NEON USER GUIDE TO MICROBIAL METAGENOME SEQUENCES (DP1.10107.001; DP1.20279.001; DP1.20281.001)

PREPARED BY	ORGANIZATION
Lee Stanish	FSU
Stephanie Parker	AQU

CHANGE RECORD

REVISION	DATE	DESCRIPTION OF CHANGE
A	10/26/2017	Initial Release
B	07/31/2019	Added new section on use of raw sequence data files in Sections 3.7-3.9
C	10/22/2020	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 3.7.1: Updated description of Associated Data Streams for bundled Soil Physical and Chemical Properties data product; Section 3.9: Clarified data relationships and joining data across related data products.
D	03/16/2022	Updated section 4.3 Data Revision with latest information regarding data release

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 DESCRIPTION

1.1 Purpose

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

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

1.2 Scope

This document describes the steps needed to generate the L1 data product Microbial Metagenomic Sequences, and associated metadata, from input data. Data from the subsamples can be found in the related data products listed below. This document also provides details relevant to the publication of the data products via the NEON data portal, with additional detail available in the file, NEON Data Variables for Soil Microbial Metagenomic Sequences (DP1.10107.001) (AD[05]), NEON Data Variables for Surface Water Microbial Metagenomic Sequences (DP1.20281.001) (AD[06]), or NEON Data Variables for Benthic Microbial Metagenomic Sequences (DP1.20279.001) (AD[07]) provided in the download package for each of the three data products.

This document describes the process for ingesting and performing automated quality assurance and control procedures on the laboratory data from samples generated by the following field sampling protocols: TOS Protocol and Procedure: Soil Biogeochemical and Microbial Sampling (AD[10]) for upland soil samples; TOS Standard Operating Procedure: Wetland Soil Sampling (AD[11]) for wetland soil samples; or AOS Protocol and Procedure: Aquatic Microbial Sampling (AD[12]) for aquatic samples. The raw data that are processed as described in this document are detailed in the file, NEON Raw Data Validation for Microbial Metagenomic Sequences (DP1.10107.001) (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

3 DATA PRODUCT DESCRIPTION

Microbial shotgun metagenomics is a technique for evaluating microbial community structure and functional potential in a sample. These data are intended to allow relationships between genomic content of samples and environmental and biogeochemical parameters to be discerned for understanding and potentially predicting long-term changes in microbial structure and function.

The Microbial Metagenomic Sequences data product provides shotgun metagenomic sequence data and metadata for soil and aquatic (surface water and benthic) microbial samples. The sampling plan implements the guidelines and requirements described in the Science Designs for TOS Terrestrial Microbial Diversity (AD[08]) and Aquatic Sampling (AD[09]). Sample collection methods differ between aquatic and terrestrial samples, but in general samples are minimally processed in order to reduce the introduction of microbial contaminants. For most samples, including soil and epipsammon, native material is processed for analysis; however, certain aquatic sample types have additional processing steps (Figure 1). After field collection, samples are frozen in the field on dry ice and transported to ultra-low freezers at the NEON field laboratories. Samples are shipped to an analytical laboratory where DNA extraction, sample library preparation and DNA sequencing occur.

The laboratory performs minimal processing of sequence data. Typically, this includes:

a. Demultiplexing, or parsing of sequence data on a per-sample basis
b. Removing sequencing indexes, which are short oligonucleotide sequences added to the sample DNA to enable analysis of many samples in a single run

The exact pre-processing steps and methodologies used may vary over time: users should refer to the Laboratory SOPs and Protocols associated with a particular sample record for more detail.

