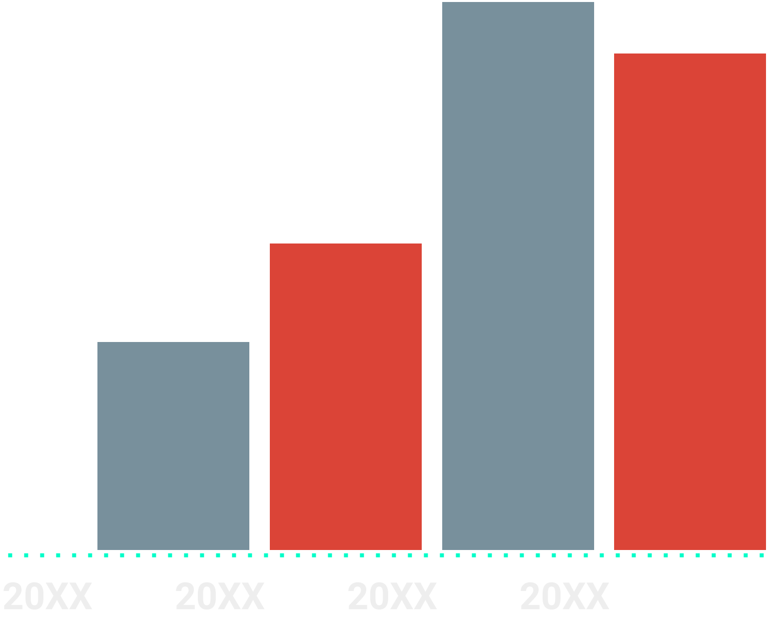
COMPARATIVE GENOMICS OF GENE FAMILIES IN RELATION WITH METABOLIC PATHWAYS FOR GENE CANDIDATES HIGHLIGHTING

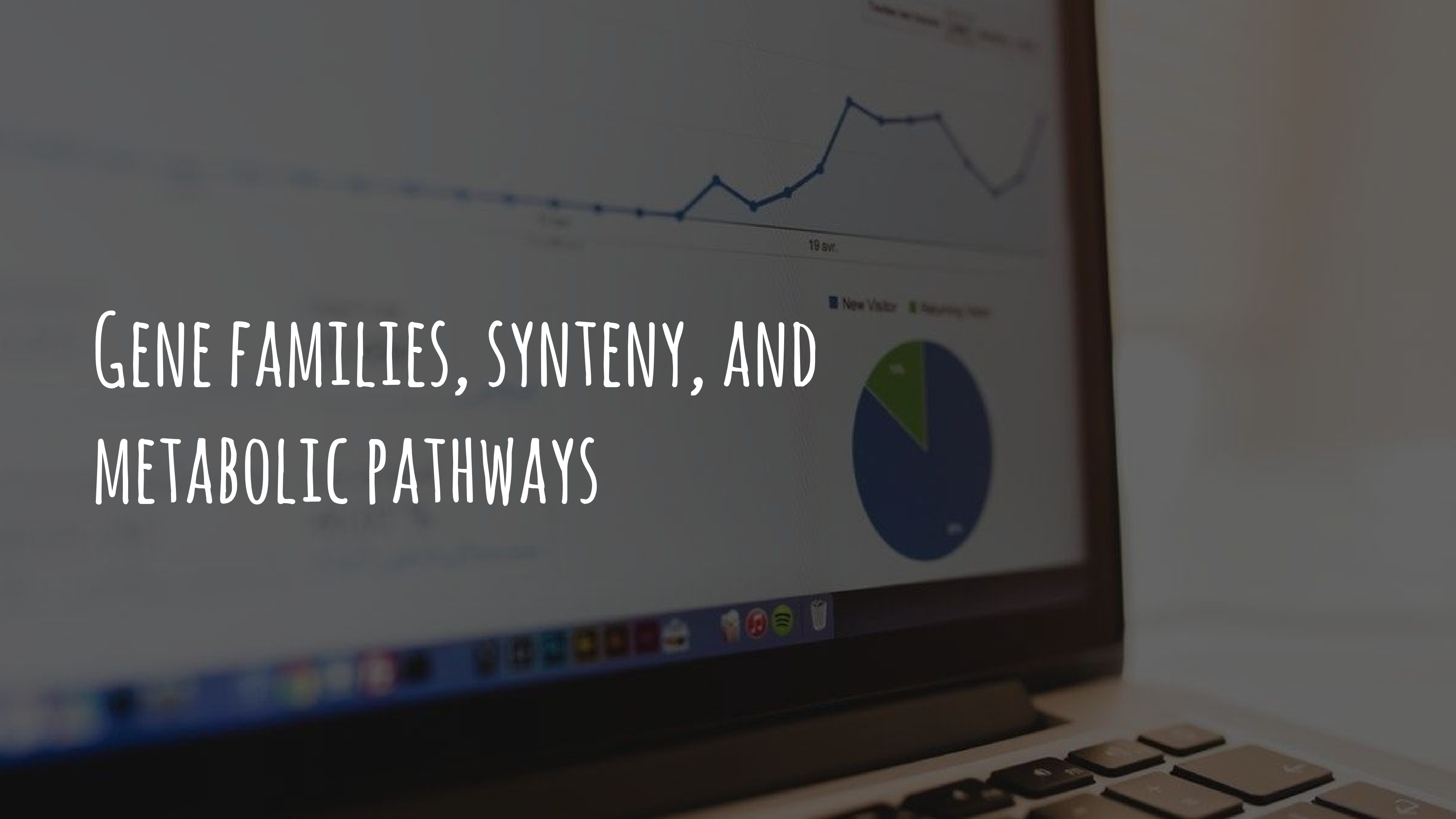
Delphine Larivière & David Couvin

Under the supervision of Dominique This, Jean-François Dufayard and Stéphanie Bocs

SUMMARY

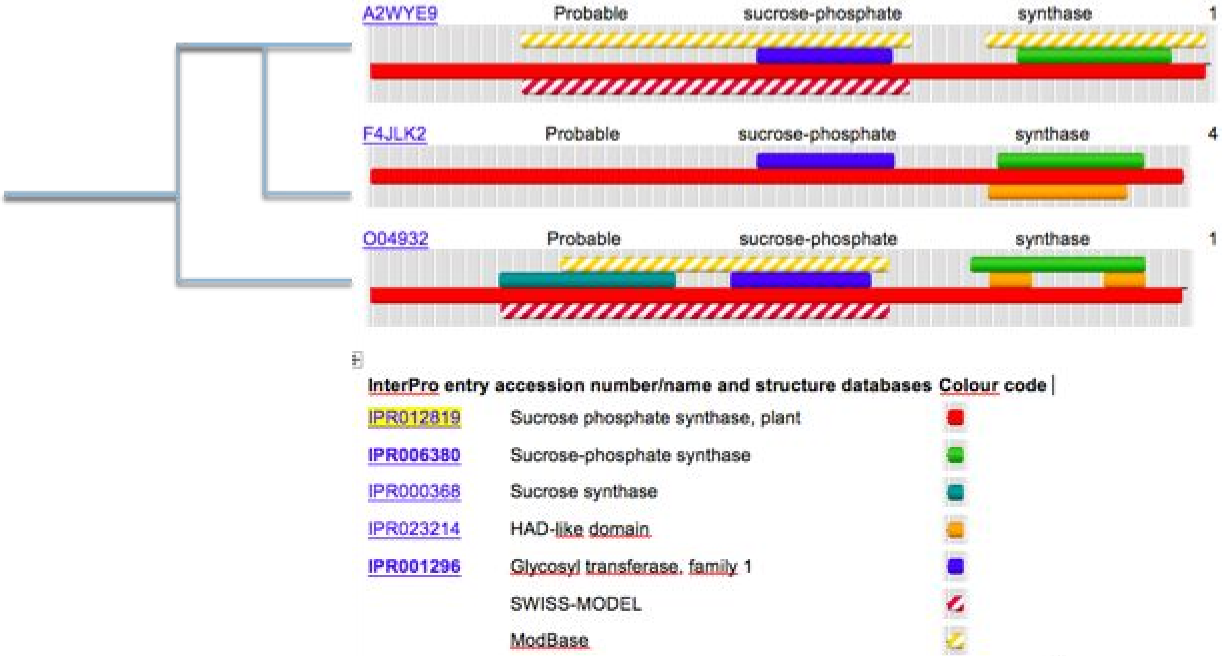
- Introduction to gene families
- Identification of evolutive events
- Metabolic pathways
- GenFam
- GenesPath
- Perspectives
- Conclusions





