TRPGR Gramene: A Platform for Comparative Plant Genomics

Introductory comments: In light of the comments from the review panel, our scientific advisory board, and budget priorities, we have worked to tighten the focus on this renewal. One of the strengths of Gramene is its leveraging of the Ensembl platform developed in the human genetics community. This platform is very feature rich, but it provides its best support for fully sequenced genomes. Therefore, the focus of this project will be to address the following question:

"*What are the functionally shared elements of plant genomes and how does diversity in these elements relate to agronomically valuable traits?*"

The three research aims of the project are designed to answer this question: (1) Using comparative genomics to identify functional elements and sequence variants that may have phenotypic consequences; (2) Annotating biological pathways in order to provide the infrastructure to understand how those sequence variants lead to phenotypes; (3) Collecting and uniformly reanalyzing QTL and diversity data in order to connect genetic diversity to phenotypic variation. These three aims form a tripod which cannot stand up if any leg is removed. The comparative maps and genome builds are essential to allow for the transfer of information from one species to the next, for the integration of QTL and diversity data, and for the dissemination of the third party functional data sets listed in specific aim 1. The pathway collection and annotation of specific aim 2 is essential for the biological interpretation of QTL and diversity data. The collection, integration and reanalysis of QTL and diversity data is necessary to identify the relationship between genotypic and phenotypic variation.

Over the next four years, only four plant species will likely have the complete genome sequence, functional and diversity data needed to address the question posed above. Therefore, this project will focus on rice, maize, sorghum, and Arabidopsis. Wheat, barley, and multiple biofuel species may provide interesting additions, but these will have to be funded through collaborations or federations with the EU, US DOE, and/or USDA support.

Additionally, we view that integration of expression profiling data is a very important area, but it will have to be accomplished through a separate proposal and through a series of collaborations, as the scope is so large. The Plant Ontology (PO) consortium has currently been absorbed by Gramene for technical funding reasons, but this excellent project should be a stand-alone project. PO will be supported by Gramene for a period of two years, during which time we (in collaboration with other community members) will apply for standalone funding

Comment [ed1]: Is this actually in the budget or not?

Specific Aim #1: Provide an infrastructure of comparative genomic data to allow for the mining and analysis of functional data on the genomes of rice and other monocots.

Rationale: The comparative mapping and integration of genomic sequence, genetic maps, and physical maps, adds significant value to existing genome sequencing and mapping studies, by allowing the transfer of information across and within species. For example, let's say that a future allelic diversity study in hexaploid wheat has identified a region that shows a clear domestication signal. By following syntenic relationships between wheat, rice and maize, a researcher can find candidate genes that are in the syntenic regions of one or both of these genomes. He can then take advantage of functional data sets that the Gramene staff has mapped to the genome to narrow the search, by filtering for genes that expressed in relevant tissues, whose knockout produces relevant phenotypes, which partake in a pathway of interest, or which are likely to be relevant due to their gene or plant ontology annotations.

Gramene is the only platform for plant research that enables both forward and reverse genetics. Researchers can leverage Gramene for traditional forward genetics to find candidate genes responsible for genetically-mapped traits, or use its collection of structural variation information (SNPs, diversity, artificial alleles) and orthology data to deduce the function of genes.

Approach: We will continue to build and maintain comparisons among the maps of the following species: *O. sativa* and 11 wild *Oryza* species (*alta, rufipogon, glaberrima, nivara, minuta, officinalis, australiensis, brachyantha, granulata, ridleyi, coarcta), Z. mays, S. bicolor, and Arabidopsis*. The comparative maps will be viewable in the following browsers:

- CMap, for aligments among multiple maps of different types.
- Ensembl Compara, for protein-based alignments between two sequenced genomes.
- Ensembl CytoView, for physical maps (e.g. fingerprint maps)
- Ensembl ContigView, for detailed nucleotide-level features
- Gene and Protein views, for detailed information on annotated gene structure and protein function.

We will integrate into this framework the following future sequence-based datasets from NSF and USDA-funded projects (subject to the funding of the proposed projects):

- Rice SNP variants and diversity information from 10,000 well distributed SNPs in 400 *O. sativa* and 100 *O. rufipogon* accessions (Susan McCouch, Cornell)
- Rice protein-protein physical interaction data, derived from yeast two-hybrid screens (David Hill, Dana-Farber Cancer Institute).
- Rice QTLs affecting response and processing of minerals, from the "Ionome" project (David Salt, Purdue University).

- Rice SNP variants and diversity information from *O. sativa* and *O. rufipogon* from seven targeted large genomic regions (Ana Caceido, University of Massachusetts).
- Optical map of rice (David Schwartz, Univ. Wisconsin)
- Annotated sorghum gene models, candidate SSR and CISP primers, strain-diagnostic DNA markers for 400 discrete sorghum mutants, photos and other phenotypic information on 3200 M3 lines, and 400 marker genotypes for two new sorghum RIL populations (Paterson and Gingle, University of Arizona).
- Resequencing diversity data on *O. sativa* and *O. rufipogon* consisting of ~650 gene fragments from a panel of *O. sativa* accessions, ~396 gene fragments from *O. rufipogon*, SNP and indel polymorphism data derived from these resequenced fragments, and 60 fully sequenced BACs from *O. rufipogon* (Michael Purugganan, NYU).
- Sequencing of the short arm of chromosome 1 from 11 wild *Oryza* species (Rod Wing, University of Arizona).
- Next generation sequencing for SNPs in 27 key maize lines (Buckler, Cornell University).

