Improved small molecule structural annotation by exploiting biochemical and biosynthetic relationships

Justin J.J. van der Hooft et al.

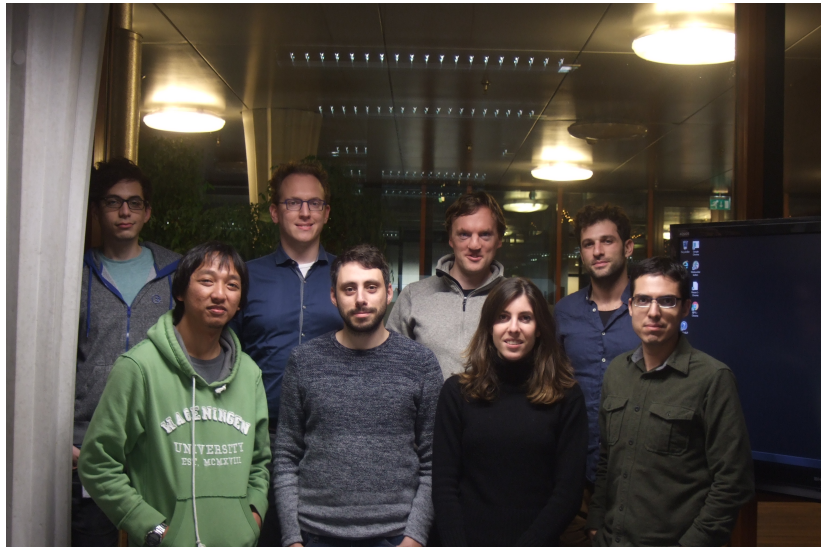
Bioinformatics Group – Wageningen University, The Netherlands

Antwerpen, 30 November 2018

WAGENINGEN
UNIVERSITY & RESEARCH

100years

Team work! ☺

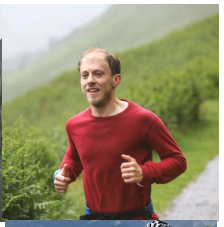


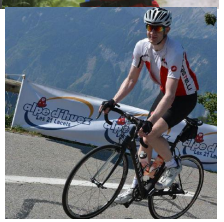
Medema lab - Wageningen UR, NL

NL eScience Center

Dorrestein lab – San Diego, USA

Glasgow Polyomics – University of Glasgow, UK





WAGENINGEN UNIVERSITY & RESEARCH

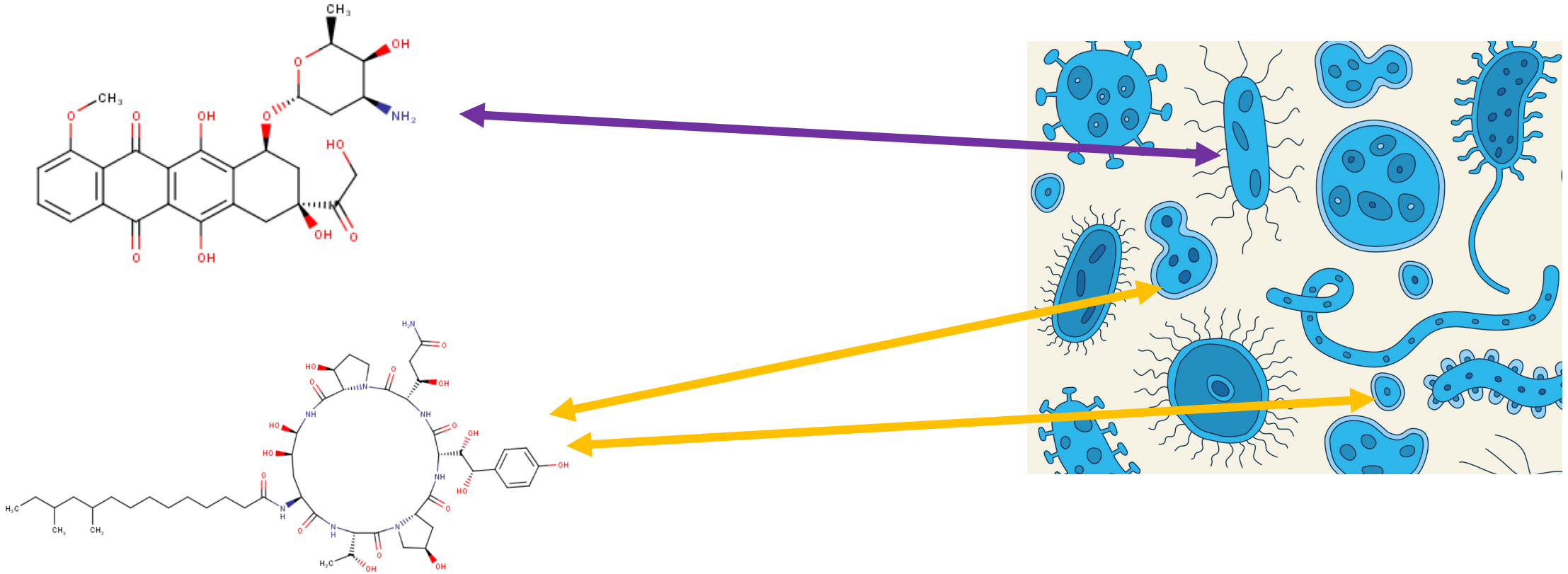
€€ Funding €€ NWO eScience center

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

University of Glasgow

University of Strathclyde Glasgow

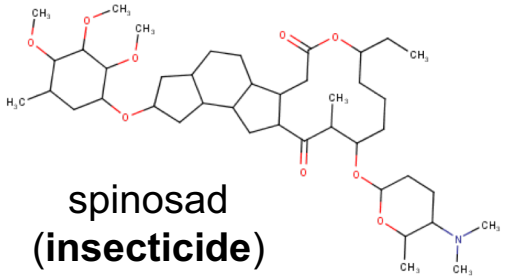
iOMEGA project:
integrated omics for metabolomics and genomics annotation



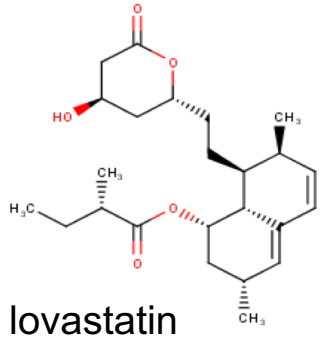
WAGENINGEN
UNIVERSITY & RESEARCH

The challenge….

Bacteria, fungi, and plants produce a large & diverse arsenal of high-value molecules:

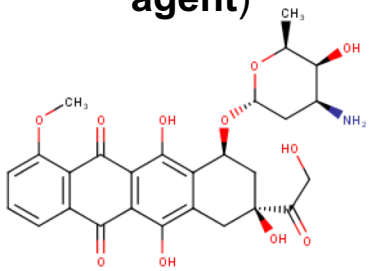


spinosad (**insecticide**)

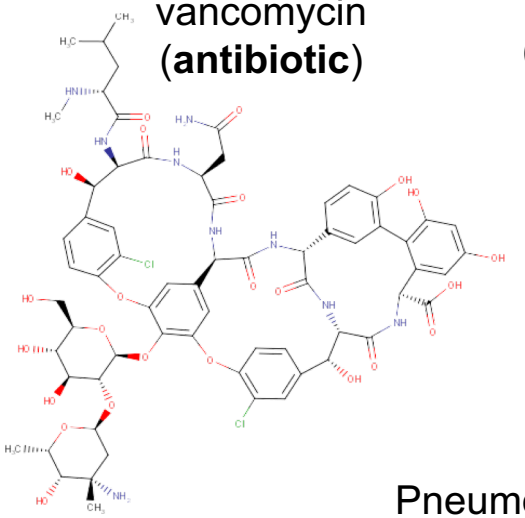


lovastatin (**cholesterol lowering agent**)

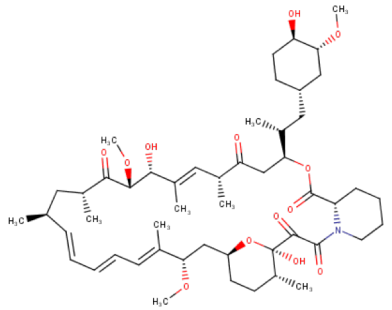
doxorubicin (**chemotherapeutic agent**)



vancomycin (**antibiotic**)

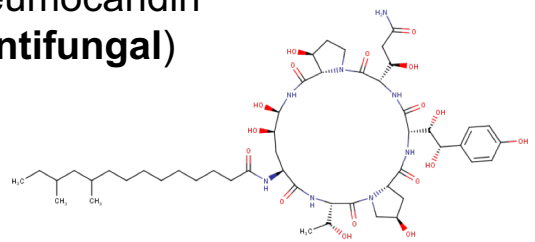


rapamycin (**immunosuppressant**)



Mass spectrometry fragmentation spectrum

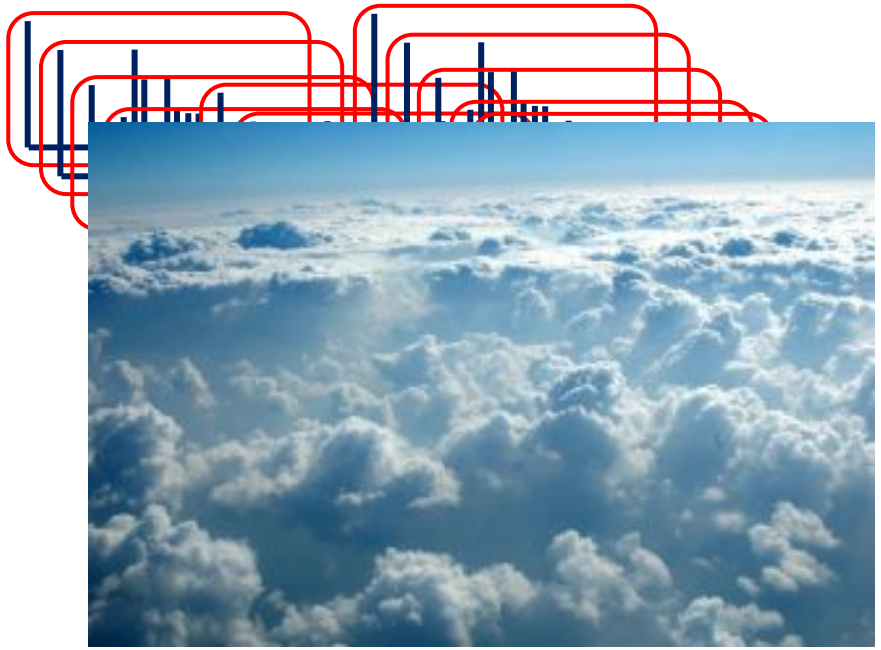
Pneumocandin (**antifungal**)



….is **large-scale coupling of spectral data to molecular structures** of known & especially **novel** natural products molecules

WAGENINGEN UNIVERSITY & RESEARCH

Motivation



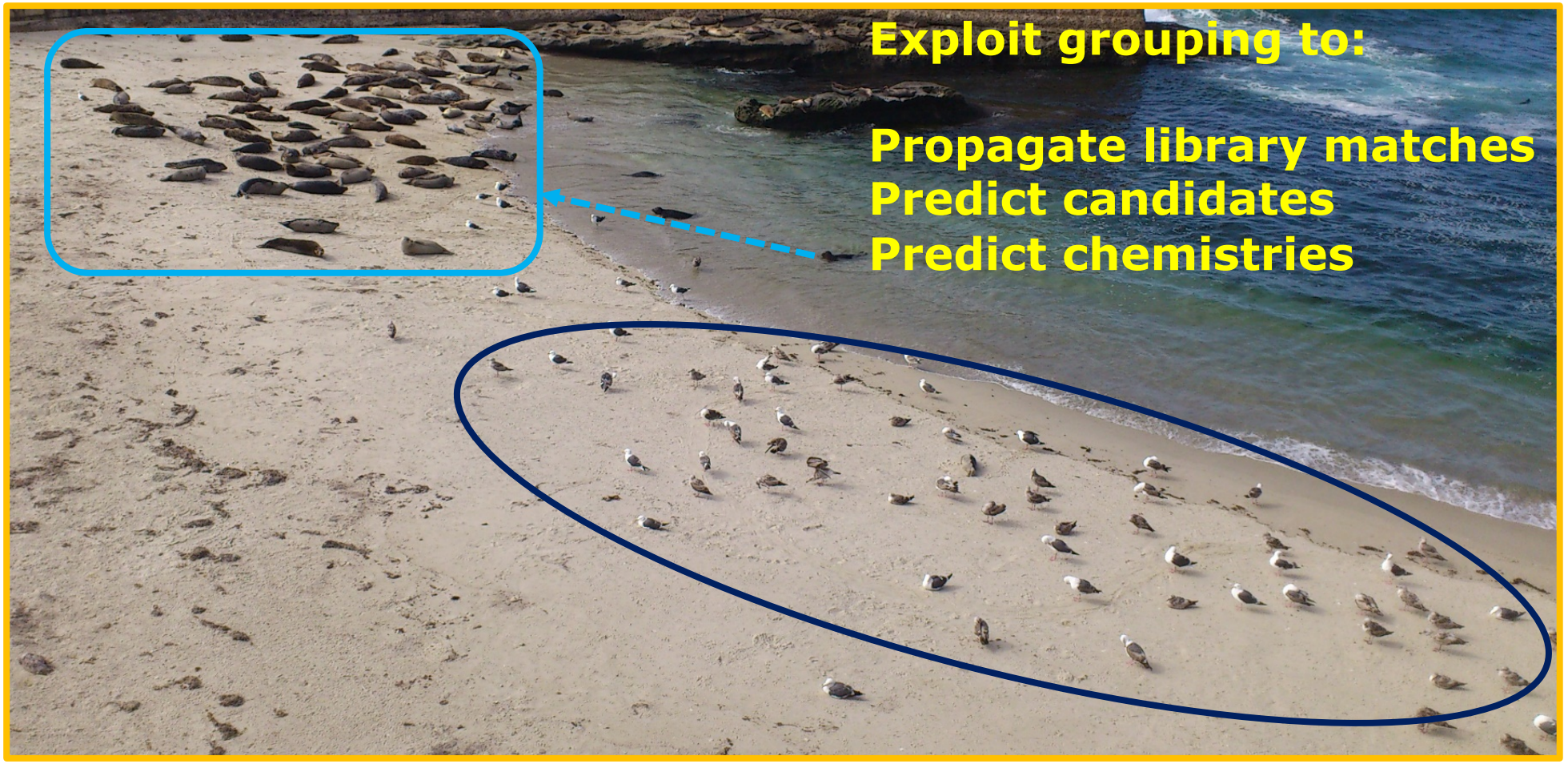
MS/MS information is key to study large sample sets with diverse chemistries

Integrated workflow exploits complementary tools to enhance interpretation

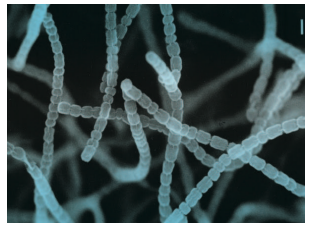
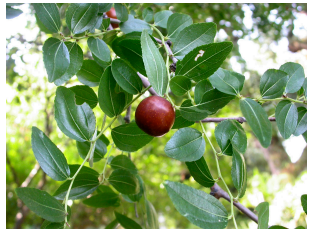
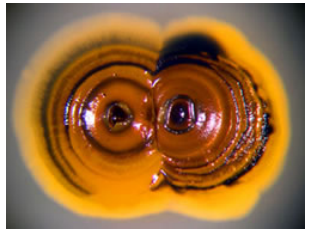
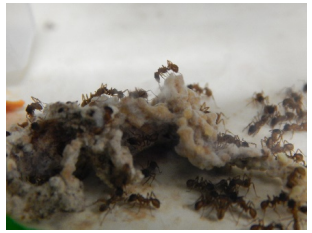
Motivation: Improved annotation power by pattern mining

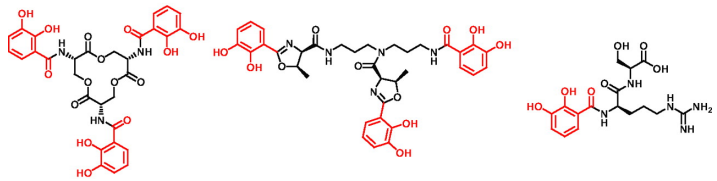
Molecular Families based on spectral similarity (Molecular Networking in GNPS)

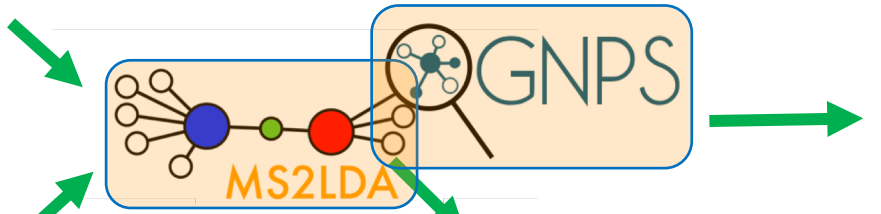
Substructures by extracting "building blocks of metabolomics" (MS2LDA)