Gene families, synteny, and metabolic pathways

Introduction to Gene Families : Definition

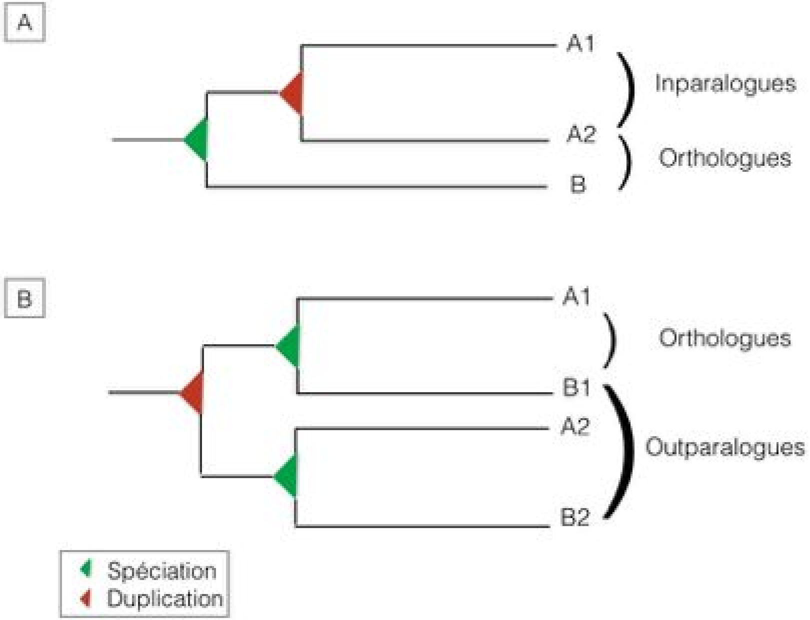


InterPro entry accession number/name and structure databases		Colour code
IPR012819	Sucrose phosphate synthase, plant	🟥
IPR006380	Sucrose-phosphate synthase	🟩
IPR000368	Sucrose synthase	🟦
IPR023214	HAD-like domain	🟧
IPR001296	Glycosyl transferase, family 1	🟪
	SWISS-MODEL	▨
	ModBase	▨

Introduction to Gene Families : Evolution

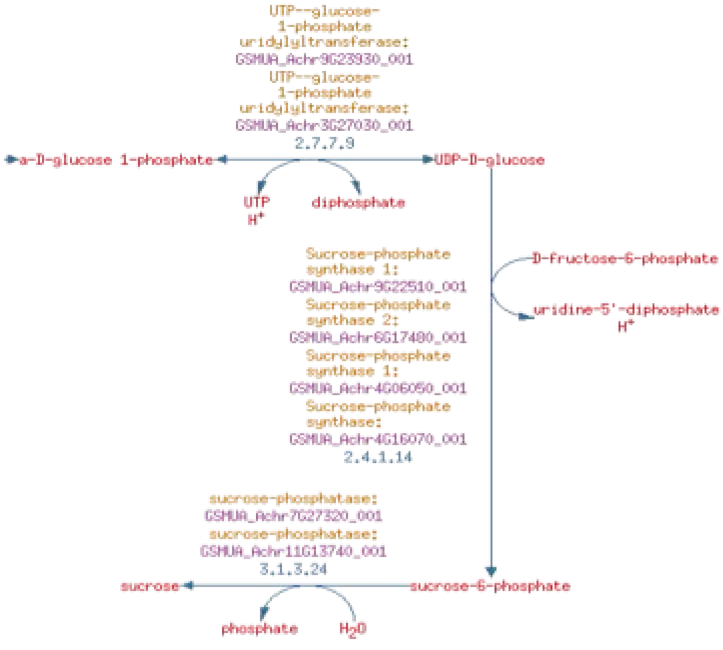
- Speciations

- Duplications :
    - Genes
    - Chromosomal segments
    - Complete Genomes (WGD, polyploïdy)



INTRODUCTION TO GENE FAMILIES : FUNCTIONAL EVOLUTION

- Type of Evolution
  - Neo-functionalization
  - Sub-functionalization
  - Pseudo-genes, Losses



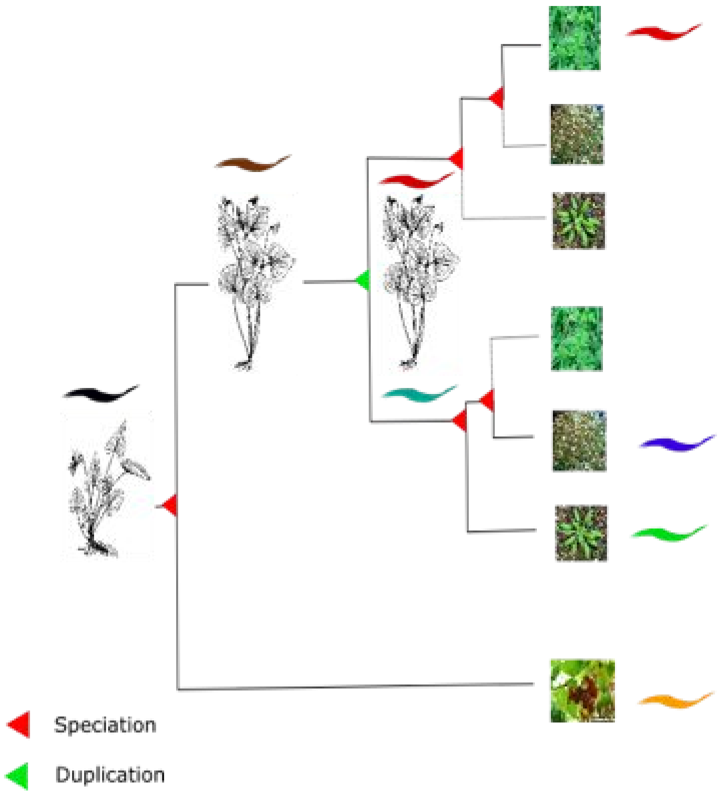
Source : Pathway tool

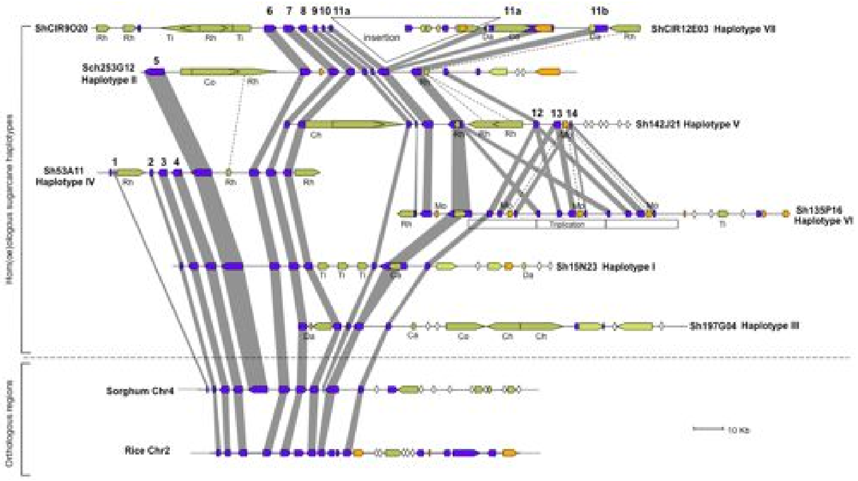
Introduction to Gene Families : Hypothesis

