Title: (IPGA) Gramene: Exploring Function Through Comparative Genomics and Network Analysis

RELEVANCE AND JUSTIFICATION

The opportunity to understand plant growth, development, and responses to environment has increased dramatically with the publication of full genome sequences of *Arabidopsis thaliana*[1], rice[2], poplar[3], grape[4], sorghum[5], *Brachypodium distachyon*[6,7], maize[7], soybean[8], strawberry[9], papaya[10], cacao[11], and the conditional access of draft assemblies such as *Arabidopsis lyrata,* sweet orange, clementine, and eucalyptus[12]. Advances in genome technology provided by industry, sequencing centers, and by several plant genome consortiums, make feasible the wholesale discovery of plant genes and their functions. Since no single plant has been studied in sufficient detail to know the function of all its genes, scientists must formulate hypotheses based on homologous relationships to best reference plant models. Although there is much information on well-studied reference models, it is scattered in printed science publications, online databases, and individual lab archives. Therefore, Gramene has developed a comparative genomics platform for identifying orthologous and syntenic mappings between genomes. This has laid the foundation for synthesizing new knowledge applicable not only to the reference plant models but also to non-model plant species using homology-driven approaches. Factors that are external (i.e., environmental) or internal (i.e., genetic and epigenetic) can cause a perturbation of a biological system. For example a gene mutation may cause an alteration of a protein function leading to a systems-level change. Such a change can be captured and deciphered only through a systems- or network-level approach. The basic tools for understanding such changes are described in the current NSF award. We propose to further develop these tools and build new resources to bring together and cross-reference sequenced genomes of important crop plant species with those of model plants. We will develop the capacity to map expression of genes in response to various environmental conditions, such as drought and salinity, and identify gene functions in order to elucidate biochemical and signaling pathways which underlie the ultimate productivity and survival of the plant of interest. The systems/network level approach we propose will not only answer fundamental biological questions such as mechanisms of adaptation and speciation, but will revolutionize the methodological approaches to crop improvement. To achieve these goals, we propose the following four specific aims:

Aim 1. Establish reference data for plant genomes and comparative annotation.

Aim 2. Establish integreted gene network analysis for plants

Aim 3. Integrate new and existing visualization/analysis tools for exploring emerging genomic information for function and phenotype associations.

Aim 4. Transform the community through communication and training opportunities

Requirements for data integration: The last 10 years have seen the development of a wide range of resources for plants including genome sequence and annotations, genetic resources, extensive molecular and non-molecular phenotypic data sets, and informatics resources such as software, databases, and algorithms. While each resource is valuable, the highest value comes from integration. To support the integration in a meaningful way requires that the information must be stored in an organized and searchable way. More detailed analyses requires knowledge of the appropriate algorithms and statistical approaches as well as the existence of shared semantic terms to support meaningful computation. Such a computational data integration framework requires that primary and derived data are annotated with standard metadata terms that describe provenance, experimental design and analysis procedures. Knowledgebases also need to store and present these data while ensuring standards of compliance, comparability and interoperability. In this proposal, we will support the development and implementation of standards to annotate, store, and access genomes and networks in the Ensembl and Reactome infrastructures, standards which have been used in the vertebrate research community and allowing us to build on more than 10 years of development while leveraging mature infrastructure and best practices, and, where commonality exists, allowing us to focus new development of plant-specific needs to support hypothesis-driven research to increase our understanding of fundamental properties in biology and agriculture.

PRIOR RESULTS

Gramene (NSF DBI 0703908), Plant Ontology (NSF DBI 0321666), An *Arabidopsis* Polymorphisim Database (NSF DEB-0723510)

Outreach: In the nine years of the Gramene project, we have served more than 5.5M page requests to more than 1.1M unique visitors located primarily in North America (over 50% of traffic), Asia (30%), and Europe (10%). We have made 32 major releases, responded to hundreds of user requests for help, published or contributed to more than 17 articles in peer-reviewed journals[7,13-28], trained more than 12 postdocs, and presented posters, talks, workshops and computer demonstrations at tens of meetings in the US and abroad.

Website: The Gramene home page was redesigned to allow easy access to our global site search, individual genomes, genetic diversity, metabolic pathways, proteins, genes, ontologies, markers and sequences, comparative maps, quantitative trait loci (QTL), BioMart, and pages detailing the significance of our main species (Fig 1).

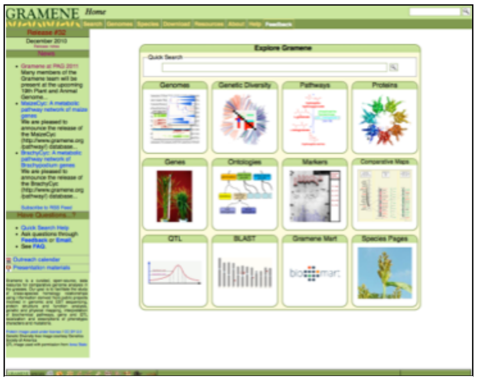


Figure 1: Gramene home Page

Genomes: Gramene hosts 10 complete plant genomes and 8 partial genomes using the Ensembl genome browser to provide users with powerful and flexible browsing of genome annotations such as gene structures, ESTs, cDNAs, proteins, genetic markers, BAC-end sequences, and SNPs. We make available 329 sources of sequence alignments to our hosted genomes using a Distributed Annotation Server (DAS). We have established a formal collaboration with the European Bioinformatics Institute's (EBI) EnsemblGenomes project to share the burden of annotating genomes and building the data mining tools needed by the plant research community (letter Kersey). The 32nd build of Gramene now includes variation data for four species integrated into our Ensembl genome browser and Plant Variation Mart builds from EBI, allowing users to easily view and download SNPs associated with any genomic region of interest. Figure 2 is a screenshot of an Ensembl browser view showing *Vitis vinifera* SNPs in the context of gene annotation. SNPs are color-coded to indicate position relative to gene features (e.g. "intronic") and consequences of the polymorphism on coding sequence (e.g. "non-synonymous") (Fig. 2).

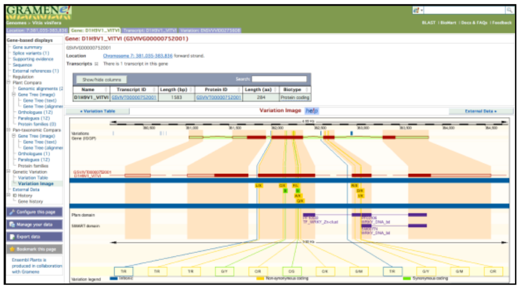


Figure 2: Display of *Vitis viniferra* variation.

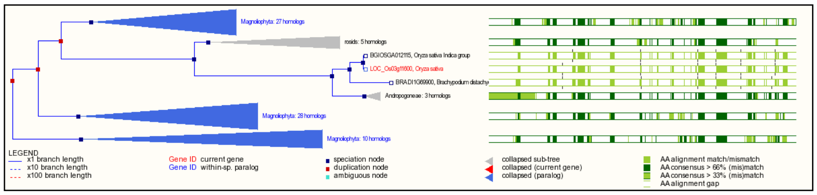

Magnoliophyta: 27 homologs
rosids: 5 homologs
BGIOSGA012115, Oryza sativa Indica group
LOC_Os03g11600, Oryza sativa
BRADI1G69900, Brachypodium distachy
Andropogoneae: 3 homologs
Magnoliophyta: 28 homologs
Magnoliophyta: 10 homologs
LEGEND
x1 branch length
x10 branch length
x100 branch length
Gene ID current gene
Gene ID within-sp. paralog
speciation node
duplication node
ambiguous node
collapsed sub-tree
collapsed (current gene)
collapsed (paralog)
AA alignment match/mismatch
AA consensus > 66% (mis)match
AA consensus > 33% (mis)match
AA alignment gap

Figure 3: Gene tree display.