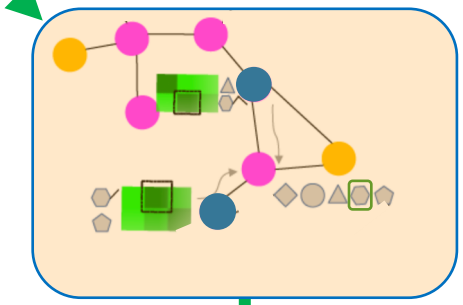


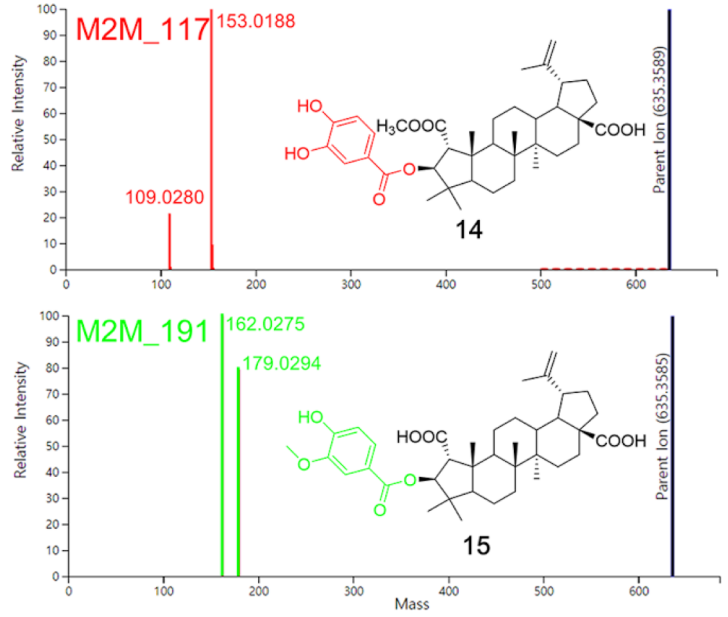
Integrated workflow

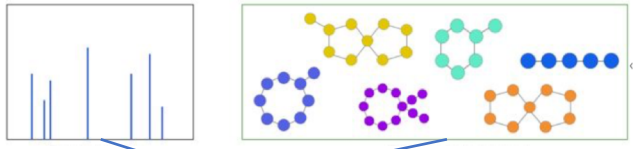





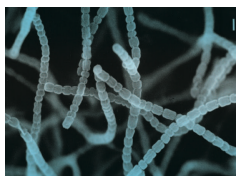


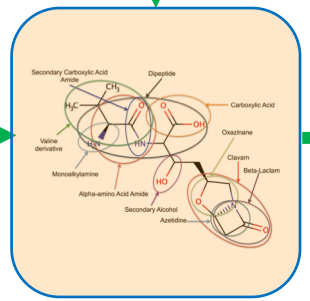


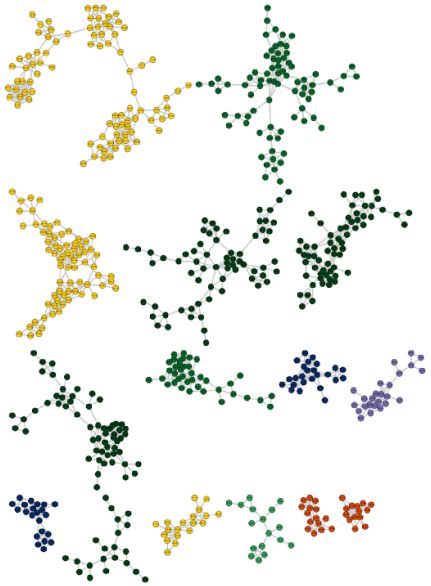






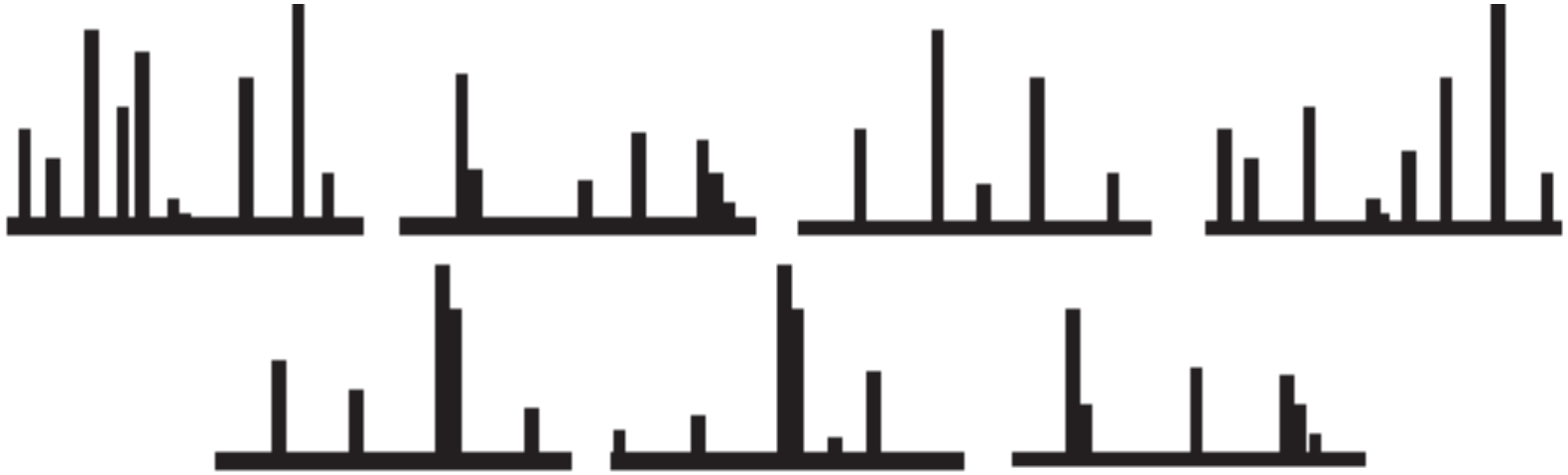






Dr Madeleine Ernst, UCSD

WAGENINGEN
UNIVERSITY & RESEARCH



Molecular Networking

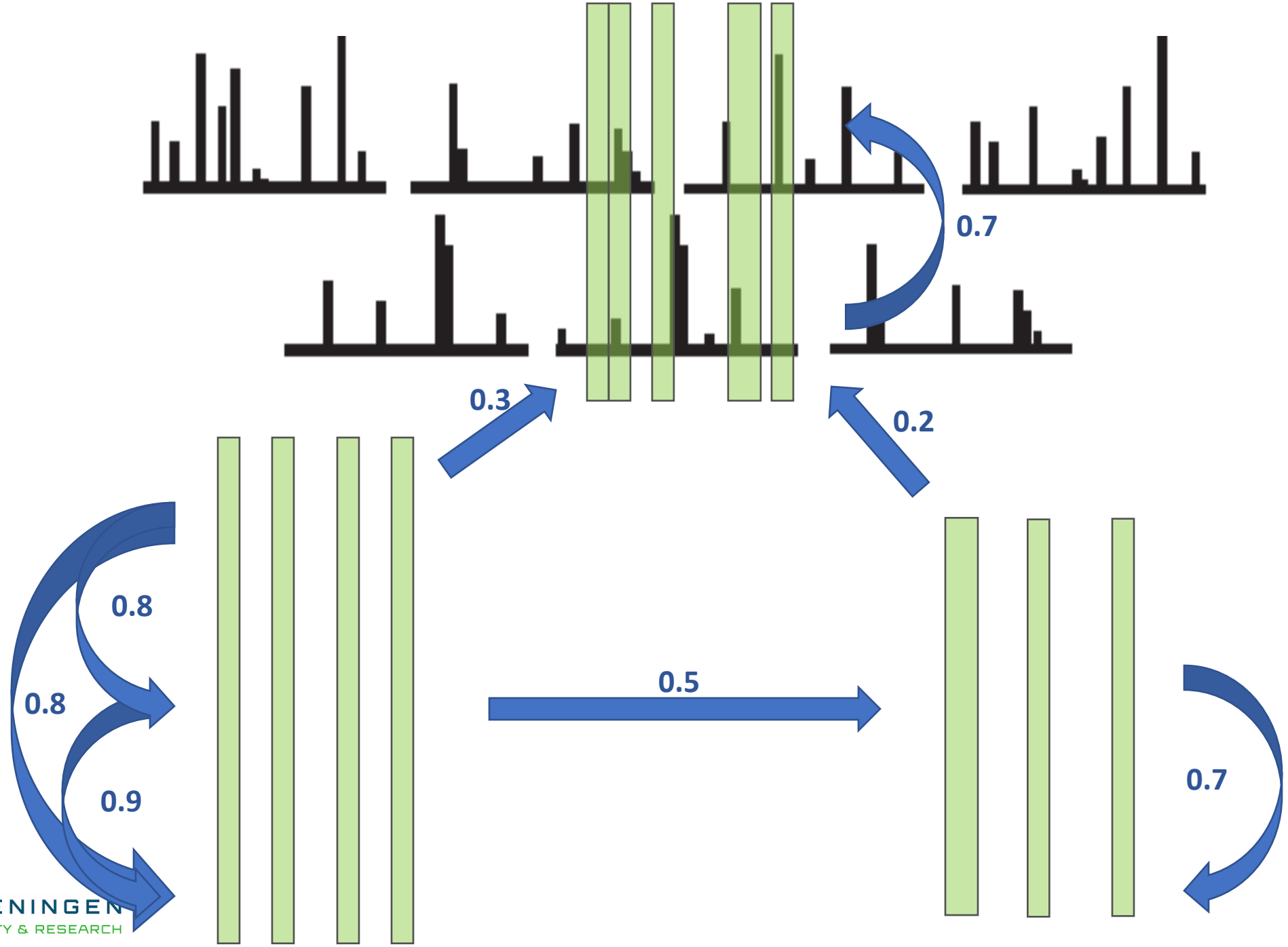
Very similar MS/MS spectra are grouped to:

- link spectra across different samples
- reduce redundancy in data set (”consensus spectrum”)

Wang, M., et al., *“Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking”* Nat Biotech (2016)

Watrous, JD *et al*. “Mass spectral molecular networking of living microbial colonies” *PNAS* (2012)