- The function is more preserved between orthologs than between paralogs [Altenhoff et al., 2012]
- Need to consider the time of divergence between two sequences; indeed, close paralogues may have undergo less changes than ancient orthologs and therefore present more similar functions [Nehrt et al., 2011]

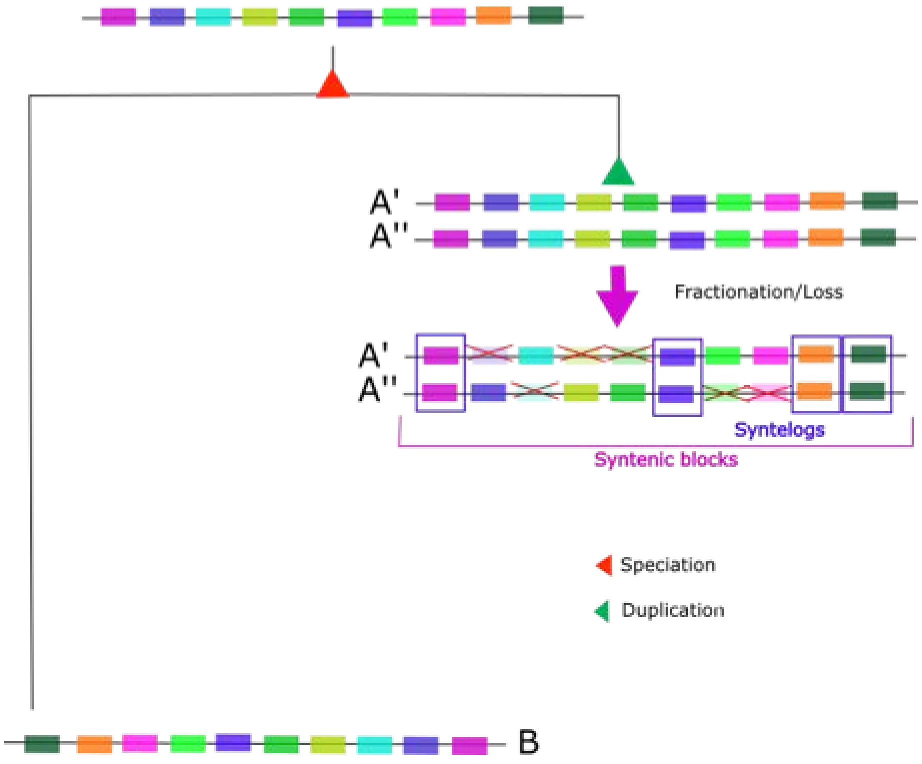
→ Knowledge of evolutive history of genes important for functional inference.

Analysis of evolutive event

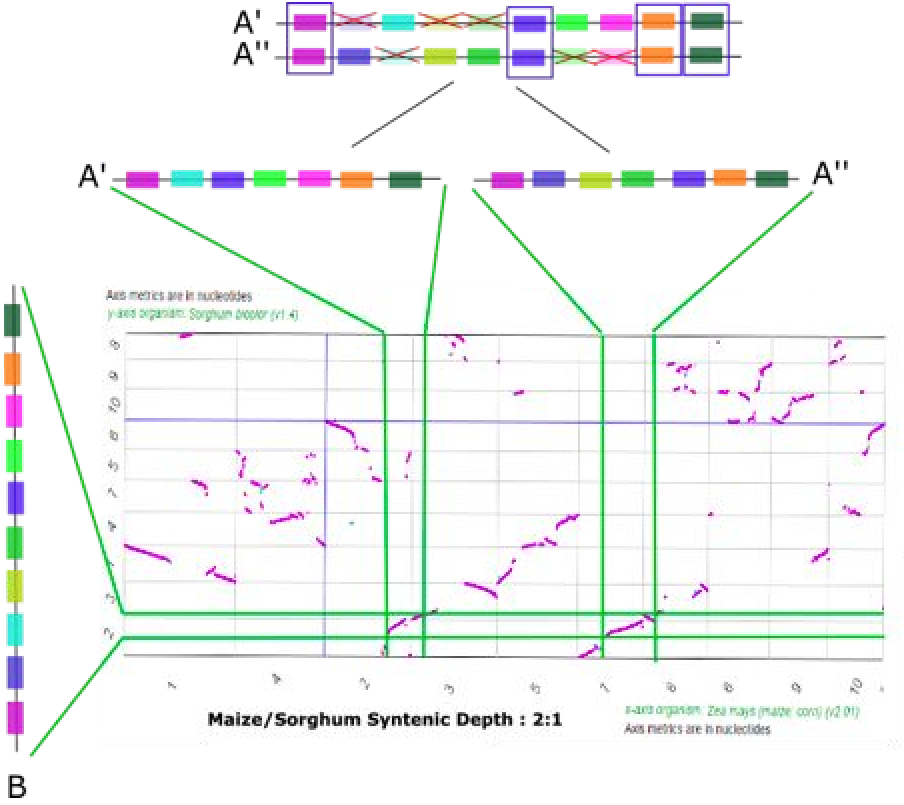




ANALYSIS OF EVOLUTIVE EVENT : SYNTENY

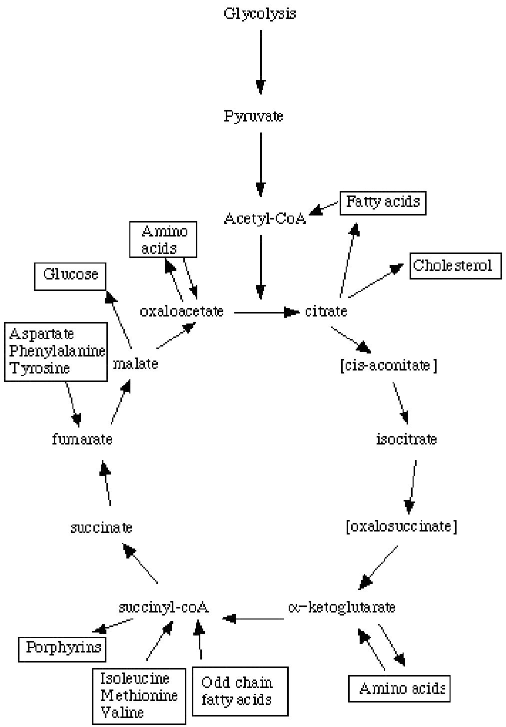


Analysis of evolutive event : Synteny



METABOLIC PATHWAYS

- In a metabolic pathway, the product of one enzyme acts as the substrate for the next. These enzymes often require dietary minerals, vitamins, and other cofactors to function.

