Gramene 2012 Annual Report
Table of Contents

Preliminary Draft of the

Gramene 2012 Annual Report

Executive Summary by Aims

Executive Summary

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.

Activities. In the past year, we expanded our collection of sequenced genomes from 12 to 19, covering 19 different plant species ranging from algae to early land plants to flowering plants, serving both research model plants such as Arabidopsis and agriculturally and energy important crops (rice, maize, sorghum, soybean, tomato, grape, and foxtail millet). In collaboration with the Oryza Genome Evolution (OGE) project, we also brought in sequences of chromosome 3 short arms of two new wild rice species (*Oryza glumaepatula* and *O. meridionalis).* In addition to the computational analysis performed on the raw sequence, which included repeatMasking, fgenesh gene prediction, interproscan protein domain prediction, and data cross-referencing, we performed genome-wide comparative analysis including gene tree building and whole genome alignments against rice and Arabidopsis, as well as an ortholog-based synteny build. We mapped and remapped array probes to the new and updated genome. We also brought in new variation datasets, mapped and remapped them to new and updated genome assemblies. We provide database dumps, a public MySQL database server, BioMart, and blast service for easy data mining and retrieving, and a DAS server for easy integration of the external data. Finally, we updated the ensembl database schema, analysis pipeline, and browser software to the most recent version.

Findings. From gene trees, we developed a method to identify split gene models and tandem duplicated genes. Split gene models could indicate gene misannotation, tandem duplicated gene models could indicate misannotation or true paralogy, both cases call for a detailed investigation to find the underlying cause. These lists were generated and provided to our collaborators for further study. A summary of the results of the analyses suggests that there is an increase in the number of misannotations and these are positively correlated with an increase in genome size.

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

Activities. We made two releases of the pathway module in the past year. RiceCyc was extensively annotated for secondary plant products such as anthocyanins, flavonoids, isoflavonoid, and chalcone biosynthesis. In collaboration with MaizeGDB, we officially released MaizeCyc version 2.0, and subsequently, a minor update 2.0.1 as part of build 35. We started working on transforming the Pathway module based on BioCyc format into the Reactome (www.reactome.org) format. This will allow curation of signaling and regulatory pathways, compared to the strengths of Metabolic pathways display and curation using BioCyc platform. Preliminary import of RiceCyc in Reactome format was successful.

Findings. The MaizeCyc 2.0 pathway module we developed was successful in projecting ~390 metabolic pathways, with 2089 enzymatic reactions associated to 8900 gene products and 1462 compounds or small small molecules. We reanalyzed the Sekhon *et al* (2011)[1] maize expression atlas data from 5 tissues namely, pooled leaves, primary root, anther, embryo and endosperm to find 10,057differentially up-regulated genes in these tissues. Of these about 1,059 differentially expressed metabolic genes mapped to 513 unique enzymatic reactions associated with 308 pathways. The MaizeCyc manuscript was submitted to the Journal Plant Physiology and is in the process of being resubmitted. A similar manuscript on RiceCyc is in preparation. The analysis of RiceCyc metabolic gene network and various gene expression experiment datasets on biotic stress available from Genevestigator[2] and diurnal photoperiod[3] add to several lines of evidence in rice Arabidopsis and other plants along with this study identifying for the first time that serotonin biosynthesis pathway is under diurnal control of the expression and possibly has a role in root system development and response to biotic and abiotic stress.

AIM 3: Capture and elucidate diversity information for each of the sequenced monocot genomes, Gramene Genetic Diversity module will acquire genotypic and phenotypic diversity data for each of the species and recalculate this data using a standardized methodology that allows us to integrate QTL values across species and to relate phenotypic diversity to candidate genes and pathway information.

Activities. The project has continued to deal with the gigantic datasets now being produced by next generation sequencing. These datasets are now scaling to 10s of millions of SNPs by 10s of thousands of taxa. This shift has involved four activities: 1) Curation of massive data sets, 2) development of software tools that use bit level operations to accelerate analysis, 3) novel statistical approaches combining GWAS and comparative analysis, and 4) pioneering data warehouse structures.

Findings. The project has delivered access to a wide range of diversity data for maize, rice, and Arabidopsis. The TASSEL analysis software has thousands of unique users every year, and it has been accelerated by these design changes by 25-fold for many analyses. The PICARA tool has been released and it is powerfully combining candidate gene information with GWAS. Finally, we decided the HDF5 data model frequently used by big data disciplines will be ideal for the large diversity datasets, and we are initiating a community shift to this standard.

AIM 5: Transform the plant research community through communication and training.

Activities. In the last year, Gramene members participated in 18 international conferences to give oral presentations, present posters, strategically meet with our collaborators, and provide hands-on user workshops at the annual PAG, ASPB, and ICAR meetings. We made contributions to seven peer-reviewed publications (five already accepted) and a book chapter. We also fostered numerous collaborations.

Findings. We have underscored the importance of social networking in the plant scientific community and have now established a weekly schedule for contributions to our news blog. We also fostered new partnerships for future collaborations which will render in increased interaction with other plant genome bioinformatics resources, the creation of educational video clips and brochures highlighting or discussing selected plant research topics, in the next year.

[1] **Sekhon RS, Lin H, Childs KL, Hansey CN, Buell CR, de Leon N, Kaeppler SM** (2011) Genome-wide atlas of transcription during maize development. *Plant J* **66:** 553-563

[2] **Zimmermann P, Hennig L, Gruissem W** (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci* **10:** 407-409

[3] **Filichkin SA, Breton G, Priest HD, Dharmawardhana P, Jaiswal P, Fox SE, Michael TP, Chory J, Kay SA, Mockler TC** (2011) Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS One* **6:** e16907

Activities and Findings

The Gramene website is a portal for comparative genomics in plants. The Gramene team members have created and integrated novel and existing software, processes, and ontologies with genomic, genetic, and comparative data sets to add value to the original data and to create an environment for discovery. In this report, we review Gramene's accomplishments during the period from May of 2011 through May of 2012 which includes one major and one minor release of the website and databases, and work that would contribute to release V35 in June. The results are presented in context to the Specific Aims on the grant, with the last section dealing with overall improvements to the website development. This is the report for the last year of funding. We have been granted a no-cost extension to continue work on the deliverables for the remainder of the funds.

Aim 1 (Genomes)

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.

In support of Specific Aim 1, Gramene has assembled a suite of tools to support generation and storage of and access to plant genomic data. Primary data come from community projects and public data banks and are subjected to various secondary computational analyses to become integrated into the Gramene databases. Third-party tools such as Ensembl and BioMart are extended and supplemented by Gramene's own developments.

Within this section we describe updates to the genomic data sets, analysis pipelines, and user interfaces.

Plant genomes hosted in Ensembl Cores

Gramene currently hosts 19 complete genomes. Of these, eight were newly added and three had primary data updates in the past year. The inclusion of *C. reinhardtii***,** *C. merolae*, *B. rapa, G.max, S. italica, S. lycopersicum,* and *S. moellendorffii* adds depth to our phylogenetic analyses, while the wild rice, *O. brachyantha*, adds detail to the *Oryza* genus. In addition to the complete genomes, Gramene hosts nine partial genomes (chromosome 3 short arms from the OGE project).

Table I: Genomes available in Release 35

Status	**Species**	**Assembly and** **Community annotation**
New	*Setaria italica*	JGIv2.1
New	*Brassica rapa*	IVFCAASv1/bra_v1.01_SP2010_01
New	*Solanum lycopersicum*	ITAG2.40
New	*Oryza brachyantha* (FF)	OGEv1.4
New	*Cyanidioschyzon merolae*	ENA1
New	*Glycine max*	Glyma1.0
New	*Selaginella moellendorffii*	ENA1
New	*Chlamydomonas reinhardtii*	ENA1
New	*Oryza glumaepatula* (AA) chr. 3S	454.pools.2012Feb/2012-04-CSHL
New	*Oryza meridionalis* (AA) chr. 3S	454.pools.2012Feb/2012-04-CSHL
Updated	*Oryza sativa Japonica*	IRGSP1.0
Updated	*Oryza glabberima* (AA)	Whole Genome Assembly v1.1 MIPS genes v1.0
Annotation Updated	*Physcomitrella patens*	JGI 1.6
Annotation Updated	*Vitis vinifera* (grape)	GGP 12X Genoscope 2010, annotation V1
Updated	*Oryza minuta* (BB) chr. 3S	CSHLv3.1
Updated	*Oryza minuta* (CC) chr. 3S	CSHLv3.1
Updated	*Oryza officinalis* (CC) chr. 3S	CSHLv3.1
Updated	*Oryza punctata* (BB) chr. 3S	CSHLv2.1
Unchanged	*Oryza nivara* (AA) chr. 3S	454 BAC pools 2009 (July 2010)
Unchanged	*Oryza rufipogon* (AA) chr. 3S	454 BAC pools 2009 (July 2010)
Unchanged	*Oryza barthii* (AA) chr. 3S	454 BAC pool 2008 (Mar 2009)
Unchanged	*Zea mays*	B73 RefGen_v2
Unchagned	*Arabidopsis thaliana*	TAIR 10

Unchanged	*Oryza sativa indica*	BGI 2005 Assembly BGI GLEAN 2008 genes
Unchanged	*Arabidopsis lyrata*	Araly1.2
Unchanged	*Brachypodium distachyon*	Barchy1.2
Unchanged	*Populus trichocarpa*	*JGI 2.0*
Unchanged	*Sorghum bicolor*	Sbi1.4

Secondary Computational Analyses

RepeatMasking

As part of our standard annotations, we perform repeat-masking analysis for each new and assembly-updated genome in preparation for subsequent gene predictions and whole-genome alignments. This includes a suit of repeat-identifying programs including TRF and DUST and uses plant-specific repeat libraries to increase specificity.