Dr Ricardo R. da Silva
Dr Madeleine Ernst
Dr Mingxun Wang
Dr Louis-Felix Nothias



WAGENINGEN
UNIVERSITY & RESEARCH

Spectral library matches from GNPS

Libraries from diverse sources

Seed node annotations for molecular families

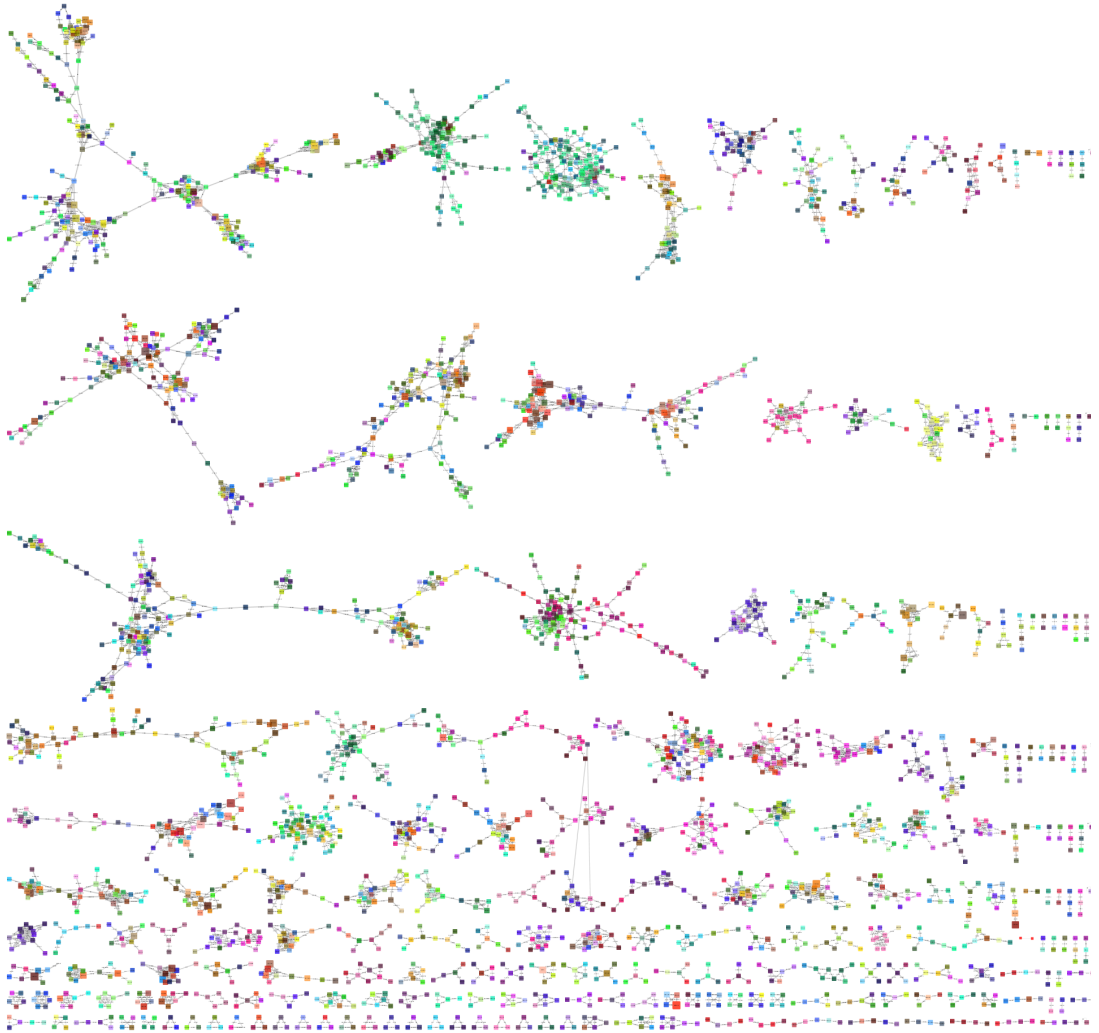
Library MS/MS Spectra

Diterpenoids

Flavonoids

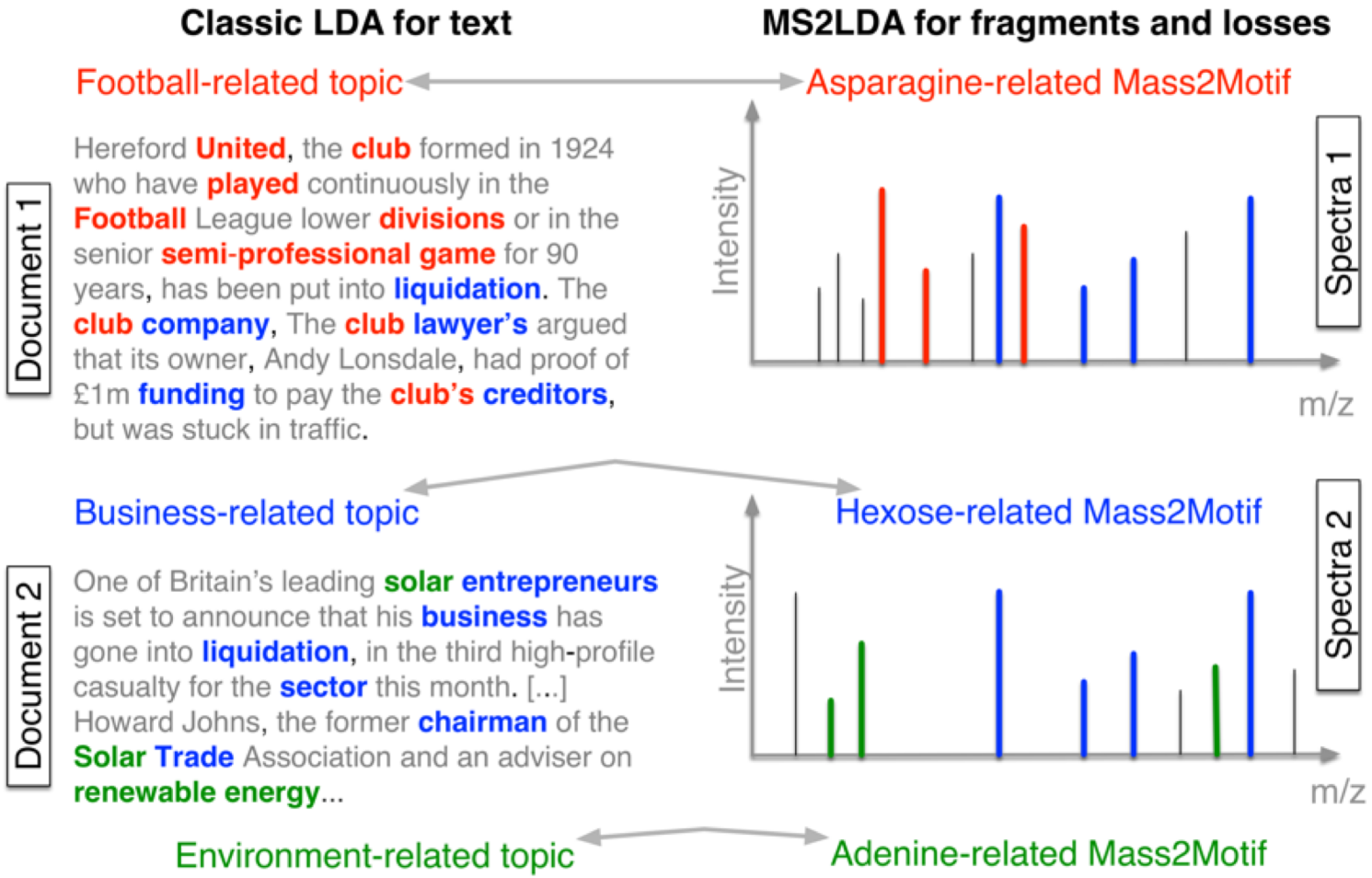
Pharmaceuticals

MS contaminants



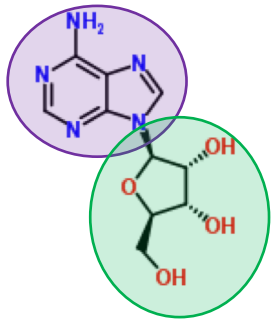
WAGENINGEN
UNIVERSITY & RESEARCH

Topic modelling: from text to molecules



MS2LDA

Unsupervised Substructure Discovery



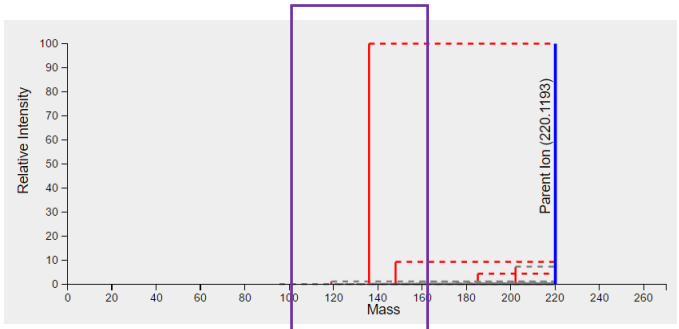
Documents <-> molecules

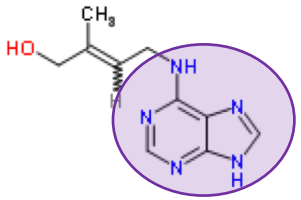
Words <-> fragments/neutral losses

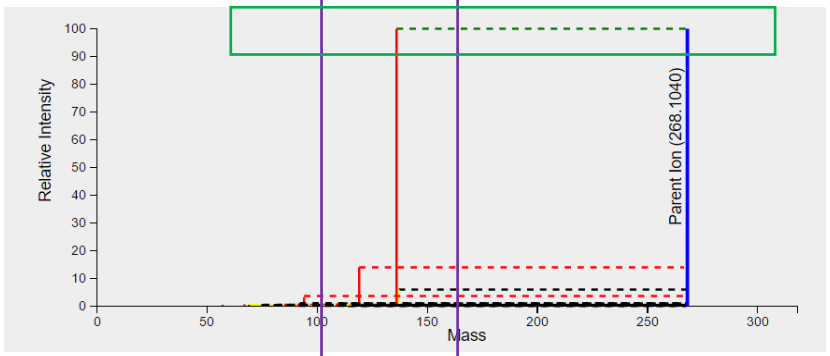
Van der Hooft et al., PNAS, 2016

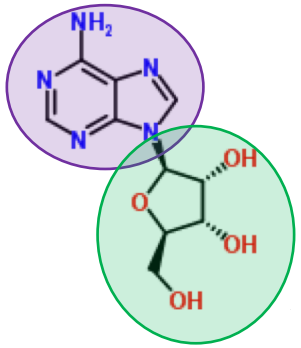
WAGENINGEN UNIVERSITY & RESEARCH

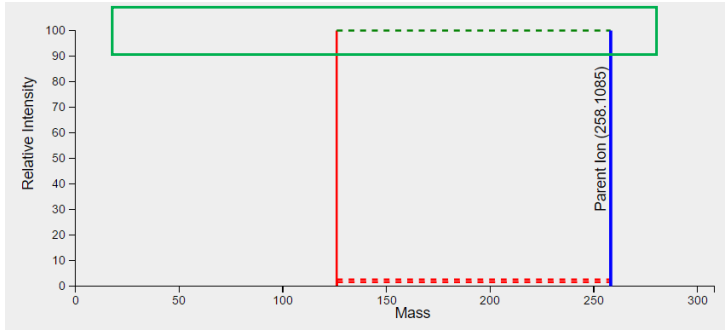
Validation: **MS2LDA** with standards

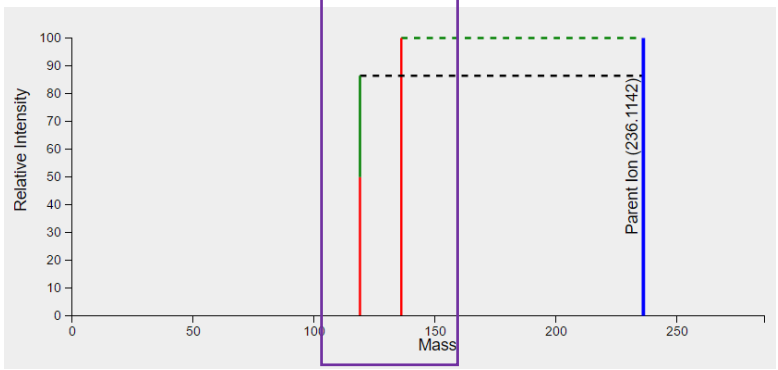


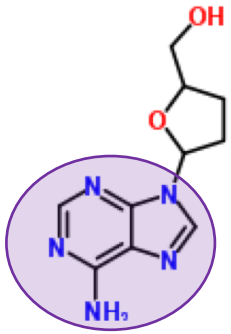


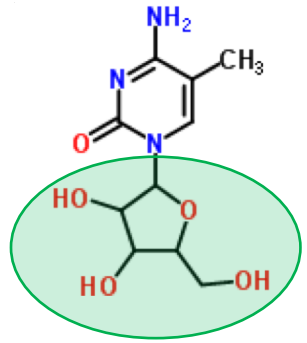


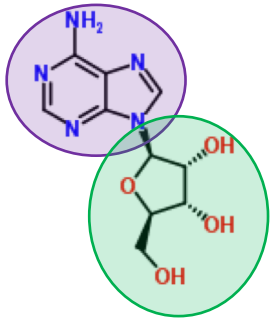












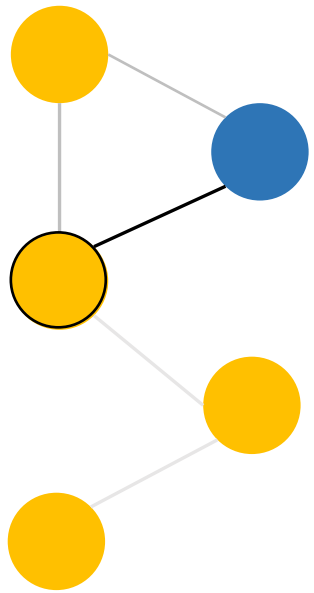
WAGENINGEN
UNIVERSITY & RESEARCH

Network annotation propagation

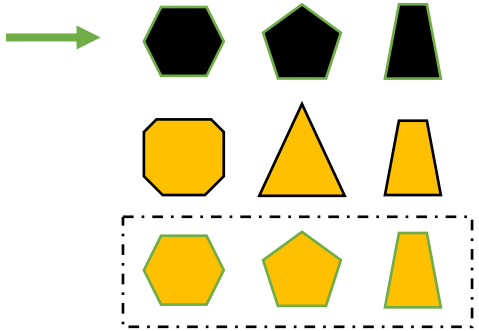
More coherent candidate structures within family

Rerank list based on **neighboring annotations**

Exploiting the **network topology**



Molecular Fingerprints

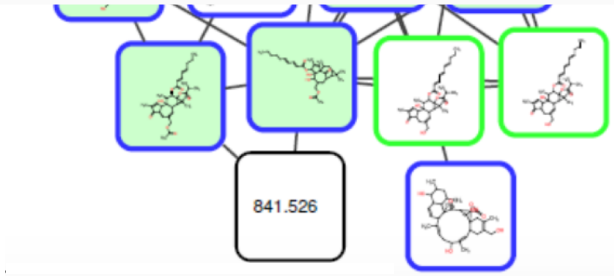


Ranked lists of in silico annotators (MetFrag)

Reranked(!) lists of in silico annotators

Library Match

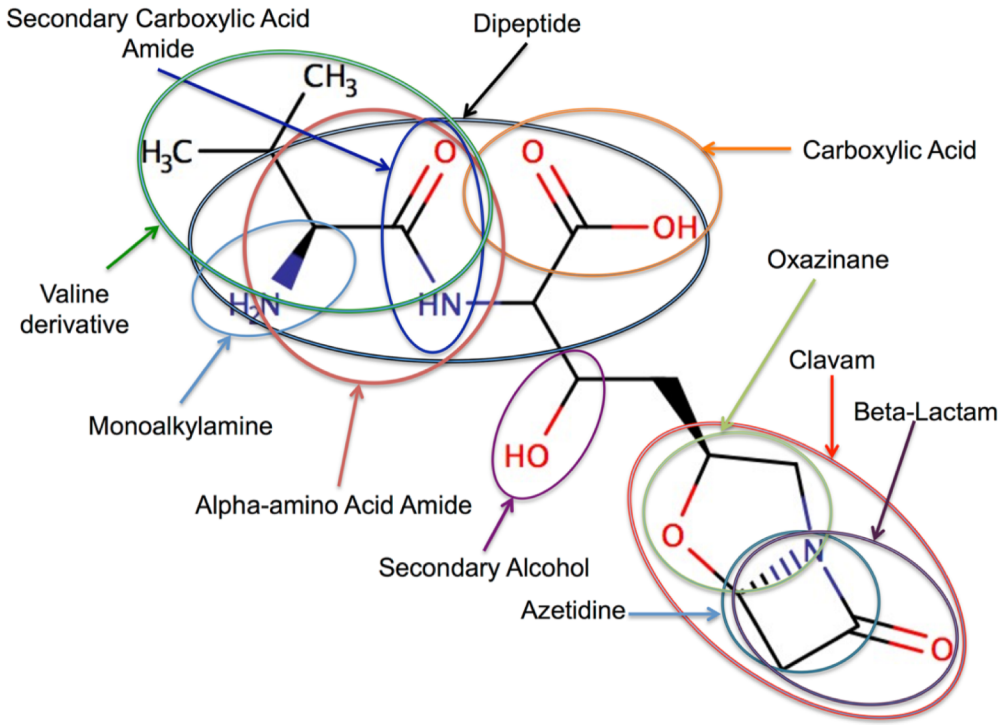
Scans	Assigned	Candidate score	Molecule	Formula	Mass	ΔMass (ppm)	Name
1	No	2.23513		C13H16N2O5S	312.07799	0.86666	acetaminophen mercapturate (83967)
1	No	2.23513		C13H16N2O5S	312.07799	0.86666	(56923619)
1	No	3.20109		C13H16N2O5S	312.07799	0.86666	AKOS019768418 (81197200)
1	No	3.22166		C13H16N2O5S	312.07799	0.86666	CHEMBL1642089 (53321314)
1	No	3.22166		C13H16N2O5S	312.07799	0.86666	CHEMBL1642098 (53318691)
1	No	3.34424		C13H16N2O5S	312.07799	0.86666	AGN-PC-0G4R64 (64889340)
1	No	3.42103		C13H16N2O5S	312.07799	0.86666	AGN-PC-0DZQ65 (61409875)
1	No	3.42875		C13H16N2O5S	312.07799	0.86666	AGN-PC-0E3QTW (61600272)

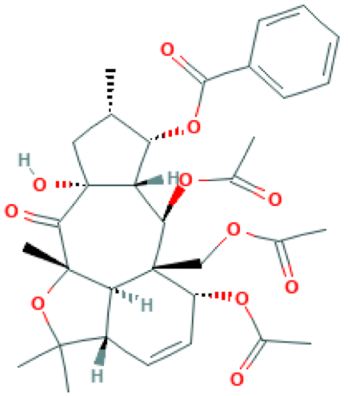


WAGENINGEN UNIVERSITY & RESEARCH

Da Silva et al., Plos Comp Biol., 2018

ClassyFire – SMILES and substituents





INPUT:

CC1CC2(C(C1OC(=O)C3=CC=CC=C3)C(C4(C(C=CC5C4C(C2=O)(OC5(C)C)C)OC(=O)C)COC(=O)C)OC(=O)C)O

OUTPUT:

ClassyFire

Kingdom: Organic compounds
Superclass: Lipids and lipid-like molecules
Class: Prenol lipids
Subclass: Diterpenoids
...

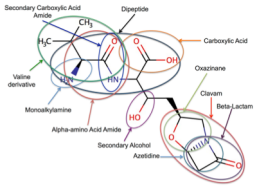
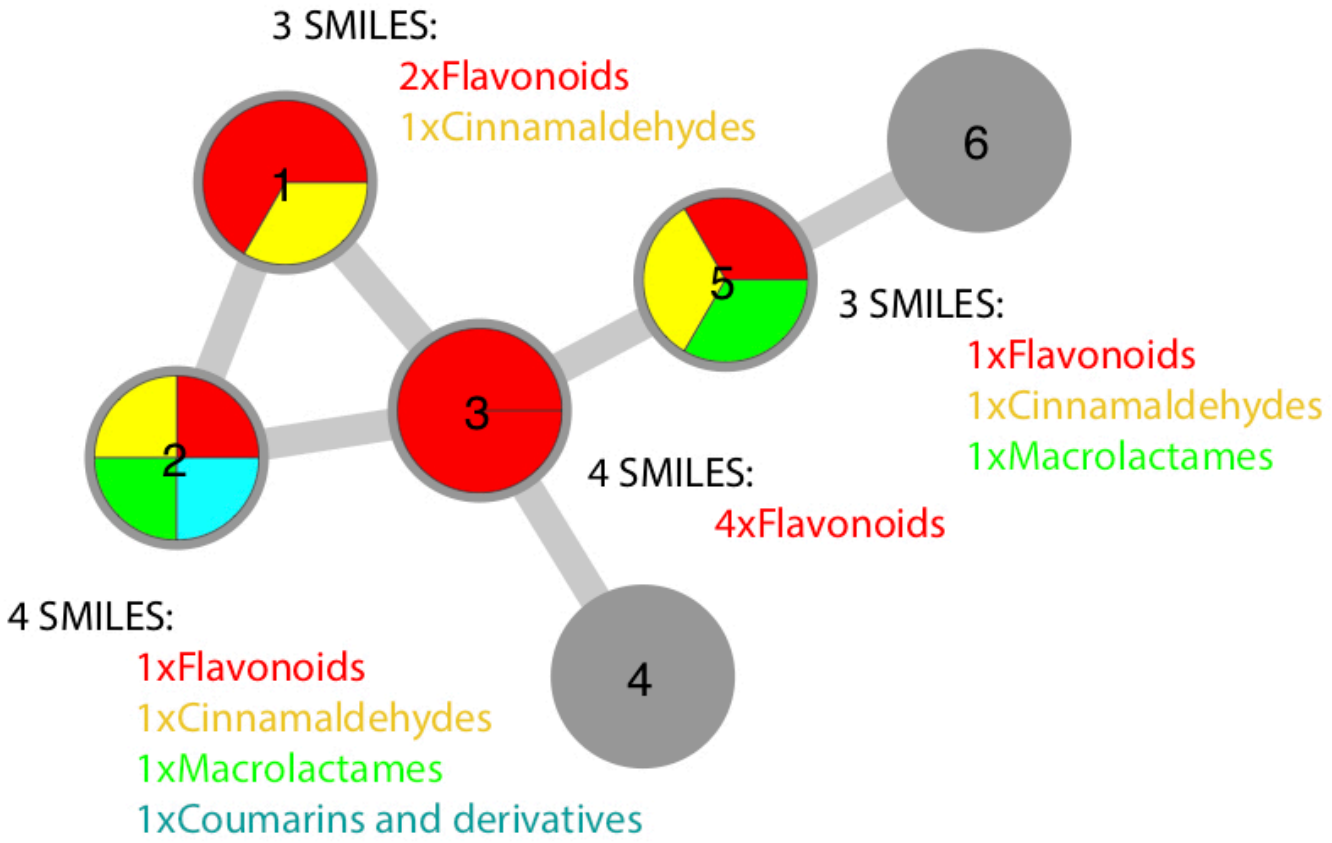


