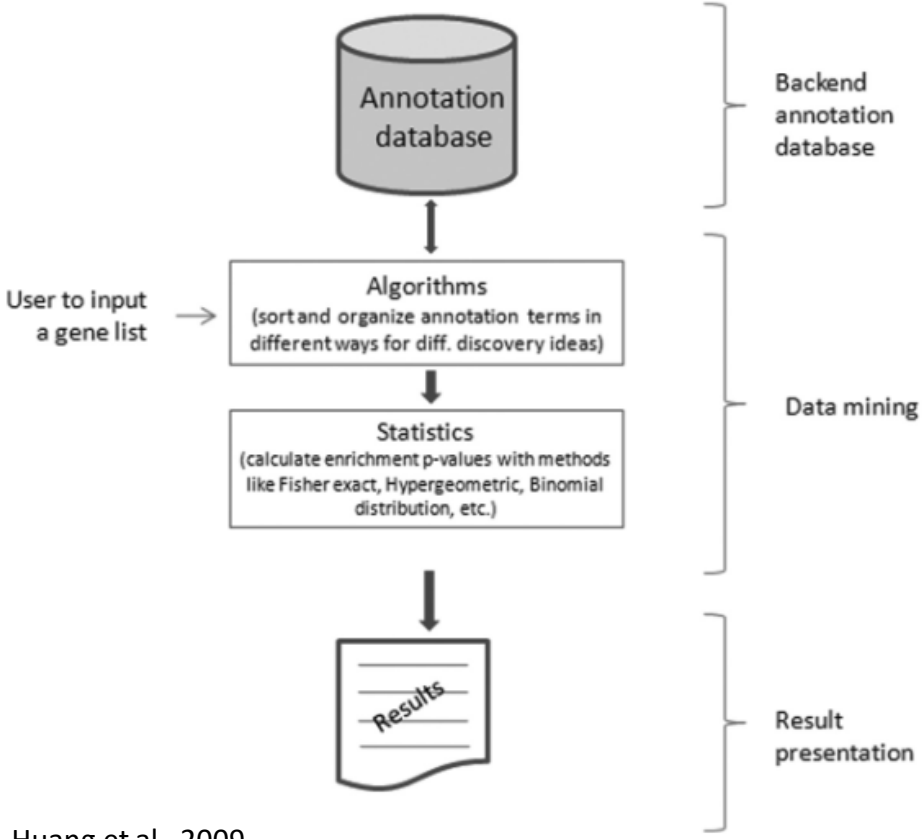
GOCOmpare: an R package to compare functional enrichment analysis results among gene lists and two species

April 1st, 2022

Chrystian C. Sosa, Diana Carolina Clavijo-Buriticá, Nicolas López-Rozo, Camila Riccio-Rengifo, Maria Victoria Diaz, Victor Hugo García Merchán, Mauricio Alberto Quimbaya

Enrichment analysis is a cornerstone in bioinformatics

How an enrichment analysis works?



Huang et al., 2009

Multiple gene lists enrichment comparison is descriptive!

How multiple gene lists are compared?

GOCompare: https://cran.r-project.org/package=GOCompare
https://github.com/ccsosa/GOCompare

Six functions in three categories:

- **Descriptive**
- **Pre-processing**
- **Graphs approach**

Enriched GO term **Functional category (Gene list)**

Inputs:

	Functional_Category	feature
1	Response to stress	AID
2	Defense response	AID
3	Immune response	AID
4	Regulation of defense response	AID
5	Immune effector process	AID
6	Activation of immune response	AID

GOCompare

Descriptive
Three functions:

- *mostFrequentGOs*

GO	freq	features
Response to stress	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Regulation of response to stimulus	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Regulation of response to stress	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Response to organic substance	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI
Multi-organism process	10	AID;AIM;DCE;ERI;EGS;GIM;IA;RCD;SPS;TPI

- *evaluateCAT_species and evaluateGO_species:*
  - *Z- proportion tests*
  - *Chi-squared tests*

