

GENES EXPRESSION MEASUREMENT BY USING THE GENOME SEQUENCES OF *O.* *CUMANA* AND SUNFLOWER

Stéphane Muñoz

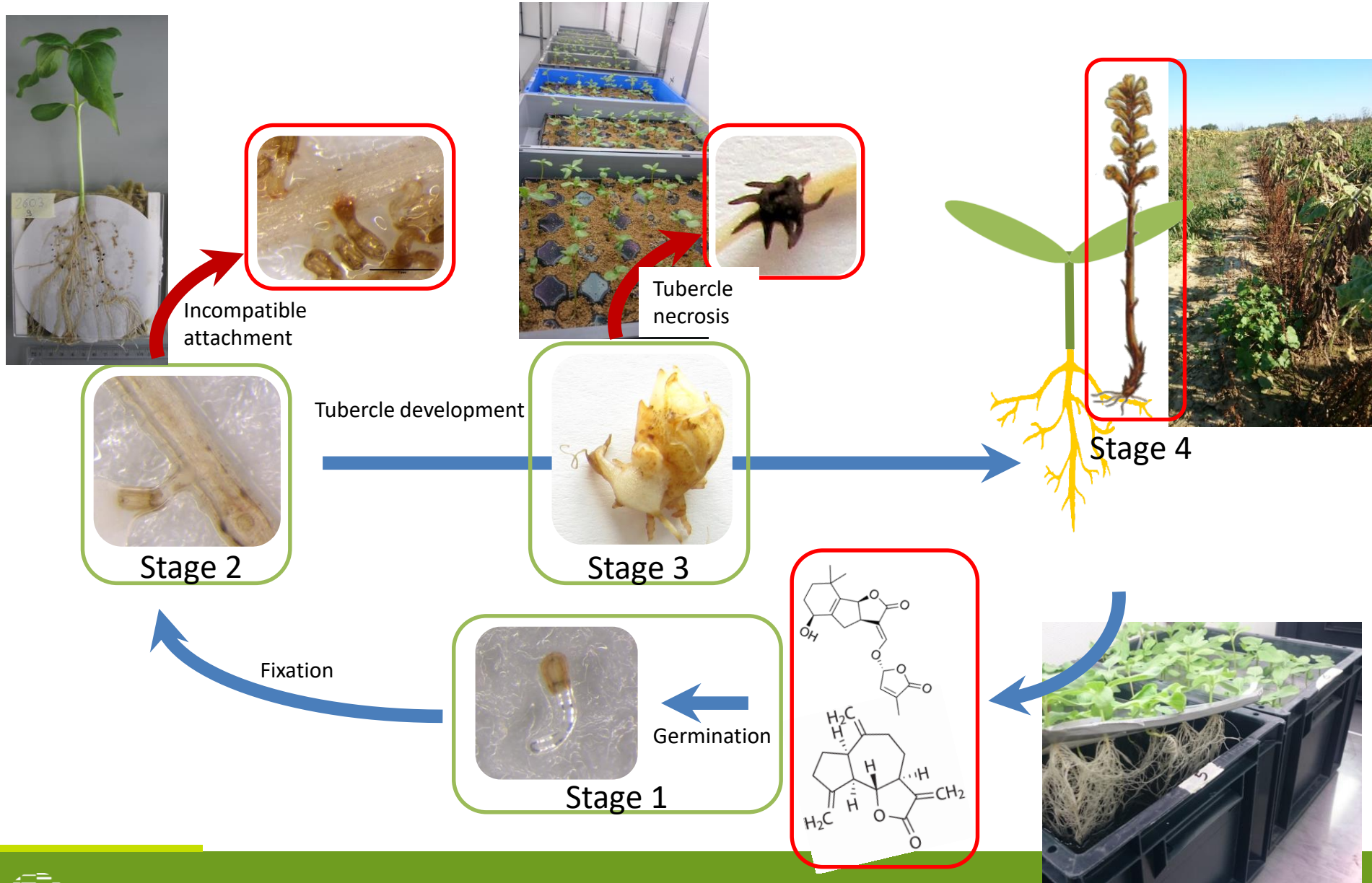
stephane.munos@inra.fr



@stephane_munos



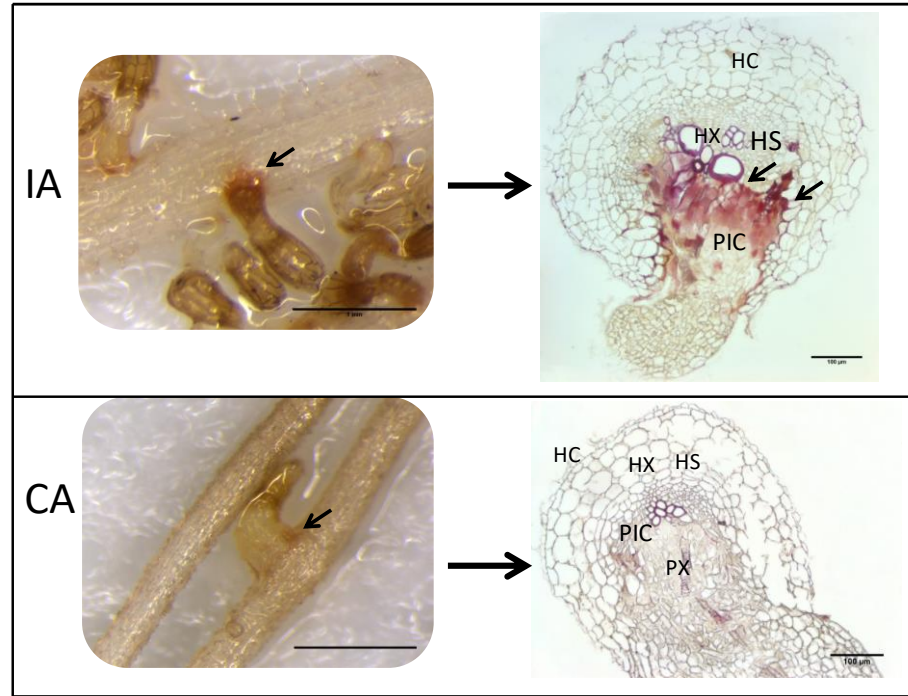
Biology cycle of *Orobanche Cumana* and resistance mechanisms



Prevent connection to the vascular system in sunflower root : *a key resistance mechanism*



Incompatible attachment



Barrier formation:
no connection to
the vascular
system of the
host

Connection to
the vascular
system of the
host

Cell wall modifications seem to be involved both compatible and incompatible attachments.

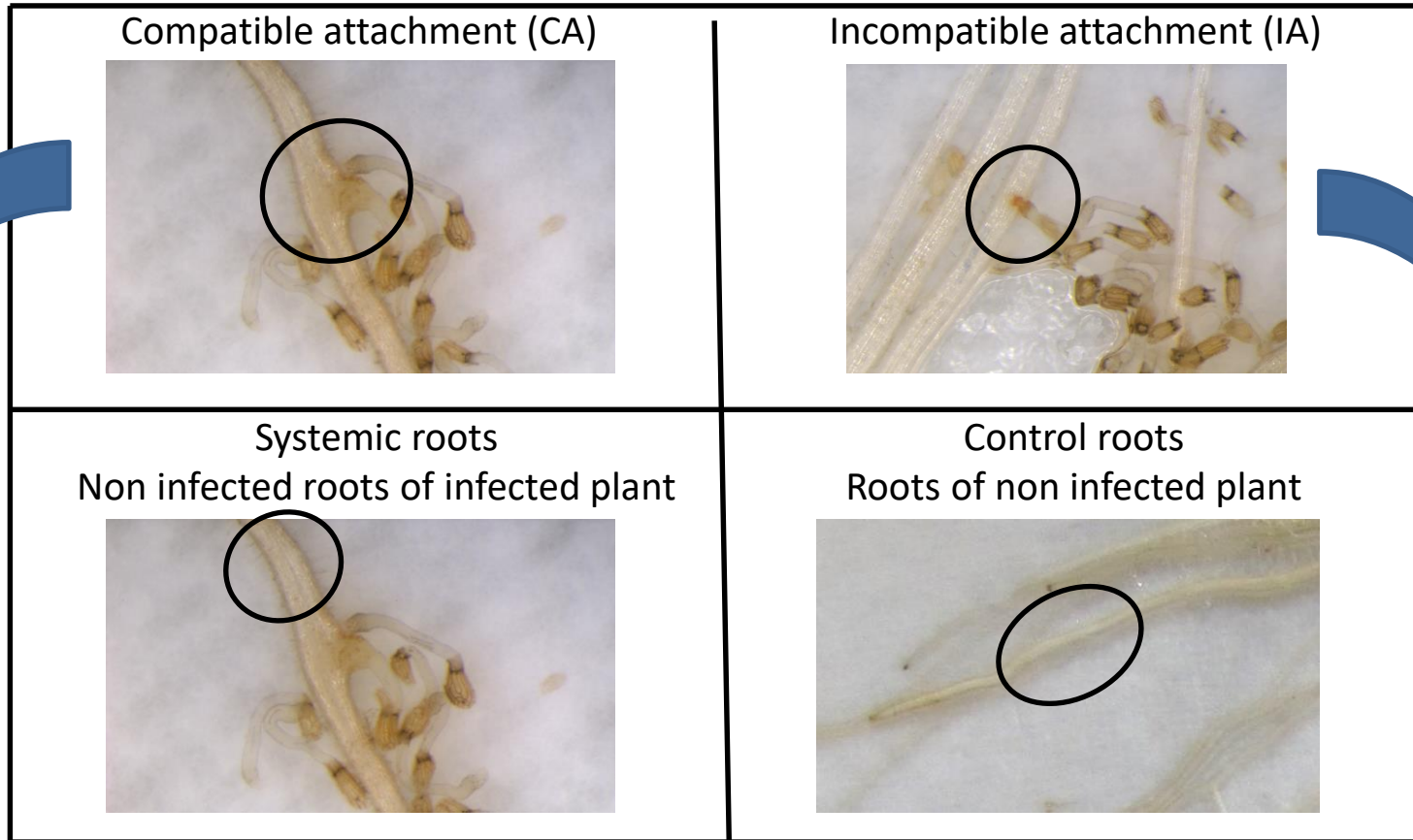
Which genes are involved?

