Integrative System For Gene Family Gathering and Analysis In A Context of Crops' Stress Response Study

Delphine Larivière[A]*, Jean-François Dufayard[A],
Stéphanie Bocs[A], Dominique This[B]

(A) CIRAD
UMR AGAP
Montpellier, FRANCE

(B) Montpellier SupAgro
UMR AGAP
Montpellier SupAgro

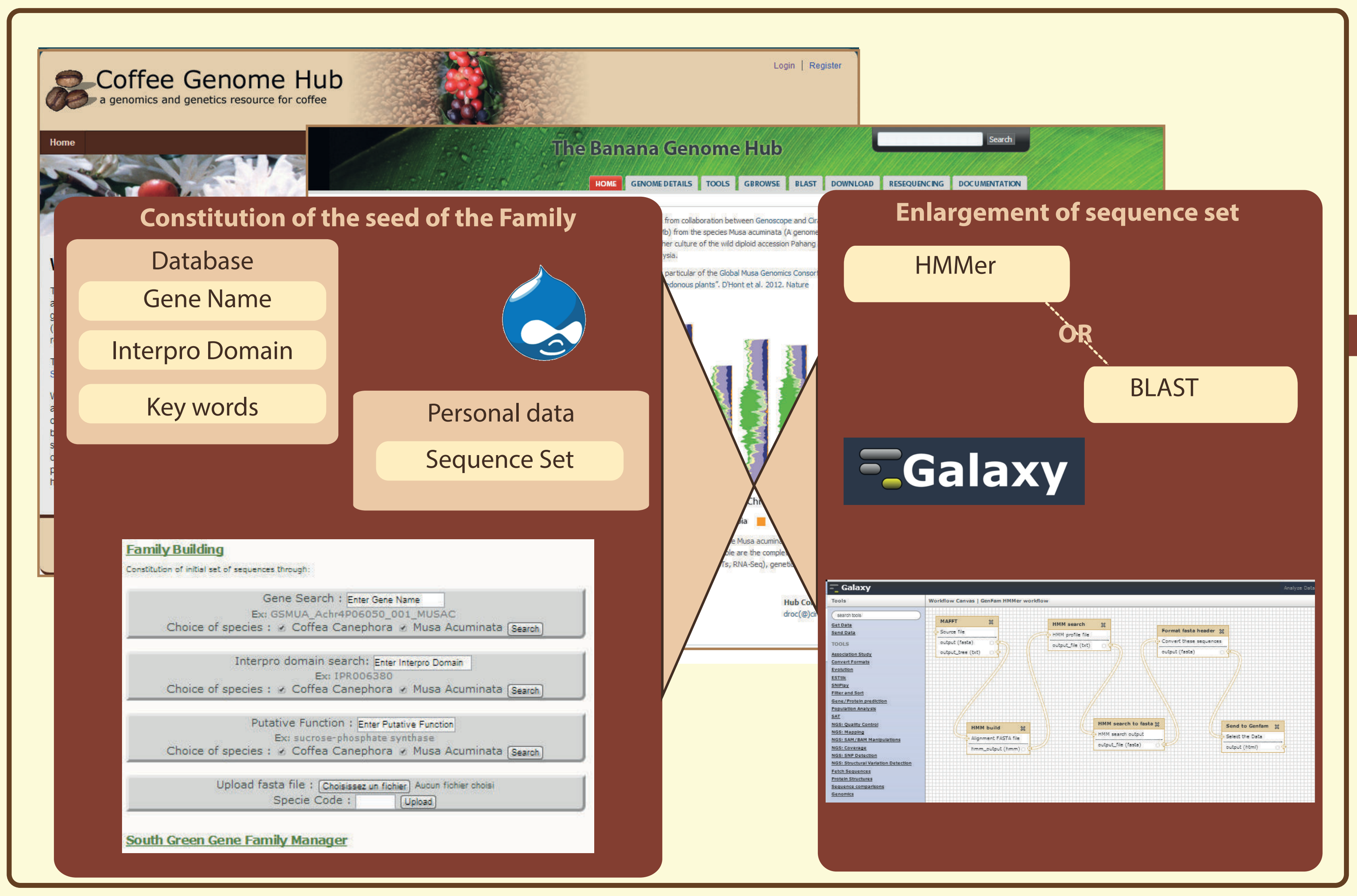
* Corresponding Author: Email: delphine.lariviere@cirad.fr

