DATA MANAGEMENT PLAN

A key component of STEPP-NET is the emphasis on data management from specimen collection through processing and archival of specimens, to generation of derivative datasets and their connection back to museum vouchers in digital databases. Rigorous data standards are critical in order to fill the gap in specimen infrastructure that currently exists for Central Asia. The main classes of data produced by the project will be: (1) physical specimens and specimen parts archived in established university museums, (2) specimen metadata (including Darwin Core fields), (3) genetic and genomic datasets, and (4) results of phylogenetic and community analysis, (5) programming code, and (6) biodiversity-themed educational materials.

Specimens – All voucher specimens (small mammals, ecto- and endoparasites) collected through this work will be housed in perpetuity at the Museum of Southwestern Biology (MSB; Division of Mammals, Division of Parasitology, and Division of Genomic Resources). The MSB is a thriving natural history museum based at the University of New Mexico with dedicated faculty curators and collections staff. The physical building contains multiple floors of dedicated collections spaces, fire suppression systems, and external and internal secure (key card) access that collectively ensure our ability to curate and care for these specimens for the long-term. These MSB Divisions also use unique barcode identifiers (scanned into the Arctos database) to track materials and preserve relational linkages (e.g., between host and parasite) originally recorded in the field.

Specimen metadata – All MSB Divisions listed above use Arctos, a web-based collection-management application comprising over 40 biodiversity collections that is a data provider to global biodiversity aggregators including Global Biodiversity Information Facility (GBIF), VertNet, and iDigBio. Specimen occurrences and associated metadata will be incorporated in digital format into Arctos, typically within one month following accession. Metadata types will include established, Darwin Core fields such as collection locality (including geocoordinates for point of capture at < 10 m precision), exact collection date, catalog and field numbers, collector, preparator, associated tissue types and preparations, and other individual-level observations (age class, reproductive status, external measurements [for mammals], infection intensity and in-host location [for parasites]). These data allow specimen-level traits and observations to be analyzed at high spatial and temporal resolution by current and future researchers. In addition, we will maintain permanent reciprocal links between all parasite and host specimen records within Arctos, a unique functionality that facilitates exploration of host-parasite networks and allows those host-parasite links to be exported to data aggregators as well.

Arctos data are held at the Texas Advanced Computing Center (TACC), to which they are uploaded locally by collection personnel through web browsers or secure file transfer protocols to the TACC Corral system. TACC is an NSF XSEDE facility providing 4 petabytes of replicated storage capacity accessible over 8 server nodes connected to national research and education networks at 10 gigabits and to the backend storage controllers at 40 gigabits. The Oracle database is backed up nightly, and Arctos media are stored in the iRODS preservation environment and are replicated to at least two storage systems. TACC services provide highly reliable, high-performance, scalable, and secure infrastructure for the entire Arctos system (Oracle database, ColdFusion applications). Users are given access for transferring data and media either directly through Arctos or through projects established on the TACC portal. Collaboration with TACC is through a signed Memorandum of Understanding. As a result of Arctos/TACC digital infrastructure, data from this project will immediately become a functional part of global biodiversity informatics.

Specimen Data Use Policy – The default condition of data in Arctos and most public databases that will house derived data (e.g., GenBank) is that they are fully web-accessible upon deposition. Specimen data that are sensitive or proprietary can be temporarily encumbered in Arctos. MSB maintains guidelines for the use of museum specimens, including destructive use of tissues by researchers other than project PIs, on its website. These guidelines follow generations of specimen-based research in explicitly considering specimens and associated data as permanent, open repositories. Arctos' policy statement on Ownership and Use of Data is prominently posted on that website. In all cases, data and media are the property of the originating institution, with all rights reserved.

Genetic and genomic datasets – STEPP-NET will generate reduced representation genomic data (DNA barcodes [mammals, parasites], ultraconserved element (UCE) loci [mammals]) and whole genome assemblies (parasite OTUs). All sequence data will be deposited in databases of the National Center for Biotechnology Information (NCBI) and linked to voucher specimen records by direct and reciprocal links using Darwin Core triplets. Barcodes and UCEs will be deposited as quality-checked, unaligned sequences directly in NCBI's Nucleotide database. Whole genomes will be deposited as raw Illumina read data on the Short Read Archive, and assemblies of these genomes will be deposited in the Genome archive.

Phylogenetic and Community Analyses – In the spirit of open science and data re-use, phylogenetic hypotheses (gene trees from DNA barcodes, multilocus phylogenetic/omic hypotheses from UCEs and whole genomes) resulting from this project will be made available in Newick or Nexus format in the Dryad digital data repository or TreeBase immediately upon publication of results. In addition, we will make (1) raw community data and (2) cleaned community data products publicly available in Dryad upon publication as well. Datasets will be curated such that data tables contain specimen- and species-level resolution, specifically including Darwin Core triplets and links to Arctos specimen records.

Programming Code – Bioinformatic workflows will be based on a mixture of standalone software packages and new Python code. All code for genomic data processing will be deposited in the Dryad digital data repository upon publication to facilitate reproducibility. Other analytical workflows (genetic distances calculation, community analyses, etc) will be based in the R language, and all code will be deposited in Dryad. In addition, the provenance of all specimen data (geocoordinates, host associations, phylogenies) and other inferred datasets we develop and use (climate, grazing intensity) will be recorded using the newly developed *recordr* package and linked to any publications that emerge from the proposed research. All of our code will be made available under the CC0 copyright license (open access, no citation required).