Understanding the molecular and cellular mechanisms involved in the incompatible attachments could help in identifying new candidate genes for the resistance

Genotype	IA rate	# of Healthy tubercles	# of necrotic tubercles	Total number of tubercles
LR1	66.5	2.4	0	2.4
HA89	7.1	3.4	0	3.4
RIL334	73.4	0.2	0.2	0.4
2603	0	13	0	13

O. cumana race F used for infection

Sampling for transcriptomic analysis



Genes from both broomrape and sunflower are expressed in the attachments samples

RNASeq experiment is the main method to measure the expression of thousands of genes simultaneously.

But reference sequences of all genes are needed

Genomics of parasitic plants and of their hosts

Feature	<i>Mimulus guttatus</i>	<i>Triphysaria versicolor</i>	<i>Striga hermonthica</i>	<i>Orobanche aegyptiaca</i>
Nutrition	Autotrophic	Hemiparasite	Hemiparasite	Holoparasite
Dependence on host	Free living	Facultative	Obligate	Obligate
Genome size (Mb/1C)	430	1975	1672	3900
Chromosome number (2N)	28	22	38	24
Hosts with abundant sequence information (model hosts)	N/A	<i>Arabidopsis</i> , <i>Medicago</i> , tomato	Maize, rice, sorghum	<i>Arabidopsis</i> , tobacco, tomato

Illustration from Westwood *et al.*, 2010

TRENDS in Plant Science

Parasitic Plant Genome Project

<http://ppgp.huck.psu.edu>

Transcriptome libraries using
454 sequencing (Roche)

Fully sequenced (Hellsten *et al.*,
2013)

<https://phytozome.jgi.doe.gov>

Sanger shotgun sequencing

scaffolds: 1507

Sequence size total: 312.7 Mb
(72.7% of the genome, 7.3% N)
N50: 21.2Mb (7 scaffolds)

33573 protein-coding transcripts

Genomics of parasitic plants and their hosts

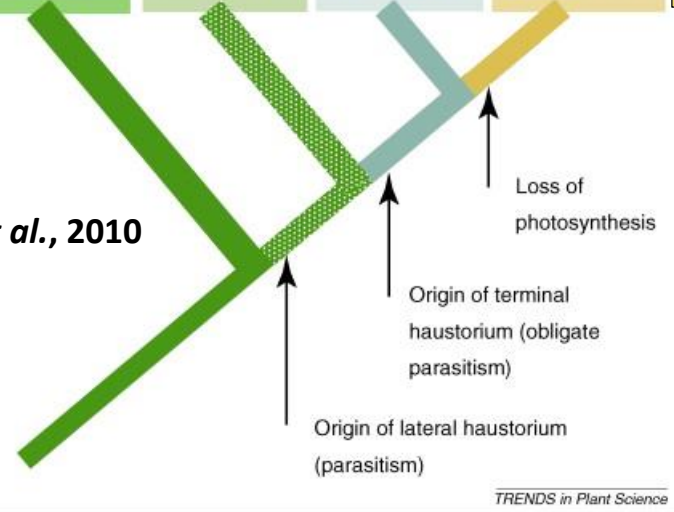
Feature	<i>Mimulus guttatus</i>	<i>Triphysaria versicolor</i>	<i>Striga hermonthica</i>	<i>Orobanche aegyptiaca</i>	<i>Orobanche cumana</i>
Nutrition	Autotrophic	Hemiparasite	Hemiparasite	Holoparasite	Holoparasite
Dependence on host	Free living	Facultative	Obligate	Obligate	Obligate
Genome size (Mb/1C)	430	1975	1672	3900	1420
Chromosome number (2N)	28	22	38	24	38
Hosts with abundant sequence information (model hosts)	N/A	<i>Arabidopsis</i> , <i>Medicago</i> , tomato	Maize, rice, sorghum	<i>Arabidopsis</i> , tobacco, tomato	sunflower



→ Fully sequenced

Fully sequenced

Illustration from Westwood *et al.*, 2010



LETTER

OPEN
doi:10.1038/nature22380

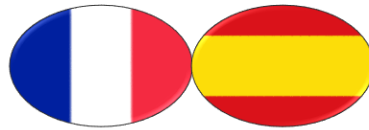
The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution

Hélène Badoin^{1*}, Jérôme Gouzy^{2*}, Christopher J. Grass^{3,4*}, Florent Murat¹, S. Evan Staitoni¹, Ludovic Cottret¹, Christine Lehmann¹, Eriwei⁵, Gregory L. Owens², Sebastian Carreno¹, Baptiste Mayjonade¹, Ludovic Legrand¹, Navdeep Giff¹, Nolan C. Kane^{2,6}, John E. Bowers⁷, Sarel Hubner^{2,8,9}, Arnaud Bellec¹⁰, Aurelie Berard¹¹, Helène Berges¹², Nicolas Blanchet¹, Marie-Claude Boniface¹, Dominique Brunel¹, Olivier Catrice¹, Nadia Chaidir¹³, Clotilde Claudot¹⁴, Céclie Donnadieu¹⁵, Thomas Faraud¹⁶, Christian Fievel¹, Nicolas Holsinger¹⁷, Matthew King¹⁸, Steven J. Knapp¹⁹, Zhao Lan²⁰, Marie-Christine Le Paslier¹, Yannick Lippé¹, Lolita Lorenzon¹, Jennifer R. Mandel²¹, Gwenola Marage¹, Ouwennelle Marchand¹, Elodie Marquand¹, Emmanuelle Bret-Mestries², Evan Morien¹, Savitri Nambossan²², Thuy Nguyen²³, Prane Nogue-Edapaguet¹, Nicolas Pouxilly¹, Francois Ralho¹, Erika Sabel¹, Thomas Schiex²⁴, Justine Thomas¹, Céline Vandecasteele¹, Didier Varot¹, Felicity Year¹, Sonia Vaurin¹⁰, Martin Crespi²⁵, Brigitte Mangin¹, John M. Burke¹, Jérôme Sabe¹, Stéphanie Muñoz¹, Patrick Vincourt¹, Loren H. Rieseberg²⁶ & Nicolas R. Langlade¹

148 | NATURE | VOL 546 | 1 JUNE 2017

Genomic resources available make *O. cumana*–sunflower a unique pathosystem

The HeliOr project



A public-private joint French-Spanish consortium



**Sunflower team:*

Sunflower genetics and genomics

O. cumana diversity

**Bioinformatics team:*

Genome assembly and annotation

SNP calling

Informatic tools for biologists

Sequencing platform

BAC libraries

Genomic tools (optical map...)

Sunflower genetics

O. cumana genetics and diversity

Fundings:



O. cumana biology



Sunflower genetics
O. cumana diversity

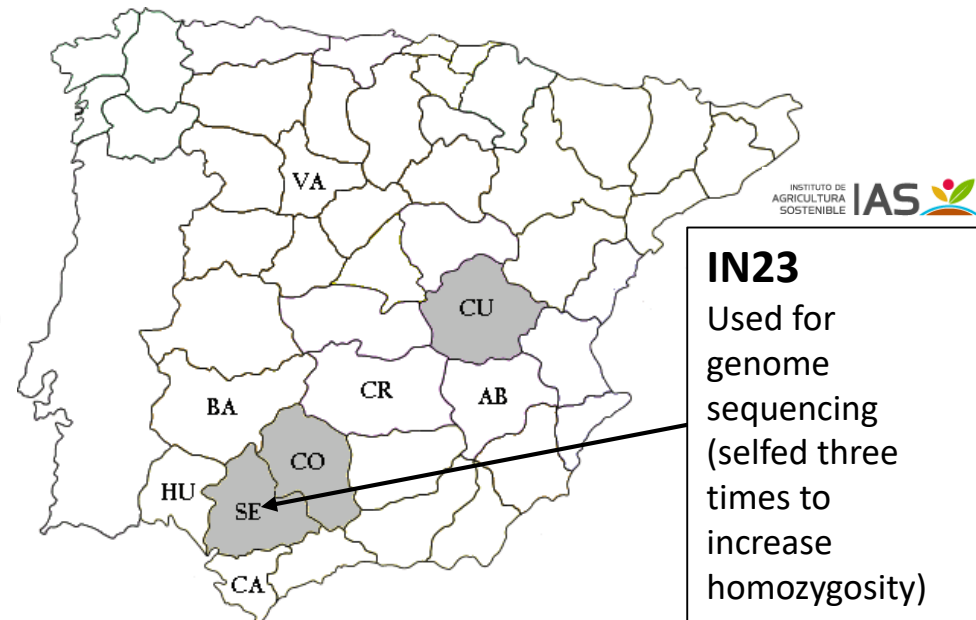
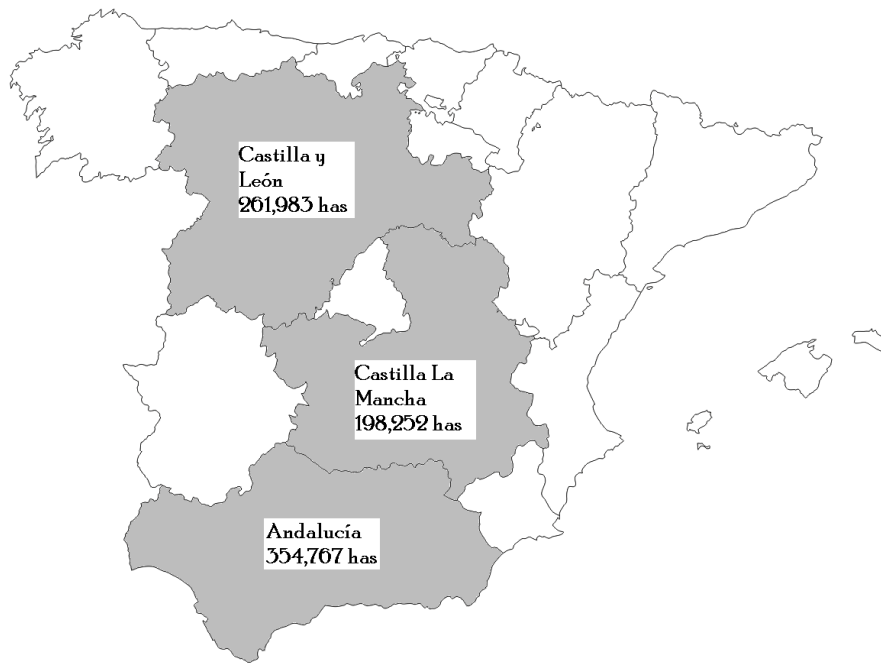


