.gz or .tar.gz) format. Files may require de-compressing prior to use.
c) In some cases, downloaded files may contain sequence data from an entire sequencing run, including data for non-target samples. The field **rawDataFileDescription** provides details into whether the sequence data for a record are at the per sample or per sequencing run level of resolution.
d) NEON currently performs bidirectional sequencing, meaning that two sets of sequence data, one in the 5' or forward direction and one in the 3' or reverse direction, are generated. Merging of forward and reverse sequence reads may be necessary.

3.10.2 From External Sequence Repositories

A subset of NEON sequence data has been uploaded to the data repository MG-RAST (http://metagenomics.anl.gov, Meyer et al., 2008), which 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

There are a number of ways to search and retrieve marker gene sequence data from external sequence repositories:

- From the NEON data portal: Links to the MG-RAST Data Repository are provided from the NEON Data Portal. Once at the MG-RAST site, entering "NEON" (case-insensitive) to the search query will bring up a list of all known NEON data in MG-RAST. This list can be further refined using the existing filtering functionality on the MG-RAST website. For example, using the 'Advanced Search" options and filtering the'study' field for the term 'surface' filters to the surface water sequence data.
- 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. Registering for a free user account is recommended.
- 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: New data are not currently being published on external sequence data repositories. The NEON data portal is the primary repository for NEON sequence 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 Marker Gene Sequences (DP0.10108.001), 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[18]).

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

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. Data fields containing the per sample QA results are published as part of the basic 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[16]).

4.3 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.

4.4 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 1: 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

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

Data analyses conducted on marker gene sequencing data conform to the current data quality standards used by practitioners. The data tables ***mmg_dnaExtraction***, ***mmg_pcrAmplification***, and

mmg_sequencing include a variable, called **qaqcStatus**, in which the laboratory can indicate sample processing issues arising during DNA extraction, PCR amplification, or DNA sequencing, respectively. Records that pass the QAQC criteria described in the associated Laboratory SOP (listed in the data field **testProtocolVersion** and available for download from the NEON Data Portal) will have a qaqcStatus = "Pass". Any records with a qaqcStatus = "Fail" should also be accompanied by free-form notes in the "remarks" variable. Typically, a sample that fails a QAQC step will not undergo downstream processing, although exceptions do exist. Users should review the QAQC criteria used by the analytical laboratory as described in the Laboratory SOP and determine whether to retain or remove records with a failing **qaqcStatus**.

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.