3-20-2019

CAREER: An integrated approach to understanding selection and evolution in heterogeneous environments

Zachariah Gompert
Utah State University

Data Management Plan

1. Data Collected, Formats, and Standards

   The proposed research will produce raw DNA sequence data, host plan and insect abundance data, experimental measurements, and cuticular hydrocarbon (CHC) profiles. I will generate the DNA sequence data on an Illumina HiSeq 2500/4500 and store the data in a series of `fastq` files, with one file per sequenced DNA pool (PoolSeq) or individual butterfly (GBS). These data will include sequences from ∼10,000 wild-caught *Lycaeides* butterflies and 2000 experimental butterfly larvae. My lab and I will process these raw sequence data by filtering out low-quality or non-target sequences. We will then align the remaining sequences to the *Lycaeides melissa* genome. We will store the processed alignments in the standard `bam` format (a binary representation of the `sam` format). We will further process the sequence data by identifying variable nucleotides and calculating genotype likelihoods and store this processed data as a single `vcf` or `bcf` file per project or experiment. We will store all scripts used to process the data and relevant command line options in 'readme' files with the data files and on `GitHub`.

   We will generate biodiversity count data at 10 focal sites each year of the project. These data will include the UTM coordinates for each site and presence or abundance counts of adult *Lycaeides* butterflies, other insects (including ants) and arachnids, and the local host plant (*Lupinus argenteus* or *Astragalus miser*) identified at each site. We will store these data in `csv` formatted ASCII text files. We will also retain the original data sheets and field notebooks used to record these data.

   Experimental data generated by the proposed research are measures of larval weight for experimental larvae from 2000 caterpillars (Obj. 2). We will store these data in `csv` formatted ASCII text files with the experimental results and metadata, such as source population, family, temperature treatment, and collection and rearing dates. We will also retain laboratory and field notebooks with these original data.

   The CHC data will consist of retention time, flame ionization, and mass/charge ratio determined through electron ionization for compounds separated via solid phase chromatography. These will include absorption and mass spectra for 2000-3000 caterpillars. These data, and the instrumentation and processing methods, will be stored on the computer associated with GC at the Nevada Institute of Chemical Ecology and on my storage space at the University of Utah Center for High Performance Computing. Data on samples (sample ID, population ID, weight) will be stored as `csv` files. We will also retain laboratory and field notebooks with original sample data as well as detailed steps completed in the extraction process.

2. Data Storage and Preservation

   We will store all primary and processed DNA sequence data on central computer servers with RAID arrays at the University of Utah Center for High Performance Computing (CHPC). I currently have 72 TB of storage space on these computers, and I have requested funds for additional storage to phase out older storage and build new capacity for the data generated by the proposed project. Data on these storage

servers are regularly and automatically backed-up. This includes two sets of quarterly tape back-ups (for all data) and weekly back-ups of active projects and data to an off-site ceph storage device (20 TB is available). As an extra precaution, raw sequence data will also be mirrored to USU's google drive account. I will also deposit the raw sequence data at NCBI's sequence read archive (SRA).

All other data will also be stored on local computers in the Gompert lab and at the University of Utah CHPC to guard against loss of data. We will also deposit these data on DRYAD for long-term storage. The DRYAD submissions will include details of the procedures used to generate the data and information about analyses of the data as metadata.

3. Dissemination Methods and Policies for Data Sharing

   We will make all sequence data and other data freely available through the open access repositories: NCBI's SRA and DRYAD. These data will become available as soon as the first associated manuscript is published and they will be made available indefinitely. Any data not published by the end of grant period (April 2024), will be made available at that time. The metadata in these repositories will contain all of the information needed for other scientists to reuse these data and recreate our analyses of them.

   My commitment to free and open exchange in science will extend beyond data sharing. Software developed as part of this project will be made freely (i.e. open-source and no monetary fee) available to academic and non-academic users under the GNU General Public License v2. This ensures that the programs and their source code will be available to all to use, modify and reuse so long as they grant the same rights to others. Source code and binaries will be deposited on SourceForge and be available via my website. Manuals and examples will be provided with the software. I will also prioritize publishing papers from this project in open access journals or those that have an open access option (which I will take advantage of; funds have been requested for this). Course materials from my courses will also be made freely available online for self-paced study by students anywhere in the world (USU has a long-standing commitment to distance education and has the resources necessary to ensure the on-line course is successful).

4. Roles and Responsibilities

   I will take the lead and primary responsibility for coordinating computer data storage and access. I will meet regularly with the post-doc and graduate students working on the project to discuss data management and ensure that all protocols for proper data back-up and access are being followed. While I will make all decisions about local data storage and access, decisions regarding long-term data storage at NCBI's SRA and on DRYAD will be made by the individuals in charge of those repositories.