Figure part of Fig. 1 from Djoumbou Feunang et al., J. Cheminform, 2016

Djoumbou Feunang et al., J. Cheminform, 2016

ClassyFire – Chemical predictions for MFs



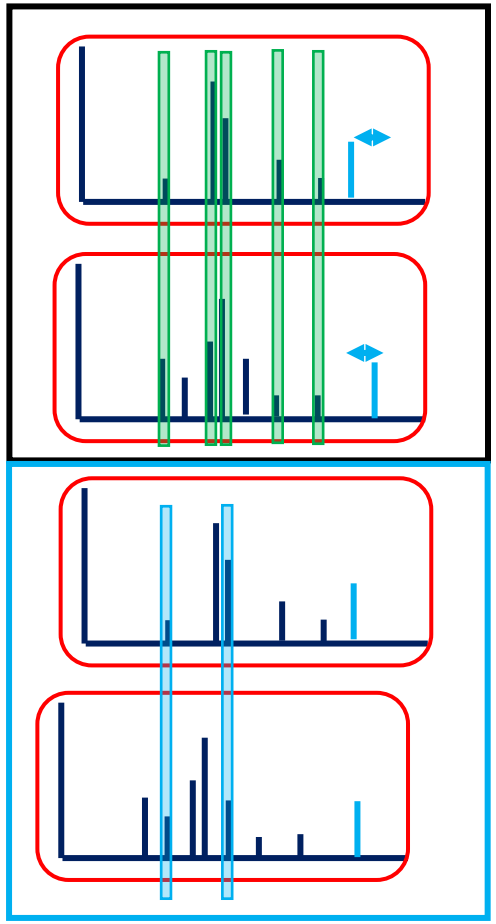
ClassyFire Scoring:

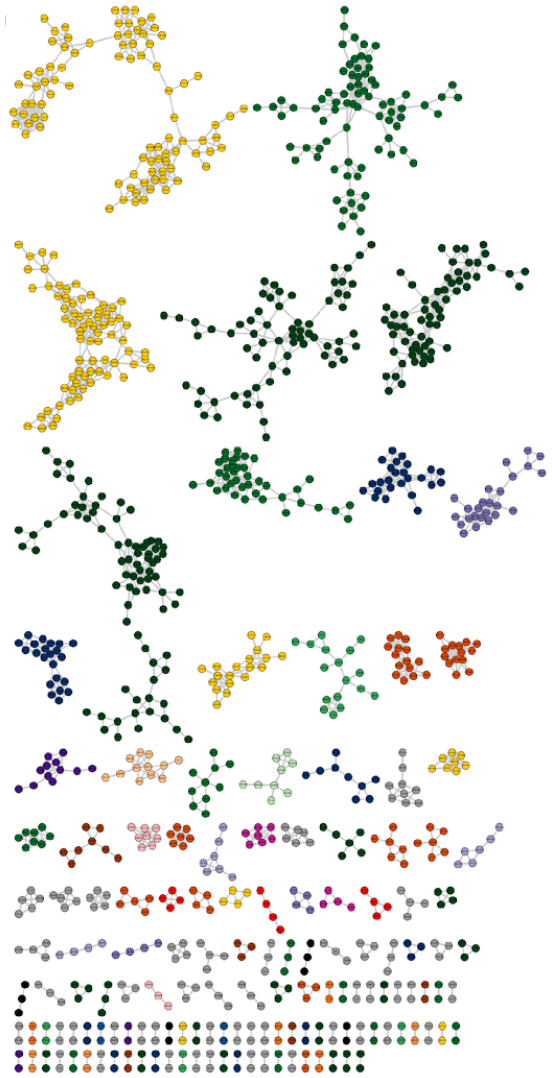
Flavonoids:	Database hits:
2.25 nodes/	4 nodes/
6 nodes =	6 nodes =
0.375	**0.67**

Disclaimer: scores dependent on structural database contents!

Illuminating the Rhamnaceae chemistry

Molecular Networking





- triterpenoids
- triterpenoidal saponins
- steroidal saponins
- flavonoid 3-*O*-glycosides
- flavonoid 7-*O*-glycosides
- flavonoid *O*-glycosides
- flavones, flavonones, flavonols
- 8-methylated flavonoids
- (hydroxyl)anthraquinones
- xanthones
- hydrolyzable tannins
- phenolic glycosides
- coumaric acid and derivatives
- lignan glycosides
- iridoid glycosides
- peptide
- hybrid peptide
- oligosaccharides
- long-chain fatty acids
- others / no matches

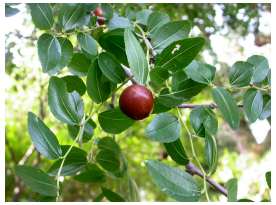
plant related classifications:

different flavonoids

phenolic glycosides

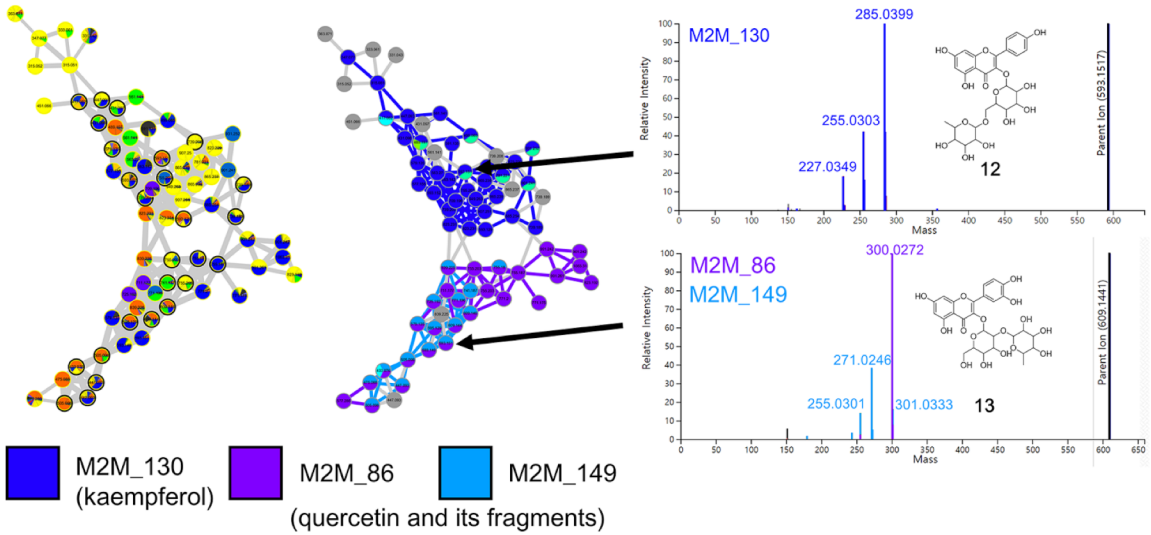
triterpenoids

Dr Kyo Bin Kang, UCSD



WAGENINGEN
UNIVERSITY & RESEARCH

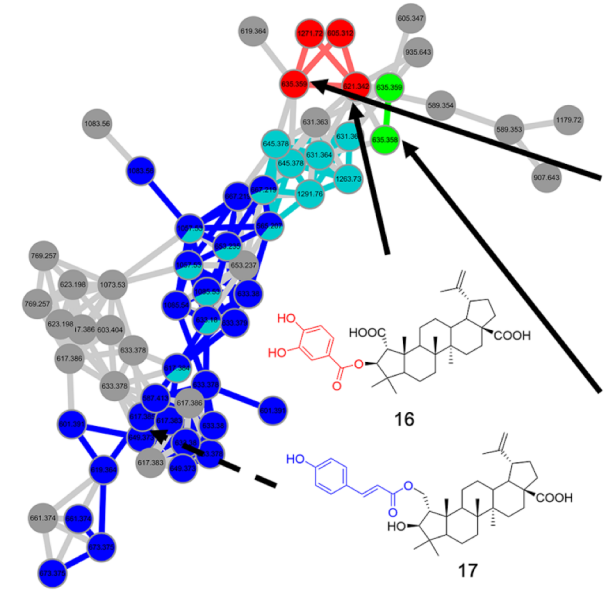
Annotating plant molecular families

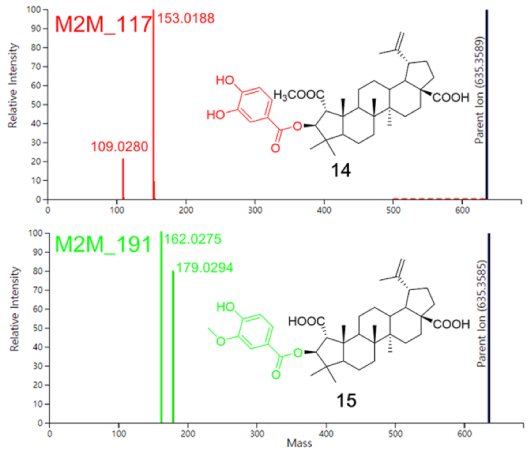


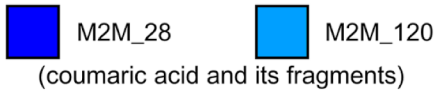
Flavonoid-3-*O*-glycoside Molecular Family:

Differentiation of subfamilies

Kaempferol and Quercetin based





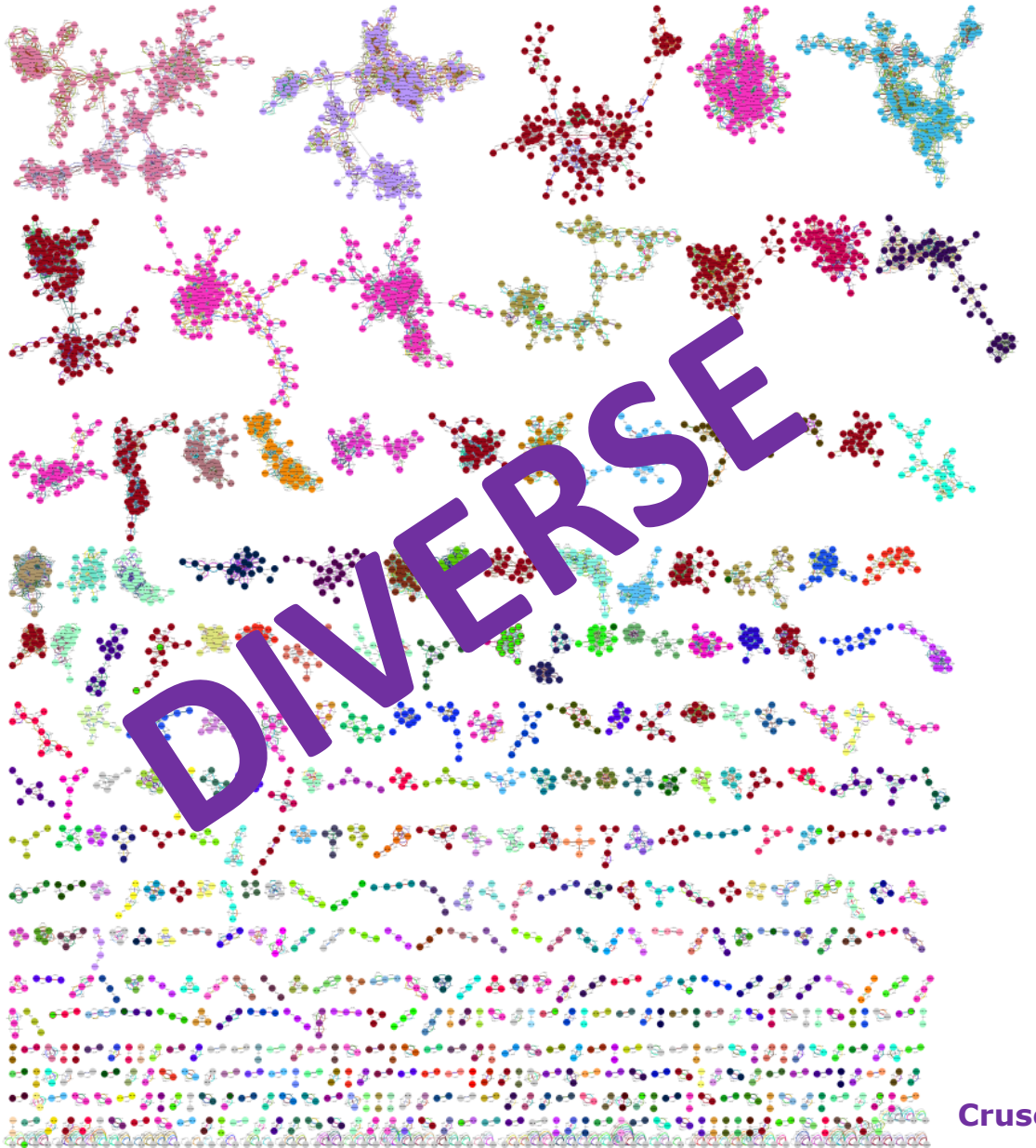


Triterpenoid Molecular Family:

Differentiation of modifications

Protocatechuic acid and Vanillic acid based

Illuminating bacterial chemistry



Molecular Network of
146 *Streptomyces* and *Salinispora sp*.

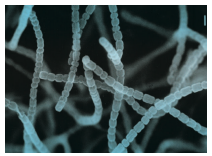
Triterpenoids

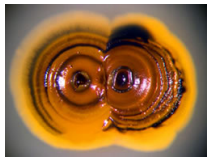
Lipids - PE

Cyclic peptides

Alpha amino acid esters

and many more….





Crusemann et al., J. Nat. Prod., 2017

Annotation of bacterial family I

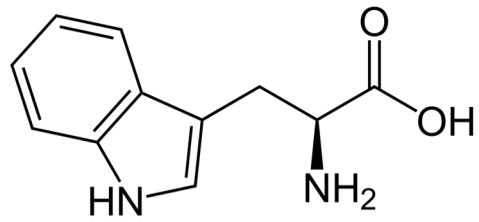
Mass2Motif Annotation

You can assign a label (annotate) this Mass2Motif from the **Annotation** field below. Additionally, a shorter annotation can also be assigned through the **Short Annotation** field. This will be used in the network visualisation.

Annotation: Tryptophan related Mass2Motif

Short Annotation: Tryptophan related Mass2Motif

Save

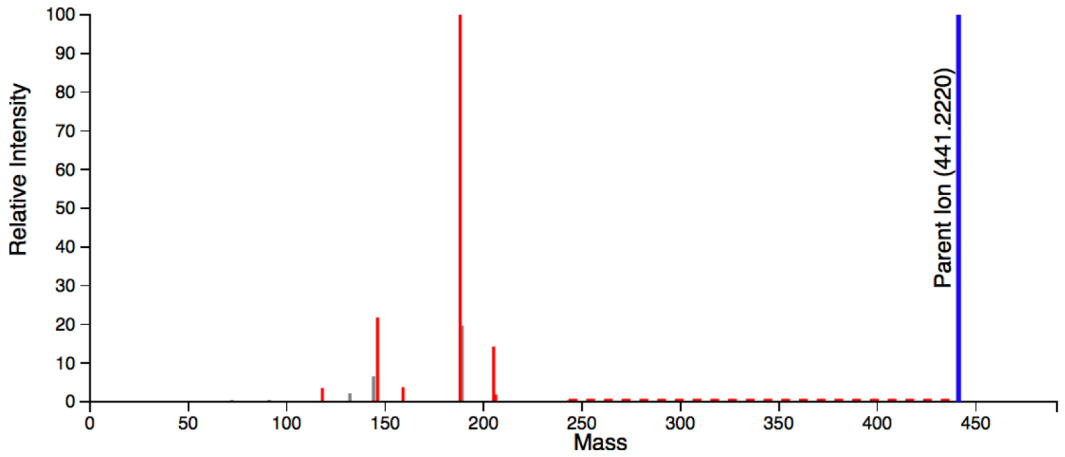


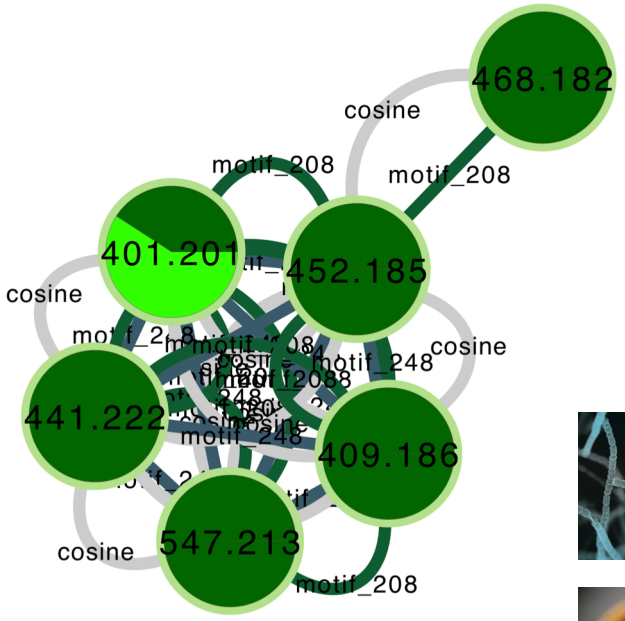
Motif 208

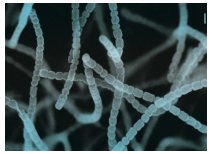
Previous

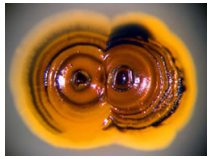
motif_208
Parent: document_1202, null (2/19)
Probability: Probability: 0.832888449166, overlap: 0.356384148972, Over

