Funded Research Records

2-7-2018

Maternal effects on offspring development in crop-pollinating bees, Megachile rotundata

Karen M. Kapheim

Utah State University, karen.kapheim@usu.edu

Maternal effects on offspring development in crop-pollinating bees, *Megachile rotundata*

Data Management Plan
Proper management of data will be key to the success of this project. The PDs' labs have regular discussions of the primary literature in data science and data management. We strive to use best practices to ensure accessibility of our data and reproducibility of our analyses.

a. Expected Data Types
1. *Specimen data*. This includes all data pertaining to our experimental bees.
2. *Digital data*. This includes videos of bee nesting to assign nest ownership in field cages.
3. *Sequence data*. This includes messenger RNA and small RNA sequences.
4. *Analytical workflows*. All data analyses will be documented in code and syntax.

b. Data Format
    Specimen data. All bee data will be captured in hard copy paper worksheets designed to be filled-in during the experiments. Each bee will be given a unique identifying code, and this code will accompany all records, including sample tubes stored in the -80°C freezer.
    Digital data. Videos of bee nesting behavior will be recorded directly to video cameras with built-in hard drives as MTS files.
    Sequence data. RNA sequences will be received as digital files in the FASTQ format via secure file transfer protocols (SFTP).
    Analytical workflows. This research will also generate a public repository of data analytical workflows written in markdown syntax (MD files).

c. Data Storage and Preservation
We will have a dedicated project within the Open Science Framework (OSF) that will serve as the central hub for all data, metadata, workflows, and communication about all aspects of the project. We will run a project wiki through OSF, and will link this wiki directly to the BOX directory where all of the specimen data is stored, the Google Drive clone of our UHPC sequence data, the GitHub repository where our data analytical workflows are created/stored, and the Harvard Dataverse, where our materials are archived.
    Specimen data. All worksheets will be scanned and archived in BOX after the data has been transcribed into spreadsheets within BOX and checked for accuracy. BOX is backed up to the cloud daily, and provides version control. All specimens used for genetic sampling and the genetic materials generated from these samples will be archived indefinitely in a -80°C freezer in my laboratory, equipped with a low/high temperature alarm and a backup CO_2 system that becomes activated in the event of high temperatures. Specimen data, including experimental data, nesting videos, and data analytical workflows will be stored in BOX and archived in the Harvard Dataverse, which is contained in a physically secured, continually-monitored facility, and system backups are made on a daily basis.
    Digital data. MTS files will be transferred from the video cameras to the lab-owned storage space at the University of Utah High Performance Computing Center (UHPC). PD Kapheim has 32 TB of storage space with a guaranteed life of 5 years, which will expire during the project funding period. The budget thus includes additional storage nodes and back-up tapes to avoid any data loss. The data storage nodes require authentication for access, and data transfers are encrypted. All team members will be able to access these data upon request. This space is backed up to alternating sets of tapes stored in a separate location on a quarterly basis. This space is

cloned to Google Drive, for which we have unlimited space, on a weekly basis. These files will be archived on the Harvard Dataverse.

Sequence data. FASTQ files will be stored on the lab-owned storage space at the UHPC, following the same protocol as described for digital data. Sequence data will be archived in the NCBI Small Read Archive (SRA).

Analytical workflows. MD files will be stored in GitHub, and archived at the Harvard Dataverse and from GitHub through Zenodo.

d. Data Sharing and Public Access

We will follow the "FAIR (Findability, Accessibility, Interoperability, and Reusability) guiding principles for scientific data management and stewardship" (Wilkinson et al. 2016 *Sci Data*).

Findability. All data and metadata generated from this project will persist with unique identifiers in the Harvard Dataverse, Zenodo (GitHub), or NCBI SRA. Data can be easily searched, accessed, and cited from each of these repositories.

Accessibility. Data and metadata will be accessible at public repositories. Data may be embargoed until publication or 3 years after the end of the funding period, whichever is first.

Interoperability. The project OSF wiki will include instructions on standards for file naming, spreadsheet organization, file management, and coding, and these will be based on recommendations from the USU Research Data Management Systems. We will also continue regular discussions about data standards at least once a quarter in regular lab meetings. When archiving data, we will conform to metadata standards of each repository.

Reusability. Each of the archival repositories we have selected perform validity checks and version control, thus enhancing long-term reusability. Metadata will accompany data, and workflows will accompany all descriptions of data analysis. These data are not subject to protection of privacy, confidentiality, security, or intellectual property rights.

e. Roles and Responsibilities

PD Kapheim will be primarily responsible for implementation of this DMP. Kapheim will ensure quality and integrity for sequence data and analytical workflows. Both Kapheim and Pitts-Singer will ensure quality and integrity for specimen and digital data. The students on the project will report primarily to Kapheim, but will coordinate with both of the program directors.

Table 1 Summary of data management plan

Data type	Format	Storage location	Backup schedule	Maintained	Archive location
Bees	n/a	-80°C freezer	CO_2 backup	Indefinitely	n/a
Data collection sheets	PDF	BOX	Daily	Minimum 5 years past the end of the project	None
Specimen data	spreadsheets				Dataverse
Video data	MTS	UHPC	Google Drive weekly; Tapes quarterly		Dataverse
Sequence data	FASTQ				NCBI SRA
Analytical workflows	MD	GitHub	Daily		Zenodo, Dataverse