DATA MANAGEMENT, ANALYSES AND INTEGRATION

Data will be managed among the research partners within a framework designed to integrate measurements, disseminate results, scale to the ecosystem level, and enable hypothesis testing. We feel that dissemination and integration of the above components is critical, for both understanding key physiological processes at the cellular level, determining the influence of environmental variables, and resolving differences in endogenously vs exogenously supplied microcystin. Our overall goal is to optimize the availability and utility of data to all members of the research team, facilitating communication and ultimately the publication of results. We also note that we will start managing our data and supplies at the beginning of the project.

Data and samples will come in several forms. For sample management, a dedicated -80°C facility will be set up and samples bar-coded and stored for ready access and cross referencing to metadata. Other metrics such as growth conditions in the laboratory or climate data (temperature, PAR, precipitation) from Environment Canada (while on the Limnos) or from the TLLER MET-station will be collected and catalogued using the same barcoding system. These data will be important as they will form the base-line for seasonal variation and provide information on climate variance during our study. Data integration will start with the development of a web-driven repository for investigators that will feed into an all-project database on a password protected FTP server. Automated backup to an offsite server (e.g., at the University of Tennessee all biological databases are backed up daily and stored for a period of weeks). For individual databases, PIs will be responsible for backing up data as it is generated. As an example, the Wilhelm lab uses a 3 level back up system (1 local, one campus server and one offsite) for sequence information.

Strains generated through this study will be made available to the scientific community upon publication. We will deposit cyanobacterial strains in the National Center for Marine Algae and Microbiota (formally the CCMP). Although the word marine is in the name, this collection does take freshwater and other microbial strains. Wilhelm is on the board of advisors for this center and has already discussed depositing strains with the director (Dr WH Wilson). We feel sharing strains is a critical component of research. For genetic constructs, these will be catalogued and stored at -80°C in the Wilhelm lab and made available to qualified researchers upon request. On average the Wilhelm lab shares 30-50 cultures annually.

Data analyses will involve comparisons within and between measurements and experimental data sets. An all-PI meeting at the onset of the project will confirm statistically relevant sample frequency and replication are built into designs. Analyses will begin with manual curation (e.g., basic parametric examinations to ensure data are within historical ranges) and then examination of data sets to ensure even variable distribution. Multivariate approaches will consider both biological and environmental variables (e.g., Rinta-Kanto et al. 2009, Wilhelm et al. 2011, Rowe et al. 2012) to resolve drivers of variance.

Initial relationships within the data set will be examined by project investigators using standard parametric and non-parametric tools within our suite of statistical packages (e.g., Canoco, Primer-E, NCSS, SPSS, etc). Application of these tools will be ongoing and pipelined from the all-project matrix so that emerging properties of the data are apparent as samples are collected. This latter part is important, as it may allow us to augment activities so that previously unidentified metrics relevant to system function can be captured. Moreover, it will allow us to quickly identify road blocks (e.g., problems with genetic mutants, missing sequence information or missing metabolites standards from databases).

PI workshops will be held annually at the University of Tennessee. One of our challenges is to determine how to best integrate these disparate data sets – although we are already working on this
for other studies (see data presented within the study).

During year 4, the all-project database (including all chemical and physical analyses) will be sent with relevant metadata to the **Biological and Chemical Oceanography Data Management** (BCO-DMO) repository; we have previously contacted them and received a welcoming response (from Cyndy Chandler). With respect to distribution to the scientific community, sequence data will be deposited in the **National Center for Biotechnology Information** as well as within the **Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis** (CAMERA) or the **Metagenomics Analysis Server** (MG-RAST) which will allow us to couple molecular and site specific data for use by other researchers. Taxonomic composition (molecular and morphological) will be included in all documentation (*i.e.*, often as supplemental data sets to papers). Data for all metabolomics analyses will be archived online in Microsoft Excel spreadsheet, although we anticipate a community driven repository will appear in the near future. If the community ratifies a standard format for these data, the spreadsheets will be reformatted to meet standards outlined by the Metabolomics Society. The availability of all data sets will be communicated through publications as well as to peers at national conferences. After year 4 of the study, we will make residual nucleic acids available upon request through email. In 2013 we have already shared nucleic acids / biological samples with 6 other labs from previous NSF-funded work.

**Toxin Analysis (SUNY-ESF)**

Since the presence of cyanotoxins in public waterways can be of extra concern, we include special considerations in data management for this. We note that if toxin levels are excessive in any of our samples we will notify the appropriate authorities (*e.g.*, local EPA contacts) and immediately provide them with all of our data.