Next



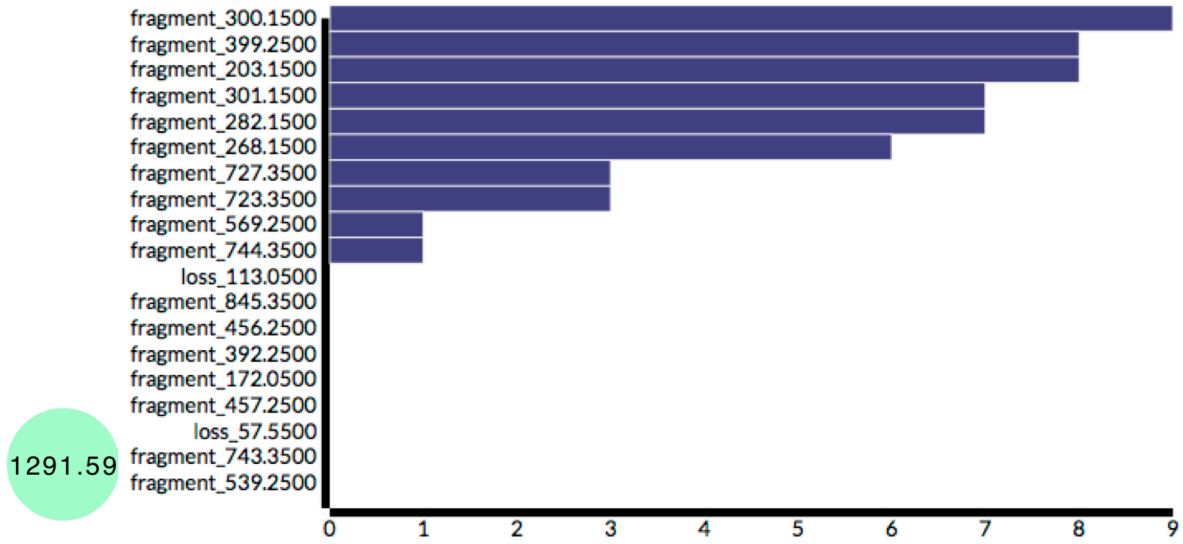


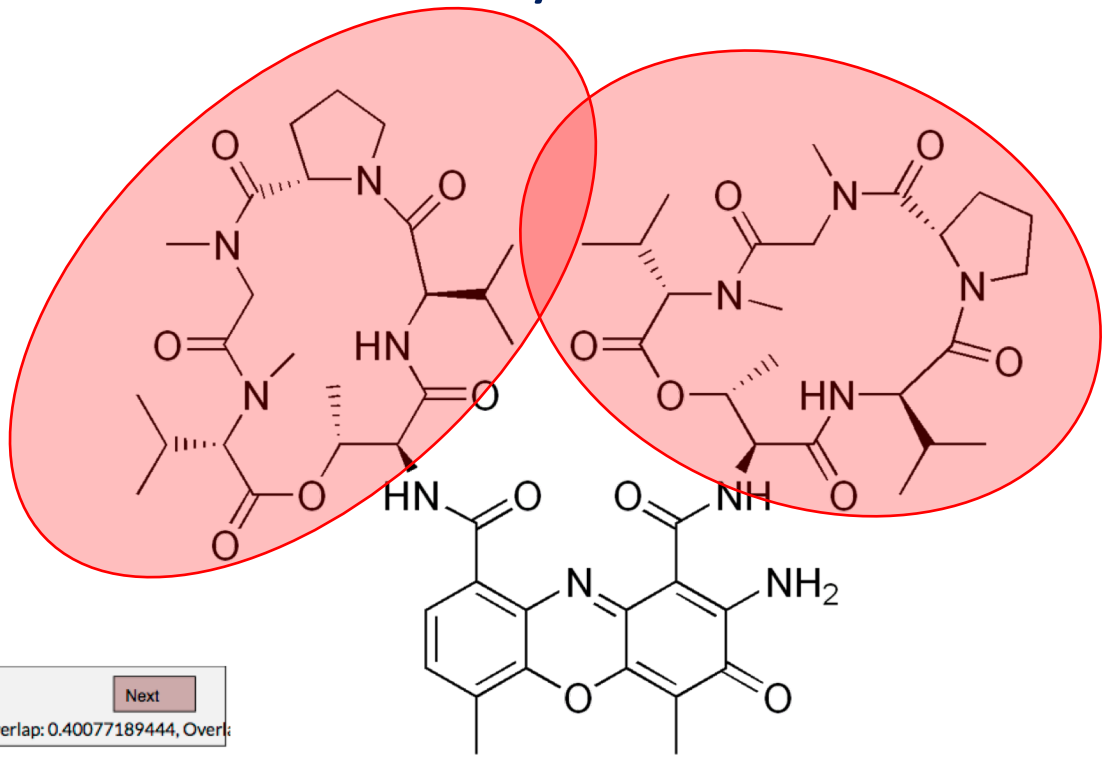




WAGENINGEN
UNIVERSITY & RESEARCH

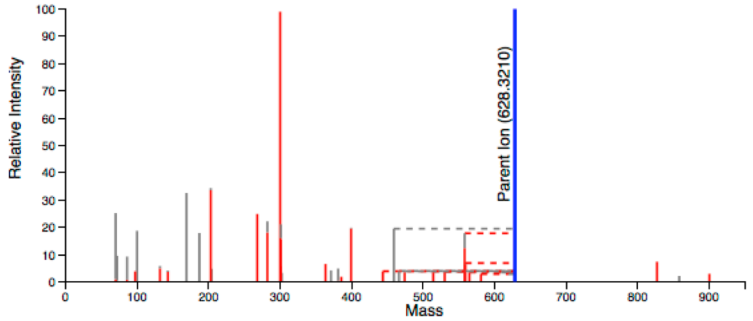
Annotation of bacterial molecular family II

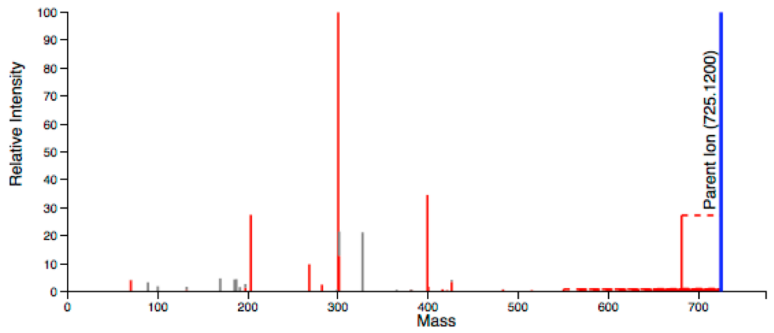




Previous motif_10 Next
Parent: document_3482, null (6/9)
Probability: Probability: 0.614024838127, overlap: 0.367052478451, Over

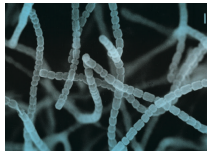
Previous motif_10 Next
Parent: document_4872, null (1/9)
Probability: Probability: 0.829939871031, overlap: 0.40077189444, Overl

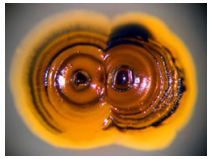




Actinomycin D

5	PeptideMass
.3E-7	1414.84
7E-10	1268.57
.0E-7	1290.57

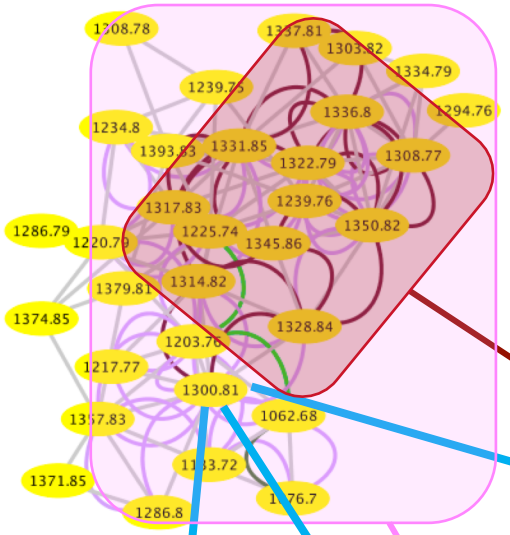




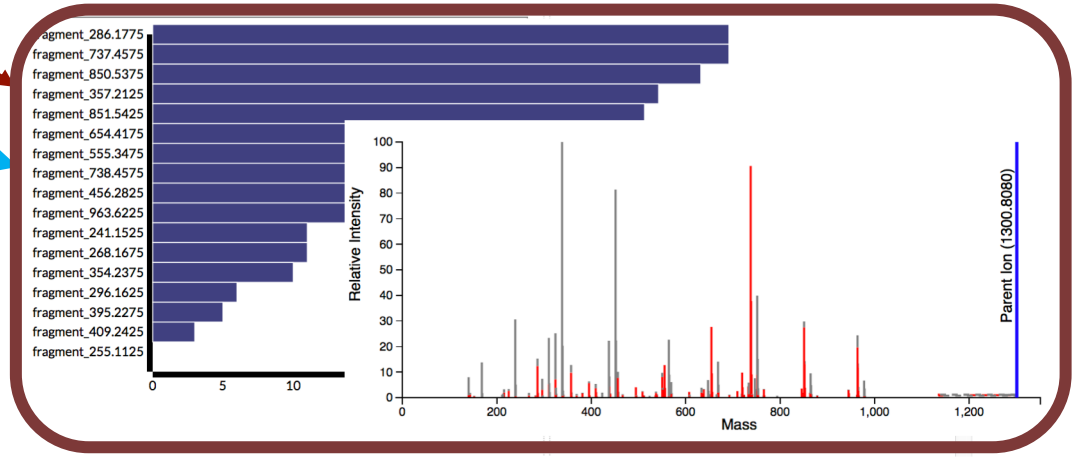
WAGENINGEN UNIVERSITY & RESEARCH

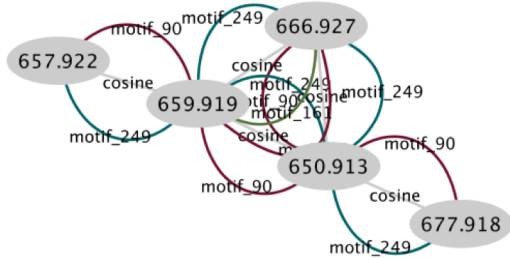
Mohimani et al., Nat. Chem. Biol., 2017

Crnovcic et al., RCS Advances, 2014

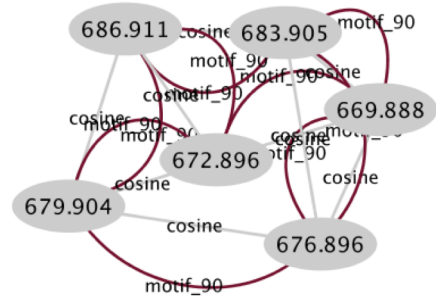


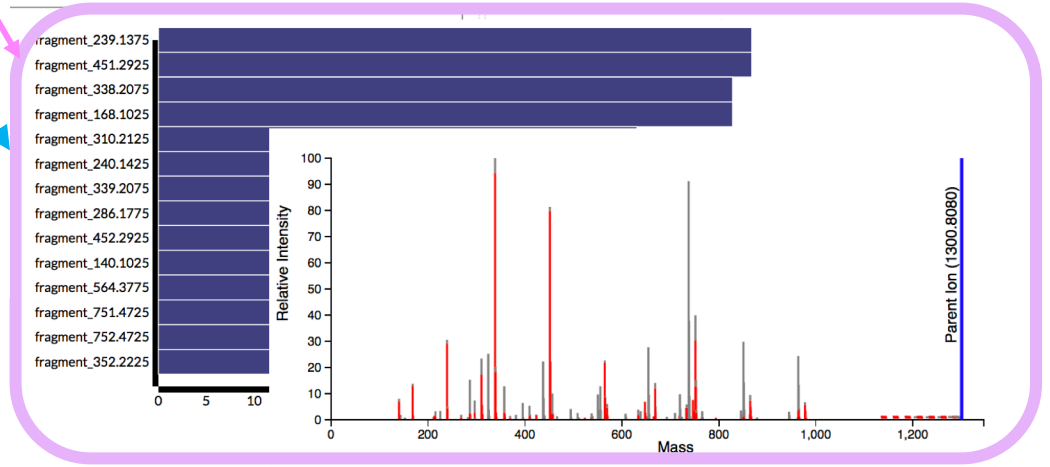
Annotation_Dereplicator_0_005bin	Annotation_Dereplicator_0_05bin	Annotation_VarQuest_0_005bin	Annotation_VarQuest_0_05bin
Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1
Xenoamicin_A_structure_0	Xenoamicin_A_structure_0	Xenoamicin_B_structure_1	Petriellin_A_Petriellin_A
		Xenoamicin_B_structure_1	Pentadactylin
		Xenoamicin_B_structure_1	Gramicidin_S2
		Xenoamicin_B_structure_1	Petriellin_A_Petriellin_A
Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1
Xenoamicin_A_structure_0	Xenoamicin_A_structure_0	Xenoamicin_A_structure_0	Xenoamicin_B_structure_1
	Cyclosporin,_9Cl_2-[O-(2-Hydroxyethyl...	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1
		Xenoamicin_B_structure_1	Xenoamicin_B_structure_1
		Xenoamicin_B_structure_1	Xenoamicin_B_structure_1
Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1	Xenoamicin_B_structure_1



