



Secondary metabolites: paths to discovery

Sandra Godinho Silva

sandragodinhosilva@tecnico.ulisboa.pt

28 April 2021

Natural Products Discovery

Why is this still a relevant field?

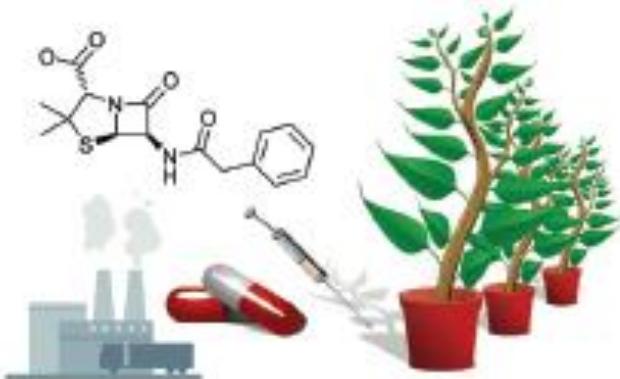
Rise of **multidrug-resistant** pathogens

+

Acute and long-term side effects of widely used drugs

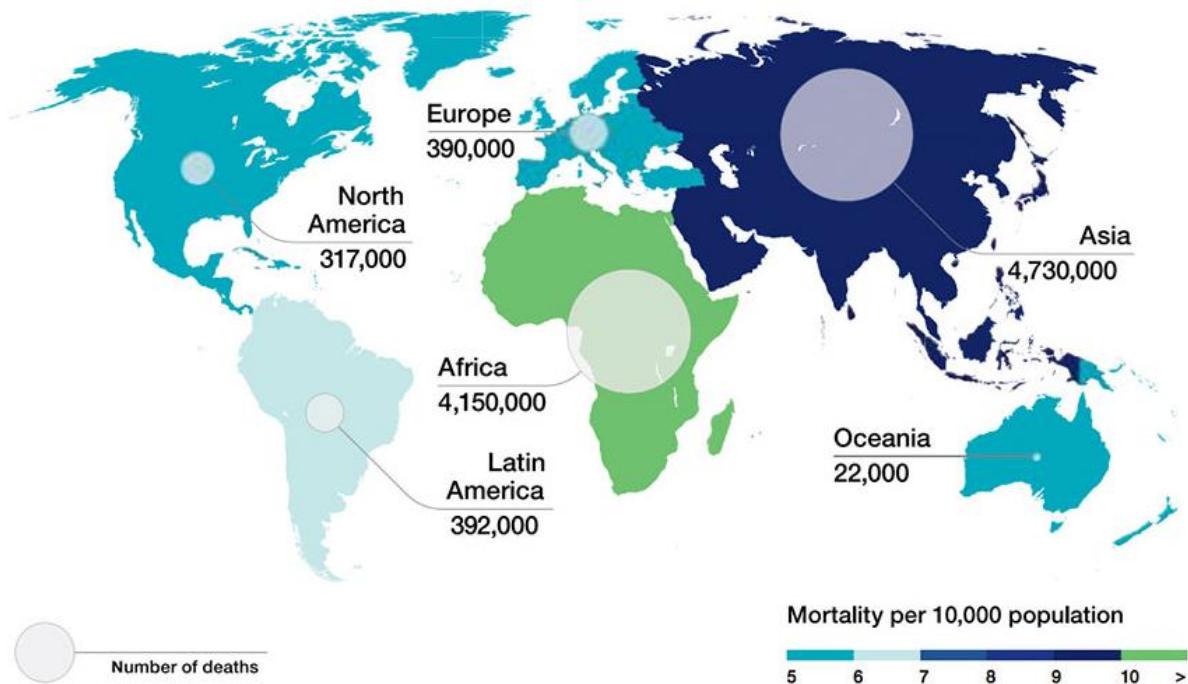
=

Urgent need for **new therapeutic agents**



Natural Products Discovery

Deaths attributable to AMR every year by 2050



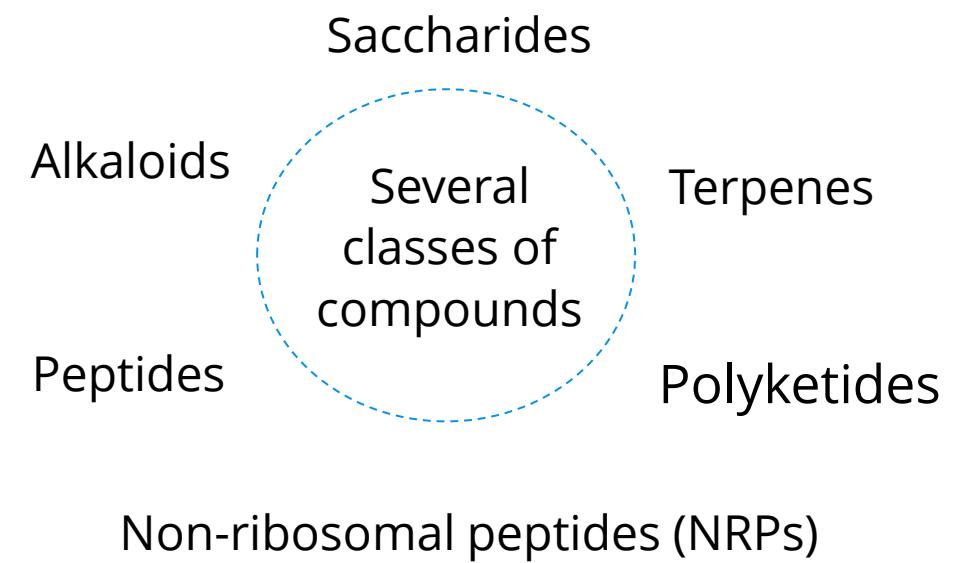
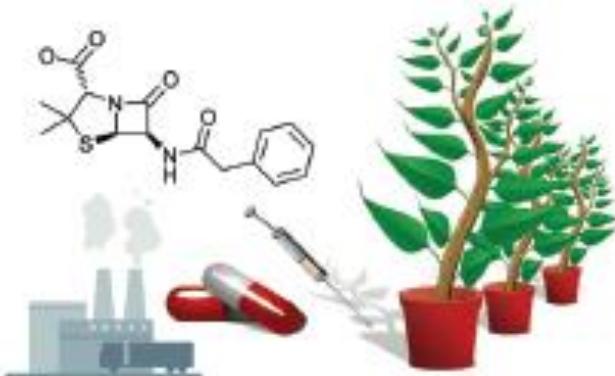
- **Antibiotic resistance** accounts for at least 50,000 deaths each year in Europe and the US.
- It is predicted that drug resistant infections will be responsible for the deaths of 10 million people worldwide by 2050.
- **Cancer** is a leading cause of death worldwide with 7.6 million deaths each year with numbers continuously rising.

