

# COMPUTATIONAL TOOLS FOR BIOSYNTHETIC PATHWAY DISCOVERY IN PLANTS

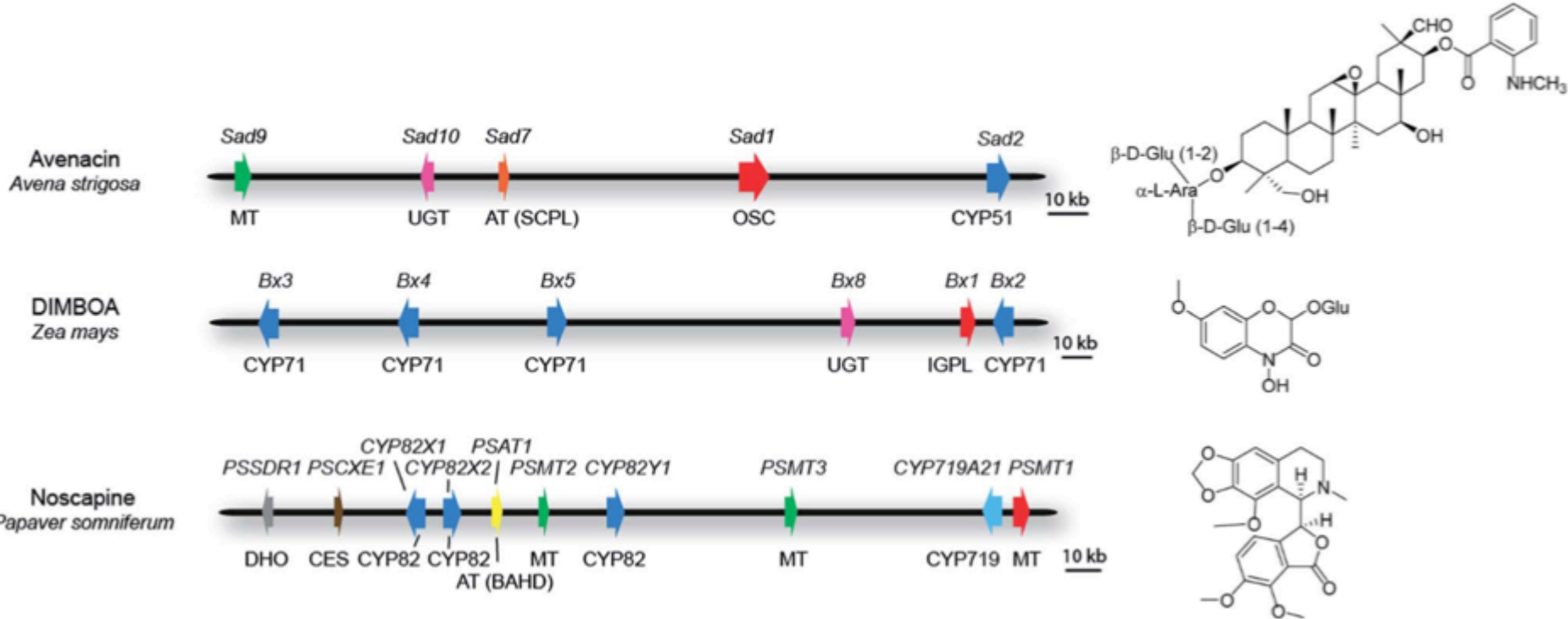


DR. MARNIX H. MEDEMA  
BIOINFORMATICS GROUP  
WAGENINGEN UNIVERSITY, THE NETHERLANDS

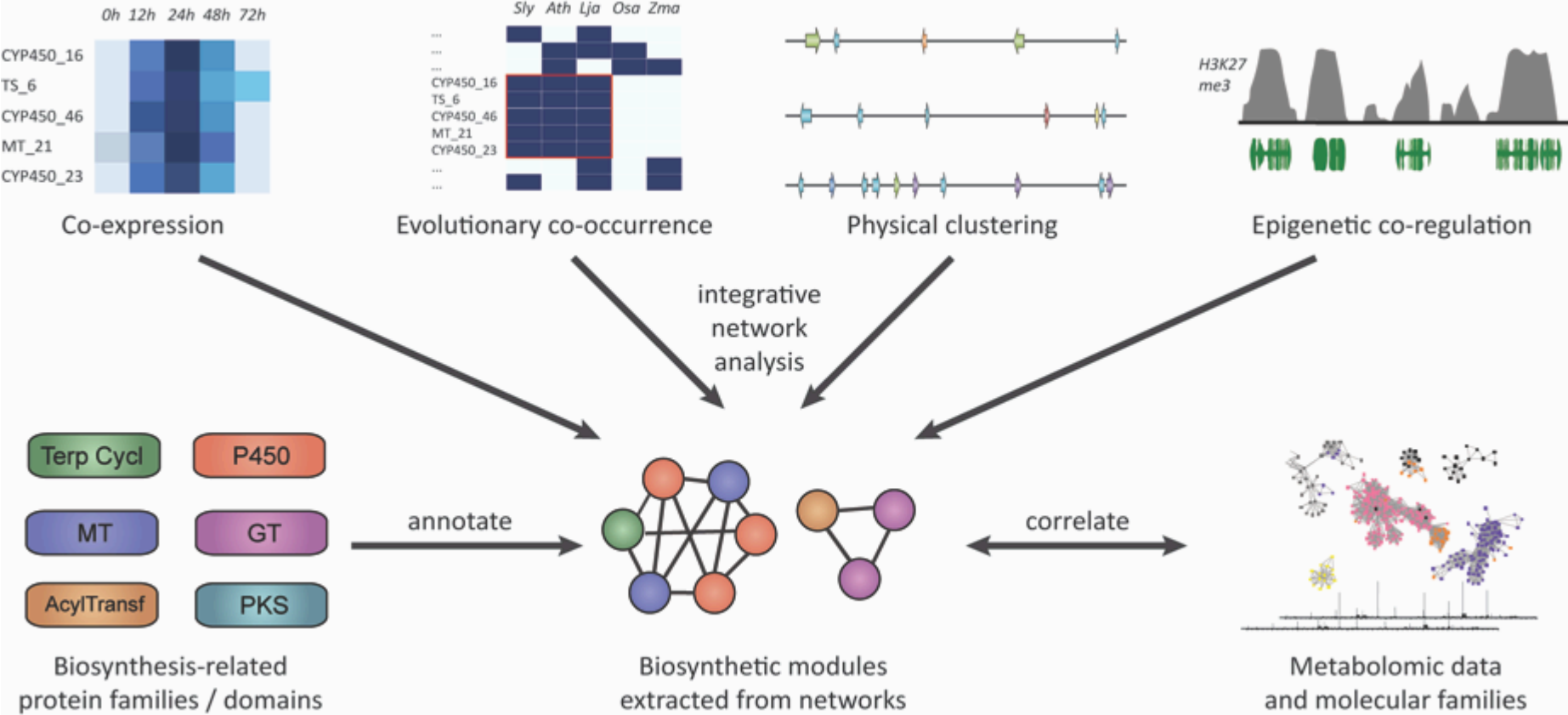
COMPUTATIONAL APPLICATIONS IN SECONDARY METABOLITE DISCOVERY (CAISMD)  
TUESDAY MARCH 9<sup>TH</sup>, 2021



# NOT INFREQUENTLY, PLANT BIOSYNTHETIC PATHWAYS ARE ENCODED IN GENE CLUSTERS



# PLANT BIOSYNTHETIC PATHWAY DISCOVERY REQUIRES AN INTEGRATIVE APPROACH



# PLANTISMASH: A SOFTWARE TOOL TO IDENTIFY BGCS IN PLANT GENOMES

The screenshot shows the Plant SMASH web interface. At the top, the browser address bar displays 'plantismash.secondarymetabolites.org'. The page header includes the 'planti SMASH' logo and the title 'Plant Secondary Metabolite Analysis Version 1.0.0-beta'. Below the header is a navigation bar with icons for home, help, error, and download. A 'Select Gene Cluster' section contains 45 numbered buttons, with button 40 highlighted in red. The main content area is titled 'CP002688 - Cluster 40 - Terpene' and features a 'Gene cluster description' section. This section provides details about the gene cluster (CP002688 - Gene Cluster 40, Type = terpene, Location: 19411187 - 19467569 nt) and includes a 'Download cluster GenBank file' link. Below the description is a genomic map showing various genes represented by colored boxes. A specific gene, 'AT5G48010', is highlighted in green. A tooltip for 'thalianol synthase 1' is displayed, showing its locus-tag (AT5G48010), protein-ID (AED95610.1), location (19457001 - 19461538), and signature pHMM hits (SQHop\_cyclase\_C and SQHop\_cyclase\_N). The interface also includes a 'Show:' section with checkboxes for 'biosynthetic genes' and 'other genes', and a 'Legend:' section with color-coded boxes for different gene types: red for Cytochrome 450, green for Terpene synthase, blue for BAHD acyltransferase, brown for (Other) Biosynthetic Genes, grey for Other Genes, light red for biosynthetic genes, light blue for transport-related genes, light green for regulatory genes, and light grey for other genes. At the bottom right, there are two small portrait photos of individuals.