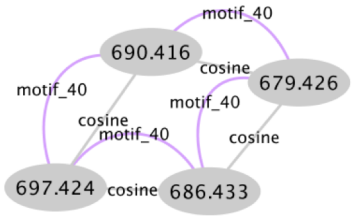


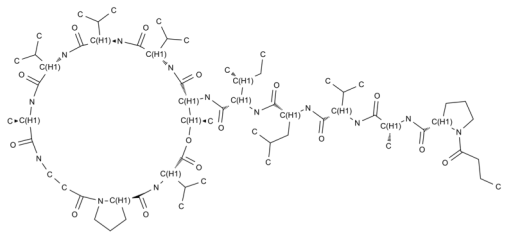
Ring



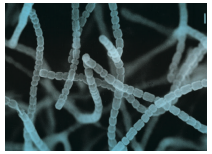


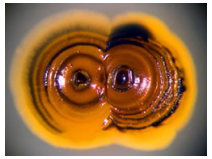
Tail





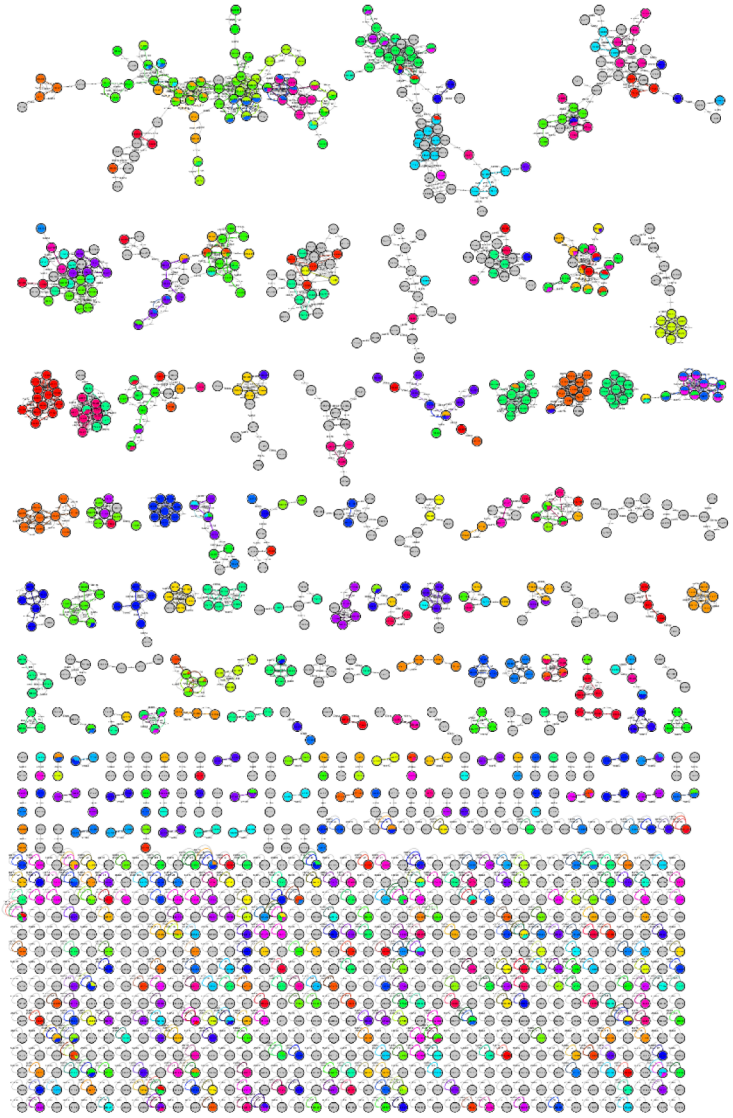
Xenoamicin A

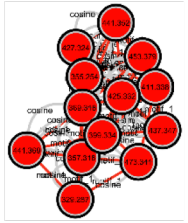


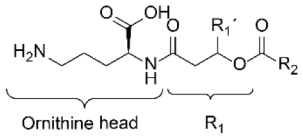


WAGENINGEN
UNIVERSITY & RESEARCH

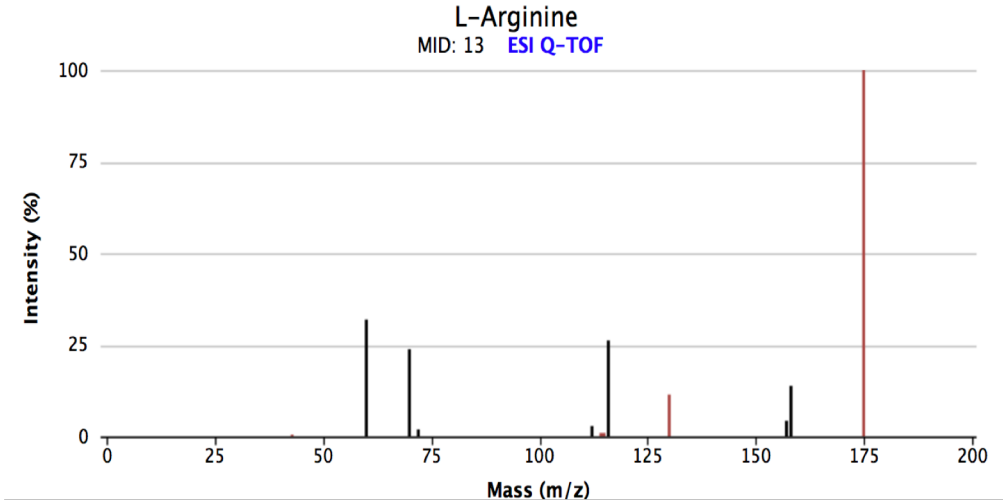
Annotation of fungal garden





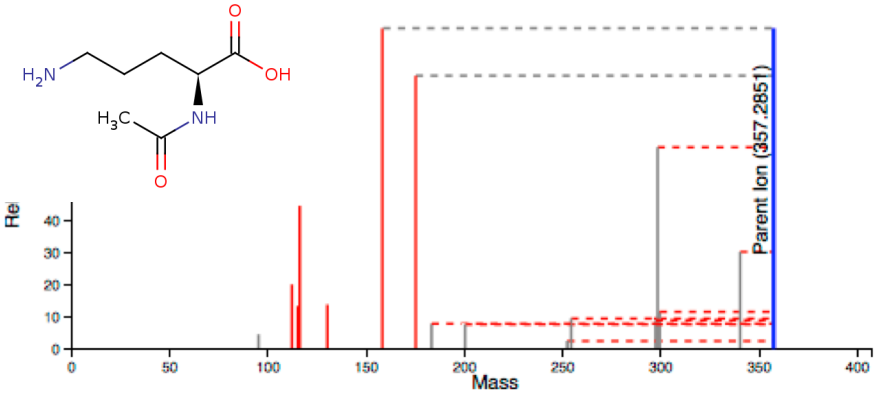


Unknown

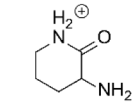


COLOR and SHAPE CODE

- **Motif 1 - N2-acetylornithine related**

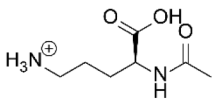


Fragment 115.0875

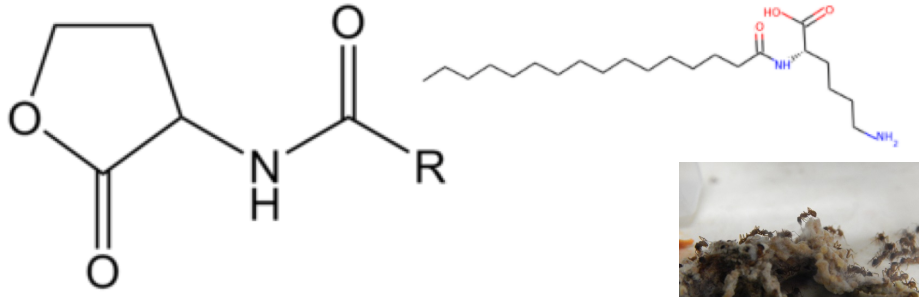


Exact Mass: 115.09

Fragment 175.1175



Exact Mass: 175.11

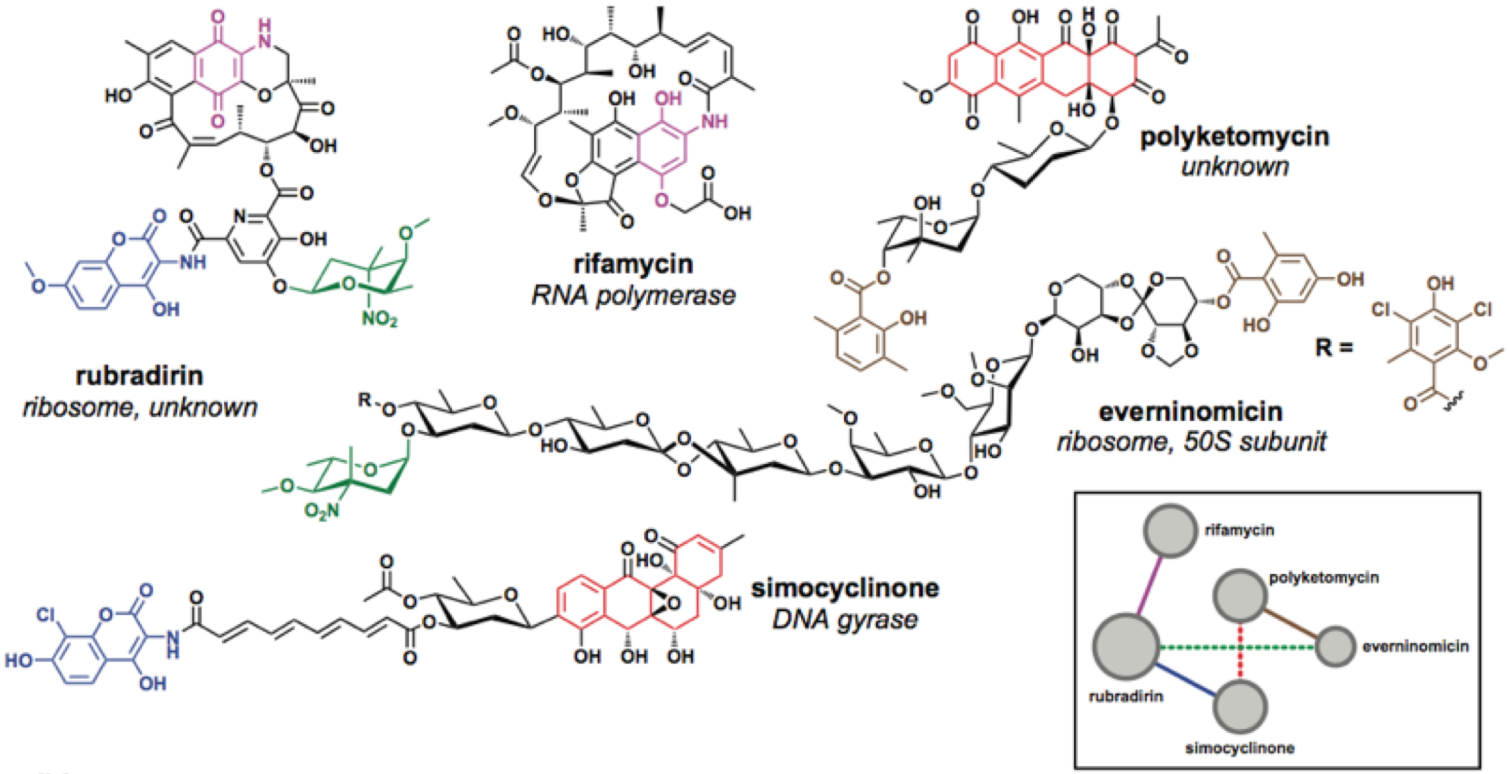


N-acyl homoserine lactone

WAGENINGEN UNIVERSITY & RESEARCH

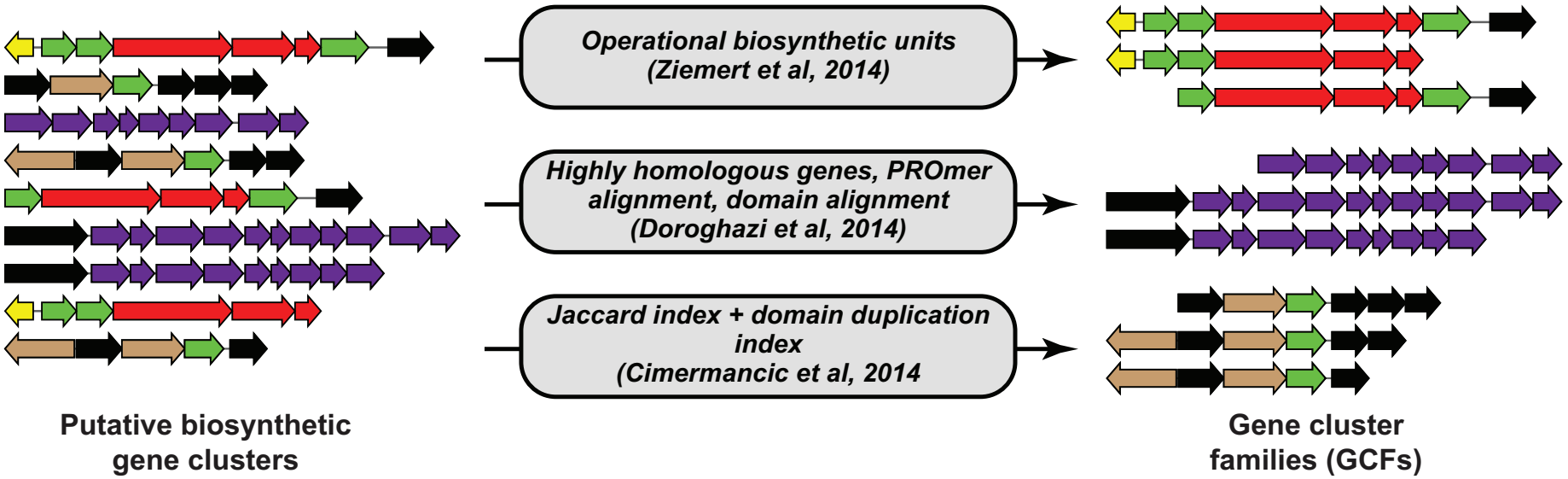
https://metlin.scripps.edu

Biosynthetic gene clusters: the key to mine genomes for novel natural product chemistry



Medema, Cimermancic et al., *PLoS Computational Biology* 10: e1004016.

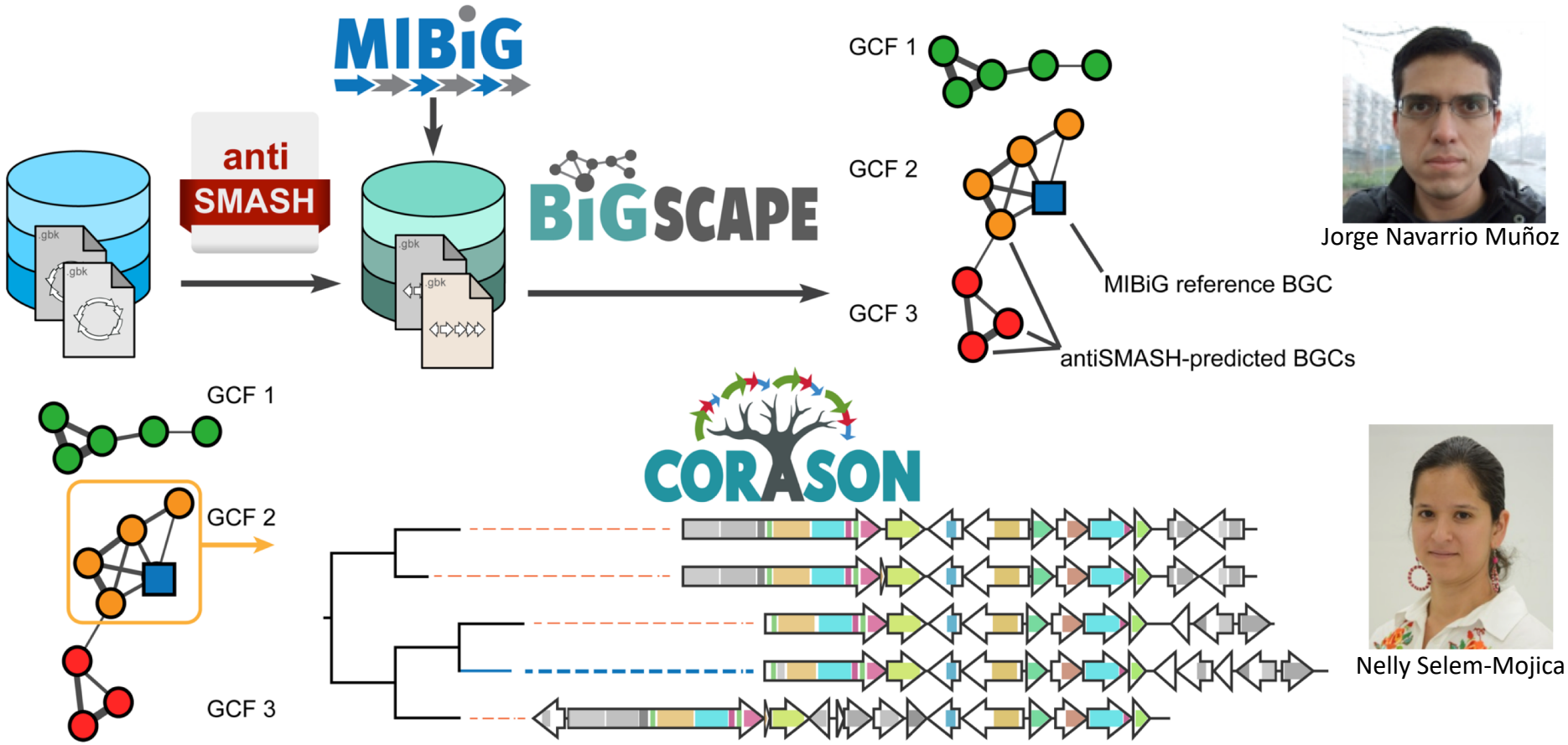
Grouping Biosynthetic Gene Clusters into Families



WAGENINGEN
UNIVERSITY & RESEARCH

Medema & Fischbach (2015) *Nat. Chem. Biol.* 11: 639-648.

