Lake Sampling DMP

Types of data

Data acquired during this proposal includes field data, sequencing data, analytical data, imaging data, and data from biological experiments. Field data will be recorded by hand and translated to spreadsheets. Analytical data will include ASCII or CSV output files and spreadsheets. Imaging files (jpeg or tiff) encompass the images themselves, and metadata files describing imaging conditions. Data from biological experiments will include what is recorded manually in notebooks, and analytical and imaging data.

Amplicon and metagenomic sequence data will be produced using an Illumina HiSeq 2500 instrument. The total size of this data is expected to be approximately 200 Gigabyte. All file formats used in this proposal will be in flexible, text-based, non-propriety format. Raw data files generated by high throughput sequencing machines are in FASTQ file format. Raw sequence data can be assembled resulting in another FASTQ file or aligned to an assembly resulting in a Sequence Alignment/Map (SAM) alignment formatted file.

Data and metadata standards

Field data will be metadata about time and conditions of sampling. The output files for analytical instruments and imaging software contains metadata about the specific conditions under which samples will be run or images. Files will be named with the appropriate ISO 8601, the international standard for data and time (e.g. 2015-08-16) along with specific sample and analytical identifiers.

ISUGIF utilizes DokuWiki as an online notebook to document data analyses and all information pertaining to the project in a centralized location. DokuWiki is a simple to use, highly versatile and Open Source wiki, where access granted to collaborators via login and passwords. A typical wiki page entry includes a detailed project description, the download location of the original data and how it was transferred, an initial original data characterization (size, number of reads), the location and version of all accessory data required to perform analyses (database version and download location, etc.), and a logical progression of commands that steps the original data through the analyses to final products (raw reads to assembly, annotation etc). With this documentation it is straightforward to reproduce or replicate our analyses since all methods, exact commands and versions of programs used are documented.

Policies for access and sharing

Physical and geochemical data from sampling or direct analysis of environmental samples will be made available through publication of supplementary datasets. In the case the journal is not open access, these datasets will be made freely available by uploading to author profiles on self-curations sites such as Research Gate. The data will be in spreadsheet formats.

Upon project completion, all sequenced amplicons, metagenomes, and associated manuscripts will be deposited into the National Center for Biotechnology Information's Sequence Read Archive (SRA), which is a public, open access genomic database. Additionally, the amplicon 16S/18S rRNA sequences will be submitted to GenBank and the Earth Microbiome Project (http://www.earthmicrobiome.org/).

Policies and provisions for re-use, re-distribution

Data will be made available immediately upon publication. In the case of data acquired in the laboratory of collaborators in Germany (specifically isotope data), any datasets acquired for sampling periods that are published will be made available, but they retain the right to any data for sampling trips where additional data outside the scope of this project, etc. are collected and analyzed.

The sequencing data will be deposited as soon as the manuscripts are published, or within two years of the project finishing, in the case that additional data exists, but is not a part of the intended manuscripts.

Plans for archiving and preservation of access

Field, analytical, imaging, and experimental data will be backup up on personal hardrives, and also on ISU's Research Files service. Prior versions of files can be accessed for up to 12 weeks, based on a system of 31 checkpoints. Data is accessed through each individual's network drive on-campus, and through a VPN while off-campus.

Sequence data is redundantly backed up usingRAID 6 storage boxes that are on a private IP address and therefore only accessible to people with accounts and passwords on the ISU campus or through the campus Virtual Private Network (VPN). Additionally, only members of the Genome Informatics Facility group will have access to the folder containing the raw and analyzed data during the exploratory to analysis phases of the project. Data will be provided to other members of the project team through password-protected access to folders containing the data.

Raw data will be managed by the Genome Informatics Facility (GIF) and stored on redundantly backed up 144 Terabyte RAID boxes maintained by the High Performance Computing (HPC) facility on campus. Raw data will be backed up immediately upon receipt. Scripts that generate analyzed data will be backed up daily and analyzed data will be backed up weekly. This model ensures rapid data recovery after the unlikely catastrophic loss of the primary RAID box. An additional copy of the raw data along with the scripts used to generate results will be kept in the GIF, which is located in a separate building from the HPC facility.

GIF has an archive server with 132 TB of storage. This machine serves as tertiary backup of all raw data and scripts that generate data analyses. This secondary site backup ensures the safety and integrity of the raw data and analyses in case of catastrophic failure at the primary site of data analysis in the High Performance Computing Facility at Iowa State University.