Ab-initio gene predictions

All genome assemblies are subjected to *ab initio* gene predictions with FGENESH to provide a standard set of genes as well as to establish an initial set of gene models where the higher quality, evidence-based, community gene annotations are not available.

Feature sequences aligned to the genomes

We have periodically mapped biologically significant DNA features such as ESTs, EST clusters, mRNAs, cDNAs, genetic markers, insertion sites, and BAC-end sequences to the hosted genome assemblies using an internally developed BLAT pipeline. These alignments provide an aggregated comprehensive context of the region and can be viewed as tracks positioned below the genomic regions on the genome browser.

Database Cross References

Cross-reference (Xref) pipelines link alternate identifiers through sequence comparisons from different data sources such as Swiss-prot and Uniprot.

Protein Domains

InterProScan pipelines are run against new and updated gene annotations to identify functional domains of the proteins.

Ontology Annotations

We use terms from two Open Biomedical Ontologies projects, Gene Ontology (GO) and Plant Ontology (PO) for our functional annotation efforts. Terms are variously mapped using four separate methods: 1) using existing gene assignments from the GO and PO projects that are mapped directly to our genes via common identifiers; 2) using gene assignments from

UniProtKB-GOA that are mapped to our genes through sequence similarity to UniProt proteins; 3) using protein domain assignments from Interpro2GO that are mapped to our genes through protein domains annotated by InterproScan; and 4) by transferring assignments between species via orthology relationships as determined by the Ensembl Compara gene tree method.

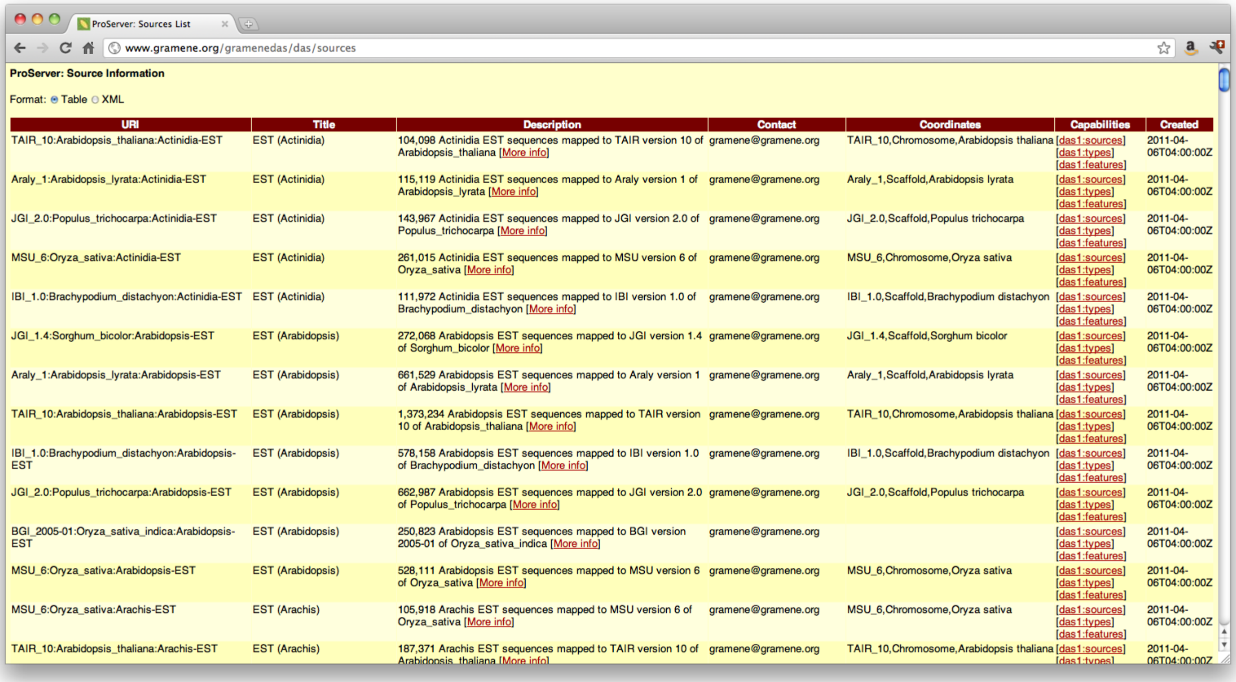
Software

Each release of Gramene is accompanied by a software update to the latest Ensembl version. In releases 34 and 34b of Gramene, we upgraded to Ensembl versions 64 and 65, respectively, and contributed software, documentation and bug reports to that project. For the coming release (build 35) in June, we are upgrading the software to Ensembl version 67, which includes major software/pipeline and schema changes.

Transcriptome sequencing (RNA-seq) has become an important means of gene discovery, annotation, and expression profiling. The Ensembl infrastructure provides several avenues for the data to be displayed including the display of the data through a DAS track. Additionally a user can upload his data directly in a GFF and BED formats. We currently host tracks for publicly available RNA-seq data on the maize browser and anticipate adding such data to our rice and *Arabidopsis* browsers in the future.

Distributed Annotation Server (DAS)

Gramene provides a DAS server with 595 tracks of sequence alignments. These can be used by the Ensembl genome browser as well as outside applications with a need to display sequences annotations on our various genome assemblies. Substantial improvements to the architecture have improved performance and usability considerably.



ProServer: Source Information

Format: ● Table ○ XML

URI	Title	Description	Contact	Coordinates	Capabilities	Created
TAIR_10:Arabidopsis_thaliana:Actinidia-EST	EST (Actinidia)	104,098 Actinidia EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [More info]	gramene@gramene.org	TAIR_10,Chromosome,Arabidopsis thaliana	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
Araly_1:Arabidopsis_lyrata:Actinidia-EST	EST (Actinidia)	115,119 Actinidia EST sequences mapped to Araly version 1 of Arabidopsis_lyrata [More info]	gramene@gramene.org	Araly_1,Scaffold,Arabidopsis lyrata	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
JGI_2.0:Populus_trichocarpa:Actinidia-EST	EST (Actinidia)	143,967 Actinidia EST sequences mapped to JGI version 2.0 of Populus_trichocarpa [More info]	gramene@gramene.org	JGI_2.0,Scaffold,Populus trichocarpa	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
MSU_6:Oryza_sativa:Actinidia-EST	EST (Actinidia)	261,015 Actinidia EST sequences mapped to MSU version 6 of Oryza_sativa [More info]	gramene@gramene.org	MSU_6,Chromosome,Oryza sativa	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
IBI_1.0:Brachypodium_distachyon:Actinidia-EST	EST (Actinidia)	111,972 Actinidia EST sequences mapped to IBI version 1.0 of Brachypodium_distachyon [More info]	gramene@gramene.org	IBI_1.0,Scaffold,Brachypodium distachyon	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
JGI_1.4:Sorghum_bicolor:Arabidopsis-EST	EST (Arabidopsis)	272,068 Arabidopsis EST sequences mapped to JGI version 1.4 of Sorghum_bicolor [More info]	gramene@gramene.org	JGI_1.4,Scaffold,Sorghum bicolor	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
Araly_1:Arabidopsis_lyrata:Arabidopsis-EST	EST (Arabidopsis)	661,529 Arabidopsis EST sequences mapped to Araly version 1 of Arabidopsis_lyrata [More info]	gramene@gramene.org	Araly_1,Scaffold,Arabidopsis lyrata	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
TAIR_10:Arabidopsis_thaliana:Arabidopsis-EST	EST (Arabidopsis)	1,373,234 Arabidopsis EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [More info]	gramene@gramene.org	TAIR_10,Chromosome,Arabidopsis thaliana	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
IBI_1.0:Brachypodium_distachyon:Arabidopsis-EST	EST (Arabidopsis)	578,158 Arabidopsis EST sequences mapped to IBI version 1.0 of Brachypodium_distachyon [More info]	gramene@gramene.org	IBI_1.0,Scaffold,Brachypodium distachyon	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
JGI_2.0:Populus_trichocarpa:Arabidopsis-EST	EST (Arabidopsis)	662,987 Arabidopsis EST sequences mapped to JGI version 2.0 of Populus_trichocarpa [More info]	gramene@gramene.org	JGI_2.0,Scaffold,Populus trichocarpa	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
BGI_2005-01:Oryza_sativa_indica:Arabidopsis-EST	EST (Arabidopsis)	250,823 Arabidopsis EST sequences mapped to BGI version 2005-01 of Oryza_sativa_indica [More info]	gramene@gramene.org		[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
MSU_6:Oryza_sativa:Arabidopsis-EST	EST (Arabidopsis)	528,111 Arabidopsis EST sequences mapped to MSU version 6 of Oryza_sativa [More info]	gramene@gramene.org	MSU_6,Chromosome,Oryza sativa	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
MSU_6:Oryza_sativa:Arachis-EST	EST (Arachis)	105,918 Arachis EST sequences mapped to MSU version 6 of Oryza_sativa [More info]	gramene@gramene.org	MSU_6,Chromosome,Oryza sativa	[das1:sources] [das1:types] [das1:features]	2011-04-06T04:00:00Z
TAIR_10:Arabidopsis_thaliana:Arachis-EST	EST (Arachis)	187,371 Arachis EST sequences mapped to TAIR version 10 of Arabidopsis_thaliana [More info]	gramene@gramene.org	TAIR_10,Chromosome,Arabidopsis thaliana	[das1:sources] [das1:types]	2011-04-06T04:00:00Z

