Gramene Scientific Advisory Board Report and Responses

Synopsis

This is the report of the 2009 Gramene Scientific Advisory committee from the 6 hour WebEx conference held December 16, 2009, and submitted January 2010. Gramene responses to the report are included inline.

The agenda of the meeting included reports on the progress of each of the 5 specific Aims, internal discussion among the SAB, and General discussion with the team and the PIs individually.

SAB Report

Introduction

The SAB have been impressed with the progress made over the last year, in particular, given the constraints that the team has faced with respect to available funding, delays in recruitment and the volume of new information. The Gramene project plays a significant role in and acts as a primary focus for the integration of rice information and its relationship with key cereal and more general plant species. With developments in new high throughput technologies for sequencing, genotyping and expression analysis and their application to a broader range of species and genotypes we anticipate that the value of this type of integration activity will increase dramatically in the next few years.

Specific Aim 1 (Genomes)

With the growing list of major additions to the set of sequenced plant genomes a major challenge for Gramene is how to prioritize the addition of new genomes and balance this against the need to actively curate the existing portfolio. The SAB appreciates the issues involved and would suggest that the following genomes should be considered:

1. The Brachypodium genome will be publically released shortly and, in the near term, provide the closest model genome to a number of key crop plants including wheat, barley and forage grasses.

Gramene response: The Brachypodium is currently in production within Gramene and will continue to be hosted by the project. In addition to hosting of the genome, there are currently on going discussions for collaboration on curating pathways.

A number of legume species are currently sequenced or in production, and one of these should be identified and supported. The options are to make the choice based on either their use as a model system (e.g. Medicago) or because of their inherent agronomic value (e.g. soybean – though this is clearly complicated by polyploidy).

> The sequencing of the related Solanaceous crops, potato and tomato, are both at a relatively advanced stage and the choice of one of these would provide a route into an extremely valuable crop taxonomic group.

Gramene Response. We agree that there will be value in hosting a legume and Solanaceaous species within the Gramene framework. For this to move forward we will need to work with both the communities and funding agencies to coordinate this effort. We anticipate the ability to bring in at least one genome in the next year and will target one or both of these genomes.

> In terms of lower plants, Selaginella appears to have some attraction because of its role as a vascular plant model.

Gramene Response: Gramene would like to support the integration of one lower species to support the evolution analysis. We will be reviewing the Seleaginella and Physcomitrella for a release in the next year.

> With the latest methods of transcriptome profiling utilizing next generation sequencing technology, it would be valuable to add an additional track to the genome browser to display this type of data and to encourage the community to provide such data to Gramene for improving current genome annotation.

Gramene response: We believe that it will be possible to host these tracks as DAS tracks and exploring requirements to streamline this process.

With the expansion in the number of sequenced plant genomes that is now, or will shortly become available, it is of major importance that Gramene is able to deliver to the user community facile and robust tools that will enable them to efficiently exploit comparative information. Currently both gene trees and CMap provide support for this activity but it is important that Gramene monitors whether these tools are evolving sufficiently fast and have the capacity to meet user needs.

Gramene response: We acknowledge the SAB recommendations regarding the improvements in tools and the need for usability as the data scales.

Specific Aim 2 (Pathways)

Gramene has ambitious aims with respect to the curation of metabolic, regulatory and other pathways. While generally congratulating Gramene's work in this field, the SAB must express some concerns. The first is that the Gramene group is working in the context of three quite different systems with respect to pathway curation. The first is the Cyc family of databases maintained by the SRI and requiring the use of proprietary tools. The second is within the context of the Reactome family of databases maintained at CSHL and the EBI; the third is within the context of the community WikiPathways efforts. Each of these has particular advantages and disadvantages (e.g. the Cyc databases

cannot handle regulatory pathways). However, with limited resources this diversity means a dilution of effort. It is our strong recommendation that, as a matter of priority, Gramene should settle on one system (and our preference would be for Reactome) and simply allow the Cyc and WikiPathway databases to be populated by automated exports using BioPax.

Gramene response: We acknowledge SAB suggestions and concerns regarding the focus of our efforts. In this reference we have established collaboration with the Reactome database and carried out a couple of test cases of importing the curated pathways from the existing RiceCyc database. After a careful adjustments to the Reactome's BioPax based retrieval of pathway information tools, we are looking forward for a bulk import of the current curated information in the Reactome. We expect that after a successful import and quality checks we will be able to release the first Rice Reactome by the end of the fall of 2010. During this time the curators will be mentored by the Reactome staff for their training and curatorial help. After the release of Rice Reactome we will look forward to integrate the additional cereal plant reactomes by making use of the gene product annotations based on the species specific manual curation and those derived largely by the phylogeny and synteny based methods applied in specific aim-1. During this move we expect that our users will have to learn the changes to the user interface and content. We will make every effort by way of outreach and tutorials to help our users get through the transition.

