

# **GOST**

**(GreenPhyl Orthologs Search Tool)**

**A phylogenomic tool for plant comparative genomics**



**Mathieu CONTE**



**Mathieu ROUARD**

SP4 project: Application and development of web service technology

# PLAN

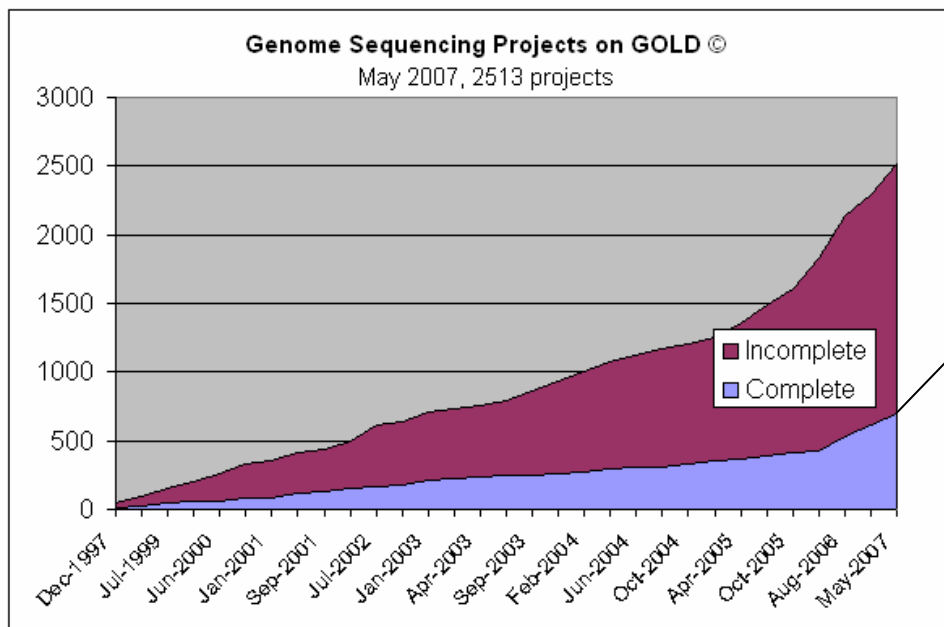
1. **Sequencing projects and comparative genomics**
2. **GreenPhylDB: A phylogenomic platform for plant comparative genomics**
3. **GOST: GreenPhyl Orthologs Search Tool**

# Sequencing Projects

Genome projects are generating vast amounts of sequences

The objective is now to determine the function of predicted genes

Computational methods are needed to help annotation transfer and functional prediction



More than 500  
genomes fully  
sequenced

# Comparative genomics

Predict gene function for one species  
using information available from other species

