Metagenome and Metatranscriptome DMP template

**Data Type**

**Types and amount of scientific data expected to be generated in the project: Summarize the types and estimated amount of scientific data expected to be generated in the project.**

**Describe data in general terms that address the type and amount/size of scientific data expected to be collected and used in the project (e.g., 256-channel EEG data and fMRI images from ~50 research participants). Descriptions may indicate the data modality (e.g., imaging, genomic, mobile, survey), level of aggregation (e.g., individual, aggregated, summarized), and/or the degree of data processing that has occurred (i.e., how raw or processed the data will be)**

This project will produce metagenomic or metatranscriptomic data from <insert experiment-specific info regarding species and conditions>. Data will be collected from <insert number of samples> specimens, generating datasets totaling approximately <insert expected total data size: estimate ### per sample for MGX and ### per sample MTX> Gb in size. The data files produced in the course of the project includes:

1. raw FASTQ sequence files
2. adapter trimmed and host/microbial rRNA depleted FASTQ files
3. FASTA files of metagenome assembled genomes (MAGs)
4. FASTA files containing dereplicated/binned MAGs
5. tables of functional (genes and metabolic pathways) and taxonomic abundances in TSV or BIOM format
6. BAM alignment files from functional and taxonomic abundance profiling

**Scientific data that will be preserved and shared, and the rationale for doing so: Describe which scientific data from the project will be preserved and shared and provide the rationale for this decision.**

The following data produced in the course of the project will be preserved and shared:

1. Raw sequencing files
2. FASTA files for dereplicated/binned MAGs
3. Tables of gene family, metabolic pathway, taxonomic abundances in TSV or BIOM format

These files will allow other researchers to use the data and reproduce the analysis.

**A brief listing of the metadata, other relevant data, and any associated documentation (e.g., study protocols and data collection instruments) that will be made accessible to facilitate interpretation of the scientific data.**

To facilitate interpretation of the data, experimental metadata describing the experimental condition for each sample as well as relevant information associated with the samples will be included. The metadata will be included in tab separated file (TSV). <Insert list of experiment-specific metadata categories here such as the experimental conditions, batches, and/or relationships between the samples>.

**Related Tools, Software and/or Code**

**State whether specialized tools, software, and/or code are needed to access or manipulate shared scientific data, and if so, provide the name(s) of the needed tool(s) and software and specify how they can be accessed**

Raw sequencing files in FASTQ format will be made available. FASTQ files do not require any specialized software to access. Illumina adapters and host and microbial rRNA sequences are removed using KneadData. The filtered read files then undergo taxonomic and functional classification using MetaPhlAn and HUMAnN. These abundance tables will be made available in both BIOM and TSV file formats. BIOM files can be manipulated using the biom software package or numerous package for the R and Python programing languages. The TSV files require no specialized software to access. Any BAM alignment files generated from MetaPhlAn or HUMAnN can be accessed with a variety of open source, UNIX command line tools. These filtered reads are assembled using MEGAHIT to generate the initial metagenome assembled genomes (MAGs). The MAGs are dereplicated using dRep and MetaBAT2. Genes and functional annotations are extracted from the dereplicated MAGs using DRAM. Called genes and associated annotations are stored as FASTA format files that can be manipulated without any specialized tools. The dereplicated MAGs are assigned a taxonomic classification and the taxonomic abundances quantified using MetaWrap. Metadata detailing relevant sample information will be made available in TSV format and will not require the use of specialized tools to be accessed.

**If applicable, specify how needed tools can be accessed, (e.g., open source and freely available, generally available for a fee in the marketplace, available only from the research team) and, if known, whether such tools are likely to remain available for as long as the scientific data remain available.**

Listed below are the tools and versions used are used in this analysis. All of the following tools are freely available:

|  |  |  |
| --- | --- | --- |
| **Tool**  | **Version**  | **URL**  |
| KneadData | 0.7.4 | https://github.com/biobakery/kneaddata |
| MetaPhlAn | 3.0.5 | https://github.com/biobakery/MetaPhlAn |
| HUMAnN | 3.0.1 | https://github.com/biobakery/humann |
| MEGAHIT | 1.1.3 | https://github.com/voutcn/megahit |
| MetBAT2 | 2.12.1 | https://bitbucket.org/berkeleylab/metabat/src/master/ |
| dRep | 2.6.2 | https://github.com/MrOlm/drep |
| MetaWrap | 1.2.1 | https://github.com/bxlab/metaWRAP |
| DRAM | 1.2.2 | https://github.com/WrightonLabCSU/DRAM |

**Standards**

**State what common data standards will be applied to the scientific data and associated metadata to enable interoperability of datasets and resources, and provide the name(s) of the data standards that will be applied and describe how these data standards will be applied to the scientific data generated by the research proposed in this project. If applicable, indicate that no consensus standards exist**

**Data Preservation, Access and Associated Timelines**

**Repository where scientific data and metadata will be archived: Provide the name of the repository(ies) where scientific data and metadata arising from the project will be archived; see Selecting a Data Repository)**

**How scientific data will be findable and identifiable: Describe how the scientific data will be findable and identifiable, i.e., via a persistent unique identifier or other standard indexing tools.**

**When and how long the scientific data will be made available: Describe when the scientific data will be made available to other users (i.e., no later than time of an associated publication or end of the performance period, whichever comes first) and for how long data**
**will be available.**

**Access, Distribution, or Reuse Considerations**

**Factors affecting subsequent access, distribution, or reuse of scientific data: NIH expects that in drafting Plans, researchers maximize the appropriate sharing of scientific data. Describe and justify any applicable factors or data use limitations affecting subsequent access, distribution, or reuse of scientific data related to informed consent, privacy and confidentiality protections, and any other considerations that may limit the extent of data sharing. See Frequently Asked Questions for examples of justifiable reasons for limiting sharing of data.**

**Whether access to scientific data will be controlled: State whether access to the scientific data will be controlled (i.e., made available by a data repository only after approval).**

**Protections for privacy, rights, and confidentiality of human research participants:**
**If generating scientific data derived from humans, describe how the privacy, rights, and confidentiality of human research participants will be protected (e.g., through de- identification, Certificates of Confidentiality, and other protective measures).**

**Oversight of Data Management and Sharing**

**Describe how compliance with this Plan will be monitored and managed, frequency of oversight, and by**
**whom at your institution (e.g., titles, roles).**