Specific Aim 3 (Diversity)

The Gramene Diversity group clearly faces particular challenges with respect to the growing volume of genotype and diversity sequence data that is becoming available for its major target organisms. There is a need to clarify and advertise just what datasets are, or will become, available and then to identify what likely users of the data will require. A major complication of this type of diversity data is that many of the potential users will aim to exploit it in conjunction with some further biological analysis of the germplasm from which the genotype data has been derived. This leads in turn to the need to grasp and resolve the complications that arise from the relationship between accessions in genebank terminology and true-breeding genetic stocks. It is important that precise seed sources directly associated with the biological entity that was sequenced or genotyped are documented and ideally available as a "set" from a single source. Recent changes in International Law with regard to the movement of germplasm may complicate the issue further. However we note that this issue is not a Gramene specific one but it is one that should be acknowledged.

Serving the diversity data is also a complex problem. Many users may not be well prepared to receive the large data sets and may well lack the software tools to hold and analyze or visualize the data they download. We welcome the moves to provide key software tools to support this activity but we also anticipate that many users will require significant support in their deployment.

Gramene response. We are working on getting these tools working and there becoming better integration. Direct launch now works, but needs to be deployed. Wizards have been piloted.

It is also not yet clear what the end game is for Gramene with respect to the analysis of diversity data sets. Is it the intention to not only store the genotype data but also phenotype data and the results of analysis?

Gramene response: We already store genotypic and phenotypic data, although recent curation by Gramene curators has focused on genotypes. A shift to more phenotypic curation will occur over the coming year. The database schema is being modified for storing results, but we have not implemented yet. We expect implement by summer and fall of 2010.

If this is the case then considerable thought will need to be given to how the results of this analysis are stored and displayed. There are a number of interesting options. For example, for the key target character of flowering time displaying the relationship of the genotype by trait analysis with potential candidate genes and response pathways could provide an interesting challenge.

Specific Aim 5 (Outreach)

The main issue with regard to the Gramene outreach activities appears to be one of identifying and supporting one or more core audiences. Is the target audience molecular geneticists or is it plant breeders? Is it both? For US breeders it would appear that the USDA CAP projects or comparable projects and their associated informatics may be the most appropriate route to serve much of the breeding community. Are plant breeders a reasonable target audience? If they are, then what would a plant breeder really need? Is the ability to identify allelic variants associated with a desirable phenotype sufficient? Should this be pursued through a complementary funding mechanism? It might be good to evaluate the stakeholder communities and determine what they use. If the plant breeders, particularly the commercial plant breeders in the US, are conducting data dumps, then user interfaces and analytical tools may not need to be developed by Gramene. This is true for maize breeders, but not for other crops supported by Gramene and not for the international maize breeders.

Gramene response: This is an excellent point raised by the SAB and we acknowledge their concern. In order to build a database portal that is applicable to a wide array of scientists, Gramene has strived for the last several years to stand out from rest of the community plant genome databases, most of which are species-specific. This has been done by careful development, curation and integration of information from both reverse and forward genetics studies to create a confluence of information streams. This way, different, research communities get a chance to link the pieces of their puzzles, and move in either direction to strengthen and validate their hypotheses. This effort has been acknowledge by users in both the breeding and genetics communities. For example, Yamamoto et al (http://news.gramene.org/?p=482 In 2009). acknowledged Gramene's

QTL resources in a paper in DNA Research entitled "Towards the Understanding of Complex Traits in Rice: Substantially or Superficially?", and in 2010, a paper published in BMC Plant Biology entitled "Transcriptional regulatory network triggered by oxidative signals configures the early response mechanisms of japonica rice to chilling stress" noted Gramene's usefulness to understanding rice transcription factors. Similarly Muylle (2005) identified of four QTLs that determine crown rust (Puccinia coronata f. sp. lolii) resistance in a perennial ryegrass (Lolium perenne) population and Armstead et al (2004 and 2005) employed fine-mapping strategies to identify candidate genes for crown rust (Puccinia coronata) resistance from meadow fescue (Festuca pratensis) which were intriogressed into Italian ryegrass (Lolium multiflorum). Both of these studies used the rice genome annotation and comparative map tools provided by Gramene as well as the species-specific markers and maps provided by its collaborators to identify the molecular markers that were used in their breeding strategies. Thus, by delineating regions, genes and traits of interest within a comparative framework, Gramene has consistently contributed essential tools and knowledge to the development of novel approaches in genetics and breeding, and has facilitated the use of integrated comparative genomics and genetics approaches towards the betterment of our US agriculture. With these words we would like to say that despite all the clutter of data set deluge, Gramene has been able to provide a clean and well annotated reference datasets and was able to identify or rather has avoided identifying any precise core users by letting the communities create a self learned as well as contributed by our extensive outreach programs lead to the adapted niche among its users for its presence and value for not just the reference model crop plants but for others with lesser known resources.