Gramene's Compara gene tree builds provide phylogenetic reconstructions and ortholog prediction. Improvements in the Ensembl user interface now allows users to easily expand and collapses branches by clicking on the tree image in the

Ensembl interface. The gene tree view in Figure 3 above is of the *O. sativa* gene “DL” showing the expandable sub-trees (grey), paralogs (blue), speciation events (blue nodes) and duplication events (red nodes). In build 32 we created over 35K trees with almost 400K genes. Compara orthology is also employed to construct blocks of synteny among genomes, the results of which are displayed from our Ensembl browser and in CMap (Fig. 4). The new multi-species view shows alignments with gene annotations across multiple species (Fig 8).

Maps: Gramene’s custom-built and widely used comparative map viewer, CMap, now holds a total of 215 maps from 32 species, an increase of over 60 maps in the last four years. It now displays three of our hosted sequenced genomes, providing links to many physical and genetic maps from non-sequenced plants and our collaborators online resources. We also display cross-species syntenic blocks created from our Compara build to show comparisons of our sequenced genomes such as in the example at right where it can be seen that sorghum's chromosome 1 has synteny to rice chromosomes 3 and 10 (Fig. 4). The software is freely available at Sourceforge and used in several web sites including GrainGenes, the LIS, CottonDB, and several projects at the Centre for Comparative Genomics at Murdoch University, Australia.

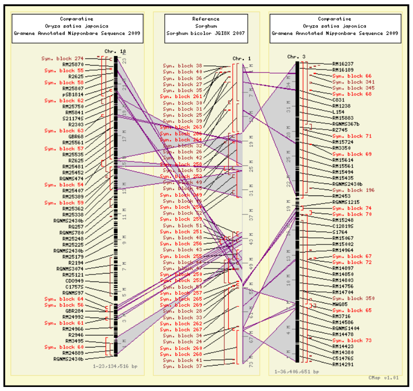


Figure 4: CMap display of rice and sorghum.

Functional Annotation of metabolic and regulatory pathway: In the past 3 years, we have released the constructed metabolic networks for rice (RiceCyc), sorghum (SorghumCyc), *Brachypodium* (BrachyCyc), and maize (MaizeCyc) and provided mirrors for seven more species-specific pathways allowing users to compare two or more datasets. Manual curation was emphasized for rice for which we have released 5 incremental versions. We developed a tool to map microarray probe IDs against the reference gene IDs to overlay expression data on the cellular overview of the metabolic network. Our manually curated genes are also present in a community curation portal called “Plant Gene Wiki.” The goals of Aim 2 to create plant networks is a natural extension of this work.

Support the Plant Ontology: The Plant Ontology, a standard vocabulary of terms to describe flowering plant anatomic structures and their developmental stages[20,29] is a key tool for connecting phenotypic traits in one species with corresponding traits in another. The development of these standards[20,26,29] was done in collaboration with public and private community collaborators, including but not limited to MaizeGDB, TAIR, Pioneer/Dupont, and Monsanto. Structured ontologies will play a key role in the functional annotation described in Aim 2.1 and training in Aim 4.

Diversity: We have curated several large SNP studies from rice, maize and *Arabidopsis* (NSF DEB-0723510) into our genetic diversity database and integrating these into our Ensembl genome views (Fig. 5) to show genome-wide associations. The ability to store and quickly query this high volume of data was made possible by a novel packing scheme that uses binary large object (BLOB) fields in a relational database. Both our web interfaces and Tassel, our desktop Java software package that evaluates genotype and trait associations with the tools of population and quantitative genetics, can now handle millions of data points. Tassel saw a major release (3.0) in the last year, and we provide direct web-start links to launch Tassel with data from rice, maize, and *Arabidopsis*. We developed the web-based “SNP Query” tool to search for genetic variation in genomic coordinates from one or more accessions. In addition, we also leverage the Flapjack tool from SCRI to allow our users to graphically browse SNPs. All variation data is made available in multiple standardized formats (e.g., PLINK, HapMap, MySQL exports) so that our users can more easily carry out their own analysis.

Genome entry points for phenotypic diversity: As there is no one standard in the plant community for storing genome-wide association studies (GWAS) information, we have worked with collaborators at EBI, Cornell, GMI and SCRI on a flexible architecture to store and efficiently serve fine-grained experimental details along with the genotypic and phenotypic information. An initial implementation for displaying a GWAS was to store the Atwell[30] *Arabidopsis* data in the Ensembl variation schema and to write custom DAS adaptors for display in our genome browser (Fig. 5). To meaningfully categorize and summarize the data, particularly complex phenotypic trait descriptions, it was necessary to map textual descriptions onto standard ontologies for biological information (e.g., Trait, Environment, Phenotypic Quality and Plant Ontologies), developing new terms as needed. This allowed us to capture quantitative and categorical experimental values as well as geographic information for downstream analysis. We explored various methods for dynamic interactive display of GWAS data and prototyped a promising new tool using existing open-source, JavaScript frameworks (Fig. 6). Moving forward, we are working with the major data providers in maize, wheat, rice, and *Arabidopsis*, who have established a consortium that will focus on developing data handling and storage standards for emerging GWAS data (letter Buckler, Nordborg, McCouch). Aim 1.2's goal of storing, analyzing and visualizing diversity data will be supported by prior work in these last two sections.

Maize Genome Sequencing (DBI 0910642) **& Maize Diversity (**DBI 0638566, 0321467)**:** In the last 4 years, we have participated in the delivery of the maize genome sequence, annotations and diversity including five major updates to the web portal and three evidence-based gene build releases, all of which are available through the maize project website (www.maizesequence.org). The resources from these projects have contributed to more than 12 publications [7,12,31-42] which have made significant contributions to our knowledge of plant genome structure, evolution, and the genetic architecture of maize, which is the most complex and diverse plant genome sequenced to date.

Of significance to this grant has been the development of pipelines to manage and visualize genome sequence, annotations, and variation as well as the insight gained from developing graph-based approaches needed to model multiple versions of a genome. Recently our group has developed expertise in *de novo* assemblies, and we are working on analyses that will allow *de novo* contigs to be anchored to the reference genome by genetic maps in order to infer their mutual order. We have developed several approaches for deriving potential gene scaffolds using transcript-guided assemblies and to resolve complex branching scenarios. We are presently working with Mike McMullen (ARS, Univ. of Missouri) and Ed Buckler (ARS, Cornell) to place novel contigs, gene scaffolds, and novel full-length cDNAs (FLcDNA) on the integrated genetic/physical map of maize. By applying genotyping-by-sequencing (GBS) technology to genotype an intermated mapping population, McMullen and Buckler have improved the maize genetic map to unprecedented resolution. All of these resources will be extended in Aim 1 of the current proposal.

