



# GreenPhylDB v2.0

**An improved comparative genomic database for  
crop functional genomics**



**Matthieu CONTE**

# Plant Genome database increase

## Plant Genome Databases

### Alfalfa

[AlfaGene](#) A genetic database of the alfalfa (*Medicago sativa*) genome (mirror site at UK CropNet)

### Arabidopsis

[TAIR](#) A searchable relational database - a comprehensive resource for the scientific community working with *Arabidopsis thaliana*

### Banana

[MGRC](#) A banana (*Musa*) genomics information resource established by the Global Musa Genomics Consortium (GMGC)

### Barley

[BarleyDB](#) A barley genome database containing barley maps, traits, and sequences (by UK CropNet)

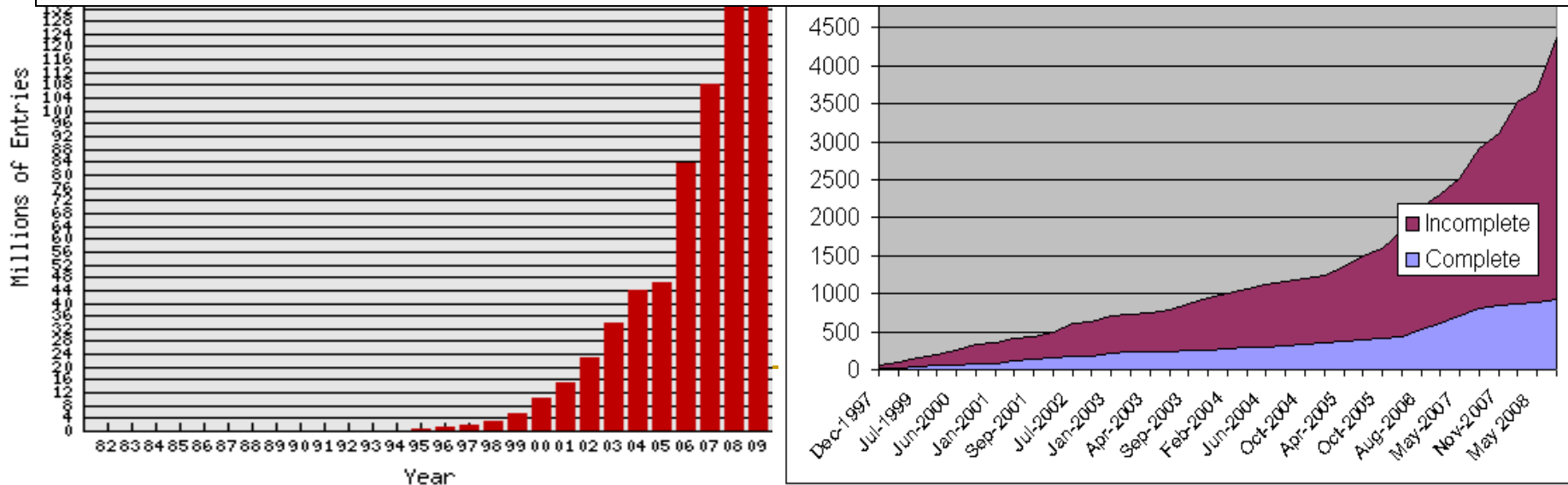
### Bean

[BeanGenes](#) A plant genome database containing information relevant to *Phaseolus* and *Vigna* species