For future releases, Gramene aims to address some specific requests that have come from the US breeding and genetics communities, while at the same time, encouraging the independent development of more specialized applications and user interfaces for specific user communities that may "dock" to Gramene or extract information from Gramene. Based on conversations with the US rice breeding community during 2009, we are developing ways for Gramene users to search on phenotype and find links to specific alleles and to germplasm resources known to carry those alleles. To develop this capacity, we have recently hired a new curator who will address this issue during 2010. Similarly, geneticists expressed their interest in being able to search on allele, and find links to genes, germplasm and phenotype, and we also aim to implement this capability during 2010. These improvements, along with enhanced visualization tools, statistical approaches and curated datasets will provide Gramene users with additional opportunities to link phenotype with genotype and aim to enhance the utility of Gramene as an information resource for the breeding and genetics communities, while continuing to support the powerful comparative framework previously developed for genes, proteins, phenotypes and pathways.

With the resources available to Gramene it is important that they are focused on realistic targets and a better understanding of the actual audience may help to achieve this. A related issue is to clarify relationships with other projects, for example, MaizeGDB, the MSU rice project, PlantGDB and PlexDB. To many in Gramene's external audience it is not at all clear what the relationships and responsibilities are. Perhaps in conjunction with

PI's from such projects it would be a good time to begin to develop a forward looking white-paper that explores the needs and solutions for these and other related projects at both a national and international level. Such a framework would help clarify the investment in resources across a broad spectrum of activities and, with respect to the Gramene group, help clarify the balance in effort and future funding applications between a single fully integrated project and individual more focused projects working within an "agreed" framework. Reference to such a document would also help grant reviewers enormously in understanding the context of any future applications.

Gramene Response: For each of the specific Aims within the Gramene proposal objectives support basic research as well as providing a framework for integration of community data sets. For each of these objectives there was identified existing as well as anticipated data sets that would be utilized in the project. The Gramene project also works directly with new projects as they are being developed with the objective of identifying appropriate data set for future integration and the resources to support these. In a better effort to communicate our existing relationships with projects, we will actively review the collaborators page and make an effort to highlight these collaborations in news posts. With regard to exploring the existing as well as future challenges of coordination and stewardship of plant related data sets the PIs on this project are actively participating in several workshops addressing these challenges at the national and international level. We anticipate that these workshops will result in a series of recommendations to address many of the concerns expressed by the SAB.

In more specific terms, we strongly endorse the re-evaluation of the Open Helix collaboration and the change in the way Gramene users are to be supported through a change in format to mini tutorials, perhaps structured around an FAQ and linked to short Flash or comparable video shorts, dealing with specific tasks. We also believe that individual module developers should, when possible, work on outreach directly, both because of the unique aspects of each area of activity and the value of feedback from their specific target audience.

General Issues

The SAB have some concerns that too much is being expected, by Gramene, from community curation. This has been a major concern in the model organism database community for some years now and no great progress has been made, other than the collaboration made between TAIR and some journals. There seems to be a wish to emulate this within Gramene, however, no evaluation of the TAIR experiment has been made, as far as we can determine. If database quality is to be maintained (as we are sure it will be) any community annotation must be curated and we question whether, in fact, this would really save effort compared with direct curation from the literature. In addition there are concerns as to how members of the community will become the owners of curated information and how credit/benefit from such activities be realized.

Gramene response: According to our discussions with the TAIR staff, the submissions from the publisher's site (data is primarily collected by the publisher), have increased

gradually. Also the submission form allows users (authors mainly) to self select the ontology based annotations in the form. These annotations are often accurate by using either the precise Gene Ontology and Plant Ontology term or a generic by selecting a top level parent term. According to the TAIR curators, the major time consuming step is the curation of dependencies which are unique to a given database structure, e.g. adding the citation to the literature database, users name, etc. prior to its usual import from PubMed in the database on our/their end, creating map positions, etc. Based on TAIR's experience and our discussions with ASPB, we think a substantially improved data flow can be generated for more automation and responsibility on part of the authors.

This SAB report is a good opportunity to register our concern that NSF is backing off on infrastructure support for essential community databases and stock centers (e.g., TAIR and others). iPlant is not going to fill this need! A greater investment rather than a reduced investment should be the priority, as all research becomes more dependent on networked datasets in need of curation. It is unrealistic to transition to fee-based or private centers. This should be a public-supported library-type system accessible to all. The apparent new directive for NSF to focus more on "mission-based" projects makes the deterioration of infrastructure support an even greater concern. Clearly it may also be reasonable to look to collaboration with other international funding agencies to support some aspects of this type of activity. However there is a growing need to develop and support the type of resource that Gramene represents and both mid and long term funding solutions need to be developed.

With the major increase in volume and complexity of sequence-based data sets that are already available and the rapid growth in scale that Next Generation and further novel sequencing technologies bring, both the challenge and value of data integration will only increase. Funding agencies need to take on board the realization that the value of their investment in these technologies will be significantly reduced unless comparable resources are also invested in the necessary informatics infrastructure and integration.