The maize diversity project addressed the need of a high-density variation map for driving high-resolution QTL mapping in maize. In collaboration with Buckler, we designed an informatics workflow for aligning high-throughput sequencing reads and deriving high-confidence polymorphisms that segregate in the *Zea* genus. A total of over 50M single-nucleotide polymorphisms (SNPs) and small insertion-deletions were scored in a panel of 103 inbred varieties (Chia et al in preparation). Furthermore, from sliding windows of contrasting read-mapping densities, we ascertained segregating copy number variations (CNVs). From these datasets, we identified 2,000 genomic loci that were potentially selected for during the domestication and improvement of maize (Ross-Ibarra et al in preparation). The combined set of SNP and copy-number variations have been used as markers for GWAS against five key agronomic traits. Significant QTLs are enriched for CNVs and genic SNPs and include markers that lie in the vicinity of previously published markers. The association results demonstrate the validity and utility of the variation map, and also underscore the role that structural variations such as CNVs play in affecting phenotype.

Results from this study will also provide primary data sets to support the pan genome representaion, maize stewardship, and functional association objectives in Aim 1 and 2.

EXPERIMENTAL PLAN

"What are the functionally shared elements of plant genomes, how do these elements work together, and how does the diversity of these elements relate to agronomically valuable traits"

The three research aims are designed to answer this question (1) Using comparative genomics to identify functional elements and sequence variants that may have phenotypic consequences; (2) Annotating gene networks in order discover metabolic pathways and biological mechanics that lead to phenotypes; (3) Collecting, integrating and presenting, genome, network and phenotypic data as an integrated framework in order to connect genetic diversity to phenotypic variation. The three aims together provide the foundation for synthesis essential for the biological interpretation of population-based data that is necessary to identify the relationship between genotypic and phenotypic variations. The outreach component supports the training of current and future scientists in data management, specifically in the use of the standards and in the need to integrate diverse data sets for discovery of basic fundamental principals in biology.

Aim 1. Establish a reference resource for plant genomes and comparative annotation.

In collaboration with EBI and other community members, we propose the development and stewardship of up to 20 plant genome reference sequences and annotations in the Ensembl[43] framework. This will provide a common standard platform for comparative genomic analysis and visualization. The enriched genome annotations will include controlled vocabularies to describe metadata and primary data associated with comparative phylogenomics, epigenetics, and population-based phenotypes. The genome assembly and data structures will allow researchers to shift their point of reference from a single species to a taxonomic clade representation of a genome.

1.1 *Establish an integrated reference genome resource.* We will host a minimum of 20 complete reference genomes over the next 4 years and will apply standard and species-appropriate annotations and analyses that will be updated through semiannual releases. Genomes will also be hosted at PlantEnsembl through our collaboration with EBI. The candidate list of additional reference species includes wheat, barley, soybean, potato, tomato, strawberry[9], the basal angiosperm *Amborella* (NSF #0922742) and the basal vascular plant *Selaginella*. Factors influencing selection will include sequence availability within GenBank, completeness of assembly/annotation, value as a crop and/or model for functional genomics, evolutionary significance (taxon portfolio), collaborations, and the recommendations and priorities suggested by our scientific advisory board (SAB) and PIs.

Base-line Annotations: In addition to importing annotations from respective community projects (Letters: Itoh, Lawrence, Huala), we will apply our own standardized annotation pipelines. At the DNA level, these include repeat-finding using both curated libraries[7,18,44-46], *de novo* methods[47,48], and feature alignments[49] consisting of a wide variety of regularly-updated public sequences, such as expressed sequence tags (EST). At the protein level, Gramene will continue to employ InterproScan[50] and other modeling software to predict protein functional domains, secondary structures, and subcellular localizations[51-62]. These data, as well direct community annotations, Uniprot best matches, and those suggested by GOstruct (Letter:Ben-Hur) will be used as input to our in-house protocols for assigning Gene Onotology (GO)[63,64], Plant Ontology (PO)[29] and Trait Ontology (TO)[16]. We will continue to use the Ensembl XRef pipeline to cross-reference gene names and identifiers from external databases [54,65-74]

The Gramene Gene Build and ncRNA pipelines: We will apply our in-house evidence-based gene-prediction method which, along with *ab-initio* methods [75], has demonstrated high sensitivity and specificity when applied to rice, *Arabidopsis*, and maize[7,27]. Future development will incorporate available RNA-Seq data [11,76-79] to provide more sensitive detection of expressed genes and will augment existing models with alternatively spliced or transcribed forms[9,18,76,80-85]. For non-coding RNA genes (ncRNA), we will employ our miRNA detection method that was developed for maize and sorghum[5,33] and will apply RFAM to annotate a variety of other ncRNA gene classes [86,87].

Deliverables: Release of new software quarterly and data and analysis in Ensembl semiannually with the addition of 2-3 genomes a year (Appendix 1-2).

1.2 Visualization of Epigenomic, Genotypic, and Functionally Phenotypic Diversity.

Epigenomic Variation: Annotation of gene regulatory elements has been historically absent from genome databases because of the lack of experimental evidence or reliable prediction methods. This predicament is rapidly changing with technologies such as ChIP-Chip, ChIP-Seq, and bisulfite sequencing, which are capable of defining genome-wide transcription factor binding sites and epigenetic markings from histone modifications and DNA methylation[84,88-91]. Such information is now contributing to predictive models of chromatin state, gene expression, and regulatory networks[80,92-94]. Gramene will integrate available data from publically funded studies in *Arabidopsis*, rice, and maize in order to build a catalogue of regulatory sites for each genome (Letter:Chen, Wagner). This research will build upon the Ensembl infrastructure as used in the Encode project[43].

Genome Variation: Gramene has consolidated sequence polymorphism data in one central portal for *Arabidopsis*, rice, maize and grape as part of the Ensembl Variation schema. These are presented to scientists as a consistent catalogue of SNPs and are classified on the basis of location within gene features and impact on coding sequence. In this proposal, we will continue to import genomic variation datasets into our Ensembl-based infrastructure and promote submission to NCBI's dbSNP resource (Letters: Brady, Patterson). To support the functional annotation of core model organism (Aim 2), we will target well-developed population-based genotype/phenotype studies in *Arabidopsis*, rice and maize via collaborations with funded projects (Letters: Buckler, McCouch, Nordborg). This is a change from our previous strategy, which directly supported the curation activities of population data. In this instance of Gramene, the curation activity will be supported as part of the species-specific collaborations. We will continue to extend our preliminary work to catalogue and display phenotypes in genomic context (Fig. 5).

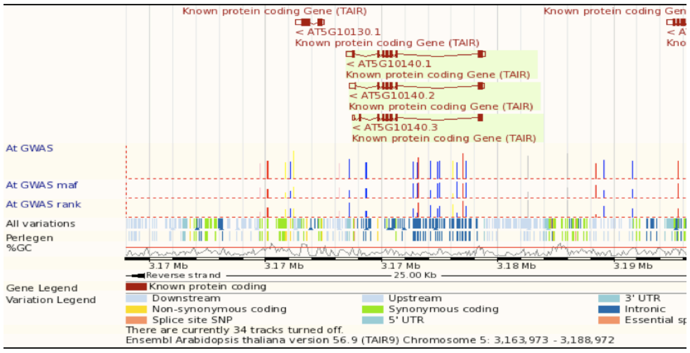
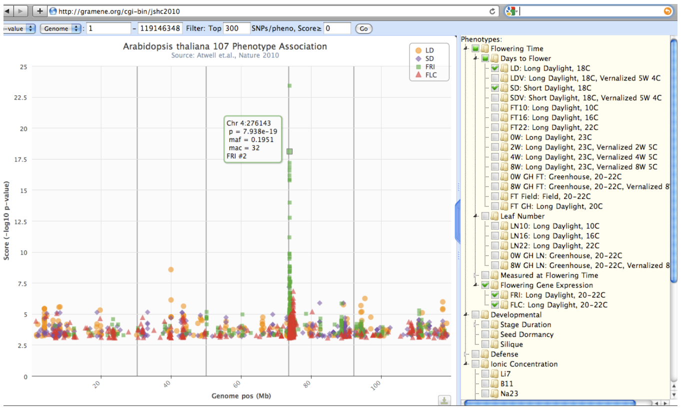


Figure 5. DAS GWAS tracks showing association p-value, allele frequency and phenotype rank on Gramene.