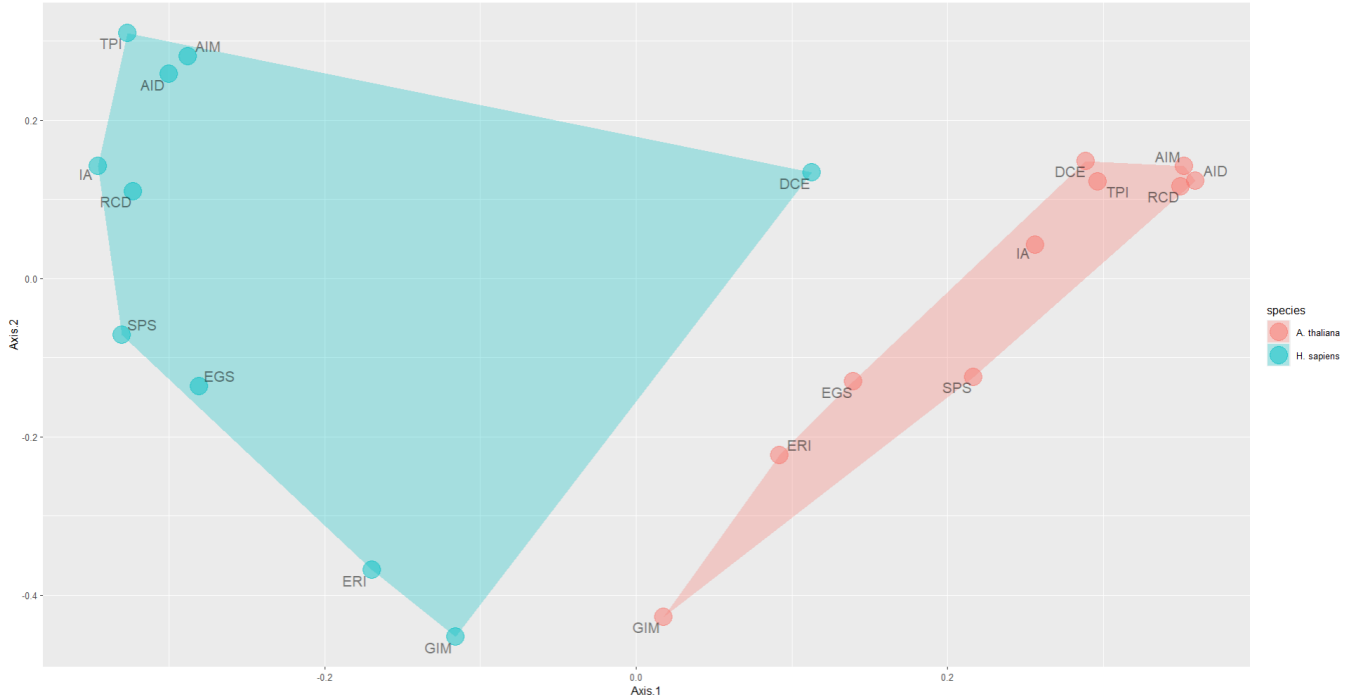
CAT	pvalue	FDR
IA	5.242296e-25	5.242296e-24
EGS	1.979177e-20	9.895884e-20
TPI	4.723417e-05	1.574472e-04
AID	5.581390e-04	6.201545e-04

* Functions designed for two species

GOCompare

- *compareGOspecies**
  - Jaccard
  - PCoA
  - Unique and shared GO

Pre-processing
One function:

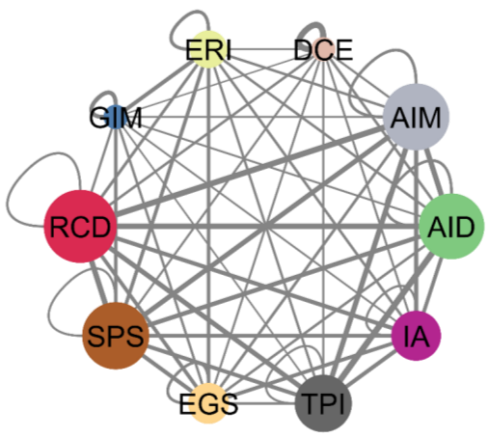


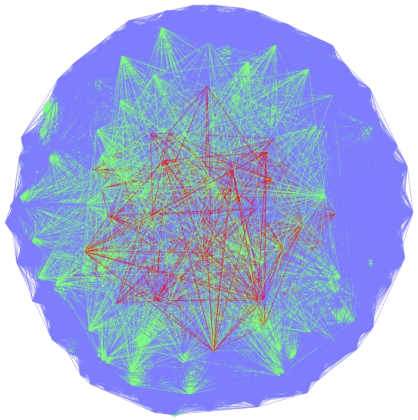
* Functions designed for two species

GOCompare

Graph approach two functions:

- *graphGOspecies and graph_two_Gospecies* :*
  - *Categories* (gene lists)
  - *GO*