Figure 1: Sample display of DAS tracks available at http://www.gramene.org/gramenedas/das/sources.

Plant expression array probes available through the Ensembl Functional Genomics module

Our collaborators in EBI have provided the array-based probes sets as part of the functional genomics databases allowing users to visualize the data in genomic context as well as integrating into BioMart-based queries. Given any list of genes, for example, a user could sort based on the availability of array data. This would allow the user to know which list of genes may have expression data available and further filter to narrow a candidate gene list for genetic marker development.

Table II: Genomes with plant expression array probes mapped to genes

Arabidopsis thaliana	Unchanged (CATMA/Agilent)
Oryza sativa indica	Unchanged (Agilent/NSF)
Oryza sativa japonica	Unchanged (Agilent/NSF)
Populus trichocarpa	Unchanged (Affy array)
Vitis vinifera	Updated mapping on Grape V1 genes (Affy array)

Plant variation hosted in Ensembl Variation

During the past year, the collaborations between Gramene, EBI, and the OGE project have brought the *Oryza glaberrima* resequencing data into the Ensembl Variation framework. The collaboration with Genetic Architecture of Maize and Teosinte (NSF 0820619 PI Buckler) project provided Gramene with HapMap2 data. This work has mainly entailed the development of loading tools of data from standard formats and reference resources including dbSNP. Support of the data in the variation modules provides the plant community with robust, web-based views for visualizing the data in genomic context. A preliminary view from release 35 showing *O. glaberrima* variations (OGE NSF 1026200 PI Wing) in their genome context is given below.

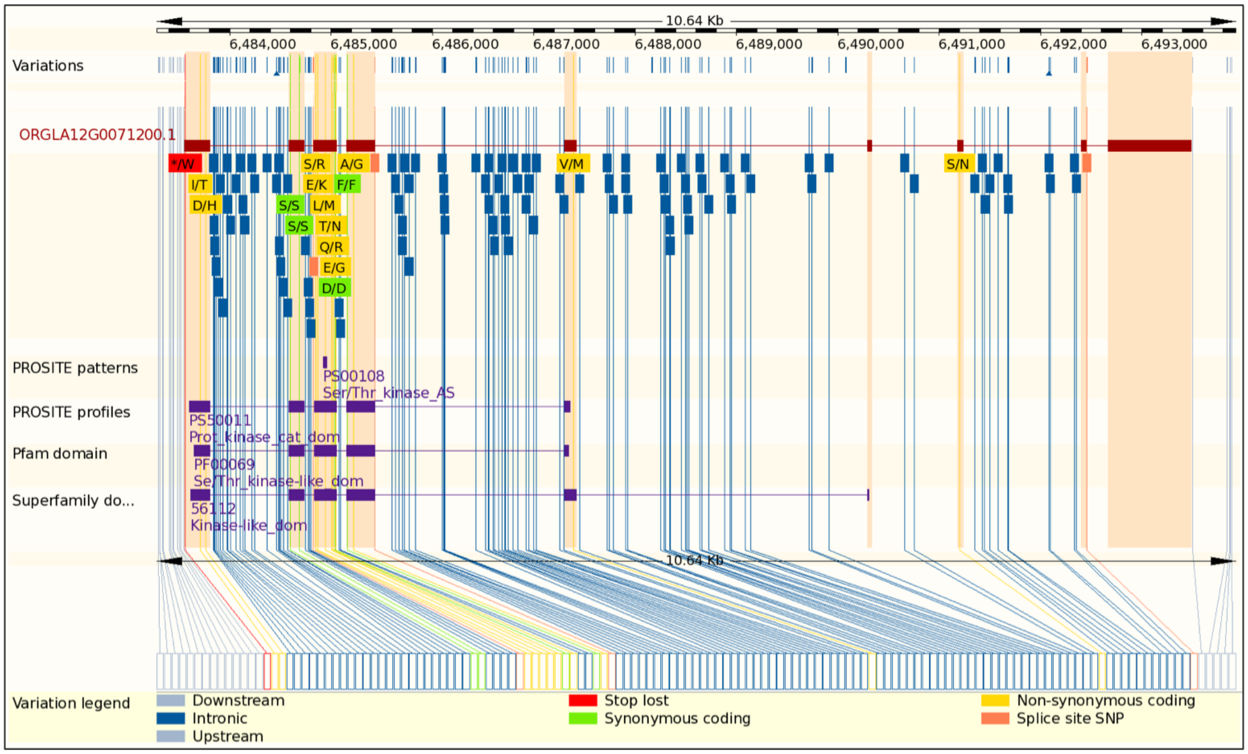


Figure 2: The functional consequences of the SNP is color coded to support information on genomic context and function.

Additional views of the data are available in table format and through BioMart queries that enable users to identify loss of function (stop codon) and transcription factors for a specific genome. The table below shows the data sets that will be released in the variation module in release 35.

Table III: Variation data sets housed in variation module

Rice	160,000 SNPs x 21 varieties (incl. Nipponbare ref.) from OryzaSNP, MSU6

Maize	1.6 million SNPs x 27 NAM founder lines from Panzea, AGPv2
Maize	55 million SNPs & indels x 103 pre-domesticated and domesticated *Zea mays* varieties, including related *Tripsacum dactyloides* (Eastern gamagrass); HapMap v2
Arabidopsis	2010 Project SNP Discovery: 637,522 SNPs x 21 ecotypes (incl. Col-0 ref.), TAIR9
	2010 Project 250K SNP chip genotypes v3.04, 214,000 SNPs x 1179 ecotypes, TAIR9
	1001 Genomes/WTCHG SNPs from dbSNP, 2.7 million SNPs, 17 ecotypes, TAIR9
Grape	71K SNPs (Myles *et al.*)
O.glaberrima	828K Variants from OGE project

BioMart

As part of Gramene's collaboration with Ensembl Genomes (EG) project, the Ensembl Mart databases and interfaces are built at EBI. In addition to these, we provide our own Marts for QTL and markers/sequences database. Plant variation Mart allows users to download SNPs in genomic regions.

Comparative Genome Analyses (Ensembl Compara)

Gene Tree Prediction and Synteny Analysis

We updated the gene trees using the Ensembl Compara protein pipeline. The GeneTree database was rebuilt for releases 34 and 35, as part of the twice-a-year data update. In interim release 34b, we added an *Oryza*-centered gene tree build from whole genome and chr3 short arm genes. For the plant tree, there are a total of 35,571 individual trees and 523,406 genes in the latest release; for the *Oryza* tree, there are a total of 17,435 individual trees and 221,563 genes in the latest release.

In the past, the phylogenetic analyses has been used to support gene-centered synteny analyses. The automated pipeline uses Compara orthologs to find collinear mappings (via DAGchainer) and defines syntenic blocks. In the previous release, the build was limited to monocots; in build 34 we added the Maize AGPv2 *vs* Rice and Maize APG2 *vs* Sorghum ortholog-based synteny data. The table below provides a summary of the analyses available.

Table IV: Showing all the Synteny analysis in Gramene35

Oryza sativa Japonica	***O.jap***						
Brachypodium distachyon	YES	***B.dis***					
Sorghum bicolor	YES	YES	***S.bic***				
***Zea mays* (AGPv2)**	YES		YES				
Arabidopsis thaliana	-	-	-	***A.tha***			
Arabidopsis lyrata	-	-	-	YES	***A.lyr***		
Vitis vinifera	-	-	-	YES	YES	***V.vin***	
Poplar trichocarpa	-	-	-	YES	YES	YES	***P.tri***

Automated Detection of Putative Split Genes

During the last year, Gramene applied its split-genes workflows to support detection of mis-annotated split genes in reference genomes hosted at Gramene. The split gene models are commonly related to an annotation artifact wherein a single gene is annotated as two or more genes due to incomplete evidence, but could also result from legitimate evolutionary processes. The Compara Gene Tree method predicts a special class of within-species paralogs called "contiguous_gene_split". A contiguous_gene_split is called when the two apparently paralogous genes lie on the same strand and in close proximity (<1MB) but have no (or little) overlapping sequence. The gene tree figure and whole genome alignment view between rice and sorghum, highlighting an example of a split gene in rice is given below.