Phenotypic variation: High-throughput sequencing and the increased utility of GWAS has led to an explosion in available pheno-genotypic data (e.g. the Atwell dataset has 1,179 ecotypes, each having 250k variation loci associated with 107 phenotypes). It is necessary to build metadata-driven views that can categorize and summarize this information. Population- and phenotype-based metadata will serve as a high-level aggregator based on high-priority traits such as flowering time, cold-stress response, and drought-stress response.

We aim to link these traits across plant genomes and derive gene function from other genomes through the union of published gene expression (through EBI Atlas), biological networks, and population structure. To this aim, we will need to develop a semantic vocabulary to tag analyses, re-annotate existing structures based on derived function, and to assign levels of confidence to these annotations based on how closely related are the species or other evidence of similar gene function via orthologous genes.


http://gramene.org/cgi-bin/jshc2010

-value Genome : 1 - 119146348 Filter: Top 300 SNPs/pheno, Score≥ 0 Go

Phenotypes:
- ☑ Flowering Time
  - ☑ Days to Flower
    - ☑ LD: Long Daylight, 18C
    - □ LDV: Long Daylight, 18C, Vernalized 5W 4C
    - ☑ SD: Short Daylight, 18C
    - □ SDV: Short Daylight, 18C, Vernalized 5W 4C
    - □ FT10: Long Daylight, 10C
    - □ FT16: Long Daylight, 16C
    - □ FT22: Long Daylight, 22C
    - □ 0W: Long Daylight, 23C
    - □ 2W: Long Daylight, 23C, Vernalized 2W 5C
    - □ 4W: Long Daylight, 23C, Vernalized 4W 5C
    - □ 8W: Long Daylight, 23C, Vernalized 8W 5C
    - □ 0W GH FT: Greenhouse, 20-22C
    - □ 8W GH FT: Greenhouse, 20-22C, Vernalized 8
    - □ FT Field: Field, 20-22C
    - □ FT GH: Long Daylight, 20C
  - □ Leaf Number
    - □ LN10: Long Daylight, 10C
    - □ LN16: Long Daylight, 16C
    - □ LN22: Long Daylight, 22C
    - □ 0W GH LN: Greenhouse, 20-22C
    - □ 8W GH LN: Greenhouse, 20-22C, Vernalized 8
  - □ Measured at Flowering Time
  - ☑ Flowering Gene Expression
    - ☑ FRI: Long Daylight, 20-22C
    - ☑ FLC: Long Daylight, 20-22C
- □ Developmental
  - □ Stage Duration
  - □ Seed Dormancy
  - □ Silique
- □ Defense
- □ Ionic Concentration
  - □ Li7
  - □ B11
  - □ Na23

Figure 6: Interactive visualization of four flowering-time related phenotype associations on Chromosome 4.

Going forward, we propose to incorporate algorithms and software modules for downstream analysis (collaboration Yu), using knowledge of pathways, genes, QTL localization, and mutations, all of which are accessible within Gramene. We explored various methods for dynamic interactive display of GWAS data and prototyped a promising new tool using existing open-source JavaScript framework on the Atwell *Arabidopsis* dataset[30] (Figure 6). Selected data points from views can be exported or automatically fed to trait analysis pipelines such as Tassel and Flapjack. Notable results and detailed user-annotations will be fed back into Gramene's data store for future reference in a publically shared setting.

Deliverables: We will update software on a quarterly basis. Standard descriptors for epigenetic will be released twice per year starting in year 2 (Appendix 1-2).

1.3 *Representation of the pan-genome*. Re-sequencing efforts in maize, rice, *Arabidopsis*, and other plants are now revealing the large degree of intra-species structural variation among different germplasms[78,95,96]. Following the trend in microbial genomics, the concept of a single reference genome is giving way to that of the "pan-genome" in both animals and plants[97,98] in order to describe the full-complement of genes and variants in a species by capturing both the conserved, "core" genome as well as the "dispensable" genome that is specific to populations or single individuals. We will explore computational strategies to handle the diversity among the sequences while providing the abstraction of intact common genome sequences. Specifically, graph-based data structures provide robust semantics for storing a comprehensive set of unique sequences (SNP's, structural and copy number variation) and representing whole genomes as paths through the graph. Such a representation will not only reduce storage requirements and improve performance, but will also allow for new kinds of comparative modeling studies. In collaboration with EnsemblPlants (Collaboration Kersey), we will first produce a *de novo* assembly of each of at least 5 genotype accessions of each species using a common algorithm, where one accession will represent the current reference (i.e. *Arabidopsis* Col-0, rice *Nipponbare*, and maize B73). For each accession, we will evaluate the completeness of the assembly as well as the shared and novel gene space to determine the potential for sequence integration in the pan genome graph. We will make use of iPlant's compute resources, including discovery environments (DE) for assembly (Letter: Stanzione). We will provide recommendations for representing and accessing the core and dispensable genome sets within a single pan-genome graph structure. The analysis will also form a framework for objective assessment of quality and completeness of assemblies generated by different algorithms.

Deliverables: Recommendation for graph-based approaches to store and display "core" and "dispensable" genomes of a species (Appendix 1-2).

1.4 *Comparative genomics: analysis of plant genomes and visualization informed by evolutionary histories*

Objective: Perform and display comparative genomic analyses derived from phylogenetic gene-trees and whole genome alignments. Identify gene-level syntenic relationships and perform secondary analyses to *annotate gene duplications, movements, losses and other changes from ancestral state. Support mining of gene families based on taxonomic origin. Apply standardized vocabulary or ontologies to this information to allow it to be referenced within various interfaces described in Aims 2 & 3.*

Phylogenomics promotes the study of gene structure and function by describing origins, histories and relationships among genes and their function. Gramene will implement the EnsemblCompara GeneTree pipeline to infer large-scale phylogenetic reconstructions[99]. The method yields accurate prediction of ortholog and paralog relationships. It also classifies nodes with respect to taxon of origin and thus indicates the timing of duplication events and gene family expansions. This resource was used extensively in characterizing the evolutionary dynamics of the maize genome[7,35], and the methods we developed are now being adapted by Gramene for application across its genomes[25]. Our gene-level synteny build uses ortholog designations to define ancestral gene positions and reveal gene movements, duplications, and losses[7] (Fig. 7). We have also developed methods to identify tandemly duplicated genes and whole genome duplications with gene-level classification of homeologs [7,35].