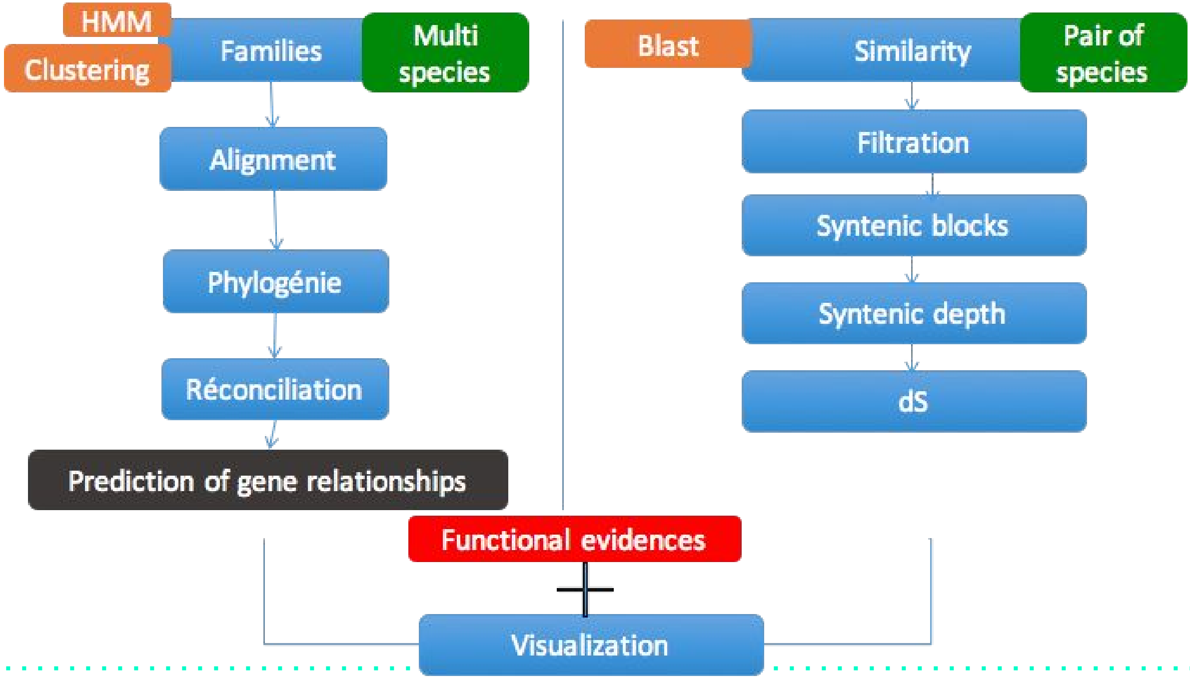


Example: Amphibolic Properties of the Citric Acid Cycle

- In our studies, the link between metabolic pathways and gene families is important to better understand the specificities of some genes, and their involvement in identified biological processes.

