Data Management Plan.

1. The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project

The PI and steering committee (Senior Personnel) will have overall responsibility for data management over the course of the project and will monitor compliance with the plan. The types of data to be generated by the network implementation will include educational modules, procedural manuals, tutorials, and assessment tools specific to the educational modules. The project will also generate yeast strains with specific gene deletions and fluorescently tagged ORFs. The types of data to be generated by the scientific activities supported by the network include high-throughput sequence data from next-generation sequencing technologies. The output generated from sequencing facilities will be Level 1 data, which consists of sequence data ranging from 35 to 300 base pairs. Each read will include strain information and unique barcode identifiers.

2. The standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)

The project data and products (as listed in 1. above) will be available in industry standard formats such as Microsoft Word, Excel, Access, HTML, and Adobe PDF formats. These documents will be securely stored on one of Juniata's servers until completion and will include all necessary headers and explanations of tabular data headings. All sequence data and metadata associated with each project will follow the Genomic Standards Consortium (GSC). All projects uploaded to Juniata's cluster will be required to submit metadata formatted in accordance with these standards. This will ensure that all sequence projects generated during the course of this project have captured the appropriate metadata and or contextual data, thus standardizing information collection and analysis. Complying with these standards is necessary for submission to publicly available databases and for publication.

3. *Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements*

The data from this project will be consistent with Juniata College's standard best practices for accessing and sharing research data. No data will be personally identifiable, and all relevant copyright and other intellectual property notifications will be clear. The project data and products (as listed in 1. above) will be made available through a secure server as soon as possible, within one month of the issuance of any interim or final reports to NSF. This information, without exposing network vulnerabilities, will be open to any interested party. The knowledge acquired, especially regarding the effectiveness of the implementation design, will be presented and otherwise made available to other institutions who are members of the network. Security of sequencing data stored on Juniata's cluster will be ensured as the cluster resides on a dedicated internal VLAN and are behind an Internet firewall which blocks direct access to these servers, with the exception of SSH, FTP, and RDP protocols for external access to partners. These firewall exceptions are monitored by Anti-Virus and Intrusion Prevention scanning services for all traffic that passes through. Access to the compute cluster is controlled by local authentication and users are required to change passwords every six months.

4. *Policies and provisions for re-use, re-distribution, and the production of derivatives*

The implementation knowledge put forth in the published data from this project will be available to researchers and educators. Juniata will produce documentation of the infrastructure design, manuals, tutorials, and performance outcomes. The de-sensitized design information will be made available for dissemination free of charge as an electronic download from a secure server. The scientific sequencing data will be stored on hard drives in RAID 5 configuration within the HHMI cluster for up to one year. Juniata's system also has redundant power supplies and battery backup. The sequence data will be trimmed and/or removed if it is below a certain quality threshold; these operations can be conducted using open source tools such as FASTX (http://hannonlab.cshl.edu/fastx_toolkit/)ref. The small reads are then

aligned to larger sequences. In the event of server corruption or failure both remote and local external hard drive strategies are being implemented to protect against data loss.

5. *Plans for archiving data, samples, and other research products, and for preservation of access to them.* All yeast strains generated in this project will be stored as glycerol stocks in a -80º C freezer at Juniata College. Workshop participants will keep glycerol stocks of constructed yeast strains at their perspective institutions. All project implementation products, metadata, protocols, software, and curriculum materials generated over the course of this project will be shared and disseminated to the public as an extension of the web and google sites outlined in the proposal. Sequence data will be submitted to publicly available databases such as NCBI, CAMERA, GOLD, MG-RAST, or IMG. Additionally, raw sequence data and associated metadata will be stored on local servers for up to one year and protected by user authentication, RAID 5 hard drives, and a battery backup. Each user will choose their own username and passwords protecting their sequence data on Juniata's cluster. The password must be changed every six months, and must not contain the user's account name or part of the user's full name that exceed two characters in length, must be at least six characters in length, and contain characters from three of the following categories: English uppercase characters (A through Z), English lowercase characters (a through z), base 10 digits (0 through 9), and non-alphabetic characters (for example, !, $, #, %) All users also must agree to our EagleNet Usage Policy. Each investigator retains the right to use the data before opening it up to wider use. However, all data generated in this study will be made publicly available by users prior to publication acceptance. If any users have ethical and/or privacy issues related to sharing the data, all personally identifying information will be removed from the data, and will comply with institutional ethics committees.