8-30-2019

Collaborative Research: RoL: Using reaction norms to link genomic and phenotypic variation with regional-scale population responses to environmental change

Peter Adler
Utah State University, peter.adler@usu.edu

Data Management Plan

The proposed research will produce four types of data, Genomics, Phenology and Physiology, Demography, and Environmental, along with extensive computer code. Lead PI Adler will take responsibility for ensuring that this Data Management plan is followed, and will directly supervise handling of the Demography and Environmental data, with the assistance of a postdoc and technician at USU. PI Lasky will be responsible for the DNA sequencing data, PI Germino will handle the phenology and physiology data, and PI Hooten will oversee the management of the computer code. For each type of data, we describe the structure of the data and our digital storage plan. We then present our approach for long-term archiving and distribution.

1. Data and code description and storage

1.1. Germplasm and genomic data

The principal types of data we will generate are raw Illumina reads from RNAseq used to *de novo* assemble a cheatgrass transcriptome and from sequencing exome capture libraries. From these raw data we will generate a transcriptome assembly, read mappings, SNP calls, and an inferred kinship matrix. Raw reads will be stored and archived in NCBI's Sequence Read Archive (SRA) format, which is a binary compressed format. Appropriate metadata will be stored with each dataset. FASTQ data are easily retrieved from SRA files. Processed and aligned read data will be available in the widely used BAM format. Other results of analysis will be in comma-delimited text formats which are easily parsed and/or imported into spreadsheet software.

All raw data will be archived to assure no loss of information pre-publication. Raw digital data will be stored on RAID arrays in the Lasky lab at PSU, in addition to automated weekly backups to University cloud storage. Smaller raw datasets will be stored on Dropbox cloud servers, with monthly archiving.

Sequenced cheatgrass accessions will be stored and disseminated as seeds to the USDA GRIN germplasm bank so that any researcher in the world can study the same inbred lines.

1.2. Phenology and physiology data

The core common garden experiments will generate data on phenology (weekly observations of developmental stage) and physiology (cold tolerance, photosynthetic rates, and morphological functional traits) of known genotypes (see 1.1). Data will be managed by the technician based in Boise (yrs 1-2) and then by the USU postdoc (yrs 2-4). These data will be stored in flat .csv files, with each record (line) corresponding to one measurement on one individual plant, at one particular field site. A separate table will link each individual plant identifier to its DNA sequence identifier. Each data file will be accompanied by a metadata file (raw text). All raw data and metadata will be stored on the cloud in Google Drive, with additional weekly back up on local hard drives and external disc drives. Once QAQC is complete, the data files will be added to the Github repositories where we will store our computer code. This will also be the version of the data available to all BromeCast participants.

1.3. Demographic data

The core common garden experiments and the satellite experiments will generate data on the emergence, survival, seed production, and biomass of individual plants with known genotypes (see 1.1). Although the data will be collected by different BromeCast participants at different

sites, all will follow the same strict, uniform protocol for experimental design, data collection and data entry. Data will be submitted to a central repository managed by the USU technician (yr 1) and postdoc (yrs 2-4). These data will be stored in flat .csv files, with each record (line) corresponding to one demographic measurement on one individual plant, at one particular field site. A separate table will link each individual plant identifier to its DNA sequence identifier. Each data file will be accompanied by a metadata file (raw text). Data storage and distribution will follow the protocol described for the Phenology and Physiology data (section 1.2).

1.4. Environmental data

We will compile data on environmental covariates at each field site, including climate means based on long-term observations, monthly precipitation and temperature during the period of study (collected on-site for the core common gardens and using interpolated climate data for the satellite sites), and soil depth, texture, and total C and N. Each covariate will be stored in a separate file, all linked by unique site and year identifiers. Each data file will be accompanied by a metadata file (raw text).

As with the Demographic data, raw Environmental data and metadata will be stored on the cloud in Google Drive, with additional weekly back up on local hard drives and external disc drives. Once QAQC is complete, the data files will be added to the Github repositories where we will store our computer code.

1.5. Computer code

We will code our statistical and population models in R, using both existing R software and new R packages we will develop. Computer code will be developed with Git version control and stored on Github repositories.

2. Data access and distribution

Our primary approach for archiving data and ensuring long-term open access is to distribute a data and code package with each peer reviewed paper we publish. These packages, published to the Dryad Digital Repository (http://datadryad.org/), will contain the data, metadata, and R code needed to reproduce all analyses, results, and figures contained in the corresponding paper. Dryad is a well-funded non-profit that is widely used in biology and has partnerships with many journals. Data and code will be distributed without restrictions. In order to help other researchers find the published data packages, we will link to them from our individual lab web pages, and from all publications that use the data.

Raw sequence reads will be archived, with appropriate and complete meta-data, at NCBI's Sequence Read Archive (SRA). SRA is permanent (contingent on future funding by the US government) so there are few concerns about longevity after our project period expires. Other data types (flat text files of analysis, etc.) will be disseminated as figures, tables, or supplementary data in publications that result from the award, consistent with community standards. Large datasets, *e.g.* SNP calls, will be archived on Dryad, where they will be freely available, upon publication.

At the conclusion of the project, we will publish a comprehensive data paper in *Ecology* that describes the database of demographic, environment, and genomic data. This will serve as a centralized resource for all output from this project that will be freely available for use by future *B. tectorum* researchers.