### Brassica

**Genes of agronomic interest are probably already sequenced**

**How can we identify them?**



# Transfer information from model plants to crops

## Comparative genomics

GCP crops



Homologous  
(orthologous) genes



Gene information transfer

Model species



Gene with unknown function

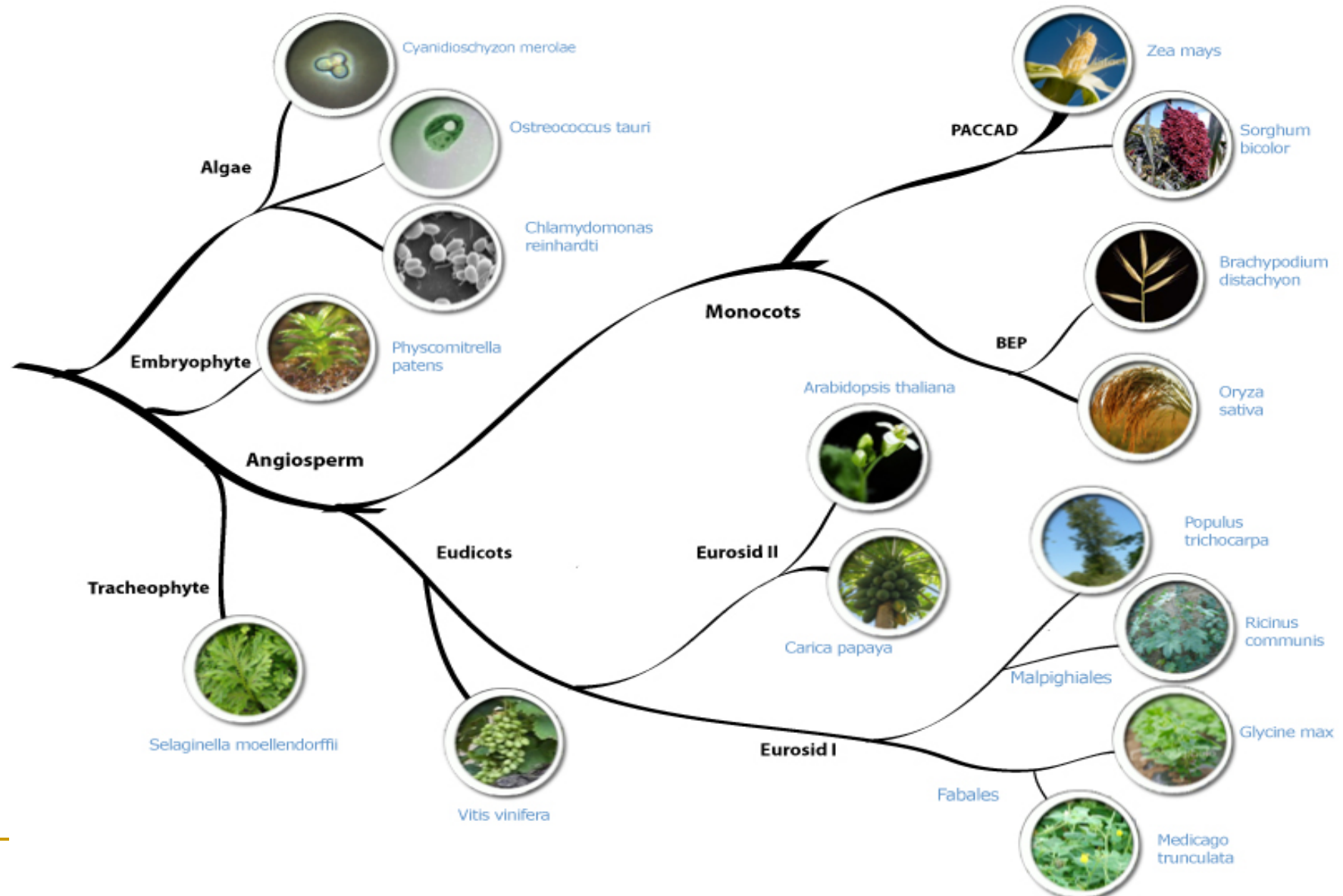
Gene with known function

---

# Our strategy:

- **Use full plant genomes**
  - **Develop a manually curated catalog of plant gene families**
  - **Provide clear homologs predictions use phylogenomic analysis**
  - **Provide external links and analysis tools**
-

# Results



---

# Results

## A plant gene family database

- **Most important plant gene family database manually annotated\***

## A phylogenomic analysis database

- **Phylogenomic analyses to identify homologous genes**

---

\* **GreenPhyl family annotation Platform**  
3rd Biocurator conference 2009, Berlin Germany

# Searching data: Different entry points

The image shows a screenshot of the GreenPhyl website's search interface. The page has a green header with the 'GreenPhyl' logo on the left and a navigation bar with 'Home', 'Clustering lists', and 'Family lists'. A search box is located on the right, with a 'Search' button and a help icon. A dropdown menu is open, showing a list of search criteria: 'Family\_name (e.g. GRAS)', 'Sequence\_id (e.g. At1g14920)', 'Family\_id (e.g. 20939)', 'Alias (e.g. GAI)', 'Annotation (e.g. GRAS)', 'Interpro (e.g. IPR005202)', 'UniProt (e.g. Q9LQT8)', 'KEGG (e.g. K00430)', and 'GO\_molecular\_function (e.g. GO:0006950)'. Several grey boxes with blue arrows point to different parts of the interface: 'Gene ID (TAIR, TIGR)' points to the dropdown menu; 'Gene name (alias)' points to the 'Alias' option; 'Gene annotation' points to the 'Annotation' option; 'Family name' points to the search input field; 'GO' points to the 'GO\_molecular\_function' option; and 'UniProt' points to the 'UniProt' option.

**Gene ID (TAIR, TIGR)**

**Gene name (alias)**

**Gene annotation**

**Family name**

**GO**

**UniProt**

Family\_name (e.g. GRAS)  
Sequence\_id (e.g. At1g14920)  
Family\_id (e.g. 20939)  
Alias (e.g. GAI)  
Annotation (e.g. GRAS)  
Interpro (e.g. IPR005202)  
UniProt (e.g. Q9LQT8)  
KEGG (e.g. K00430)  
GO\_molecular\_function (e.g. GO:0006950)

Home Clustering lists Family lists

Search ?