We use modified versions of the Ensembl pipeline, Ensembl database, and Ensembl web site to manage, analyze and display these data sets. Additional data sets will be added in future years as they arise.

We are able to support these and future data sets in a scalable manner because we have established a formal “pay as you go” system in which the data submitter agrees to provide a bioinformatics liaison on their end to cleanse the data and place it in a standardized format for submission to Gramene. These agreements are contained in written “contracts” contained within the letters of collaboration that we write for each participating project. The cost of curation of each of these projects is therefore supported in large part by the individual projects. This greatly simplifies our work and allows us to maintain a core operation to manage these data sets, align them to the genome, quality control them, and, in some cases, to develop special visualization tools to display them. Therefore we can handle increasing data loads without a concomitant increase in our curational staff.

Whole genome alignments (WGA) are a special type of comparative genomic map with nucleotide-level resolution. They are particularly useful for the identification of functional genomic elements because regions that are under purifying evolutionary selection show different patterns of conservation that neutrally evolving regions. In recent years, WGAs have been shown to be invaluable for identifying non-coding RNA genes, cis-regulatory regions, and other functional elements in animal genomes. They are also invaluable for understanding genome evolution.

We will use the Ensembl Compara pipeline to perform and publish WGAs of cereals and several model genomes to provide a multiple alignment involving the agronomically important species of rice, maize and sorghum, against the reference genomes of brachypodium, poplar, and arabidopsis. The rationale for providing alignments against arabidopsis and poplar is that the evolutionary distance between these species and the monocots will help distinguish evolutionary conserved elements from regions that are slowly evolving due to chance. The rationale for including *brachypodium distachyon* is its emerging role as a model monocot. This multiple alignment will be downloadable in multiple sequence alignment (MAF) format, and browsable via the Ensembl genome browser in a format that indicates the locations of evolutionarily conserved segments in the context of genome annotations.

In addition to the visualization of the comparative maps on the Gramene browsers, we will provide access to the data via:

- Bulk downloads of filtered and unfiltered data sets
- GrameneMart, a flexible database query and report engine based on BioMart (www.biomart.org)
- DAS, a web service for sharing genome annotations across the Internet (www.biodas.org) and used heavily by Ensembl for vertebrate annotations.
- QTL data will also be shared via SSWAP, a semantic web service developed for use in the Virtual Plant Information Network (http://vpin.ncgr.org/semanticmoby.shtml)

Release Schedule: We propose to reduce the number of Gramene releases to two per year; currently we release quarterly. This less frequent release schedule will allow curators and software developers to spend time on community outreach, support, and infrastructure maintenance.

Personnel Required to Support this Aim: We require 2.5 full time software developers, half a curator, and 25% of a system administrator to keep the core operational.

- **Map manager (100%)**; this developer is responsible for the data processing of all genetic and physical maps, QTL maps, and assignment of SNPs, markers and diversity information to the genome. This individual is also responsible for maintaining the CMap viewer and the database.
- **Map curator (75%);** this curator (a PhD-level biologist) is responsible for literature-based acquisition of QTL and genetic map information, for monitoring the processing of map data by the map manager, and for making decisions as to how the map data should be labeled and displayed. This curator is also responsible (25% effort) for monitoring and maintaining the Gramene community WIKI (described under outreach).

- **Build manager (75%);** this developer manages the genome builds, in which sequence-based datasets are aligned to the genome, integrated, and uniquified. This individual will do double duty by maintaining the web site infrastructure, the protein-based synteny maps, GrameneMart, and the DAS and SSWAP web services.
- **Website developer & WGA analyst (100%);** this developer will devote half his effort to enhancing Gramene by adding enhancing the web site and database by adding and improving user interfaces, adding new visualizations, improving data mining facilities, and working with affiliated databases to improve develop data exchange and integration protocols. The other half of his effort will be spent performing whole genome alignments among sequenced cereals and reference genomes.
- **System administrator (25%);** responsible for hardware maintenance, backing up datasets, installing security and other patches, managing the network, and maintaining user accounts.

Dr. Ware will supervise this portion of the effort.

Deliverables (Aim 1):

- Comparative genetic, physical and sequence maps of *O. sativa* and 11 wild *Oryza* species (*alta, rufipogon, glaberrima, nivara, minuta, officinalis, australiensis, brachyantha, granulata, ridleyi, coarcta*), *Z. mays, S. bicolor,* and *Arabidopsis.*
- Nucleotide-level genome annotation views of these species, including the functional consequence of SNPs, using Ensembl framework.
- Physical (contig) map views of these species, using Ensembl framework.
- Protein comparisons among these species.
- Detailed gene and protein level views of the annotations of these species.
- Incorporation of roughly 9 functional genomics data sets in various species generated by NSF and USDA-funded groups listed in the main text.
- Whole genome alignments of rice, maize, sorghum, *brachypodium*, poplar and *arabidopsis,* refreshed annually.
- Twice-yearly releases via web browser, FTP bulk downloads, GrameneMart data warehouse queries, DAS and SSWAP.

Specific Aim #2: Enhance the value of the comparative maps with pathway, phenotypic and other functional information from rice, maize, and Arabidopsis.

Rationale: The comparative maps developed in specific aim #1 are not particularly useful in isolation, but become extremely valuable when combined with information on the function of genes, proteins and regulatory regions, temporal and spatial expression patterns, natural and artificial variation in the genome, and the phenotypic consequences of this variation.