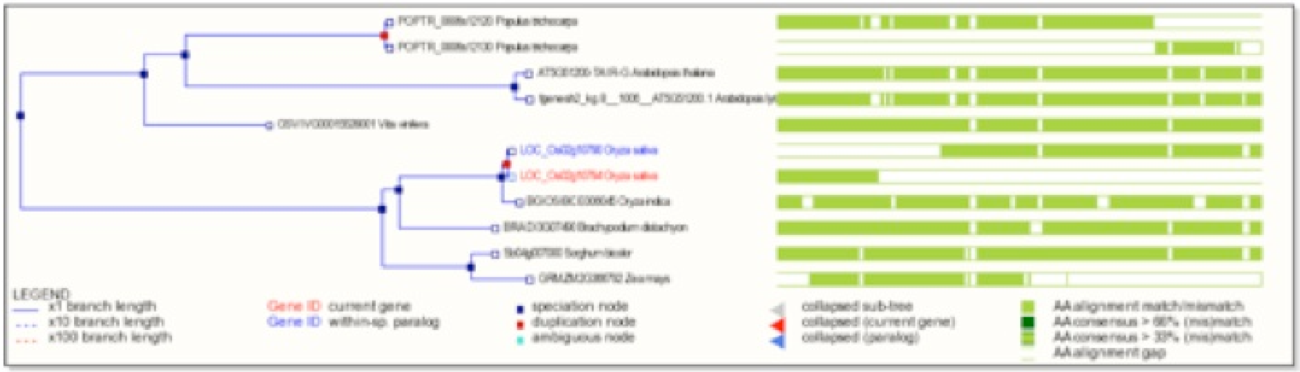


Figure 3: A misannotation of a rice gene structure observed as truncated paralog in this gene tree.

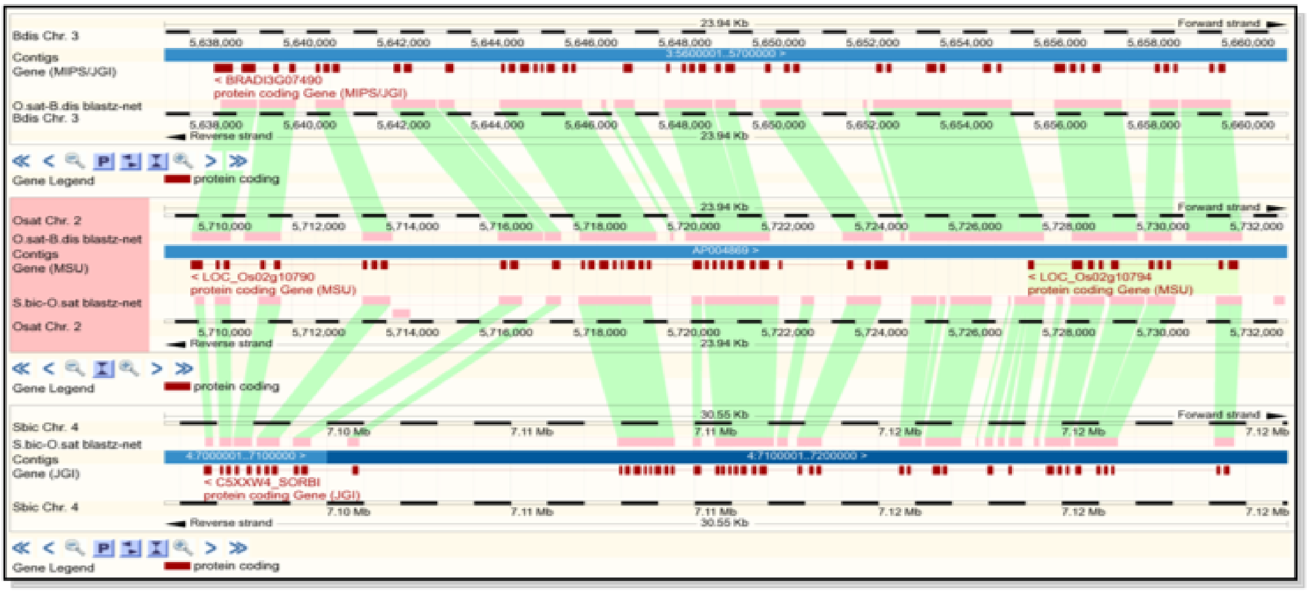


Figure 4: The whole genome alignments support the truncated rice gene trees

Table V: Predicted contiguous split gene models by species

Species	Split Genes
Arabidopsis lyrata	40
Arabidopsis thaliana	8
Brachypodium distachyon	36
Brassica rapa subsp. pekinensis	28
Chlamydomonas reinhardtii	28
Glycine max	100
Oryza brachyantha	104
Oryza glaberrima	194
Oryza sativa indica	224
Oryza sativa japonica	158
Physcomitrella patens	58
Populus trichocarpa	574
Selaginella moellendorffii	62

Setaria italica	140
Solanum lycopersicum	298
Sorghum bicolor	434
Vitis vinifera	368
Zea mays	400

Whole Genome Pairwise Alignments

In release 35, Gramene will be upgrading the raw alignment tool from BLASTZ to LASTZ, a drop-in replacement for BLASTZ that is backward compatible with BLASTZ's command-line syntax. It supports all of BLASTZ's options but also has additional ones and may produce slightly different alignment results. Gramene creates (B)LASTZ-net alignments for closely related pairs of species. The alignments are the results of post-processing the raw BLASTZ or LASTZ results. In the first step, original blocks are chained according to their location in both genomes. The netting process chooses the best sub-chain in each region for the reference species.

In release 34, we generated the following WGA alignments:

- BLASTZ-net between *Oryza sativa Japonica* and *Oryza glaberrima* whole genome
- BLASTZ-net between *Oryza sativa Japonica* and each of the 4 updated OGE/OMAP chromosome 3 short arms (*O.minuta (BB), O.minuta (CC), O.officinallis, O.punctata*)
- BLASTZ-net between *Oryza sativa Japonica* and *Glycine max*
- BLASTZ-net between *Oryza sativa Japonica* and *Chlamydomonas reinhardti*
- BLASTZ-net between *Oryza sativa Japonica* and *Selaginella moellendorffii*
- BLASTZ-net between *Arabidopsis thaliana* and *Glycine max*
- BLASTZ-net between *Arabidopsis thaliana* and *Chlamydomonas reinhardti*
- BLASTZ-net between *Arabidopsis thaliana* and *Selaginella moellendorffii*

In release 35, we are adding the following pairs:

- LAST-net between *Oryza sativa Japonica* and *Solanum lycopersicum* (tomato)
- LAST-net between *Oryza sativa Japonica* and *Setaria italica* (Foxtail millet)
- LAST-net between *Oryza sativa Japonica* and *Cyanidioschyzon merolae* (red algae)
- LAST-net between *Oryza sativa Japonica* and *Brassica rapa* (Chinese cabbage)
- LAST-net between *Oryza sativa Japonica* and *Oryza brachyantha* (wild rice)
- LAST-net between *Oryza sativa Japonica* and *Oryza glumaepatula* chr3s
- LAST-net between *Oryza sativa Japonica* and *Oryza meridionalis* chr3s
- LAST-net between *Arabidopsis thaliana* and *Solanum lycopersicum* (tomato)
- LAST-net between *Arabidopsis thaliana* and *Setaria italica* (Foxtail millet)
- LAST-net between *Arabidopsis thaliana* and *Cyanidioschyzon merolae* (red algae)

- LAST-net between *Arabidopsis thaliana* and *Brassica rapa* (Chinese cabbage)
- LAST-net between *Arabidopsis thaliana* and *Oryza brachyantha* (wild rice)

The following chart shows previous pairs of genomes have been compared.