Gene family analysis is an important way to understand complex processes underlying stress response in crops. Several tools exist to study families and propose automatically clustered families or curated published families but automatic clustering is rarely sufficient for precise studies.
Biologists need most of the time to manually constitute their families.
We propose to develop an integrative system that will allow to gather sequences and informations from different sources for the analysis and visualisation of a customized family as a synthetic and dynamic view.
This work is part of a PhD project and will eventually be integrated in the SouthGreen[1] bioinformatics platform.
http://www.southgreen.fr/

Keywords : Gene family, data integration, comparative genomics, data visualization

1 Family Constitution

For a customized gene family, the constitution is done in two steps: the constitution of the seed of the family, and the enlargement of the sequence set. The seed set can be selected through gene name, interpro domain, or keywords search. The databases that are queried for the moment are Chado databases[2] of *Musa acuminata* and *Coffea canephora*. To enlarge the datasets to similar sequences in other species, the tool is linked to HMMer and BLAST workflows in Galaxy workflow manager[3].



Constitution of the seed of the Family

Database
- Gene Name
- Interpro Domain
- Key words

Personal data
- Sequence Set

Family Building
Constitution of initial set of sequences through:

Gene Search : Enter Gene Name
Ex: GSMUA_Achr4P06050_001_MUSAC
Choice of species : ☑ Coffea Canephora ☑ Musa Acuminata [Search]

Interpro domain search: Enter Interpro Domain
Ex: IPR006380
Choice of species : ☑ Coffea Canephora ☑ Musa Acuminata [Search]

Putative Function : Enter Putative Function
Ex: sucrose-phosphate synthase
Choice of species : ☑ Coffea Canephora ☑ Musa Acuminata [Search]

Upload fasta file : [Choisissez un fichier] Aucun fichier choisi
Specie Code : [Upload]

South Green Gene Family Manager

Enlargement of sequence set

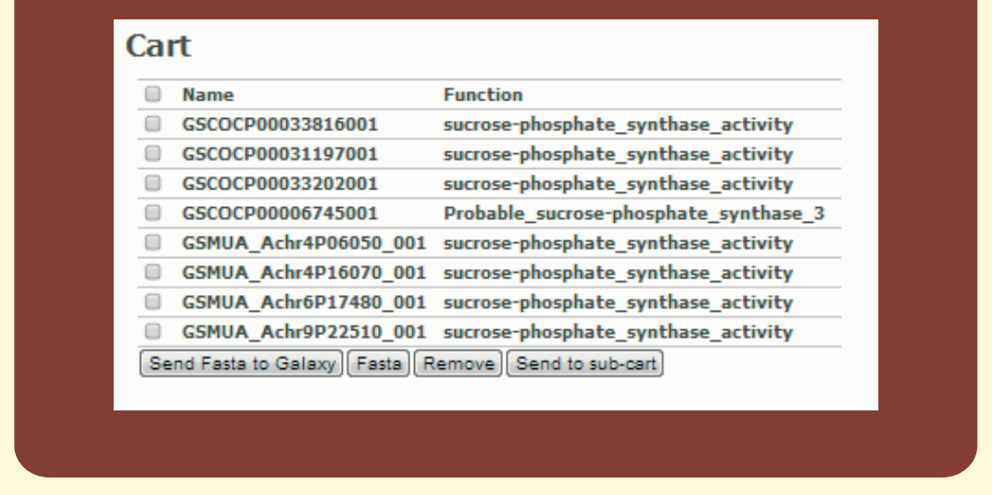
HMMer
OR
BLAST

Galaxy

Constitution of a «cart» of sequences:

Add sequences
Remove sequences
Download Fasta

At any moment



Cart

☐	Name	Function
☐	GSCOCP00033816001	sucrose-phosphate_synthase_activity
☐	GSCOCP00031197001	sucrose-phosphate_synthase_activity
☐	GSCOCP00033202001	sucrose-phosphate_synthase_activity
☐	GSCOCP00006745001	Probable_sucrose-phosphate_synthase_3
☐	GSMUA_Achr4P06050_001	sucrose-phosphate_synthase_activity
☐	GSMUA_Achr4P16070_001	sucrose-phosphate_synthase_activity
☐	GSMUA_Achr6P17480_001	sucrose-phosphate_synthase_activity
☐	GSMUA_Achr9P22510_001	sucrose-phosphate_synthase_activity

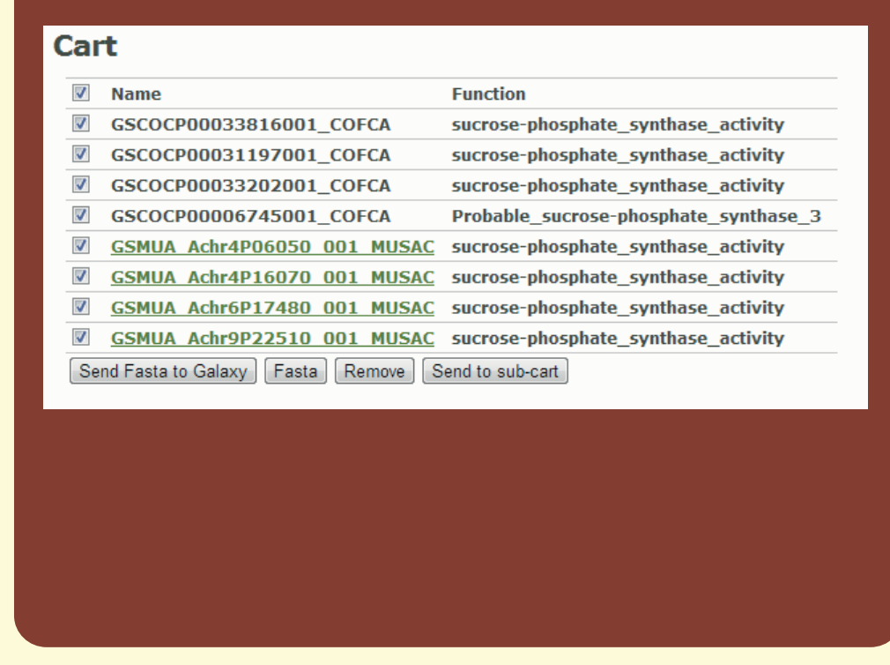
[Send Fasta to Galaxy] [Fasta] [Remove] [Send to sub-cart]

2 Family Analysis

The tool integrates several tools for gene family analysis. These tools are GMOD components or are implemented in Galaxy workflow Manager. The tool is also linked to CoGe[4] the Banana Genome Hub[5] and the Coffee Genome Hub[6] in order to get functional information on sequences.