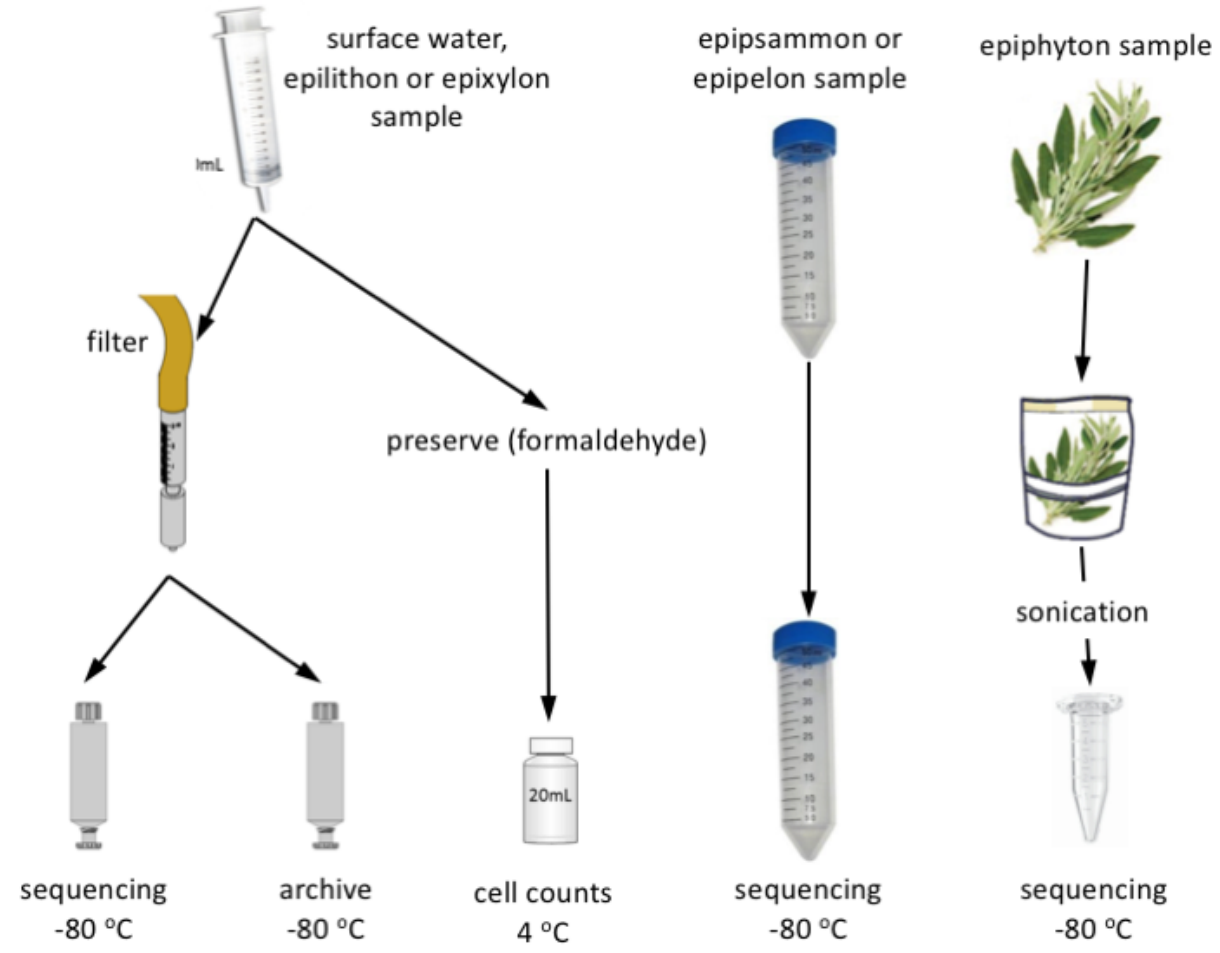

Figure 1: Overview of aquatic microbial field sample types, field processing steps, and analyses. Note that samples destined for cell count analysis are part of a different data product, DP1.20138.001.

For soils, a sample represents a plot-level composite sample of soils collected at 1-3 randomly assigned, individual X, Y locations of a particular horizon type (Figure 2). NEON designates soil horizons broadly as either organic (O) or mineral (M).

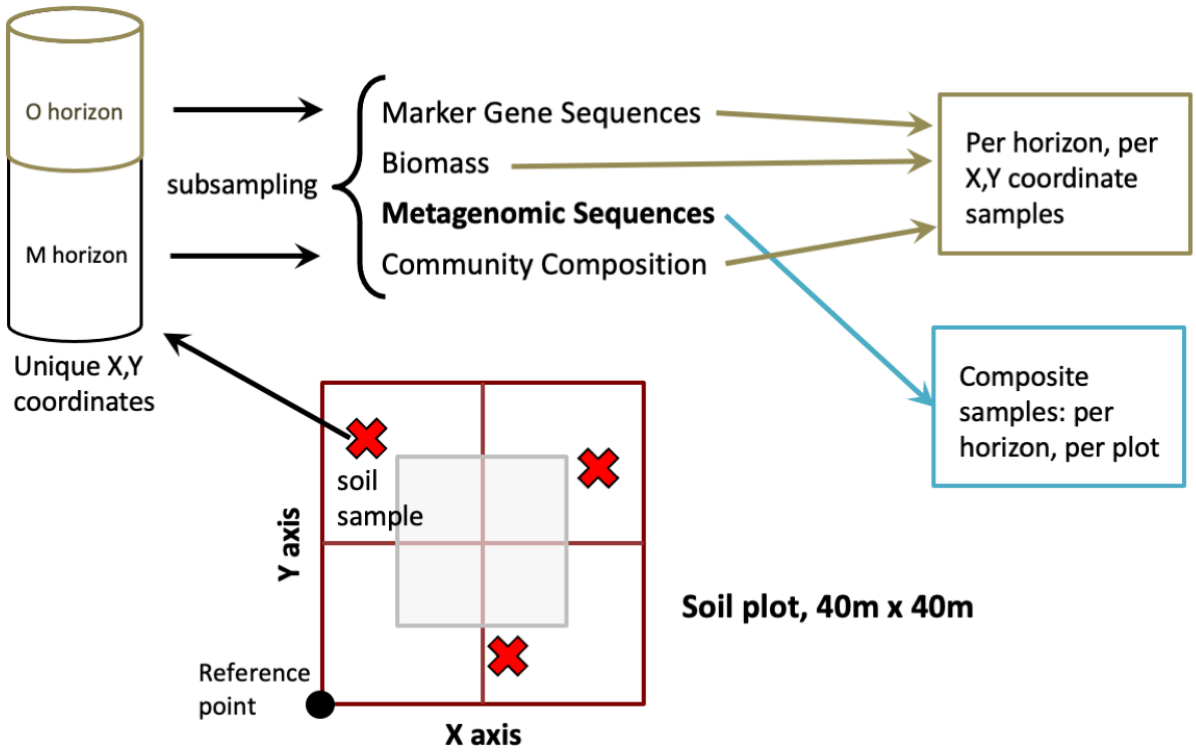

Figure 2: Overview of soil microbial field sampling and analysis workflow. The 20m x 20m grey region in center of plot is not sampled for soils to minimuze distrubance to plant communities.