**A phylogenomic database for plant comparative genomics**

# Family entry page:

Identify all genes potentially involved in the same process

Home Clustering lists Family lists Tools Documentation About us

Family name	Major Intrinsic family (MIP)
Synonym	Plasma membrane intrinsic protein (PIP) subfamily
Family ID	
Cross-reference(s)	TAIR
Curation status	NOD26-like intrinsic protein (NIP) subfamily
Family status	Phylogenomic analysis not done

Link to GreenPhyl V1.0 : [20952](#)


Family structure Family composition Protein pattern Protein-coding sequences Orthologs Gene expression

Clustering level	GreenPhyl family id (number of sequences)				
1	20952 (392)				
2	25016 (363)	27464 (29)			
3	30808 (104)	30608 (259)	34869 (21)		
4	40549 (21)	36346 (148)	36460 (104)	36570 (111)	

Small basic intrinsic protein (SIP) subfamily

Tonoplast intrinsic protein (TIP) subfamily

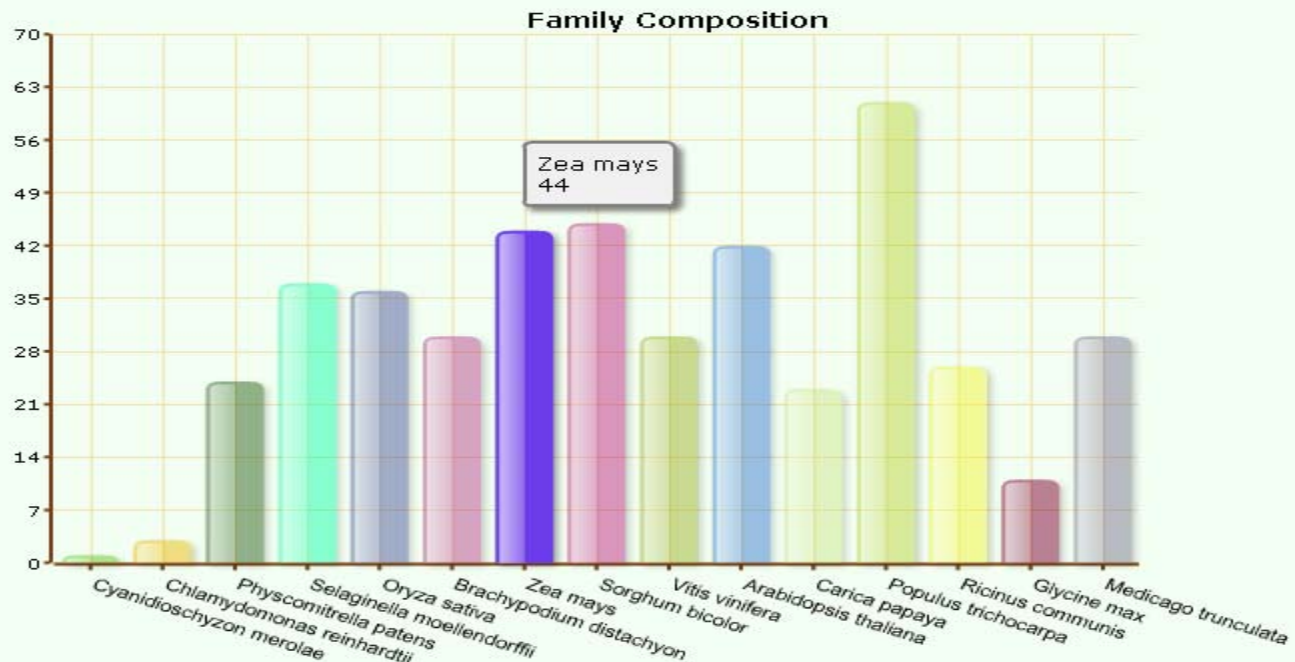
# Family entry page

Family name	Major Intrinsic family (MIP)
Synonym	
Family ID	20952
Cross-reference(s)	<a href="#">TAIR</a>
Curation status	 <a href="#">[Curate it]</a>
Family status	Phylogenomic analysis not done

Link to GreenPhyl V1.0 : [20952](#)

<a href="#">Family structure</a>	<a href="#">Family composition</a>	<a href="#">Protein pattern</a>	<a href="#">Protein-coding sequences</a>	<a href="#">Orthologs</a>	<a href="#">Chromosome position</a>
----------------------------------	------------------------------------	---------------------------------	--	---------------------------	-------------------------------------

Total : 443



Information to display :

Gene name  UniProt  KEGG  InterPro  Annotation  Gene Ontology

# Family entry page


Family name	Major Intrinsic family (MIP)
Synonym	
Family ID	20952
Cross-reference(s)	TAIR

