C-CoMP Data Management Plan

Data-types. Chemistry data will include lab- and field-based metabolite profiles by NMR spectroscopy (Edison) and by liquid chromatography mass spectrometry (LC/MS; Kujawinski), protein profiles by LC/MS (Saito), and data and metadata from the field studies (Bates). Biology data will include lab- and field-based sequence data for eukaryotes (Dyhrman), prokaryotes (Moran) and viruses (Sullivan). Modeling data will include output from genome-scale models (Segrè), flux-balance models of metabolic networks (Covert) and regional- and global-scale ecosystem models (Doney). K-12 curriculum materials will be developed by C-CoMP education and science personnel for classroom interventions and museum exhibits. Undergraduate curriculum materials (CUREs) will be developed through interactions among C-CoMP personnel and members of CUREnet. Educational data, including surveys, student assignments, and audio recordings and associated transcripts, will be generated during testing of K-12 curricula (O'Dwyer) and undergraduate courses (Dolan). Software for data integration and visualization will be developed at the University of Chicago (Eren) with C-CoMP dedicated computational nodes.

Data standards. Metabolite data will be collected in vendor software (LC/MS; Thermo Scientific; NMR; Bruker). LC/MS data will be converted to open-source formats to facilitate processing using open-source platforms (e.g., XCMS). Metabolite metadata will be collected according to the Metabolomics Standards Initiative[1]. Protein data and metadata will be collected according to best practices laid out from Ocean Metaproteomics data sharing effort[2]. Field data from BATS will be collected in ASCII and hexadecimal files and provided to users in ASCII, Matlab or Excel formats. Sequence data and associated metadata will be stored according to the standards of the National Microbiome Data Collaborative. We will coordinate with emerging standards on environmental sequence data, resulting from an NSF-supported OCB workshop on Ocean Nucleic Acids Intercalibration. For modeling data, we will follow practices in Gil et al.,[3] which specifies data, metadata and computations to be stored in public repositories.

Data access and sharing. Data generated in C-CoMP laboratories will be stored in repositories that provide DOI numbers to ensure persistent access to our products, unless otherwise indicated. Metabolite data will be deposited in either MetaboLights or the Metabolomics Workbench. Field data from BATS is stored at the BATS ftp website and at the NSF-supported Biological & Chemical Oceanography Data Management Office (BCO-DMO) repository. Processed protein data will be stored at BCO-DMO and accessed through the Ocean Protein Portal. Proteomics raw mass spectral data will be submitted to ProteomXchange via Massive or Pride. Sequence data will be stored at NCBI, iMicrobe and/or iVirus. Raw model data will be stored on local servers initially. For CESM results, we will adhere to their data policy, which includes data release within one year of generation. In general, C-CoMP will meet the NSF guidelines on data release within 2 years of generation but will release any hardened data products before then, if possible. All described repositories provide immediate access to the broader community. All software development efforts in C-CoMP will follow open-source software development practices, will be licensed through General Public License, and the source code of stable releases as well as the development branches will be accessible to the community through GitHub repositories. Stable releases of our software will also be available as Conda packages through the Anaconda Package Repository and as Docker containers through the Docker Hub for platform-independent, easy-to-install or easy-to-run scenarios.

Intermediate products such as metabolite abundance files, peptide abundances, metagenome assemblies, distilled model outputs and data visualizations will be stored in a dedicated Google Drive for access by all C-CoMP personnel. When ready, these intermediate products will be moved to BiGG, GitHub, Zenodo, and other cloud services for DOI assignments. A BCO-DMO C-CoMP project page will be created to host the field data and link to the various raw and processed repositories (e.g., NCBI, ProteomXchange, MetaboLights) to enhance data and metadata discoverability and promote data reuse. Datasets stored at BCO-DMO can be downloaded in a variety of formats. BCO-DMO data managers work

with submitters to utilize standard vocabularies for parameters leveraging an open, authoritative source of oceanographic vocabulary terms and descriptions (curated by the British Oceanographic Data Centre) for its master vocabulary terms whenever possible. Numerous environmental parameters measured in C-CoMP (e.g. metabolites and targeted peptides biomarkers) are not yet present in the curated ocean parameter vocabulary, but the data management postdoc will work with the Digital Coordinator and BCO-DMO to incorporate new parameters into the oceanographic standard vocabulary to promote discoverability and reuse.

Policies and provisions for re-use, re-distribution and the production of derivatives. C-CoMP supports any and all re-use of data deposited in online repositories. The Center will maintain a blog on its web page, promoted by social media, to share data analysis news and intermediate findings written by early-career trainees in plain language. Through continuous documentation, our observations will reach researchers and students quickly, benefit from community feedback and criticism, and create networking opportunities for young Center-affiliated scientists. We will share fully reproducible bioinformatics analyses and model calculations with Docker containers and Jupyter notebooks for reproducibility, platform independence and ease-of-use. All publications will be posted on pre-print servers to ensure prompt communication to the community. C-CoMP has set aside funds to support open-access publication for all C-CoMP manuscripts. As required by NSF, C-CoMP data and publications will be deposited into the NSF Public Access Repository (NSF-PAR).

Provisions for privacy Educational data is confidential and cannot be stored on a publicly available server, and instead will be stored on a secure local server at UGA (Dolan) or BC (O'Dwyer). We will obtain the necessary IRB approvals for educational data collection, management, storage, and archiving. During data analysis, the data will be stored on password-protected systems and will be accessible only by project personnel certified in human subjects research. To protect the confidentiality of study participants and future data dissemination, the following (or similar) language will be used in the informed consent: "The information in this study will only be used in ways that will not reveal who you are. You will not be identified in any publication from this study or in any data files shared with other researchers. Your participation in this study is confidential. Federal or state laws may require us to show information to university or government officials [or sponsors], who are responsible for monitoring the safety of this study." Anonymized interview and survey data will be disseminated in aggregate to protect individuals' identities in papers disseminated via refereed journals, research talks, and on the PIs' websites.

Data archiving policies. Each investigator is responsible for maintaining data associated with their own research group's activities, and in accordance with any institutional requirements of the PI's home institution. Hard-copy and digital notebooks will be used to record details of experiments and model simulations, as appropriate for each laboratory. These notebooks will be stored in the home institution but available for retrieval upon request. All research notebooks of each investigator are property of their home institution. C-CoMP senior personnel are responsible for due diligence with respect to short-term storage of data, using computer hardware and software that is available to their laboratory. All data shall be retrievable from primary media or back-ups, as well as reasonably protected from accidental loss due to corruption, power loss, or failure of computer hardware. All primary and processed data will reside on local servers with appropriate backup capabilities for at least five years, or until it has been successfully uploaded to and made publicly available through a nationally or internationally funded database.

1 Sumner, L. W. *et al.* Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211-221 (2007).

2 Saito, M. A. *et al.* Progress and challenges in ocean metaproteomics and proposed best practices for data sharing. *Journal of Proteome Research* **18**, 1461-1476, doi:10.1021/acs.jproteome.8b00761 (2019).

3 Gil, Y. *et al.* Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance. *Earth and Space Science* **3**, 388-415 (2016).