Gramene: A comparative genomics and pathways resource for plants
http://www.gramene.org

Marcela Karey Tello-Ruiz
telloruiz@cshl.edu
Doreen Ware
ware@cshl.edu
Cold Spring Harbor Laboratory

Recent activities and newly developed tools and/or resources

Gramene provides open access to plant reference genome assemblies, genetic variation, gene expression, gene functional annotations including curated and projected metabolic and regulatory pathways, and phylogenetic annotations including within species paralogs, lineage specific genes, and ortholog/paralog assignments between species. In collaboration with EnsemblGenomes, Gramene hosts 93 plant reference genomes (almost 4 million genes in total) including three Arabidopsis species (nearly 100,000 genes): *A. thaliana, A. lyrata,* and *A. halleri*. For each reference genome sequence, we provide structural and functional gene annotations including ontology associations and protein domain assignment, genetic and structural variants, phylogenetic trees with orthologous and paralogous gene classification, whole-genome alignments, and synteny maps.

The current gene tree build includes assignment of 122,947 gene families comprising close to 2.9 million genes, supporting homolog and ortholog assignments to the three hosted Arabidopsis species. Access to the data is provided through a graphical user and application programming interface. From the homology tab on the search interface, the gene of interest is anchored by its gene tree assignment, and contains functional and visually informative structural information (e.g., color-coded protein domains and tick marks indicating splice junctions) and interactive features (e.g., ability to select a specific GO term or InterPro domain) to support access to orthologs, paralogs, and homologs that share functional annotations. The search interface and homology view allows custom pruning of the gene trees to selected species of interest, and visualizing sequence conservation to the amino acid level. *A. thaliana* serves as an anchor species within Gramene. *A. thaliana* homologs are displayed as part of the query results within the Gramene search for all species. In addition, *A. thaliana* is used as the dicot model for pairwise DNA level whole-genome alignments collection. Within the past year, the alignments subset for *A. thaliana* grew from 66 to 84, including alignments between *A. thaliana* and each of *A. lyrata* and A. halleri.

In addition, we host alignments between *A. lyrata* and each of Medicago truncatula, *Oryza sativa* (Japonica rice), *Theobroma cacao* (cacao), and *Vitis vinifera* (grapevine); and *A. halleri* to Japonica rice, cacao, and grapevine. Our synteny collection includes synteny maps for *A. thaliana* and each the following four species: *A. lyrata, Brassica rapa*, Japonica rice, and grape; and for *A. lyrata* and grapevine. These maps are based on the gene trees and positional information within the reference assembly. We continue to host 12.9 million Arabidopsis SNPs from the 1001 Arabidopsis Genomes Project. Variants are provided in the context of gene annotation, gene regulation, and protein domain structure, along with predicted functional consequences (e.g. missense variant), and genotypes.

In our continued collaboration with the Expression Atlas project (EMBL-EBI), we provide baseline expression data for 24 species, including *A. thaliana* and *A. lyrata* through both, our Gramene Ensembl genome browser and Plant Reactome pathways interfaces. In addition, we provide direct links to differential gene expression data on the EMBL-EBI Expression Atlas website for a partially

overlapping set of 24 species, including *A. thaliana* and *A. lyrata* More recently, EBI Atlas, developed the capacity to host single-cell gene expression data; currently five data sets from four studies are available (Ryu *et al*, 2019; Jean-Baptiste *et al*, 2019; Shulse *et al*, 2019; Turco *et al*, 2019).

In collaboration with Reactome, Gramene hosts 320 metabolic and regulatory pathways curated in Japonica rice and inferred in 106 additional plant species (including the three Arabidopsis species) based on orthology. Reactome pathways are checked and peer-reviewed prior to publication to ensure factual accuracy and compliance with the data model, and a system of evidence tracking ensures that all assertions (which use community standard controlled vocabulary ontologies) are supported by primary literature. Gramene's integrated search capabilities, and interactive views facilitate visualizing gene features, gene neighborhoods, phylogenetic trees, gene expression profiles, and pathways. The views also assist cross-referencing to other bioinformatics resources, including AraPort, TAIR, and NASC for Arabidopsis. Gramene provides tools to support integration of user data sets in context to the reference data. These tools include a sequence assembly converter (which allows the conversion of genomic coordinates between the TAIR9 and TAIR10 genome assemblies), a genetic variant effect predictor, an advanced BioMart-based query interface, data analysis and visualization of OMICS data, multi-species pathway comparisons, and a BLAST/BLAT sequence aligner.

Together these reference comparative genome data and tools enable powerful cross-species comparisons among plants and reference eukaryotic species. Gramene data sets that include Arabidopsis species:

- Structural and functional annotations for 2.2 million gene models in 93 plant reference genomes including three Arabidopsis model species, *A. thaliana, A. lyrata,* and A. halleri, cereal, vegetable, and fruit crops (e.g., Brassicas, Fabaceas, Solanaceas), basal plants and algae.

- 122,947 phylogenetic tree families (built with 93 plant and 5 non-plant species), 382 whole-genome alignments (86 with Arabidopsis species), and 84 synteny maps (5 with Arabidopsis sp.).

- Almost 238 million genetic and structural variants for 14 plant species, including 12.9 million Arabidopsis SNPs from the 1001 Arabidopsis Genomes Project. The Arabidopsis SNP set includes genotypes for over 1,000 accessions, and was combined with phenotypic data (107 phenotypes associated with 95 inbred lines) from the GWAS study by Atwell *et al* (2010).