Educational Materials – Two classes of educational products will be produced by this project. First, undergraduate educational modules related to a) Palearctic mammalian biodiversity and biogeography, b) host-parasite coevolution, and c) globalization and emergent pathogens will be developed with the Biodiversity Literacy in Undergraduate Education initiative. All of these will be permanently hosted as downloadable QUBES bundles that contain data (or permanent data links) and Microsoft Word documents describing module activities. Second, K12 educational kits will be developed for use by educators and student teachers in North Carolina. These kits will include maps, images, and physical models and will be maintained as resources for check-out by UNCG Libraries.

Data Management Roles and Responsibilities – Trained field personnel (PIs, graduate students, and undergraduate students supervised by PIs) will be responsible for implementing our longstanding specimen data standards described above. Curatorial and collections staff at MSB will be responsible for implementing longstanding specimen archival procedures, including data digitization and management within Arctos, physical specimen tracking, and scanning of field notes and associated metadata. PIs McLean, Malaney, Galbreath, Greiman, and Koerner will be responsible for data standardization, quality control, and data accessibility/reproducibility during and after project completion. Prior to publication and final deposition, PI labs will archive DNA sequence datasets in triplicate (lab computers and 2 external hard drives). Preliminary project datasets being actively used in analyses will be maintained using cloud-based solutions by PI labs (Google Drive, Box).

RESOURCES, EQUIPMENT, & OTHER FACILITIES

FACILITIES

Department of Biology: The department of Biology at Georgia Southern University provides many opportunities for hands on training in many different fields, through programs such as NSF REU summer supplements. The department has recently, within the last 5 years, moved to a new 135,275 square-foot LEED certified building with sate of the art lecture rooms, teaching labs, prep rooms and research labs, providing the appropriate resources for the development of strong research programs including students at all levels.

Laboratory: Dr. Greiman has ca. 500 sq. ft. laboratory space in the Biological Sciences building, Department of Biology, GSU. Greiman laboratory includes a fume hood, 2 PCR hoods with UV lights, 1 refrigerator, 3 regular freezers, an ultracold freezer, multiple slide warmers, 2 regular thermal cyclers, 1 ABI Step One Plus ABI real-time PCR machine, 1 Eppendorf Microcentrifuge with 2X96 well rotor, 1 high end programmable ultrasonic disruptor that allows indirect (=suitable for sonication in closed tubes without loss of DNA) processing of up to 8 samples at a time, 1 Eppendorf 5810R centrifuge with 4X96 well rotor, 1 Accuspin desktop centrifuge, 1 thermomixer, a Fotodyne gel imaging system, various single- and multi-channel pipettors, Eppendorf repeater, power supplies, electrophoresis units, balances, magnetic stirers, vortex mixers, and heat blocks. Microscopes include a Nikon Eclipse Ni-U research microscope equiped with Nomarski differential interference optics, 2 Olympus and 1 Nikon field grade dissecting microscopes. One of the field grade Olympus dissecting microscopes is fitted with a C-mount and can be taken to the field for digital imaging. The Nikon Ni-U compound microscope is equiped with a 24-megapixel digital camera and image capture and measurement software.

Computer: Department computing facilities have various PCs and Mac computers. Greiman has several desktop workstations in his office and laboratory, including a large memory TS P920 workstation. All computers have wired and wireless access to the Internet. The PI possesses specialized software needed for sequence assembling, alignment and phylogenetic analyses (Geneious, MrBayes, BEAUti and BEAST, Arleguin, and other packages. For analyzing larger genomic datasets GSU has a large computing cluster (Talon Cluster) available to faculty.

Office: Offices are presently available for the faculty. All graduate student offices have been refurbished in 2013 (new building).

MAJOR EQUIPMENT

Dr. Greiman also has access to an Eppendrof 96-well real-time PCR machine, ABI 3500 Genetic Analyzer (DNA sequencer), scanning electron microscope, confocal microscope, pippen prep (size selection for whole genome library prep), Bioanalyzer, Covaris M220 Focused-ultrasonicator (genome library prep (DNA fragmentation)), BioTek Synergy H1 hybrid multi-mode microplate reader, Quibit flurometric quantitation through the Department of Biology.

OTHER RESOURCES

Vertebrate Animal facilities: Adjacent to the new Biological Sciences Building is a 14,940-square-foot animal care and research-support facility. The fieldhouse includes an aquatics research area, animal research area, an insectary with warm room, storage space for field equipment, a locker room with showers, loading dock storage, and office space. The aquatics room provides air and treated water lines, shelf space for aquaria, and space for self-contained aquatic housing systems (e.g., for zebrafish). The animal research space includes animal rooms, a gowning area, a surgery room, and a cage washer. These facilities are designed to meet the latest guidelines for the care and housing of animals.

Administrative Resources: GSU Biology provides supplies, mail delivery, telephones, and electricity, as well as support for purchasing, human resources, contracts and grants management, accounting, and event planning. Additionally, the department has departmental secretarial staff (2 people), research technician (1 person), research coordinator (1 person), collections manager (1 person), IT service (1 person), and laboratory coordinator (1 person) are full-time.