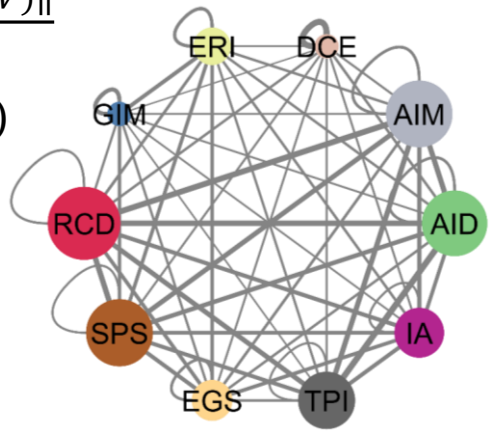
GO	GO_WEIGHT
Regulation of response to stimulus	0.5178571
Positive regulation of response to stimulus	0.5178571
Regulation of protein metabolic process	0.5178571
Positive regulation of cellular metabolic process	0.5178571
Positive regulation of metabolic process	0.5178571
Regulation of cellular protein metabolic process	0.5178571

* Functions designed for two species

Undirected-weighted graph approach

Use of Categories as nodes

- $G = \{V, E\}$
- $E = \{e = (U, V)\ BP \neq \emptyset\}$
- $w(e) = \frac{\|BP(U) \cap BP(V)\|}{\|nBP\|}$
- $K_w(U) = \sum w(U, V)$



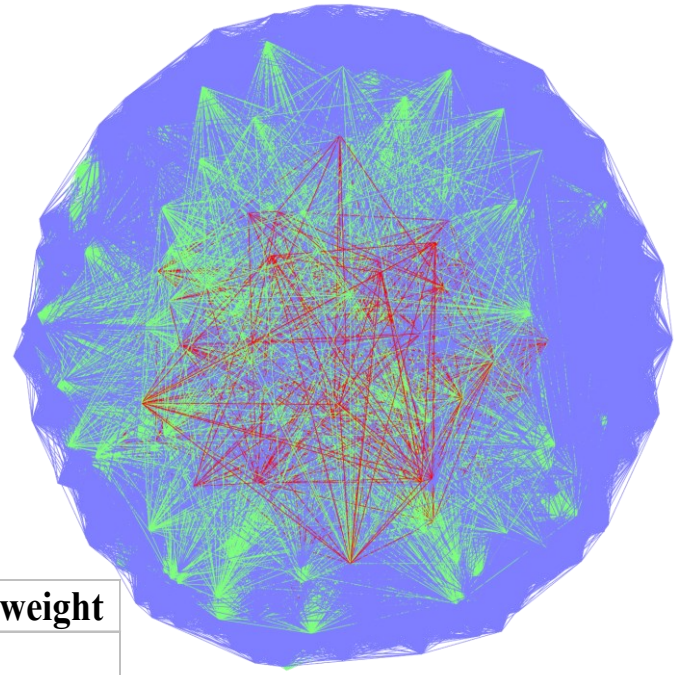
GO terms= BP

Category	Node weight
RCD	0.988
AIM	0.981
AID	0.977

GO terms as nodes

- $G = \{e = (U, V)\ \mathrm{Cat}(U) \cap \mathrm{Cat}(V) \neq \emptyset\}$
- $Cat(U) = n\ U \in M\ ;\ Cat(V) = n\ V \in M$
- $w(e) = \frac{\|Cat(U) \cap Cat(V)\|}{\|nBP\|}$
- $K_w(U) = \sum w(U, V)$

Functional groups= Cat



GO term	Node weight
Response to stress	2.494
Regulation of response to stimulus	2.494
Regulation of response to stress	2.494

A case of study: Aluminum tolerance in *Oryza sativa* subsp. *japonica*

- How functionally different are *Oryza sativa* subsp. *japonica* (OSJ) and *A. thaliana* (ARATH) for aluminum tolerance?

- What are the OSJ genes involved?
- What are the main molecular functions of proteins involved?

Literature search for aluminum tolerance genes:

- Genome-wide association studies
- Quantitative trait loci
- Transcriptomics
- QTL and SSR

Filtering :

- Orthologs using g:Orth
- Limit to OSJ genes with ARATH orthologues
- Limit to ARATH genes with GO:MF
- Removing GO:BP (“biological_process”) GO:MF (“molecular_function”)

Functional groups (gene lists based on biological processes) :

- (elim + Kolmogorov–Smirnov) + custom background
- Refinement groups using REVIGO
- Functional groups assignment for ARATH ortholog genes
- Functional enrichment (GO:MF)

g:Profiler

topGO

- How are connected the different *biological processes in terms of molecular functions*?
- How different are the *species in terms of molecular functions*?
- What are the main genes among *biological processes*?