Our past work has created a rich foundation of functional information consisting of a comprehensive one-pass annotation of rice gene products using the gene ontology, an ontology of terms to describe plant phenotype, and a well-annotated set of associations between naturally-occurring strains, mutants, and ontology terms describing alterations in their phenotypes. However, these functional descriptions are over-simplified; they consist of assertions about single gene products, whereas in the real world phenotypic effects are the result of assemblages of gene products acting and interacting in concert.

During 2006-2007, we developed an infrastructure, *RiceCyc*, for describing and annotating plant biological pathways, and used this infrastructure to create a database of intermediary metabolism in *O. sativa japonica* containing 1700 reactions and their associated enzymes in 323 pathways. While this resource is valuable, many genes of agronomic and biological interest are not involved in intermediary metabolism but in higher order processes such as regulation of growth, intracellular signaling, defense, and response to the environment.

Approach: We will extend our work on RiceCyc to higher order pathways involving growth, regulation, defense, intra- and inter-cellular signaling, transport, and response to the environment, prioritizing pathways to those most likely involved in the agronomically-important traits of flowering time, inflorescence development, response to abiotic and biotic stress, and carbohydrate metabolism.

To do this, we will make automated pathway inferences using the Reactome software system (www.reactome.org), part of an NIH-funded pathway database that focuses on cell biology. This project is a collaboration between Stein, Ewan Birney of EBI, and Peter D'Eustachio of NYU. Reactome has a mature and validated set of tools for inferring pathways based on protein orthology, a suite of authoring and curation tools, and a visualization engine that allows large data sets, such as microarray expression studies, diversity/association studies, and QTLs, to be superimposed on top of pathway diagrams for statistical analysis of overrepresentation. These inferred pathways will be integrated into the intermediary metabolism pathways currently in RiceCyc via BioPax (www.biopax.org), a file format that has become the standard for representing biological pathway data. We anticipate being able to infer pathways involving at least 2000 genes in each of rice, sorghum, and maize. (Arabidopsis is already available via AraCyc).

Subsequently, over the course of the project we will hand-curate an additional 120 biological pathways in rice among selected signaling and developmental pathways using the Reactome authoring and curational tools. These hand-curated rice pathways will then be inferred automatically in sorghum, maize and Arabidopsis. Community curation of these pathways can be performed by community members via the pathways WIKI described later.

The following are representative of the larger set of high priority pathways that we will curate:

- Starch Metabolism
- Flowering time
- Plant Hormone and derivative synthesis pathways
- Heavy metal-induced signal transduction
- Plant defense (host pathogen interaction mediated) signaling pathway

The pathways will be integrated into the Gramene infrastructure via a graphical pathway browser, a query and search engine, and the “skypainter” tool, an interactive tool for identifying pathways which are statistically overrepresented in collections of genes, such as those that are under a set of related QTLs. These tools are all part of the current Reactome software.

The pathways we develop will be integrated into other pathway datasets, thereby increasing their value to the community by allowing for comparisons among plant and animal species. Using BioPax as the standard exchange format, we will merge the pathways into the holdings of Reactome, with Pathway Commons (www.pathwaycommons.org), a shared repository for pathway data, with the Plant Pathways Database (PI, Sue Rhee, Carnegie Institute), a project currently under consideration for NSF funding, and with Arabidopsis Reactome (arabidopsisreactome.org), a pathways database at the John Innes Centre in the UK that also uses the Reactome software infrastructure. The sharing of pathways among multiple databases will give researchers great flexibility in their choice of user interfaces, search and analysis tools, and will foster the development of computational models based on molecular pathway data.

It is important to understand that the proposed pathway annotation activity is complementary to the Carnegie effort. The latter focuses on intermediary metabolism and similar biochemical pathways that involve small molecules. The Gramene effort will focus on the genetic, cellular and developmental pathways that are most likely to shed light on the mechanisms underlying QTLs.

Accurate pathway curation is highly labor intensive, due to the need to document each reaction with a primary citation, to preserve the chain of evidence when inferring reactions based on cross-species data, and to accurately identify the gene products that participate in the reaction. In order to accelerate the acquisition of useful pathways, we will open a Gramene Pathway WIKI to allow for community annotation of pathway data. The WIKI will be based on an infrastructure currently being developed for Reactome that allows community members to build pages that describe pathways and pathway components. The infrastructure is based on the MediaWiki infrastructure used in Wikipedia (www.wikipedia.org), but has extensions that allow community members to make structured assertions. One extension is a graphical reaction editor allows researchers to draw pathways using conventional representations for biochemical reactions, genetic regulatory pathways, signaling pathways, and transcriptional cascades. This information is stored in a structured manner in a backend database using the Reactome data model. Another extension is a Table Editor, recently developed by the software engineers at EcoliHub (http://www.ecolicommunity.org), that allows controlled vocabulary terms, typed fields and other structured statements to be embedded into otherwise conventionsl WIKI pages.

Community-contributed pathways will be available for browsing and searching side by side with internally curated pathways, but the two will be distinguished in the public interface so that researchers can treat them differently if they so choose. Any member of the community will be able to edit the WIKI. To avoid spam and other abuse, we will use “Turing tests” such as the reCAPTCHA system (http://recaptcha.net) to intercept robots.

Protein and gene function information. In the past, we have devoted considerable effort to the hand annotation of rice gene products using gene ontology terms, protein family relationships, functional domains, and known variants. As this work has largely been superseded by the rice genome annotation efforts at the J. Craig Venter Institute (formerly known as The Institute for Genomic Research), we will no longer perform this task, but will instead automatically import this information from the JCVI Rice Genome Annotation project. The Rice Genome Annotation project is currently slated to terminate at the end of 2007, and we do not know its renewal status; should it terminate, we will obtain gene product updates from the high volume curation provided by UniProt (http://ca.expasy.org/sprot/).