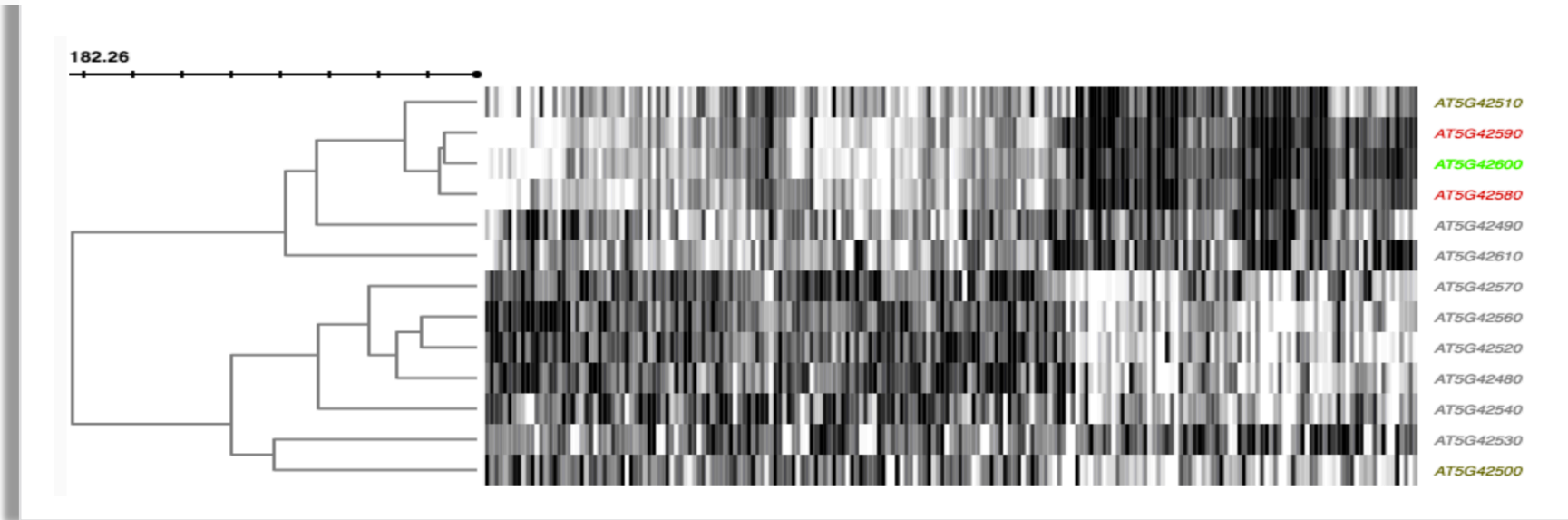


# EXPRESSION DATA CAN BE LOADED INTO PLANTISMASH TO IDENTIFY COEXPRESSION PATTERNS WITHIN CLUSTERS...

## CP002688 - Cluster 39 - Lignan-terpene

### Gene cluster description

CP002688 - Gene Cluster 39. Type = lignan-terpene. Location: 16987519 - 17064815 nt. Click on genes for more information. Download cluster GenBank file  
Show pHMM detection rules used



# ... AND BETWEEN CLUSTERS

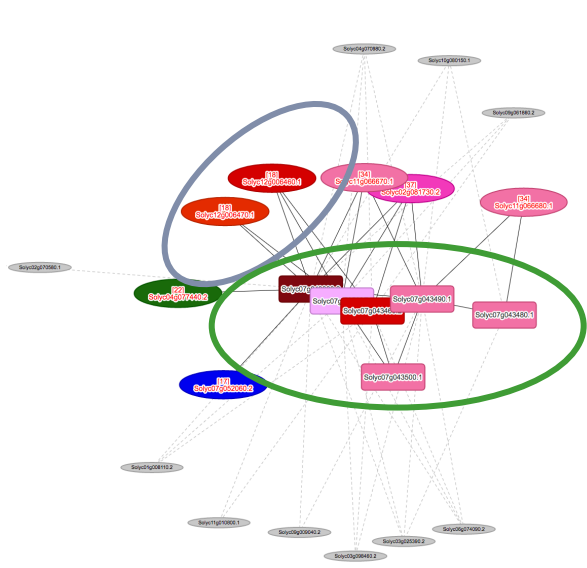
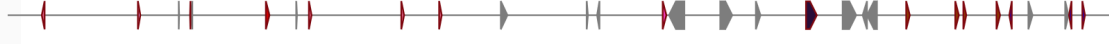
## SL2.50ch07 - Cluster 16 - Saccharide

### Gene cluster description

SL2.50ch07 - Gene Cluster 16. Type = saccharide. Location: 57075999 - 57706232 nt. Click on genes for more information.

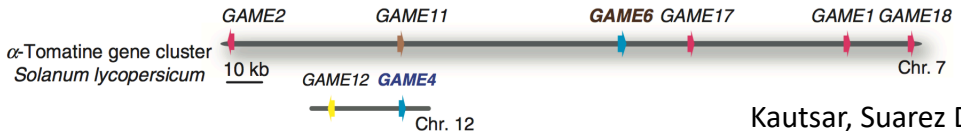
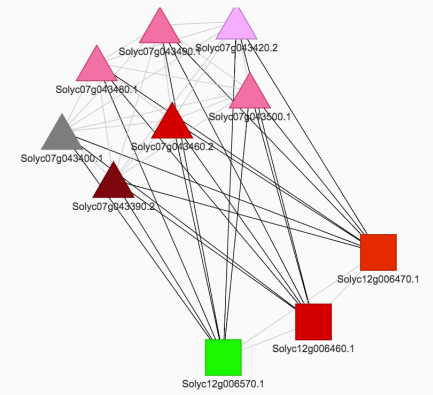
[Download cluster GenBank file](#)

Show PHMM detection rules used

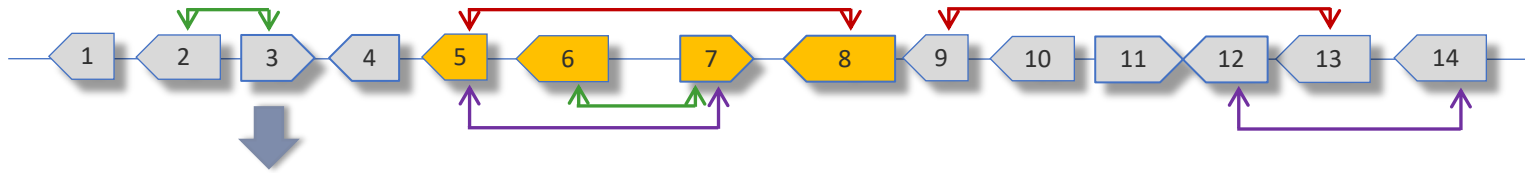


Source	Target	Attributes [?]	Details
Cluster 16	Cluster 18	{21, 7, 3}	show

### Cluster 16 (triangle) X Cluster 18 (square)



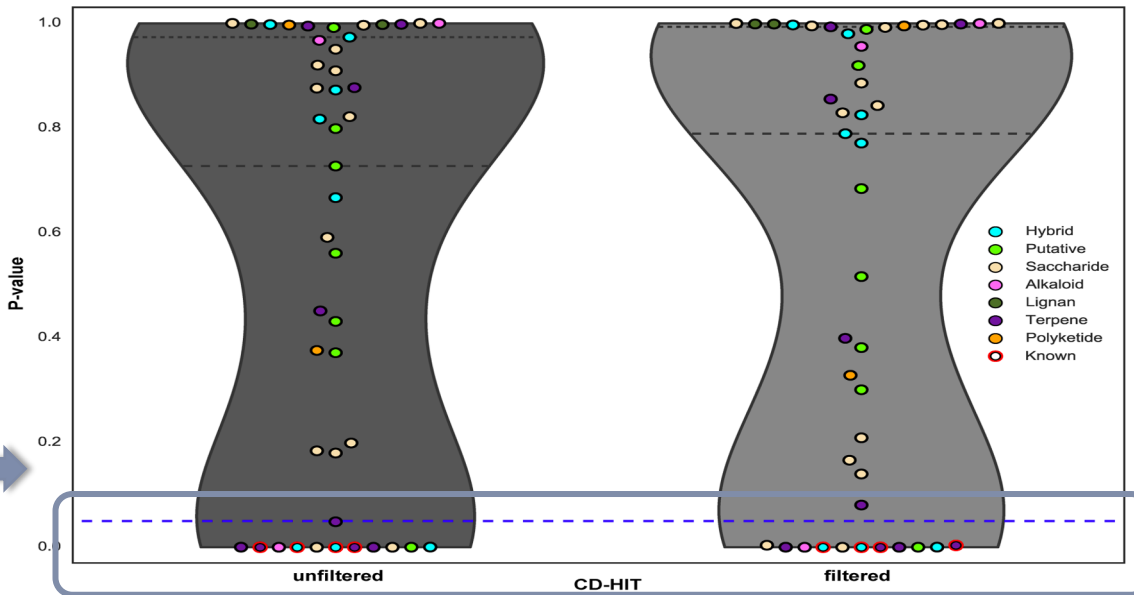
# STATISTICS INDICATE SIGNIFICANTLY COEXPRESSED GENE CLUSTERS: IN *ARABIDOPSIS*, ~25% OF ALL CANDIDATE BGCS ARE HIGHLY COEXPRESSED



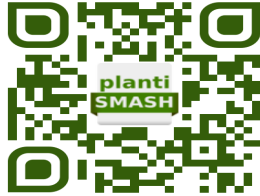
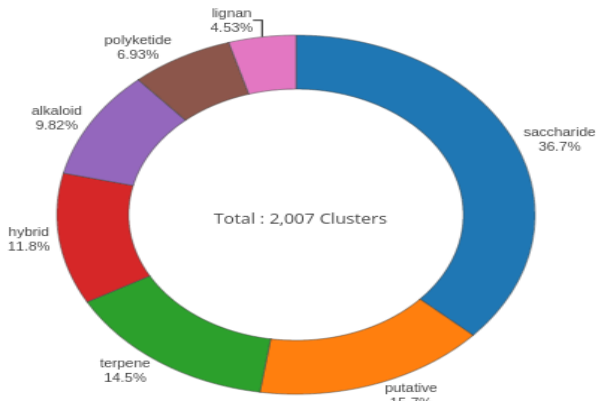
$$R_{(BGC)} = [R_{(5,8)}, R_{(5,7)}, R_{(6,8)}, R_{(5,6)}, R_{(6,7)}, R_{(7,8)}]$$

$$R_{(Background)} = [R_{(1,4)}, R_{(9,12)}, R_{(10,13)}, R_{(11,14)}, R_{(1,3)}, R_{(2,4)}, R_{(9,11)}, R_{(10,12)}, R_{(11,13)}, R_{(12,14)}, R_{(1,2)}, R_{(2,3)}, R_{(3,4)}, R_{(9,10)}, R_{(10,11)}, R_{(11,12)}, R_{(12,13)}, R_{(13,14)}]$$