Undirected weighted graphs using GOCOmpare (OSJ, ARATH, BOTH):

-GO terms graphs

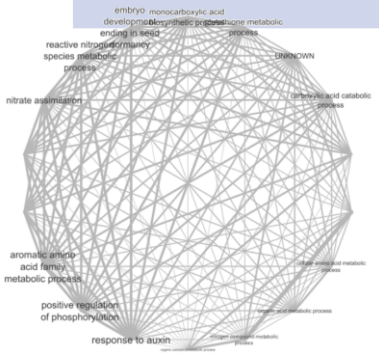
- Node weights >90th percentile

GO Jaccard distances+ clustering

Treemaps

Proportion tests for genes and MF

Undirected graphs for OSJ and ARATH genes and functional groups



Cytoscape

$$G = \{V, E\}$$

$$E = \{e = (U, V) \mid Cat(U) \cap Cat(V) \neq \emptyset\}$$

$$Cat(U) = nU \in M; Cat(V) = nV \in M$$

$$w(e) = \frac{\|Cat(U) \cap Cat(V)\|}{\|nBP\|}$$

$$K_w(U) = \sum w(U, V)$$

ccsosa/

GOCompare

This is an R package designed to analyze functio enrichment results between two species (A list of genes and their...

1 Contributor | 1 Issue | 0 Stars | 0 Forks

Prop.test

Gene summary

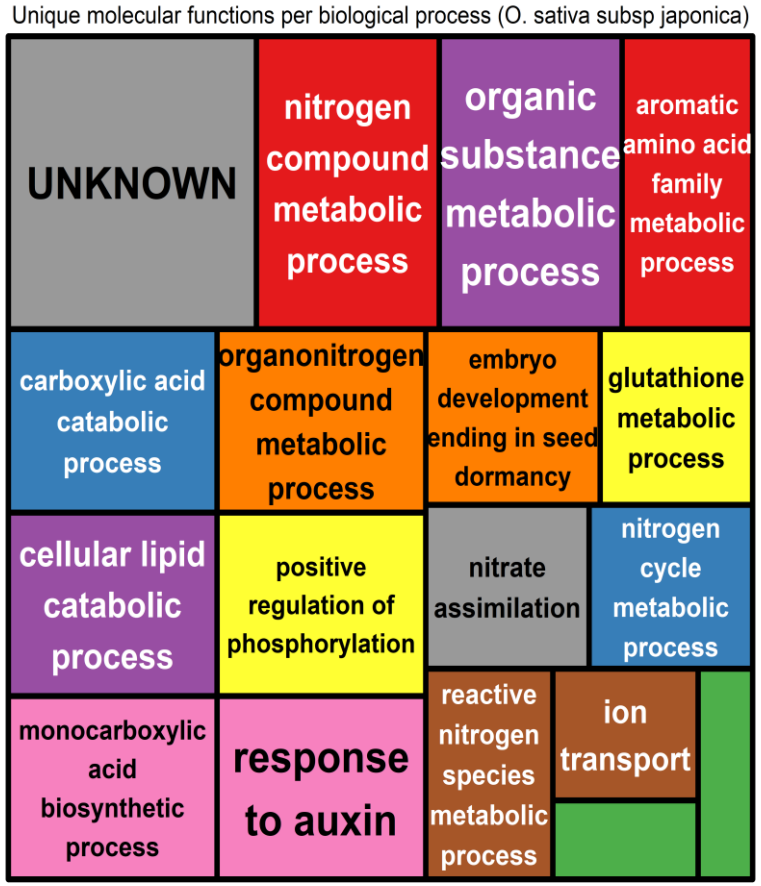
- 1198 OSJ genes
- 1267 ARATH ortholog genes
- Genes used for enrichment:
  - Unique OSJ genes: 507
  - Unique ARATH genes: 985
- 18 functional groups

Groups	OSJ	ARATH
organic substance metabolic process	261	539
UNKNOWN	212	1114
nitrogen compound metabolic process	189	402
organonitrogen compound metabolic process	133	304
cellular amino acid metabolic process	69	138
organic acid metabolic process	69	138
ion transport	37	72
carboxylic acid catabolic process	15	28
monocarboxylic acid biosynthetic process	15	27
glutathione metabolic process	9	30
cellular lipid catabolic process	6	9
response to auxin	5	9
positive regulation of phosphorylation	4	9
reactive nitrogen species metabolic process	3	4
nitrogen cycle metabolic process	3	4
aromatic amino acid family metabolic process	3	11
nitrate assimilation	3	4
embryo development ending in seed dormancy	3	3