Table VI: Showing all the blastz-chain-net pairwise whole genome alignments available in Gramene35

O. sativa japonica	***O. jap***				
O. sativa indica	YES	***O. ind***			
S. bicolor	YES	-	***S. bic***		
B. distachyon	YES	YES	-	***B. dis***	
A. thaliana	YES	YES	YES	YES	***A. tha***
A. lyrata	YES	-	-	-	YES
Vitis vinifera	YES		-	-	YES
Populus trichocarpa	YES	-	-	-	YES
Physcomitrella patens	YES	-	-	-	YES
O. barthii 3S	YES	-	-	-	-
O. brachyantha 3S	YES	-	-	-	-
O. glaberrima 3S	YES	-	-	-	-
O. minuta (BBCC) 3S	YES	-	-	-	-
O. nivara 3S	YES	-	-		-
O. officinalis 3S	YES	-	-	-	-
O. punctata 3S	YES	-	-	-	-
O. rufipogon 3S	YES	-	-	-	-

Markers database

Gramene maintains a custom MySQL database to house almost 42 million plant markers and sequences from GenBank and various mapping studies. These are used in our annotation pipelines for our completed genomes. This database also holds the results of the alignments as well as manually curated maps from the community and literature. We then build our comparative maps (CMap) from this resource as well as our DAS. Moving forward we will be evaluating the use of the relational store for these data sets.

Distributed Annotation Server (DAS)

Gramene provides a DAS server with 595 tracks of sequence alignments. These can be used by the Ensembl genome browser as well as outside applications to display sequences annotations on our various genome assemblies. Substantial improvements to the architecture have improved performance and usability considerably.

Germplasm

Our germplasm database has been updated to include 1,300 germplasm from *Arabidopsis thaliana*, 4,600 from *Oryza*, 8,200 from *Zea*, and several more from *Tripsacum* to now hold 12,927 records with links to Gramene's markers, map sets, genes and proteins.

BLAST server

Gramene continues to provide BLAST services for sequence searching against the complete genomic, predicted cDNA, and predicted peptide sequences of all genomes that we maintain.

Comparative genetic and physical maps

During last year, Gramene comparative maps module (CMap) had no major updates in data or software.

Aim 2 (Pathways)

Enhance the value of the comparative maps with pathway, phenotypic and other functional information from rice, maize, and Arabidopsis.

BrachyCyc and MaizeCyc metabolic pathway databases were officially released in Gramene release 34. MaizeCyc was built by Gramene developers and curators and was released simultaneously in collaboration with the MaizeGDB project (cf. http://news.gramene.org/?p=660), and was upgraded to official release status in an interim release after build 33.

RiceCyc was updated to version 3.2, BrachyCyc to version 2, MaizeCyc to version 2.0, SorghumCyc to version 1.1 (see table below). Several non-plant pathways originating from MetaCyc projections were removed from RiceCyc (n=27), SorghumCyc (n=26), and BrachyCyc (n=27).

Table VII. Overall counts in RiceCyc, BrachyCyc, MaizeCyc and SorghumCyc

Class	RiceCyc 3.2	BrachyCyc 2.0	MaizeCyc 2.0	SorghumCyc 1.1
Pathways	316	327	390	302
Enzymatic Reactions	2103	2057	2,109	1838
Transport Reactions	87	87	68	9
Polypeptides	47894	26633	39656	36347
Enzymes	6050	7723	8894	10636
Transporters	603	950	291	269
Compounds	1543	1641	1467	1356

MaizeCyc 2.0

MaizeCyc is a catalog of known and/or predicted metabolic and transport pathways from maize (*Zea mays* ssp. *mays*) and was developed by personnel at Gramene and the maize model organism database, MaizeGDB, in collaboration with the Maize Genome Sequencing Project (MGSC). The representation of a total of 390 pathways involving 4,106 genes in this catalog is based on the electronic and manual annotations of the B73 RefGen_v2 gene models. It includes various sequence-based associations provided by Gramene, MaizeSequence.org, and MaizeGDB to external database entries from EntrezGene, UniProt-SwissProt, and GenBank. In this round, manual annotations of genes included mapping of classical phenotype genes to sequenced genomic loci provided by Schnable and Freeling (cf. http://synteny.cnr.berkeley.edu/wiki/index.php/Classical_Maize_Genes), and proteomics-supported gene annotations from Friso et al (2010) (cf. http://www.ncbi.nlm.nih.gov/pubmed?term=20089766). The database was created using the Pathway Tools PathoLogic module developed by Peter D. Karp and co-workers at the Bioinformatics Research Group at SRI International.

A manuscript describing the creation of MaizeCyc and a test case transcriptome analysis using MaizeCyc is in the final stages of editing to be resubmitted to *Plant Physiology*.

Progress on Plant Reactome Projects

Due to limitations on extending the current metabolic pathway databases to accommodate the regulatory and signaling pathways and reactions, it was decided to use the tools developed by Reactome (http://www.reactome.org) -- built for human and metazoan communities and funded by NIH -- as a model to develop a “Plant Reactome” portal. Discussions on the design/architecture of the “Plant Reactome” portal are in progress. In order to achieve this goal, the following strategies were successfully applied:

Rice Reactome

- Started with RiceCyc import in BioPax level-2 format and built on the existing Reactome and Ensembl curated gene database resources.

- The Rice Reactome was integrated in the Reactome Central database where all the Reactome curation is stored centrally at the Ontario Institute for Cancer Research (OICR), Toronto , Canada. The central database is the curation hub for all curators and developers working on Reactome and exists independent of the species.

- Rice Reactome curation progress is listed below.

Table VIII. Progress in the pathway curation activities for Rice Reactome

Rice Reactome Curation Step	Curational Activity	Number of pathways
1	Select & Upload Reference entity properties from UniProt	80
2	Enter Catalytic property from GO	75
3	Curate gene sets of reactions in the pathway	35
4	Assign compartment to each reaction and gene product	75
5	Write summations for Reactions and Pathways	75
6	Layout pathway diagrams	75

To expedite steps 1 and 2 outlined in the above table we are in the process of implementing a bulk UniProt data retrieval and Rice-Reactome-DB upload strategy.

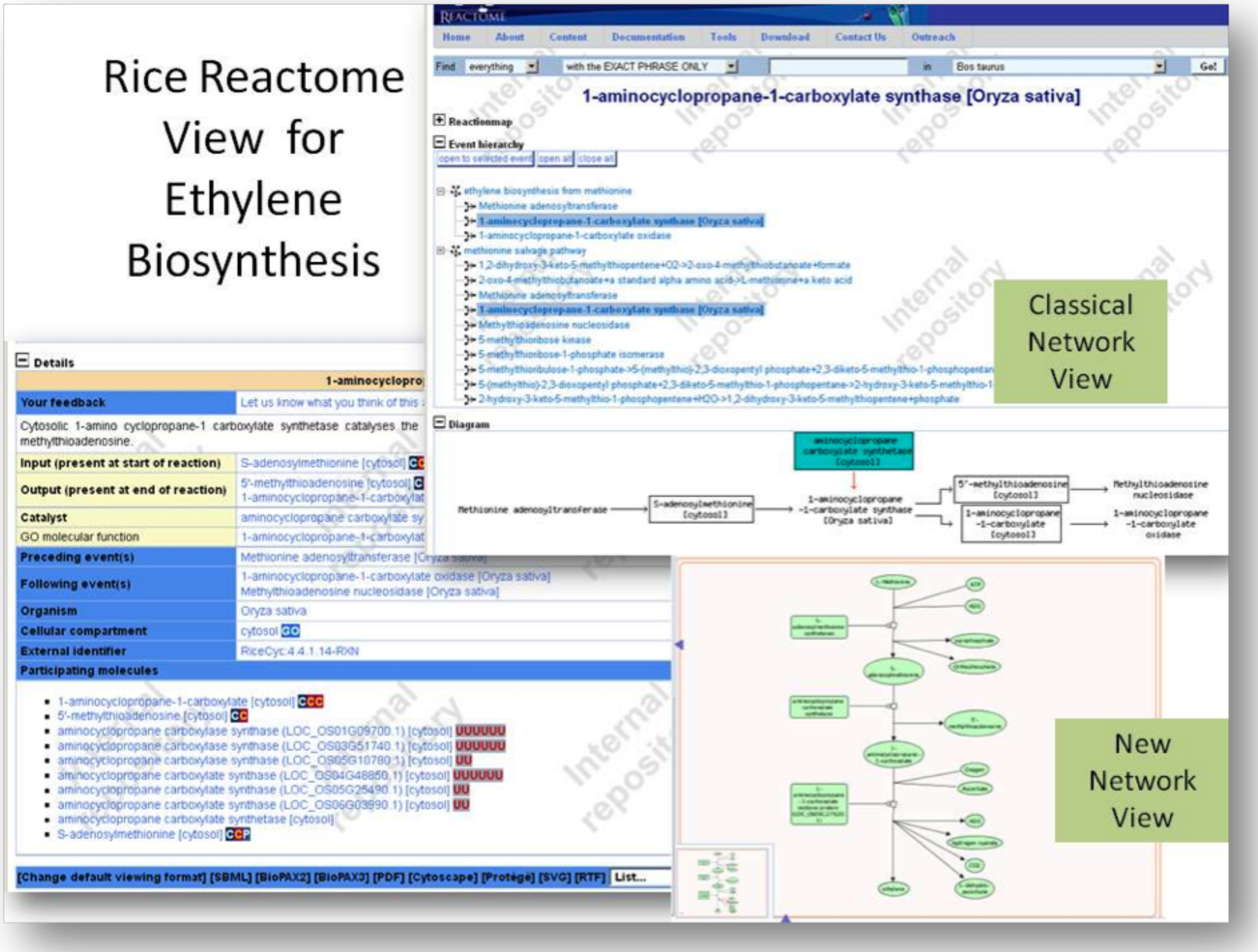
Arabidopsis Reactome

- After consulting with Nick Provart’s group at the University of Toronto, Canada; the Reactome project (EBI, US and Canada); and Henningt Hermjakob, who is responsible for maintaining the current Arabidopsis Reactome group at EBI (UK), it was decided to subsume the Arabidopsis Reactome project into the Plant Reactome project proposed by Gramene. International labs/projects of Provart and Hermjakob groups will be the primary curators for Arabidopsis Reactome, with all the major development and sublime curation support coming from Gramene.

