Data Management Plan

I. Types of data

Transcriptome measurements will be made with custom designed Agilent multi-plexed arrays (Ashworth et al. 2013), or with paired-end sequencing using Illumina Genome Analyzer II. For 2x35 reads, this corresponds to around 180-200 million reads per flow cell or ~1.6-1.75 GB high quality data output per day. Sequencing, transcriptome profiles and experiment metadata will all feed into the network inference pipeline. In addition to new data generated in this proposal we will also use public data including sequence (NCBI-nr, RefSeq, PFAM etc.), metabolic networks (KEGG) and functional associations (STRING) to infer GRNs. We will also generate education kits with wet lab and classroom material (PowerPoint presentations, Excel spreadsheets, Test templates, Handout materials etc.), curriculum, and lesson plans.

II. Data and Metadata Standards

Meta-data in the form of automated input from sensors, user input and algorithm log files will be stored in different formats (such as tab-delimited, fasta, fastq, BAM, xls, nexus, msf etc.) to ensure compatibility between different programs and databases. This will allow us to create key-value pairs to integrate relational database structures, spreadsheets and xml files. We will collect contextual details such as sampling time, horizontal and vertical location, technology (e.g paired end vs single end), parameters for sequence analysis, alignments (e.g log files) and search databases along with other metadata standards. We will adapt meta data standard that is compliant with MIGS/MIMS (Minimum Information About a (Meta)Genome Sequence) standards (http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIMARKS). Established by Genomic Standards Consortium to efficiently capture information from sequencing samples, this standard is now widely adopted by many algorithms and databases. All education material developed in this project will be aligned to state and national standards.

III. Policies for access and sharing and provisions for appropriate protection/privacy

All raw and processed data in the form of web accessible databases, R packages, Java packages and applications and in other forms will be available on through indexed public websites and databases, our own website (http://baliga.systemsbiology.net/) and freely accessible through several mechanisms upon publication. Data privacy, confidentiality and security will be adjusted according to publication timeline.

IV. Policies and provisions for re-use, re-distribution

There will be no permission restrictions on data and they will be open for sharing and re-use upon publication.

V. Plans for archiving and Preservation of access

Collected and generated data will be curated both automatically, manually and computationally before made available. Periodic data backups and submission to online repositories such as GEO will be done according to suggested journal and archive guidelines. All captured metadata and protocols will be documented and available in the form of references and research papers. Data preservation beyond the life of the project will be determined on case-by-case bases and will be backed up on tapes for long-term storage. We also publish all of our data on our website: http://baliga.systemsbiology.net.

VI. Algorithms and Interpreted Data

Machine learning algorithms developed *for Thallassiosira pseudonana* will be published as part of conference proceedings or research papers. Code and data will be made available on the website, as we have done with previous software. Model results and predictions generated in this study will be made available on the Baliga lab website (http://baliga.systemsbiology.net/) alongside or integrated with existing EGRIN models, as part of a "Network Portal".

VII Education Kits, Curriculum, and Lesson Plans

The education materials developed in our laboratory have been made freely available at http://baliga.systemsbiology.net/drupal/education/. The modules/kits and are typically disseminated free-of-cost during teacher training, pilot tests, and optimization. Subsequently, the kits are available for purchase at-cost of production. These funds are used to purchase reagents for continued low-cost production of kits. That said, we have always supplied kits free-of-cost to disadvantaged public schools, and teachers with limited resources.

morellan 8/2/2013 12:38 PM
Formatted: Left, Indent: Left: 0", Line spacing: multiple 1.15 li