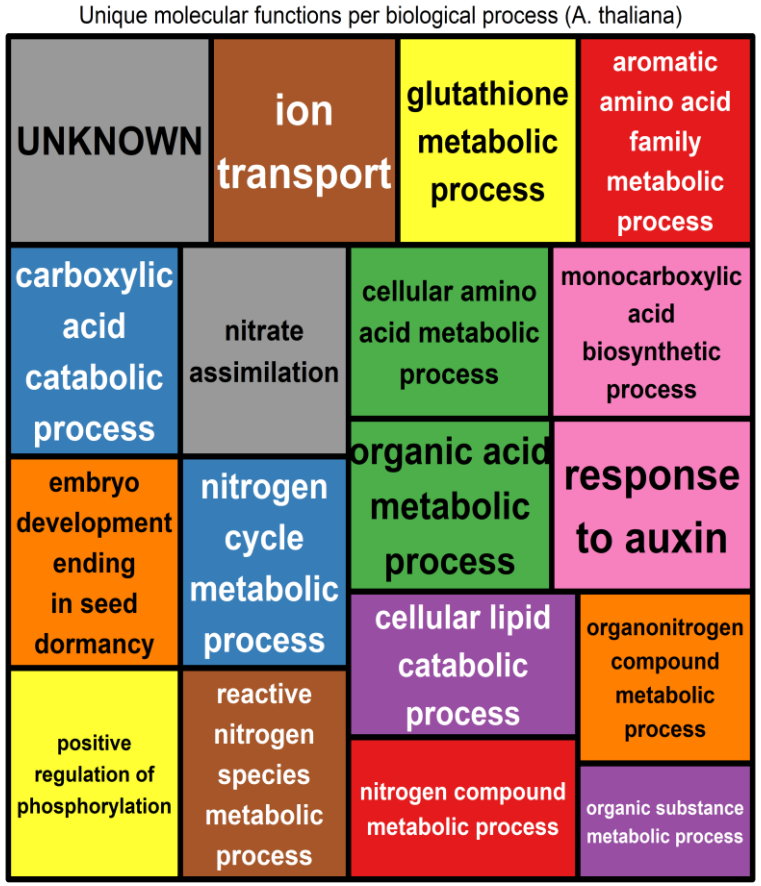
compareGOspecies + treemap

Oryza sativa subsp. *japonica* *Arabidopsis thaliana* Both species

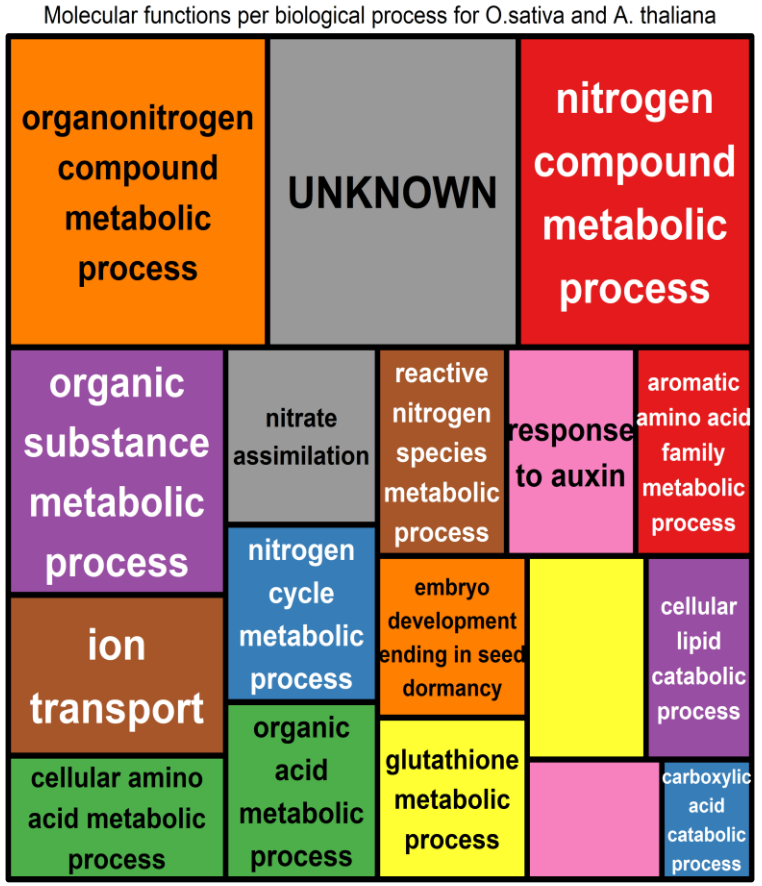
A.