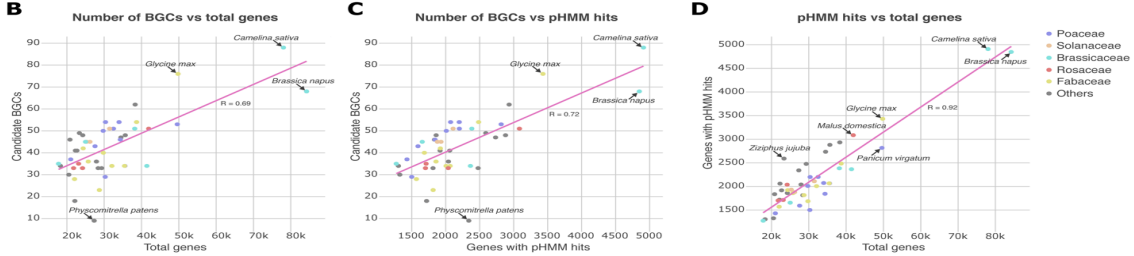
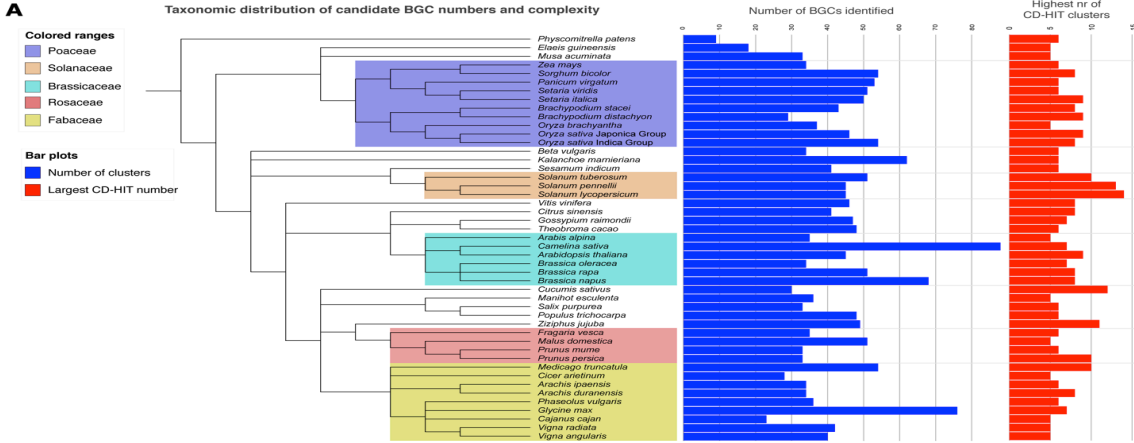
Wilcoxon rank sum  
(Mann-Whitney U) test



# GENE CLUSTER REPERTOIRES ARE DIVERSE THROUGHOUT THE PLANT KINGDOM



[plantismash.secondarymetabolites.org/precalc](http://plantismash.secondarymetabolites.org/precalc)



Plant Secondary Metabolite Analysis

[plantismash.secondarymetabolites.org](http://plantismash.secondarymetabolites.org)

# PLANTISMASH IDENTIFIES MANY COMPLEX PLANT BIOSYNTHETIC GENE CLUSTERS

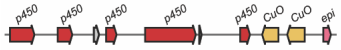
***Aquilegia coerulea* putative triterpene biosynthesis gene cluster**  
 chromosome 3 - 123 kb



***Medicago truncatula* putative triterpene biosynthesis gene cluster**  
 chromosome 6 - 256 kb



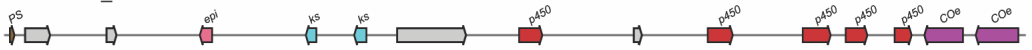
***Theobroma cacao* putative alkaloid biosynthesis gene cluster**  
 chromosome 3 - 55 kb



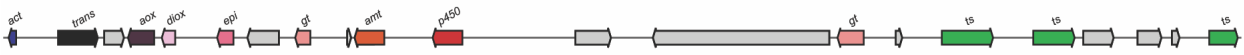
***Citrus sinensis* putative hybrid terpene biosynthesis gene cluster**  
 chromosome 4 - 173 kb



***Sesamum indicum* putative polyketide biosynthesis gene cluster**  
 scaffold NC\_026154 - 123 kb

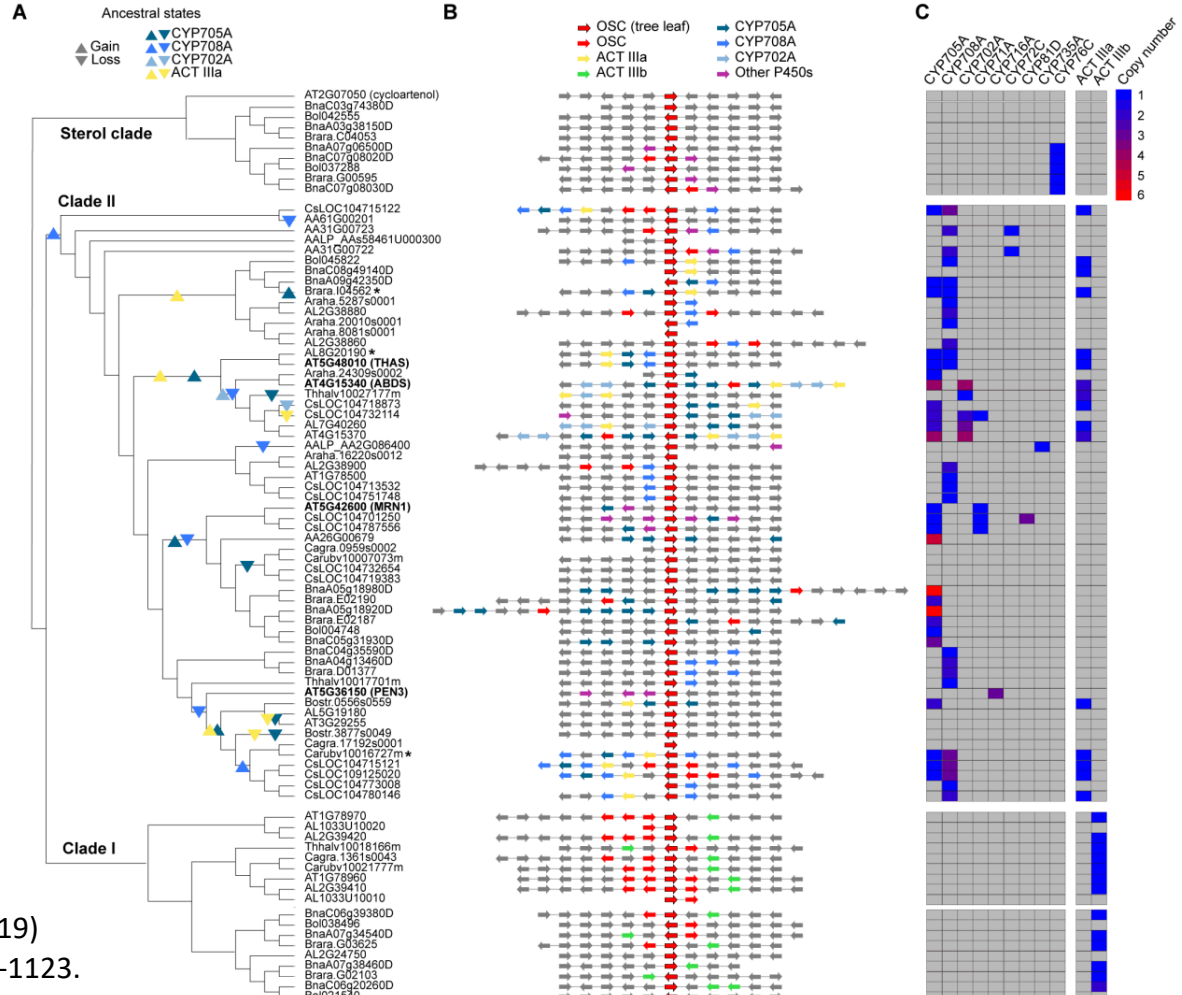


***Solanum pennellii* putative hybrid terpene biosynthesis gene cluster**  
 chromosome 12 - 211 kb



Cytochrome P450	Glycosyltransferase	Methyltransferase	Oxidoreductase	CoA-ligase
Terpene synthase	Ketosynthase	BAHD acyltransferase	Dehydrogenase	Amino oxidase
Copper amine oxidase	Squalene epoxidase	Scl acyltransferase	Dioxygenase	Aminotransferase
Pictet-Spengler enzyme (Bet v1)	COesterase	Epimerase	Transporter	Other