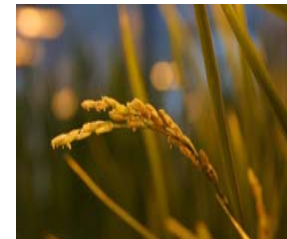
Gene with unknown function



Annotation transfer

Model species

Homologous genes

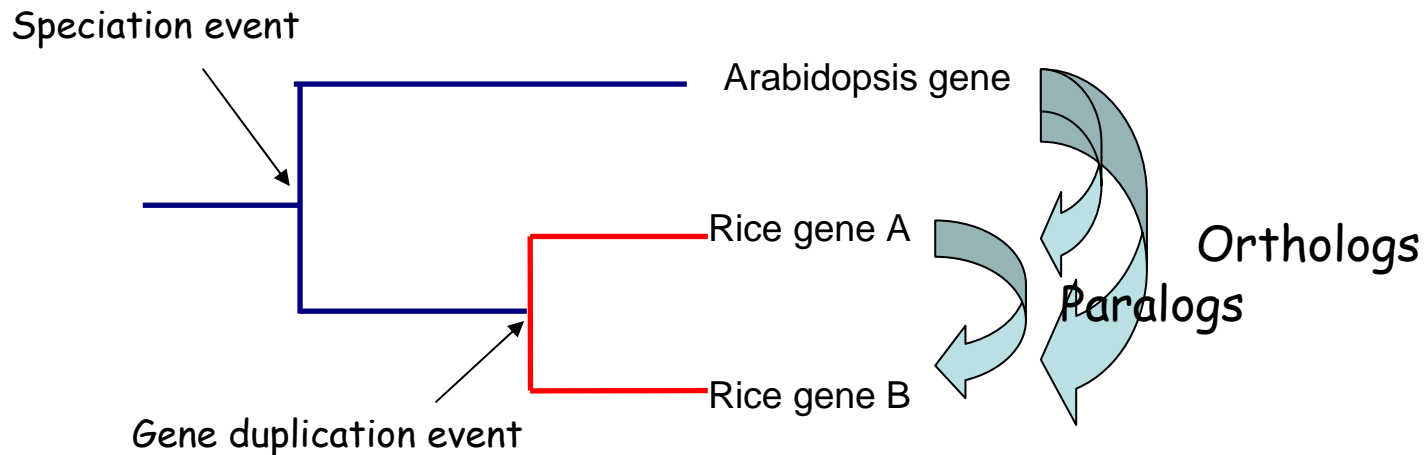


Gene with function X

# Homologous genes

## orthologous - paralogous

- **Orthologous genes** are homologous genes that are descended from the last common ancestor through speciation and most probably encode proteins with a similar function in different species



- **Paralogous genes** are referred as homologous genes that evolved through duplications and may encode proteins with more divergent functions

# How to predict homologous genes?

## Similarity vs. Homology

Similarity and Homology are **not** the same thing, even if homology is inferred from certain types of similarity

### Similar

having likeness or resemblance (an observation)

### Homolog

genetically connected (an historical fact: common ancestor)

# Function prediction by similarity?

Popular similarity methods: BLASTp, BBMH/RBH...

## ADVANTAGES:

- Easy to use
- Fast
- Directly on full genomes

## DRAWBACKS:

- How to fix E-value threshold for annotation transfer?

False positive/negative rate.

Two sequences can present some similarity without any evolutionary relationships

Real ortholog have some time low similarity score

- Cannot identify duplication events:

Tricky to predict one-to-many or many-to-many relationships  
(inparanoid, OrthoMCL, KOG)

# Function prediction by phylogeny?

## ADVANTAGES:

- Efficient for detection of duplications and speciations (paralogs and orthologs)
- Efficient to detect complete relationships (1/n, n/n) if you use complete family

## DRAWBACKS:

- Time consuming calculation
- Gene family clustering required

## Current methods:

- RIO and Orthostrapper :only on 1900 plant gene families (Pfam)
- GOST (using GreenPhylDB family : 6420 plant gene families)

# **GOST**

**(GreenPhyl Orthologs Search Tool)**

## **Objectives**

- **Identify by phylogeny methods orthologous and paralogous genes for any plant gene (GCP crops)**
- **Work on a larger set of the plant gene families (GreenPhylDB)**
- **Develop a tool as fast as similarity search (a Blastp) by using pre-computed phylogeny from GreenPhylDB data sources (GCP SP4 output)**

# PLAN

1. Sequencing projects and comparative genomics
2. **GreenPhylDB\***: A phylogenomic platform for plant comparative genomics
3. GOST: GreenPhyl Orthologs Search Tool

\* "GreenPhylDB: A database for plant comparative genomics" (Submitted)

# GreenPhylDB

A phylogenomic platform for plant comparative genomics



Developed on two plant model species



- *Oryza sativa* and *Arabidopsis thaliana* model plants of monocotyledon and dicotyledon
- Full genome available
- Gene annotation quality (TAIR release 7, TIGR release 5)
- Most of functional evidence
- Full sequenced genome of other plants exists but annotation still in progress.
- In the future, GreenPhylDB will integrate other plant genomes...