B.



C.



Similar proportions of enriched terms

evaluateCAT_species + evaluateGO_species

- **Different response to aluminum in terms of molecular functions?**

Four functional groups FDR <0.05

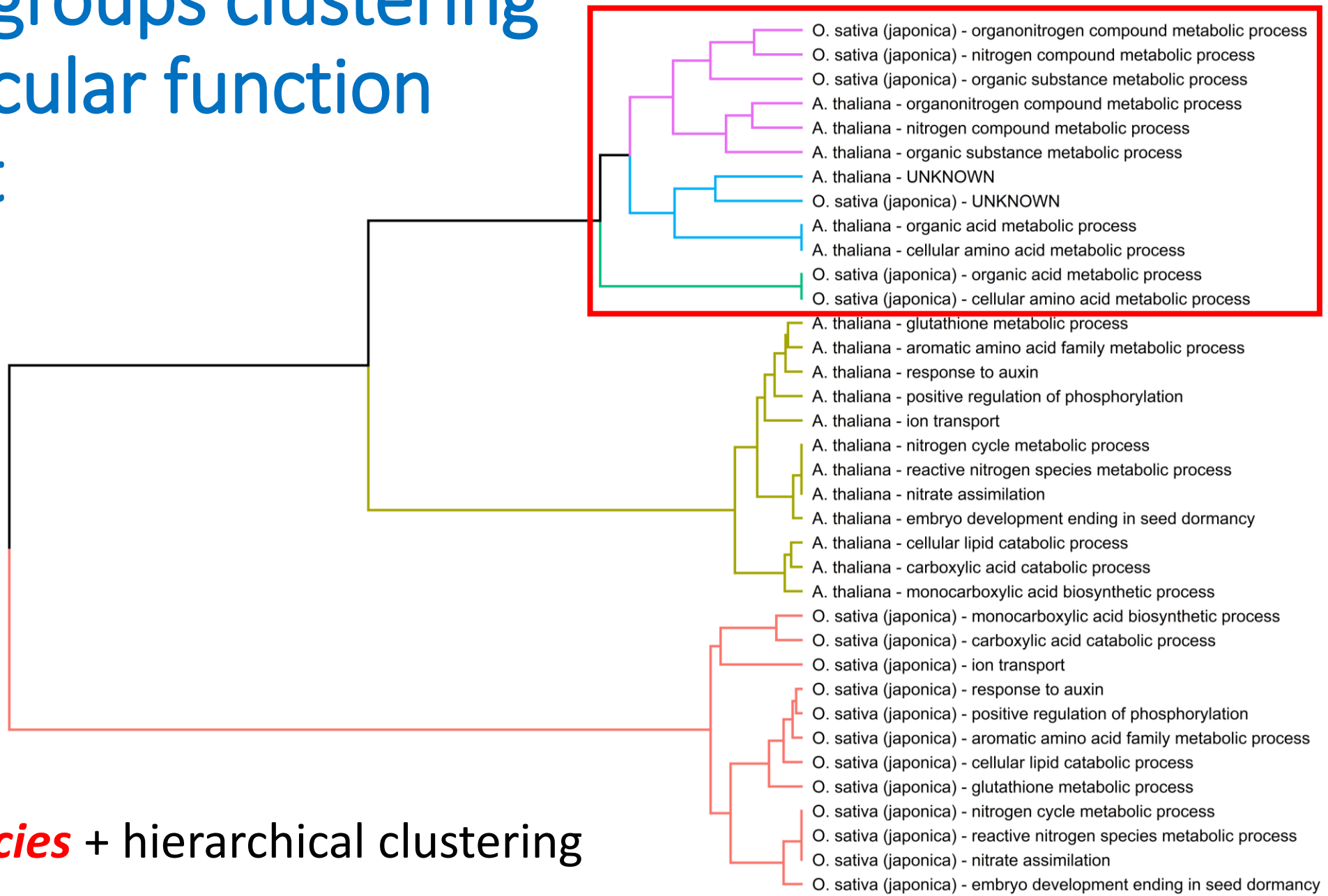
CAT
organic substance metabolic process
nitrogen compound metabolic process
organonitrogen compound metabolic process
UNKNOWN