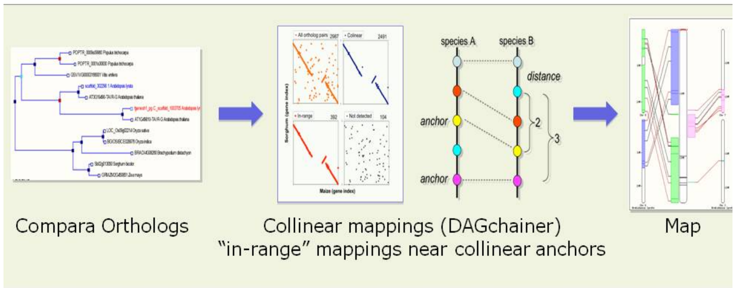

species A
species B
distance
anchor
>2
3
anchor
Compara Orthologs
Collinear mappings (DAGchainer)
"in-range" mappings near collinear anchors
Map

Figure 7: Gene-centered synteny pipeline using phylogenetically determined orthologs

Complementing phylogenetic analysis, Gramene displays pairwise whole genome alignments [49]. Multiple sequence alignments are performed using the Enredo-Pecan-Ortheus (EPO) method, which shows the history of genome rearrangements via ancestral genome reconstruction[100]. These alignments provide the basis for detecting evolutionarily constrained elements within coding and non-coding regions using GERP[101,102].

Phylogenomic descriptors: Output from phylogenomic and comparative analysis will be incorporated into a structured vocabulary that describes gene origins and inter-relationships. Classifications will include ortholog, paralog, syntelog, homoeolog, moved gene, and tandem duplications that relate one gene to another. The goal is to enable integration of these properties within other Gramene interfaces, such as the browser, phylogenetic trees, gene-centered pages, Reactome, and Atlas. An example application would be the study of regulatory subfunctionalization and neofunctionalization by integration with expression data.

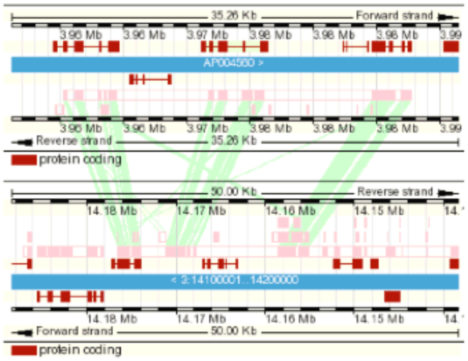


Figure 8: Two rice genes (top track) whose putative homologs were possibly missed in Brachypodium (bottom).

Comparative genomics offers great potential for cross-annotation of genomes over ancestrally derived regions but has generally been underutilized. An immediate application is the detection of overlooked genes where an aligned region is annotated in one genome but not in another. Gramene additionally employs a method to detect putative split gene models based on inconsistencies between gene tree and protein multisequence alignment. Over 4,700 putative split models were detected in Gramene Release 31, ranging from 101 in *Brachypodium* to 1,181 in poplar. We will develop pipelines to flag potential mis-annotations and provide these to community annotation projects on a collaborative basis

Deliverables: 1) Phylogenetic gene trees 2) whole genome alignments and 3) synteny analysis within the dicot and monocot clades. 4) Catalogue of mis-annotations based on comparative genome analysis 5) Standard phylogenetic descriptors and annotations of gene origin. (Appendix 1-2)

1.5 Stewardship of the maize B73 genome and assembly for one update in years 1 & 3. For this proposal, we are specifically requesting funds for an additional round of improvements to the B73 maize reference assembly[7]. We will incorporate community improvements for the genome models and assembly based on sequenced improvements collected by MaizeGDB (Letter: Lawrence). We will provide a new assembly which will include updates to contig order by using the NAM genetic maps[38,103] (Letter: Buckler) and evidence based gene builds[27]. This will be an extension of the on-going collaboration between the Maize Genome Sequencing project and MaizeGDB (see preliminary results).

Deliverables: 1 release of an evidence-based gene build for B73 and an improved B73 reference assembly supported by NAM genetic map (Appendix 1-2).

Aim 2: Establishment of integreted gene network analysis for plants

We will contribute to the development, implementation, and stewardship of gene networks for *Arabidopsis*, maize and rice using the Reactome model. We will contribute to the integration of expression profiling data sets, including the pending surge of RNAseq data for *Arabidopsis*, maize and rice using the ATLAS (EBI) infrastructure.

2.1 Functional annotation of core model organisms for reconstruction of metabolic and signaling pathway networks. This functional annotation of *O. sativa*, *A. thaliana*, and *Zea mays* genomes will be enriched and informed by addition of metabolic and regulatory network pathways, associations to mutant and QTL phenotypes, epigenetic markers, and gene expression profiles. Thus providing the foundation for comparative functional analysis in model plant genomes. These annotations will also serve as the gold standard (Fig. 9A) for whole-genome functional annotation of non-reference model plants well beyond the conventional use of only protein domain-(Interpro) based functional assignments.

Reconstruction of metabolic and signaling pathway networks: As described earlier, we used the Pathway Tools system[104] to predict and curate biological pathways in rice, sorghum, *Brachypodium*, and maize. Pathway Tools relies on a manual input and an enrichment process involving automated reconstruction of an intermediary metabolic network based on the reference database called MetaCyc. However, expanding existing pathway models for creation of regulatory pathways involving co-expressed and co-evolving network of genes poses limitations on scalability issues. Specifically, for a gene network modeled as a directed graph where labeled nodes represent molecules and edges represent regulatory interactions, the limitation lies in the number of such nodes and connecting edges. While the Cyc tool works well at the smaller level of bacterial and fungal genomes, we are estimating there will be millions of interactions for plant genomes. Therefore, we believe that the NIH-supported Reactome tool[105,106], developed as an open source project for the curation and representation of the human genome interactome, is the preferred choice. Additionally, Reactome developers are creating tools that allow data to be exchanged among various pathway development platforms such as those in use by Pathway Tools, including SBML, BioPax, and OWL formats. We also propose to extend our current metabolic pathway collection by reconstructing gene networks for developmental, genetic, and signaling pathways in rice, *Arabidopsis*, and maize.

We will import the latest versions of RiceCyc, MaizeCyc, and AraCyc (Letters: Huala,Lawrence) into the Reactome database framework (Letter: Peter D'Eustachio). We will use existing Reactome tools to curate key pathways for the agronomically important traits of flowering time; inflorescence development; root development; regulation of metabolism; transport and biosynthesis of plant growth hormones; carbohydrates; secondary metabolites; B vitamins (NSF #1025398); regulation of response to abiotic