Upon receipt, sample are immediately assigned a unique sample identification number and all available information is entered into the laboratory database. This database contains signoff spots for initial receipt, extraction, visual exam and the different individual toxin analysis. The database is backed up nightly onto a separate computer and trice weekly onto SUNY-ESF’s secured server system. Individual analysis are also held separately on the individual instrument computers and backed up weekly to the secured server. Co-PI Boyer peruses this system daily to ensure compliance. All samples are stored from receipt until analysis is completed in a dedicated –20°C freezer located in a secured room. Average time from receipt until analysis is 1-3 weeks. After analysis, voucher samples are stored for 1-5 years at -80°C to allow for reanalysis if needed. We have also made available these extracts to outside laboratories upon request.

Instrumentation is calibrated daily using commercially purified standards of anatoxin-a (Sigma), Microcystin-LR, -RR,-LA and LF (Alexis Chemical Co.), and cylindrospermopsin /deoxycylindrospermopsin. Calibration logs and curves are integrated into the laboratory database described above and backed up automatically. Integrity of any commercial standards is verified upon initial receipt using LCMS and by 600 MHz NMR. For HPLC analysis, replicate standards are run every 20 unknown samples. This allows us to detect drift in the instruments. Positive samples are always analyzed at least twice on two separate days. For the protein phosphatase assay, duplicates protein inhibition curves containing at least 8 points are always run on the same plate under identical conditions as the unknown samples. All data analysis is subject to rigorous quality control. Co-PI Boyer also routinely submits blind positive and negative samples into the system to check for operator bias. Full QA-QC protocols and our standard operating procedures (SOPs) are available upon request.
**Mentorship of post-doctoral fellow, graduate students, and other young scientists**

While this section is intended to focus on the mentoring and development of the postdoctoral fellow, we note that we intend to mentor all students and young scientists in a manner in concordance with the ideas and opportunities presented here.

Mentoring of postdoctoral fellows will occur in a manner consistent with and exceeding the guidelines suggested in the now broadly available **FASEB mentoring statement** ([www.faseb.org/portals/0/pdfs/opa/QReports/July-Sept08/MentoringRGrants.pdf](http://www.faseb.org/portals/0/pdfs/opa/QReports/July-Sept08/MentoringRGrants.pdf)). Within this project the postdoctoral fellow will be mentored to help develop skills as a scientist and future professor in areas that will benefit future research, education, and outreach endeavors. Scientific development will be emphasized as the fellow conducts specific projects in the proposed research plan. Although the fellow will be under Campagna’s direct day-to-day supervision, s/he will also be mentored by Wilhelm and Zinser. These investigators are experts in fields closely related to what we assume will be a successful applicant’s long-term research and teaching interests. At the onset of the project the fellow will develop, in conjunction with the PIs, an **Individual Development Plan** (IDP): these plans have been shown to facilitate success in postdoctoral efforts. It will contain both goals and timelines for milestones, including data collection, analysis and dissemination.

The research effort will provide an outstanding opportunity for the fellow to collaborate with American, Chinese, and Canadian scientists with a breadth of expertise, developing important contacts and peers within both the chemical and biological scientific communities. The fellow will gain experience in mentoring: all labs involved engage graduate and undergraduate students, and it is our plan to have the fellow assist with a subset of students in order for them to develop their own mentoring skills (note: as the main goal of the fellows position is research, we will be wary of having the fellow commit too much time to mentoring).

The postdoctoral fellow will be encouraged to develop a strong scientific reputation during their participation in this project and will be urged to present their results and lead special sessions at scientific meetings and workshops. They will write manuscripts for refereed journals with first authorship encouraged and supported by the senior investigators during the writing process - at the same time they will be monitored in navigating the process of documenting science, including aspects such as determining co-authorships. The fellow will be encouraged to continue project-related research if they move to a tenure-track academic position during the course of the project: Campagna and Wilhelm will plan to sub-contract funds from the grant, when appropriate, to support the fellow’s capability to continue the research effort and contribute to project goals. Beyond this, the fellow will be given hands-on training in all aspects of the project but with a focus on metabolomics and field collection, and we are committed to helping them implement this developing technology during the next phase of their career if they so desire.

The PIs, especially the senior members, have a long record of mentoring pre- and post-doctoral fellows and visiting scientists from a variety of institutions and countries. Overall, our mentoring goal is to expand their horizons concerning the interactions between microbial biochemistry, function and evolution while advancing the chosen research direction of investigating synergies between cellular processes and ecosystem ecology.