3.1 Spatial Sampling Design

Microbial metagenomics sampling is executed at all NEON sites. A summary of the spatial design for the aquatic and terrestrial sampling is provided here. More comprehensive descriptions for soil (DP1.10086) and aquatic surface water (DP1.20138 for surface water and DP0.20270.001 for benthic) sampling can be found in the associated Data Product User Guides.

At terrestrial sites, soils are sampled from three pre-determined, randomly assigned X,Y locations per 40 x 40 meter plot (Figure 2). Ten plots per site are sampled, four within the tower airshed (Figure 3) and six others distributed across the landscape, located in dominant vegetation types. The number of distributed plots within each vegetation type are proportional to the percent coverage of that type See AD[02] for further details on the NEON TOS spatial design.

Aside from the spatial data, which is reported at the plot level from the plot centroid, all accompanying field and non-metagenomic laboratory data are reported at the spatial resolution of a single sampling location, e.g., an X,Y coordinate (+/- 0.5 meters) within a NEON plot. For generating plot-level field data

to accompany pooled metagenomic soil samples, a data user should calculate average values for each individual sample used to generate the composite sample. The individual samples used to generate the pooled metagenomics samples are found as a pipe-delimited string in the field **genomicsPooledIDList** located in the data table ***sls_metagenomicsPooling***, which is part of the Soil Physical Properties (distributed periodic) data product (DP1.10086).

At aquatic sites, microbial surface water samples are collected in conjunction with water chemistry sampling (Figure 4), and the number and types of sampling locations varies by ecosystem type. At sites consisting of a flow-through lake (true inlet and outlet), up to 3 locations are sampled: the lake inlet, lake outlet, and profiling buoy. At sites consisting of seepage lakes (no true inlet and outlet), microbe samples are collected at the buoy location only. At large, non-wadeable stream sites (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. Benthic microbial samples are collected at all wadeable streams at up to 8 locations throughout the 1 km sampling reach.

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/

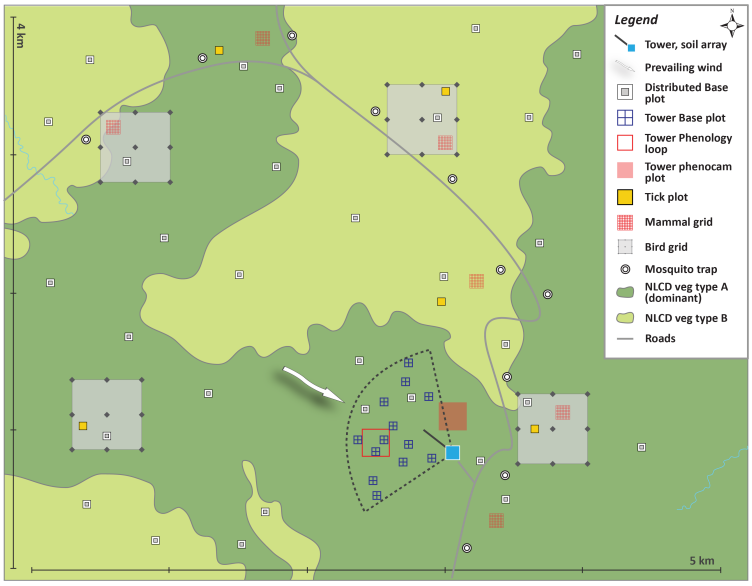

Figure 3: Representation of a NEON terrestrial site with Tower and Distributed plots shown. A subset of six (6) distributed base plots shown here are randomly selected for soil sampling, after accounting for vegetation type.