Selected Family

Send the designed family to analysis tools



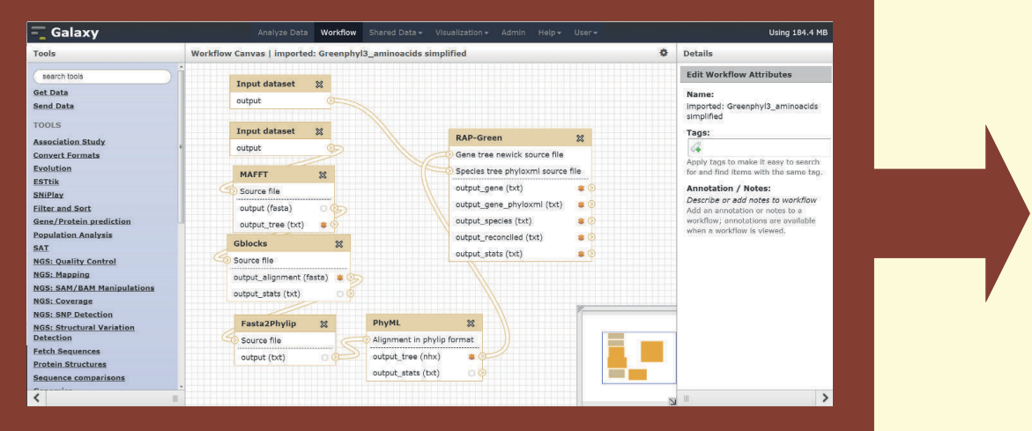
Cart

☑	Name	Function
☑	GSCOCP00033816001_COFCA	sucrose-phosphate_synthase_activity
☑	GSCOCP00031197001_COFCA	sucrose-phosphate_synthase_activity
☑	GSCOCP00033202001_COFCA	sucrose-phosphate_synthase_activity
☑	GSCOCP00006745001_COFCA	Probable_sucrose-phosphate_synthase_3
☑	GSMUA_Achr4P06050_001_MUSAC	sucrose-phosphate_synthase_activity
☑	GSMUA_Achr4P16070_001_MUSAC	sucrose-phosphate_synthase_activity
☑	GSMUA_Achr6P17480_001_MUSAC	sucrose-phosphate_synthase_activity
☑	GSMUA_Achr9P22510_001_MUSAC	sucrose-phosphate_synthase_activity

[Send Fasta to Galaxy] [Fasta] [Remove] [Send to sub-cart]

Analyse

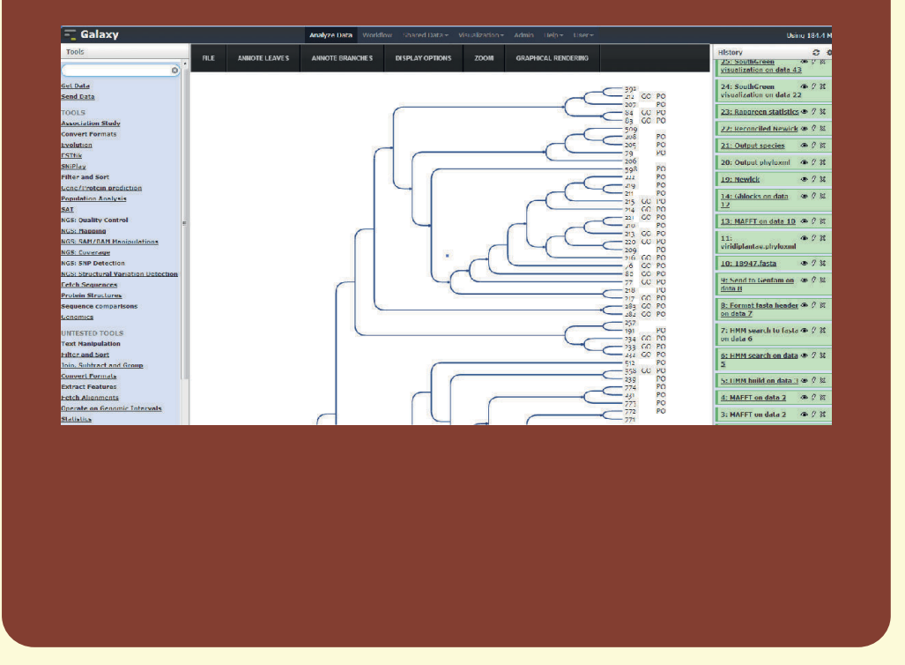
Run Analysis in Galaxy or in Gmod Tools



GMOD Galaxy

Result Visualisation

Visualise the results in Galaxy



Other Functionalities (To be done)

History

An history of analysis and family modifications wil be created so that the user can go back at any moment and keep a track of analysis parameters and why he added or removed a sequence.