BiG-SCAPE & CORASON automate reconstruction and phylogenomic analysis of BGC families

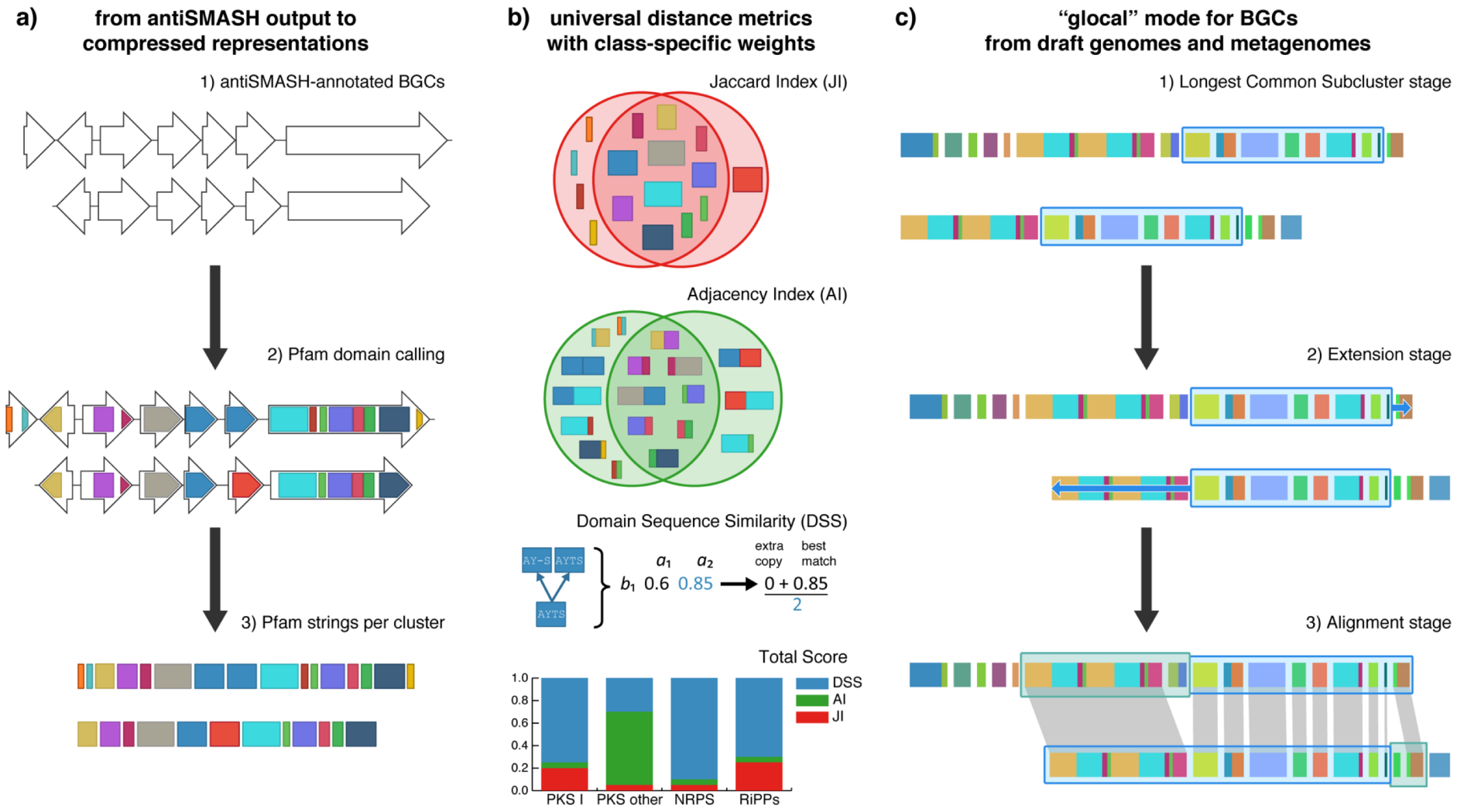


Jorge Navarrio Muñoz

Nelly Selem-Mojica

WAGENINGEN
UNIVERSITY & RESEARCH

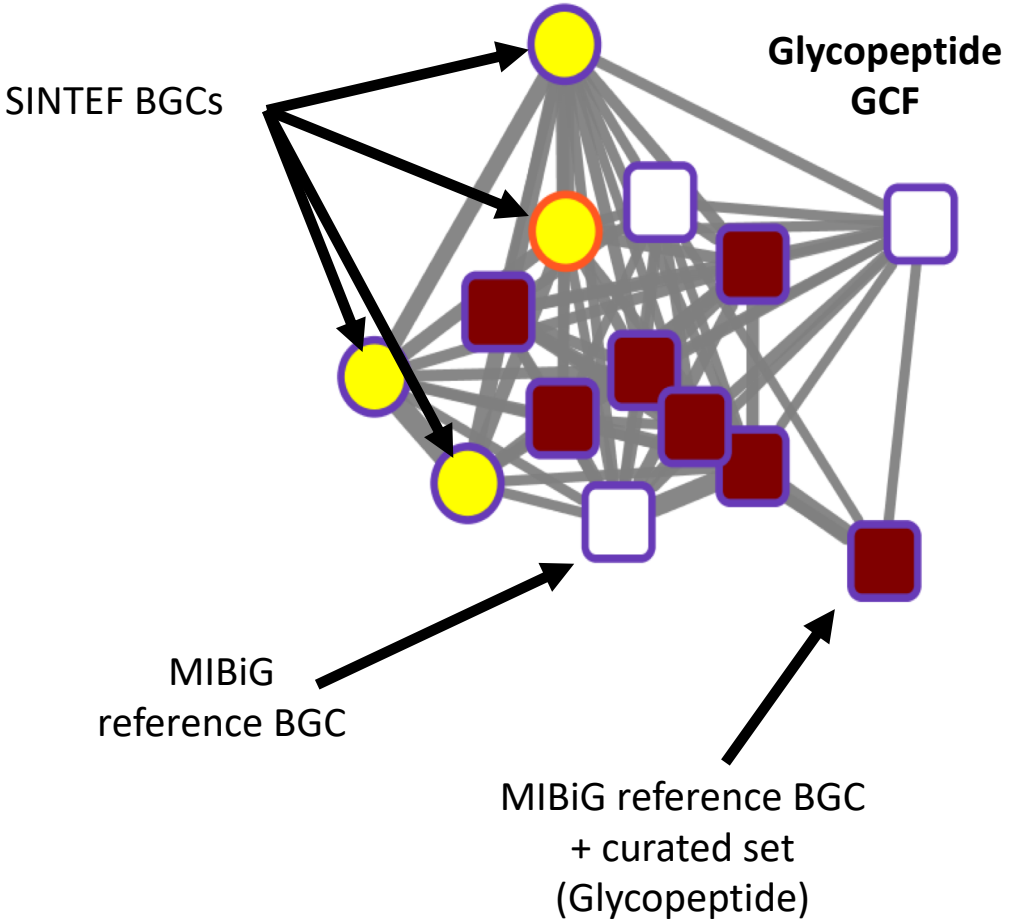
Efficient Data compression, Distance metrics and alignment modes

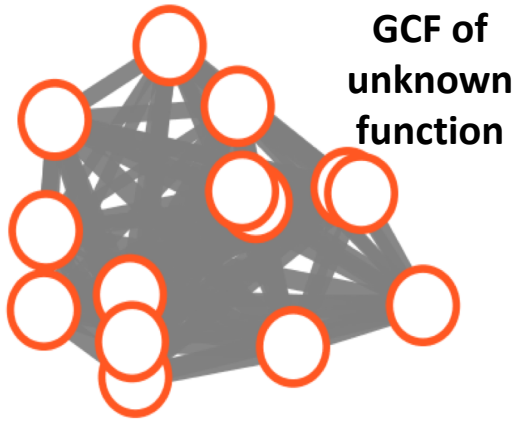


WAGENINGEN UNIVERSITY & RESEARCH

Jorge Navarrio Muñoz, Marley Yeong, Emzo de los Santos

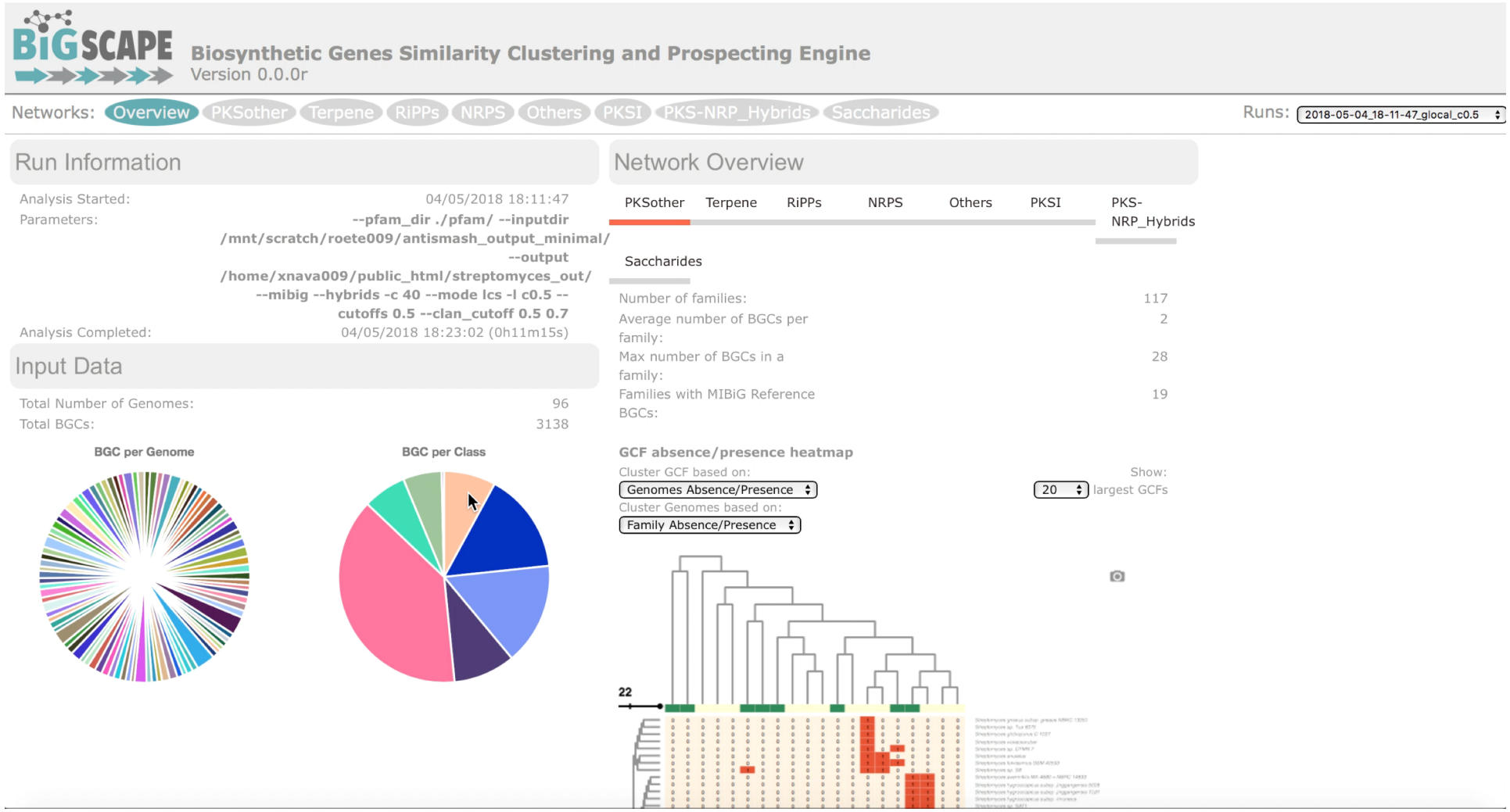
MIBig reference data allows rapid annotation propagation





https://mibig.secondarymetabolites.org

A rich web interface allows effectively exploring biG-scape outputs



Satria Kautsar

https://git.wageningenur.nl/medema-group/BiG-SCAPE

Big-scape has quickly been adopted by the scientific community

nature microbiology ARTICLES
https://doi.org/10.1038/s41564-018-0200-0

Bipartite interactions, antibiotic production and biosynthetic potential of the *Arabidopsis* leaf microbiome

Eric J. N. Helfrich[1,2], Christine M. Vogel[1,2], Reiko Ueoka[1], Martin Schäfer[1], Florian Ryffel[1], Daniel B. Müller[1], Silke Probst[1], Markus Kreuzer[1], Jörn Piel[1]* and Julia A. Vorholt[1]*

Plants are colonized by phylogenetically diverse microorganisms that affect plant growth and health. Representative genome-sequenced culture collections of bacterial isolates from model plants, including *Arabidopsis thaliana*, have recently been established. These resources provide opportunities for systematic interaction screens combined with genome mining to discover uncharacterized natural products. Here, we report on the biosynthetic potential of 224 strains isolated from the *A. thaliana* phyllosphere. Genome mining identified more than 1,000 predicted natural product biosynthetic gene clusters (BGCs), hundreds of which are unknown compared to the MIBiG database of characterized BGCs. For functional validation, we used a high-throughput screening approach to monitor over 50,000 binary strain combinations. We observed 725 inhibitory interactions, with 26 strains contributing to the majority of these. A combination of imaging mass spectrometry and bioactivity-guided fractionation of the most potent inhibitor, the BGC-rich *Brevibacillus* sp. Leaf182, revealed three distinct natural product scaffolds that contribute to the observed antibiotic activity. Moreover, a genome mining-based strategy led to the isolation of a *trans*-acyltransferase polyketide synthase-derived antibiotic, macrobrevin, which displays an unprecedented natural product structure. Our findings demonstrate that the phyllosphere is a valuable environment for the identification of antibiotics and natural products with unusual scaffolds.

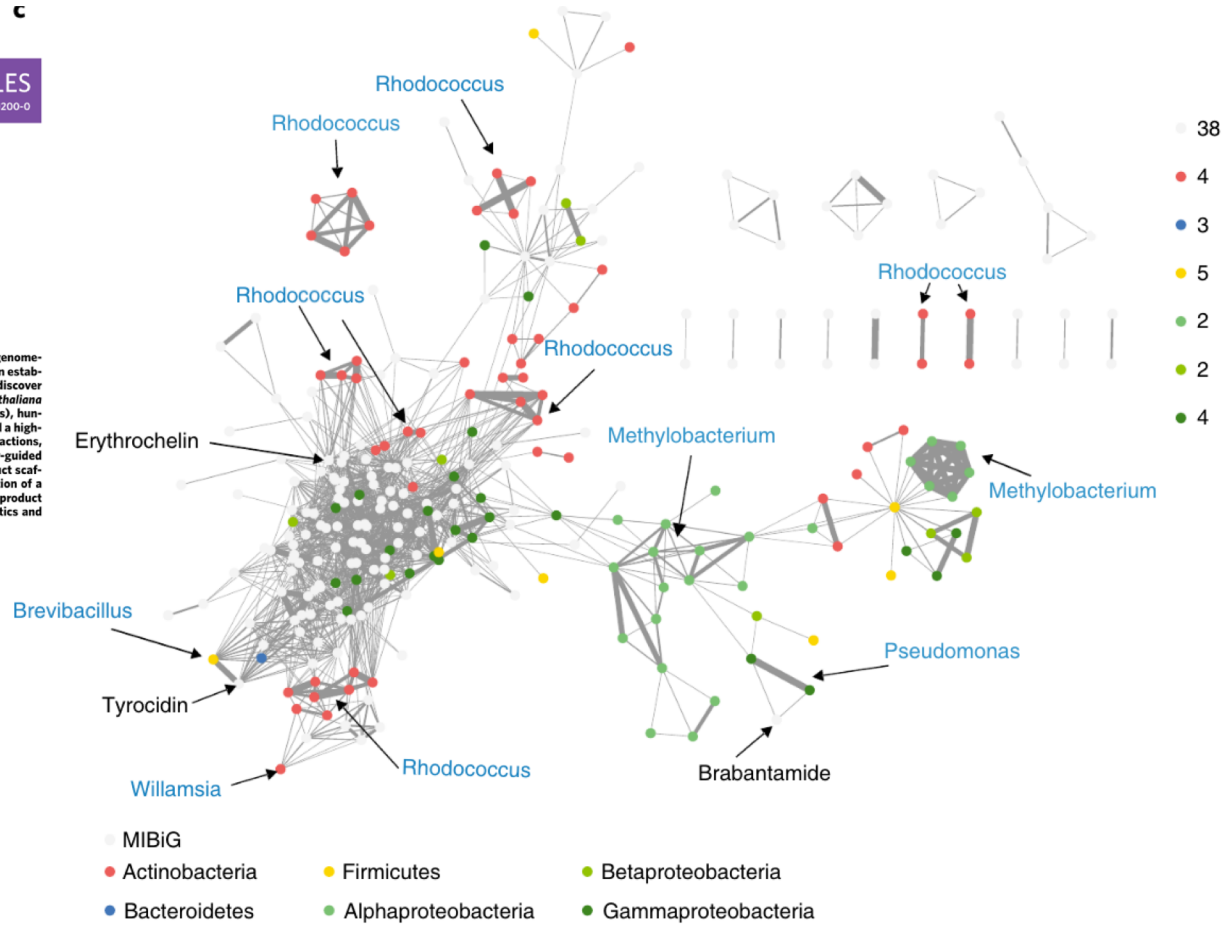
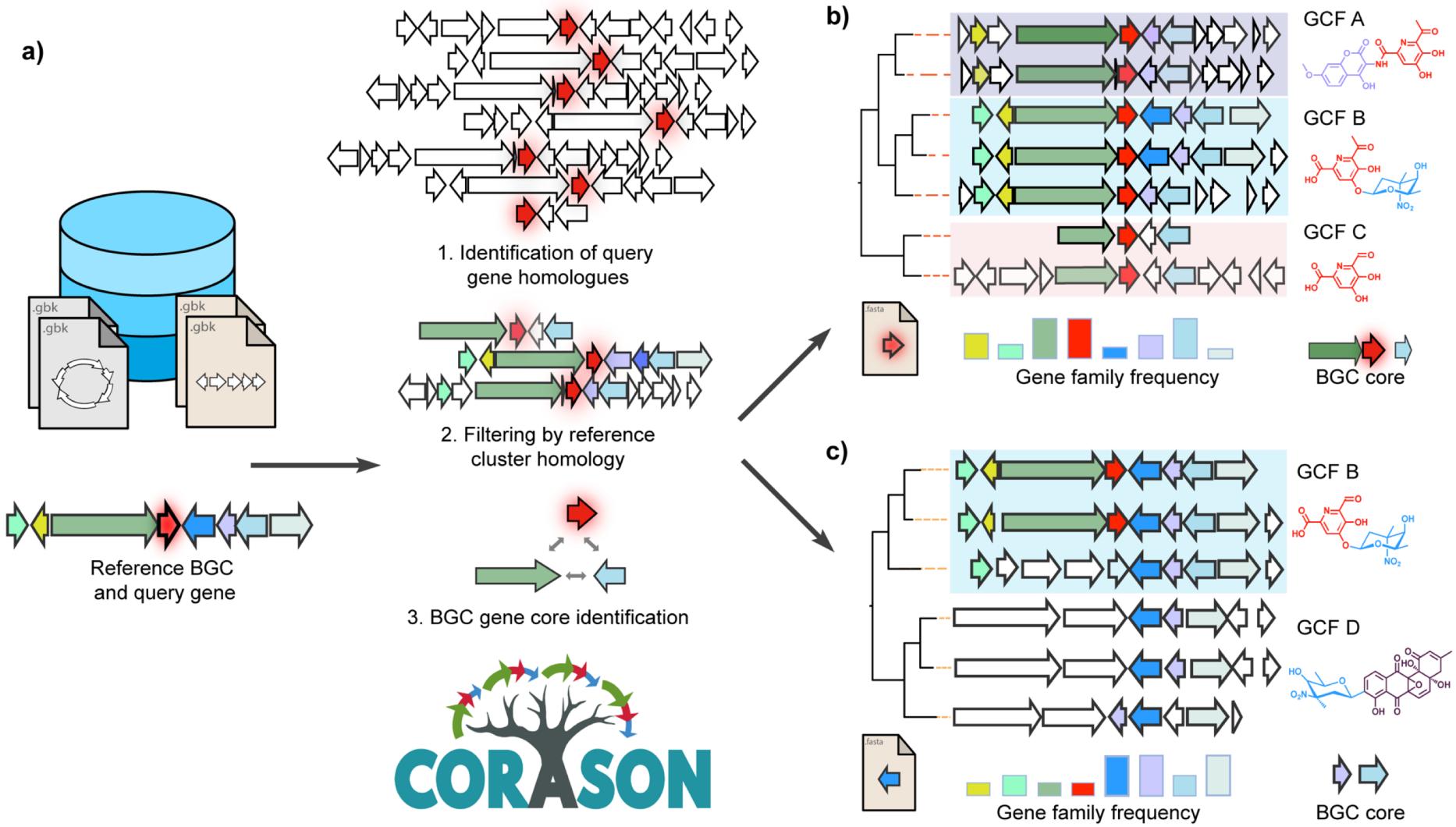


