7-21-2020

Collaborative Research: Diverse selective pressure on fruit chemical traits from mutualists and antagonists as a major driver of chemical evolution at the whole plant level

Noelle G. Beckman
Utah State University, noelle.beckman@usu.edu

DATA MANAGEMENT PLAN

1. *Types of data:* To promote collaboration across institutes and the sharing of data, models, and other products, the cyberinfrastructure utilized by the research team will support all stages of the data lifecycle, from creation, processing, documenting, preserving, and sharing to reusing data. The data generated by the project include raw LC-MS-MS data and associated metadata from metabolomics analyses, field observational data, and experimental data from bioassays and behavioral experiments. In addition, these research efforts will generate source code for data analysis.

Other materials produced: Research results will be published in peer-reviewed journals. The project will also produce educational materials, journalism stories for Utah Public Radio, science communication films, and a website to summarize and disseminate products of the proposed research.

2. *Data and Metadata Standards.* For metabolomics data, we will use standards established and peer-reviewed by metabolomics researchers for documenting and preserving chemical data and data analyses (Sumner et al. 2007). These include documentation of sample preparation and processing conditions, as well as parameters for all steps in the chemical analysis pipeline, from mass spectrometry experimental design to data interpretation. Metadata in contributed datasets will follow the Metabolomics Standards Initiative (Goodacre et al. 2007). For ecological data, we will use existing and emerging standards for documenting and preserving data and follow best practices suggested by the Data Observation Network for Earth (DataONE; Strasser et al. 2012). We will include metadata following the standardized Ecological Metadata Language. Metadata can be generated through Morpho, a free online data management tool developed by the Knowledge Network for Biocomplexity (http://knb.ecoinformatics.org/morphoportal.jsp). For documenting derived data, best practices by DataONE suggest including a description of the primary data, derived data, data processing details, and any problems that emerged and how they were dealt with.

3. *Roles and Responsibilities*: The postdoctoral scholar and PIs will develop protocols for data collection, processing (e.g., quality assurance and control), and documenting (e.g., annotation with metadata). All members of the research team collecting data will be responsible for following these protocols for data processing and documentation. The lead author for each study (PIs, the postdoctoral scholar, or the graduate student) will ensure data have appropriate documentation and metadata and deposit data in appropriate repositories at the time of publication. All PIs will assume responsibility to ensure any unpublished data are made publicly available at the close of the project.

4. *Dissemination Methods*: A website developed by Impact Media Lab will summarize products and information on how to access these products, including links to repositories that house datasets, source code, peer-reviewed publications, as well as products for communicating our scientific results to the public. We will use existing repository services to facilitate collaboration through sharing of files, data, source code, and results to stimulate the generation and development of ideas, theory, meta-analyses, and models. We will use Google Drive for file storage with weekly backups and version control of non-code documents. CyVerse will enable easy access to datasets, tools, and applications with different permissions. For example, folders and subfolders can be created with the flexibility of collaborating among participants or the public by setting different permissions. We will use GitHub, a software repository with version control, to facilitate collaboration and review of code for analysis. Raw chromatography and mass spectrometry data will be made publicly available through MetaboLights (Haug et al. 2012) and annotated metabolomics data through GNPS MassIVE (Wang et al. 2016). Source code will be made publicly available through GitHub. We anticipate that the act of compiling datasets and software and making them readily available will facilitate collaborations with the research team and the broader scientific community.

5. *Policies for Data Sharing and Public Access*: As publicly funded research, any data products created by the research team (meta-analyses, source code, datasets) must be made available to other scientists and the public. Final products will be made publicly available and published (using Digital Object Identifiers or DOI) in repositories with appropriate attribution and citation information. Following the lead of the International Long-Term Ecological Research Network (http://www.ilternet.edu/data-info-management), unpublished data and other products resulting from research activities will be available to researchers from a password-protected folder on CyVerse to ensure that use meets the ethical and attribution standards of our research team. Users can gain access to these data by contacting the PIs and agreeing to an access agreement. A draft of this agreement will be developed among the PIs, postdoctoral scholar, and graduate students during the first year of the research project. After the tenure of the research grant, intellectual property policies of products resulting from the grant will comply with requirements of the collaborating institutions.

6. *Archiving, Storage and Preservation*: The CyVerse cyberinfrastructure will allow for data preservation, sharing, and reuse among the research team and the public. CyVerse will store data, derived data, and results while in active use. For data publication and permanent archival, we will plan to use the GNPS MassIVE repository as well as MetaboLights for the mass spectrometry data and the Environmental Data Initiative or Knowledge Network for Biocomplexity for ecological data. All of these repositories accept multiple data formats, assign permanent unique identifiers (MassIVE ID, MetaboLights ID, or DOI), and ensure data are discoverable, reusable, and citable. In addition, MassIVE automatically ensures that molecular library matches remain up-to-date by periodically comparing datasets to the ever-growing libraries linked to GNPS. Source code created through the project will be assigned a DOI and made freely and publicly available using GitHub. All other products will be archived at USU's official institutional repository (DigitalCommons@USU).

References

Goodacre, R., D. Broadhurst, A. K. Smilde, B. S. Kristal, J. D. Baker, R. Beger, C. Bessant, S. Connor, G. Capuani, A. Craig, T. Ebbels, D. B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, and F. Wulfert. 2007. Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics **3**:231-241.

Haug, K., R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. 2012. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic Acids Research **41**:D781-D786.

Strasser, C., R. Cook, W. Michener, and A. Budden. 2012. Primer on data management: What you always wanted to know. A DataONE publication **http://dx.doi.org/doi:10.5060/D2251G48 (available via CDL)**.

Sumner, L. W., A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and M. R. Viant. 2007. Proposed minimum reporting standards for chemical analysis. Metabolomics **3**:211-221.

Wang, M., J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A.

Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein, and N. Bandeira. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nature Biotechnology **34**:828.