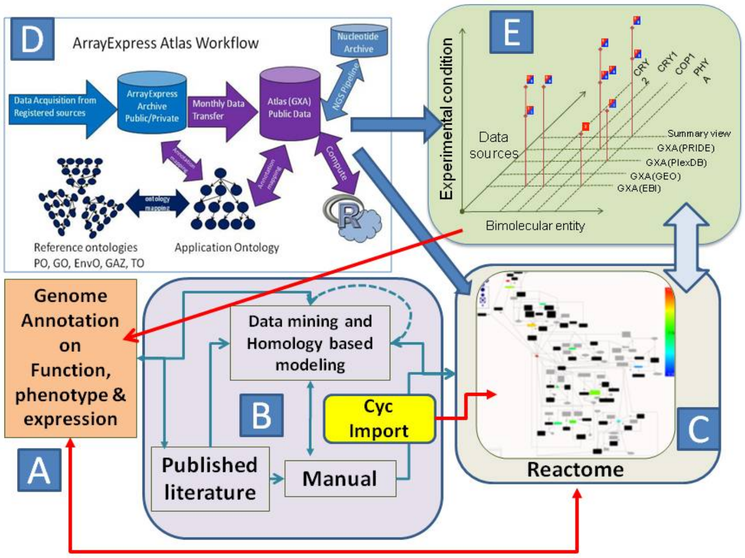


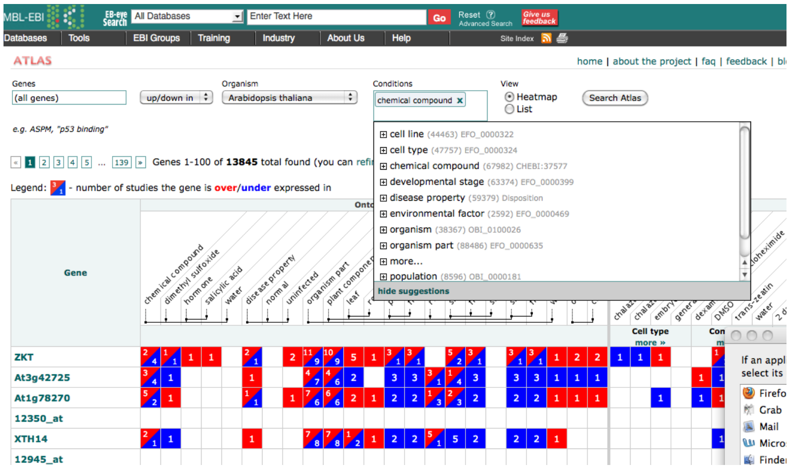
Figure 9: A model representation of the Aim 2 workflow.

stresses, drought, photoperiod and salt; and important biotic disease stresses. We will emphasize high-priority datasets currently being generated such as on C3-C4 photosynthesis in rice, maize, *Brachypodium* and *Setaria* (foxtail millet) (Letter: Nelson) or those that will be funded in the next couple of years. This effort will provide the data archive, annotation and integration capabilities required by the collaborative projects and our end users in the generation and validation of hypothesis-driven models. We plan to achieve the above by adopting existing reconstructions and data mining approaches. (1) Import the latest MetaCyc-based builds of rice, maize and *Arabidopsis* metabolic pathways (Letter: Lawrence, (Fig. 9B). (2) Enrich the metabolic pathway dataset with pathway inferences drawn from Reactome[107], primarily focusing on regulatory processes, such as cell growth, cell cycle, DNA replication/repair and circadian rhythms, which is in addition to the above mentioned priority sets. (3) Validate the pathways empirically through an integrated approach based on annotated EBI-ATLAS gene expression data described later (Fig. 9D) and phylogenomic analysis in Aim 1. (4) Introduce experimentally validated and/or predicted molecular interaction datasets, including protein-protein, Chip-Seq, biochemical affinity, epigenomics, and genetic interaction data (Letter: Wagner, Huala, Lawrence, Provart). The majority of these would be imported from curated databases like BAR, IntAct, BIND, DIP, MINT, etc., and/or would be requested from authors of the publications in a standard format (Letter: Crispin/ASPB) as described in Aim 3. (5) Retrieve functional gene annotations and molecular annotations from legacy publications using data mining tools such as Textpresso and BioCreative[108-110] methods that make use of Natural Language Processing (Fig. 9B). Though these methods are constantly evolving, we plan to create a reference library that includes known and predicted gene names, symbols, functions, phenotypes, and pathway annotations in the three target species. This library, reviewed by manual intervention, will constitute the training – and eventually the experimental – dataset for genome-wide analysis on legacy and future large-scale annotations of published data sets. We anticipate bottlenecks to access full-text data from publishers, therefore we will implement a test set on select publications from the two ASPB journals, *Plant Physiology* and *Plant Cell*. We anticipate that the data mining methods that we develop in collaboration with ASPB will be applicable to any journal. In our proof of concept work with the rice Reactome, we were successful in importing all 374 RiceCyc metabolic networks. In addition we were able to derive ~130 enriched set of regulatory/signaling pathways including DNA replication, DNA repair and cell cycle based on human gene orthology based projections in the Reactome central database maintained by Lincoln Stein's Lab at OICR (Letter: Stein) by way of the BioPax exchange format. We are currently refining the datasets and are developing standardized methods for curating such networks. Upon full Reactome availability, the current MetaCyc based views of rice, sorghum, and *Brachypodium* served principally from Gramene will be retired and mirrored from iPlant resources (Letter: Stanzione).

In addition to the model species Reactome databases, we will create a pilot reference catalog of plant Reactome network. We have found that often experimental support for pathways in the three model species may come from non-model plants[111-117]. Therefore, we will initiate a pilot project to curate a reference plant Reactome including experimentally-determined pathways from both non-model and model plant species. We anticipate that, while creating a plant kingdom wide reference networks for a given biological process, the evidence may not necessarily come from the three model plants of choice and we may also encounter reports suggesting variant networks of genes/ interactors compared to the model plants. In order to capture these subtle variations in gene networks we will create super-pathways and interaction networks that includes all the variant components (reactions and interactors) identified in the plant kingdom with empirical support, thereby creating intelligent partial projections into a kind of "plant superset" Reactome. Once developed, this dataset will provide a strong basis for projections into new species based on their taxonomic proximity and the evolution of orthologous genes (Aim 1).

Deliverables: Reactomes for rice, maize and Arabidopsis with priority on flowering time, inflorescence development, root development, regulation of plant growth hormone metabolism, transport and synthesis, carbohydrate metabolism, vitamin-B biosynthesis (NSF 1025398), regulation of secondary plant metabolism in response to abiotic and biotic stress, C3-C4 carbon assimilation, photosynthesis, photorespiration(Appendix 1-2).

2.2 *Meta data analysis and annotation of the gene expression datasets.* An important aspect of any functional annotation and network analysis with reference to an experimental condition or an organism's phenotype is determined by the involvement of genes that regulate the response and the intricate molecular network of biological processes. External (i.e. environmental) or internal (i..e genetic and epigenetic) factors can cause a perturbation of a biological system. For example, a gene mutation may cause an alteration of a protein function leading to systems-level changes in that plant leading to a phenotype. In order to improve the annotation of gene function, we will use a hypothesis-driven model to study the impact of genetic variation (Aim 1) on the molecular network (Aim 2.1) as well as the gene expression profile(s) to conduct a research project for inferring the global gene expression map of rice, maize and *Arabidopsis,* analogous to the human map[118]. The gene expression profiles will be annotated for their expression and metadata across all validated experimental datasets. As a usual practice many microarray datasets are submitted prior to publication in the Gene Expression Omnibus (GEO). It is a database repository of expression data from high-throughput array and RNA-seq experiments. However, the way this repository has been setup the list of plant GEO expression data sets are organized as separate experiments and, as such, is not extractable to query all plant anatomical parts, development stages, and possible experimental treatments. However, if these data sets are subjected to meta-analyis, they could be queried to extract expression characteristics. For example looking at the whole genome transcriptome level of inter and intra-specific comparison under a common set of environmental conditions or the expression profile of a common plant part such as root specific expression under drought. In order to provide plant biologists answer some of these questions, we will collaborate with EBI's Array Express and ATLAS projects (Letter: Parkinson). The datasets will be imported from the GEO for the three species,



Genes: (all genes) up/down in Organism: Arabidopsis thaliana Conditions: chemical compound ✕ View: ● Heatmap ○ List

e.g. ASPM, "p53 binding"