20 Molecular functions FDR <0.05

GO
primary active transmembrane transporter activity
identical protein binding
vitamin binding
vitamin B6 binding
glutathione binding
transferase activity, transferring alkyl or aryl (other than methyl) groups
4 iron, 4 sulfur cluster binding
transferase activity
ATPase-coupled transmembrane transporter activity
ion transmembrane transporter activity
endopeptidase activity
metal cluster binding
iron-sulfur cluster binding
glutathione transferase activity
lysophospholipid acyltransferase activity
hydro-lyase activity
pyridoxal phosphate binding
hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides
carbon-oxygen lyase activity
organic acid transmembrane transporter activity

Functional groups clustering using Molecular function enrichment

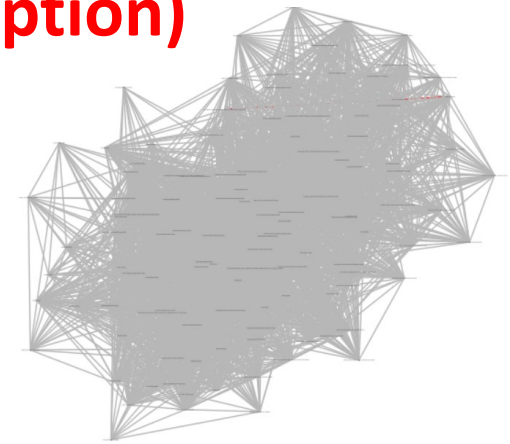
There are functional groups similar between species!

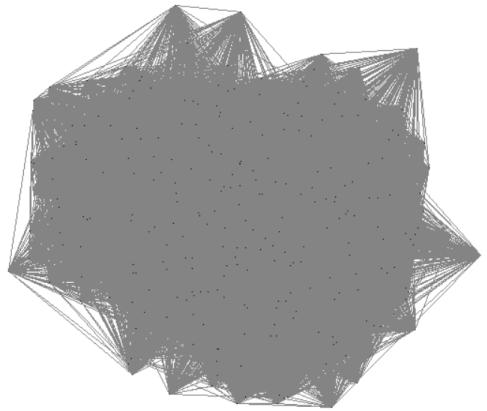


compareGOspecies + hierarchical clustering

graphGOspecies (GO option)

Molecular functions related with signaling and transporters!





Molecular functions related with lignin and ion binding processes!

Top 10 OSJ MF node weights

GO	WEIGHT
ATPase-coupled transmembrane transporter activity	2.76316
hydro-lyase activity	2.42982
lysophospholipid acyltransferase activity	2.34211
endopeptidase activity	2.34211
primary active transmembrane transporter activity	2.26316
glutathione binding	2.21053
glutathione transferase activity	2.21053
vitamin binding	2.13158
identical protein binding	2.0614
vitamin B6 binding	2.01754

Top 10 ARATH MF node weights

GO	WEIGHT
glucan endo-1,3-beta-D-glucosidase activity	6.15472
lysophosphatidic acid acyltransferase activity	6.15472
1-acylglycerol-3-phosphate O-acyltransferase activity	6.15472
lysophospholipid acyltransferase activity	6.15472
acylglycerol O-acyltransferase activity	6.15472
hydroquinone:oxygen oxidoreductase activity	6.15472
indole-3-acetic acid amido synthetase activity	6.06415
mechanosensitive ion channel activity	6.01509
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	5.92453
monooxygenase activity	5.92453

agronomy

MD

Article

Phenylalanine Ammonia-Lyase (PAL) Genes Family in Wheat (*Triticum aestivum* L.): Genome-Wide Characterization and Expression Profiling

Fatima Rasool [1,2], Muhammad Uzair [2], Muhammad Kashif Naeem [2], Nazia Rehman [2], Amber Afroz [3], Hussain Shah [4] and Muhammad Ramzan Khan [1,2,*]