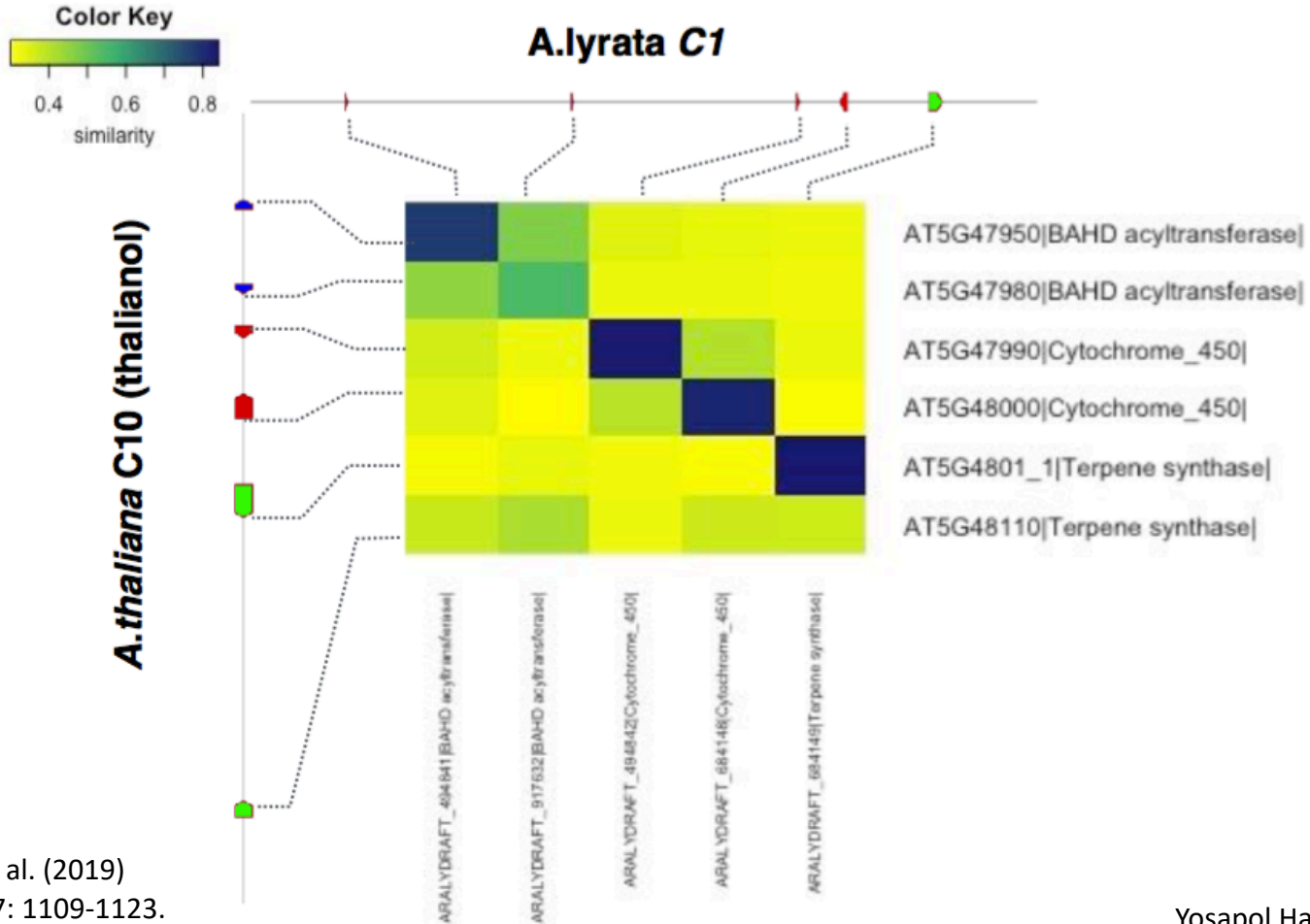
# PLANT TRITERPENE BIOSYNTHETIC LOCI ARE HIGHLY DYNAMIC



Hernando Suarez, Zhenhua Liu, Anne Osbourn

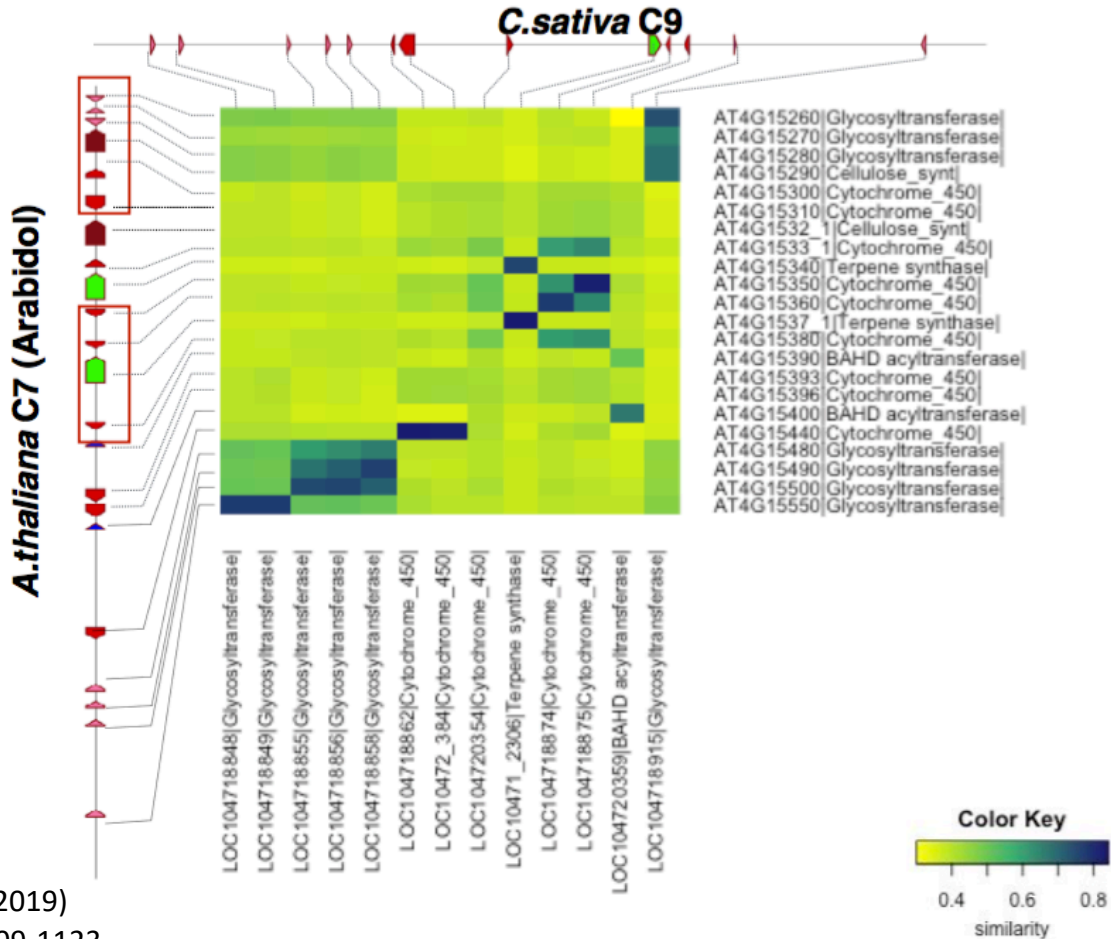
Liu, Suarez Duran et al. (2019)  
*New Phytologist* 227: 1109-1123.

# WITHIN THE SAME PLANT GENUS, SOME GENE CLUSTERS ARE CONSERVED



Liu, Suarez Duran et al. (2019)  
 New Phytologist 227: 1109-1123.

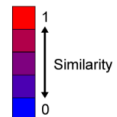
# ACROSS CLOSELY RELATED GENERA: LARGE DIFFERENCES!



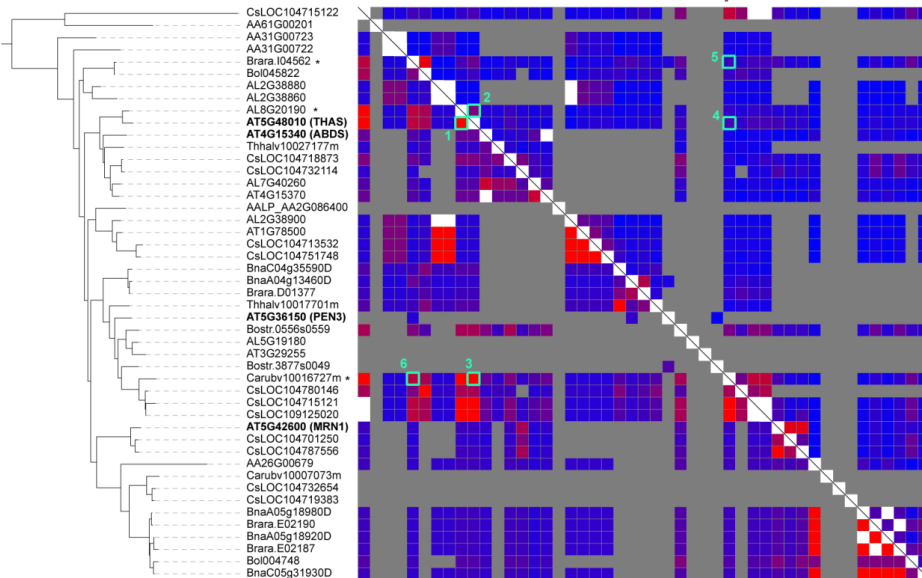
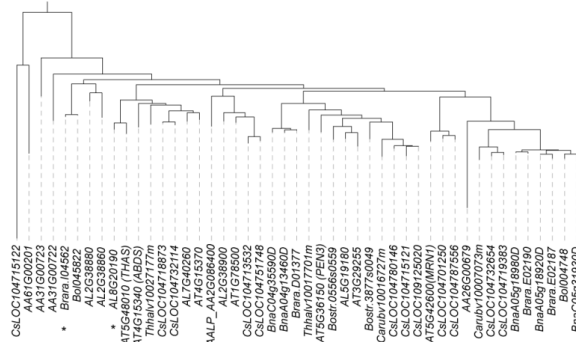
Liu, Suarez Duran et al. (2019)  
*New Phytologist* 227: 1109-1123.