# GreenPhylDB workflow



**30500 Arabidopsis genes**  
**TAIR**



**50200 rice genes**  
**TIGR**

**21386 automatically generated clusters**

**6420 manually validated as gene families**

**4400 phylogenetically analysed gene families**

# **GreenPhylDB**

## **Important aspects**

### **A plant gene family database**

- **Most important plant gene family database: 6420 manually annotated**

### **A pre-computed phylogenomic analysis database**

- **Total number of gene families analyzed: 4400**

**Rice and Arabidopsis...ok**

**"But I'm working on maize or banana?"**

# PLAN

1. Sequencing projects and comparative genomics
2. GreenPhyl DB: A phylogenomic platform for plant comparative genomics
3. **GOST: GreenPhyl Orthologs Search Tool**

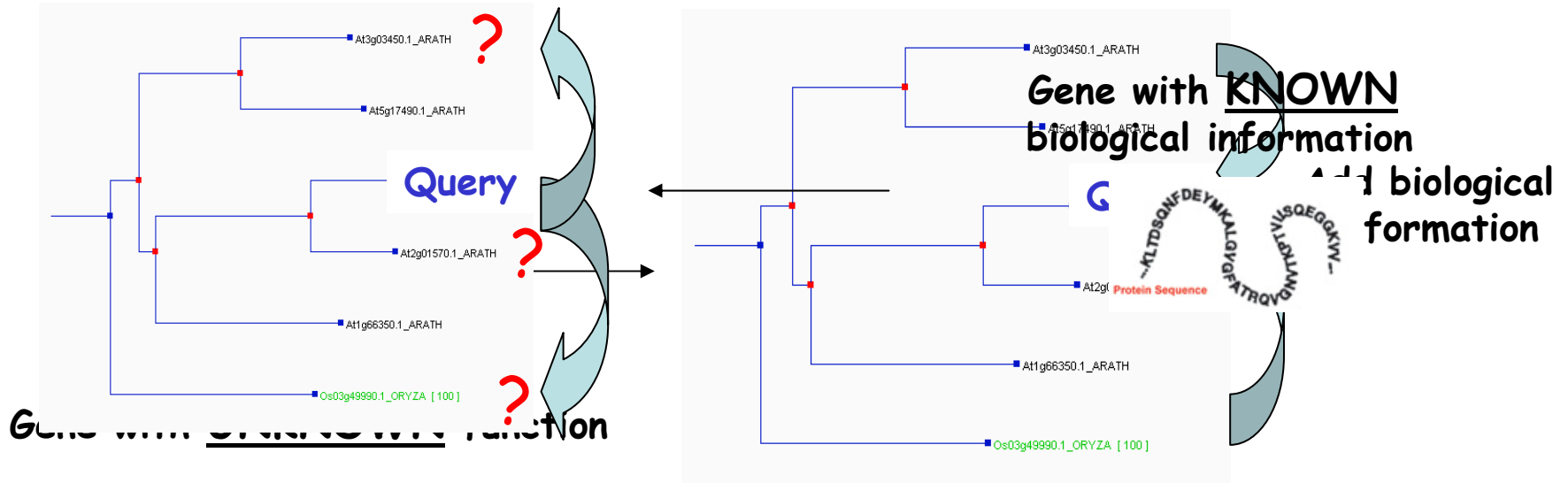
# GOST

(GreenPhyl Orthologs Search Tool)\*

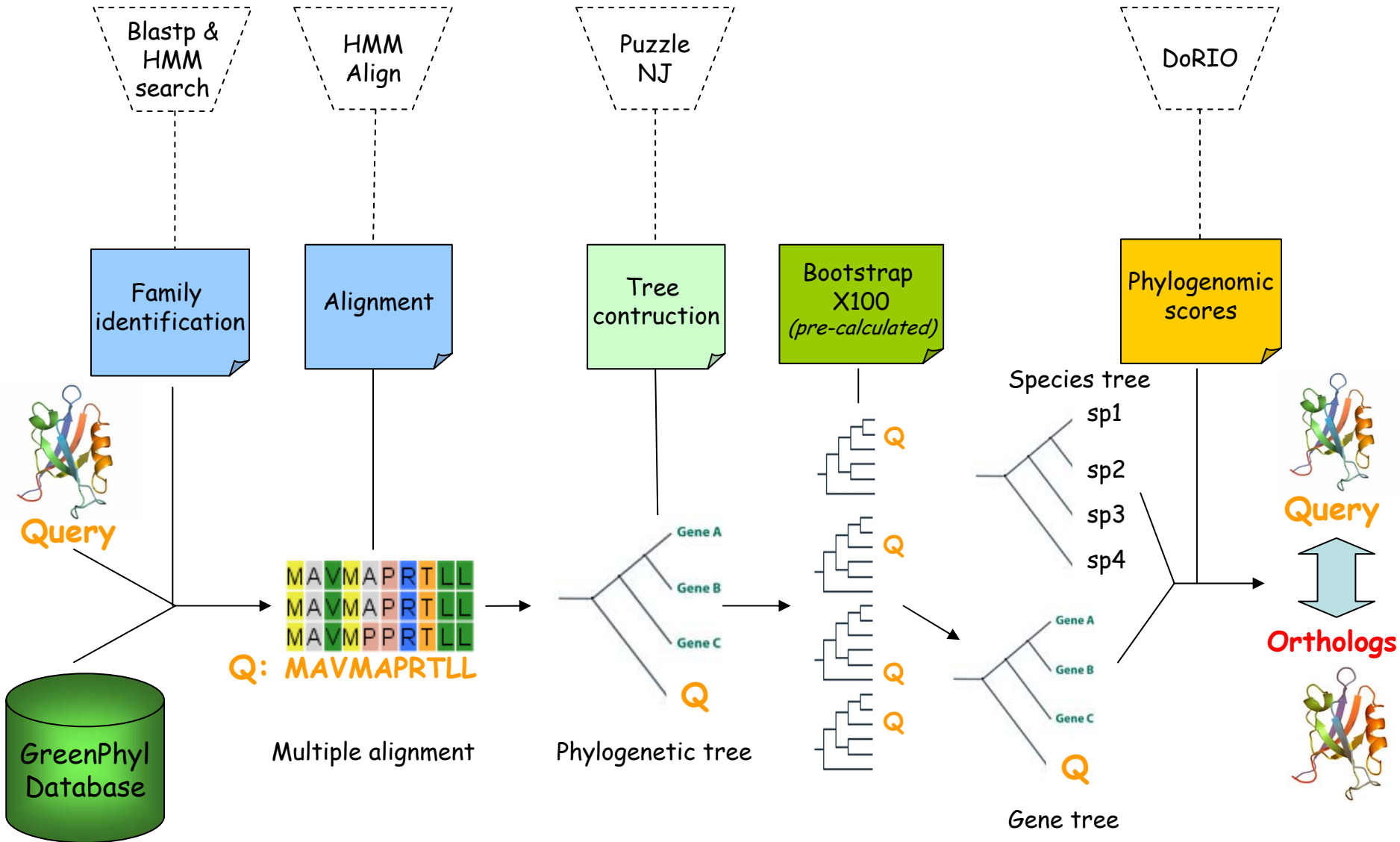
A Phylogenomic Tool for plant comparative genomics

2 different use cases

2. Add information to rice versus arabidopsis with genes from other species particularly studied or arabidopsis



# GOST pipeline



# Sequence submission

## GOST (GreenPhyl Orthologs Search Tool)

GOST is a powerful tool developed to predict phylogenomic relationships between any plant gene and *O.sativa* / *A.thaliana* gene(s). new sequence into a model phylogeny developed on *O.sativa* and *A.thaliana* to infer orthologs relationships.

