GreenPhylDB: Phylogenomic resources for comparative and functional genomics in plants

Rouard M[1], Guignon V[1], Walde C[1], Droc G[2], Zmasek C[3], Dufayard JF[2] and Conte M.G[1,4]

[1] Bioversity International - Commodities for Livelihoods Programme Parc Scientifique Agropolis II, 34397Montpellier - Cedex 5, France
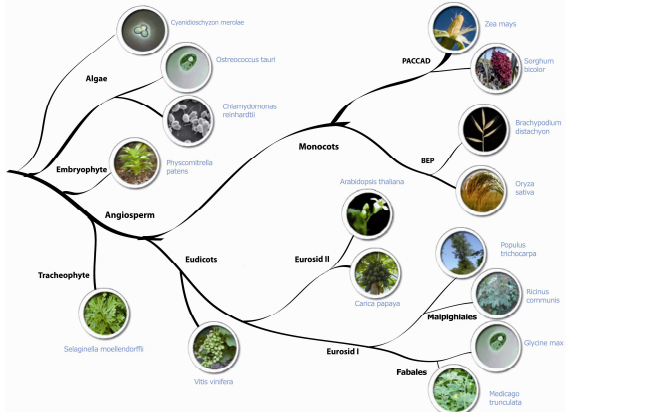[2] CIRAD, Department BIOS, UMR DAP - TA40/03, 34398 Montpellier, France
[3] Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA.
[4] SYNGENTA Seeds SAS, 12 Chemin de l'Hobit, 31790 St Sauveur, France

GreenPhylDB is a web resource designed for comparative and functional genomics in plants. Here, we present a pre-release of the version 2, which contains a catalogue of gene families for 16 full genomes, covering a broad taxonomy of green plants. Results of our automatic clustering are first manually annotated and then analyzed by a phylogeny approach to predict orthologs. This ortholog prediction is particularly useful for functional genomics and candidate gene identification of genes affecting agronomic traits of interests.

Plantae genomes

Figure 1: 15 full plant genomes including economically and socially important crops like rice, sorghum and maize, (and 1 red algea as an outgroup), were clustered in GreenPhylDB. By integrating genomes based on a broad taxonomic sampling, we defined a consistent and extensive set of homeomorphic plant gene families. This is useful for accurate orthologs prediction through phylogenetics that helps to unravel some evolutionary aspects of the green plant.

Figure 2: Approximately 8250 gene families can be explored using keywords, various identifiers such as locus tags, InterPro or KEGG. A gene family ontology browser was also developed to display list of genes families associated to Gene Ontology terms. Lists of plant-specific or lineage-specific gene families are also available.

From orthology to function

Orthology Confidence score in %

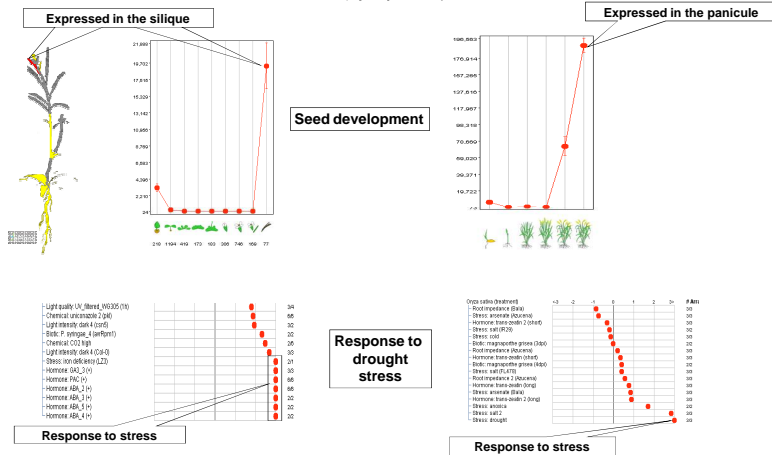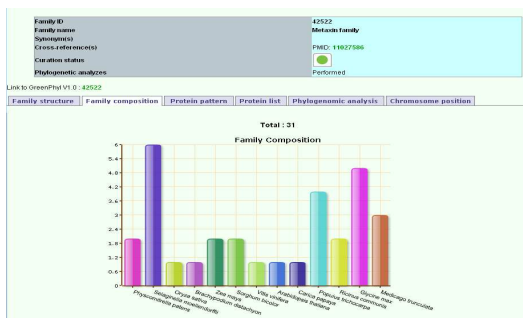2 genes of *A. thaliana*   1 gene of *O. sativa*

Figure 5: An example of *O. sativa* sequence belonging to the Tonoplast intrinsic protein (TIP) subfamily (fid:36570) that has 2 predicted orthologs in *A. thaliana* with a score of 100%. GreenPhylDB proposes for instance cross-references to the public version of *Genevestigator* (https://www.genevestigator.com), an expression database allowing studies of gene-regulation in order to study the gene level of expression in different tissues and under different conditions. In this example, we can see that these genes are over-expressed in the panicle (rice) and in the silique (*arabidopsis*) and under drought conditions. For Arabidopsis, we provide also cross-references to a 3D viewer tool called ePlant (Nicholas Provart : http://bar.utoronto.ca/eplant/) containing also gene expression data.

Figure 3: Example of family composition of the Metaxin gene family (fid: 42522). The gene members (31 sequences) are present in the land plants and have no homolog in the 3 *algae* genomes comprised in the database.
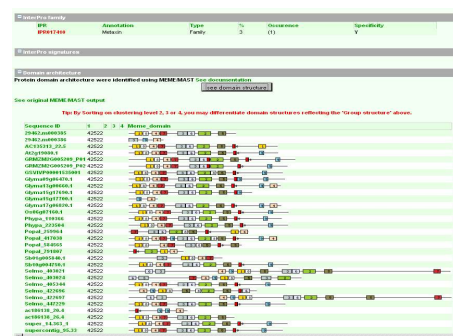

Figure 4: Protein domains
Protein domain architecture of the gene families were searched using InterProscan (http://www.ebi.ac.uk/Tools/InterProScan) and the Meme suite (http://meme.sdsc.edu). Interpro signatures specific to gene families were calculated.
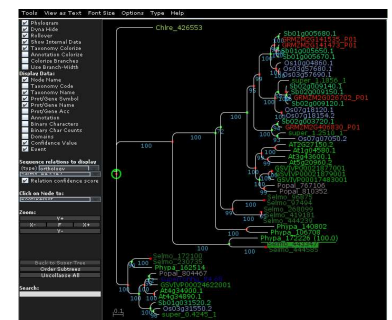

Figure 5: Phylogenetic analyses
Annotated gene families are analyses through a phylogenetic-based pipeline that allow orthologous/paralogous predictions and to highlight gene duplications that are frequents in higher plants. Gene trees are available online via the archaeopteryx applet (http://www.phylosoft.org/archaeopteryx/)

References: (http://greenphyl.cirad.fr)
- Rouard M, Guignon V, Laporte MA, Aluome C, Walde C, Zmasek CM, Périn C and Conte MG (2010) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Research, 1–8, doi:10.1093/nar/gkq811
- Conte MG, Gaillard S, Droc G, Périn C (2008). Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. BMC Genomics, 2008, 183

Contact: Mathieu Rouard (m.rouard@cgiar.org)