Advices to farmers
List of infected fields in
France



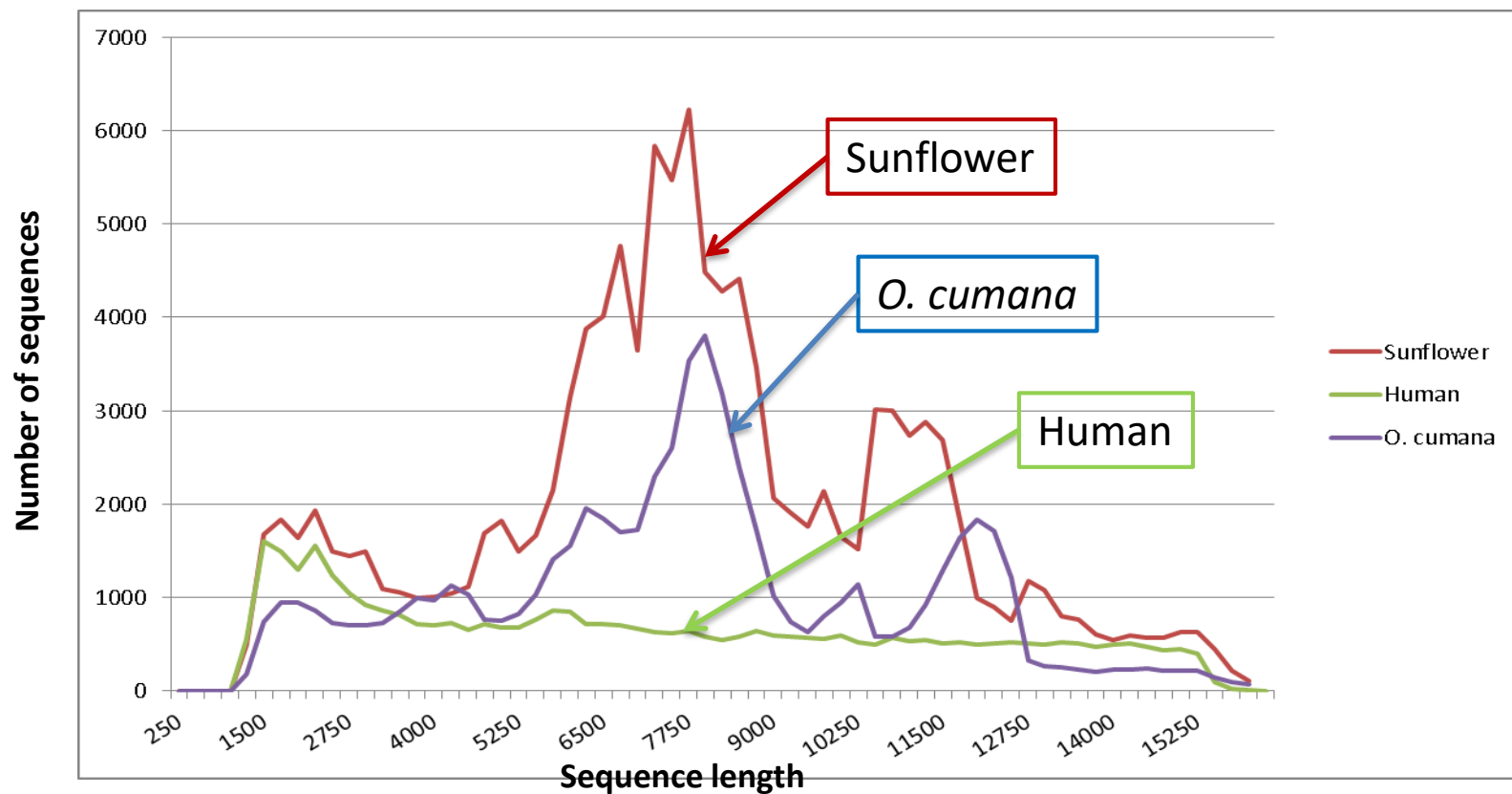
Sunflower and *Orobanche cumana* in Spain

- Race F has been the most virulent race until recently in both main areas of occurrence of sunflower broomrape: Cuenca (CU) and the Guadalquivir Valley: two distinct genetic pools



O. cumana: a nightmare not only for sunflower. For bioinformaticians too!!!

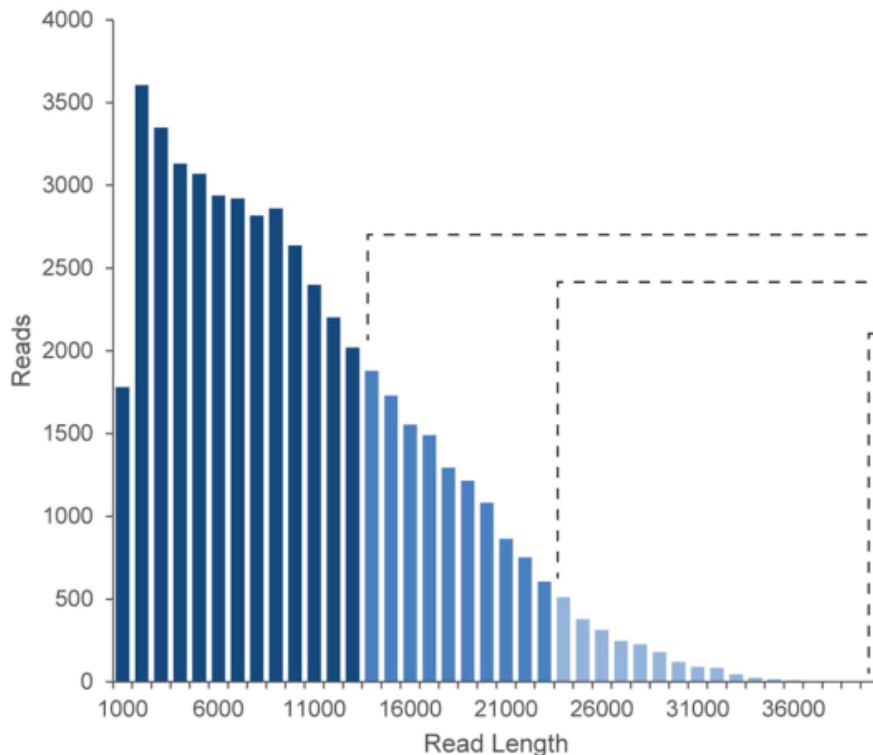
J. Gouzy



A lot of repeats (like sunflower, i.e. 33% of the genome) but longer than in sunflower

Long read sequencing using PacBio RSII (Pacific Biosciences)

P6-C4: Read Length Performance



Half of data in reads: > 14 kb

Top 5% of reads: > 24 kb

Maximum read length: > 40 kb

Data per SMRT® Cell: 500 Mb – 1 Gb (in 4 hours)

PacBio produces sequences longer than the known repeats

P6-C4, 4-hr movie, 20-kb BluePippin™ size-selected *E. coli* library (1 SMRT Cell)

PacBio data



N. Pouilly

- All data produced at GeT-Plage (INRA)



- Produced from October 2015 to February 2016
 - 100X depth expected
 - 126 SMRT Cells (mean: 1.19 Gb/SMRT Cell)

Statistics of contig sequences after assembly of the raw data

Steps	NUM	MAX (Mp)	N50 (Mp)	NUM >=N50	MEAN (bp)	MEDIAN	Total (Gb)
Raw data (subreads) 126 SMRT Cells	13.2M	85.05					149.9
Corrected reads (CANU)	7.04M	55.53	13.98	2.01M	10651	9777	75.06
	NUM	MAX (Mb)	N50 (Mb)	NUM >=N50	MEAN (Mb)	MEDIAN (Mb)	Total (Gb)
Genome assembly (CANU)	905	16.88	3.57	107	1.53	6.49	1.388
Remove spurious + Sequence based Scaffolding + polishing (QUIVER)	793	16.98	4.21	96	1.74	7.43	1.380

/2

From contigs to chromosomes sequences!

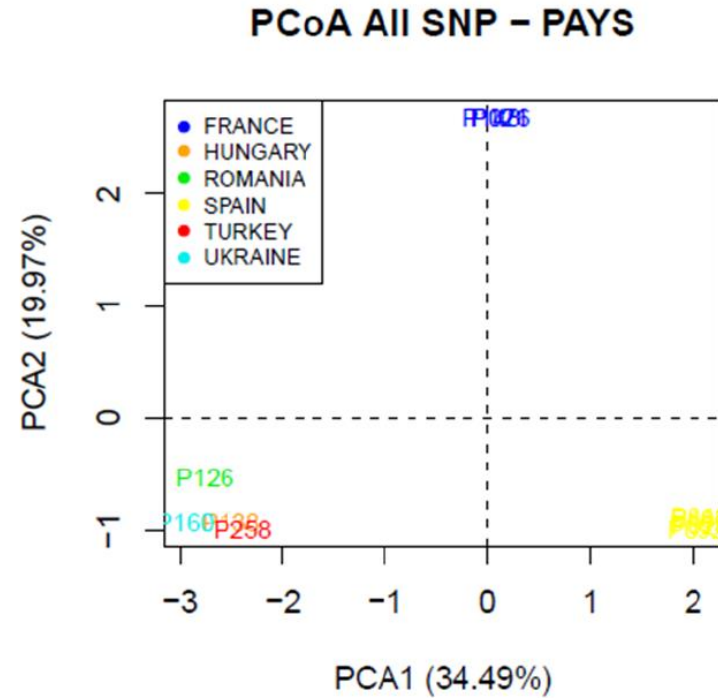
Using genetic map and optical map!