Source: Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations (2014)

The need for new therapeutical drugs is real

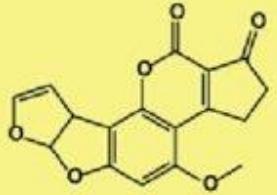
Natural Products Discovery

- Small organic molecules produced by living organisms;
- Normally are secondary metabolites:
 - Not essential for growth and reproduction;
 - Provide **survival advantage**.



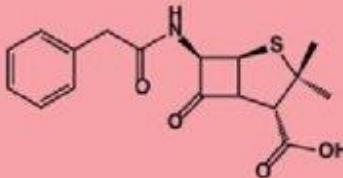
Secondary metabolites – classes:

polyketide

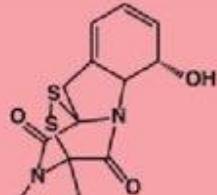


aflatoxin B1

non-ribosomal peptides

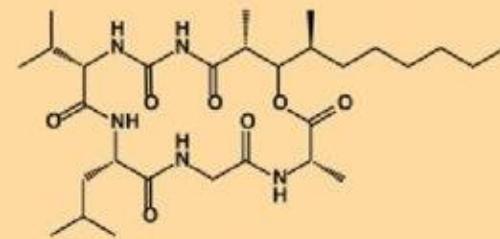


penicillin G



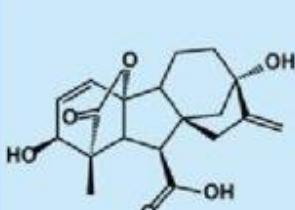
gliotoxin

polyketide/non-ribosomal peptide hybrids



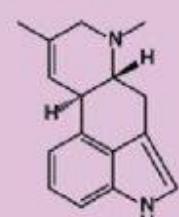
emericellamide A

terpene



gibberellin A3

prenylated tryptophan derivative



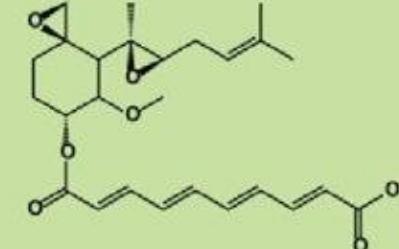
agroclavine

non-canonical

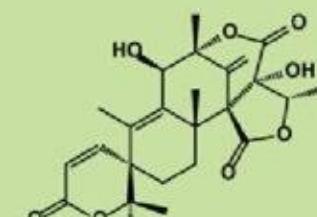


norloline

meroterpenoids



fumagillin



austinol

Secondary metabolites encoding genes are organized in:

Biosynthetic Gene Clusters (BGCs)

Physically clustered group of two or more genes in a particular genome that together encode a biosynthetic pathway to produce a specialized metabolite.



A BGC represents a biosynthetic and evolutionary unit.

Encodes for:

- Biosynthetic enzymes;
- Resistance enzymes;
- Enzymes to produce unusual building blocks;
- Regulatory machinery.

Horizontal gene transfer (HGT)

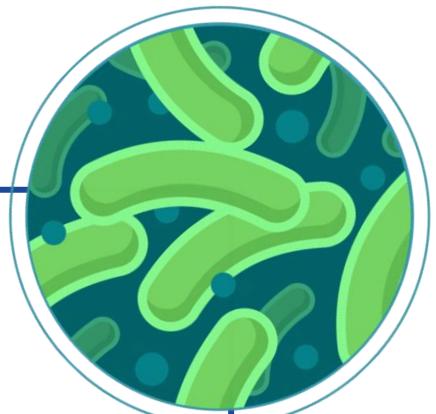
BGCs are prone to **horizontal gene transfer** (HGT)

- Evidenced by:
- Their clustering;
 - Frequent linkage with mobile genetic elements;
 - Detection on plasmids.

Mutation
Recombination
Gene gain
Gene loss
Gene duplication
Successive merge of smaller subclusters

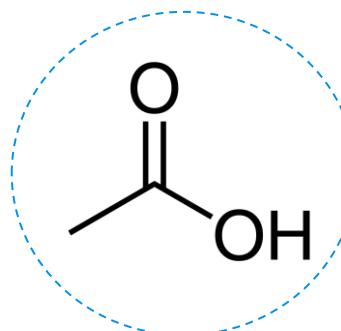
BGC diversification mechanisms

- Guided by:
- selective pressures;
 - opportunities for genetic exchange.



