The Development of *Toxoplasma gondii* Systems Biology Data Management and Dissemination Core using XSEDE

David Rhee
Albert Einstein College of Medicine
Bronx, NY 10461
david.rhee@einstein.yu.edu

Matthew Croken
Albert Einstein College of Medicine
Bronx, NY 10461
matthew.croken@phd.einstein.yu.edu

Joseph Hargitai
Albert Einstein College of Medicine
Bronx, NY 10461
joseph.hargitai@einstein.yu.edu

Gos Micklem
University of Cambridge
Cambridge CB2 3EH, United Kingdom
g.micklem@gen.cam.ac.uk

Kami Kim
Albert Einstein College of Medicine
Bronx, NY 10461
kami.kim@einstein.yu.edu

Aaron Golden
Albert Einstein College of Medicine
Bronx, NY 10461
aaron.golden@einstein.yu.edu

ABSTRACT

The advent of massively parallel sequencing technology has provided researchers with the ability to interrogate the genome at a base pair resolution. As proteomics and metabolomics technology has matured in parallel, more researchers are turning to "omics" approaches to ask interesting biological questions on a much larger scale. As a result, we are experiencing a deluge of data that calls for proper data management, dissemination and analysis. This need for a smart 'big data' management solution becomes acute when consortia of individual laboratories work collectively – generating substantial and diverse data sets in multi-faceted investigations using diverse molecular assays. As with any collaborative effort, the viability of such an undertaking is critically dependent on the ability to manage and control the information flow generated. This requires a data management strategy that would not only collate and curate, but also aid in an integrative exploration of these diverse data products providing an 'end-to-end' solution for collaborative projects. The challenges of implementing such strategy require optimally managing the complexity of the different experimental systems and conditions, and the variety of data types that will be generated.

Important issues include: the orderly transfer and processing of raw data; the collection of sufficient associated experimental meta-data to facilitate consortium analysis, and also later use by the broader research community; the integration of all data types to facilitate the exploratory and hypothesis generating work of the consortium participants; the integration of other relevant public datasets; the timely deposition of datasets to public repositories; the timely release of a integrated database; and finally the maintenance of a public website for the project to include access to protocols, datasets and the integrated data warehouse.

We present the *Toxoplasma gondii* Systems Biology Data Management and Dissemination Core, which will operate within the NSF's XSEDE grid computing environment to provide the informatics scaffold and processing environment to coordinate the generation, articulation and storage of the various and diverse data products that will be produced as part of an ongoing -omics collaboration elucidating the role of epigenetics and genetics in regulation of pathogenicity of the parasitic pathogen *Toxoplasma gondii*. The Data Core will automate the absorption of the continually generated data and metadata products from project groups into the proposed data space in a structured and controlled fashion using a custom CRUD system. By adapting the existing and well-proven data warehousing InterMine infrastructure, sophisticated and complex interactions with these data products will be immediately available to the collaborative partners, including the ability to interrogate appropriate bench-related metadata plus existing public datasets. Thus this database provides complementary resources to established community databases such www.toxodb.org, which can only display a fraction of the many high-though put datasets available and does not serve as a repository for data or metadata. This warehousing approach will be further extended by allowing deeper exploratory analysis of the data and metadata products by taking advantage of the Grid computing resources available to the consortium with elements of Einstein's advanced LIMS software infrastructure (WASP System). In this way, additional integrative analyses based on the application of segmentation or clustering algorithms to the originally produced datasets will be possible, and will be immediately re-incorporated into the project's data space. Situated in the center is the Einstein Genome Gateway, a web-based gateway portal, which will serve as the lynchpin that will hold all the components together and make the Data Core fully accessible to all collaborating participants.

Our fundamental system design encapsulates a 'big data' management solution using XSEDE resources. We believe our work can provide a functional model that can be adapted by anyone interested in developing an integrative informatics environment to effectively manage, disseminate and analyze big data sets that are diverse in nature using XSEDE grid computing environment.