Diversity analysis in *O. cumana*



M. Coque

	Total
FRANCE	3
HUNGARY	1
ROMANIA	1
SPAIN	5
UKRAINE	2
Total	12



Exome capture from the 12 populations : 362285 SNPs



1536 SNPs selected to maximize the diversity of the whole set

Fundings:  **SOFIPROTEOL**
L'engagement durable

A segregating population for the first genetic of *O. cumana*



B. Pérez-Vich L. Velasco



IN12

INA



IN12

F1

INA

IN23 used for genome sequencing

The first genetic map of *O. cumana*



X. Grand

1536 SNPs + 168 SSR x 91 F2 and parental lines have been genotyped

509 SNPs + 18 SSR were polymorphic and did not show any distortion of their segregation in the full population

Genetic map built using CarthaGène software (INRA) with a high stringency

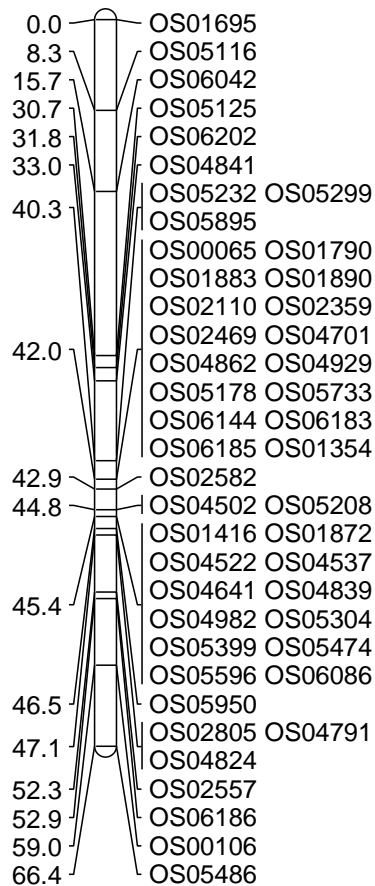


1479cM

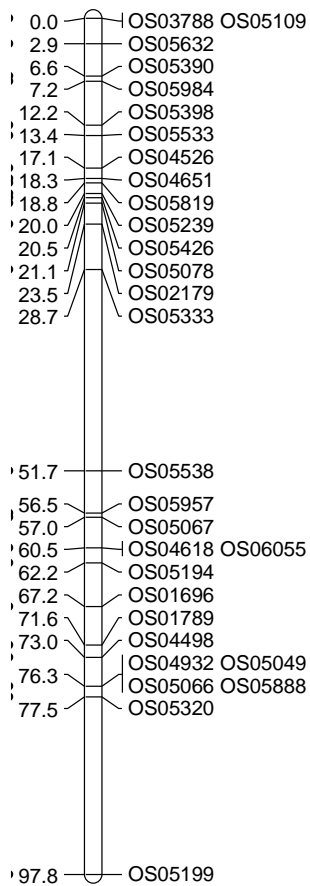
28 linkage groups for the 19 chromosomes

The first genetic map of *O. cumana*

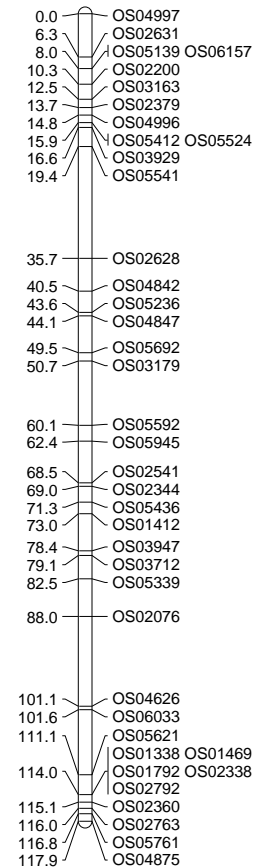
.KHv2_LG28



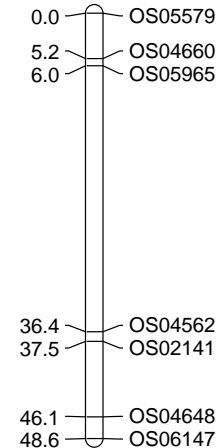
_KHv2_LG24



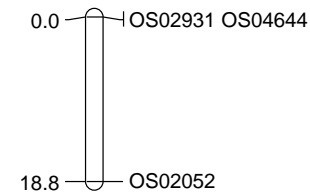
LKHv2_LG22



LKHv2_LG11



LKHv2_LG21



Colinearity between the genetic map and the genome

Almost perfect colinearity between physical map and genetic map.

But strong variations between physical distances and genetic distances according to the regions.

This genetic map enabling the anchoring of 95 contigs (593Mb) from the 256 contigs representing 90% of the genome assembly.

161 contigs remain unmapped to anchor 90% of the genome!

ReSequencing the 2 parental lines of the segregating population



A. Calderon

The genome of the 2 parental lines used to produce the F2 segregating populations have been resequenced (HiSeq, Illumina)



40 912 accurate polymorphic SNPs between the two parental lines

Improving the genetic map



L. Hu

Not enough contigs anchored to the genetic map (95 contigs)

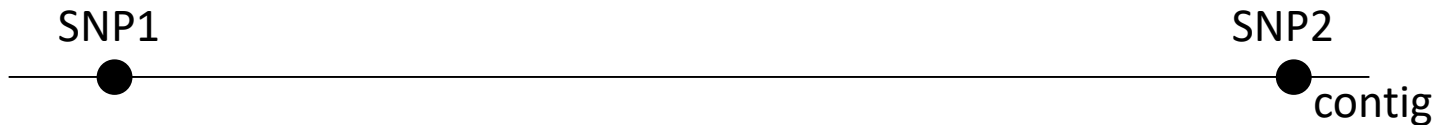
40 912 polymorphic SNPs between the two parental lines



13511 SNPs located on 145 contigs from the 161 contigs that remain unmapped



278 SNPs will be genotyped in the full segregating population to anchor the 145 more contigs



All these data should anchor 90% of the genome assembly

Transcriptomic data to annotate the *genome*



Fundings **P.R.O.M.O.S.O.L.**
ASSOCIATION
POUR LA PROMOTION
DE LA SÉLECTION
DES PLANTES OLÉAGINEUSES

P. Delavault

60 RNASeq libraries sequenced (data obtained on 26 April 2016)

Corresponding to **20 broomrape development stages** from seeds to flowering (3 replicates/stage)





Use of an automatic annotation pipeline

- EuGene Plant pipeline (egn-ep)

Total number of genes: **55726**

Number of protein coding genes: 46447

Mean gene length (bp): 3568.75

Per cent genes with introns : 63 **genes with 5' UTR: 82%** **genes with 3' UTR: 83%**

- Exons	Mean number per gene: 3.63	Mean length (bp): 472.95	GC %: 40.94
---------	----------------------------	--------------------------	-------------

- Introns	Mean number per gene: 2.63	Mean length (bp): 706.35	GC % 35.40
-----------	----------------------------	--------------------------	------------

-CDS	Mean length (bp): 732.32	GC % 45.25
------	--------------------------	------------

Experimental expression evidence added!

Number of non protein coding genes 9279

* Tephra – repeat annotation - <https://github.com/sestaton/tephra>

11 456 genes (8 256 protein coding + 3 200 ncRNA) are fully included in repeat_region annotated by tephra

Assessment of the annotation quality using « BUSCO » (a set of conserved genes in plants)

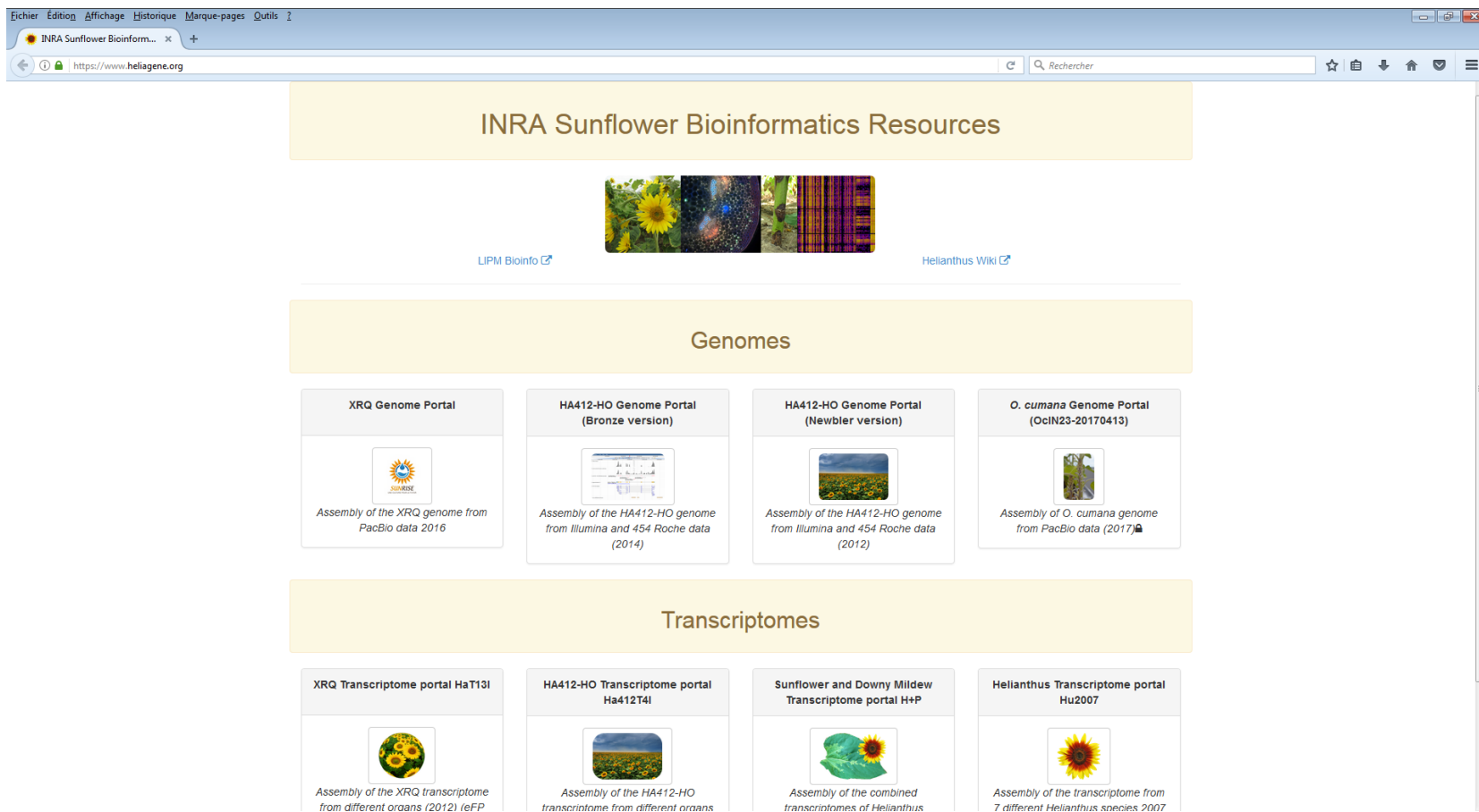
	Complete and unique	Complete but Duplicated	% complete	partial	Missing
Sunflower	1153	196	93.7%	21	70
<i>O. cumana</i>	999	47	72.7%	72	322