Polyketides

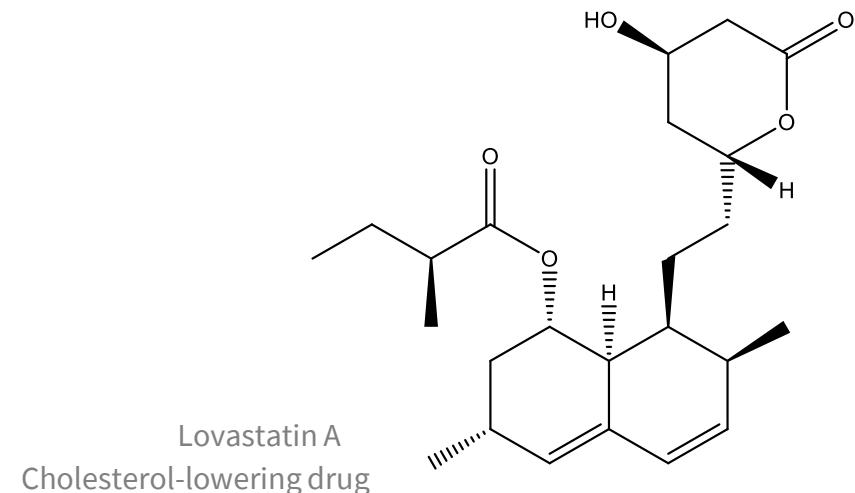
- One of the largest classes of natural products.
- Synthesized by large multifunctional enzymes: **Polyketide Synthases (PKS)**.
- Extremely high structural diversity.
- Important applications in medicine and pharmaceutical industry.



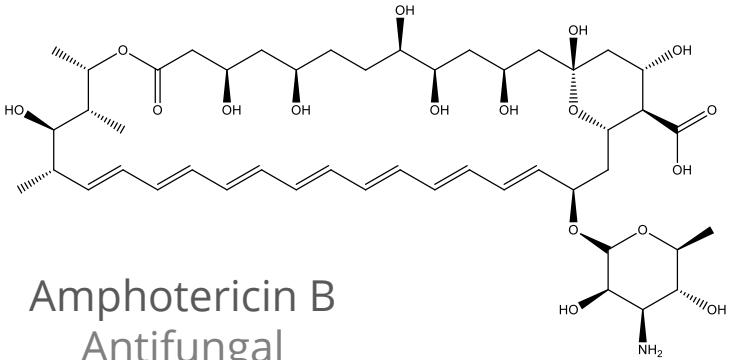
The building blocks used are derived from one of the simplest molecules available in nature:
acetic acid



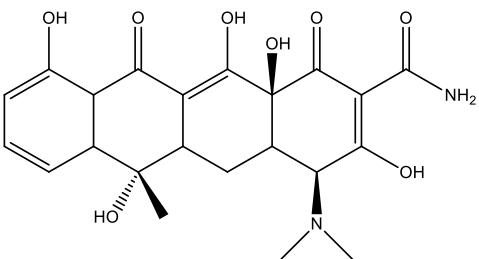
Most common: { Malonyl-CoA
Methylmalonyl-CoA