Load this in Excel format

Get selected sequences

	Species code	Sequence id	Locus Alias	UniProt	KEGG	InterPro	GO Term ID
<input checked="" type="checkbox"/>	ARATH	At1g01620.1	PIP1-3	Q08733	K09872		plasma membrane intrinsic protein 1C (PIP1C) / aqu
<input checked="" type="checkbox"/>	ARATH	At1g17810.1	TIP3-2	O22588	K09873	IPR000425 IPR012269	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At1g17810.2	TIP3-2	O22588	K09873		major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At1g31885.1	NIP3-1	Q9C6T0	K09874	IPR000425 IPR012269	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At1g52180.1				IPR000425	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At1g73190.1	TIP3-1	P26587	K09873	IPR000425 IPR002197 IPR012269	tonoplast intrinsic protein, alpha / alpha-TIP (TI
<input checked="" type="checkbox"/>	ARATH	At1g80760.1	NIP6-1	Q9SAI4	K09874	IPR000425 IPR005829	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At2g16835.1				IPR000425	water channel protein, putative, similar to MipC (
<input checked="" type="checkbox"/>	ARATH	At2g16850.1	PIP2-8	Q9ZVX8	K09872	IPR000425 IPR012269	plasma membrane intrinsic protein, putative, very
<input checked="" type="checkbox"/>	ARATH	At2g25810.1	TIP4-1	O82316	K09873	IPR000425 IPR012269	tonoplast intrinsic protein, putative, similar to
<input checked="" type="checkbox"/>	ARATH	At2g29870.1			K09874	IPR000425	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At2g34390.1	NIP2-1	Q8W037	K09874	IPR000425 IPR012269	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At2g34390.2	NIP2-1	Q8W037	K09874		major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At2g36830.1	TIP1-1	P25818	K09873	IPR000425 IPR012269	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At2g37170.1	PIP2-2	P43287	K09872	IPR000425 IPR012269	plasma membrane intrinsic protein 2B (PIP2B) / aqu
<input checked="" type="checkbox"/>	ARATH	At2g37180.1	PIP2-3	P30302	K09872	IPR000425 IPR012269	plasma membrane intrinsic protein 2C (PIP2C) / aqu
<input checked="" type="checkbox"/>	ARATH	At2g39010.1	PIP2-6	Q9ZV07	K09872	IPR000425 IPR012269	aquaporin, putative, similar to plasma membrane aq
<input checked="" type="checkbox"/>	ARATH	At2g45960.1	PIP1-2	Q06611	K09872	IPR000425 IPR012269	plasma membrane intrinsic protein 1B (PIP1B) / aqu
<input checked="" type="checkbox"/>	ARATH	At3g04090.1	SIP1-1	Q9M8W5	K09875	IPR000425	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At3g06100.1	NIP7-1	Q8LAI1	K09874	IPR000425	major intrinsic family protein / MIP family protei
<input checked="" type="checkbox"/>	ARATH	At3g16240.1	TIP2-1	Q41951	K09873	IPR000425 IPR012269	delta tonoplast integral protein (delta-TIP), iden

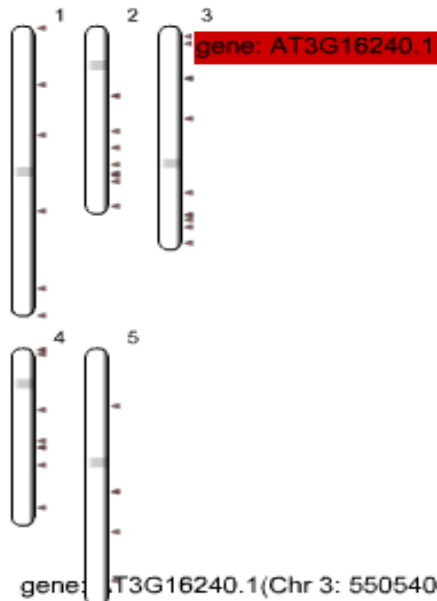
# Family entry page

Family name	Major Intrinsic family (MIP)
Synonym	
Family ID	20952
Cross-reference(s)	<a href="#">TAIR</a>
Curation status	 <a href="#">[Curate it]</a>
Family status	Phylogenomic analysis not done