Fig. 3 | BiG-SCAPE analysis of BGCs detected by antiSMASH in 207 genomes of the *At*-LSPHERE strain collection and comparison with the MIBiG database of characterized BGCs. a, RiPP BGCs ($n=246$). **b**, Type I PKS BGCs ($n=277$). **c**, NRPS BGCs ($n=338$). Nodes and singletons are colour coded according to their origin (MIBiG database or phylum/class) and represent individual BGCs. The widths of interconnecting edges indicate the degree of relatedness between two BGCs, with connections up to a raw distance of 0.75 retained (Supplementary Table 12). Numbers of singletons are indicated. Black labels denote compounds associated with selected MIBiG BGCs. Blue labels highlight clusters of related BGCs or individual BGCs of interest by phylogenetic distribution.

CORASON facilitates finding gene clusters around homologues of gene of interest



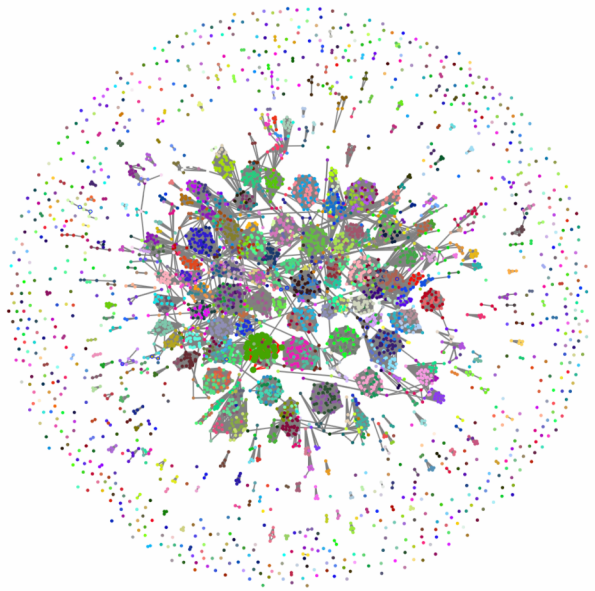
https://github.com/nselem/corason/wiki

WAGENINGEN
UNIVERSITY & RESEARCH
Nelly Selem

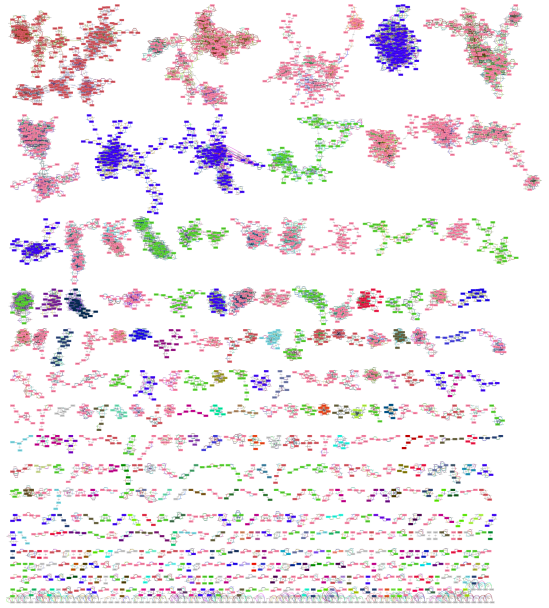
Linking substructures to genetic elements

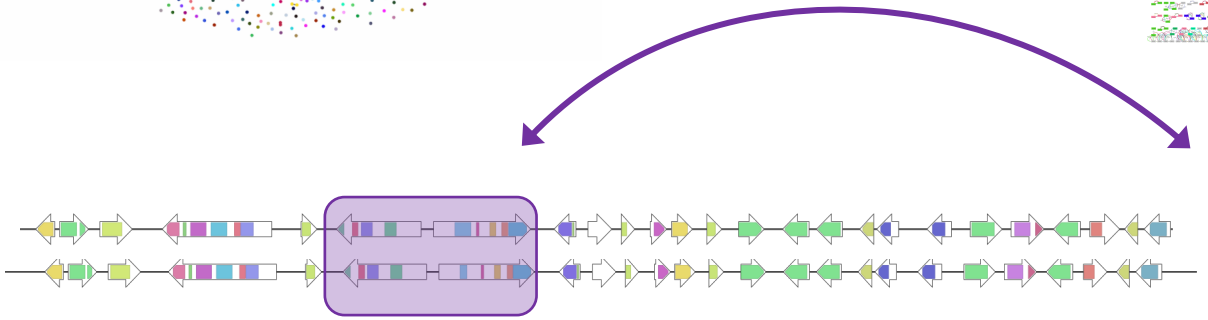
iOMEGA: Integrated Omics for MEtabolomics and Genomics Annotation

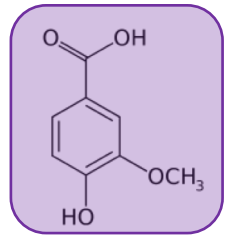
Gene Cluster Families & Metabolite Families

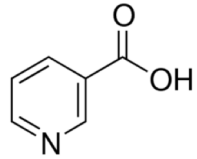


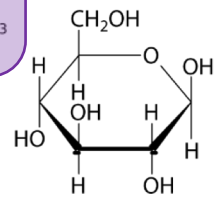
<--->





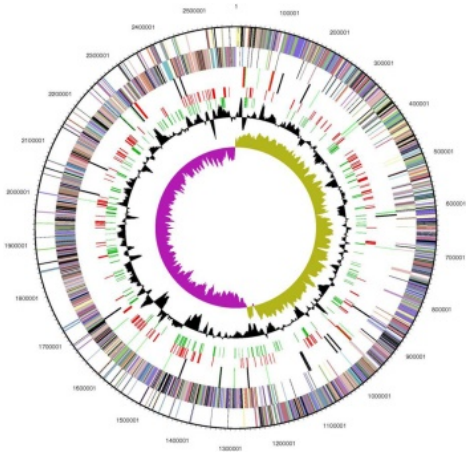


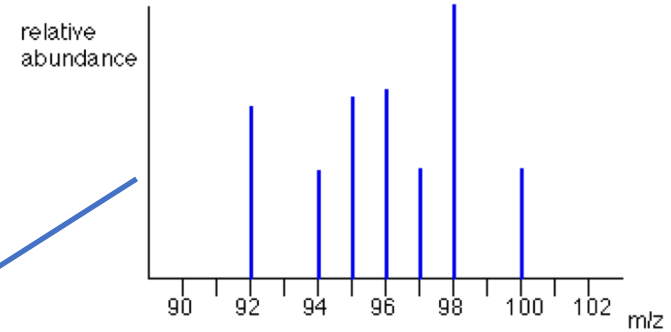




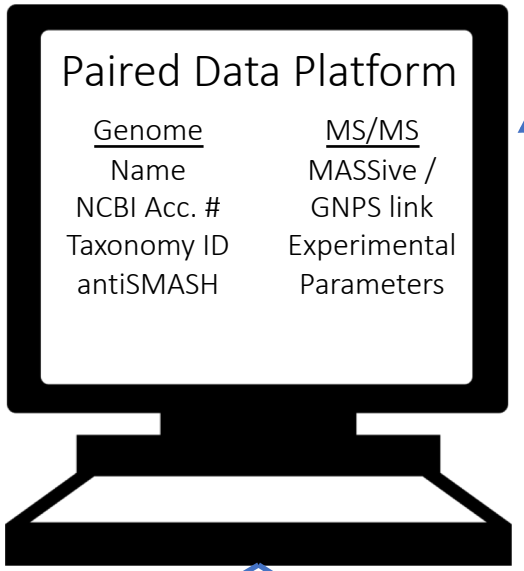
WAGENINGEN
UNIVERSITY & RESEARCH

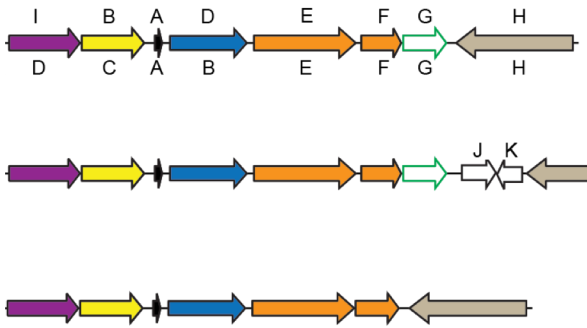
Paired Data Platform





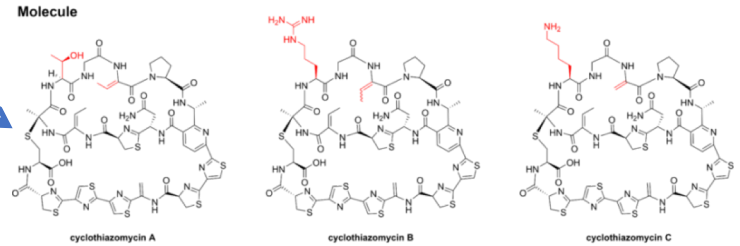
Michelle Schorn





Gene clusters
Gene cluster families

LINK!



Molecules
Molecular families

WAGENINGEN
UNIVERSITY & RESEARCH

Paired Data Platform

iOMEGA paired data platform schema JuNo

This is the JSON schema for paired genomic / metabolomic data.

version *

1

Personal data*

Name of contact for correspondence

This person will be the point of contact for any communication related to this entry.

Academic institution or company name

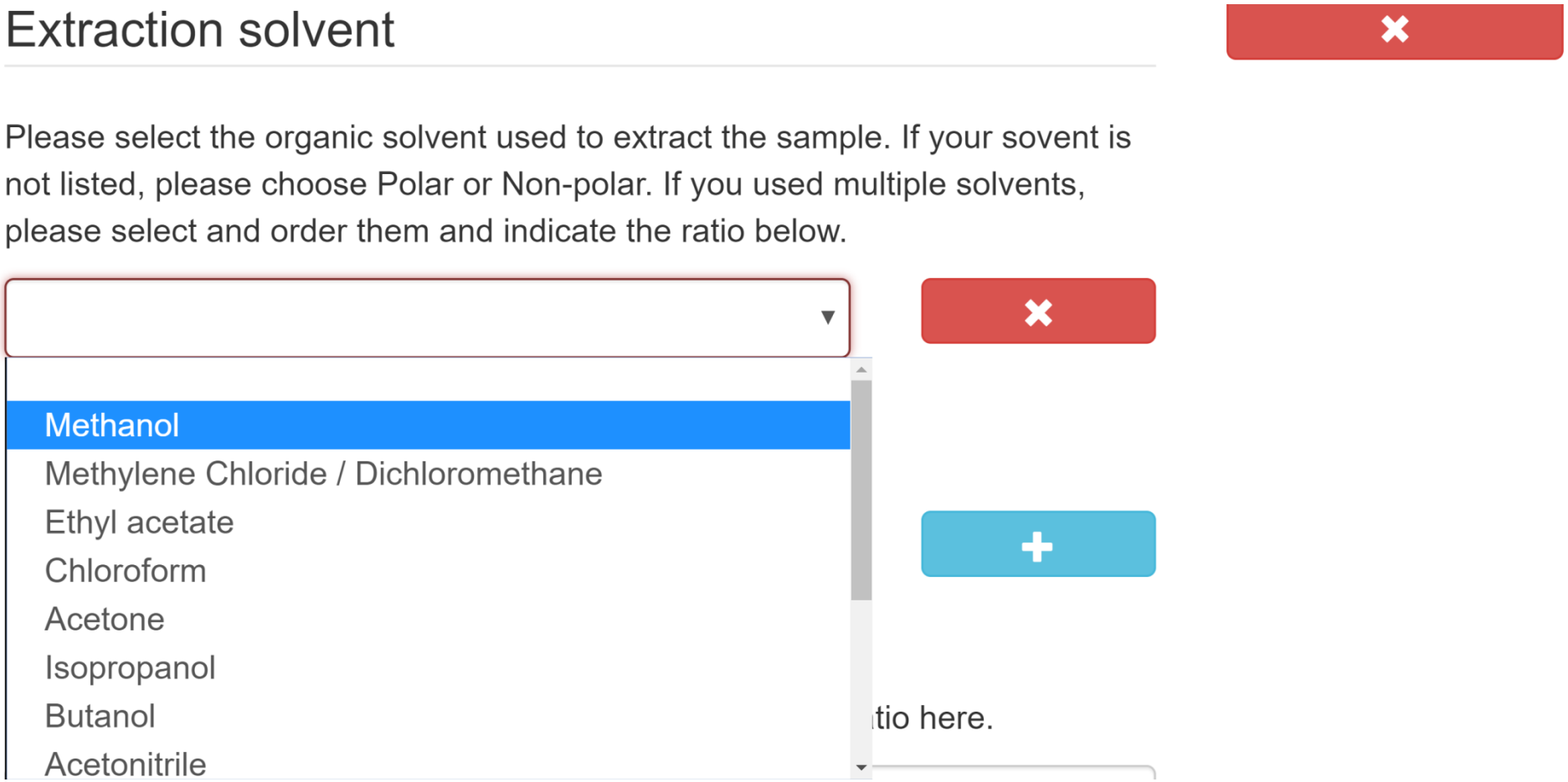
Please use the full, official name of your institute in English. E.g., 'Harvard University'.

Submitter contact e-mail address

Name of the principal investigator of the submitter

This person is contacted in case the submitter has moved institution

Paired Data Platform



WAGENINGEN
UNIVERSITY & RESEARCH

Thank you!

Special Issue: Metabolomics Data Processing and Data Analysis—Current Best Practices

Deadline for manuscript submissions: **28 February 2019**

Guest Editors

Dr. Kati Hanhineva
University of Eastern
Finland
Twitter: @KatiHanhineva

Dr. Justin van der Hooft
Wageningen University
Netherlands
Twitter: @jjjvanderhooft

Keywords
Metabolomics data processing data interpretation annotation and visualization data analysis

WAGENINGEN UNIVERSITY & RESEARCH

MDPI OPEN ACCESS

metabolites
an Open Access Journal by MDPI