

DATA MANAGEMENT PLAN

Types of Data and Materials Produced

We will generate a number of types of data during this project, some as biological samples or videos, others as researcher-observations or instrument-generated values.

Spatial Coordinates: Capture locations for animals will be recorded via GPS in Universal Transverse Mercator (UTM) coordinates and saved as .csv files.

Biological Samples: Plasma samples from whole-blood will be kept in a -80°C freezer which is connected to an alarm that communicates directly with the PI's cell phone. RNA from organ tissue and DNA from swabs for pathogen will be stored in a similarly alarmed freezer at the co-PI's laboratory.

Field and Lab Methods: Standard operating procedures for animal capture and transport; laboratory assays performed in all laboratories; primers for qPCR assays; housing of animals; and workflows for data analysis will all be saved as text or spreadsheets (.txt and .csv).

Videos of animal behavior: Digitally encoded videos of animals before and during infection experiments will be saved as .mp4 files.

Infection data (researcher observations and instrument generated data): Spreadsheets of numerical values of the severity of inflammation, mass change, locomotion, and feeding (researcher observed); and antibody levels, cytokine mRNA expression, amounts of pathogen DNA (instrument generated); all organized by individual animal ID and day of infection will be saved as .csv files.

Nucleotide sequences (instrument generated): raw RNA sequences will be generated by an Illumina HiSeq 2500 and saved in FASTA and FASTQ format (text-based, non-proprietary).

Standards, Formats and Metadata

Whenever possible, our data will conform to international standards. For example, dates will be encoded using the ISO8601 standard (e.g. 2016-08-05) and locations will be entered as Universal Transverse Mercator (UTM) coordinates (e.g. 12U 298257 m E 5410201 m N). In addition, we will maintain a data dictionary for all measurements recorded, including non-standard terms. For example, our labs use "DPI" to abbreviate "days post-infection," which will be defined in the data dictionary as "the number of days following initial experimental infection with X, with this date of initial infection defined as day 0."

File formats (all non-proprietary) are given in the description of each type of data type above.

Metadata for raw data will include descriptions of methods and standard operating procedures used to collect said data (.txt files). Metadata will be provided for each processed data set, drawn from our data dictionary. These metadata will include times, dates, and methods of data processing, relevant experimental procedures, column and row name definitions, units and measurement standard. Datasets, associated metadata, and analysis workflows will be packaged using Ecological Metadata Language (EML) via Morpho¹.

File formats (all non-proprietary) are given in the description of each type of data type above.

Roles and Responsibilities

PI will be primarily responsible for the design, implementation, and updating of this data management plan, in collaboration with co-PIs, the postdoctoral associates, and the graduate students supported by

¹ Morpho. <https://knb.ecoinformatics.org/#tools>

this award. The postdoctoral associate will be responsible for periodic checks on adherence to the data management plan, consulting with PI to resolve issues.

Should PI be unable to complete this primary role for any reason, co-PI will step into this role. We will ensure that all co-PIs have full access to this plan and cloud-based data and metadata so that if such a transition must occur, it will be seamless.

Dissemination Methods during the Award

During the course of the award, processed data and metadata will be shared among all members of the research team using CyBox, ISU's secure, password-protected, version-controlled, cloud-based file-sharing program, provided through Box². This service uses multiple data centers ensuring files are not stored in one physical location. All files will be synced locally to a password-protected PC in the PI's lab, which is backed up via CrashPlan PROe³, an offsite backup service provided by ISU. In keeping with ISU Health and Safety Policy, this lab remains locked at all times and is keyed separately from all other labs in the building. Raw RNA sequence data will be stored by the Genome Informatics and stored on redundantly backed up RAID boxes consisting of 144 Terabytes of storage, maintained by the High Performance Computing (HPC) facility on campus. Project personnel will be given access to this data through password-protected access provided through the GIF and HPC.

Upon publication of manuscripts, data and metadata (encoded as described above, unless otherwise specified by journals) pertaining to those manuscripts will be made publicly available through Figshare⁴ and referenced within the publication. Any sequence data from publications will also be made available through the National Center for Biotechnology Information's specific databases, such as the Gene Expression Omnibus (GEO)⁵ for RNA-seq data.

PI Policies on Sharing and Public Access

All PIs are committed to sharing data with other researchers and making raw data publicly accessible. No data collected will require protection of privacy or confidentiality. As such, upon publication of each paper resulting from this work, we will publish and archive the associated data in a publicly accessible, searchable database (see above and below). In addition, within 24mo after results from our RNAseq experiment are published, even if all data are not utilized for that paper, we will deposit the entire sequence dataset into GEO. Published and deposited data will be reusable and redistributable by others. In addition, funds have been requested for publication costs, so open access journals will be preferred.

PIs will retain priority access to excess biological samples for 12mo following publication of results, exploration of novel hypothesis generated during the course of this award. After 12mo, PIs will make excess biological samples available for sharing upon request and validation of the requester's legal permissions to possess such samples, until no material remains.

Archiving, Storage and Preservation

All data from this award, including any not resulting in a publication, will be archived and made freely available via Figshare within three years of the termination of this award. Data will be accompanied by metadata using the EML standard, as described above. Biological samples will be stored in PI and Co PI's labs for two years following the termination of the award.

² Cybox. <https://www.it.iastate.edu/services/storage/cybox>

³ CrashPlan PROe. <https://store.crashplan.com/store/>

⁴ Figshare. <https://figshare.com/>

⁵ Gene Expression Omnibus (GEO). <http://www.ncbi.nlm.nih.gov/geo>