Phytozome Comparative Plant Genomics Portal

Authors: David Goodstein[1], Sajeev Batra[1], Joseph Carlson[1], Richard Hayes[1], Jeremy Phillips[1], Shenqiang Shu[1], Jeremy Schmutz[1,2], Daniel Rokhsar[1]

[1] U.S. Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA
[2]. Hudson-Alpha Institute for Biotechnology – Huntsville, AL

** To whom correspondence may be addressed. David Goodstein, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA. dmgoodstein@lbl.gov*

September 9, 2014

ACKNOWLEDGMENTS:

Work by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Phytozome Comparative Plant Genomics Portal

David M. Goodstein

Sajeev Batra[1], Joseph Carlson[1], Richard D. Hayes[1], Jeremy Phillips[1], Shengqiang Shu[1], Jeremy Schmutz[1,2], Daniel S. Rokhsar[1]

[1]U.S. Dept. of Energy Joint Genome Institute [2]HudsonAlpha Institute for Biotechnology

Plant Science at the JGI

The Dept. of Energy Joint Genome Institute is a genomics user facility supporting DOE mission science in the areas of Bioenergy, Carbon Cycling, and Biogeochemistry. The Plant Program at the JGI applies genomic, analytical, computational and informatics platforms and methods to:

- Understand and accelerate the improvement ("domestication") of bioenergy crops
- Characterize and moderate plant response to climate change
- Use comparative genomics to identify constrained elements and infer gene function
- Build high quality genomic resource platforms of JGI Plant Flagship genomes for functional and experimental work
- Expand functional genomic resources for Plant Flagship genomes

The current JGI Plant Flagship genomes are:

JGI commitment to Flagship Genomes:

1. Multiple rounds of genome and annotation improvement
2. Development of related resources – diversity, functional assays, comparative tools – in collaboration with respective plant communities
3. Long-term accessibility of data, analyses and tools

These genomes have been through multiple rounds of assembly and annotation improvement, and are currently the focus of extensive transcriptomics efforts to develop a standardized, reproducible, and updateable comparative functional resource (Gene Atlas project), as well as deep and broad resequencing efforts to capture and elucidate the extent and consequences of natural variation.

The Phytozome Group – Annotation, Analysis, Access

The Phytozome group consists of biologists, computational scientists, and software developers with extensive experience in human, plant and eukaryotic model organism genomics and associated data systems. Our group is responsible for developing, applying and maintaining accurate, reproducible and scalable methods for:

1. **Genome Annotation** – transcript assembly, multi-method gene calling, annotation forward mapping, proteome annotation.
2. **Gene Family construction and analysis** – family construction, orthology and paralogy analyses.
3. **Reseq and RNA-Seq analysis** – standardization of collaborator-contributed analyses, co-expression and comparative expression analyses.
4. **Uniform programmatic and interactive access**, **analysis** and **visualization** tools across multiple data types and data sets.
5. **Providing a "home" for JGI-generated and related community plant genomic data.**

28 of the 48 plant genomes available in the current Phytozome v10 were annotated by the group, the most recent being the genome of the shrub willow, *Salix purpurea*, a poplar comparator and a bioenergy feedstock candidate in its own right.

Contact: phytozome@jgi-psf.org

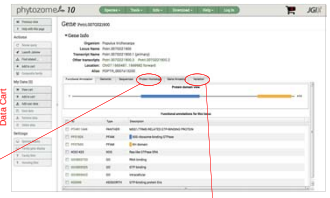
Phytozome

First released in 2006, Phytozome is accessed by ~120k users making approximately ~400k visits in a typical year. Our largest users communities are in the United States, China, and Japan. Brazil, Germany, France, Canada, India, the U.K., and South Korea round out the top ten communities.

Initial access to genes and gene families is provided via keyword, BLAST or BLAT search from the home page, or by browsing genomic features in JBrowse.

Gene and Genome-anchored views

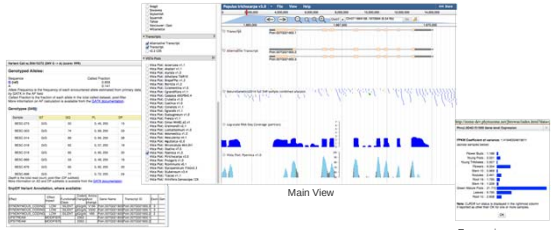
Every gene has an associated **Gene Page**, with location, alternative splicing, homology, proteome classification, and diversity information. Functional annotations, SNPs, homologs, and gene-related sequences can be added to the Phytozome cart and uploaded to InterMine or Jalview for analysis, and saved indefinitely in user accounts.



Peptide homologs amongst all other Phytozome proteins are calculated via dual affine Smith-Waterman. Alignment gaps and inserts are shown graphically, while detailed pair format alignments are also available. Browser and Gene page links to each homolog are provided.

If variant data is present, SNPs from the gene locus and 5kb upstream and downstream are shown. Variants are color-coded by predicted effect, with visual summaries of population allele and genotype frequencies.

JBROWSE Genome View – Fully populated JBrowse instances for all 48 Phytozome genomes provide gene structure detail, genome-aligned peptide homology, SNP tracks, RNA-Seq read alignments, and whole-genome alignments via VISTA. SNP feature pop-ups have detailed genotype information, while gene and transcript level expression pop-ups show differences across libraries.



SNP pop-up

Main View

Expression pop-up (*Phaseolus*)

Gene Family-centric Resources

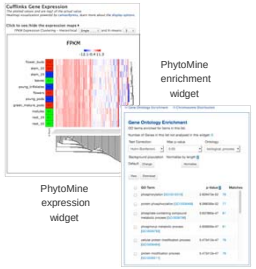
Phytozome gene families are hierarchically constructed via single-linkage-clustering, with thresholds based on intra- and inter-species similarity and coverage distributions. Family pages show membership, phylogentic distribution, the near syntenic environment of each member gene, the local domain and exonic structure, and the results of various automated classification and naming pipelines. Users can traverse each family's history to examine its imputed ancestral composition, and can search for related families via family consensus sequences or shared annotation.

Data Access: Bulk and Query-based

Phytozome is now integrated into JGI's Genome Portal, for fast downloads of pre-compiled bulk datasets. PhytoMine, our InterMine implementation, can take lists built in Phytozome or directly in PhytoMine and perform Enrichment analysis, expression visualization, and build custom data sets with precisely the fields and formats desired.

PhytoMine lists can be saved, augmented, reduced, intersected with other lists, converted between various types, and exchanged with Phytozome. Gene, transcript, protein and SNP lists are supported. PhytoMine also supports a web service and Perl, Ruby and Java APIs for programmatic access to Phytozome data.

PhytoMine home



PhytoMine enrichment widget

PhytoMine expression widget

Coming up in v10.1

v10.1, expected at the end of September, will include

- Access to ortholog and paralog calls in both Phytozome and Phytomine.
- Improved SNP viewers
- Richer PhytoMine templates to support common custom data requests
- Help pages

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.