Specific Orientation (stress response)

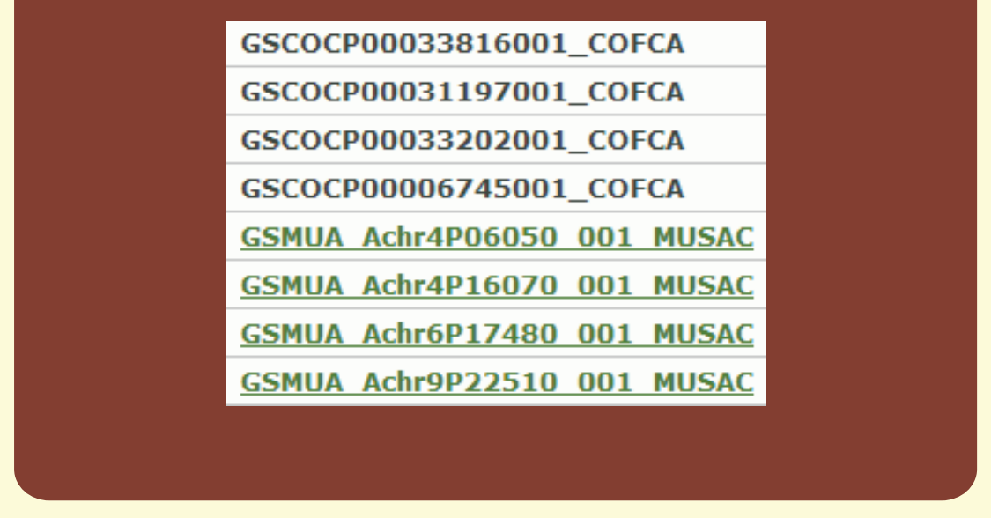
Stress response specific information will be linked to the tool with controlled vocabulary, known annotations, bibliography. These informations will come from stress response specific databases.

3 Data Visualization (To be done)

Visualization is a crucial step for data understanding. The tool will display an integrative and dynamic view of the results of the gene family analysis and information about the members of the family. The integrative visualization will allow interaction between differents views that will contain for example phylogenetic tree, syntenic dotplot, pathway information, genome browser, evidences for sress response...

Members of the family

Display each member of the family and their structural annotation and allow to select one or more sequences and to remove an unwanted element.



GSCOCP00033816001_COFCA
GSCOCP00031197001_COFCA
GSCOCP00033202001_COFCA
GSCOCP00006745001_COFCA
GSMUA_Achr4P06050_001_MUSAC
GSMUA_Achr4P16070_001_MUSAC
GSMUA_Achr6P17480_001_MUSAC
GSMUA_Achr9P22510_001_MUSAC