**A lot of missing genes!
Due to the biology of *O. cumana*?**

All genome information are available in a Genome Browser

<https://www.heliagene.org>

S. Carrère



The screenshot shows a web browser window displaying the INRA Sunflower Bioinformatics Resources website. The browser's address bar shows the URL <https://www.heliagene.org>. The website has a yellow header with the text "INRA Sunflower Bioinformatics Resources". Below the header is a navigation bar with links for "LIPM Bioinfo" and "Helianthus Wiki". The main content area is divided into two sections: "Genomes" and "Transcriptomes".

Genomes

- XRQ Genome Portal**: Assembly of the XRQ genome from PacBio data 2016.
- HA412-HO Genome Portal (Bronze version)**: Assembly of the HA412-HO genome from Illumina and 454 Roche data (2014).
- HA412-HO Genome Portal (Newbler version)**: Assembly of the HA412-HO genome from Illumina and 454 Roche data (2012).
- O. cumana Genome Portal (OciN23-20170413)**: Assembly of O. cumana genome from PacBio data (2017).

Transcriptomes

- XRQ Transcriptome portal HaT13I**: Assembly of the XRQ transcriptome from different organs (2012) (eFP).
- HA412-HO Transcriptome portal Ha412T4I**: Assembly of the HA412-HO transcriptome from different organs.
- Sunflower and Downy Mildew Transcriptome portal H+P**: Assembly of the combined transcriptomes of Helianthus.
- Helianthus Transcriptome portal Hu2007**: Assembly of the transcriptome from 7 different Helianthus species 2007.

All genome information are available in a Genome Browser

The screenshot shows a web browser window displaying the OclN23-20170413 genome portal. The browser's address bar shows the URL <https://www.heliagene.org/OclN23-20170413/>. The website's navigation bar includes a logo, the site name "OclN23-20170413", and menu items: "Genome browser", "Sequence Tools", "Downloads", "Help", "keyword, accession", "Submit", "Login", and "Logout".

A green banner at the top of the page contains the following text:

- Welcome to OclN23-20170413 genome portal
- More information about access protocol [HERE](#)
- If you need any other feature to be displayed, feel free to contact us at lipm.info@toulouse.inra.fr

The main content area is divided into four sections:

- Genome Browser:** "Navigate through the genome assembly. Multiple tracks are available:"
 - Structural annotation
 - Transcript alignments
 - Protein alignments
 - Repeats
 - ncRNAsA "GO" button is located at the bottom left of this section.
- Search tools:** "Find results and annotations based on pre-computed analyses. These analyses are based on:"
 - Blastp vs. model plants
 - Blastp vs. NR database
 - InterPro scan
 - Blast2GOInput fields for "keyword, accession" and a "Search" button are provided.
- Sequence tools:** "Use blast suite to find feature similarities and extract features. Available databases are:"
 - Genome assembly
 - Genes
 - mRNA, Proteins, CDS
 - ncRNAs
 - Promotor sequences"Blast" and "Extract sequences" buttons are located at the bottom.
- Download section:** "Download complete datasets." with a "GO" button.

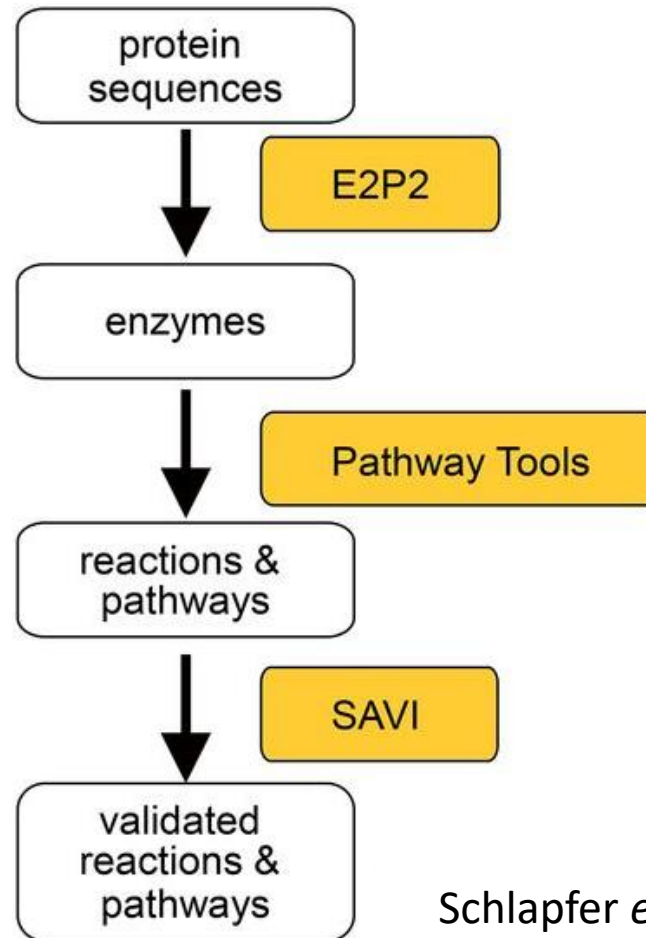
The Lipm logo is centered at the bottom of the page. A black arrow points from the "GO" button in the "Genome Browser" section to a box labeled "MTA for the access".

Biochemical informations from the genome sequence



L. Cottret

Automatic process to build the metabolic database



Schlapfer *et al.*, Plant Physiology. 2017.

Metabolic database

LOGIN | Why Login? | Create New Account

Enter a gene, protein, metabolite or pathway...
Searching *Orobanche cernua cumana* [change organism database](#)



Home Search Genome Metabolism Analysis Help

Summary of *Orobanche cernua*, Subspecies *cumana*, version 1.0

Authors: LIPM Bioinfo, INRA

Summary:

This Pathway/Genome Database (PGDB) was generated by the PathoLogic [Karp10, Dale10, Caspi14] component of Pathway Tools software version 20.0 and MetaCyc version 20.0 on 26-May-2017 13:31:49.

Taxonomic lineage: cellular organisms, Eukaryota, Viridiplantae, Streptophyta, Streptophytina, Embryophyta, Tracheophyta, Euphyllophyta, Spermatophyta, Magnoliophyta, Mesangiospermae, eudicotyledons, Gunneridae, Pentapetalae, asterids, lamiids, Lamiales, Orobanchaceae, Orobancheae, Orobanche, *Orobanche cernua*, *Orobanche cernua cumana*

Unification Links: [NCBI-Taxonomy:78542](#), [NCBI-Taxonomy:78542](#)

Replicon	Total Genes	Protein Genes	RNA Genes	Pseudogenes	Size (bp)	NCBI Link
1 Undisplayed Contigs/Replicons	0	9801	0	0	0	
Total:	9801	9801	0	0		
Genes without a physical map position:	9801					

Pathways:	405
Enzymatic Reactions:	3393
Transport Reactions:	81
Polypeptides:	9803
Protein Complexes:	0
Enzymes:	9781
Transporters:	283
Compounds:	2566
Transcription Units:	0
tRNAs:	0

447 pathways

3458 reactions

1517 classified into pathways

Genetic Code Number:

1 -- Standard

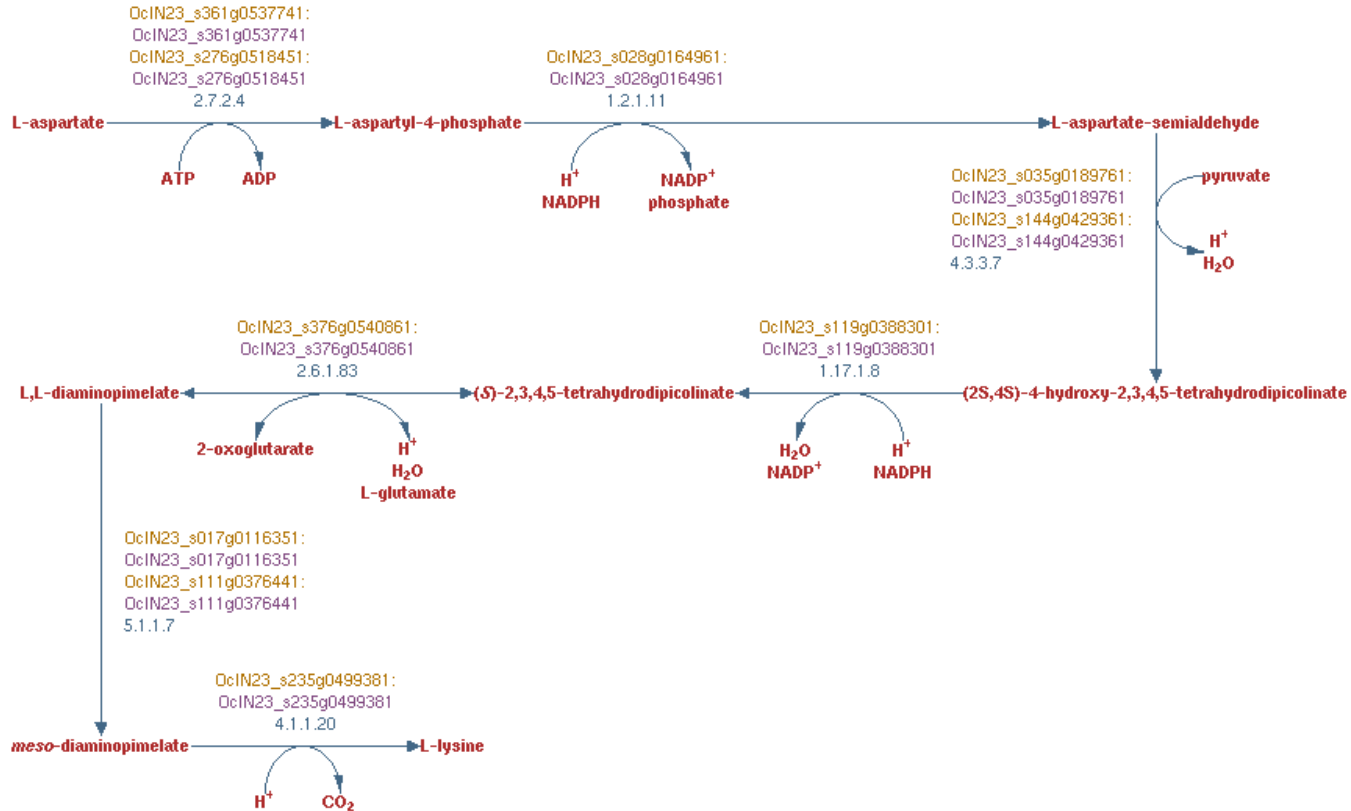
PGDB Unique ID: 1E90

Restricted access but collaborations are welcome

Link between pathways, reactions, enzymes and genes

Orobanche cernua cumana Pathway: L-lysine biosynthesis VI

Show Predicted Enzymes ▾ More Detail Less Detail

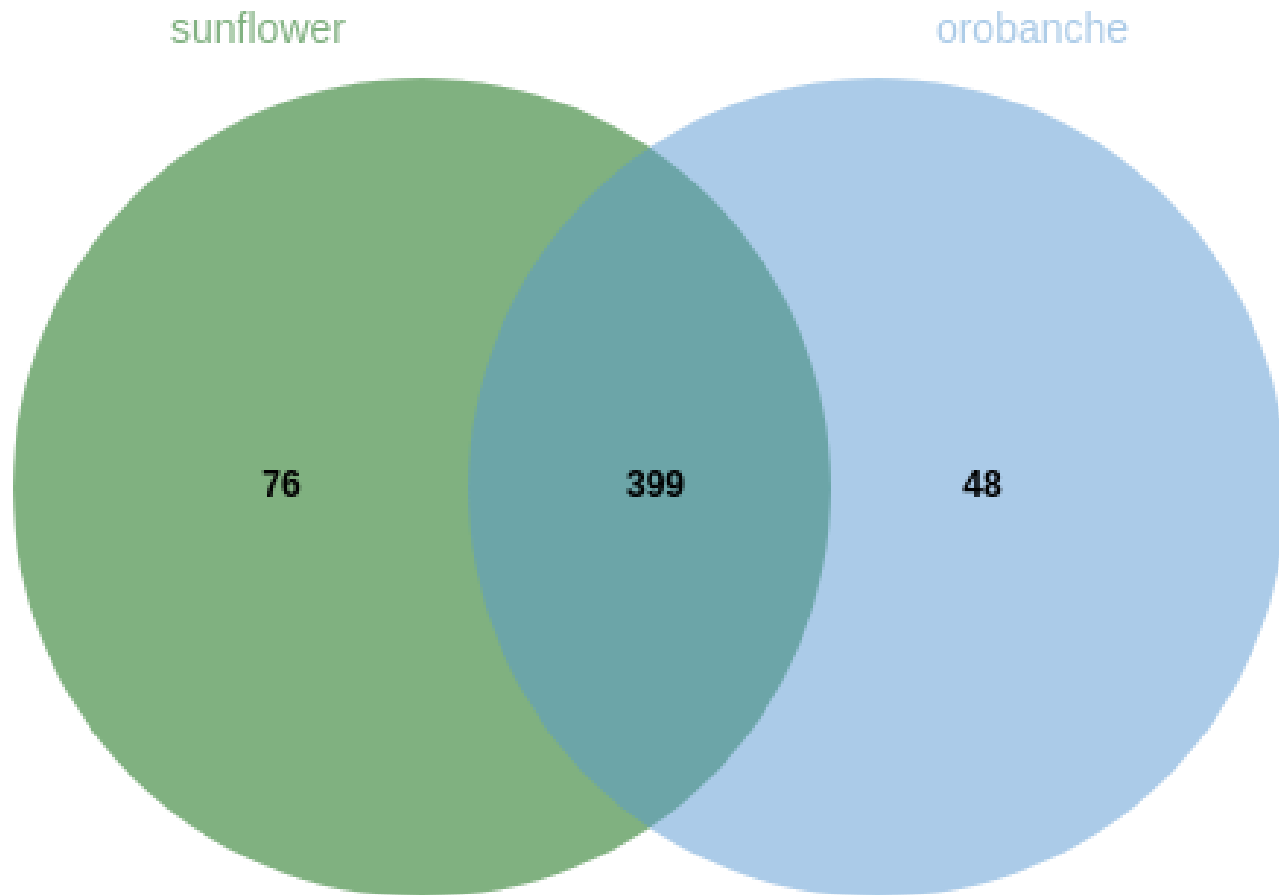


If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

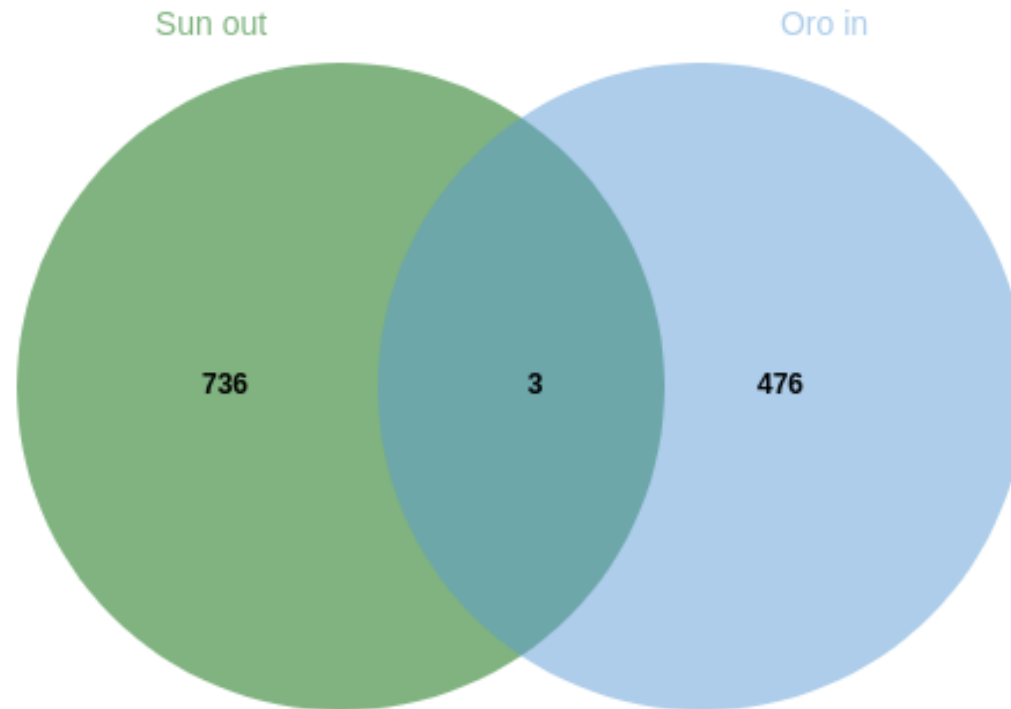
Superclasses: Biosynthesis → Amino Acids Biosynthesis → Proteinogenic Amino Acids Biosynthesis → L-lysine Biosynthesis

Pathway Summary from MetaCyc:
General Background

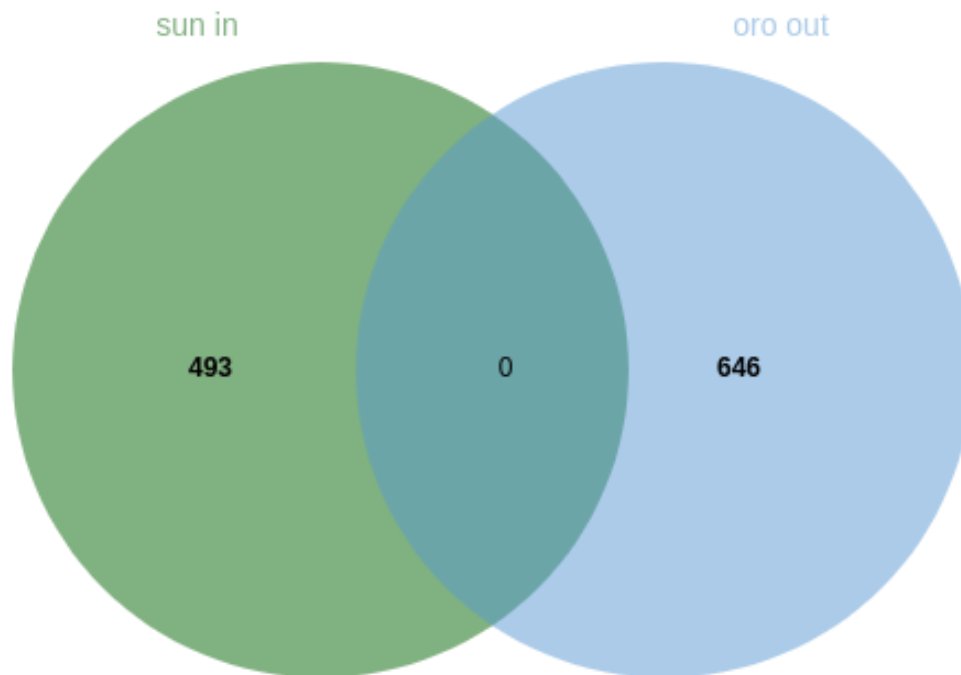
Comparison of sunflower and *O. cumana* pathways



Intersection between Orobanche input metabolites and sunflower output metabolites



Intersection between broomrape output metabolites and sunflower input metabolites



Use of the annotated genome for RNASeq analysis

We used the annotated mRNA sequences from XRQ (sunflower) and IN23 (*O. cumana*) to map the RNASeq data

Preliminary results, more analysis need to be performed to mak then more accurate!

