Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1S050

Title: KBase Educators: Microbiome Workforce Development Program

Creator: Elisha Wood-charlson - **ORCID:** 0000-0001-9557-7715

Affiliation: Lawrence Berkeley National Laboratory (lbl.gov)

Principal Investigator: Adam Arkin

Project Administrator: Ellen Dow

Contributor: Elisha Wood-Charlson

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: NSF 22-522

Grant: 2316244

Template: NSF-BIO: Biological Sciences

Project abstract:

Undergraduate educators strive to incorporate useful and meaningful skills that engage their students in current research and prepare their students for jobs in science. However, access to resources that support training in microbiome research limit the ability of many institutions to effectively teach these skills. The KBase Educators: Program for Microbiome Workforce Development Research Coordination Network (RCN) Incubator is focused on developing a more

inclusive and accessible microbiome research curriculum that enables a diversity of institutions to train students with the skills necessary to successfully enter the workforce. In addition to teaching resources, the RCN will bring together a peer support network that reaches across five types of institutions, with the goal of forming equitable partnerships that allows each educator to contribute meaningfully to develop a comprehensive microbiome research-based curriculum that accurately reflects realistic access to resources and standards.

Participating educators will form working groups to develop a modular curriculum with the guiding question of how microbiomes respond to climate change across ecosystems and potential impacts for humans that culminates in student-led data publications. The RCN will support a workshop where educators can co-create and trial-run modules with their peers, and receive training from partner organizations, such as the Agricultural Microbiome RCN, National Ecological Observatory Network, and National Microbiome Data Collaborative. This will prepare them to pilot the modules in their classrooms, across a range of institutions, and to collect feedback and outline recommendations to scale the program in subsequent years. The goal of training an inclusive workforce is to enable students to become the next generation of researchers capable of answering current questions and generating open data that will address the grand challenges of our future. This project is being jointly funded by the Directorate for Biological Sciences, Division of Biological Infrastructure, and the Directorate for STEM Education, Division of Undergraduate Education as part of their efforts to address the challenges posed in Vision and Change in Undergraduate Biology Education: A Call to Action (http://visionandchange/finalreport/).

This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

Start date: 06-15-2023

End date: 05-31-2024

Last modified: 06-07-2023

the creator(s) endorse, or have any relationship to, your project or proposal

---

KBase Educators: Microbiome Workforce Development Program

Data and Materials Produced

Describe the types of data, physical samples or collections, software, curriculum materials, and other materials to be produced in the course of the project. (For collaborative proposals, the DMP must cover all the various data types being collected by each collaborator.)

Physical samples and sample collection protocols: As part of the initial program testing, instructors may choose to include a module on sample collection. During the workshop, teachers will be trained in robust sample collection methodology by National Ecological Observatory Network (NEON) staff. Collection protocols will be posted to the KBase protocols.io group: https://www.protocols.io/workspaces/doe-kbase, and given a Digital Object Identifier (DOI). All samples will be collected alongside robust metadata, leveraging the Genomics Standard Consortium's (GSC) Minimum Information about x Sequence (MIxS) environmental packages, as appropriate to the samples being collected (i.e., soil, water, etc). During the RCN workshop, educators will also be trained in sample metadata collection using the GSC MIxS templates by the National Microbiome Data Collaborative (NMDC) staff. MIxS sample metadata fields are compliant with the iSamples System for Earth Sample Registration (SESAR) metadata fields, so all samples will be registered for International Geo Sample Number (IGSN). IGSNs tracking and management recently moved to DataCite, so we will continue to watch for changes/updates to the metadata schema/fields as that integration solidifies.

Sample processing protocols: The modules also include sample processing protocols, which will be standardized based on feedback during the workshop around equity and accessibility with respect to supplies and laboratory instrumentation. All sample processing protocols will be posted to the KBase Protocols.io group and given a DOI.

Genomics and metagenomic sequence data: All raw genomic and metagenomic sequence data will be 150bp, paired-end Illumina short-read (~400 Mbp/2.7 Million reads), which will be uploaded directly into a shared KBase Google drive as FASTQ files by the SeqCoast sequencing facility. We anticipate 10 sequencing runs to be funded by this incubator. We are capable of handling many more, if additional data can be generated. Instructors and students then import raw FASTQ sequence data into KBase for analysis (details below). Upon completion of the course, raw sequence data and sample metadata will be archived in NCBI Sequence Read Archive (SRA) under a BioProject specific to the yearly cohort's science questions.