Functionnal and relationnal annotation

Display informations about selected sequence(s):
Pathway, Annotation

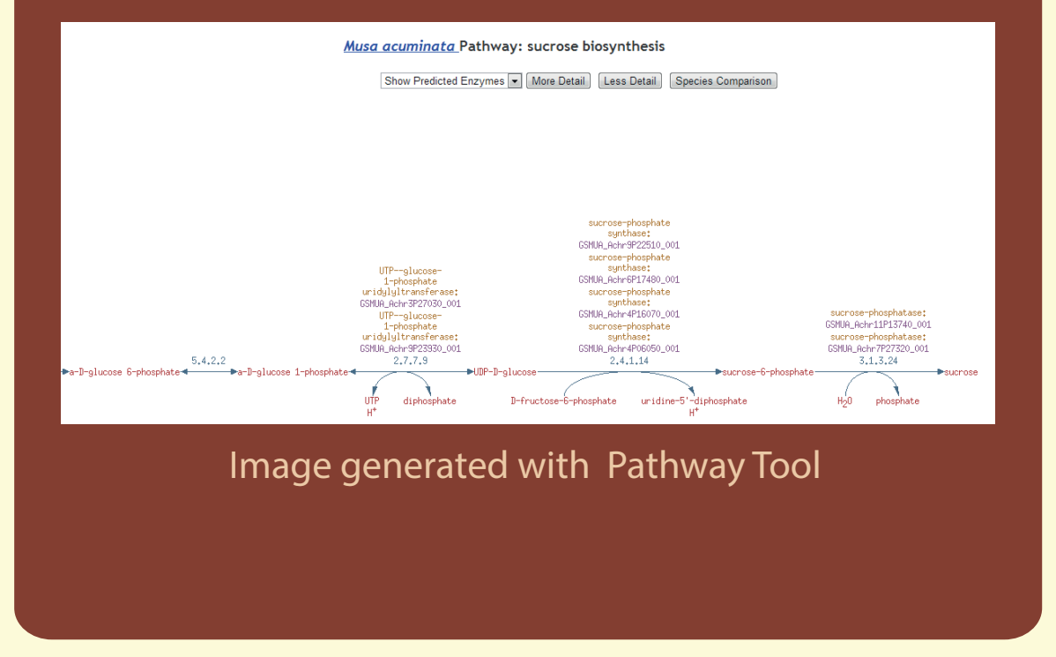


Image generated with Pathway Tool

Genome Browser

Display tracks allowing to visualise gene structure, genomic environnement and further informations if available : SNPs, ESTs ...