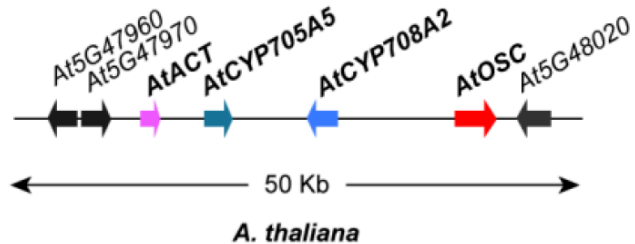
# PLANT TRITERPENE BIOSYNTHETIC LOCI ARE HIGHLY DYNAMIC



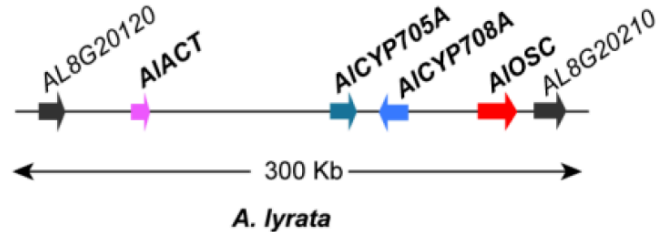
Same locus  
No domains in common



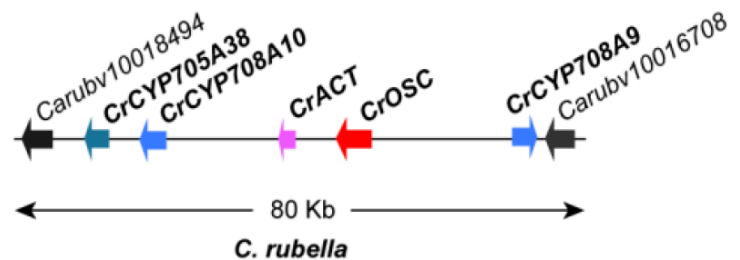
Domain Composition Similarity (Jaccard Index)



*A. thaliana*



*A. lyrata*

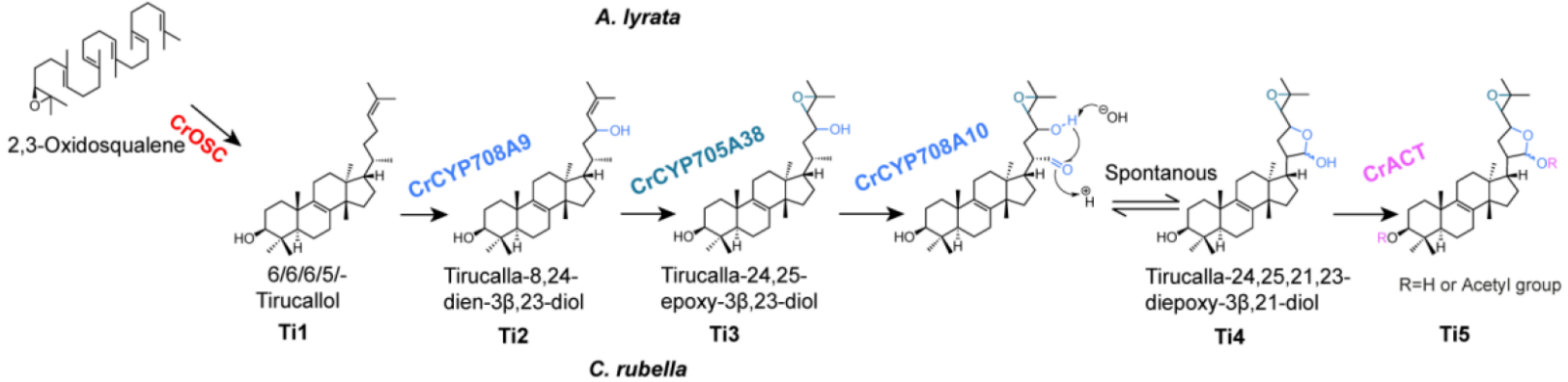
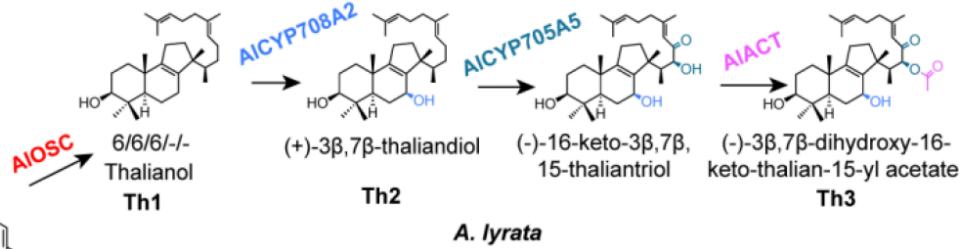
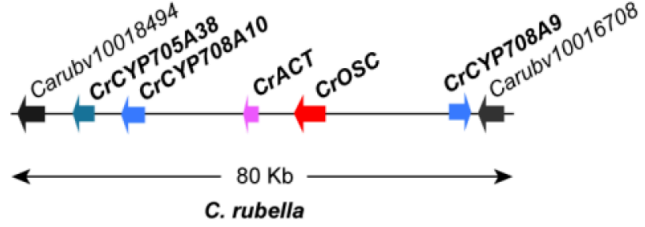
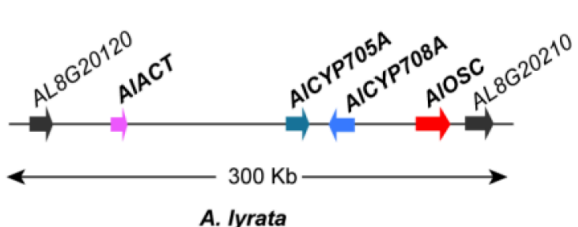


*C. rubella*

Sequence Similarity (DSS Index)

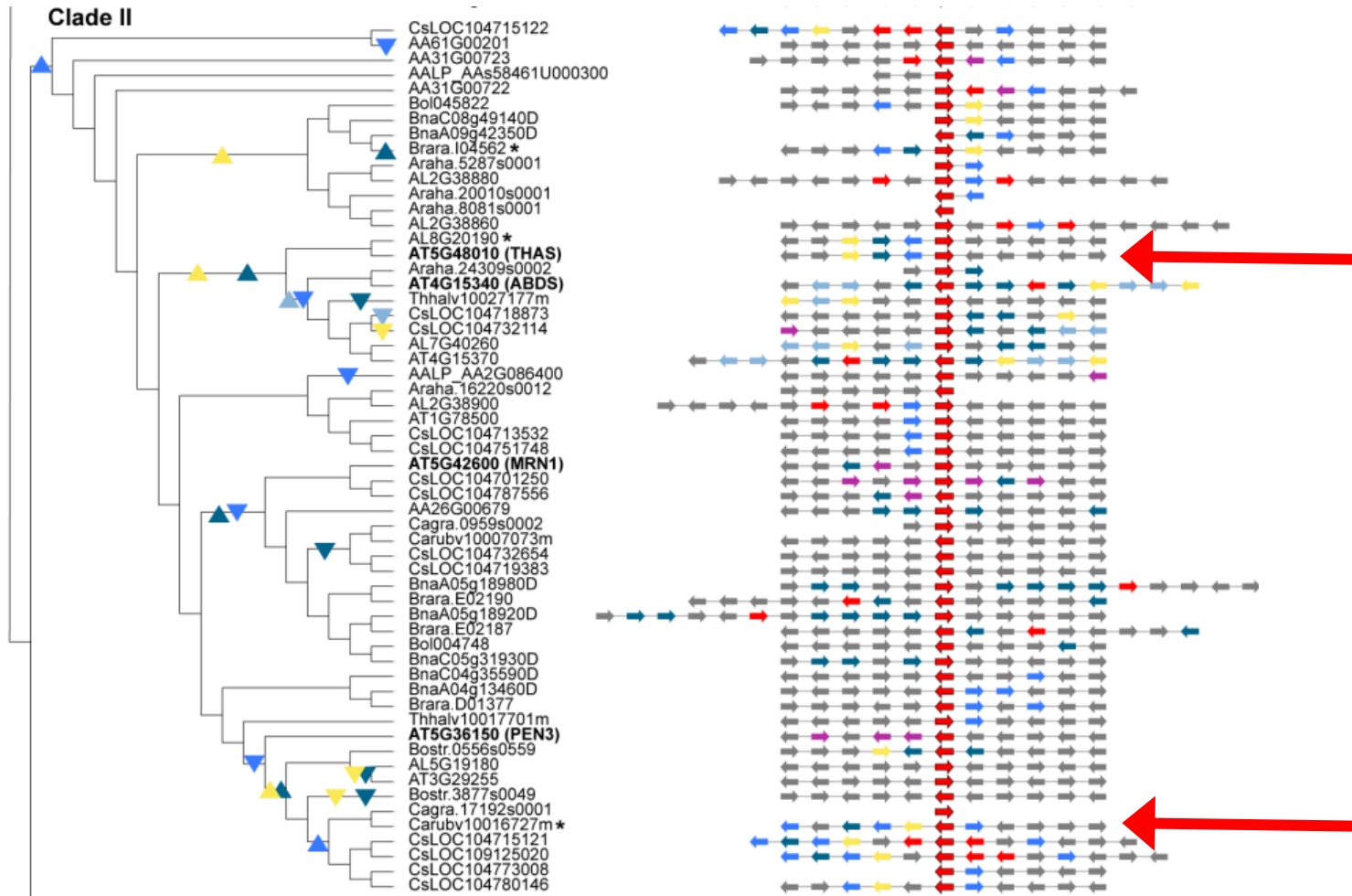
Liu, Suarez Duran et al. (2019)  
New Phytologist 227: 1109-1123.