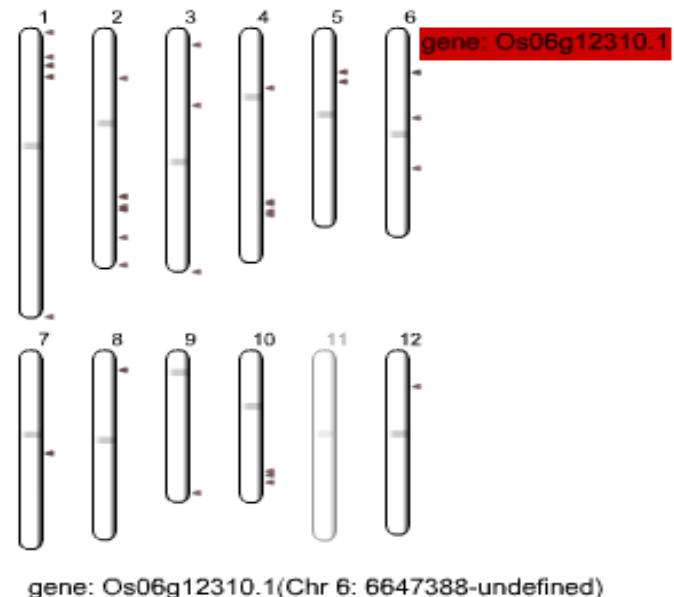
Link to GreenPhyl V1.0 : [20952](#)

[Family structure](#) [Family composition](#) [Protein pattern](#) [Protein-coding sequences](#) [Orthologs](#) [Chromosome position](#)

*Arabidopsis thaliana*



*Oryza sativa*



---

# Results

## A plant gene family database

- Most important plant gene family database manually annotated

## A phylogenomic analysis database

- Phylogenomic analyses to identify homologous genes
-

# Sequence entry page

## Identify genes with potential similar function

Home

Clustering lists

Family lists

Tools

Documentation

About us

Sequence id	At1g73190.1
Species	Arabidopsis thaliana
Alias	TIP3-1
Gene annotation	tonoplast intrinsic protein, alpha / alpha-TIP (TIP3.1), identical to SP:P26587 Tonoplast intrinsic protein, alpha (Alpha TIP) (Arabidopsis thaliana) (Plant Physiol. 99, 561-570 (1992))
Gene Ontology	transporter activity
Cross-references	TAIR entry: <a href="#">At1g73190.1</a> OryGenesDB entry: <a href="#">At1g73190.1</a> UniProt entry: <a href="#">P26587</a> Geneinvestigator entry: <a href="#">P26587</a> T-DNA Express entry: <a href="#">At1g73190</a> Germplasm entry: <a href="#">At1g73190</a>

Link to GreenGenes V1.0: [At1g73190.1](#)

Gene classification

Gene model

Gene sequence

Domain pattern

Phylogenomic predictions

Tree - alignment

Orthology (o) <sup>?</sup> Subtree-neighbor (n) <sup>?</sup> SuperOrthologs (s) <sup>?</sup> Distance (D)

[Os10g35050.1](#)

UniProt  
[Q9FWV6](#)

Alias  
TIP31

o  
100

n  
99

s  
0

D  
0.4294

UltraParalogy (p) <sup>?</sup> Distance

[At1g17810.1](#)

UniProt  
[Q22588](#)