Sunflower DEG during incompatible attachment

Incompatible Attachments



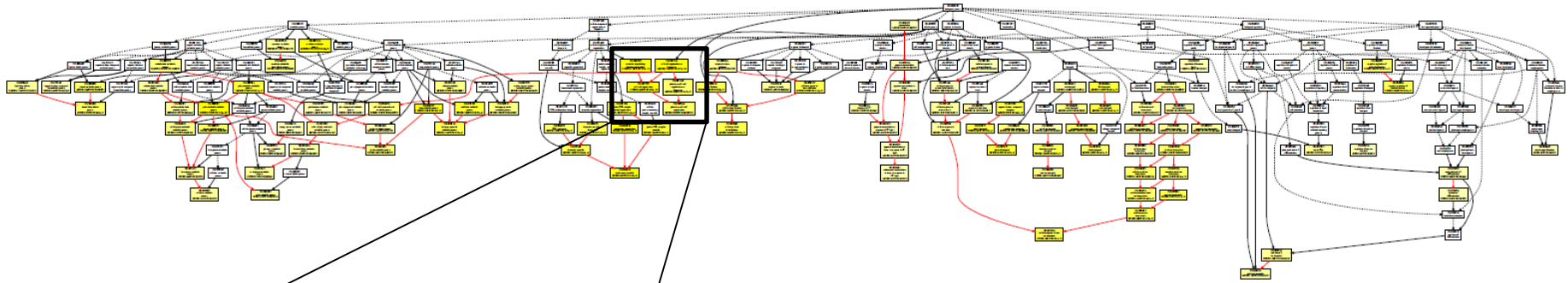
Control



LR1: 3444 DEG

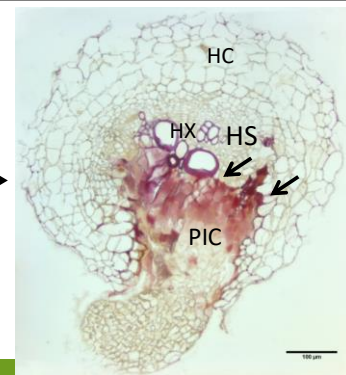
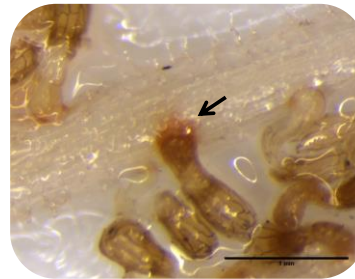
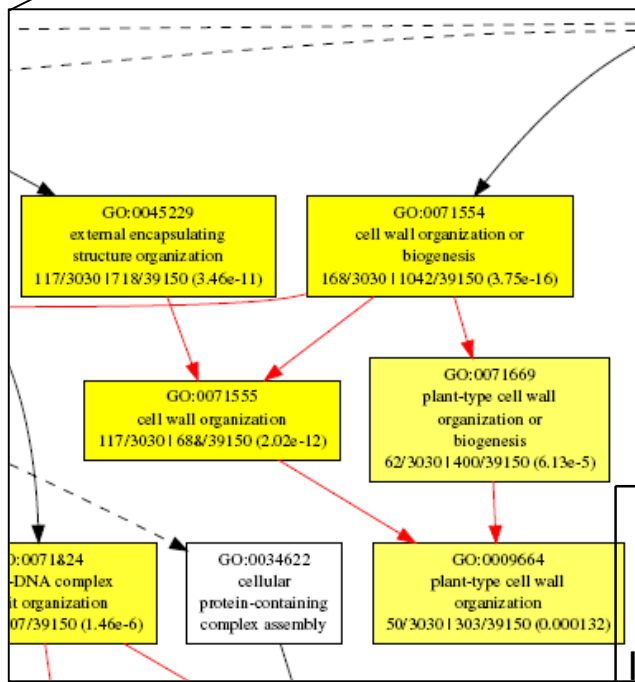


Gene Ontology terms enrichment analysis using GOEAST



A more complex response during incompatible attachment

DEG involved in cell wall biogenesis are enriched to prevent connection of the vascular system of the host



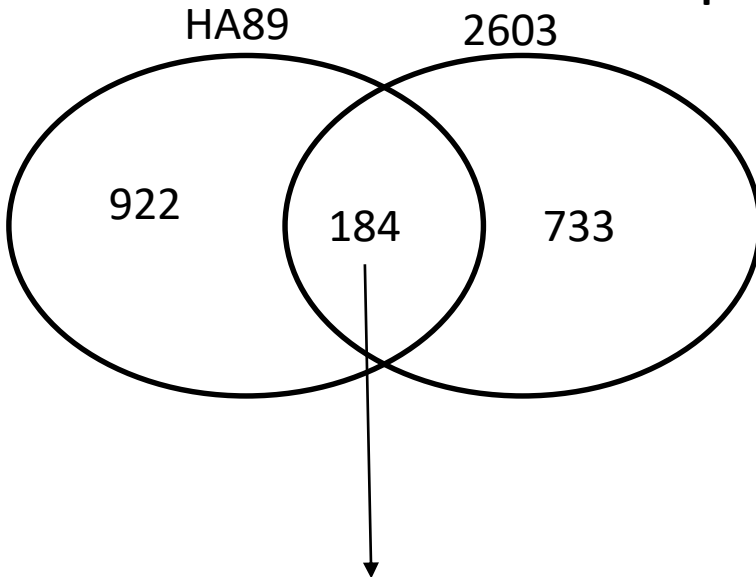
Barrier formation:
no connection to the vascular system of the host

Sunflower DEG during compatible attachment

HA89: 1106 DEG

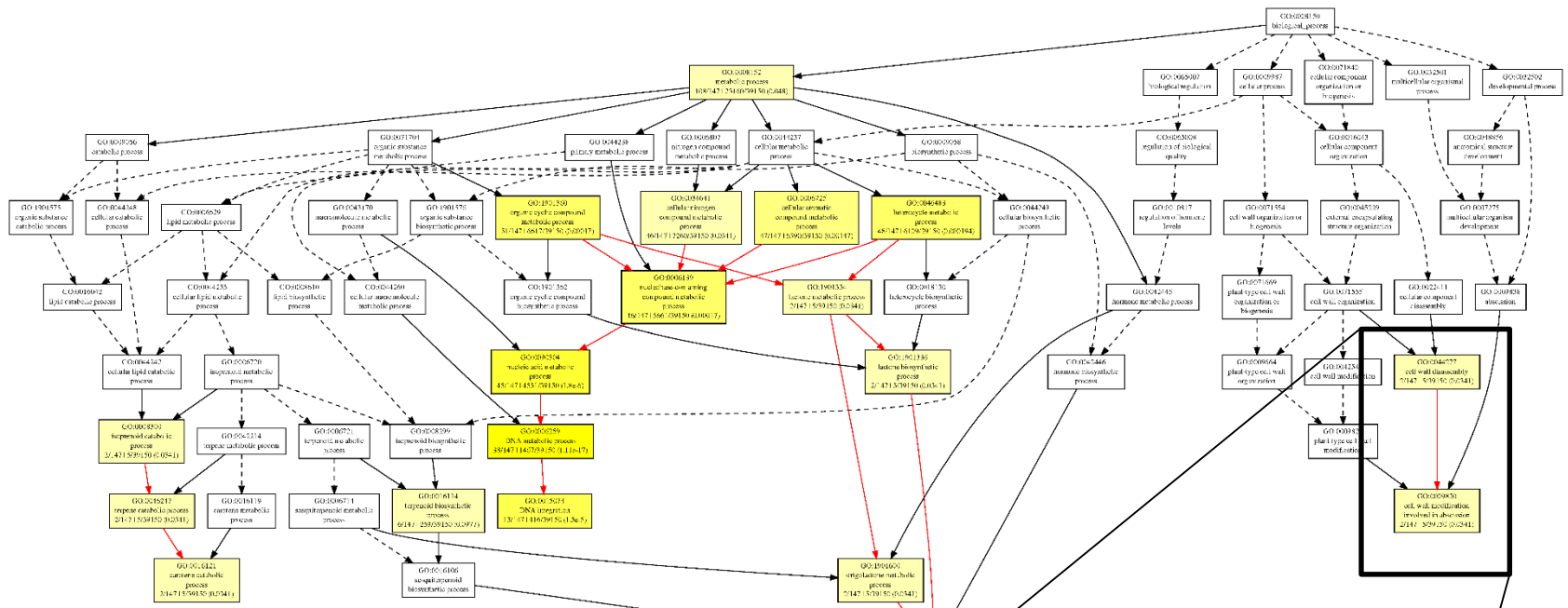
CA 2603: 917 DEG

Compatible Attachments



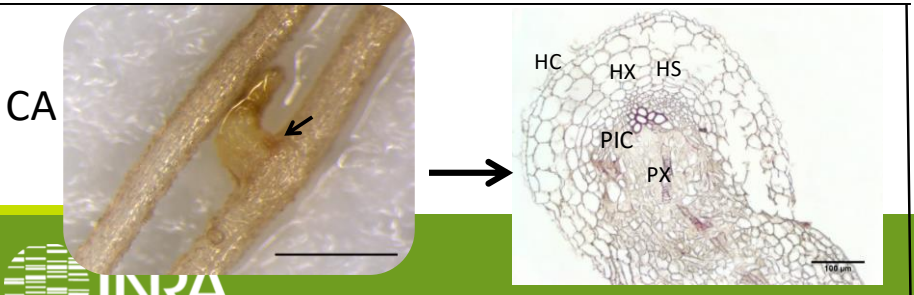
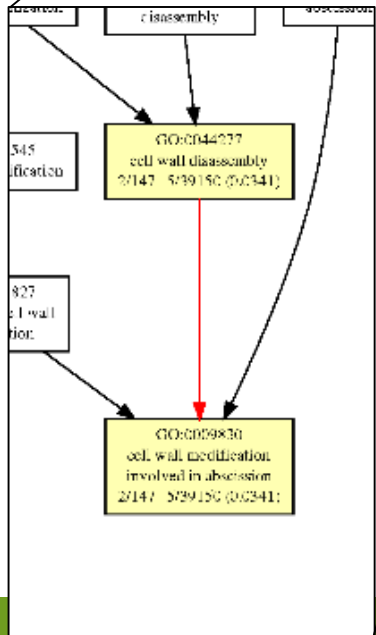
Gene Ontology terms enrichment analysis using GOEAST





DEG involved in Cell wall disassembly are enriched.