There is considerable collective community knowledge about the function of gene products that never makes it into UniProt, however. Therefore, we propose, as an adjunct to the Gramene Pathway WIKI, a self-sustaining community forum to describe other aspects of gene structure and function. The Gramene Pathway WIKI will include a series of gene product pages that describe their subcellular location, biologic function, evolutionary relationships, variations, and the phenotypic effects of those variations. The community-supported gene pages will use the Table Editor in order to capture and store controlled vocabulary terms and other structured assertions into the Gramene database. Like the rest of the Gramene Pathway WIKI, this facility will be open to all members of the community, but monitored to control spam and other abuse.

We currently import rice gene models (exon structure and alternative splicing information) from the JCVI Rice Genome Annotation project and will continue to do so until that project terminates. It is obviously of considerable importance to have high quality gene models and for there to be a mechanism by which the community can report errors in models and make corrections. Unfortunately, we cannot support this activity.

Personnel Required to Support this Aim: We require a full time bioinformaticist and a curator in order to acquire and maintain biological pathways and to manage the gene function WIKI.

- **Pathway database developer (100%)**; this bioinformaticist is responsible for the pathway database, the automatic pathway inference pipeline, maintaining the curational tools and the user interfaces, ensuring seamless connectivity to the diversity module, with particular attention to QTLs. This developer will manage the import from and export to affiliated pathway databases. She will also be responsible for administering the Gramene Pathway and Gene Function WIKI.
- **Pathway curator (75%)**; this curator (a PhD-level biologist) is responsible for prioritizing, acquiring, and updating biological pathways in rice from the literature and from comparative genomics sources. He or she will act as the primary liaison to the Plant Pathway Database, The Pathway Commons, Reactome and Arabidopsis Reactome. The curator's remaining 25% effort will be devoted to maintaining the Gramene WIKI Pathway pages, and has been placed in the Outreach section.

Dr. Jaiswal Pankaj will act as the pathway curator for this effort in addition to supervising the pathway database administrator.

Deliverables (Aim 2):

- Automated pathway inferences of pathways involving greater than 2000 genes in each of rice, sorghum and maize.
- Web-based tools to visualize similarities and differences among pathways of rice, sorghum, maize and *Arabidopsis*.
- Hand-curation of at least 120 rice pathways involving agronomically significant traits.
- A community WIKI for curating additional rice, sorghum and maize pathways.
- A community WIKI for annotating the function of rice, sorghum and maize genes.

Specific Aim #3: We will acquire genotypic and phenotypic diversity data for each of the sequenced monocot genomes. We will recalculate this data using a standardized methodology that allows us to integrate the QTL values across species and to relate phenotypic diversity to candidate genes via pathway information.

Rationale: Until recently, most complex trait dissection was able to resolve 15-20cM regions containing 500-700 genes. While numerous studies have shown that there are more homoeologous QTL regions than would be expected by chance (Paterson et al 1995), it has not been possible to address this at the gene level until now. Intermated linkage populations (Lee et al 2002), positional cloning (Yan et al 2003, 2004; Konishi et al 2006), association mapping (Yu et al 2006a), joint linkage-association mapping (Blott et al 2003), and nested association mapping (Yu et al 2006b) provide high resolution avenues to dissect these complex traits. With the application of high resolution mapping studies to complex traits in grasses, it is now possible to identify either key genes or small genomic regions controlling complex traits. In sequenced genomes, it will be possible to estimate the probability that a gene contributes to a given trait for each gene in the genome. In addition, high resolution approaches allow complex trait association probabilities to be calculated at the individual SNP level across the entire genome.

Approach: To make this vision a reality, we will collaborate with groups working on diversity and complex trait genetics to acquire diversity data for the sequenced grass genomes. We will bring these data sets into the existing Gramene diversity module, subject them to stringent quality control, and provide researchers with a standard set of visualization, query and analysis tools.

To complement the diversity datasets, we will acquire and curate raw genotypic and phenotypic segregation data for selected QTL studies. Gramene already provides information on 10,791 QTL identified for 315 agronomic traits from 9 species of cereal crops. This comprises virtually all the QTL reported over the last 12 years in rice (~8,000 QTL) and nearly all high-yield studies in non-rice cereals. Our first-pass curation provides basic information about the name, map position, and PO associations of each QTL, and links it to the genomic sequence and genetic map. However, it is difficult to relate one study to another because of differences in study design and statistical methodology. Ideally one wishes to combine multiple QTL studies of the same trait in order to increase statistical power and narrow the region of linkage; in practice one cannot do this without access to the raw segregation data.

Therefore we propose to supplement our first-pass QTL data by systematically mapping QTL for all the above key datasets using raw genotypic and phenotypic segregation data obtained directly from selected high-value studies in rice, maize and sorghum. These datasets will be curated, consistency checked, refactored into the standard GDPDM schema, and made available for download, visualization, analysis and meta-analysis.

In consultation with Drs. McCouch and Buckler, Gramene curators will prioritize QTL studies based on the size of the population and the number of traits studied. We will also give higher priority to studies related to plant development, central metabolism, and abiotic stress tolerance. These particularly trait areas have been chosen to match the focus of our pathway curation, which together with QTL and diversity-based associations will accelerate making the connection between a set of genetic linkages and a set of candidate genes.