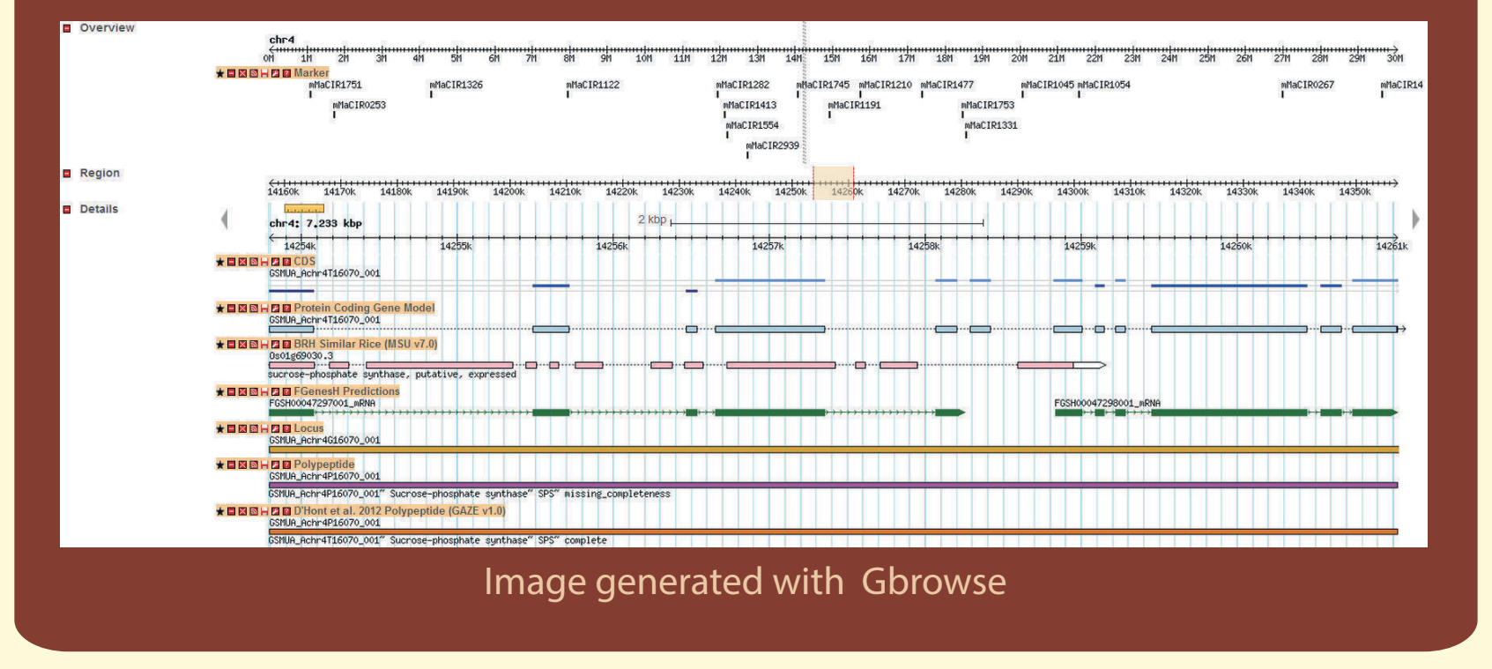


Image generated with Gbrowse

Phylogeny

Phylogenetic tree allowing te visualisation of other data available such as expression data, domains, gene structure