# GENE CLUSTERS WITH IDENTICAL ENZYME (SUB)FAMILY CONTENT MAY PRODUCE DIFFERENT CHEMISTRY

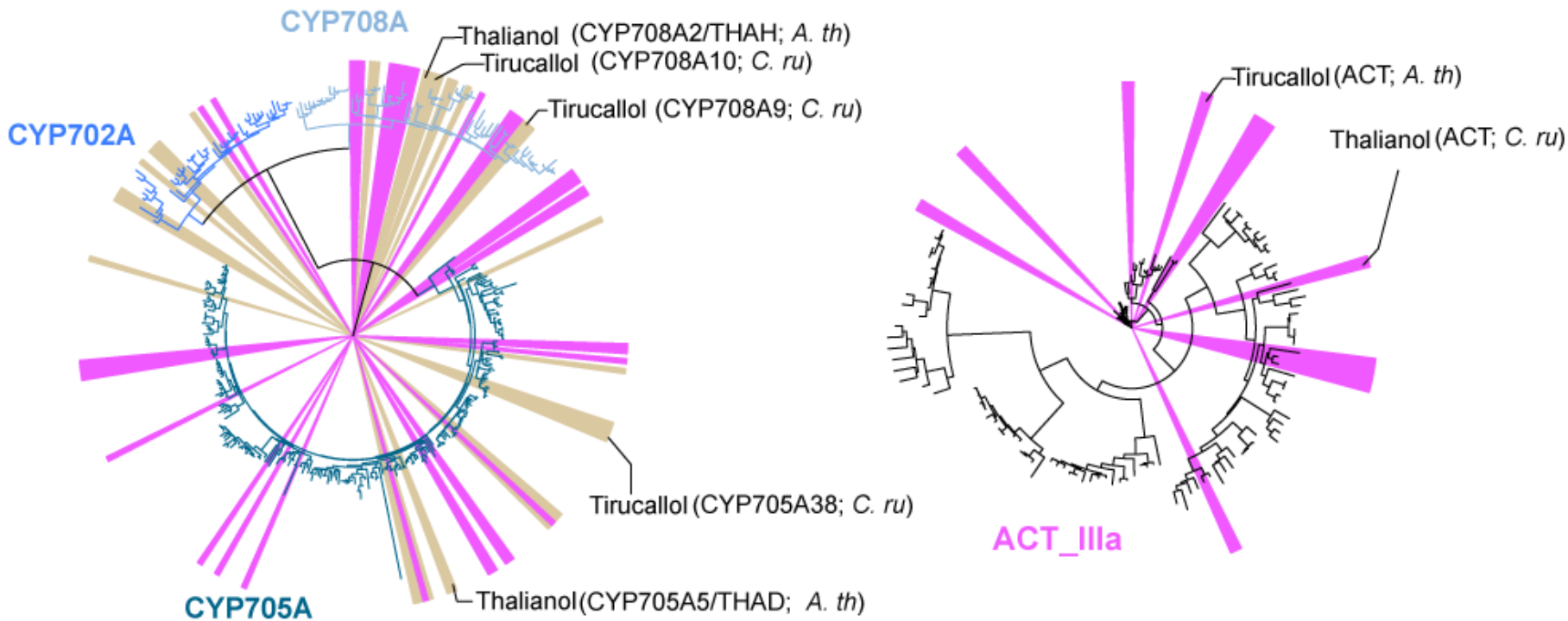


# TRITERPENE SYNTHASE PHYLOGENY SUGGESTS POSSIBLE INDEPENDENT EVOLUTION

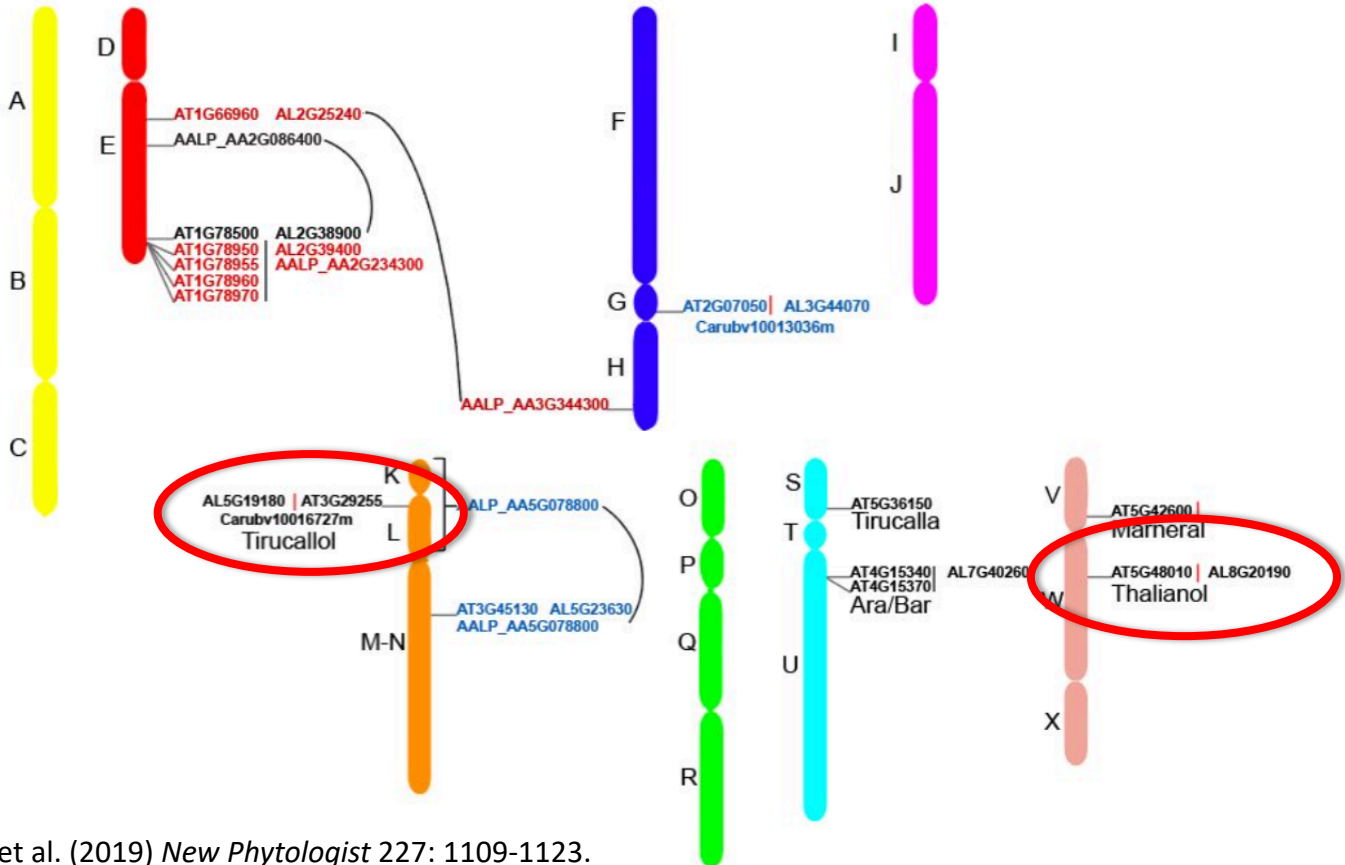
Liu, Suarez Duran et al. (2019) *New Phytologist* 227: 1109-1123.



# THE TIRUCALLOL AND THALIANOL BGCS CONTAIN P450S / ACYLTRANSFERASES FROM DIFFERENT CLADES OF THE SAME ENZYME SUBFAMILY

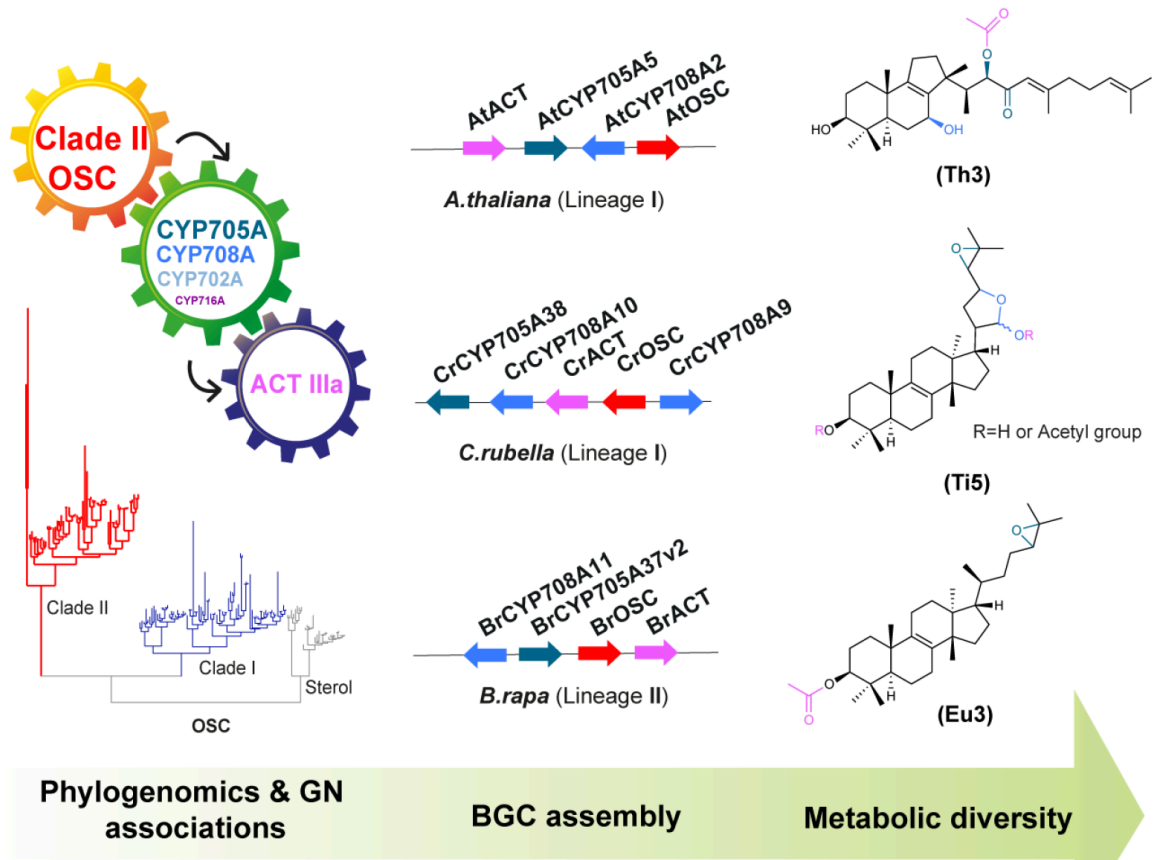


# TRITERPENE SYNTHASE GENES MAP BACK TO DISTINCT CHROMOSOMAL LOCATIONS ON THE ANCESTRAL CRUCIFER KARYOTYPE



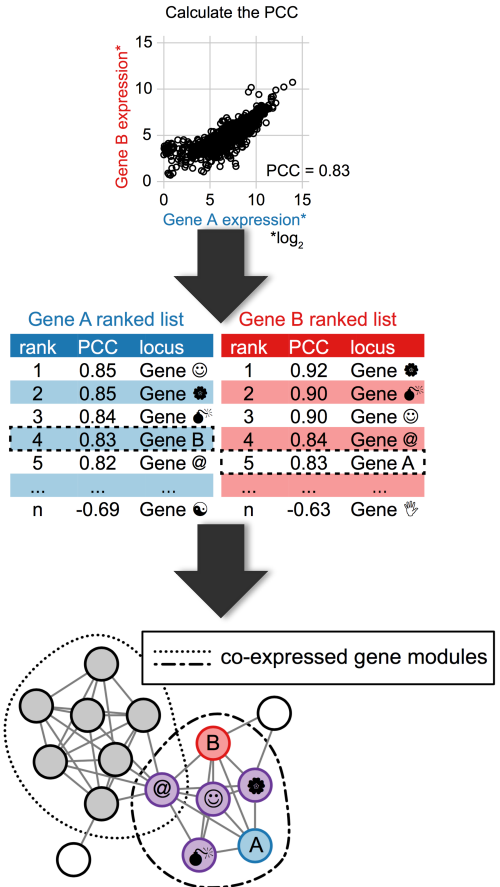
Liu, Suarez Duran et al. (2019) *New Phytologist* 227: 1109-1123.