GenFam: a tool dedicated to gene families study

GenFam : Analysis of gene Families



HMM

Clustering

Families

Multi species

Alignment

Phylogénie

Réconciliation

Prediction of gene relationships

Blast

Similarity

Pair of species

Filtration

Syntenic blocks

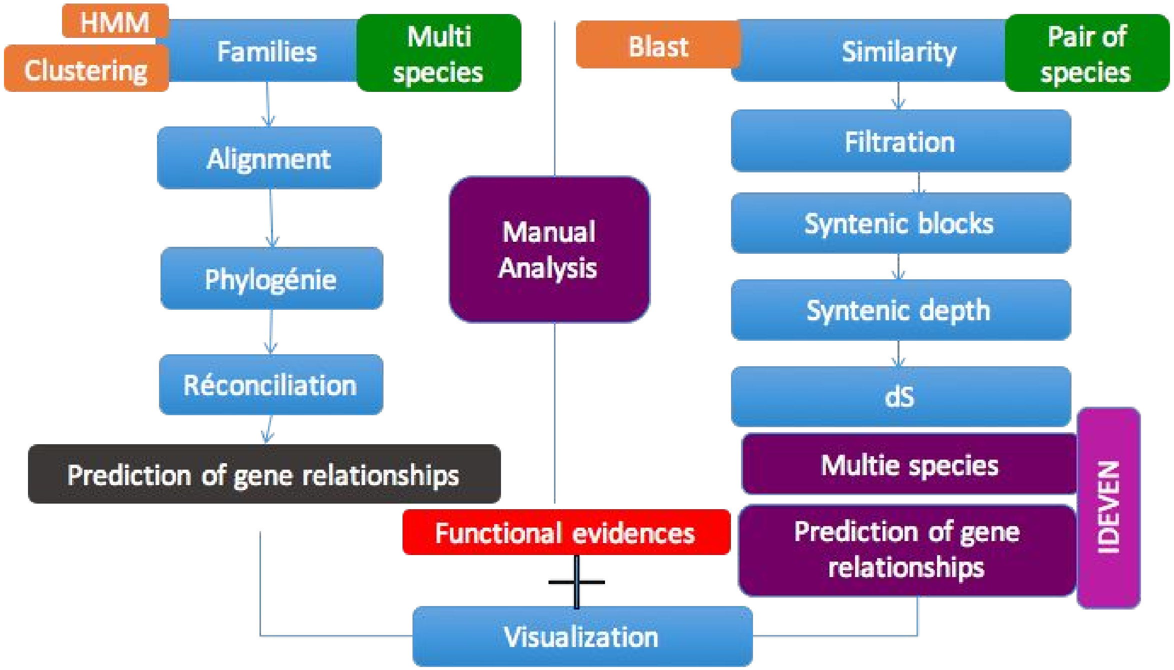
Syntenic depth

dS

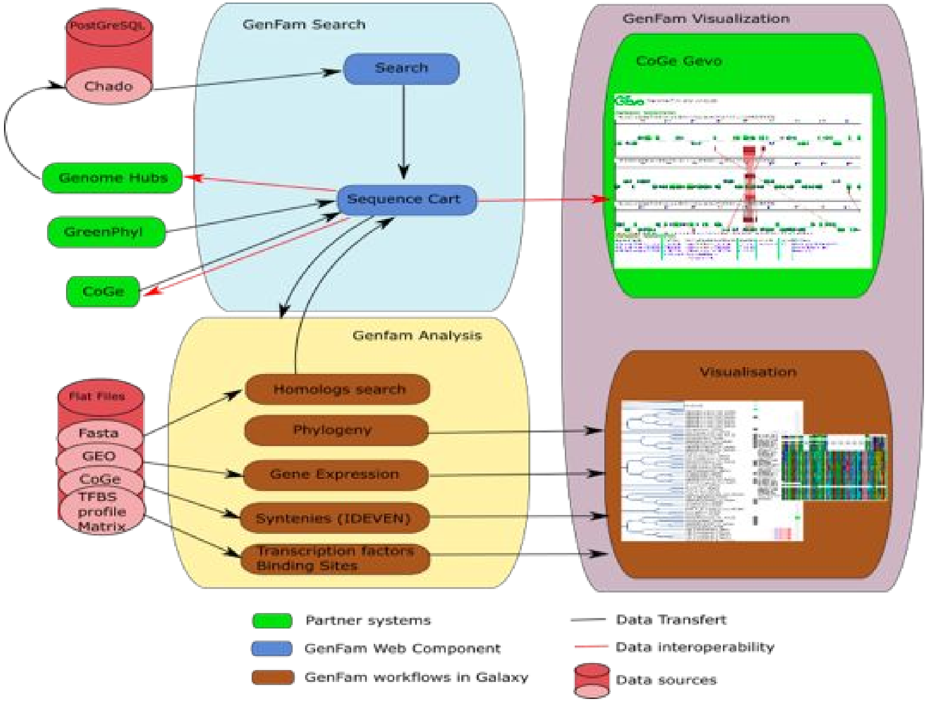
Functional evidences

Visualization

GenFam : Analysis of gene Families



GenFam : Architecture

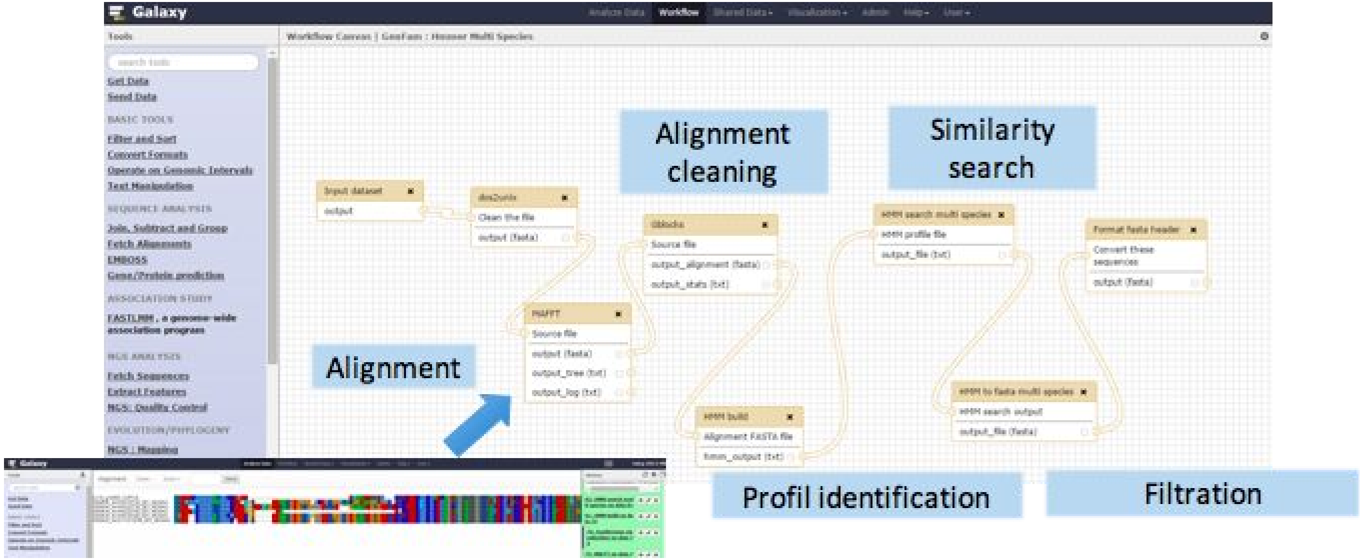


GenFam : Web Portal

- Creation of initial set of sequences
- Import through fasta or json format
  - GreenPhyl families
- Access to Chado Databases
  - Banana
  - Coffee
- Cart of sequences to manage the family

GenFam is available at: http://genfam.southgreen.fr/

GenFam : Homologs search



Alignment cleaning

Similarity search

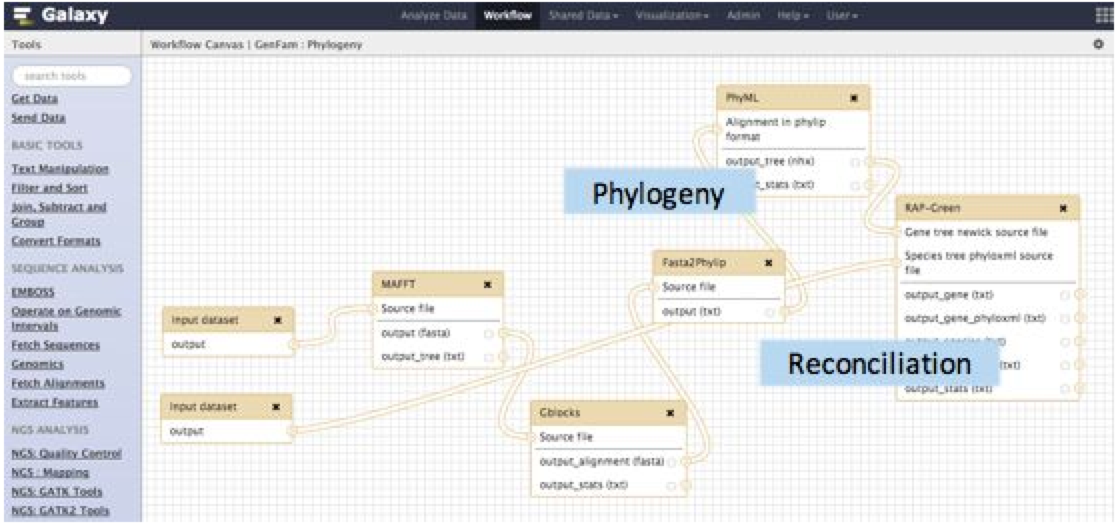
Alignment

Profil identification

Filtration

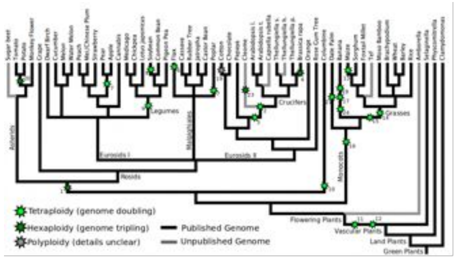
20XX 20XX 20XX 20XX

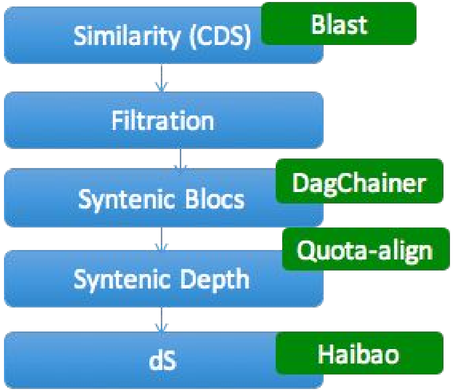
GenFam : Phylogeny



GenFam : Ideven for Synteny Integration

- 25 species
- Synteny identification through CoGe SynMap suite.
- dS Calculation (Haibao Tang)
  - Yang Nielson



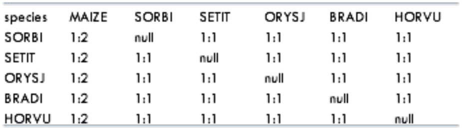

Similarity (CDS) — Blast

Filtration

Syntenic Blocs — DagChainer

Syntenic Depth — Quota-align

dS — Haibao



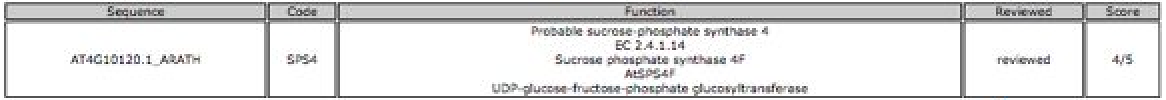
species	MAIZE	SORBI	SETIT	ORYSJ	BRADI	HORVU
SORBI	1:2	null	1:1	1:1	1:1	1:1
SETIT	1:2	1:1	null	1:1	1:1	1:1
ORYSJ	1:2	1:1	1:1	null	1:1	1:1
BRADI	1:2	1:1	1:1	1:1	null	1:1
HORVU	1:2	1:1	1:1	1:1	1:1	null