Step1. Your sequence will be attributed to a GreenPhyl family

Step2. Your sequence will be integrated into the phylogenomic tree of the selected family

Paste your PROTEIN sequence in fasta format

```
>gi|75207630|sp|Q9ST59|RHT1 WHEAT DELLA protein RHT-1  
protein 1) (Protein Rht-B1/Rht-D1)  
MKREYQDAGGSGGGGGMGSSSEDKMMVSAAGEGEEVDELLAALGYKVRASDM  
GVGAGAAPDDSFATHLATDTVHYNPTDLSSWVESMLSELNAPPPPLPPAPQLNA  
PSVDSSSSIYALRPIPSGATAPADLSADSVRDPKRMRIIGSSSTSSSSSSSS  
AAAANATPALPVVVVDTQEAGIRLVHALLACAEAVQQENLSAAEALVKQIPLLE  
ALARRVFRFRPQPDSLLDAAFADLLHAHFYESCPLYLKFAHFTANQAILEAFAC  
WPALLQALALRPGGPPSFRLTGVGPPQPDETDALQQVGWKLQFAHTIRVDFQ  
QPEGEEDPNEEPEVIAVNSVFEMHRLLAQPGALEKVLGTVRAVRPRIVTVVEQ  
EANHNSGTFLDRFTESL  
HYSTMFDLSLEGGSSGGGPSEVSSGAAAAPAAAGTDQVMSEVYLGRQICNVVACEGA  
ERTERHETLGQWR  
NRLGNAGFETVHLGSNAYKQASTLLALFAGGDGYKVEEKEGCLTLGWHTRPLIATSA  
WRLAGP
```

'Green revolution' genes encode mutant gibberellin response modulators.  
Peng J and all. Nature 2000

Find family

Clear

**Note: Optimal performance with COMPLETE sequence**

# Family identification and species selection

Your sequence was inferred to be part of this GreenPhyl family:

Clustering level	Family Name	Family id
MCL_I1.2	GRAS transcription factor family	20939
MCL_I2		25010
MCL_I3		30566
MCL_I5	DELLA subfamily	37810

The phylogenomics analysis for the Cluster '20939' named 'GRAS transcription factor family' is available [here](#)

Select the specie of your query in the list hereunder:

**WARNING!** Please select cautiously the species of your query (or the most similar). This step might have a significant influence on the final results.

- Solanum bulbocastanum (SOLBU)
- Solanum lycopersicum (SOLLC)-Tomato
- Sorghum bicolor (SORBI)
- Spinacia oleracea (SPIOL)
- Staurastrum punctulatum (STAPU)
- Takakia ceratophylla (TAKCE)
- Theobroma cacao (CACAO)
- Triticum aestivum (TRIAE)- Wheat**
- Vicia faba (VICFA)-Fava bean
- Vigna unguiculata (VIGUN)
- Vitis vinifera (VITVI)
- Volvox carteri (VOLCA)
- Zea mays (MAIZE)

[Our plant tree of life](#)

[Link to NCBI Taxonomy](#)

**Requirement:**

- to indicate species

[Search homologs](#)