The curator will contact the corresponding author, seek permission to import the raw dataset into Gramene, and then work with the author to reformat the dataset appropriately. We will remap each QTL systematically using statistical approaches appropriate for the different populations, for example - linkage populations will be analyzed by regression, composite and multiple interval mapping (Zeng 1993, 1994; Kao *et al* 1999), association populations and joint linkage-association studies by Q+K mixed model approach (Yu *et al* 2006a), and nested association by regression based approaches (Yu *et al* 2006b). The remapping information will be stored in the GDPDM schema for retrieval, display and further analysis.

Ultimately geneticists studying the molecular basis for a trait wish to connect a region of QTL association to a specific gene or list of genes. To facilitate this connection, we will make it possible to relate QTLs to their genomic context in annotated genomes, in orthologous regions of the annotated genomes of related plants, to Gene Ontology terms that are overrepresented by genes lying underneath the QTLs, and to pathways that are similarly overrepresented.

Currently Gramene displays the results of curated QTL studies as intervals on genetic maps and annotated genomes, but does not provide support for displaying the detailed QTL distributions across the genome. We will enhance the genome browser and CMap to display the entire QTL distribution across the genome and to show the magnitude of allelic effects as a quantitative graph. We will also make curated QTL studies available for download and viewing using CMTV (Comparative Mapping and Trait Viewer; cmtv.sourceforge.net).

Using comparative mapping data (SA1), we will allow QTL from one species to be projected onto orthologous regions and genes of other species and visualized with CMap. In addition, we will provide tools that allow researchers to identify GO terms and pathways that are significantly overrepresented among the genes underneath a QTL, using hypergeometric distribution statistics (Subramian *et al* 2005). As described in SA2, we will display significant QTL and the magnitude of allelic effects on the SkyPainter pathway viewer.

In contrast to geneticists, plant breeders do not need to know which genes are responsible for a trait of interest. Breeders' goal is to selectively combine positive QTLs in a particular line; therefore they need molecular markers for a region of QTL, and an estimate of the germplasm breeding value. Gramene already provides molecular marker data via its genome annotation (SA1a), and diversity (SA3a) modules. We will supplement this with an estimate of the germplasm breeding value. We will determine which lines contain the optimal combinations of genes either by estimating their additive genetic components in contrast to relatives or by looking at their QTL constitution. For available datasets, we will estimate additive genetic relationship matrices based on molecular markers (Yu *et al* 2006a; Kennedy *et al* 1992; Lynch *et al* 1998), then for available traits we will make BLUP estimates of their breeding values. We will present the data in tables and in pedigree views in the Gramene diversity pages. If diverse germplasm are scored for large numbers of markers at the QTLs, we will incorporate both sets of information into our displays.

We will also expand the TASSEL open source software project (sourceforge.net/projects/tassel) to accommodate linkage and nested association mapping analysis and Gramene data sets.

Diversity and QTL data sets: The following data sets are available as starting points for this effort. Together they comprise more than 1000 individuals and over 100 traits, some mapped at single gene resolution:

- Maize
  - NSF Maize Diversity Project (Doebley, Wisconsin)
  - Intermated IBM Population (Community work with MaizeGDB)
  - Sequencing of diverse maize by next generation approaches (Buckler, Cornell)
- Rice
  - RiceCap (Jim Correll, Univ. Arkansas), QTL mapping data related to disease resistance and milling quality in four populations
  - Rice Diversity Project (Susan McCouch, Cornell Univ), 500 diverse *O. sativa* and *O. rufipogon* accessions genotyped with 10,000 SNPs and phenotyped for 41 traits
  - Perlagen project (Jan Leach, Colorado State Univ) Re-sequenced 100 MB of low-complexity DNA (~25% of the rice genome) on 20 diverse *O. sativa* accessions using Perlagen technology
  - Rice “Ionome” QTLs (David Salt, Purdue University)
  - Rice SNP variants and diversity information from *O. sativa* and *O. rufipogon* from seven targeted large genomic regions (Ana Caceido, University of Massachusetts).
- Arabidopsis
  - 2010 projects on Arabidopsis diversity (Nordborg)

We believe we can acquire the raw data for a significant subset of rice QTL and diversity data sets in ongoing collaborations between plant geneticists, rice breeders and Dr. McCouch. We queried 15 key rice researchers (as outlined in the table below) and within 24 hours, obtained 12 commitments to send raw genotypic and phenotypic QTL datasets to Gramene curators within 3 months. This suggests that a significant proportion of rice researchers who have published QTL papers will agree to deposit their raw datasets in Gramene. In the future, the Diversity Curator will solicit additional datasets for which papers have already been published, as outlined in the attached document (Supplemental file on QTL and diversity datasets for rice).

	Collaborators	Affiliation	Country	Email
1	Dr. Adam H. Price	University of Aberdeen	UK	a.price@abdn.ac.uk
2	Dr. David J. Mackill	IRRI, Philippines	Philippines	d.mackill@cgiar.org
3	Dr. Gu Xingyu	South Dakota State University	USA	GUX@fargo.ars.usda.gov
4	Dr. Henry T. Nguyen	University of Missouri	USA	NguyenHenry@missouri.edu
5	Dr. James Oard	Louisiana State University Agricultural Center	USA	joard@agcenter.lsu.edu
6	Dr. Lihuang Zhu	Chinese Academy of Sciences, Beijing	China	lhzhu@genetics.ac.cn
7	Dr. Qifa Zhang	Huazhong Agricultural University, Wuhan	China	qifazh@mail.hzau.edu.cn