20XX 20XX 20XX 20XX

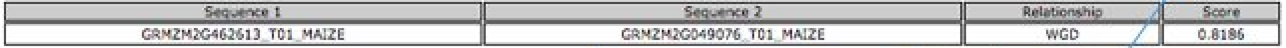
GenFam : Ideven for Synteny Integration

#Nom gene1	Nom gene 2	Event	dS	Mean dS	Block size
AT4G02280.1-PROTEIN_ARATH	BD01_PF61360.1_BRADI	ortholog	4.8013	3.7568285714285707	10
AT5G11110.1-PROTEIN_ARATH	BD03_PF06280.1_BRADI	ortholog	2.4897	2.3304153846153843	6
AT5G11110.1-PROTEIN_ARATH	BD03_PF20090.1_BRADI	ortholog	3.8082	3.1849600000000007	7
AT3G43190.1-PROTEIN_ARATH	GM09_PF06450.1_GLYMA	ortholog	1.4188	2.36896	7
GM13_PF08280.1_GLYMA	GM15_PF15670.1_GLYMA	WGD	0.6661	0.295960409556314 1	146

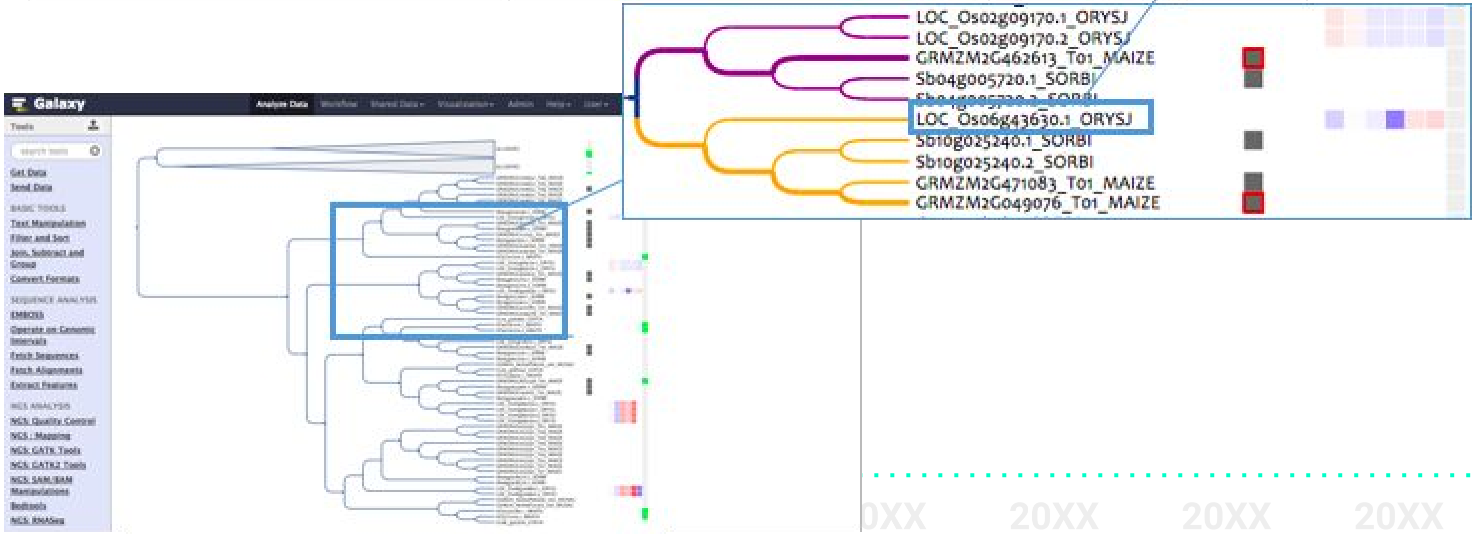
GenFam : Visualization



Sequence	Code	Function	Reviewed	Score
AT4G10120.1_ARATH	SPS4	Probable sucrose-phosphate synthase 4 EC 2.4.1.14 Sucrose phosphate synthase 4F AtSPS4F UDP-glucose-fructose-phosphate glucosyltransferase	reviewed	4/5

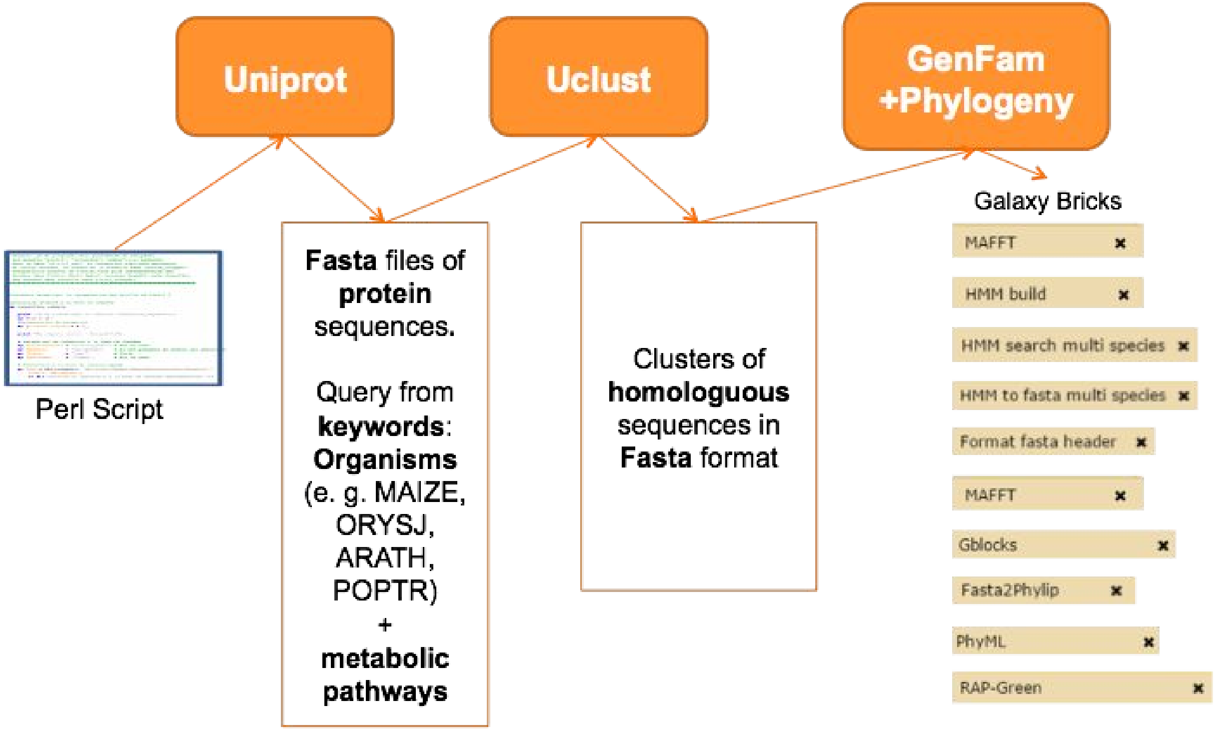


Sequence 1	Sequence 2	Relationship	Score
GRMZM2G462613_T01_MAIZE	GRMZM2G049076_T01_MAIZE	WGD	0.8186



GenesPath: a tool using GenFam, and focused on metabolic pathways

GenesPath: the workflow



Fasta files of **protein** sequences.