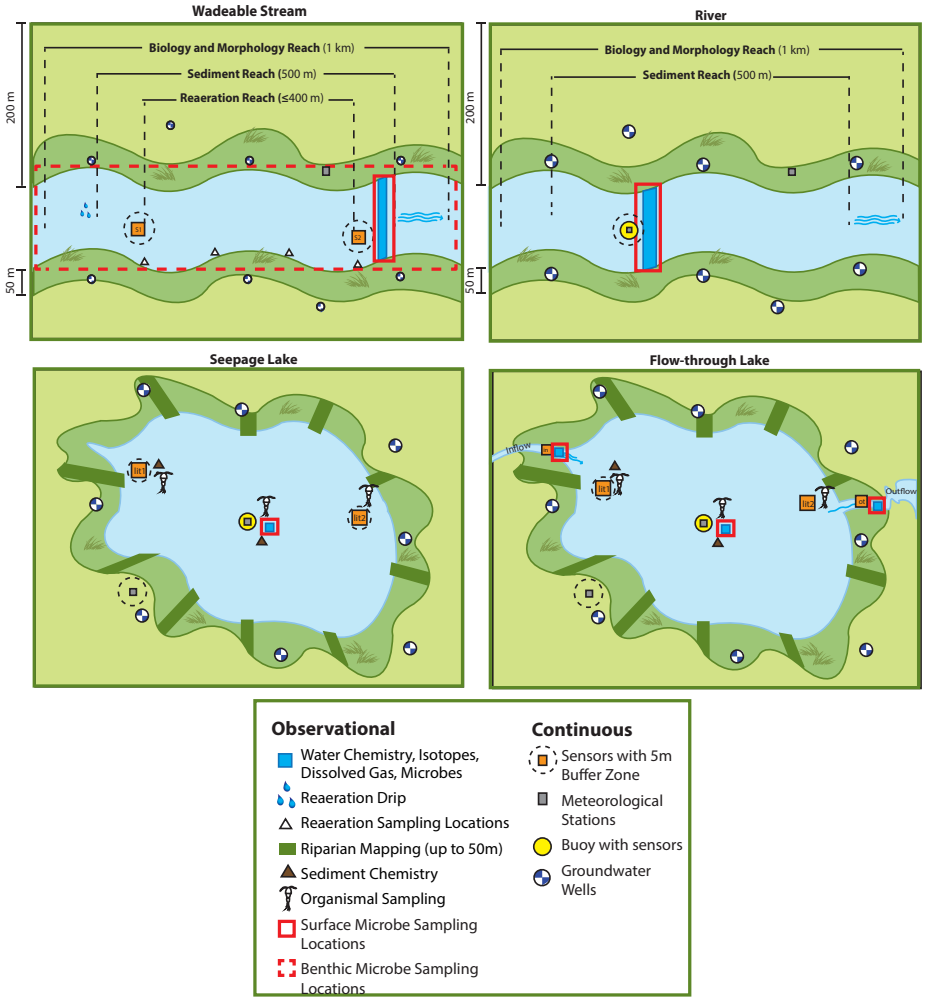

Figure 4: Generic NEON aquatic site layouts with microbial sampling locations highlighted in red boxes.

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), aquatic surface water (DP1.20138), and aquatic benthic habitats (DP0.20270.001) can be found in those respective Data Product User Guides.

3.2.1 Soils

At terrestrial sites, soil metagenomic sampling occurs annually at a minimum of one site per domain during the period of peak greenness and in conjunction with the soil physical and chemical properties data product (DP1.10086).

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 metagenomics analysis to a maximum depth of 30 cm.

3.2.2 Aquatics

At aquatic sites, surface water sampling and metagenomic analysis occurs annually and approximately during the period of peak productivity. Benthic microbial sampling occurs only in wadeable streams, and is on a similar schedule to periphyton sampling, which occurs 3 times per year (roughly equating to Spring, Summer, and Autumn). The Summer sampling event ("Bout 2") is analyzed for metagenomics.

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 - 2014: Subsamples were collected for metagenomic sequencing analysis during every soil sampling bout, at all sites, and subsamples from unique sampling locations were not pooled at the plot level.
- Jan 2014 - Sept, 2014: Pooling subsamples at the plot level was tested, resulting in a subset of samples collected during this time period being pooled at the per-plot, per-horizon level by the analytical laboratory prior to analysis.
- 2015 - current: Subsamples from the same plot and for the same horizon type are pooled prior to analysis.
- 2015 - 2019: Samples are collected for shotgun metagenomics analysis once per year during the period of peak greenness/productivity at all sites.
- 2019 - current: Samples are collected for shotgun metagenomics analysis once per year during the period of peak greenness/productivity at one site per domain and during all 'coordinated' bouts at any site.

3.3.2 Aquatics

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

- 2018 - current: Surface water samples are collected at only the buoy sensor station at seepage lake sites (lacking a true inlet and outlet).

3.4 Variables Reported

All variables reported from the analytical laboratory (L0 data) are listed in the file, NEON Raw Data Validation for Microbial Metagenomic Sequences (DP1.10107.001) (AD[04]). All variables reported in the published data (L1 data) are also provided separately in the following files:

- NEON Data Variables for Soil Microbial Metagenomic Sequences (DP1.10107.001) (AD[05])
- NEON Data Variables for Surface Water Microbial Metagenomic Sequences (DP1.20281.001) (AD[06])
- NEON Data Variables for Benthic Microbial Metagenomic Sequences (DP1.20279.001) (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), and the VegCore data dictionary (https://projects.nceas.ucsb.edu/nceas/projects/bien/wiki/VegCore; accessed 16 February 2014).

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

NEON TOS spatial data employs the World Geodetic System 1984 (WGS84) for its fundamental reference datum and GEOID09 for its reference gravitational ellipsoid. 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*, which should be sufficient spatial resolution for plot-level metagenomic samples. For samples that represent a single X,Y location within a plot, a more accurate measurement may be desired. Refer to the User Guide for Soil Physical and Chemical Properties, periodic, for more information and instructions.