« 1 2 3 4 5 ... 139 » Genes 1-100 of **13845** total found (you can refi

Legend: 3/1 - number of studies the gene is over/under expressed in

⊞ cell line (44463) EFO_0000322
⊞ cell type (47757) EFO_0000324
⊞ chemical compound (67982) CHEBI:37577
⊞ developmental stage (63374) EFO_0000399
⊞ disease property (59379) Disposition
⊞ environmental factor (2592) EFO_0000469
⊞ organism (38367) OBI_0100026
⊞ organism part (88486) EFO_0000635
⊞ more...
⊞ population (8596) OBI_0000181

ZKT
At3g42725
At1g78270
12350_at
XTH14
12945_at

Figure 10: Gene expression summary prototypes for A. thaliana provided by EBI ATLAS. Views allow single gene to a set of genes across multiple experiments and experimental conditions.

which currently stands at ~2,700 experiments. Then we will annotate the metadata using reference ontologies Followed by a validation process that evaluates data quality and experimental design the successful experiments coming out of the validation step will undergo metadata analysis which we will improve to work on the plant datasets. According to Parkinson (personal communication), about 20-25% of the mouse and human experiments listed in the ArrayExpress failed the validation step, and we expect the same for plants. The current ATLAS workflow accepts microarray and RNA-seq, and the proteomics datasets from the PRIDE, a proteomics data archive. After successful analysis and integration in the ATLAS project, datasets will be embed in the expression data views as gene features on respective gene(s) (see a prototype from Arabidopsis in Fig. 10). For advanced users interested in working with large scale data, we will provide the analyzed results and raw experiment datasets from EBI and iPlant's Cloud archives. However, before we embark on analyzing the targeted datasets from GEO in bulk, we will prioritize a smaller set of experiments to show the reference expression atlas of a model species and attempt to answer compelling biological questions in areas such as flowering time, flower and seed development and abiotic and biotic stress responses.

Deliverables: Expression data sets; Annotate the complete GEO datasets (about 3,000 experiments or a priority set as defined above) for rice, maize and Arabidopsis using the meta-analysis and annotation pipelines developed by the EBI-ATLAS project and those modified under this collaboration to suit the annotation required for plant-based datasets (collaboration Huala, Lawrence, Chen) (Appendix 1-2).

Aim 3. Integrate new visualization/analysis tools for exploring emerging genomic information for function and phenotype associations.

We will support the development and implementation of new interfaces through the synthesis of existing tools including Ensembl, Reactome, ATLAS, and Biomart. These tools support hypothesis-driven interrogation of genome- and phenome-associated data sets to allow comparative *functional* genomics. We will provide multiple methods for programmatic access to Gramene data resources.

One of Gramene's strengths has been to combine multiple data types and tools, developed both internally and externally, under one platform[25]. As described under Aim 2, we will incorporate new tools and visualization software (Reactome, ATLAS, and our internally developed GWAS viewer), each with its own special capabilities and visual interfaces for exploring function. It is not enough to merely combine these in one web site, but rather to provide an integrated analysis system where data of different experimental origins are merged to establish biological relationships such as between expression and chromatin state. Comparative genomics has limited value if one cannot easily determine how two evolutionarily defined orthologs compare in terms of expression profile, phenotype association, or membership within a network or genomic position. Integrating such information means the ability to project it across many data types. Users looking at a phylogenetic trees should be able to see InterPro and GO annotations simultaneously and to access expression and pathway information (and vice versa) readily without searching for individual gene identifiers. As described in Aims 1 and 2, we intend to develop and adhere to standardized ontologies and metadata classifications to allow systematic search, exchange, and representation of data within different software modules. In addition, complex search and mining capabilities across these dimensions must be efficient and flexible enough to meet any different users' needs, perspectives, and approaches. Finally, information must be programmatically accessible to users and external automated tools. In Aim 3 we describe the software, data architecture, and tools to make this possible.

3.1 *Build a high-capacity data warehouse that promotes flexible implementation of software modules*. This goal addresses two fundamental problems. First, the process of interconnecting multiple software components and database schemas, contributed by each tool and added over time, is labor-intensive, inefficient, and does not scale well. Second, the sheer quantity of data and its diversity has

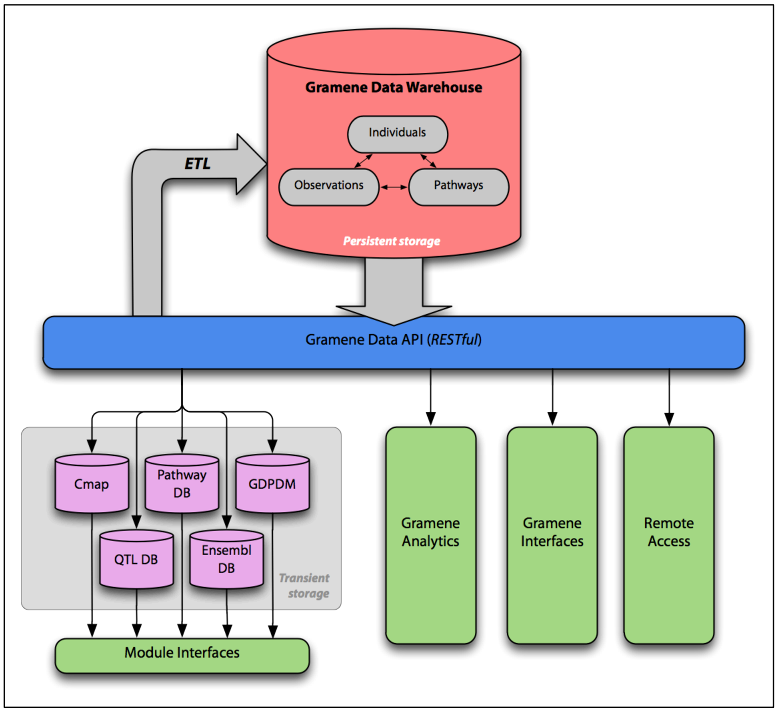


Figure 11: The proposed Gramene Data Warehouse and its relationship to other components and services.

outgrown typical relational database architectures. A robust solution is to build a generalized framework into which new tools and interfaces can be plugged in as required. To accomplish this, we will build a uniform back-end data warehouse to a canonical representation of the data with a flexible and open application-programming interface (API) for management and access. The API will allow adaptors to generate on-demand instances of the data in formats required by each of the modules. For example, Ensembl-style core databases will be instantiated from the canonical Gramene data warehouse for use by genome browsers. The API will also allow us to develop custom user interfaces and will provide channels for programmatic remote access as a web service for community access.

To make the warehouse scalable in terms of data quantity and complexity, we will use a combination of relational databases (MySQL) to capture metadata and strict relationships among components, and non-relational key-value stores to handle very large non-normalized data sets such as next-generation sequences and genotypes. We have explored various options that allow efficient storage and querying for real-time analytics including low-level coordinate-based systems include Fastbit[119], HDF5[120], and Tabix[121] and higher-level technologies such as Hbase[122], MongoDB[122],and CouchDB[123] that include more robust features such as parallel file storage, data integrity, and structured query (e.g., SQL) interfaces. Such technologies are complementary to relational databases.