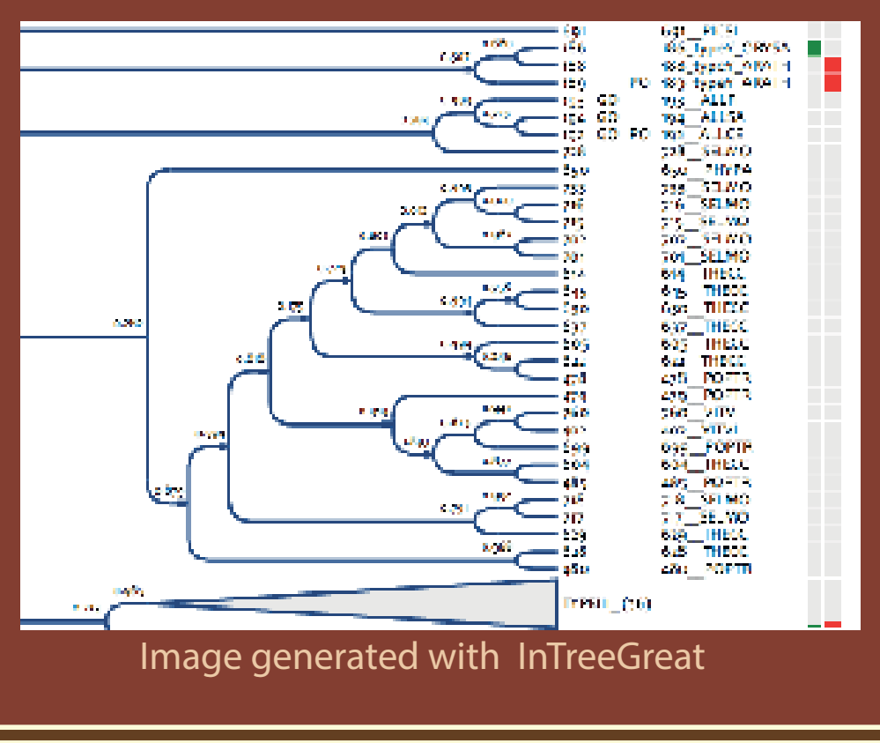
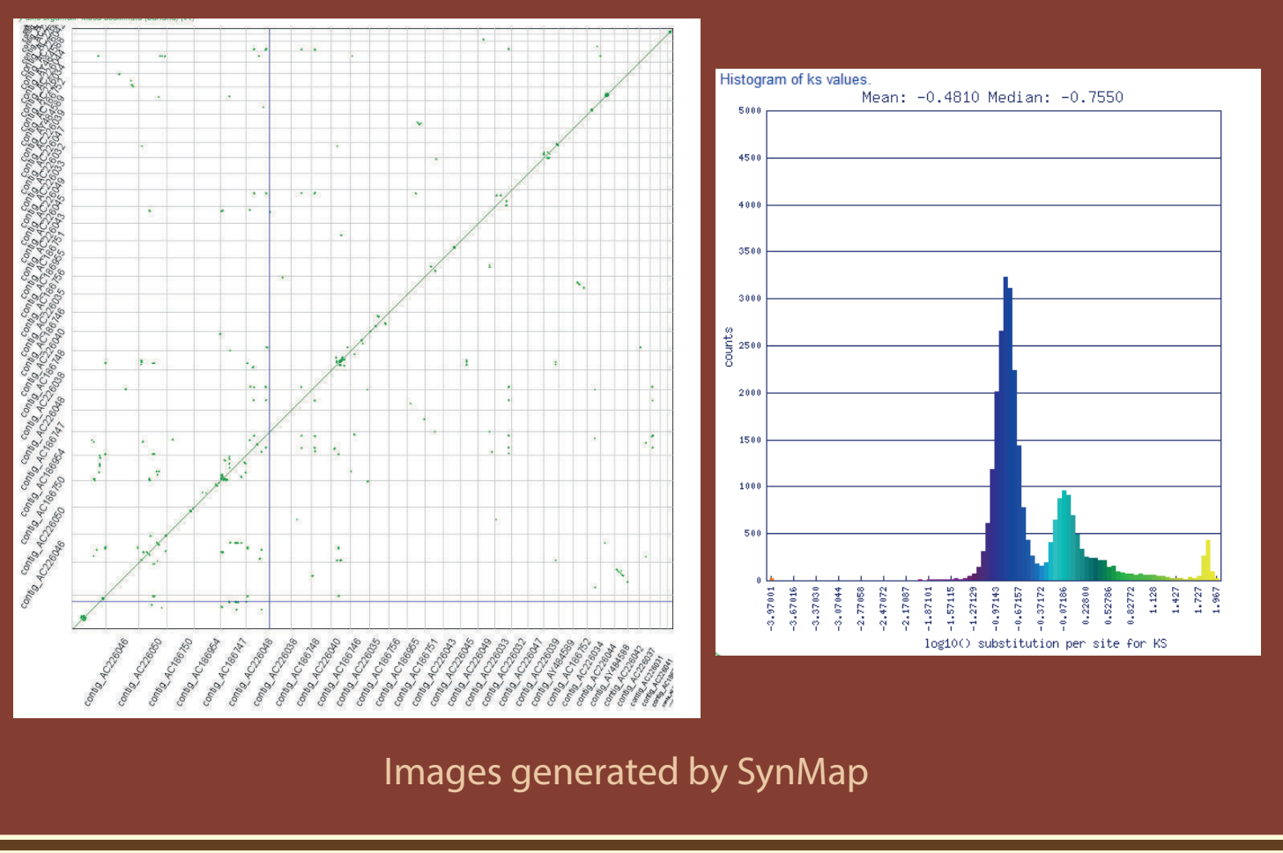


Image generated with InTreeGreat

Synteny

Graph and dot plot allowing tho visualize the syntheny of the region around our sequences of interest.



Images generated by SynMap

(1) Rouard et al. The South Green Bioinformatics Platform #P988 , PAG XXII
(2) Mungall et al. 2007 Bioinformatics,
(3) Goecks et al. 2010 Genome Biol.
4) Lyons et al . 2008 The Plant Journal
Lyons et al. 2008 Plant Phys.
(5) Droc, Larivière et al. 2013 Database: the journal of biological databases and curation
(6) Dereeper et al., Coffee Genomics #2125, PAG XXII

Montpellier SupAgro

cirad

South Green bioinformatics platform