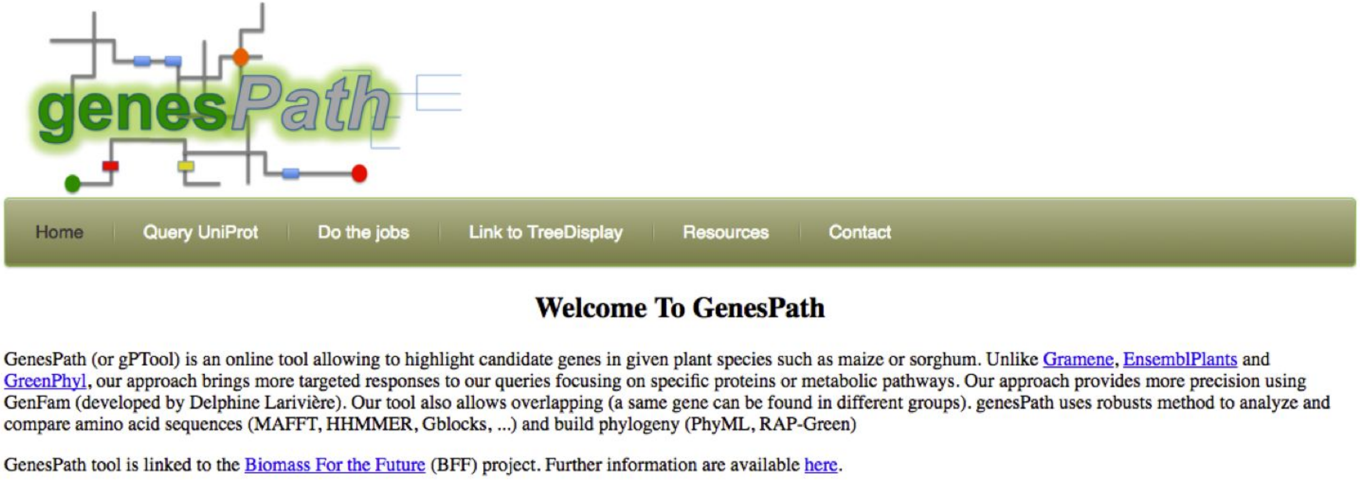
Query from **keywords**: **Organisms** (e. g. MAIZE, ORYSJ, ARATH, POPTR) + **metabolic pathways**

Clusters of **homologuous** sequences in **Fasta** format

Galaxy Bricks

- MAFFT ✖
- HMM build ✖
- HMM search multi species ✖
- HMM to fasta multi species ✖
- Format fasta header ✖
- MAFFT ✖
- Gblocks ✖
- Fasta2Phylip ✖
- PhyML ✖
- RAP-Green ✖

GenesPath: web interface



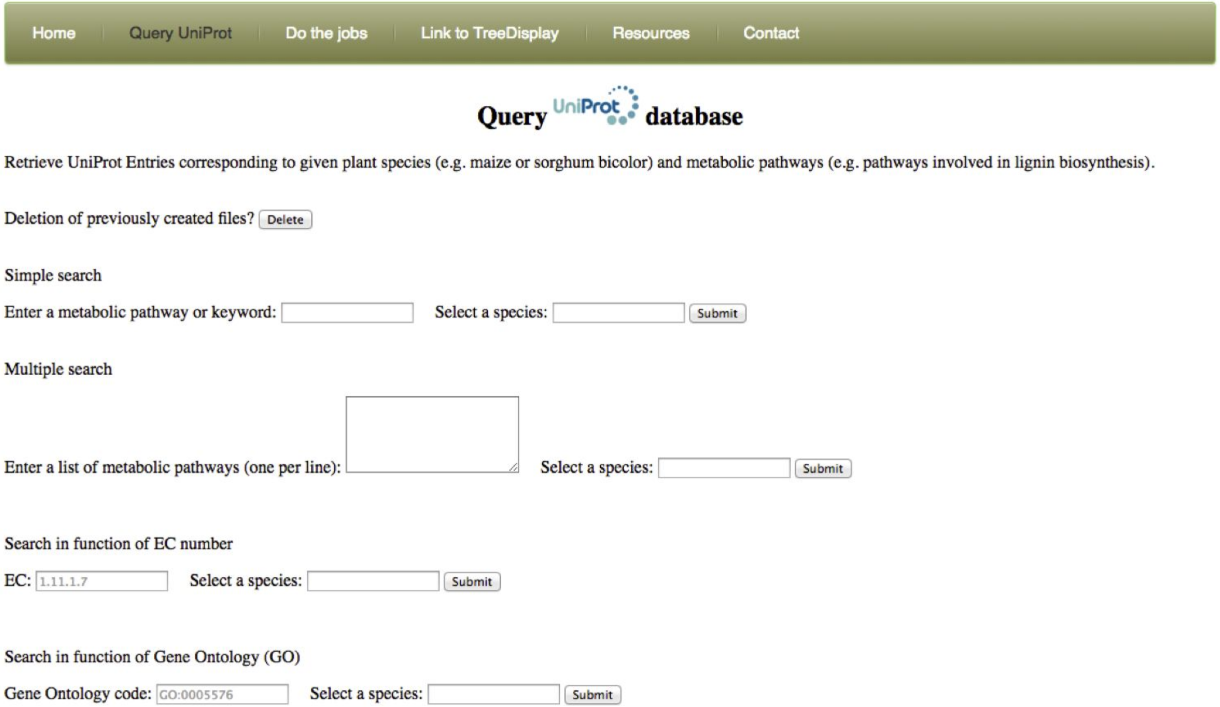
Welcome To GenesPath

GenesPath (or gPTool) is an online tool allowing to highlight candidate genes in given plant species such as maize or sorghum. Unlike Gramene, EnsemblPlants and GreenPhyl, our approach brings more targeted responses to our queries focusing on specific proteins or metabolic pathways. Our approach provides more precision using GenFam (developed by Delphine Larivière). Our tool also allows overlapping (a same gene can be found in different groups). genesPath uses robusts method to analyze and compare amino acid sequences (MAFFT, HHMMER, Gblocks, ...) and build phylogeny (PhyML, RAP-Green)

GenesPath tool is linked to the Biomass For the Future (BFF) project. Further information are available here.

GenFam BFF Biomass For the Future ANR Investissements d'avenir cirad SouthGreen bioinformatics platform

GenesPath: query UniProt database



Query UniProt database

Retrieve UniProt Entries corresponding to given plant species (e.g. maize or sorghum bicolor) and metabolic pathways (e.g. pathways involved in lignin biosynthesis).

Deletion of previously created files?