3.5.2 Aquatics

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

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

- Surface water samples: Surface water microbe cell count (DP1.20138), in the table ***amc_fieldSuperParent*** and ***mms_fieldSurfaceMicrobes***.
- Benthic samples: Benthic microbe field data (DP0.20270.001), in the table ***amb_fieldParent***.

3.6 Temporal Resolution and Extent

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

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

3.7 Associated Data Streams

This section describes the data products that are directly linked or closely related to the metagenomics sequencing data products.

3.7.1 Soils

Soil metagenomic 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) - This data product bundle includes field data, soil moisture and pH, laboratory measurements of soil carbon and nitrogen concentrations (DP1.10078.001) and stable isotopes (DP1.10100.001), and inorganic nitrogen measurements derived by field incubations of soil (DP1.10080.001). Note that not all measurements are made on every corresponding sample measured for shotgun metagenomics analysis, and vice-versa. To join soil physical and chemical data with these data, join the linking field **genomicsSampleID** in the tables ***mms_metagenomeDnaExtraction*** and ***sls_metagenomicsPooling***.
- Soil microbe community composition (DP1.10081.001): Microbial community composition data derived from marker gene sequencing. To merge these data sets, first join the community composition data with the soil field data table ***sls_soilCoreCollection***, as described in the corresponding Data Product User Guide, then join with the metagenomics data using the individual **sampleID(s)** listed in the field **genomicsPooledIDList** in the table ***sls_metagenomicsPooling***, which corresponds to a **genomicsSampleID** in the ***mms_metagenomeDnaExtraction*** table.

- Soil microbe group abundances (DP1.10109.001): Bacteria/archaeal and fungal abundances as measured by quantitative PCR (qPCR). To merge these data sets, first join the group abundance data with the soil field data table ***sls_soilCoreCollection***, as described in the corresponding Data Product User Guide, then join with the metagenomics data using the individual **sampleID(s)** listed in the field **genomicsPooledIDList** in the table ***sls_metagenomicsPooling***, which corresponds to a **genomicsSampleID** in the ***mms_metagenomeDnaExtraction*** table.
- Soil microbe marker gene sequences (DP1.10108.001): Microbial 16S and ITS sequence data. To merge these data sets, first join the marker genes data with the soil field data table ***sls_soilCoreCollection***, as described in the corresponding Data Product User Guide, then join with the metagenomics data using the individual **sampleID(s)** listed in the field **genomicsPooledIDList** in the table ***sls_metagenomicsPooling***, which corresponds to a **genomicsSampleID** in the ***mms_metagenomeDnaExtraction*** table.
- Soil microbe biomass (DP1.10104.001): Microbial biomass as measured by PLFA. To merge these data sets, first join the biomass data with the soil field data table ***sls_soilCoreCollection***, as described in the corresponding Data Product User Guide, then join with the metagenomics data using the individual **sampleID(s)** listed in the field **genomicsPooledIDList** in the table ***sls_metagenomicsPooling***, which corresponds to a **genomicsSampleID** in the ***mms_metagenomeDnaExtraction*** table.

3.7.2 Aquatics

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

- Surface water microbes field data are part of the download package for the metagenomics data product and do not require additional downloading. The field **geneticSampleID** within the table ***mms_fieldSurfaceMicrobes*** can be used to link these data products.
- Benthic microbes: The field data are part of the download package for the metagenomics data product and do not require downloading additional data products. Tables in this data product 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 ***mga_fieldSuperParent*** can be used to link these data to metagenomics data.
- 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***.
- 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***.
- 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 the metagenomics data.

- Benthic (DP1.20277.001) and surface water (DP1.20278.001) microbe group abundances: Bacteria/archaeal and fungal abundances as measured by quantitative PCR (qPCR). Link using the field **geneticSampleID** in the tables ***mga_benthicGroupAbundances*** and ***mga_swGroupAbundances***.
- Benthic (DP1.20280.001) Microbial 16S and ITS marker gene sequences data. The field **geneticSampleID** in the tables ***amb_fieldParent*** and ***mmg_benthicDnaExtraction*** can be used to link these data to the metagenomic data.
- Surface water (DP1.20282.001) Microbial 16S and ITS marker gene sequences data. The field **geneticSampleID** in the tables ***mmg_swDnaExtraction*** can be used to link these data to the metagenomic data.
- Depth profile at specific depths (DP1.DP1.20254.001): Secchi depth measurements taken at lakes and non-wadeable streams. Information in **eventID** can be used to link these data to the surface water metagenomic data.

3.8 Product Instances

Soil metagenomic samples are collected at all terrestrial NEON sites. A maximum of 10 plots will be sampled at a site within every NEON domain once per year during peak greenness. Most years, the surface soil horizon (organic or mineral) will be collected, while during a coordinated microbes/biogeochemistry bout (which occurs once every 5 years), up to 2 soil horizons will be collected to a maximum depth of 30cm. For each soil horizon sampled, 3 samples per plot are collected. Currently, all of the samples of the same horizon and from the same plot are composited. Thus at most sites, there will be 10 metagenomics samples generated per site per year at sampled sites, with up to 20 samples generated during a coordinated soil microbes/biogeochemistry bout.