8	Dr. Rebecca Nelson	Cornell University	USA	RJN7@cornell.edu
9	Dr. Susan McCouch	Cornell University	USA	srm4@cornell.edu
10	Dr. Yunbi Xu	CYMMIT	Mexico	y.xu@CGIAR.ORG
11	Dr. Hei Leung	IRRI	Philippines	H.Leung@cgiar.org
12	Dr. Glenn Gregorio	IRRI	Philippines	G.gregorio@cgiar.org
13	Dr. SN Ahn	Chungnam National University	Korea	ahnsn@cnu.ac.kr
14	Dr. Ed Redona	PhilRice	Philippines	edredona@mozcom.com
15	Dr. H. W. Cai	National Institute of Genetics	Japan	hcai@jfsass.or.jp

Personnel Required to Support this Aim: The statistical analysis and visualization portion of the effort will be led by Dr. Ed Buckler. The Buckler group developed the leading association mapping algorithm for diverse populations and pedigrees that will be the heart of the analysis. Two USDA statistical geneticist programmers are associated with the Buckler group to help implement methods (Drs. Peter Bradbury and Zhiwu Zhang). Jean-Luc Jannink, an international leader in plant quantitative genetics, is joining the USDA in Ithaca and will be a great resource. The Buckler group also has ongoing collaborations with Jianming Yu at KSU (USDA-NRI grant) and Carlos Bustamante at Cornell to help with additional statistical concerns.

Dr. Susan McCouch will be the liaison to plant breeders and to groups developing rice diversity data sets. Her relationship to these data producers are deep and extensive.

In addition to Drs. Buckler and McCouch, the following staff are necessary to meet this goal:

- **Scientific programmer (100%)**; A senior scientific programmer will be responsible for developing the QTL integration and cross-study analysis software, as well as the scientific visualization systems described earlier.
- **Diversity & QTL curator (100%)**; This curator (a PhD-level biologist) will be responsible for importing, QC'ing, and releasing QTL and diversity data sets, annotating phenotypic traits with plant ontology terms for cross-referencing and collaborating with the data analyst on the design of visualizations that will be compelling to breeders and geneticists.
- **Data analyst (25%);** A junior scientific programmer will be responsible for using the QTL analysis software developed under this project to reanalyze and remap raw rice diversity and QTL data sets.
- **Postdoctoral Associate (100%);** A postdoctoral Associate (biologist with strong quantitative genetics background) will be responsible for improving the integration of modules (as described in aims 1, 2 & 3) and the querying capability of Gramene. S/he will actively investigate genotype-phenotype relationships in the grasses and provide

constructive feedback to Gramene developers & curators to improve the biological relevance and agronomic utility of the db.

Deliverables (Aim 3):

- Curated diversity and QTL data from more than 1000 individuals and 100 traits among rice, maize and *Arabidopsis*, released in standardized, reanalysis-ready format.
- Estimation of germplasm breeding values for each complete diversity data set.
- Estimation of additive genetic components for the traits for each complete diversity data set.
- Determination of optimal combination of genes for each studied trait in complete diversity data sets.
- Software that facilitate this type of meta-analysis, integrated into visualization and data mining facilities of Gramene.

Specific Aim #4 Support the Plant Ontology (PO) *Two Years Only*

Since 2004 Gramene has been integral participants in the The Plant Ontology Consortium (POC; www.plantontology.org) (DBI-9978564), a standard vocabulary of terms to describe flowering plant anatomic structures and their developmental stages . The PO is a key tool for connecting phenotypic traits in one species with corresponding traits in another. The PO web site currently lists 15734 associations to the anatomy ontology and 9608 associations to the growth and development stage ontology, contributed by Gramene, MaizeGDB, TAIR and NASC. The POC has recently been joined by curators at SolDB (Solanaceae), LIS (Legume), BarleyBase (Barley) and the CGPDB (Compositeae), as well as a number of industrial partners including Pioneer Hi-Bred Inc. and Monsanto Inc. The association database allows a researcher to find germplasms, mutant genes, and QTLs that cause common phenotypic effects among two or more species.

Since the expiration of the original POC grant (August 2006), Gramene has maintained the basic infrastructure for the POC, including the database and the POC web site, as a supplement to the main Gramene grant. These funds also support a full-time ontology editor who handles the essential tasks of coordinating community-proposed changes, additions to the ontology, and community outreach.

For practical reasons arising from NSF’s policies on coPI-ship on multiple Plant Genome Initiative grants, none of the POC co-PIs we able to submit a separate grant application to support the POC in 2007, and Gramene was encouraged by program staff t0 incorporate POC support into its own renewal.

We propose to continue our support for this endeavor by hosting the POC web site, the underlying ontology and the association database. To keep the ontology up to date, we will actively solicit reviews from experts, and manage proposed changes and additions to the ontology. We will accept both large-scale association submissions from organism-specific databases, as well as individual submissions from community researchers.

It is important to understand that we treat the Plant Ontology project as a standalone project. Even though the funding for the project is requested jointly via Gramene, we will ensure that the POC project's activities are not governed by, or compete with, Gramene objectives. We request support for a period of two years (FY08-09), which will provide us time to apply for independent funding for the POC project during FY09 or, should first proposal fail, FY10. Should we be funded for this activity in FY09, we will consult with the NSF as to whether to reallocate or relinquish the unneeded portion.

