GREENPHYL: An optimised phylogenomic pipeline for Ortholog prediction between two model plant species *Arabidopsis thaliana* and *Oryza sativa*

Matthieu CONTE, Sylvain GAILLARD, Christophe PERIN*. Corresponding author: **perin@cirad.fr**

CIRAD UMR PIA, Avenue Agropolis, 34398 Montpellier Cedex 5, France.

➢ Overview

The increasing amount of sequence data provided by full or partial genome sequencing projects urgently need a way to transfer information from model species to new sequenced ones. Orthologous and paralogous genes identification is now a major objective for gene function prediction as orthologous sequences are more likely to share the same function than paralogous sequences.

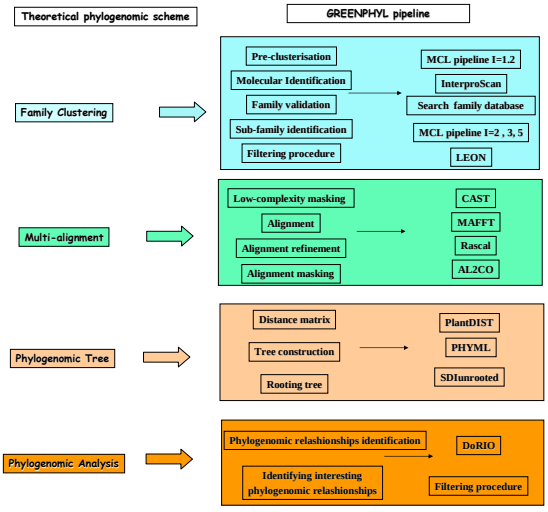
➢ Objectives

Obtain a catalogue of orthologous genes between *Oryza sativa* and *Arabidopsis thaliana*

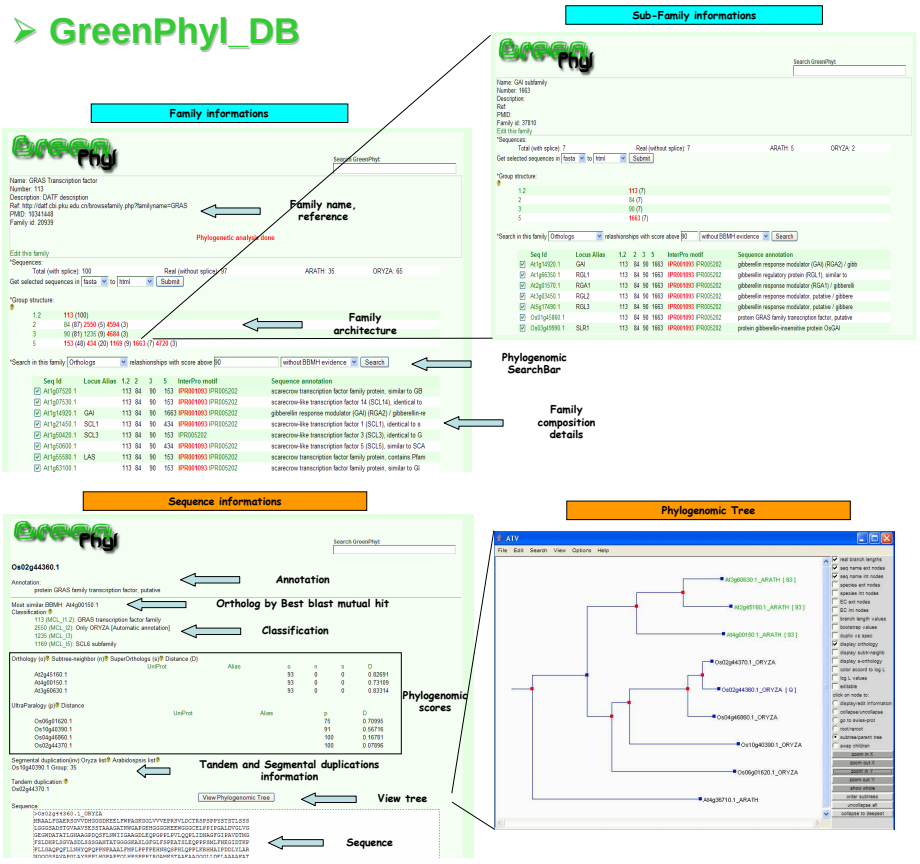
➢ Results

We developed GREENPHYL, an optimized phylogenomic pipeline combining genome and phylogenetic analyses to reconstruct the evolutionary history of genes for each family and identify orthologs and paralogs. Moreover, contrasting with most other phylogenomic analysis pipelines, GREENPHYL includes an automatic analysis of the generated tree. GREENPHYL is actually applied to TIGR *Arabidopsis thaliana* (Version 5) and *Oryza sativa* (Version 3) (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/). GREENPHYL output data are automatically loaded in a dedicaced database (GreenPhyl_DB).

➢ Pipeline



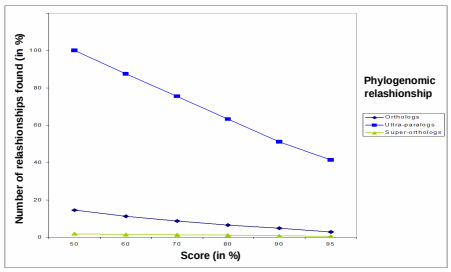
➢ GreenPhyl_DB



➢ GreenPhyl_DB statistics

Clustering: Total number of clusters: 21,038. Manually curated : 3,612.
Validated by TAIR, IPR , DATF and DRTF families (Manual clustering)

Phylogenomic:
Currently 64 transcription factor families have been analysed using the GreenPhyl pipeline



➢ Greenphyl Prediction validation

We evaluated GREENPHYL performances against a set of published genes already functionally characterized in the two plant model *Arabidopsis thaliana* and *Oryza sativa*. Here we present a short illustration of sensitivity and selectivity of Greenphyl phylogenomic analysis using the GRAS transcription factor family.

The GAI sub-family is involved in gibberellin signal transduction pathway and belongs to the GRAS transcription factor family. AtRGA2 and AtRGA, functionally redundant in *Arabidopsis,* shared the same function with *Oryza* gene Os03g49990.1 (SLR1) [PMID:11340177].
AtRGA2 and AtRGA are linked by a significant paralogous scores (100% UltraParalogs. Not show).

Moreover, AtRGL1, 2 and 3 (At1g66350.1, At5g17490.1 and At3g03450.1) are linked with AtRGA2 and AtRGA with SubtreeNeighbor scores. These three genes belong effectively to the same DELLA family, they are all involved in gibberellin transduction, but share a distinct function compared to AtRGA2 and AtRGA (neo-functionalization) [PMID:15173565]. They are also ortholog to SLR1.