- After establishing the quality checks, the processed Arabidopsis Reactome was integrated in the Reactome Central database (Ontario Institute for Cancer Research - OICR)

- Improvements made towards expediting the rice Reactome release would also be applied to Arabidopsis Reactome

- Nick Provart’s group at the University of Toronto and Jaiswal lab (OSU, Gramene) will carry out development and curation of Arabidopsis pathways following the protocols established for Rice Reactome



Rice Reactome View for Ethylene Biosynthesis

1-aminocyclopropane-1-carboxylate synthase [Oryza sativa]

⊞ Reactionmap

⊟ Event hierarchy

- ethylene biosynthesis from methionine
  - Methionine adenosyltransferase
  - 1-aminocyclopropane-1-carboxylate synthase [Oryza sativa]
  - 1-aminocyclopropane-1-carboxylate oxidase
- methionine salvage pathway
  - 1,2-dihydroxy-3-keto-5-methylthiopentene+O2->2-oxo-4-methylthiobutanoate+formate
  - 2-oxo-4-methylthiobutanoate+a standard alpha amino acid->L-methionine+a keto acid
  - Methionine adenosyltransferase
  - 1-aminocyclopropane-1-carboxylate synthase [Oryza sativa]
  - Methylthioadenosine nucleosidase
  - 5-methylthioribose kinase
  - 5-methylthioribose-1-phosphate isomerase
  - 5-methylthioribulose-1-phosphate->5-(methylthio)-2,3-dioxopentyl phosphate+2,3-diketo-5-methylthio-1-phosphopentan
  - 5-(methylthio)-2,3-dioxopentyl phosphate+2,3-diketo-5-methylthio-1-phosphopentane->2-hydroxy-3-keto-5-methylthio-1
  - 2-hydroxy-3-keto-5-methylthio-1-phosphopentene+H2O->1,2-dihydroxy-3-keto-5-methylthiopentene+phosphate

Classical Network View

⊟ Diagram

aminocyclopropane carboxylate synthetase [cytosol]
Methionine adenosyltransferase
S-adenosylmethionine [cytosol]
1-aminocyclopropane -1-carboxylate synthase [Oryza sativa]
5'-methylthioadenosine [cytosol]
1-aminocyclopropane -1-carboxylate [cytosol]
Methylthioadenosine nucleosidase
1-aminocyclopropane -1-carboxylate oxidase

⊟ Details

	1-aminocyclopro
Your feedback	Let us know what you think of this
Cytosolic 1-amino cyclopropane-1 carboxylate synthetase catalyses the methylthioadenosine.	
Input (present at start of reaction)	S-adenosylmethionine [cytosol] CC
Output (present at end of reaction)	5'-methylthioadenosine [cytosol] C 1-aminocyclopropane-1-carboxylat
Catalyst	aminocyclopropane carboxylate sy
GO molecular function	1-aminocyclopropane-1-carboxylat
Preceding event(s)	Methionine adenosyltransferase [Oryza sativa]
Following event(s)	1-aminocyclopropane-1-carboxylate oxidase [Oryza sativa] Methylthioadenosine nucleosidase [Oryza sativa]
Organism	Oryza sativa
Cellular compartment	cytosol GO
External identifier	RiceCyc:4.4.1.14-RXN
Participating molecules	

- 1-aminocyclopropane-1-carboxylate [cytosol] CCC
- 5'-methylthioadenosine [cytosol] CC
- aminocyclopropane carboxylase synthase (LOC_OS01G09700.1) [cytosol] UUUUUU
- aminocyclopropane carboxylase synthase (LOC_OS03G51740.1) [cytosol] UUUUUU
- aminocyclopropane carboxylase synthase (LOC_OS05G10780.1) [cytosol] UU
- aminocyclopropane carboxylate synthase (LOC_OS04G48850.1) [cytosol] UUUUUU
- aminocyclopropane carboxylate synthase (LOC_OS05G25490.1) [cytosol] UU
- aminocyclopropane carboxylate synthase (LOC_OS06G03990.1) [cytosol] UU
- aminocyclopropane carboxylate synthetase [cytosol]
- S-adenosylmethionine [cytosol] CCP

[Change default viewing format] [SBML] [BioPAX2] [BioPAX3] [PDF] [Cytoscape] [Protégé] [SVG] [RTF] List...

New Network View

Figure 5: A modified view of the Rice Reactome view, showing reactions from the ethylene biosynthesis pathway.

Aim 3 (Diversity)

To capture and elucidate diversity information for each of the sequenced monocot genomes, Gramene Genetic Diversity module will acquire genotypic and phenotypic diversity data for each of the species and recalculate this data using a standardized methodology that allows us to integrate QTL values across species and to relate phenotypic diversity to candidate genes and pathway information.

Data curation activities in the Genetic Diversity module (http://www.gramene.org/db/diversity/diversity_view) for the last year have focused on large diversity and genome-wide association studies in Maize, Rice and *Arabidopsis* being produced by the NSF Maize Diversity Project (www.panzea.org), the NSF Rice Diversity Project (www.ricediversity.org), and the NSF *Arabidopsis* project, as well as from a few smaller projects. We have concentrated on curation of single-gene rice phenotype studies in order to capture the natural variation among the accessions of both wild type and cultivated species of *Oryza* with a focus on domestication related traits such as grain quality, yield, flowering time and disease resistance. An emphasis was placed on curating studies that employ large germplasm panels in their analyses. Data sets added to the module during the last year are summarized in the table below.

During last year, we started making the shift to supporting genotypic data sets with billions of datapoints. We have finished schema and analysis tool support for these massive datasets. However, the technology keeps moving extremely rapidly and we expect the maize and rice communities to have genotypic data sets with a trillion data points by the end of the year. In collaboration with the Maize Diversity Project and iPlant, we have developed a NetCDF4/HDF5 based genotypic data structure, and server/client software to write and read NetCDF4/HDF5 files. The system is currently be tested on Maize GBS genotypic data with 20 billion data points. We will continue to explore and develop community approaches for storing, accessing, and analyzing trillion datapoint data sets.

The SNP Query tool (http://www.gramene.org/db/diversity/snp_query) can be used to retrieve and filter SNP data by chromosome or cultivar subgroups and can now display genes and QTLs overlapping with SNPs of interest. The tool also now allows users to search by a gene, QTL, or trait name or ID. Included in Gramene release (Build 35), the Phenotype Study web interface (built using the REST web services architecture), which displays phenotype measurement data and all relevant ontological information. The interface displays related genes and QTLs. The Phenotype Study interface will be a useful new tool for plant breeders.

TASSEL has undergone significant development in the past year resulting in the major point release, TASSEL 4.0. It includes many performance, stability, and efficiency improvements, related to the following: Consistent handling of data types, bitwise storage and algorithms, an alignment viewer, QQ/Manhattan plots, progress monitoring, LD display, taxa and site filtering, genotype summary, and a command line interface. Many bugs were fixed and the user guide was updated. There have over 1,000 downloads of the source code by power users, tens of thousands downloads of the various versions of the software, and we estimate there are move than 3,000 active TASSEL users. TASSEL is now available on iPlant Discovery Environment and Atmosphere.

Research in the Genetic Diversity module in the past year includes the development of *PICARA*, an analytical pipeline designed to systematically summarize observed SNP/trait associations identified by genome wide association studies (GWAS) and to identify candidate genes involved in the regulation of complex trait variation. The pipeline provides probabilistic inference about *a priori* candidate genes using integrated information derived from genome-wide association signals, gene homology and conserved synteny searches, and curated gene sets embedded in pathway descriptions. The results of such analyses allow for a summarization of heterogeneous biological information and encourage the cross reference of expert knowledge in database biology. We also demonstrate the performance of *PICARA* using flowering time variation and discover a regulatory feature that is characteristic of these *a priori* flowering time candidates in maize. The publication of *PICARA* and its results on flowering time variation has been submitted for peer review.