Outreach: Since the usage of ontology in data search and annotation odf datasets from new species specific interest groups requires a lot of outreach and intreaction component, we propose to organize two workshops yearly. One such workshop is proposed to be organized annually at the Plant and Animal Genome meetings at San Diego and the second workshop one will be organized at venues coinciding with the ASPB's plant Biology meetings or species specific meetings such as maize genetics conference. We also plan to give presentations at two additional meetings a year organized by Botanical Society of America or Crop Science Society of America and one international meeting of relevance.

Personnel Required to Support this Aim: The current POC editor Dr. Chih-wei Tung will be the liaison to the plant community and to groups interested in contributing their species specific plant anatomy and growth stage vocabulary terms.

The following staff is necessary to actively reach out to the plant researchers and develop the vocabularies:

- **Software Developer (50%)**; This scientific programmer will be responsible for maintenance of the POC database, website, ontology development and web browser tools.
- **Ontology Editor (100%)**; The editor (a Ph. D. level scientist Dr. Chih-wei Ting) will work closely with the ontology developers from the previous NSF project (DBI 0321666) and additional developers from new species groups aiming to integrate new terms. Editor will also work with various user groups that are integrating ontologies into their annotations and database tools they develop. The editor's other significant responsibilities will include presenting the PO project and utilities at various national and international symposia, workshops and site visits, interacting with the Gene Ontology's (GO) developer group to learn and update the ontology development SOPs and software improvements.

Dr. Pankaj Jaiswal will supervise this activity.

Deliverables (Aim 4):

- A continually improved community-driven ontology of plant developmental stages and anatomical components.
- Infrastructure for community contribution of associations between ontology terms and mutants, germplasms, and genes.
- The plant ontology web site, providing searchable access to the ontology and associations.

Specific Aim #5: Education, Outreach and Diversity

*Rationale:*As a community database, education, outreach and diversity are at the core of Gramene's mission. Our aim is to involve researchers from the disciplines of pure science (e.g. plant geneticists), translational science (e.g. plant breeders), educators (e.g. plant biology teachers at the high school and college levels), and students at all levels.

*Approach:*Our original approach was heavily dependent on a full-time outreach coordinator to actively push Gramene and its services to the community via presentations, tutorials and workshops as well as a K-12 outreach component via the Dolan DNA Learning Center. While we still consider this to be an effective route, it is also an expensive one both in terms of travel costs and personnel.

For this reason, we propose a different strategy that provides significant cost savings without reducing the effectiveness of our engagement with the community. The three components of this strategy are 1) a community-supported WIKI targeted to pure and translational plant sciences researchers; 2) traditional "push" presentations at meetings, staffed by the coPIs and senior curators; 3) virtual and physical tutorials sponsored by a commercial partner, OpenHelix, targeted at students in under-served institutions and underrepresented minorities.

1) Community WIKI. As described in the earlier sections, we will develop a community WIKI for Gramene consisting of two main set of pages: one set for annotating gene structure and function, and another set for describing biological pathways. These community-supported pages will initially be seeded with information from the current Gramene gene and pathway pages, as well as from affiliated projects like the TIGR Rice Gene Annotation effort, and then will be open to community members to add to, edit and enhance following the tremendously successful social model pioneered by Wikipedia. In addition to these data-rich pages, we will move our current community outreach pages into the WIKI. This includes roughly 130 layperson-oriented pages that describe the economics, biology and history of the world's major monocot crops.

We believe that this WIKI will become a focus for Gramene community outreach and engagement. The bar for participating will be low, and the value of participating will be high as community members begin to make connections among WIKI pages and between WIKI pages and structured database pages on the main Gramene site. To ensure high value to the community, however, WIKIs need to be monitored by editors to prevent spam, to encourage consistency of style, and to step in when necessary to diffuse the occasional flame war. For this reason, the map curator (specific aim #1) and the pathway curator (specific aim #2), will each spend 25% of their effort monitoring and managing the WIKI.

2) Presentations at meetings. We will present Gramene at two national or international meetings per year. One meeting will be chosen for its high visibility to the target audience of researchers likely to use the resource, for example Plant and Animal Genomes. The other will be chosen for its likelihood to reach students, women and underrepresented minorities, for example, a research seminar at a two-year college or traditionally minority-serving university. Each of the coPIs (Stein, Ware, Jaiswal, McCouch & Butler) ,will be responsible for giving two of these presentations over the lifetime of the project.

3) OpenHelix sponsored workshops and tutorials. OpenHelix, LLC (www.openhelix.com) is a for-profit bioinformatics education endeavor which provides virtual and on-site training for educational institutions, academic research laboratories, and companies. OpenHelix has three main types of products. Its virtual products consist of high production-value, self-run web-based tutorials that are available on a subscription basis. Subscriptions can be purchased on an individual basis (typically $89 for one week of access), as as part of a “classroom package” designed for the use of educators in their classrooms. The classroom packages allow up to 30 students to access the tutorial for a limited time for a total typical cost of $249.

Live trainings take the form of regional seminars and on-site training. Regional seminars are hands-on three-hour workshops taught by PhD-level instructors. These are scheduled in advance and have a graded tuition scale ranging from $89 for students to $149 for commercial participants. Finally, OpenHelix provides onsite training to research institutions under a flexible arrangement that allows for half-day, full-day or multi-day training seminars.

OpenHelix also has several sample tutorials that are available free of charge. Readers are encouraged to run these tutorials to get a sense of the OpenHelix product.

Although OpenHelix does not currently track the gender or ethnic background of its students, half of the attendees of live trainings are women, and students are ethnically diverse. According to Mary Mangan, President of OpenHelix, attendees are “a great and cosmopolitan bunch...lots of non-native English speakers, and scientists from many cultures.”