Alias  
TIP32

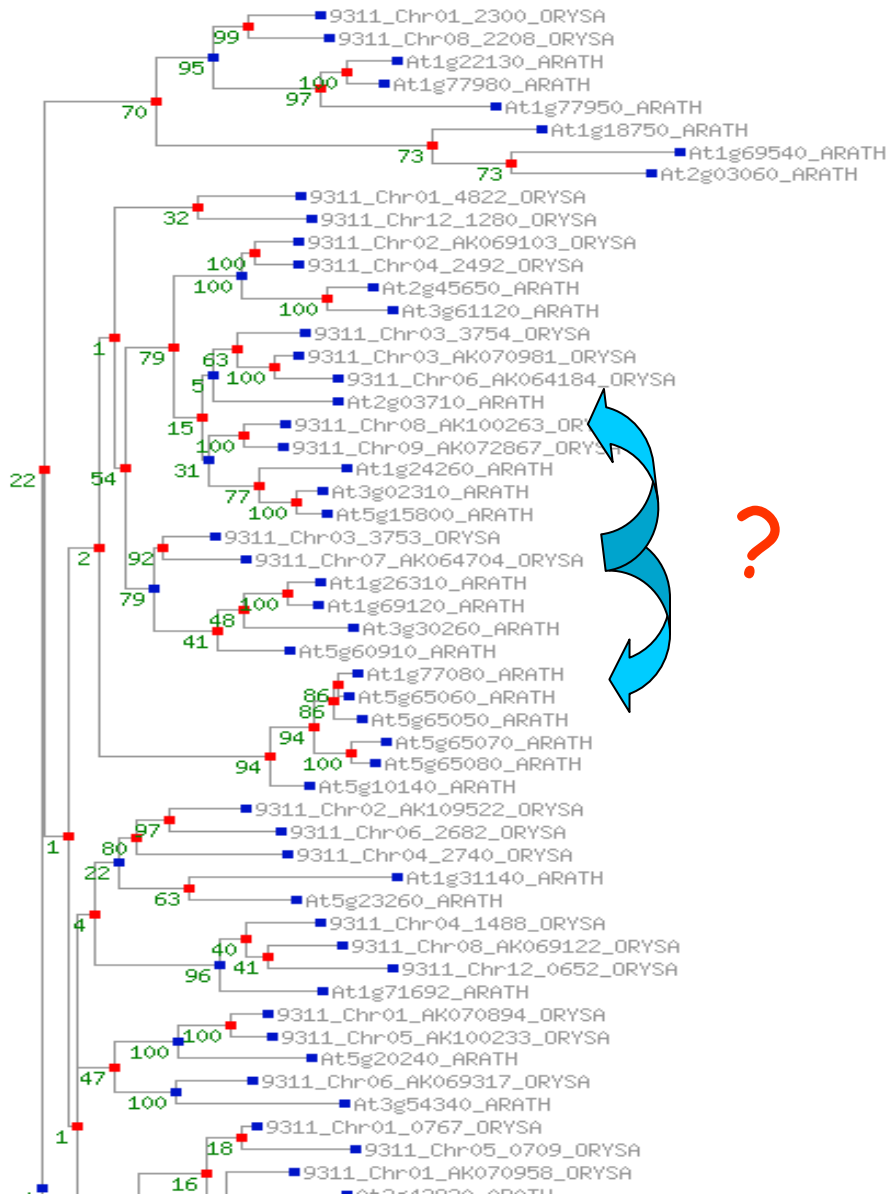
p  
100

D  
0.14341

Segmental duplication(inv): Oryza list <sup>?</sup> Arabidopsis list <sup>?</sup>

[At1g17810.1](#) Group: 2

# Orthologs prediction in other databases



How to identify easily the orthologs relationships?

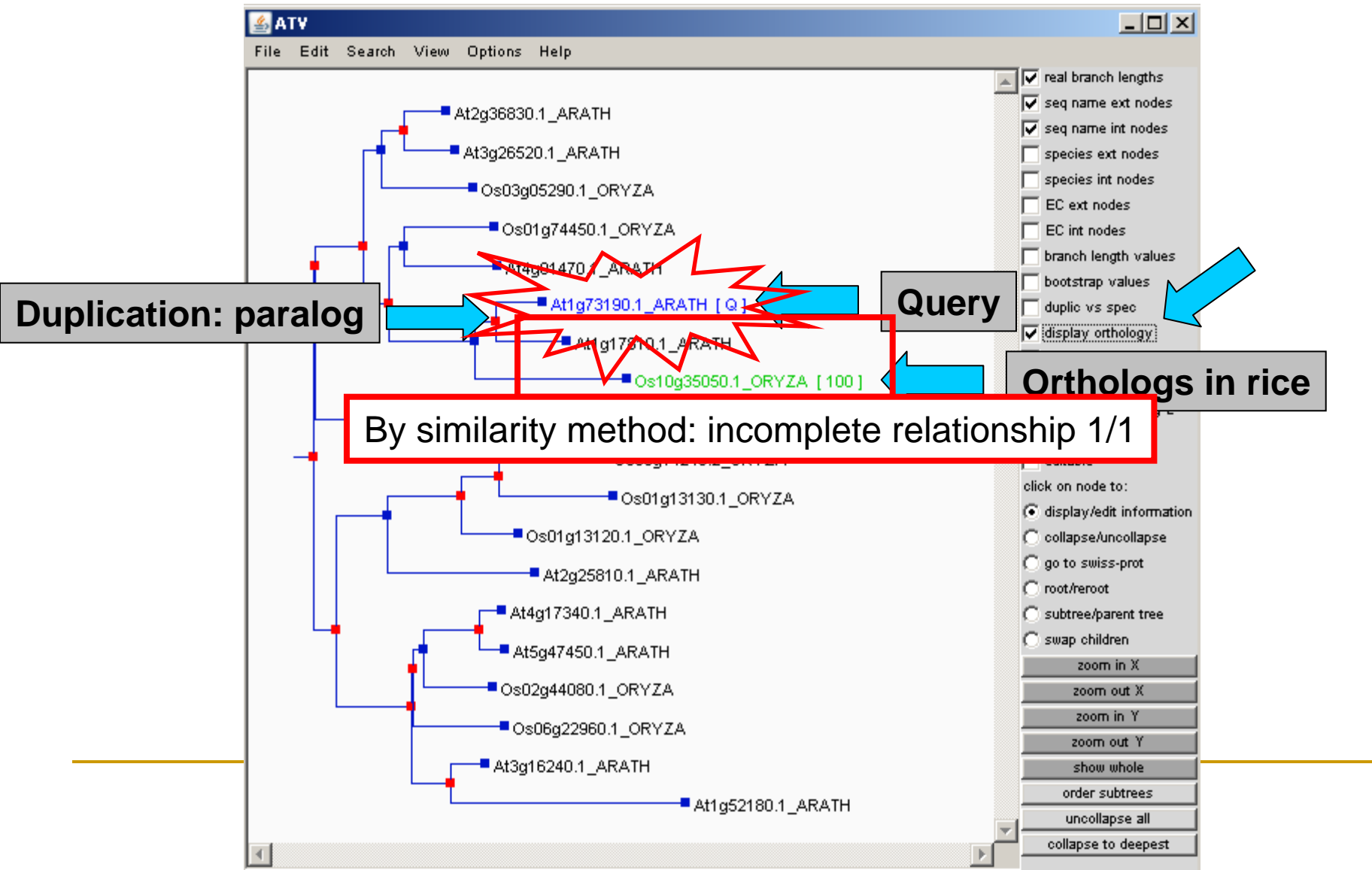
# Ortholog scoring system to facilitate interpretations