Simple search

Enter a metabolic pathway or keyword:

Select a species:

Multiple search

Enter a list of metabolic pathways (one per line):

Select a species:

Search in function of EC number

EC: 1.11.1.7

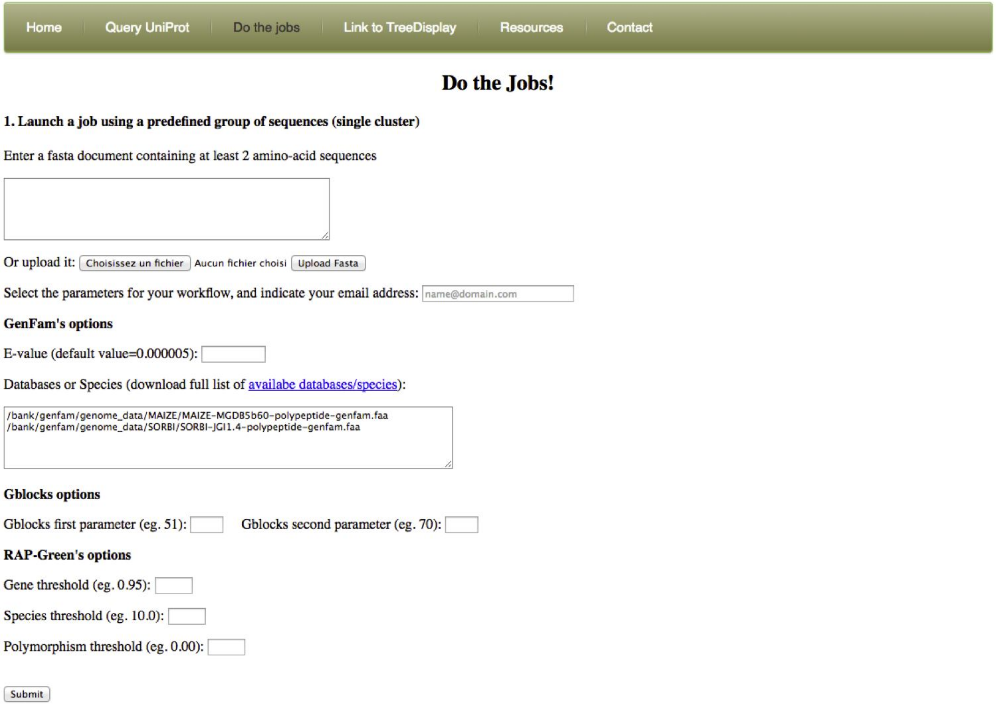
Select a species:

Search in function of Gene Ontology (GO)

Gene Ontology code: GO:0005576

Select a species:

GenesPath: launch workflow through the web



Do the Jobs!

1. Launch a job using a predefined group of sequences (single cluster)

Enter a fasta document containing at least 2 amino-acid sequences

Or upload it:

Select the parameters for your workflow, and indicate your email address:

GenFam's options

E-value (default value=0.000005):

Databases or Species (download full list of availabe databases/species):

```
/bank/genfam/genome_data/MAIZE/MAIZE-MGDB5b60-polypeptide-genfam.faa
/bank/genfam/genome_data/SORBI/SORBI-JGI1.4-polypeptide-genfam.faa
```

Gblocks options

Gblocks first parameter (eg. 51): Gblocks second parameter (eg. 70):

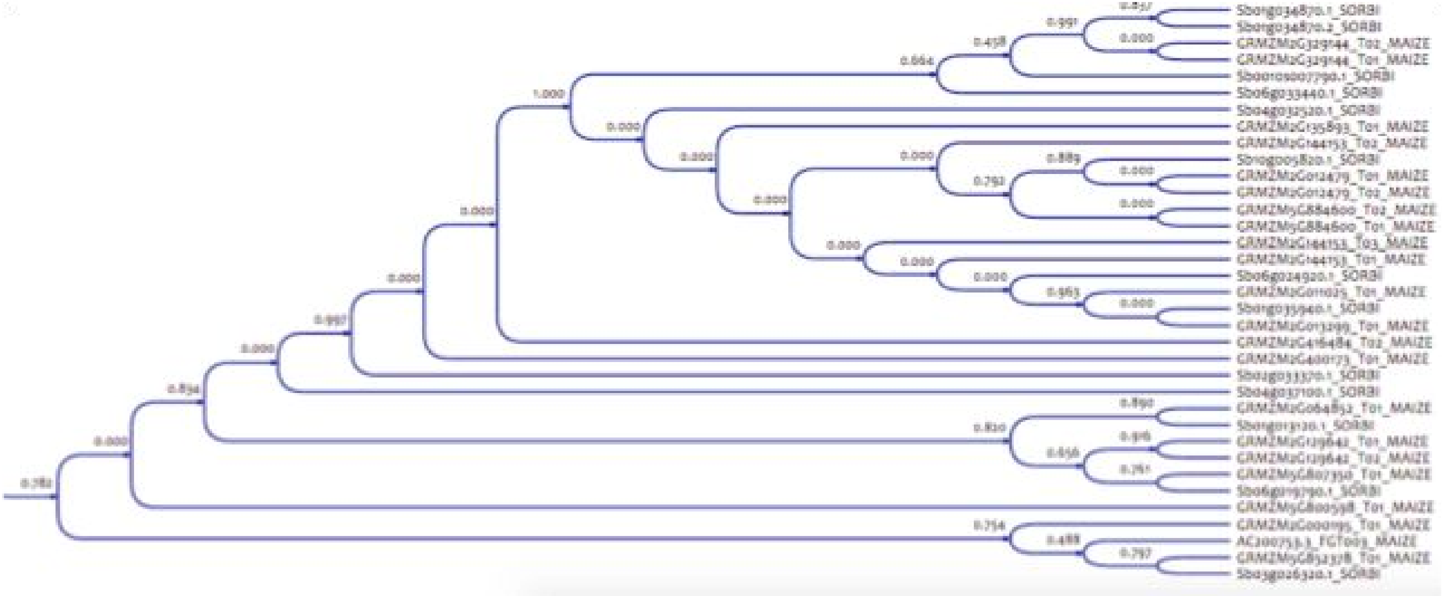
RAP-Green's options

Gene threshold (eg. 0.95):

Species threshold (eg. 10.0):

Polymorphism threshold (eg. 0.00):

GenesPath: phylogenetic tree



GenesPath: CSV files representing orthologous genes

Cluster name	Gene ID_Species	Gene Name	Locus Tag	Chromosome	Start	End	Orthologous group
Cluster 1	OS02G49230.1-PEP_ORYSJ	LOC_OS02G49230.1-PEP	OS02_PF33150.4_ORYSJ	ORYSJ02	30096306	30098580	1
Cluster 1	SB10_PF10650.1_SORBI	SB10G010860.1	SB10_PF10650.1_SORBI	SORBI10	14421774	14424602	1
Cluster 1	ZM05_PF39000.2_MAIZE	GRMZM2G075562_T01	ZM05_PF39000.2_MAIZE	MAIZE05	204171997	204174388	1
Cluster 1	OS07G47140.1-PEP_ORYSJ	LOC_OS07G47140.1-PEP	OS07_PF28660.1_ORYSJ	ORYSJ07	28185202	28187762	2
Cluster 1	ZM07_PF29680.1_MAIZE	GRMZM2G414423_T02	ZM07_PF29680.1_MAIZE	MAIZE07	171822613	171824015	2
Cluster 1	SB02_PF34770.1_SORBI	SB02G042230.1	SB02_PF34770.1_SORBI	SORBI02	75932722	75934993	2

CONCLUSION AND PERSPECTIVES : GENFAM

Web system for manual analysis of gene families

Integration of syntenic analysis through IDEVEN for gene relationship prediction

Integration of heterogeneous data

Synthetic visualization in IntreeGreat

Use of Galaxy API to allow analysis through the GenFam website

Integration of dS based WGD identification

Integration of other types of functional evidences (TFBS)

Integration of syntenic and domains visualization to IntreeGreat

CONCLUSION AND PERSPECTIVES : GENESPATH

Take into account gene families and metabolic pathways

Link with a specific project named “Biomass For the Future” (BFF)

GenFam and GenesPath are complementary tools

Launch the workflow of analysis without using Galaxy

Integration of phylogenetic analysis and statistics on genes (CSV) using IntreeGreat

GenesPath will be improved (it does not contain yet the IDEVEN tool)

THANK YOU FOR YOUR ATTENTION!

genfam.southgreen.fr