IEP Data Management Plan

Basic Information

Year: 2020; PEN: 344; Date Updated: 2019-10-18; Start Date: 2019-12-01

Study Title

Developing an eDNA metabarcoding protocol to improve fish and macroinvertebrate monitoring in the San Francisco Estuary

Principal Investigator

Individual(s) responsible for the project. Include name, agency, e-mail, & phone.

Andrea Schreier, Adjunct Assistant Professor, University of California, Davis, amdrauch@ucdavis.edu, 530 752 0664

Point of Contact

Individuals who data users should contact for access to the data or questions about the data. Include name, agency, e-mail, & phone number or write "same as above."

Ravi Nagarajan, Assistant Project Scientist, University of California, Davis, rpnagarajan@ucdavis.edu, (530) 752-0175

Data Description

A very brief description of the information to be gathered; the nature and scale of the data that will be generated or collected. Include approximate size (in MB) of the resulting data set.

1) Reference sequence database (~15 MB)

2) Sampling metadata (~0.5 MB)

3) eDNA metabarcoding sequencing data (~1000000 MB)

Related Data

Optional. Existing datasets that you incorporate into analysis and reporting for this program element, existing data that are relevant to your study, or data that are collected simultaneously.

Not Applicable

Metadata

A description of the metadata to be provided along with the generated data, including the metadata standards used. Provide the file name and information on how users can access the metadata (e.g., a link).

All sampling dates and locations will be recorded. Water quality parameters such as turbidity, temperature, pH, DO, and salinity will be measured with a YSI and recorded. Microhabitat variables such as presence of vegetation or depth will be recorded. These data will be digitized in an Excel spreadsheet (~500 KB), archived on Box and the 1 TB external hard drive, and made available on Data Dryad. Species detected and numbers of individuals captured for each conventional survey will be accessed online (e.g. EDSM; rotary screw trap:) or obtained from the Suisun Marsh Fish Study, either in person from T. O'Rear or at https://watershed.ucdavis.edu/project/suisun-marsh-fish-study.

Storage and Backup

A description of the short-term storage methods and backup procedures for the data, including the physical and electronic resources to be used for the short-term storage of the data.

Raw Sanger sequence data (~4 MB) and the reference sequence database (up to ~10 MB) will be archived in the online platform Box, on Data Dryad, and backed up on a 1 TB external hard drive. Metadata will be archived on Box and the 1 TB external hard drive and made available on Data Dryad. Raw sequence reads from the Illumina MiSeq will be stored on the UC Davis Genome Center computing cluster, will be backed up on two 1 TB external hard drives stored in different locations, and will be made available on Data Dryad.

Archiving and Preservation

The procedures for long-term archiving and preservation of the data, including succession plans for the data should the expected archiving entity go out of existence.

All data and metadata described below will be shared and archived on Data Dryad (https://datadryad.org/) or a similar open access data sharing platform. Except for raw Illumina data, all data will be backed up on 1 TB external hard drives as well as the online platform Box if possible. All analysis files (~ 3 GB) will be stored on the UC Davis Genome Center computing cluster, archived on Box, and backed up on a 1 TB external hard drive. Raw Sanger sequence data and the reference sequence database will be archived in the online platform Box, on Data Dryad, and backed up on a 1 TB external hard drive.

Access and Sharing

A description of how data will be shared. Include (1) access procedures, (2) embargo periods, (3) technical mechanisms for dissemination (e.g., website addresses, listserv information), (3) whether access will be open or granted only to specific user groups, and (4) a timeframe for data sharing and publishing.

All Data will be made fully public and accessible within three months of project completion. All data and metadata described below will be shared and archived on Data

Dryad (https://datadryad.org/) or a similar open access data sharing platform. All co-PIs will agree upon the format of data and metadata collection and storage and all support personnel will be trained in data collection methodology. QA/QC will be performed to ensure all data is recorded accurately. These data and metadata, along with our reference sequence database, will be uploaded to Data Dryad. Links to datasets will be made available on the GVL website (gvl.ucdavis.edu) and in open access publications for use by San Francisco Estuary (SFE) resource managers, interested researchers, and the general public.

Format

Formats in which the data will be generated, maintained, and made available. Include BOTH general data type (e.g., spreadsheet, relational database) and file format (extension).

All data will be machine readable and the general data type is "spreadsheet". Sanger sequencing data will be plain text files (.txt). Illumina sequencing reads will be in the FASTQ format. Each entry in a FASTQ files consists of 4 lines: 1) A sequence identifier with information about the sequencing run and the cluster. 2) The sequence (the base calls; A, C, T, G and N). 3) A separator, which is simply a plus (+) sign. 4) The base call quality scores. These are Phred +33 encoded, using ASCII characters to represent the numerical quality scores.

Quality Assurance

Brief description of procedures for ensuring data quality. Provide links to Quality Assurance Project Plan and/or QA/QC Standard Operating Procedures.

All Sanger sequence data will be manually QC'd. For Illumina sequencing data, quality control statistics for MiSeq reads and bioinformatic pipelines will be thoroughly described in reports and publications. Analyses will be conducted with programs freely available on the internet (e.g. QIIME 2 or R). For filters/genetic samples, each sample will be entered into the GVL ItemTracker archive database with unique identifiers that will stay with the sample throughout the analysis process. The exact location of each filter and DNA extract will be tracked with ItemTracker. Environmental metadata collected during eDNA surveys will be checked against environmental metadata collected by concurrent trawl surveys.

Rights and Requirements

A link to or instructions to locate the agency's rights and requirements for data use.

Not Applicable