OpenHelix focuses on the most powerful and popular web-based bioinformatics resources, including two other databases that Stein works on: the HapMap database and Reactome. This effort was highly beneficial to both these efforts: we saw usage of both resources increase at the same time that the number of basic support requests decreased. Recently, due to repeated community inquiries, OpenHelix designed and created an online tutorial and accompanying curriculum for Gramene. Under a non-exclusive arrangement, the tutorial materials were developed by OpenHelix with input and review from the Gramene staff. The Gramene tutorial is anticipated to go live in summer 2007 and can be expected to attract between 50 and 400 views per month. Live trainings are based on customers' requests and OpenHelix cannot predict the demand at this time. However, Gramene is their first plant-centric offering, and has already been requested by several customers.

We propose to leverage this ongoing effort by providing tuition support to women and underrepresented minorities for both the virtual and live tutorials. To qualify for tuition relief, educators will fill out a scholarship request form administered by OpenHelix. They will be considered qualified for relief if their institutions meet the requirements for a "minority-serving college or university" (underrepresented minorities represent at least 25% of enrollees). Organizers of live tutorials will be eligible for scholarships if they can certify that at least 25% of the attendees will be from these target groups.

OpenHelix will initially offer Gramene tutorial scholarships for online classroom packages and on-site tutorials. This is the most efficient use of funds, as the proposed annual scholarship budget of $20,000 will be able to reach as many as 2,400 students. It also reduces the temptation to students to fraudulently report their ethnicity or gender. However, if insufficient group packages request scholarships during the first year of the project, then OpenHelix will open up the process to individuals, who will be asked to self-report their gender and ethnicity.

To reach out to educators who can benefit from this program, OpenHelix will direct advertising to minority-serving colleges and universities, as well as to agricultural schools and land grant institutions. Gramene will also publicize this service on its web site and via the presentations described in the previous section.

To measure the impact of this program, OpenHelix will collect anonymous and voluntary demographic information on the individuals who subscribe to the Gramene tutorial and provide it to Gramene quarterly.

Susan McCouch will act as liaison between OpenHelix and Gramene, and will monitor the allocation of scholarship funds.

Personnel Required to Support this Aim:

- **WIKI gene page moderator** (25%). This is the same PhD-level curator who is responsible for maps and markers.
- **WIKI pathway page moderator** (25%). This is the same PhD-level curator who is responsible for pathway curation.
- **DAS and SSWAP Web-Services** (25%). This is the same PhD-level developer who is responsible for managing the bi-annual database builds.

Other Costs: Travel to two meetings per year for outreach presentations, one to a major scientific meeting, and the other to an institution with a high proportion of women and underrepresented minorities.

Deliverables (Aim 5):

- Scholarships for up to 2,400 students to attend Gramene tutorials sponsored by OpenHelix. Scholarship awards will be weighted towards underrepresented minorities, and historically minority-serving educational institutions.
- Community-supported WIKIs for cereal gene function and pathways.
- Two presentations at two domestic or international meetings per year, one of which will be at a historically minority-serving institution.

Literature Cited

Blott, S., J.-J. Kim, S. Moisio, A. Schmidt-Kuntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, *et al.* Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* 163(1): 253-266 (2003).

Paterson, A. H., J. E. Bowers and B. A. Chapman. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A.* 101(26): 9903-8. Epub 2004 May 25. (2004).

Lee, M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt, D. Blair and A. Hallauer. Expanding the genetic map of maize with the intermated b73 x mo17 (ibm) population. *Plant Molecular Biology* 48(5-6): 453-61 (2002).

Yan, L. L., A. Loukoianov, A. Blechl, G. Tranquilli, W. Ramakrishna, P. SanMiguel, J. L. Bennetzen, V. Echenique and J. Dubcovsky. The wheat vrn2 gene is a flowering repressor down-regulated by vernalization. *Science* 303(5664): 1640-1644 (2004).

Konishi, S., T. Izawa, S. Y. Lin, K. Ebana, Y. Fukuta, T. Sasaki and M. Yano. An snp caused loss of seed shattering during rice domestication. *Science* 312(5778): 1392-1396 (2006).

Yu, J., J. Wang, W. Lin, S. Li, H. Li, J. Zhou, P. Ni, W. Dong, S. Hu, C. Zeng, *et al.* The genomes of oryza sativa: A history of duplications. *PLoS Biol* 3(2): e38. Epub 2005 Feb 1. (2005).

Yu, J. M., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38(2): 203-208 (2006).

Ilic K, Kellogg EA, Jaiswal P, Zapata F, Stevens PF, Vincent LP, Avraham S, Reiser L, Pujar A, Sachs MM, Whitman NT, McCouch SR, Schaeffer ML, Ware DH, Stein LD, Rhee SY (2007) The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol. **143:** 587-599. Epub 2006 Dec 2001.

Jaiswal P, Avraham S, Ilic K, Kellogg EA, Pujar A, Reiser L, Seung RY, Sachs MM, Schaeffer M, Stein L, Stevens P, Vincent L, Ware D, Zapata F (2005) Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. Comparative and Functional Genomics **6:** 388-397

Pujar A, Jaiswal P, Kellogg EA, Ilic K, Vincent L, Avraham S, Stevens P, Zapata F, Reiser L, Rhee SY, Sachs MM, Schaeffer M, Stein L, Ware D, McCouch S (2006) Whole Plant Growth Stage Ontology for Angiosperms and its Application in Plant Biology. Plant Physiol **25:** 25