IPG2P DATA INTEGRATION WORKING GROUP STATUS REPORT JANUARY 2010

CHRIS JORDAN AND DOREEN WARE

1. Summary

This document presents the initial efforts of the iPlant Genotype to Phenotype Data Integration(DI) working group during the year 2009, along with current plans for the year 2010 as of January. The initial approach defined by Doreen Ware and Chris Jordan is described, along with the results of participation and consultation with various other working groups within the iPlant G2P effort and numerous conversations and e-mails over the course of 2009. It has been prepared by Christopher Jordan in consultation with Doreen Ware, and presents the results of efforts by the entire Working Group. The authors particularly wish to acknowledge the contribution of Jerry Lu with regard to management of the user survey.

2. Goals and Approach

The initial approach taken by the working group co-leads was to base our efforts around the progress and goals of the other working groups within the G2P effort; specifically, due to the fact that the Next Generation Sequencing working group had made the most progress at the time the DI group began deliberations, we chose to focus on ensuring that we met the needs of that working group, and to follow up with efforts to meet the needs of other working groups as their agendas became more clear. In addition to this, we chose to start a parallel effort to focus on archiving data and preserving provenance information, based on the priorities identified by Steve Welch and Goff after their visits to the various sites involved in the iPlant G2P effort.

Chris Jordan took the lead on the task of monitoring the developments in the various working groups, as well as leading the provenance and archiving task. Chris participated in the regular discussions and/or kept track of the presentations and working documents developed by the Next Generation Sequencing, Visual Analytics, Statistical Inference and Modeling working groups over the course of 2009, as well as participating in Steering Committee calls and various discussions with working group leads and participants.

The primary goals of the working group, at a very high level, are to provide input on the general needs related to data infrastructure and data management for the G2P project, and to help select and define data models to ensure that iPlant-developed cyberinfrastructure works with the full range of data types necessary to accomplish its goals, and that the data utilized by various working-group defined components is interoperable to the greatest extent possible. As such, it is assumed that success within the Data Integration working group will be critical to the overall success of the G2P effort, particularly with regard to the ability to draw on reference and user-contributed data sources, and to exchange data amongst the components and tools selected by the various working groups.

3. Primary activities reports

3.1. **Data Source Survey.** The developing workflow definitions from the Next Generation Sequencing effort, as well as the early discussions of other working groups, made it clear that an ability to handle both reference and user-contributed genome, sequence, and annotation data would be a critical first step for the overall effort, as well as being a reasonable entry point into the entire G2P workflow. The DI working group early on determined that it was necessary to understand the landscape of reference data sources, particularly in the area of genomics, and to develop an understanding of the use and variation of format standards for the exchange of such data, in order to make effective recommendations to the community. For this reason, the working group initiated development and distribution of a simple data sources survey, to collect information from some of the primary data sources for G2P workflow users on how they provide data, both in terms of the mechanisms used for selection and access of data and the formats in which data is provided.

Jerry Lu took the lead in developing this survey, and the DI working group solicited input both within the working group and from the project participants in general on the data sources and types of most interest. This resulted in a still-growing list of data sources, which is being maintained by Chris Jordan, and an initial round of survey results from some of the largest and most important genome data providers (SolGenomics, Gramene, TAIR, and MaizeGDB). Initial survey responses made it clear that the overall survey results would be more useful if a longer and more specific survey was offered to a broader set of data resources, with more specific prompts for the information we are seeking to gather. This longer survey is currently in development, as well as the list of data providers to be surveyed, and input is being actively sought from the entire G2P community on both the survey contents and the data providers to be surveyed. The spreadsheet listing data sources is expected to continue to grow, and to include various relevant forms of information as the Data Integration group defines its goals and requirements in more detail.

3.2. **Workflow Support.** The Next Generation Sequencing workflows that have been presented thus far to the data integration group will almost exclusively work with tools designed for various types of genome-related data, using relatively standard formats such as GFF3. For these types of tools, there will be limited ability to modify the format of data coming into the pipeline, and provenance and other metadata would have to be associated with the data files through an external mechanism, potentially either XML files or a database table associating files and/or URLs with metadata. However, the Visual Analytics workflows that have been discussed thus far will use more general types of algorithms and tools used to visualize numeric data in N-dimensions; wrapper scripts can be utilized to implement some of the required semantic interpretation, but in many cases it will eventually be required to present simplified versions of various biological data types to enable their use by visualization and statistical analysis tools in particular.