# INDEPENDENT, DYNAMIC COMBINATORIAL EVOLUTION OF TRITERPENE DIVERSITY IN BRASSICACEAE: A MODEL FOR SYNTHETIC BIOLOGY

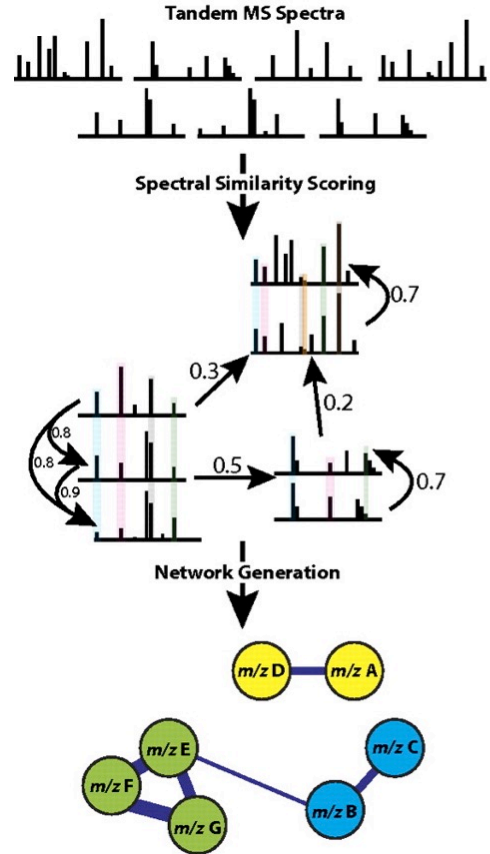


Liu, Suarez Duran et al. (2019) *New Phytologist* 227: 1109-1123.

# RECENT ADVANCES IN TRANSCRIPTOME/METABOLOME ANALYSIS PROVIDE FURTHER OPPORTUNITIES FOR UNSUPERVISED DISCOVERY

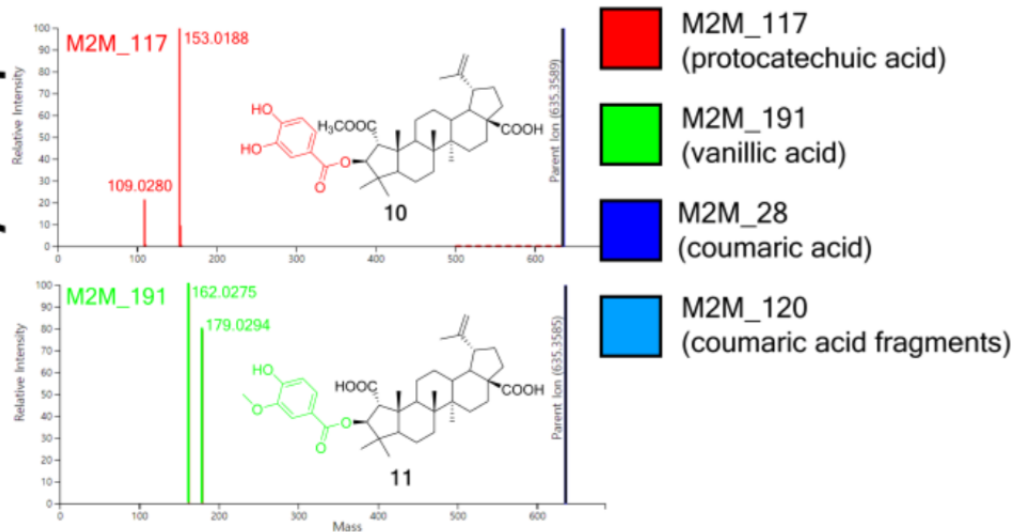
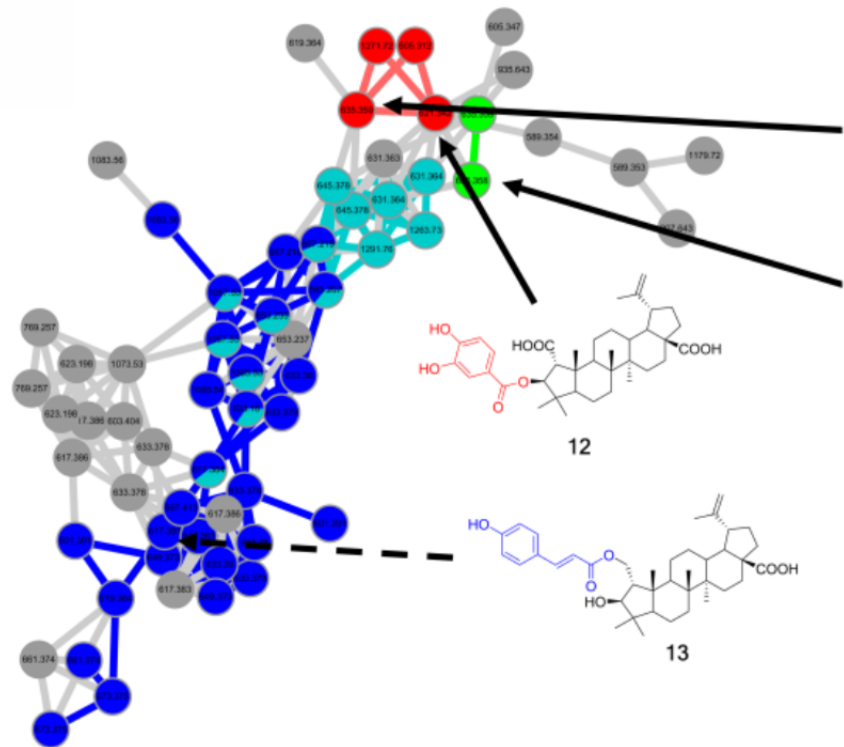


Wisecaver et al. (2017) *The Plant Cell*. 29 (5) 944-959.



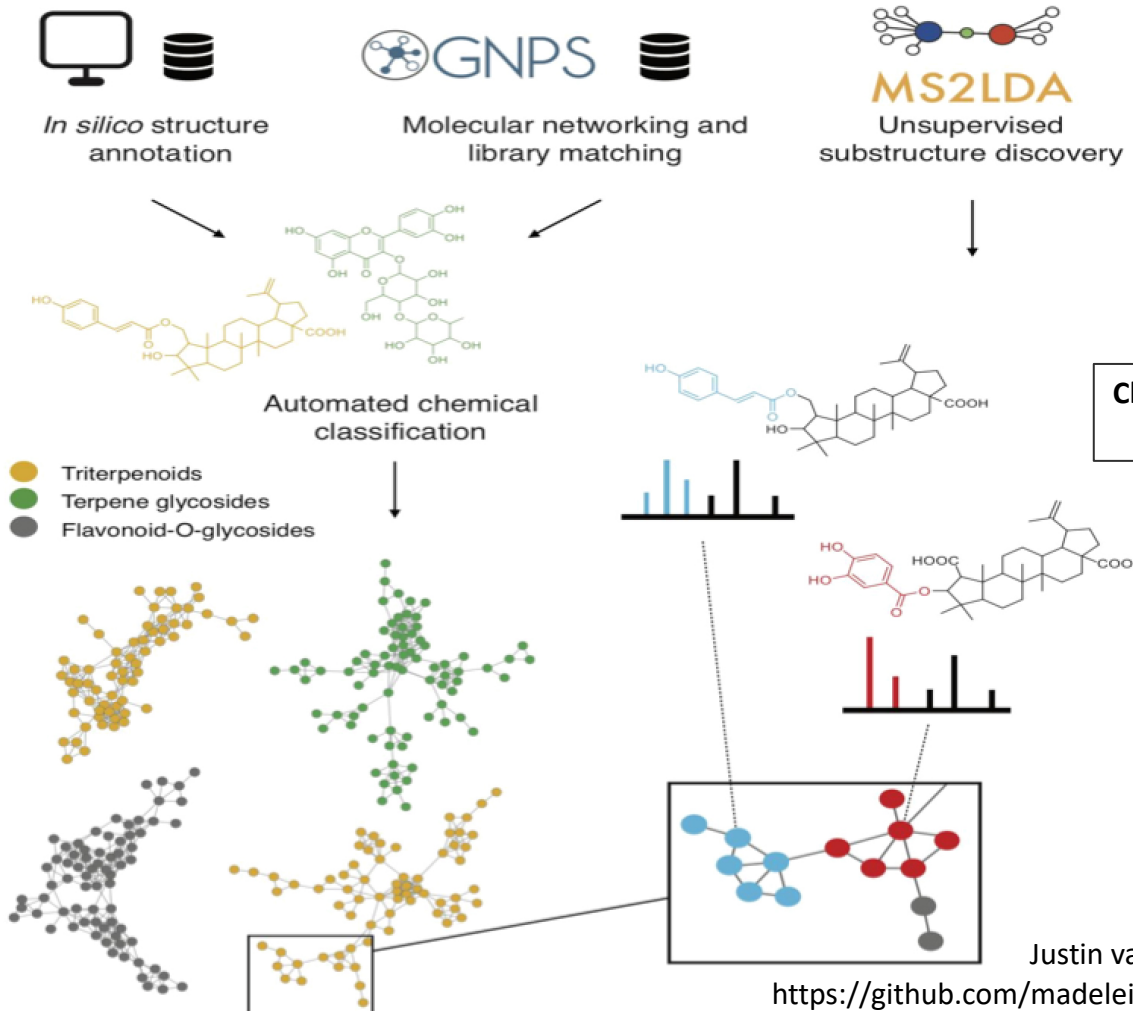
Watrous et al. (2012) *PNAS*. 109: E1743–E1752.