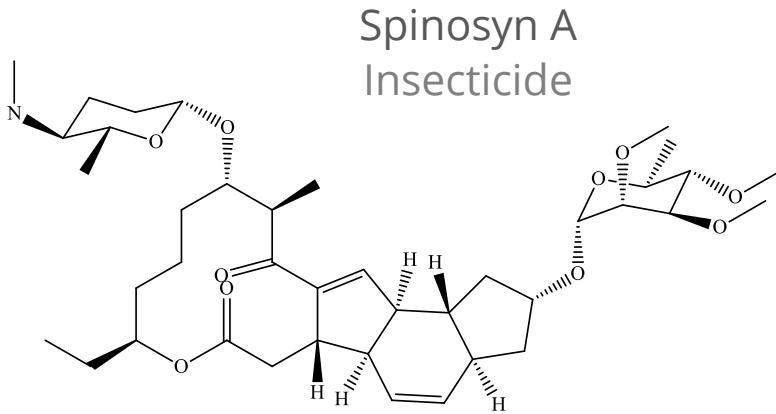
Famous Polyketides



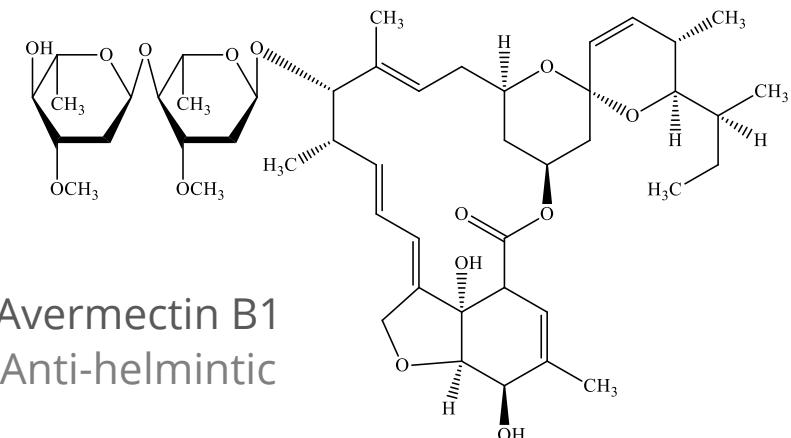
Amphotericin B
Antifungal



Tetracycline
Antibiotic



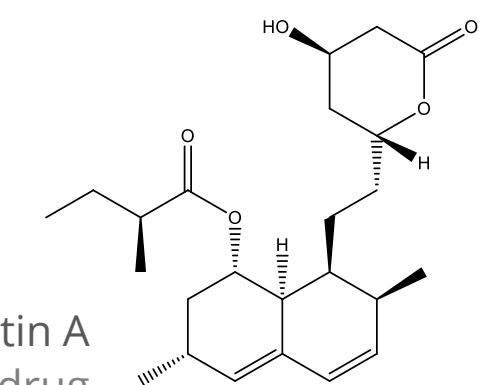
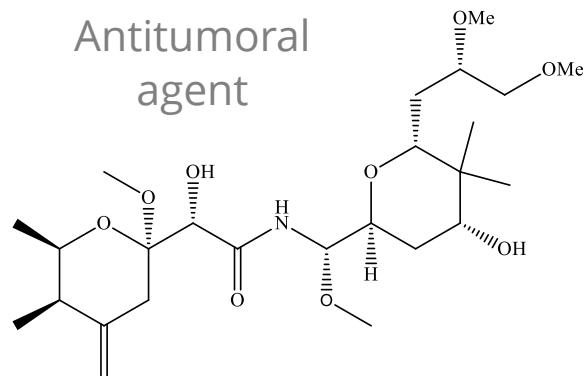
Spinosyn A
Insecticide