Aquatic samples are collected at all aquatic NEON sites. For surface water metagenomics sampling, a maximum of 3 sample locations will be sampled at every site once per year, for a maximum of 3 metagenomics samples collected per site per year. At wadeable stream sites where benthic microbial sampling occurs, up to 8 samples are collected for metagenomics once per year, for a maximum of 8 metagenomics samples per site per year.

3.9 Data Relationships

3.9.1 Soils

Figure 5 provides an overview of the relationships among data products that are closely related to soil metagenome sequences. 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 (DP1.10086). Depending on the type of bout and time of year, a record from ***sls_soilCoreCollection*** may have zero or one child records in (Soil Physical Properties, DP1.10086) ***sls_metagenomicsPooling***.

Up to three soil samples from ***sls_coreCollection*** may be composited into a single sample for metagenomics analyses. The list of **sampleID**s from ***sls_soilCoreCollection*** that comprise a composited metagenomics sample (called the **genomicsSampleID**) is provided in the Soil Physi-

cal Properties product as the **genomicsPooledIDList** in the table ***sls_metagenomicsPooling***. Each **genomicsSampleID** is sent for DNA extraction, generating one or more records in ***mms_metagenomeDnaExtraction*** (i.e. the **genomicsSampleID** in ***sls_metagenomicsPooling*** = **genomicsSampleID** in ***mms_metagenomeDnaExtraction***). For each **genomicsSampleID** occurring in the tables **sls_metagenomicsPooling** and ***mms_metagenomeDnaExtraction***, the composited samples are denoted by the string 'comp'.

In some instances, the soil sample is not composited but instead represents an individual X,Y location. This is a subsample of the parent **sampleID** in the table ***sls_soilCoreCollection***, and is sent for DNA extraction (i.e. **geneticSampleID** in ***sls_soilCoreCollection*** = **genomicsSampleID** in ***mms_metagenomeDnaExtraction***). For each non-composited **genomicsSampleID**, sample names contain the X and Y coordinates.

One or more **dnaSampleID**s are expected per **genomicsSampleID**, depending on the number of DNA extractions that occur on a sample provided to the lab. 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.

Table **mms_metagenomeSequencing** describes the sequencing preparation and analysis metadata. One record is expected per **dnaSampleID**.

Unprocessed (e.g. no quality filtering, although demultiplexing and barcode sequence removal has typically been performed) sequence data are available on the NEON data portal. Table **mms_rawDataFiles** provides URL links that initiate downloading these data. There is typically one sequence file per sample and per read direction (NEON performs bidirectional sequencing), although data from a single sample and direction may be split over multiple files due to the large file size. As such, one record per combination of **dnaSampleID** and **rawDataFileName** is expected.

In addition, a subset of early NEON metagenomics sequence data may be 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 aberrations have occurred; *users should check data carefully for anomalies before joining tables*.