# MOLECULAR NETWORKS CAN BE ANNOTATED USING MOTIF FINDING AND OTHER ALGORITHMS FOR SUBSTRUCTURE PREDICTION





# THE MOLNETENHANCER WORKFLOW GENERATES ANNOTATED MOLECULAR NETWORKS



# THE PAIRED OMICS DATA PLATFORM CONNECTS GENOMICS WITH METABOLOMICS



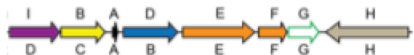
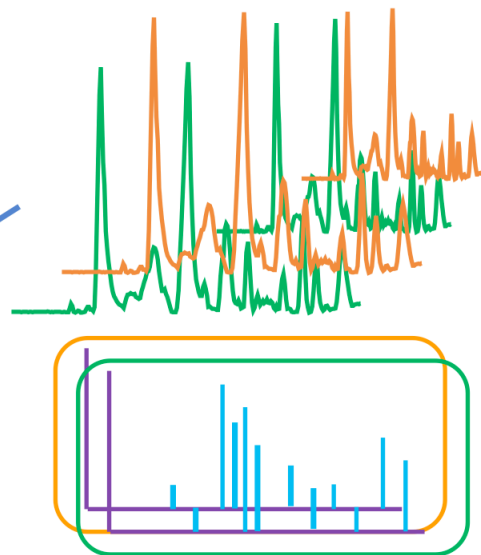
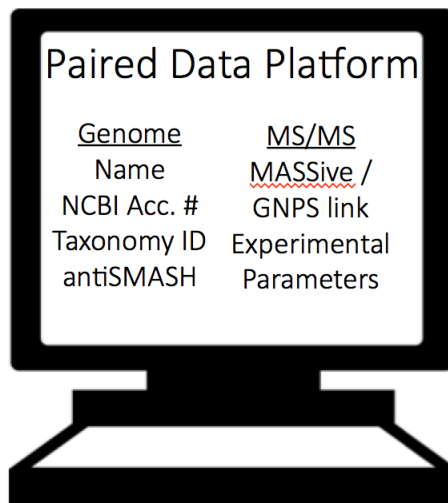
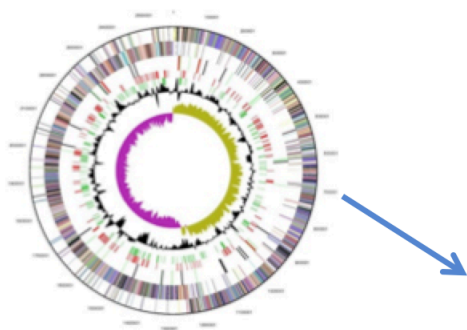
Dr. Justin van der Hooft



Dr. Michelle Schorn



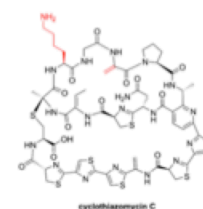
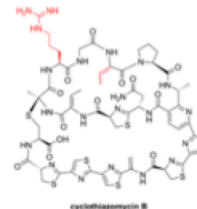
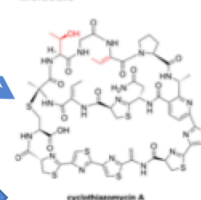
Dr. Stefan Verhoeven



Gene clusters  
Gene cluster families



Molecule



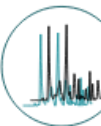
Molecules  
Molecular families

# THE PAIRED OMICS DATA PLATFORM CONNECTS GENOMICS WITH METABOLOMICS



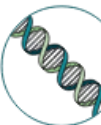
## 1. Submitter information

- Name
- Contact details



## 2. Metabolomics information

- GNPS-MassIVE/MetaboLights ID
- Molecular network ID
- PMID



## 3. (Meta)genomic information

- Genome ID
- BioSample

Genome label



## 4. Experimental details



### Sample growth conditions

- Medium
- Growth temperature, duration, OD
- Aeration

Sample growth condition label



### Extraction methods

- Solvent (ratio)
- Extracted material
- Resins

Extraction methods label



### Instrumentation methods

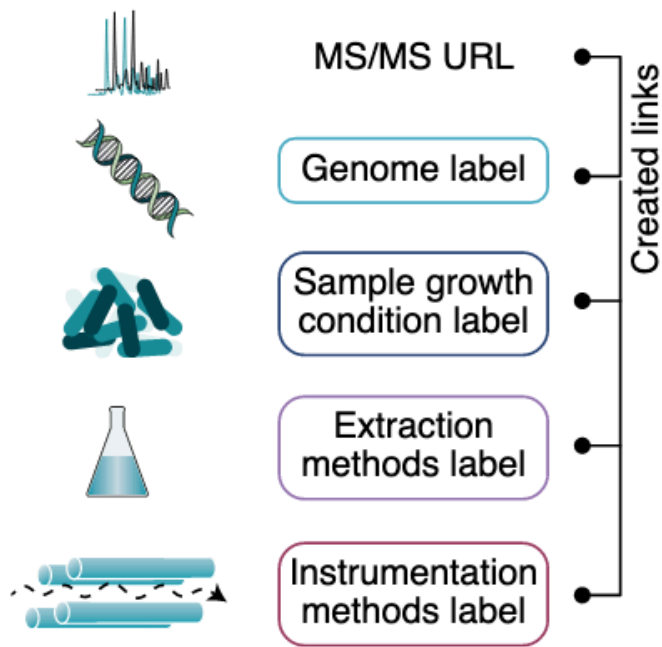
- Instrument type
- Ionization mode
- Mass range, CE

Instrumentation methods label

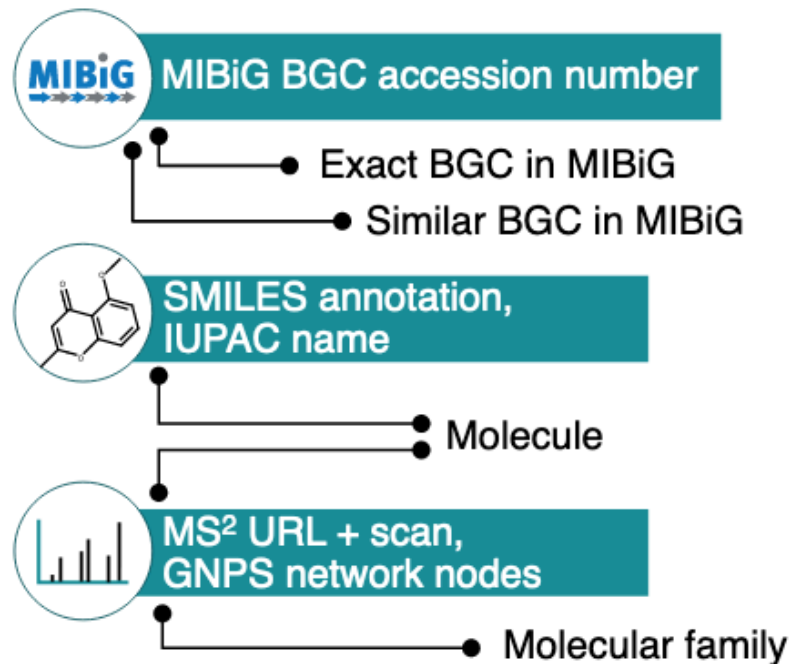
# THE PAIRED OMICS DATA PLATFORM CONNECTS GENOMICS WITH METABOLOMICS



## 5. Genome–metabolome



## 6. BGC–MS/MS



# THE MEDEMA RESEARCH GROUP @ WAGENINGEN UR

## Principal Investigator:

- > Marnix Medema

## Postdocs & PhD students:

- > **Hernando Suarez Duran** – until Dec 2019
- > **Satria Kautsar** – until Dec 2020
- > **Michelle Schorn** – until Feb 2019
- > Vittorio Tracanna
- > Victòria Pascal Andreu
- > Barbara Terlouw
- > Lotte Pronk
- > Hannah Augustijn
- > Zachary Reitz
- > Mohammad Alanjary
- > Huali Xie



## WU Bioinformatics colleague PI:

- > **Justin van der Hoof**

Part of the Wageningen UR Bioinformatics Department,  
headed by Prof. Dick de Ridder

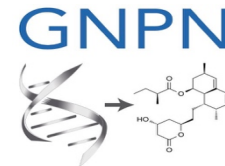
## MSc students:

Jeanine Boot, Bram van Wersch, Liana van Grieken,  
Matthias van den Belt, Thijs Finnegan

## Funding:



European Research Council  
Established by the European Commission



# MANY THANKS TO COLLABORATORS



- > Anne Osbourn
- > Zhenhua Liu
- > Charlie Owen



- > Pieter Dorrestein
- > Kyo Bin Kang



- > Tilmann Weber
- > Kai Blin
- > Simon Shaw

netherlands



- > Lars Ridder
- > Florian Hüber
- > Stefan van der Hoeven