# Phylogenomic predictions for the query

Orthologs found

Orthology Subtree

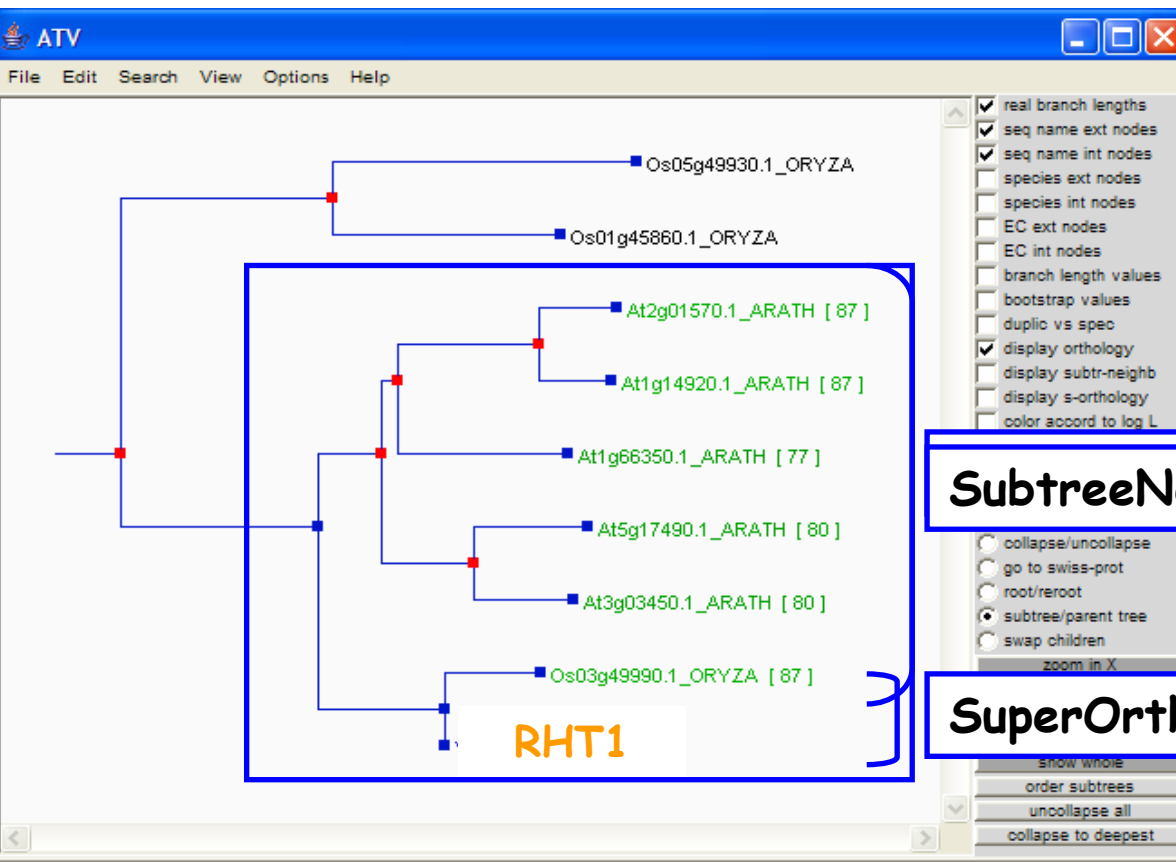
SuperOrthologs score %  
One-to-One relationship