3.2 *Improve visual interfaces and site navigability.* User interfaces to Gramene's data need to integrate more data and improve access to it. Integration is key to ensure that all relevant data are accessible from a single, predictable view of a given object, such as a gene, transcript, QTL, or genetic marker. Improved access means that it is easy for the user not only to retrieve and visualize results but also to further manipulate the data, such as viewing it in different formats (e.g., HTML, XML, JSON, and plain text), saving the object(s) locally or in a cart, or launching new queries based on what has already been accumulated. The iterative rollout of our open APIs to the Gramene data warehouse will serve to streamline the process of unifying views and interactions with the data.

Mining data. The Gramene Quick Search tool, developed internally and shown ubiquitously across the website, is a popular entry point into Gramene interfaces. It will leverage the warehouse API for faster and more accurate queries across larger and more diverse data sets. The API will also enable rich and contextualized search results, providing a coherent and pleasant user experience. Biomart, generated for Gramene in close collaboration with EBI and which has its own API, will continue to serve as a data-mining and download tool for core genes, functional genomics, genome variation, and comparative genomics. We will add Biomart support for more data types, such as our ontologies due to their relevance to advanced queries and data retrieval. For example, a user could search for genes found within the region defined by a trait associated with a particular phenotype and then determine their orthologs in another species.

Graphical displays. User data indicates that the Ensembl genome browser is the most popular interface on our site, so we intend to focus our development on improving both the variety and scope of data we present there. For example, we intend to integrate expression, pathway, and network data sets into the gene summary pages, link synteny views to comparative map views for the inclusion of non-sequence maps, and create more cross-references to resources such as markers, QTLs, literature, and proteins.

Beyond coordinate-based genome browsing, we are exploring other visual representations of biological data that leverage dynamic, graphical, web toolkits (Aim 1.5). Such displays will allow disparate variables to be arbitrarily intermixed and superimposed, ultimately placing control in the user's fingertips. For example, Reactome's gene expression tool allows a user to color-code expression data overlaid on gene networks based on statistical significance. This gives the end-user a medium to examine underlying biological phenomena and to drive hypothesis-driven science.

3.3 ***Improve community data access***. One of our primary goals has been to provide complete transparency by sharing data sets with the community that were both incorporated from other sources as well as procured internally through our research efforts (e.g., annotations). We will continue to support dynamic access to live Gramene data as well as the download of whole data sets for offline research.

For remote programmatic access, we have provided both a public instance of our databases as well as database dumps for download. We have also provided annotations dynamically as a DAS (Distributed Annotation System) service. Leveraging our open APIs, we intend to expand the web services that Gramene provides, such as SOAP (Simple Object Access Protocol[124]) and REST (Representational State Transfer)[125]. As part of the Virtual Plant Information Network (VPIN[126]) we have published QTL ontology descriptors for the Simple Semantic Web Architecture Protocol (SSWAP[127]). While they were initial prototypes, we intend, in collaboration with iPlant, to broaden our publication of contextual information and metadata descriptors for the Semantic Web as its adoption gains momentum in the life sciences. Our long-term objective is not only to expose these access channels for the community, but to demonstrate our commitment to transparency by funneling internal data access to the Gramene core through the same open APIs.

Gramene will continue to benefit from federation with a growing number of external data and service providers. We will work with other frameworks to develop standards for real-time sharing of data. This will, for example, expose our information to online biological workbenches such as Galaxy and Bioconductor. We intend to distribute the Gramene software, packaged with relevant databases, for each point release as virtual machines (VM). Furthermore, we intend to install Gramene VM images in cloud computing environments such as iPlant (Letter: Stanzione). "Gramene in the Cloud" will provide end users a platform for isolated, high-performance instance of all our resources for their own personal use.

Aim 4. Outreach and Education: Transform the community through communication and training.

Gramene proposes community outreach activities aimed at 1) training plant biologists on the use of the bioinformatics tools we develop, 2) engaging members of the community in data annotation, 3) recruiting expert researchers in developing and supporting standards for successful integration and data exchange and 4) training undergraduates and faculty.

4.1 ***Gramene users (researchers and general public interested in plant biology)***. Gramene will use multimedia sites like YouTube's (TBA) to provide video and audio tutorials as well as FaceBook and iPlant's My-Plant and community forums to reach out to its users. We will organize quarterly online webinars on selected topics in which remote users would interact virtually with Gramene curators and outreach staff, learn about new contents and user interfaces, and seek help for analyzing their own datasets. In addition, community members will be able to directly interact with Gramene staff through presentations at prominent meetings including Plant and Animal Genome, Plant Biology, Maize Genetics, and *Arabidopsis* Research.

4.2 ***Data annotation, exchange and format standardization workshops (collaborators and experts).*** To support development and training on heterogenous data sets prioritized in Aims 1, 2, and 3, we will hold semiannual virtual workshops on community standards and develop self-guided tutorial materials on the annotation of genes, germplasm, and high-throughput sequencing data sets for expression profiling, phenotypes, metabolomics, and epigenetics. Workshop participants will include our collaborators as well as community members identified through NSF's Research Coordination Networks (RCN) projects including the Epigenomics of Plants International Consortium (Letter: Wagner) and Phenotype Ontology (Letter: Huala). The resulting standards and validators will be shared with plant community databases, iPlant (Letter: Stanzione), and our collaborators. In addition to Gramene-sponsored resources, we will participate in regular workshops and developers' meetings hosted by collaborators from EBI Ensembl.

4.3 ***Collaboration with the American Society of Plant Biologists (ASPB) publishing group.*** ASPB publishes the top-ranking journals *Plant Cell* and *Plant Physiology* and could play a critical role in the advertisement and adoption of standards by the plant community by requiring data annotation at the time of manuscript submission. Integrating data that are published in journals with data that reside in databases improves the utility of both. As Gramene hosts information on a growing variety of plant species, we will work with the ASPB journals (Letter: Crispin) and community experts to develop an improved and automated version of the manuscript and data submission form that applies to multiple species. We begin with a focus on targeted pathways (Aim 2) and enlarge our scope through an iterative process as adoption grows. The goal is to train current and future plant biologists how to provide accurate and reliable data sets as we believe that authors make the best curators.

4.4 ***Annotation jamborees (young investigators and students).*** We will organize one jamboree per year on pathway annotations. These will be held at prominent scientific meetings such as those mentioned in aim 4.1. We will use Reactome's curation tools, as well as the popular WikiPathways portal[128] (Letter: Pico) to engage members of the community in our pathway curation efforts.

4.5 ***Public Lecture series at Cold Spring Harbor Laboratory.*** CSHL hosts this series to share the wonders of art and science with the surrounding community. The lectures are free, and offered on campus to an audience of non-scientists (auditorium capacity = 350). Ware will participate twice over the course of the project by giving talks on the current challenges that face agriculture.

4.6 ***Summer internships for undergraduate students and faculty.*** The Ware and Jaiswal laboratories routinely train undergraduate and high school students. Two undergraduates at both CSHL and OSU will participate in summer internship programs. We also propose to host one summer faculty at OSU. Preference will be given to applicants from under-represented groups. Interns will be mentored on projects that combine molecular approaches such as gene expression profiling and computational approaches such as in gene expression and network analysis. In addition, we will provide our interns with educational materials on current plant research topics such as the effects of climate change and population expansion on agriculture and current trends in plant consumption and biofuel applications. We will distribute these materials to educational websites devoted to plant research (e.g., WeedToWonder.org). At the end of each summer, we will be able to gauge the interns' conceptual and procedural knowledge, attitudes toward science, and interest in continuing on plant research through a short survey and oral presentation in their respective host institutions. Their research will be integrated in the Gramene database and publications.