
Novel approaches for assessing biological indicators of soil health

A Data Management Plan created using DMP Assistant

Creator: Kari Dunfield

Principal Investigator: Kari Dunfield

Data Manager: Kari Dunfield

Affiliation: University of Guelph

Template: University of Guelph Agri-Food Funding Template

ORCID ID: 0000-0003-1316-0452

Project abstract:

Soil are increasingly recognized as a critical player in the fight against climate change, leading to mounting interest in regenerative agriculture and soil health. A limitation of the focus on soil health as an endpoint, is the lack of an accurate, validated biological indicator of soil health. Using multi-omics approaches to identify key microbial taxa, we will map agricultural ecosystem dynamics using a microbial network approach. We propose to identify keystone taxa in Ontario corn and soybean based cropping systems that will provide a target for microbial screening. These studies can be further integrated into a global database systematizing different outcomes, environmental conditions, cropping systems, soil types and growing seasons. This study will also include a review of current biological indicators of soil health, a cost-analysis of genomic approaches, and a KTT strategy that will focus on teaching farmers and Certified Crop Advisors how to interpret measures of soil biology.

Identifier: 7036

Last modified: 01-03-2022

Grant number / URL: UG-T1-2021-101041

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Novel approaches for assessing biological indicators of soil health

Data collection

Provide an overview of the data that will be generated, collected or acquired to support this project. If data will be acquired from a third party, specify the source.

Soil diversity can be characterized using genomic tools but quantitative linkages to predict soil health remains difficult. To address this knowledge gap, this research program will:

- (1) Assess the impact of sustainable management practices on soil biology – including bacteria, fungi and mesofauna;
- (2) Use network approaches to model inter-kingdom (bacteria-fungi-mesofauna) interactions in soil communities, and to identify keystone indicator species that relate to measures of soil health, system productivity and resilience.
- (3) Establish a soil sampling workflow and bioinformatics pipeline that can be used to identify microbial groups that are predictors of soil health
- (4) Work with industry partner (Biomemakers Inc.) to demonstrate to stakeholders the potential to use genomic analysis of soil communities as an indicator of soil health on Ontario farms.
- (5) Work with Soils@Guelph and knowledge advisors from GFO and OMAFRA to transfer knowledge about using genomic tools to understand soil biology and predict soil health.

Numeric data will be collected for Objectives 1 & 2

No data will be collected for Objectives 3-5.

What method(s) of data collection will be employed?

Primary data collected will be dna sequences aquired from MiSeq analysis of soil metagenomic DNA

What types of data will be included?

Primary data collected will be dna sequences aquired from MiSeq analysis of soil metagenomic DNA

What software or digital formats will be used to collect, manage and analyze the data?

Bioinformatic and statistical platforms.

QIIME2 pipeline v. 2019.1

FUNGUILD (Nguyen et al. 2016).

R programming environment

Gephi 0.9.2. (Bastian et al., 2009)

Provide an indication of the scope of the data?

Approx 1000 PCR amplicons will be sequenced.

Data storage

Estimate the size of data storage that will be required.

Approx 15 GB data

Where will your data be stored during the collection, collation and analysis phases of the project?

Data will be collected at the sequencing facility - Genome Quebec, and transferred to a password protected lab desktop computer. Analysis is done through remote access to the lab computer.

What backup strategy will be employed?

Analysed and raw data is backed up in a sharepoint site, and also in secondary storage through the U of G.

How will your data files be organized? What file naming conventions will you use? A brief overview or example would be adequate.

For genomic DNA sequencing data, we will use the naming conventions of the NCBI SRA database (<https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/>). The SRA publicly accessioned objects are: BioProject accession (PRJNA #), STUDY (accession in the form of SRP #), EXPERIMENT (SRX #), SAMPLE (SRS #), RUN (SRR #). For instance, a BioProjec will be created, including a title and a code provided by SRA system, and it including three Studies, one for each sequencing data set (by kingdoms Bacteria, fungi, mesofauna). Each experiment and sample will be self-explanatory (treatment + replicate number + library + sequencing strategy + layout + instrument model).

What metadata will be developed for your data? Will there be supplemental documentation prepared to assist with the interpretation and analysis of your data?

Metadata are stored in excel spreadsheets. Each column in the spreadsheet will have a simple text heading. Units of measurement will also be noted. The metadata to be uploaded public repositories (e.g., NCBI SRA database) will include a detailed explanation about the groupings of samples per experimental conditions (i.e., treatments, replicates) and origins, and also describing

the sequencing technologies, geographic coordinates, etc.

Data archiving and preservation

Will you deposit your data in the UG data repository or an external data repository? If you are opting to not archive your data in a repository, where will your data be housed after completion of your project?

Genome sequences will be published in public NCBI databases.

Soil metadata generated in this project will be published as supplemental material in publications.

Discuss any data transformations that will be needed so your data is preserved in appropriate, non-proprietary formats.

The raw data (demultiplexed paired-end sequences) will be uploaded to the SRA repository as received from sequencing facilities (i.e, demultiplexed (forward and reverse) sequences. These sequences will be in conventional "Cassava" format, compressed (e.g., SO_7139_5_105_S46_R1_001.fastq.gz; SO_7139_5_105_S46_R2_001.fastq.gz). For data analysis, including public access, the mate-pair sequences (R1 and R2) need to be denoised and merged according to the strategy selected by the researcher.

If some of your data will not be preserved, how long will you retain it? Will the non-preserved data be destroyed?

Non-published raw data will be retained for a minimum of 5 years after students finish publishing thesis and then destroyed.

Sharing and reuse

Will the data that you archive in a data repository be made available for sharing and reuse by other researchers?

Next-generation sequencing reads are submitted to the publicly accessible NCBI database

Explain which version of your data or subset of your data will be shared.

Raw sequence reads will be submitted to the sequence read archive of NCBI.

When will your data be available for discovery by other researchers? Will you impose an embargo on publication of your data? If so, please provide details on the duration of the embargo.

The data will be available after the study has been published. Accession numbers for the sequence reads will be published in each paper.

Will you limit who can access your data? If so, who will that be and why are you limiting the data's reuse?

Final genomics data will be openly available through NCBI.

Are there specific license terms you will assign to users of your data?

N/A

Restrictions/limitations

Are there limitations or constraints on how you manage your data resulting from legal, ethical or intellectual property concerns?

N/A

Would your data need to be anonymized or de-identified before being shared with others?

No.

Confidential information

What information do you want to include in your DMP that should not be publicly shared?

N/A