Distance

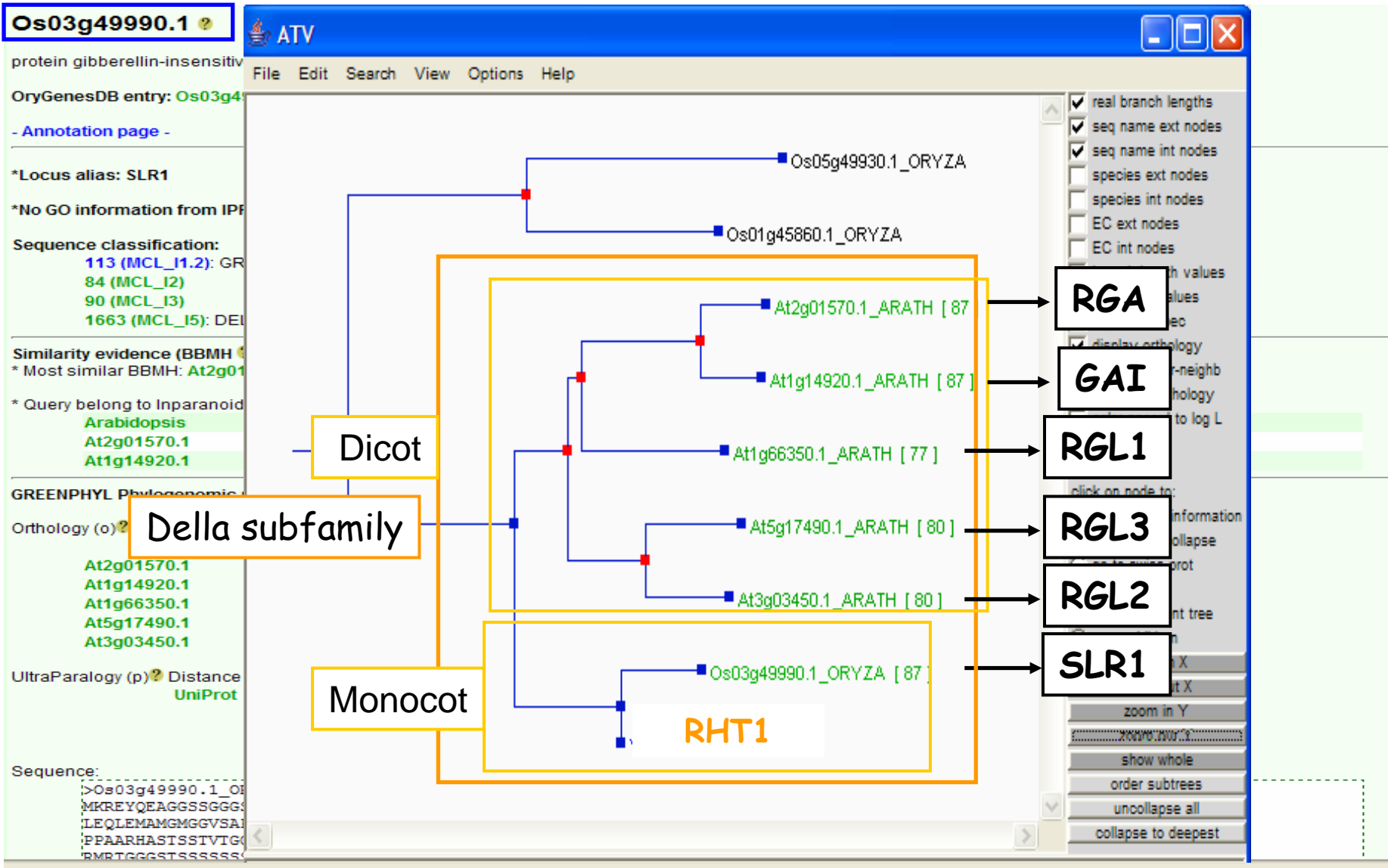
SubtreeNeighbor score %

Group into same clade

Id	Ortholog %	SubtreeNeighbor %	SuperOrthology %	Distance
Os03g49990.1	87	100	87	0.03184
At1g14920.1	87	97	0	0.27414
At2g01570.1	87	97	0	0.31311
At3g03450.1	80	90	0	0.33925
At5g17490.1	80	90	0	0.36147
At1g66350.1	77	87	0	0.34259



# Phylogenomic prediction for the query



# Web accessibility

- GOST is accessible via the GreenPhylDB website (<http://greenphyl.cirad.fr>)
- Web services:
  - Web services were developed allowing to integrate it into the **GCP platform** (e.g. <http://dayhoff.generationcp.org>)
  - Scientific community can use this service and include it for instance into an automatic workflow of genome annotation .

# GOST assets

GOST combines several advantages

1. Applied on GreenPhylDB plant families (6400 gene families)
2. Ranked ortholog and paralog prediction (phylogeny-based methods)
3. As fast as blast search
4. Can be inserted in automatic workflows

# THANKS



Dr Christophe Perin  
Gaetan Droc

The logo for IRRI (International Rice Research Institute) consists of the letters "IRRI" in a green, serif font, set against a solid yellow rectangular background.

Dr Richard Bruskiewich  
Mylah Anacleto  
Lord Hendrix Barboza