A data model is under development by Chris Jordan in consultation with Greg Abrams, Jerry Lu, Ruth Grene, Damian Gessler, and many others throughout the G2P effort, which would focus on the ability to interpret most standard biological data formats, and some degree of variation on those formats, and to convert these into a standard tabular or n-dimensional array format, with wrapper components providing any necessary semantic interpretation to workflow tools, enabling common data analysis and visualization algorithms and components to utilize this data without having to be modified to understand biological formats and ontologies. The metadata handling, and semantic processing of data in general, would be implemented as a kind of middleware layer within the iPlant workflow tools, thus enabling iPlant to utilize and/or provide any level of sophistication in provenance and other metadata, as well as enabling the use of the many rich ontologies available in the life sciences, without requiring that every component used by iPlant workflows understand all of these formats and ontologies.

A significant issue that must be addressed in order to facilitate further progress in the workflow area by the DI group is whether there will be one workflow tool/system utilized by iPlant, or each working group can/may select their own. There have been evaluation efforts and much discussion in the working groups of workflow systems such as Taverna, Pegasus, and Condor, and in the Visualization group there has been a focus on VizTrails, but none of these tools would seem to directly provide the type of flexible, iPlant-specific, web-enabled interface which is a goal of the overall iPlant project as we understand it, and the overall effort of the DI working group and the core team in particular will be much more difficult if data integration tools must be provided for multiple workflow systems. iPlant should move as quickly as possible to select a workflow tool, or select a workflow language such as BPEL to be utilized by custom-developed, web-based workflow systems which could be implemented by the core team. The DI group feels strongly that this lack of clarity on the workflow tools to be utilized by iPlant cyberinfrastructure will soon become an impediment to further progress.

3.3. **Archival and Provenance.** A separate but important concern being addressed by Chris Jordan and others within the DI working group is the need to define a process and criteria to handle archival of iPlant-generated data. Specifically, if the data model currently in development is utilized, there will be significant advantages for both performance and reproducibility if any data conversions from standard resources and formats made by iPlant tools are also archived for later reuse. This may be particularly important for the standard data resources which may be constantly in flux, and for which "versioning" is critically important. There may also be a possibility for iPlant to act as a "conduit" for user-generated data to be submitted to iPlant and through iPlant, offered to appropriate reference repositories. TACC and iPlant in combination have sufficient infrastructure to be able to handle archival and access to many terabytes of data generated or processed by the iPlant project, but policies will need to be developed to enable the iPlant cyberinfrastructure itself to make decisions about when to retain data submitted to or converted by iPlant CI based on user actions.

Related to this issue is the question of provenance metadata; a team is currently in the process of formation to focus on this issue, and iPlant G2P participants are actively encouraged to contact Chris Jordan if they would like to participate in this effort. There are multiple issues that will be addressed by this team:

(1) Definition of minimum standards for data quality, and minimum requirements for the level of provenance information required for ingest and use of data by the iPlant CI - for example, is it required that we know the location of an experiment, the specific equipment used, etc.
(2) Selection and/or definition of formats and controlled vocabularies for the description of experimental parameters and other relevant provenance information.
(3) The level of detail with which actions taken on a specific piece of data, or the digital "path" that generated a specific visualization, graph, or plant model, will be recorded. Decisions must be made on whether the goal is to enable a reconstruction of the set of operations used to generate an image, for example, or simply to enable understanding of how a certain result was obtained.
(4) The extent to which the archival system can and will be used to enable reproduction of digital "experiments" conducted within iPlant. For reference data sources where versioning is difficult or impossible to do, and data retrieved by a given query may change from day to day, iPlant can archive and provide it's own "versioning" system for given data retrievals, i.e. "Maize Genome data as of January 5th, 2010 at 8AM GMT", but clear boundaries must be defined.

Project leadership will need to be involved to some extent in defining and validating decisions on issues like data quality and mandatory minimums for provenance information. It is hoped that broad representation from across the project working groups can be achieved in this team and that consensus can be quickly reached on requirements to enable sustainable, usable, and scientifically sound cyberinfrastructure.