**DATA MANAGEMENT PLAN**

**Types of data:**

**Question 1: How can we measure and compare changes in non-stationary networks?**

This project task generates no novel data, but will yield R packages to implement analyses or graphics. Data from the study will be built into these packages as worked examples.

**Question 2: How do gene co-expression networks change in reintroduced populations?**

Data for this question will be obtained from wild-caught stickleback from our experimental source and destination lakes.

1. Illumina sequence reads from 3’TagSeq library preparations for transcriptomics.
2. Transcript relative abundance calculated from (a), for 30 fish per lake (8 source lakes and 8 destination lakes) each year for 8 years (including archived samples from 2019-2021). These data will be collected for pronephros only in 2019-2020, and for pronephros and liver in 2021-27.
3. Gene co-expression network structure in each sample.
4. We will record covariates for all sampled stickleback, including standard length, mass, sex, condition, diet, and infection load for common macroparasites.

**Question 3: Are changes in network architecture parallel among replicate populations?**

This question will draw on data generated for Question 1 and 2

**Question 4: To what extent are co-expression network changes genetic or plastic?**

Data for this question will come from three sources. (4a) will use data from Question 2, plus ddRADseq to assign individuals’ ancestry proportions. (4b) will use transcriptomic data from a transplant experiment into G Lake from four source lakes, to evaluate transcriptomic plasticity. (4b) will entail transcriptomic data from lab-reared stickleback from source lakes to control for environmental effects via common-garden rearing. The data will entail:

1. Illumina sequence reads from 3’TagSeq library preparations for transcriptomics.
2. Tables of transcript relative abundance: rows are individual fish, columns are genes, and entries are observed number of sequence reads for a given gene in a given individual fish.
3. SNP genotypes obtained from 3’TagSeq sequence reads for ancestry estimation (4a).
4. In addition to data from Question 1 (reused in 4a), for (Q4b) we will obtain expression data from 50 fish from each of four source lake populations, and 240 transplanted fish (4 enclosures, 2 time points, 40 fish per time point per enclosure), two tissues per individual. For (4c) we will generate tables of transcript relative abundance calculated from 8 source lake genotype fish reared in the laboratory (5 fish from each of 5 families per pure lake cross). In addition we obtain similar data for 4 between-lake F1 hybrid combinations (5 fish from each of 5 families per cross).
5. Gene co-expression network structure in each sample.
6. We will record covariates for all sampled stickleback, including standard length, mass, sex, infection load for common macroparasites (except for common-garden fish used in (4c)).

**Question 5: Are gene expression networks, and species interaction networks, co-evolving?**

In the final stage of this project, we track changes in species interaction networks and relate these to changes in gene expression networks (using data from Question 1). The species interaction network data will entail the following kinds of data:

1. Stomach contents counted and identified to the lowest taxonomic level, from 100 stickleback from each of the source and destination lakes in each year.
2. Macroparasite infections (count and identity) from 100 stickleback from each of the source and destination lakes in each year.
3. 16S gut microbiome sequence data from 50 stickleback from each of the source and destination lakes in years 3 and 5 of this project, supplemented with data obtained separately by Senior Personnel Kathryn Milligan-Myhre using her start-up funds at UConn. This data takes the form of read counts of microbial OTUs per individual fish.

**Question 6: Does gene network structure (or, changes) predict genomic targets of selection?**

In addition to the transcriptomic data from Question 1 and 2, this part of the study will entail poolseq data for allele frequency estimates in the F6 generation (year 4 of this funded project) to estimate allele frequency changes in each destination lake, and each source lake. The data will entail:

1. Illumina sequence reads from pools of 100 fish per lake (8 source lakes and 9 destination lakes in 2022, 2025, plus already-existing data from the 8 source lakes in 2019).
2. Tables of allele frequencies, by population, calculated from the data in (a)
3. Estimates of allele frequency changes in each population for each SNP or indel, and corresponding estimates of selection strength.

**Standards for data and metadata format:** All data above will be recorded in a relational database (SQL server) with separate tables for each type of data, linked via keys for fish identity, population, year, sample date, sample location. The database will contain pointers to the locations and names of raw data files (e.g., Illumina sequence outputs) and processed data (e.g., SNP genotype inferences). Each table within the SQL database will be accompanied by a text file detailing the names of each column, its meaning and intended use, and where applicable explaining methods to obtain the data in a given column, including measurement units.

**Policies for data sharing:** We will share all data in public data repositories (e.g., Dryad) upon submission for publication. We will include R code for any analyses alongside the data tables, to facilitate replication of any particular results. Illumina sequence data will be archived on the Short Read Repository, gut microbiome data on the Earth Microbiome Project repository, and all other data deposited in Dryad. All raw files will be accompanied by metadata following the standards described above. Data generated from this project will also be included as tutorials in R packages or other software developed for this project.

**Policies for data re-use and re-distribution:** We encourage colleagues to re-use data generated by our work for their own analyses and publications, either using our raw files to extract their own data, or using our extracted data for re-analysis or meta-analysis. We commit to sharing files, data, and statistical or graphical code in a timely manner upon request. We do not expect co-author status on papers arising from this data re-use, except where we become deeply involved in the conceptual design of the derived study, analysis, or writing.All DNA or RNA samples will be retained at -80C , with RNA stored as tissue in RNA later, or extracted RNA in Trizol buffer. An aliquot of DNA or RNA samples will be available upon request from the PI responsible for those samples, as long as supplies last.

**Plans for data archiving:** All genetic and genome sequencing data sets, as well as any genotype-phenotype datasets will be deposited in public databases upon publication. All data used in graphics or statistical analyses will be archived on Dryad or similar repositories on publication.

**Data backup:** All files will be backed up in real-time, both on cloud storage (GitHub) and monthly on hard drives kept in separate buildings from the computers holding data for analysis, to ensure the data cannot be lost or damaged without having a recoverable copy in another location.

**Protocol sharing:** All laboratory protocols used in this project will be written in detail as instruction files on Protocols.io to facilitate protocol sharing and retention of key methods.