Table IX. Genetic Variation Data Sets in Arabidopsis, Rice and Maize

Species	**Data Sets**
Arabidopsis	Arabidopsis 2010 Nordborg v3.06 - 214,052 SNPs x 1,307 germplasm, TAIR10 coordinates
Rice	Updated coordinates (MSU6->MSU7) for the following rice diversity datasets: * Zhao K *et al*, *PLoS* May 2010, Rice Diversity 1536 SNP array: 1311 SNPs x 395 varieties * Zhao K *et al*, *Nat Comm*, 2011. Rice GWAS study. 413 diverse accessions, Rice Diversity 44K SNP array x 34 phenotypes

Maize	Chia *et al*, *Nat Genet* 2012. HapMap v2 set. 55 million SNPs and indels x 103 pre-domesticated and domesticated *Zea mays* varieties and *Tripsacum dactyloides*

Aim 4 (PO)

Support the Plant Ontology (Two Years Only)

The Plant Ontology (PO) portion of the original Gramene plan has been moved to a separate project headed by Pankaj Jaiswal at Oregon State University.

Aim 5 (Outreach)

Education, Outreach and Diversity

Gramene's outreach activities encompassed an array of activities for sharing the information by creating tutorial materials, publicizing our work and news items at different platforms, organizing Gramene workshops at several meetings, establishing collaborations with researchers to integrate their work into our analyses, and much more. A summary of our efforts to connect with our users and stakeholders is given below.

Collaborations

The Gramene team maintains collaboration with several domestic and international researchers. Here is the list of some of the major groups with whom we are working:

- EnsemblGenomes (EG): Since 2009, Gramene shares the burden of producing the "core" Ensembl genome databases with the EG team at the EBI (UK). Gramene builds the Compara databases, and EG builds the Mart databases. A standard set of quality checks is performed on all databases, and our releases are coordinated to complement each other. This collaboration has increased the productivity of both teams by efficiently dividing the work needed to annotate and compare plant genomes as well as to generate the data mining tools needed by plant researchers.

- Plant Ontology: Gramene continues to be an active participant in the Plant Ontology consortium. Gramene makes use of the PO as a controlled vocabulary for describing anatomy and developmental stages as part of the automated annotation process. With each release Gramene submits our annotation file gene, and QTL associations to PO CVS repository.
- Gene Ontology: Gramene continues to be an active participant in the GO consortium. We use the gene ontology from GO cvs repository as a standard vocabulary within the project website. We submit our annotation files of gene ontology and rice protein associations to GO CVS repository at each Gramene release.
- Uniprot: We get all Poaceae proteins from Uniprot website at each gramene release. Uniprot links back to gramene protein pages as cross references.
- iPlant (Steve Goff, Dan Stanzione, Matt Vaughn): Gramene has provided reference genome sequences and annotations for the hosting in the DE.
- Brachypodium genome (JGI)
- Poplar genome (JGI)
- Soybean genome (JGI, Soybase)
- MaizeGDB: As mentioned earlier, Gramene closely coordinated our initial release of the MaizeCyc pathways database with the MaizeGDB group. In addition, we have regular monthly conference calls with their group to discuss our efforts to improve the quality of the maize genome and various related projects.
- MaizeSequencing.org: At each release, Gramene works to build and release the Ensembl version of the maize genome.
- MetaCyc, BioCyc, Pathway Tools (Peter Karp): Until a full transition to the Reactome infrastructure for pathways, Gramene is still invested in the Pathway Tools from SRI Internation and works with them to keep our software current with their latest releases.
- Reactome (Lincoln Stein, Peter D'Eustachio): As described earlier, Gramene is working closely with the Reactome group to produce a Plant Reactome portal.
- Arabidopsis Reactome (Nick Provart, Henning Hermjakob)
- PlantCyc (Sue Rhee)
- SolCyc and Solanaceae Genome Network (Lukas Mueller)
- Phenote curation tool (Nomi Harris, Suzi Lewis)'
- BrachyBase (Todd Mockler)
- Sorghum Biofuel and Bioenergy Project (John Mullet)
- Maize Pathways (Andrew Hanson, Chris Henry)
- C3-C4 project (Tim Nelson, Tom Brutnell, Chris Myer, R. Bruskiewich)
- WikiPathways
- Expression data (Todd Mockler, Tim Nelson, Tom Brutnell)
- Arabidopsis 2010 (Magnus Nordborg, NSF #0723510) (Diversity): Gramene has been the recipient of genetic diversity data from *A. thaliana, A. lyrata,*
- OryzaSNP Project (Diversity)
- Panzea (Diversity)
- Scottish Crops Research Institute (SCRI) (Diversity)

- Jan Dvorak and the Wheat D genome project (NSF): Gramene has dedicated resources as part of the project for the integration of the data sets. Data sets are still in preparation and will be integrated in the next release
- NASC
- TAIR
- Oryza Genomes Project (Rod Wing, NSF PGI #0638820): Formerly known as the Oryza Mapping Alignment Project (OMAP), Gramene has dedicate resources to support the integration of the data sets, and has worked closely with the OGE project to expand our annotation of the wild rices.
- USDA ARS Grape end 2009
- NSF #0701916 PGI PI Dvorak end 2011 (wheat)
- NSF PGI PI Wilson end 2010 (maize)
- NSF PGI PI #0723510 Scanlon end 2012 (maize)
- NSF PGI PI Springer to start this year (maize)
- NSF PGI #1032105 PI McCombie end 2012 (wheat)
- EBI BBRSC Paul Kersey (travel for coordination participants)
- NSF PGI PI McCouch end 2014 (rice)
- Rice diversity (Susan McCouch, NSF #0606461, #1026555): The Rice Diversity project (http://ricediversity.org/) provides Gramene with genetic variation data from rice.
- Oryza Genome Evolution (Rod Wing, NSF #1026200)
- Rice diversity (Michael Purugganan, NSF #0701382)
- Rice diversity (Olsen, NSF #0638820)
- Genetic Architecture for Maize and Teosinte (Ed Buckler, NSF #0820619)

Letters of Support

Gramene PIs have provided letters of support to the following projects in the last year:

- Oswald Crasta (Chromatin Inc., Functional genomics approaches for improvement of sweet and lignocellulosic sorghums for bioenergy)
- Siobahn Brady (UC Davis, Characterization of genetic, molecular and stress response variation in the energy sorghum secondary cell wall biosynthesis gene regulatory network)
- Oswald Crasta (Chromatin Inc., Engineering sorghum for improved nutrient use efficiency)
- Eva Huala (TAIR/Stanford, Building a corpus of experimental results on plant gene function for translational research)
- Palitha Dharmawardhana (OSU, Poplar Interactome for Bioenergy Research)
- Z. Jeffrey Chen (U Texas, NSF/PGRP: Genomic and Functional Analysis of Circadian Rhythms and Growth Vigor in Maize)
- Julia Bailey-Serre (UC Riverside, Integrative analysis of plasticity in cell fate determination in plants)

- Michael Purugganan (NYU, Environmental Gene Regulatory Interaction Networks in Rice)
- Todd Mockler (DOE, Modulation of Phytochrome Signaling Networks for Improved Biomass Accumulation Using a Bioenergy Crop Model)
- Henning Hermjakob (EBI/BBSRC, Arabidopsis Reactome)
- Nick Provart (Univ. of Toronto, Bio-Array Resource)
- Andrew Hanson (IU, Comparative genomics-driven discovery of B vitamin pathways in maize)
- Sergei Filichkin (OSU, Mechanisms and cellular factors modulating gene expression in plants through nonsense-mediated mRNA decay)

Meetings, Courses and Teaching

Following is a list of the meetings and courses which Gramene members attended in the last year:

Meeting Name	*Location*	*Attendee*	*Presentation*
Plant Genomics European Meeting	Istanbul, Turkey	William Spooner	Poster
Biology of Genomes	Cold Spring Harbor, NY	Marcela Monaco	Poster
Grape RCN	Lake Tahoe, NV	Joshua Stein	Gramene resources for grape
ICAR	Madison, WI	Doreen Ware, Joshua Stein, Pankaj Jaiswal	Poster, Gramene workshop
IBC	Melbourne, Australia	Pankaj Jaiswal	Poster
ISMB	Vienna, Austria	William Spooner	Poster
Metabolic Pathways Annotation Jamboree on Arabidopsis	Boston, MA	Palitha Dharmawardhana, Marcela Monaco	
Gordon Plant Metabolic Engineering	Waterville Valley, NH	Doreen Ware	Talk about Kbase