Avermectin B1
Anti-helmintic

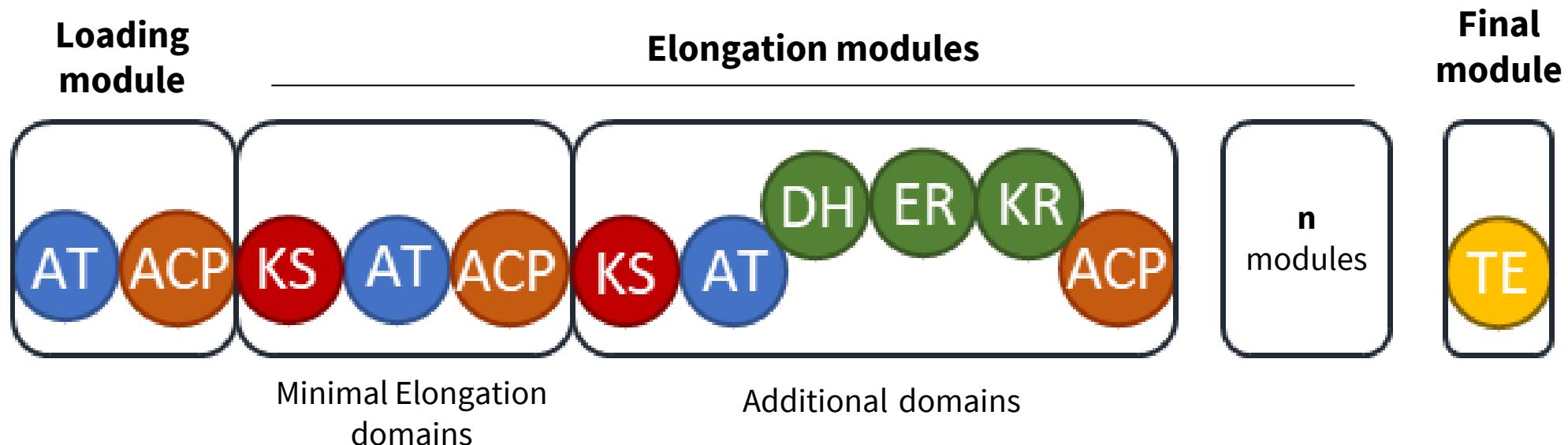
Important therapeutic drugs

Pederin
Antitumoral
agent



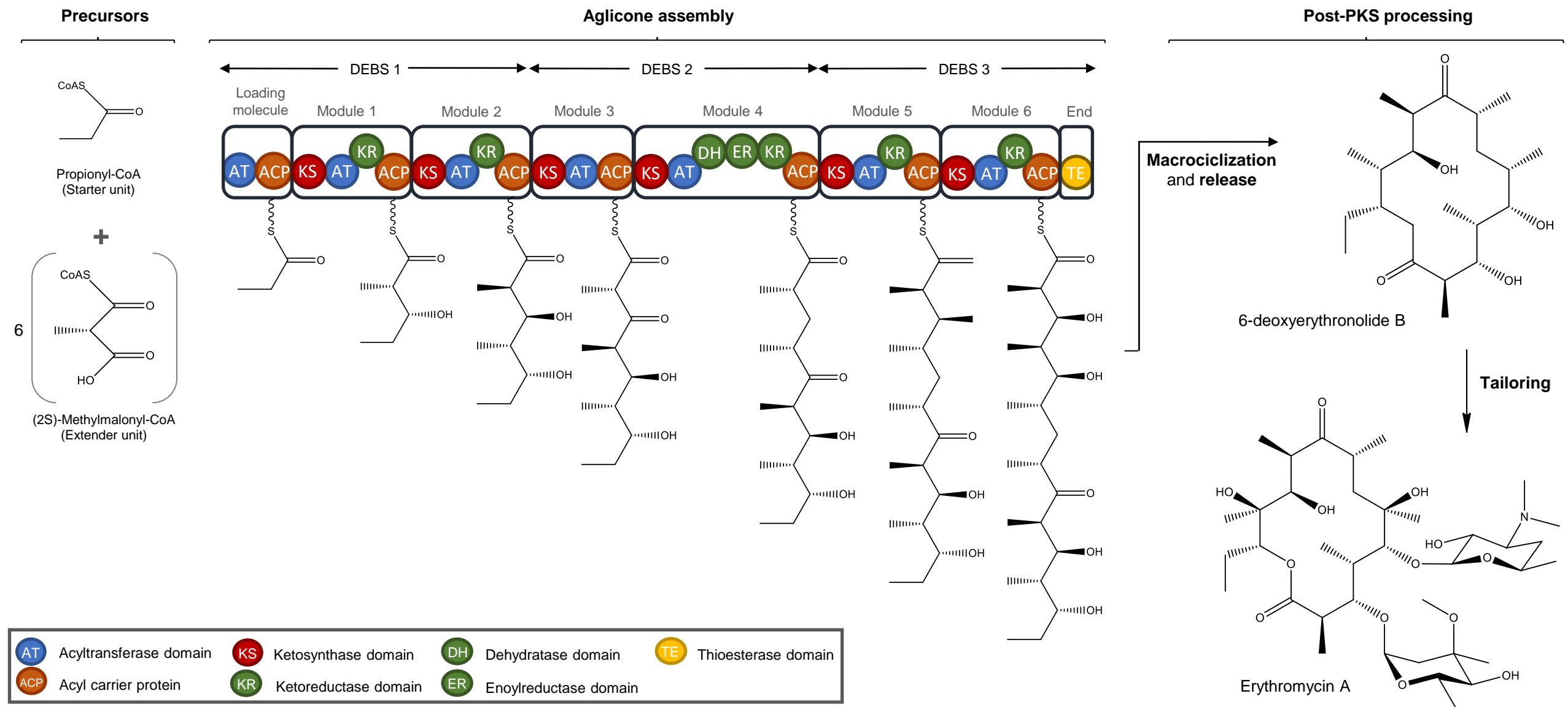
Lovastatin A
Cholesterol-lowering drug

General PKS composition



AT	Acyltransferase domain	KS	Ketosynthase domain	DH	Dehydratase domain	TE	Thioesterase domain
ACP	Acyl carrier protein	KR	Ketoreductase domain	ER	Enoylreductase domain		

“DEBS”: the prototype of type I PKS

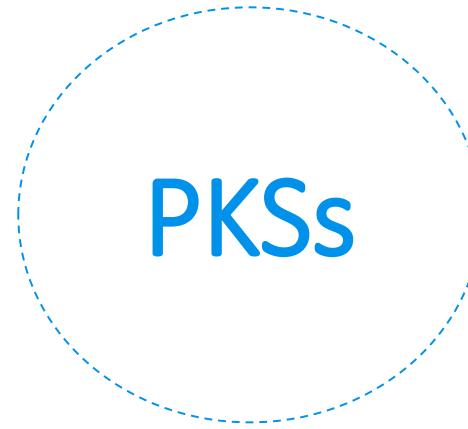


PKSs classification



Based on
enzyme architecture:

Type I
Type II
Type III



Based on
domain organization:

Iterative
Modular (only type I)

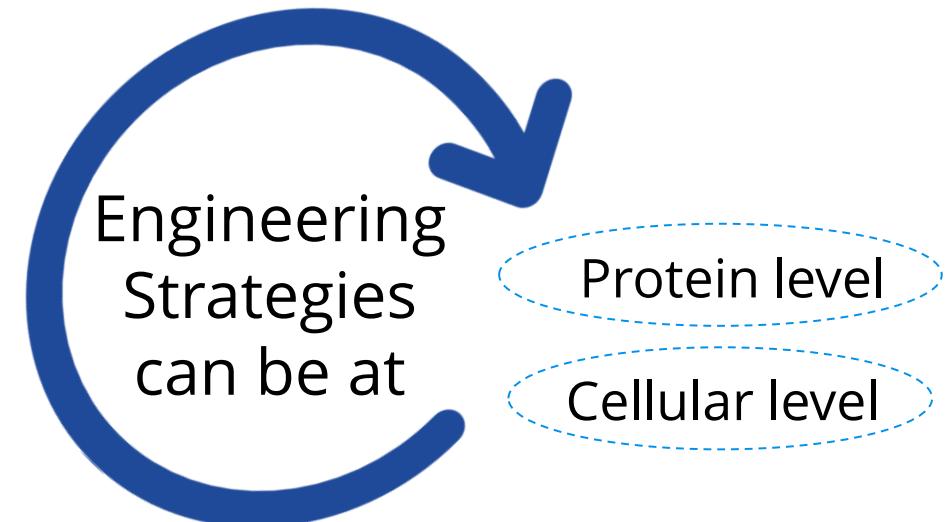
There are also diverse **PKS-NRPS hybrids** worth to mention.

Engineering Polyketides

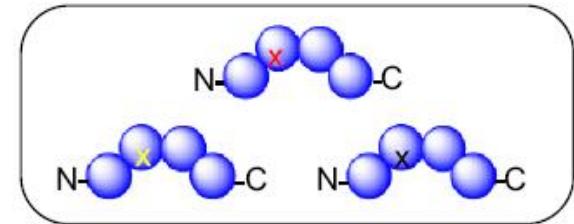
Polyketides are promising targets for **synthetic biology**:

- Highly modular architecture;
- Clinical relevance;
- High abundance.