The interaction (*O. cumana*?) seems to target cell wall degradation gene to enable *O. cumana* to connect to the vascular system of the host

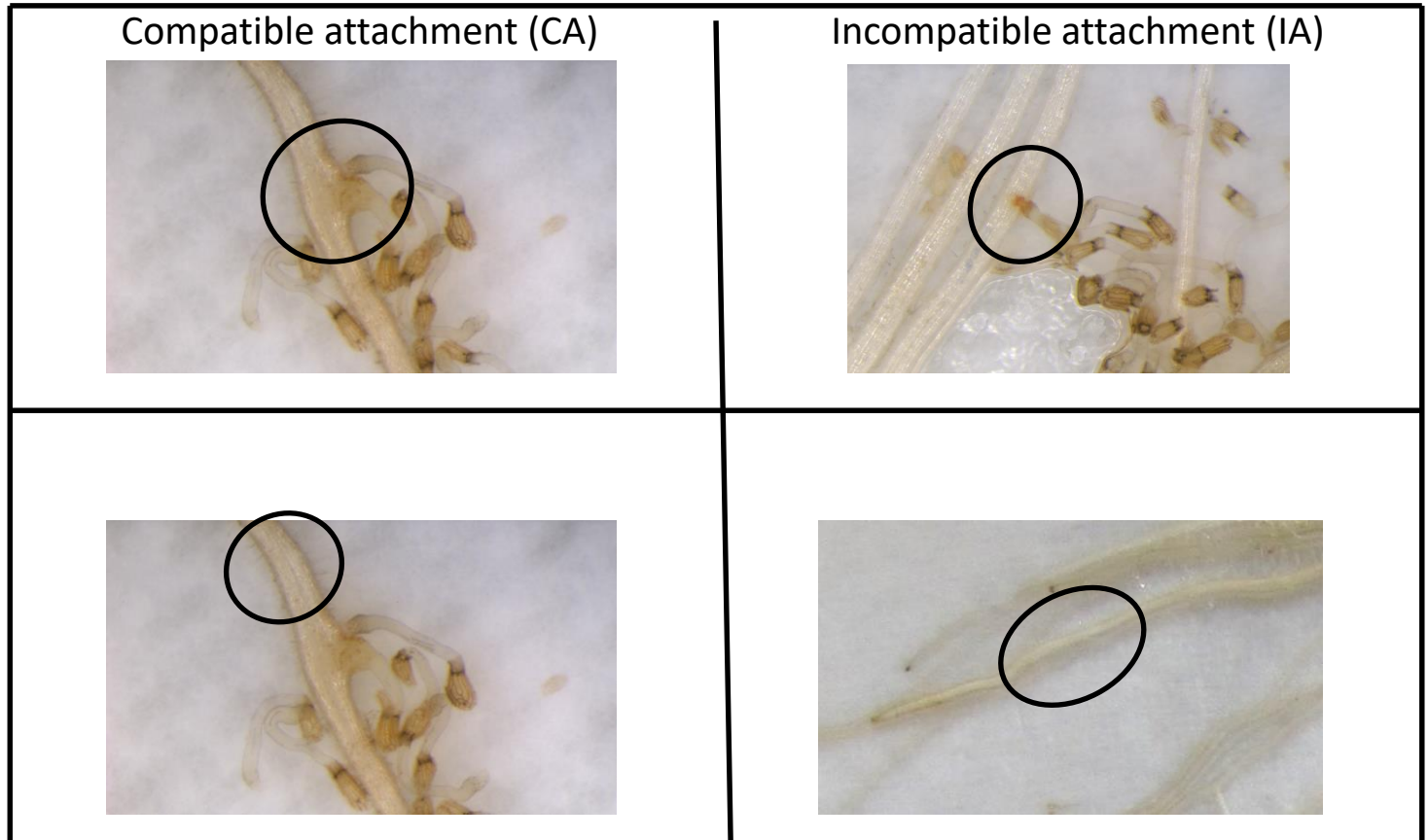


Connection to the vascular system of the host

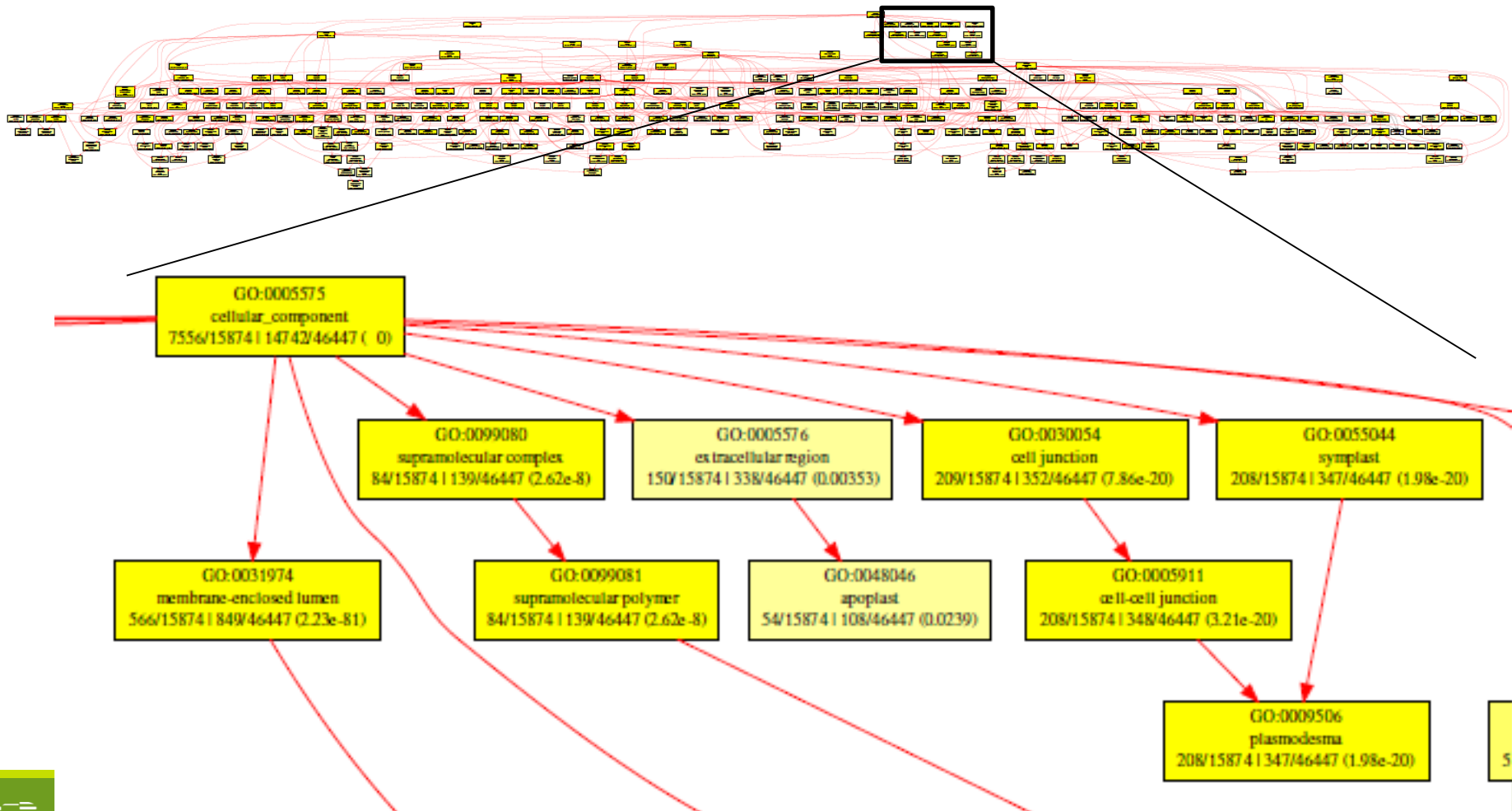
O. cumana DEG

IA vs CA LR1: 15874

IA vs CA Other Resistant line : 742



O. cumana DEG



Summary

A high quality genome sequence of *O. cumana*
produced

www.heliagene.org

Usefull for functional and genetic analysis

Many Thanks to Collaborators

IAS-CSIC:

Álvaro Calderón González
Begoña Pérez-Vich
Leonardo Velasco

LBPV, Nantes University :

Philippe Delavault
Marc-Marie Lechat
Philippe Simier

Biogemma:

Clotilde Claudel
Marie Coque
Sébastien Faure
Xavier Grand
Nicolas Ribière

CNRGV:

Hélène Bergès
Stéphane Cauet
Céline Jézioriski
William Marande

Get-Plage:

Cécile Donadieu
Olivier Bouchez
Maarten Pirson

Terre Inovia:

Christophe Jestin



LIPM:

Julia Bazerque	Nicolas Langlade
Nicolas Blanchet	Marion Larroque
Marie-Claude Boniface	Ludovic Legrand
Fanny Bonnafous	Johann Louarn
Sébastien Carrère	Anne-Sophie Lubrano
Olivier Catrice	Brigitte Mangin
Mireille Chabaud	Gwenola Marage
Ludovic Cottret	Prune Pegot-Espagnet
Alexandra Dühnen	Nicolas Pouilly
Pauline Duriez	David Rengel
Louise Gody	Erika Sallet
Florie Gosseau	Camille Tapy
Jérôme Gouzy	
Luyang Hu	