Nodes: 507

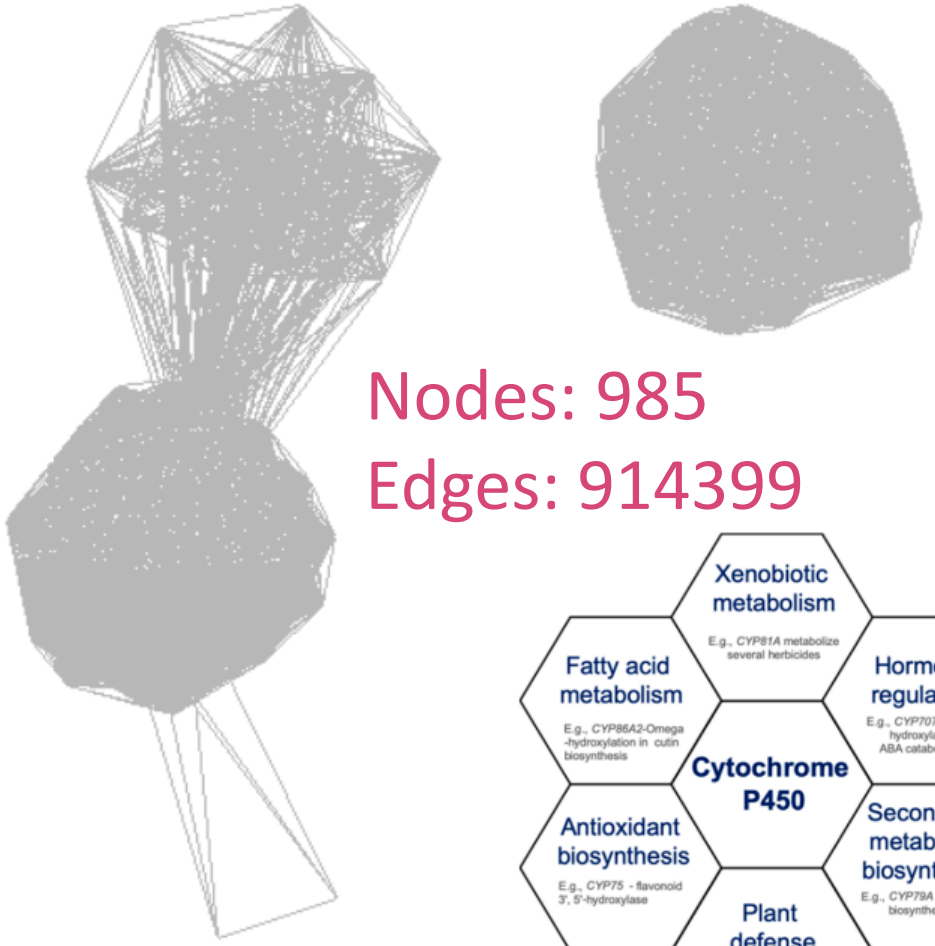
Edges: 96404

Genes with the highest degree value for OSJ

Name	Description	Degree
Os02g0626100	Phenylalanine ammonia-lyase	983
Os02g0626600	phenylalanine ammonia-lyase, putative, expressed	983
Os02g0187100	Similar to cyclase	963
Os03g0738400	Serine hydroxymethyltransferase	914
Os03g0794500	Glutamate dehydrogenase	914
Os04g0623800	Aminomethyltransferase Stress and defense	914
Os07g0188800	Aldehyde dehydrogenase	914
Os11g0455500	erythronate-4-phosphate dehydrogenase	914
Os01g0192900	aminotransferase, classes I and II, domain containing protein, expressed	820
Os01g0760600	Aspartate aminotransferase	820

Name	Description	Degree
AT1G17060	putative cytochrome P450 brassinosteroid binding	6678
AT3G14610	putative cytochrome P450	6678
AT3G14620	putative cytochrome P450	6678
AT3G14630	putative cytochrome P450	6678
AT3G14640	putative cytochrome P450	6678
AT3G14650	putative cytochrome P450 The mRNA is cell-to-cell mobile.	6678
AT3G14660	putative cytochrome P450 The mRNA is cell-to-cell mobile.	6678
AT3G14680	putative cytochrome P450	6678
AT3G14690	putative cytochrome P450 The mRNA is cell-to-cell mobile.	6678
AT1G51340	Encodes a root citrate transporter which together with the root malate transporter ALMT1 are the primary mechanism of aluminum tolerance.	6678
AT3G08040	Encodes a member of the MATE (multidrug and toxin efflux family), expressed in roots but not shoots.	6678
AT5G15180	Peroxidase superfamily protein;(source:Araport11)	6678
AT2G46950	cytochrome P450, family 709, subfamily B, polypeptide 2;(source:Araport11)	6678
AT2G46960	member of CYP709B	6678
AT4G27710	member of CYP709B The mRNA is cell-to-cell mobile.	6678

Genes with the highest degree value for ARATH



Nodes: 985

Edges: 914399

Pandian et al., (2020)

Figure 1. Diverse roles of cytochrome P450s in plants.

Conclusions

- We present a straightforward pipeline to compare two species and functional categories
- GOCompare allowed recognize processes related to organic and metabolic process
- GOCompare allowed finding different roles of molecular functions between species suggested by their node graph weights
- Key genes and molecular functions are related to signaling cascades

¡Gracias!

VI Congreso Colombiano de Biología Computacional y Bioinformática

Pontificia Universidad JAVERIANA Cali

ómicas

- Evolución, Ecología y Conservación -

ccsosaa@javerianacali.edu.co