Pathways can be manipulated/redesigned
to produce new molecules.



PKS protein modifications



AT, KR
mutagenesis

Possible
strategies

Reductive
loop swaps

Domain
fusion

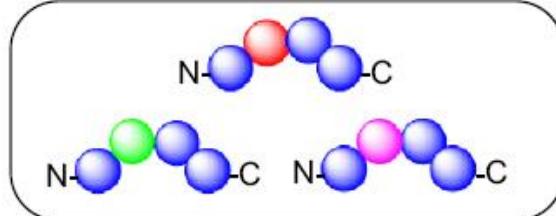
Modification
of active sites

Domain
swapping

Different substrate
specificity or higher
substrate promiscuity.

Change of pool of
building blocks.

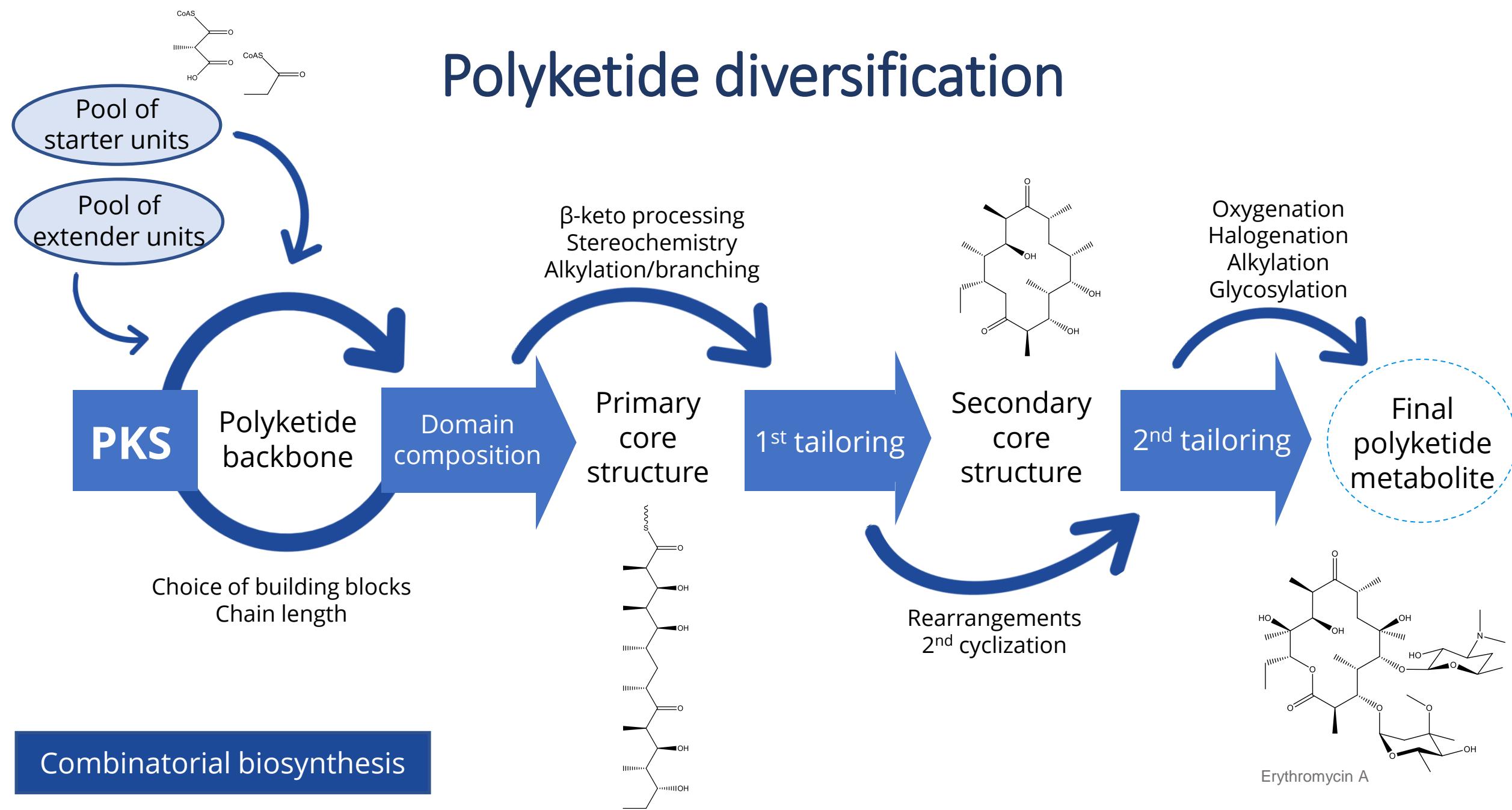
Entire modules or
single domains (TE, AT).



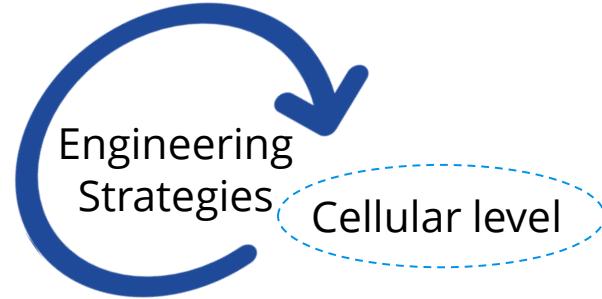
Engineering
Strategies

Protein level

Polyketide diversification



Heterologous expression of BGCs



Strategy to produce compounds

- Too complex to be chemically synthesized.
- Produced by complex, slow-growing microorganisms.