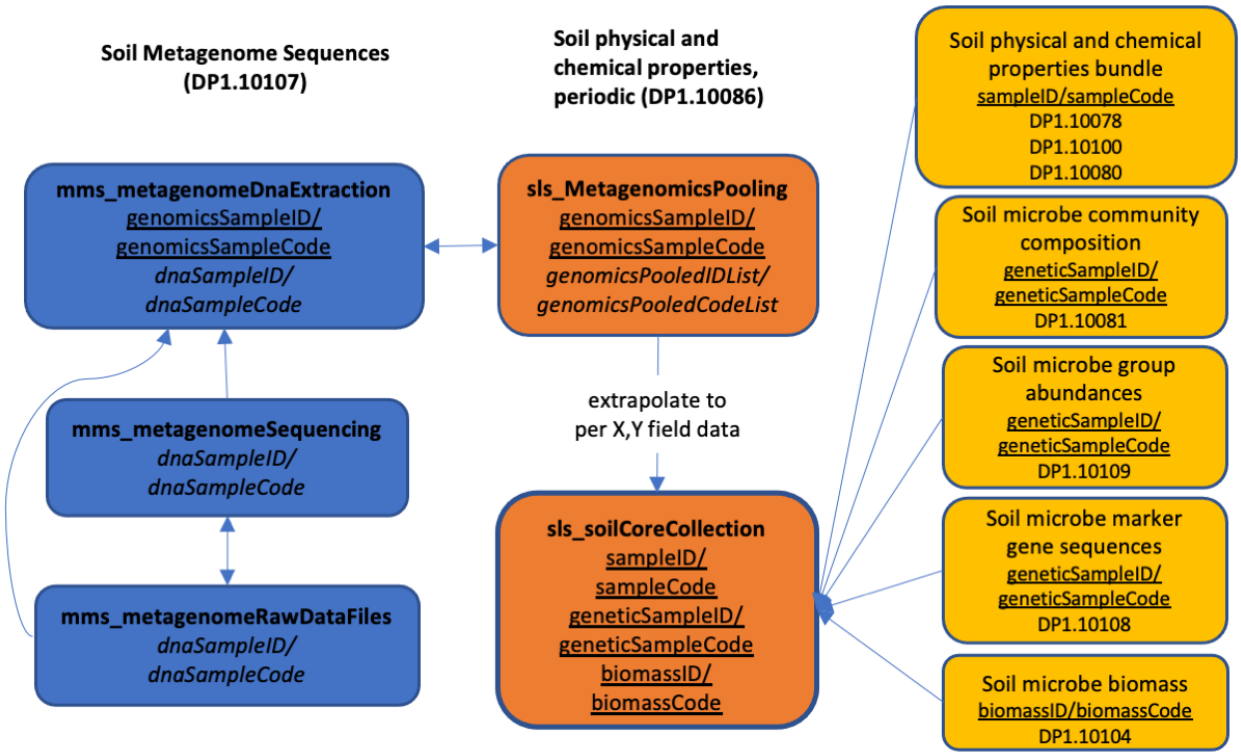

Figure 5: Diagram showing the relationships between the soil metagenome sequencing tables and related soils tables and data products. The mms tables are shown, as well as the key sls data tables in orange and related soils data products in gold. soilCoreCollection is the critical linking table for joining mms with other related soils data products. Within each box, the bolded text is a data table, the underlined text represents fields that link across data products, and the italics text represents fields that link tables within a data product. Data product IDs are provided for related soils data products.

Soil Physical Properties (NEON DP1.10086)

sls_soilCoreCollection.csv - > One record expected per sampleID. Depending upon boutType and whether samples are composited. Generates samples used in Soil microbe community composition (DP1.10081.001), Soil microbe group abundances (DP1.10109.001), Soil microbe marker gene sequences (DP1.10108.001), 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). If soils are not composited, the **geneticSampleID** generated here corresponds to the Soil microbe metagenome sequences (DP1.10107.001) mms_metagenomeDnaExtraction **genomicsSampleID**.

sls_metagenomicsPooling.csv - > One record expected per plotID per horizon per collectDate (day of year, local time). Record represents a mixture of the samples collected in a plot (listed in **genomicsPooledIDList**). Each record generates a single **genomicsSampleID**, corresponding to the **genomicsSampleID** in Soil microbe metagenome sequences (DP1.10107.001) mms_metagenomeDnaExtraction.

Soil Microbial Metagenome Sequences (DP1.10107.001)

mms_metagenomeDnaExtraction.csv -> One record expected per dnaSampleID. A genomicsSampleID will represent only one extraction per plot/horizon combination and per collectDate (day of year, local time). Generally there will be only one extraction per genomicsSampleID (i.e. one record per collectDate (day of year, local time)), but in some cases multiple extractions will be necessary and will generate multiple **dnaSampleID**s for the same **genomicsSampleID**. Each record generates a single dnaSampleID, corresponding to the mms_metagenomeSequencing dnaSampleID. *Important Note*: The 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 metagenomics samples, filter the records in the ***mms_metagenomeDnaExtraction*** table to include only those with a **dnaSampleID** that is also contained in the ***mms_metagenomeSequencing*** table.

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 metagenomics samples, it may be helpful to filter the records in the mmg_soilDnaExtraction table to include only those with a value of 'metagenomics' or 'marker gene and metagenomics' in the variable **sequenceAnalysisType**.

mms_metagenomeSequencing.csv -> One record expected per **dnaSampleID**. Each record generates a single **dnaSampleID**, corresponding to the mms_metagenomeDnaExtraction **dnaSampleID**.

mms_rawDataFiles.csv -> Two records expected per **dnaSampleID**, one for the forward sequencing read and one for the reverse sequencing read. Ancillary files related to the raw data may also be provided in this table. Each record generates a single **dnaSampleID**, corresponding to the dnaSampleID in the upstream tables mms_metagenomeDnaExtraction and mms_metagenomeSequencing. One record per combination of **dnaSampleID** 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 Surface water microbe cell count (DP1.20138), in the table ***amc_fieldSuperParent***. Each **parentSampleID** may be subsampled into one **geneticSampleID** that undergoes microbial analyses, and an archive sample, described in the table ***mms_fieldSurfaceMicrobes*** within the Surface water microbe cell count product. These **geneticSampleIDs** are sent for DNA extraction (i.e. **geneticSampleID** from ***mms_fieldSurfaceMicrobes*** = **genomicsSampleID** in ***mms_swMetagenomeDnaExtraction***).

One or more **dnaSampleID**s are expected per **genomicsSampleID**, depending on the number of DNA extractions that occur on a sample provided to the lab. 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.

Table **mms_swMetagenomeSequencing** describes the sequencing preparation and analysis metadata. One record is expected per **dnaSampleID**.

Raw/minimally processed sequence data are available on the NEON data portal in the table **mms_swRawDataFiles**. In addition, 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 aberrations have occurred; *users should check data carefully for anomalies before joining tables*.

Surface Water Microbe Cell Count (DP1.20281.001)

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

Surface Water Microbe Field Data (DP1.20281.001)

mms_fieldSurfaceMicrobes.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 geneticSampleIDs, corresponding to the Surface water microbe metagenome sequences (DP1.10107.001) variable **genomicsSampleID** in the table ***mms_swMetagenomeDnaExtraction***.