Sequence id	At1g73190.1
Species	Arabidopsis thaliana
Alias	TIP3-1
Gene annotation	tonoplast intrinsic protein, alpha / alpha-TIP (TIP3.1), identical to SP:P26587 Tonoplast intrinsic protein, alpha (Alpha TIP) (Arabidopsis thaliana) (Plant Physiol. 99, 561-570 (1992))
Gene Ontology	transporter activity
Cross-references	TAIR entry: <a href="#">At1g73190.1</a> OryGenesDB entry: <a href="#">At1g73190.1</a> UniProt entry: <a href="#">P26587</a> Genevestigator entry: <a href="#">P26587</a> T-DNA Express entry: <a href="#">At1g73190</a> Germplasm entry: <a href="#">At1g73190</a>

Link to GreenPhyl V1.0: [At1g73190.1](#)

Gene classification	Orthologs found	sequence	Domain pattern	Phylog	Confidence score	Annotation	
Orthology (o) ? Subtree-neighbor (n) ? SuperOrthologs (s) ? Distance (D)	<a href="#">Os10g35050.1</a>	UniProt <a href="#">Q9FWV6</a>	Alias TIP31	<b>o</b> 100	<b>n</b> 99	<b>s</b> 0	<b>D</b> 0.4294
UltraParalogy (p) ? Distance	<a href="#">At1g17810.1</a>	UniProt <a href="#">O22588</a>	Alias TIP32	<b>p</b> 100			<b>D</b> 0.14341
Paralogs found ? Arabidopsis list ?							Confidence score

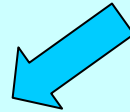
# Orthologs prediction visualisation



# Orthologs prediction

Sequence id	At1g73190.1
Species	Arabidopsis thaliana
Alias	TIP3-1
Gene annotation	tonoplast intrinsic protein, alpha / (Plant Physiol. 99, 561-570 (1992))
Gene Ontology	transporter activity
Cross-references	TAIR entry: <a href="#">At1g73190.1</a> OryGenesDB entry: <a href="#">At1g73190.1</a> UniProt entry: <a href="#">P26587</a> Genevestigator entry: <a href="#">P26587</a> T-DNA Express entry: <a href="#">At1g73190</a> Germplasm entry: <a href="#">At1g73190</a>

**GENEVESTIGATOR**  
shaping biological discovery  
**expression data access**



tonoplast intrinsic protein, alpha (Alpha TIP) (Arabidopsis thaliana)

Link to GreenPhyl V1.0 : [At1g73190.1](#)

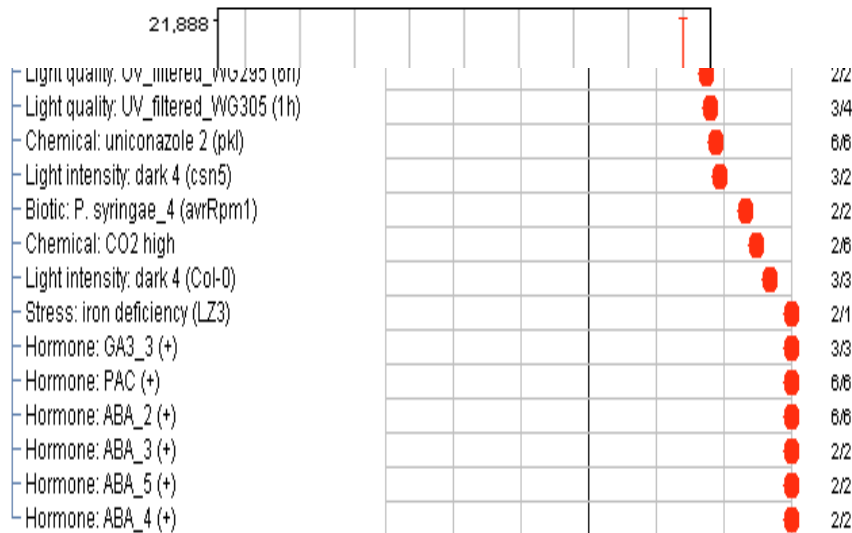
Gene classification	Gene model	Gene sequence	Domain pattern	Phylogenomic predictions	Tree - alignment
<b>Orthology (o)</b> <sup>?</sup> Subtree-neighbor (n) <sup>?</sup> SuperOrthologs (s) <sup>?</sup> Distance (D)					
<a href="#">Os10g35050.1</a>		<b>UniProt</b> <a href="#">Q9FWV6</a>		<b>Alias</b> TIP31	<b>o</b> 100 <b>n</b> 99 <b>s</b> 0 <b>D</b> 0.4294
<b>UltraParalogy (p)</b> <sup>?</sup> Distance					
<a href="#">At1g17810.1</a>		<b>UniProt</b> <a href="#">O22588</a>		<b>Alias</b> TIP32	<b>p</b> 100 <b>D</b> 0.14341
Segmental duplication(inv): Oryza list <sup>?</sup> Arabidopsis list <sup>?</sup>					
<a href="#">At1g17810.1</a> Group: 2					