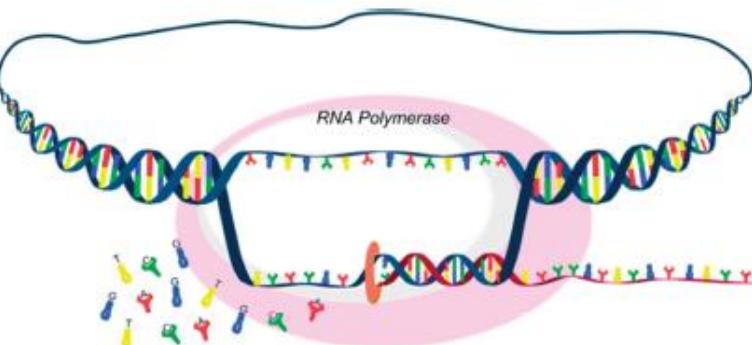
- Allow the expression of cryptic (silent) BGCs.
- Overproduction of target compounds.

Possible hosts:

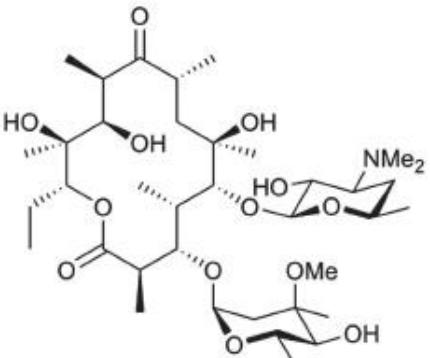
- *Streptomyces*
- Myxobacteria
- *Escherichia coli*
- *Saccharomyces cerevisiae*

Heterologous production of polyketides was first demonstrated with *Streptomyces parvulus* in 1984.

Challenges in Heterologous BGC expression

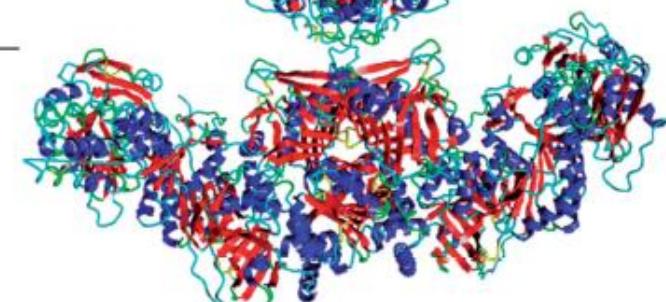


Solutions: Promoter replacement
Transcription from native promoters

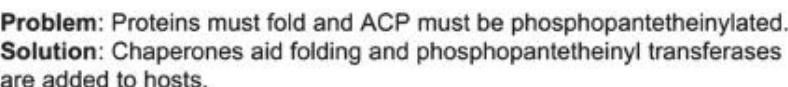


Erythromycin A

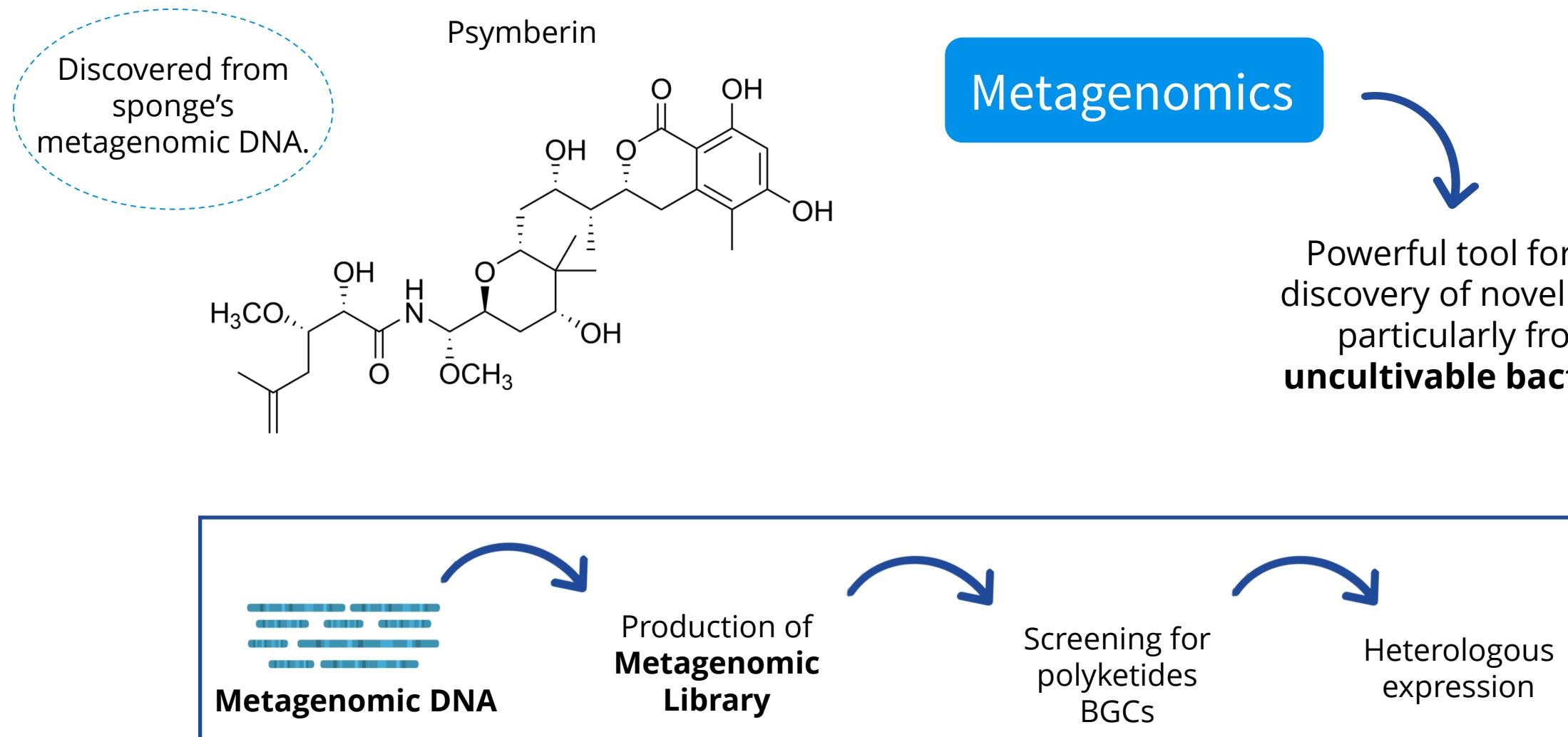
Solution: Co-expression of resistance pathway/Sensitive host has not prevented mg/L of product formation.