- Experimental baseline and differential expression data for 972 experiments in 28 plant species, including *A. thaliana* and *A. lyrata*

- v320 reference metabolic and regulatory pathways curated in rice and inferred in 106 additional plant species (including the three Arabidopsis species in Gramene).

- Integrated search capabilities and interactive views to query and visualize gene features, gene neighborhoods, phylogenetic trees, gene expression profiles, pathways, and cross-references to other bioinformatics resources (e.g., AraPort, TAIR, and NASC).

- Analysis tools to support comparative analyses of our data as well as user-provided data (e.g., BLAST/BLAT sequence aligner, sequence assembly converter for TAIR9/TAIR10 genomic coordinates, genetic variant effect predictor, BioMart, Reactome pathways analysis/visualization of OMICS data and multi-species pathway comparisons).

Gramene is committed to open access and reproducible science based on the FAIR (Fair, Accessible, Interoperable and Reusable) data principles. We are a phylogenomic resource, built upon best-of-class open source software, Ensembl, Reactome, and Expression Atlas infrastructure platforms. Gramene has developed a powerful and flexible document-based architecture that enables advanced searching via a web-service accessible by a variety of programming languages; each platform supporting web-based and programmatic access through application programming interfaces (APIs). Extensive use of ontologies, database cross-references, common data formats, metadata, community engagement and open-source software promotes interoperability within the ecosystem of informatics data and services.

Gramene's genome portal utilizes the Ensembl infrastructure and is developed in collaboration with the Ensembl Genomes project (EMBL-EBI); the pathway portal, Plant Reactome (http://plantreactome.gramene.org) utilizes the Reactome infrastructure, and is developed in collaboration with OCIR; the baseline expression data from both, our genomes and pathway browsers, is a collaboration with the Expression Atlas project (EMBL-EBI). Integration across these platforms in Gramene was supported by NSF grant IOS-1127112. More recent work for Gramene has focused on the development of species Pan-genome sites providing access to reference assemblies for multiple accessions of the same species, which facilitates the identification and characterization of common and variable genome regions. Importantly, all our species-specific pan-sites include *Arabidopsis thaliana* as a reference.

Currently the project is solely supported by the USDA-ARS (1907-21000-030-00D).

Ryu KH *et al*. (2019). Single-Cell RNA Sequencing Resolves Molecular Relationships Among Individual Plant Cells. Plant Physiol. 179(4):1444-1456. doi: 10.1104/pp.18.01482.

Jean-Baptiste K, *et al*. (2019) Dynamics of Gene Expression in Single Root Cells of *Arabidopsis thaliana*. Plant Cell. 31(5):993-1011. doi: 10.1105/tpc.18.00785.

Shulse CN, *et al*. (2019) High-Throughput Single-Cell Transcriptome Profiling of Plant Cell Types. Cell Rep. 27(7):2241-2247.e4. doi: 10.1016/j.celrep.2019.04.054.

Turco GM, *et al*. (2019) Molecular Mechanisms Driving Switch Behavior in Xylem Cell Differentiation. Cell Rep. 28(2):342-351.e4. doi: 10.1016/j.celrep.2019.06.041.

Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Tello-Ruiz MK, Naithani S, Gupta P, Olson A, Wei S, Preece J, Jiao Y, Wang B, Chougule K, Garg P, Elser J, Kumari S, Kumar V, Contreras-Moreira B, Naamati G, George N, Cook J, Bolser D, D'Eustachio P, Stein LD, Gupta A, Xu W, Regala J, Papatheodorou I, Kersey PJ, Flicek P, Taylor C, Jaiswal P, Ware D. Nucleic Acids Res. 2021 Jan 8;49(D1):D1452-D1463.
doi: 10.1093/nar/gkaa979. PMID: 33170273

Planned future activities

With future support, we will continue to maintain and build the Gramene resource aiming to have a minimum of two releases per year: 1) update and expand our reference data collection of plant genomes, genetic variation, gene expression, and standardized comparative annotations, and orthology-based projections; 2) curate single-cell expression data and curate metabolic pathways to Expression Atlas and enrich our Plant Reactome pathways; 3) develop pan-genome resources for maize, rice, sorghum and grapevine; and 4) transform the community through communication and

training opportunities.

Please provide a paragraph describing the general impact of the COVID19 pandemic on your activities

Prior to 2020, we held three genome annotation jamborees in conjunction with a major scientific meeting to train plant researchers and members of the plant community—particularly faculty at Primarily Undergraduate Institutions (PUIs)—to curate gene models and develop curriculum modules for their students to partake in Course-based Undergraduate Research Experiences (CUREs). Because of COVID-19 travel restrictions, in 2020, we had to rethink how to host this type of events that would otherwise have been in-person.

In record time, we set up our first virtual annotation jamboree for March 10-12 with the participation of more than two-dozen participants from 13 higher education institutions and the USDA ARS. Participants learned about gene and genome structure, natural variation, gene expression, sequencing technologies, and bioinformatics to eventually assess the quality of computationally predicted gene models using gene annotation tools from their homes. While COVID has presented challenges for traditional outreach, it has also prompted us to rethink opportunities for virtually learning, and created new opportunities to support higher education and remote work.