ASPB	Minneapolis, MN	Ken Youens-Clark, Sunita Kumari, Palitha Dharmawardhana, Pankaj Jaiswal, Doreen Ware	Posters, Gramene/Plant Ontology workshop
Plant Genome Evolution	Amsterdam, The Netherlands	Susan McCouch	
Genome Informatics	Cold Spring Harbor, NY	Ken Youens-Clark, Jim Thomason, Joshua Stein, William Spooner, Sunita Kumari, Marcela Monaco	Posters
Plant Genomes and Biotechnology	Cold Spring Harbor, NY	Palitha Dharmawardhana, Marcela Monaco	Posters
9th International Symposium on Rice Functional Genomics	Taipei, Taiwan	Joshua Stein, Charles Chen	Poster
PAG XX	San Diego, CA	Terry Casstevens, Ken Youens-Clark, Palitha Dharmawardhana, Pankaj Jaiswal, Doreen Ware, Marcela Monaco, Joshua Stein	Posters (P0856, P0858, P0951, P0952, P0993), Gramene workshop, Plant Database outreach booth
Maize Genetics	Portland, OR	Marcela Monaco	Poster
New Mexico BioInformatics, Science, and Technology Symposium	Santa Fe, NM	Ken Youens-Clark	
Biology of Genomes	Cold Spring Harbor, NY	Marcela Monaco	Poster
77th Cold Spring Harbor Symposium on Quantitative Biology: The Biology of Plants	Cold Spring Harbor, NY	Marcela Monaco	Poster

Publications

Following is a list of publications to which Gramene has contributed in the last year:

- Gramene database: a hub for comparative plant genomics (Jaiswal): Andy Pereira (ed.), Plant Reverse Genetics: Methods and Protocols, Methods in Molecular Biology, vol. 678, DOI 10.1007/978-1-60761-682-5_18 PDF pubmed:20931385
- Filichkin SA, Breton G, Priest HD, Dharmawardhana P, Jaiswal P, et al. 2011 Global Profiling of Rice and Poplar Transcriptomes Highlights Key Conserved Circadian-Controlled Pathways and cis-Regulatory Modules. PLoS ONE 6(6): e16907. doi:10.1371/journal.pone.0016907
- QlicRice: a web interface for abiotic stress responsive QTL and loci interaction channels in rice Smita S, Lenka SK, Katiyar A, Jaiswal P, Preece J, Bansal KC. Database (Oxford). 2011 Sep 30;2011:bar037. Print 2011.
- BioMart Central Portal: An Open Database Network for the Biological Community. Kasprzyk A, Arnaiz O Baran J, Blake A, Baldock R, Chelala C, Croft D, Cros A, Cutts R, Forbes S, Fujiwasa T Goodstein DM Gundem G, Haggarty B, Haider S, Hall M, Harris T, Haw R, Hubbard S, Hsu J, Iyer V, Jones P, Kinsella R, Katayama T, Kong L, Lawson D, Liang Y, Lopez-Bigas N, Lush M, Mason J, Moreews F, Ndegwa N, Oakley D, Perez-Llamas C, Primig M, Rivkin E, Shepherd R, Simon R, Skarnes W, Smedley D, Sperling L, Spooner W, Stevenson P, Stone K, Teague J, Wang J, Wang J, Whitty B, Wong-Erasmus M, Youens-Clark K, Yung C, Zhang J, Gadaleta E. Database (Oxford). 2011 Sep 18;2011:bar041
- GrameneMart: The BioMart Data Portal for the Gramene Project. Spooner W, Youens-Clark K, Staines D, Ware D. Database. (Accepted, print 2012).

Submitted

- *PICARA*, a probabilistic inference on functional implication of a priori candidates and its application on genome-wide associations of flowering time variation in maize. Chen C, DeClerck G, Tian F, Stein J, McCouch S, Buckler E. (submitted, *Genome Research*).
- Maize Metabolic Network Construction and Transcriptome Analysis. Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, Amarasinghe V, Thomason J, Harper L, Gardiner J, Cannon EKS, Lawrence CJ, Ware D, and Jaiswal P (submitted, *Plant Physiology*)

Manuscripts in preparation

- RiceCyc: Metabolic Networks in Rice

Website

Data and code sharing

Gramene allows free and open access to all data and code via the following:

- **Code repository:** The public Subversion repository is available at http://svn.warelab.org/gramene

- **FTP:** Point releases of software and data are available for current and several archived versions at ftp://ftp.gramene.org
- **Public MySQL server:** All databases from the current release are available on a public MySQL server that can be directly queried by our users at the host gramenedb.gramene.org

Additional Websites

Gramene maintains the following websites:

* http://www.gramene.org/ (public website)
* http://dev.gramene.org/ (development website open to the public)
* http://news.gramene.org/ (news blog)
* http://docs.gramene.org/ (documentation Wiki)
* http://www.plantontology.org/
* http://www.gramene.org/gramenedas/
* http://www.maizegenetics.net/ (TASSEL software)
* http://gmod.org/wiki/CMap
* http://search.cpan.org/dist/Text-RecordParser/
* http://search.cpan.org/dist/SQL-Translator/
* http://search.cpan.org/dist/Bio-GenBankParser/

Social networks

Gramene maintains a news web log ("**blog**") at news.gramene.org. During last year, we posted 31 entries announcing new developments on the Gramene website, publications, gave various presentations and tutorials at several meetings and conferences, as well as news developments and job openings of interest to our users. We are also exploring other social media outlets such as **Facebook**, where we have recently initiated a Gramene page, and on **Twitter**, where our handle is "GrameneDatabase."

Other outreach

Gramene prepared posters and brochures for the following public outreach events:

1) Global collaborations for USDA members (*e.g.*, PI Ware) at G8 Summit event on food security. Highlighted collaborations for agriculture and science together for G8 representatives, illustrating the importance of Gramene for crop and plant researchers.

2) OSEC-organized event. Developed materials designed to inform the lay person about resources like Gramene. *E.g.,* What is Gramene? Why is Gramene important?

3) Howard Hughes Medical Institute (HHMI)-organized event. HHMI is interested in understanding how basic research is leveraged in crop plants by ARS and collaborators.

Software development

Integration Database

Over the 11 years of Gramene's history, new data sets that were introduced over the years (QTL, genes, proteins, pathways, diversity) were developed in separate databases and were integrated into user views via the use of common accession strings (*e.g.*, ontology terms). Moving forward, we would like to integrate these separate structures into a single database. This presents an opportunity to streamline our models and rewrite our software layers to take advantage of advancements in software tools and best practices refined in the last few years. We will also move to a heterogenous data storage model that incorporates relational databases such as MySQL where appropriate, but also incorporate key-value databases like CouchDB or Cassandra for non-relational data, and indexed flat-files like tabix/samtools/FastBit for alignments, all the while looking for ways to introduce modern caching engines such as memcached to reduce query times for our users.

Ensembl

Gramene stays current with the most recent versions of Ensembl at each of our major releases as well as during interim releases when an upgrade is necessary to address problems. We work closely with Ensembl developers to fix bugs, contribute code, and suggest software enhancements. In the last year, we have updated Ensembl to versions 62, 64, and 65 with our one interim and two major releases.

BioMart

The Ensembl group has built our Mart databases based on the Ensembl supported modules.

GDPDM

Gramene's diversity data is stored in the Genomic Diversity and Phenotype Data Model (GDPDM) database schema (http://www.maizegenetics.net/gdpdm). The last year saw one point release (4.3) to address genotyping values.

TASSEL (also refer to specific Aim 3)

Tassel 4.0 was released November 2011. Improvements include the following:

- Additional Web-launch TASSEL files created to allow diversity data sets to be loaded automatically
- Consistent Implementation of all data types (SNP, Haplotypes, etc.)
- Bitwise Data Structures for Speed (40 - 200 fold increase) and Memory Efficiency
- Many improvements to software architecture and design
- New QQ and Manhattan Plots
- 70% speed improvement to Cladogram Function
- Improved LD results display
- Many improvements to Progress Monitoring
- Much improved Taxa and Site Name Filtering
- Tassel on iPlant Discovery Environment and Atmosphere

- New Genotype Summary Function
- More User Friendly Alignment Viewer
- Improved Error Messages
- GLM and MLM:
    - GLM interface simplified
    - Compression and faster P3D implemented for MLM resulting in reduced runtime
    - Matrix Algebra library wrapper written to make switching to newer, faster libraries easier
    - EJML Matrix Algebra library interface implemented
- Pipeline (Command Line Interface)
    - Automates complex loading/analysis pipelines
    - Does not need Java coding to create
    - Has simultaneously executing pipeline segments
    - Works from web site launch, command line, and GUI
    - Self-Describing Plugins
    - Integrated GBS Pipeline with existing Tassel Pipeline
    - Added more parameters to command line interface (i.e. LD, Site Filtering, Numerical Transform)
    - Can Define Pipelines in XML

CMap

CMap was initially created by Gramene in the early years of the project and has remained stable at version 1.01 since its release in 2008. This version has been downloaded over 1000 times from Sourceforge, where the software is hosted, and is used extensively in the community by groups studying plants, insects, bacteria and animals. CMap is freely available from the Generic Model Organism Database (GMOD) project.

Reactome

See Specific Aim 3

Management

Project meetings

The project has moved from weekly to monthly “all-hands” meetings and now uses the weeks in between to follow up on specific aims and collaborators meetings. All-hands meetings are used to present progress toward tasks, organize build and release schedules, coordinate with collaborators, and present results of work. The PIs meet every other month to discuss project progress, and address risks associated with the current project. The project did not meet this year for a retreat, nor did we convene the Scientific Advisory Board.