Surface Water Microbial Metagenome Sequences (DP1.20281.001)

mms_swMetagenomeDnaExtraction.csv -> One record expected per dnaSampleID. Generally there will be only one extraction per genomicsSampleID (i.e. one record per collectDate (day of year, local time)), but in some cases multiple extractions will be necessary. Each record generates a single dnaSampleID, which corresponds to the **dnaSampleID** in the table ***mms_swMetagenomeSequencing***.

mms_swMetagenomeSequencing.csv -> One record expected per **dnaSampleID**. Each record generates a single dnaSampleID, corresponding to the mms_metagenomeDnaExtraction **dnaSampleID**.

mms_swRawDataFiles.csv -> Two records expected per **dnaSampleID**, one for the forward sequencing read and one for the reverse sequencing read. Ancillary files related to the raw data may also be provided in this table. Each record generates a single **dnaSampleID**, corresponding to the dnaSampleID in the upstream tables mms_swMetagenomeDnaExtraction and mms_swMetagenomeSequencing. One record per combination of **dnaSampleID** and **rawDataFileName** is expected.

3.9.2.2 Benthic Habitats The protocol dictates that each namedLocation sampled yields a unique **sampleID** per collectDate (day of year, local time) in Benthic microbe metagenome sequencing (DP1.20279.001), in the table ***amb_fieldParent***.

One or more **dnaSampleID**s are expected per **genomicsSampleID**, depending on the number of DNA extractions that occur on a sample provided to the lab. 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.

Table **mms_benthicMetagenomeSequencing** describes the sequencing preparation and analysis metadata. One record is expected per **dnaSampleID**.

Raw/minimally processed sequence data are available on the NEON data portal in the table **mms_benthicRawDataFiles**. In addition, 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 aberrations have occurred; *users should check data carefully for anomalies before joining tables*.

Aquatic Benthic Microbes Field Data (DP1.20279.001 and DP0.20270.001)

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

Benthic Microbe Metagenome Sequences (DP1.20279.001)

mms_benthicMetagenomeDnaExtraction.csv -> One record is expected per **dnaSampleID**. Generally there will be only one extraction per genomicsSampleID (i.e. one record per collectDate (day of year, local time)), but in some cases multiple extractions will be necessary. Each record generates a single dnaSampleID, which generally corresponds to the **dnaSampleID** in the table ***mms_benthicMetagenomeSequencing***, depending on the number of DNA extractions that occur on a single genomics sample provided to the lab. Duplicate records for an individual dnaSampleID should not exist.

mms_benthicMetagenomeSequencing.csv -> One record expected per **dnaSampleID**. Each record generates a single dnaSampleID, corresponding to the mms_metagenomeDnaExtraction **dnaSampleID**.

mms_benthicRawDataFiles.csv -> Two records expected per **dnaSampleID**, one for the forward sequencing read and one for the reverse sequencing read. Ancillary files related to the raw data may also be provided in this table. Each record generates a single **dnaSampleID**, corresponding to the dnaSampleID in the upstream tables mms_benthicMetagenomeDnaExtraction and mms_benthicMetagenomeSequencing. One record per combination of **dnaSampleID** and **rawDataFileName** is expected.

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

3.10 Special Considerations: Obtaining Sequence Data

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

3.10.1 From the NEON Data Portal

Information on raw sequence data are in the mms_rawDataFiles publication table, which is available by selecting the 'expanded' package during download. From this table, a URL listed in the NEON field **rawDataFilePath** provides the link to the raw sequence file. 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) Each raw data file may be gigabytes (GB) in size. Ensure you have sufficient space prior to downloading many files. *NOTE*: Due to the large file size, some of the files from a single sequencing run may be split into separate files (same file name appended with 'A', 'B', 'C', etc). Data from a single sample will be contained within the file associated with that record: merging of split files is not required. If merging was desired, however, simply concatenate the un-compressed files.
b) Downloaded files are typically in a compressed (.tar.gz or .gz) format. Files may require uncompressing prior to use.
c) Downloaded files may contain sequence data from an entire sequencing run, including data for non-target samples.
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 metagenomics 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 minimally processed metagenomics sequence data.

- 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 search functionality on the MG-RAST website.
- 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 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 NEON Raw Data Validation for Microbial Metagenomic Sequences (DP1.10107.001), provided with every download of this data product. Contained within this file is a field named 'entryValidationRulesParser', which describes syntactically the validation rules for each field built into the data ingest validation. Data entry constraints are described in Nicl syntax in the validation file provided with every data download, and the Nicl language is described in NEON's Ingest Conversion Language (NICL) specifications (AD[16]).

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

4.2 Automated Data Processing Steps

Metagenomics sequencing data are generated in batches of multiple samples and QA/QC is performed by the analytic facility. For each sample, minimum quality criteria must be met in order to accept the data for the sample. The general criteria include 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 in the metagenomeSequencing table.

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]).

4.3 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 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 metagenomics 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 arising during DNA extraction or 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

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.

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.

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.

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.