Data analysis workflows: KBase user interface supports reproducible analysis workflows. Instructors have a variety of KBase analysis modules available (see proposal Figure 1). In addition, KBase has an analysis workflow based on the American Society for Microbiology (ASM) Microbiology Resource Announcements (MRA) for isolates or metagenomes (e.g., isolate - https://kbase.us/n/122280/22/). This workflow has been approved by the MRA editorial staff, and successfully used by students to publish data in MRA (McLoon et al. 2022).

All workflows in KBase can be made public - open to the KBase community after login - and "published", which generates a static HTML snapshot of the analysis workflow that is available outside the login and index by search engines. This "static Narrative" gets a DOI, which is included in the MRA data availability statement, as well as the reference section. This connects the data set directly to the publication, and enables students to generate and share FAIR datasets for the broader research community.

Standards, Formats and Metadata

Describe the standards to be used for all the data types anticipated, including data or file format and metadata. [Note: Where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies.]

integrated in other sections

Roles and Responsibilities

Describe the roles and responsibilities of all parties with respect to the management of the data (including contingency plans for the departure of key personnel from the project).

Responsibilities and Resources: Adam Arkin (PI) will be responsible for overseeing the project. Ellen Dow (Co-PI) will be responsible for designing, collecting, and analyzing student surveys and coordinating the implementation of protocols with participants and associated IRB. Elisha Wood-Charlson (Co-PI) will be responsible for FAIR data - sharing and publication.

Dissemination Methods

Describe the dissemination methods that will be used to make data and metadata available to others during the period of the award, and any modifications or additional technical information regarding data access after the grant ends.

All genomes and metagenomes funded by the RCN will be published by students as MRA data sets, with the incubator covering most of the publication costs for 10 sequencing runs. In addition, all data generated by students as part of the KBase Workforce Development Program will be stored inside a KBase Organization, as a way to build a collection of student-generated genomes/metagenomes. All data are searchable inside the KBase platform, but an Organization allows for aggregation of a collection of data in a more user-friendly, browsable format.

Curriculum materials: Novel curriculum materials generated by the RCN will be made more broadly available to the community via Quantitative Undergraduate Biology Education and Synthesis (QUBES) platform (https://qubeshub.org).

Policies for Data Sharing and Public Access

Describe the PI's policies for data sharing, public access and re-use, including re-distribution by others and the production of derivatives. Where appropriate, include provisions for protection of privacy, confidentiality, security, intellectual property rights and other rights.

Program assessment, student assessment: Demographics from participating institutions, qualitative data around module implementation and effectiveness, as well as instructor and student feedback, will be collected through surveys and interviews during the 2023-2024 school year. All survey and interview protocols will be approved by the Institutional Review Board prior to Fall 2023. Surveys will be anonymous and administered through an online platform (Google Form), and managed in Google Sheets within a protected Lawrence Berkeley National Laboratory Google Drive account. For students, questions will focus on self-assessment on learning and confidence, as well as accessibility of material and relevance to their professional and personal goals. Questions will be answered using a 5-point Likert scale. For instructors, formal feedback will be collected also using anonymous online surveys, alongside informal interviews and group discussions. Informed consent will be obtained from all instructors.

Archiving, Storage and Preservation

Where relevant, describe plans for archiving data, samples, software, and other research products, and for on-going access to these products through their lifecycle of usefulness to research and education. Consider which data (or research products) will be deposited for long-term access and where. (What physical and/or cyber resources and facilities (including third party resources) will be used to store and preserve the data after the grant ends?)

This data management plan will be submitted to DMPtool and assigned a DOI. All research outputs from the RCN will be connected to the DMP DOI using relationship_type standard vocabulary, as defined by DataCite. We are leveraging the data citation framework, as published by Co-PI Wood-Charlson et al. (2022, https://doi.org/10.1371/journal.pcbi.1010476), to establish a network of persistent identifiers that connects all research outputs to a funded project.

All research outputs will be deposited in the places listed above, but will have a centralized website landing page on the www.kbase.us website. KBase has an extensive website and documentation to support the KBase Educators program, so we will expand on that site with additional resources for the Workforce Development Program.