# From Orthologs prediction to function

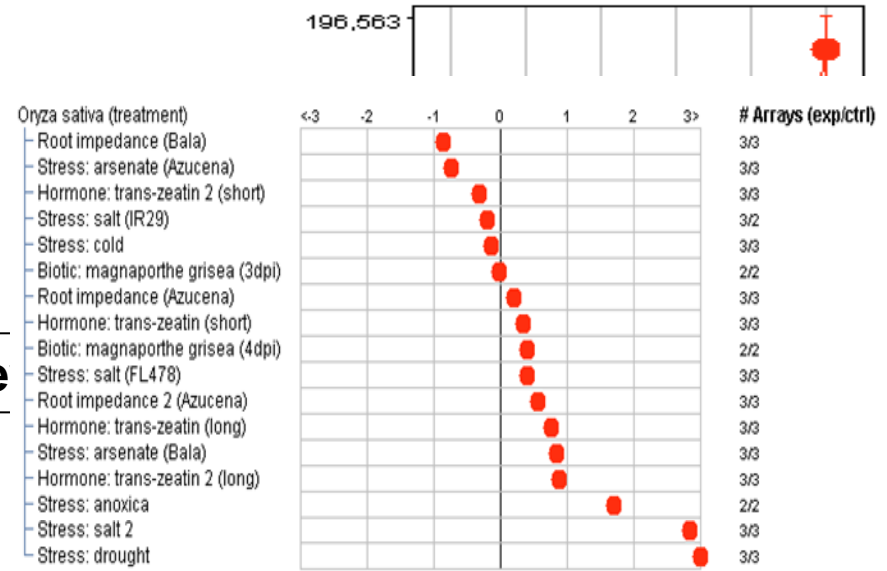
2 arabidopsis genes



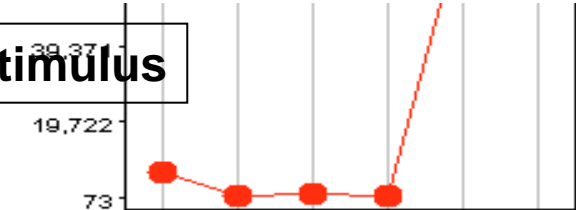
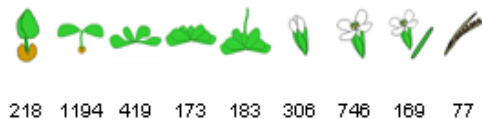
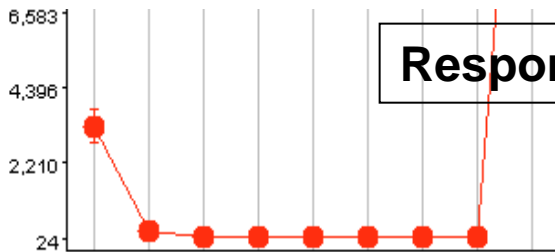
1 rice gene



de



Response to drought stress stimulus



# Outputs and future

- **Comprehensive list of plant gene families**  
*(Families involved in stress, species and phylum specific families...)*
  - **Tools to study gene families and protein coding genes in plants**  
*(Chromosome position, Protein patterns, Dynamic statistics, GOST: GreenPhyl Orthologs Search Tool, annotation platform ...)*
  - **List of homologous genes between crops with/without expression data**
  - **Manual curation and bio-analysis of gene families:**
    1. **Classification of gene families using Gene Ontology (GO plant slim)**
    2. **Link gene and family annotation: collaboration with the Swiss Institute of Bioinformatics (Uniprot)**
-

# Acknowledgements



**Mathieu Rouard (PI)**



**Christophe Perin  
Nadège Lanau  
Gaetan Droc  
Valentin Guignon**



---

# THANKS

**Contacts:**

**Mathieu CONTE: M.CONTE@CGIAR.ORG**

**Mathieu ROUARD: M.ROUARD@CGIAR.ORG**

---