Problem: Precursors can be missing.
Solution: Add precursors or their biosynthetic pathways to hosts.



Discovery methods of new BGCs



Discovery methods of new BGCs

The Genomics Era



Has led to a shift in Natural Products Research.

Genome Mining



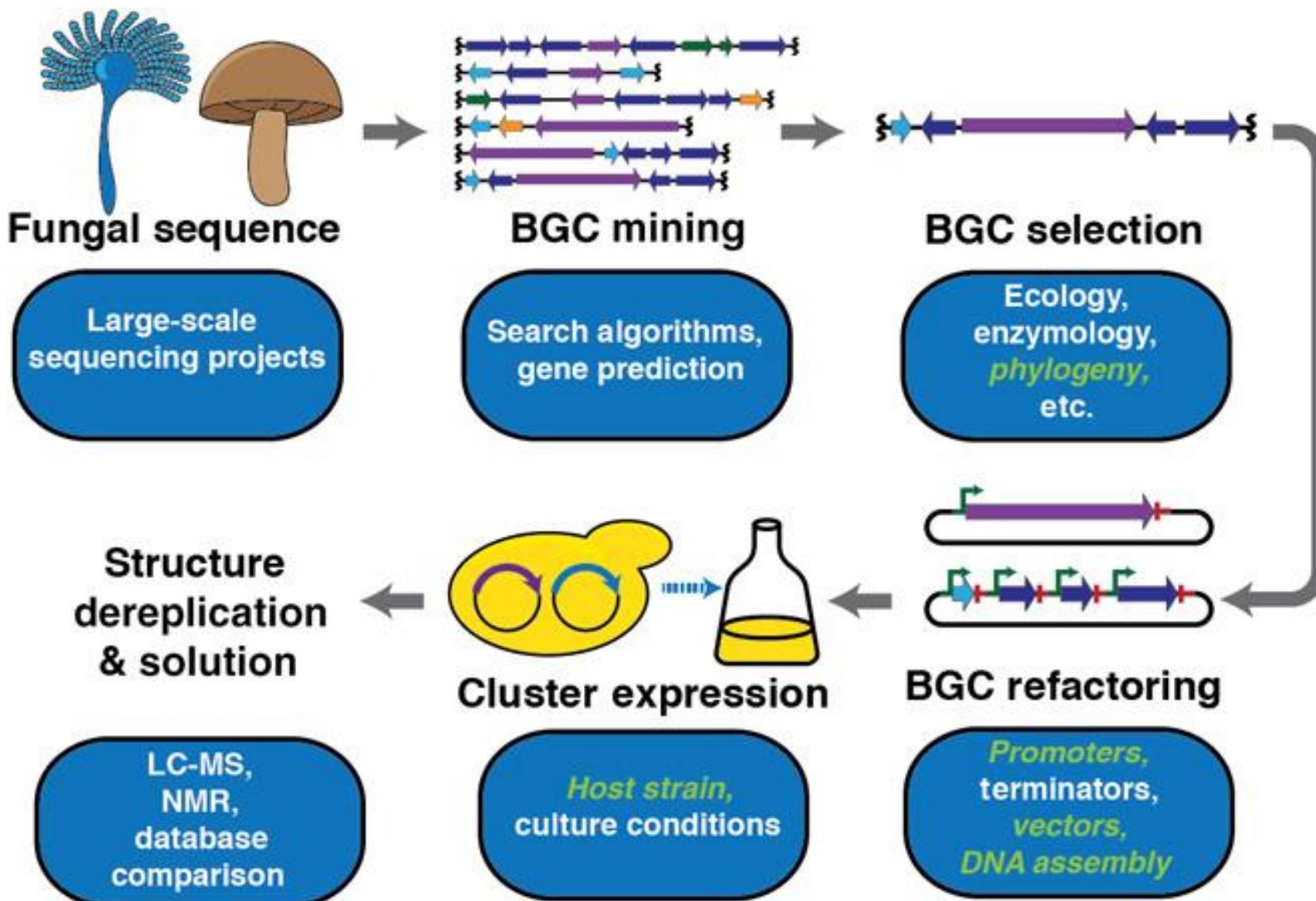
In silico discovery of novel BGCs.



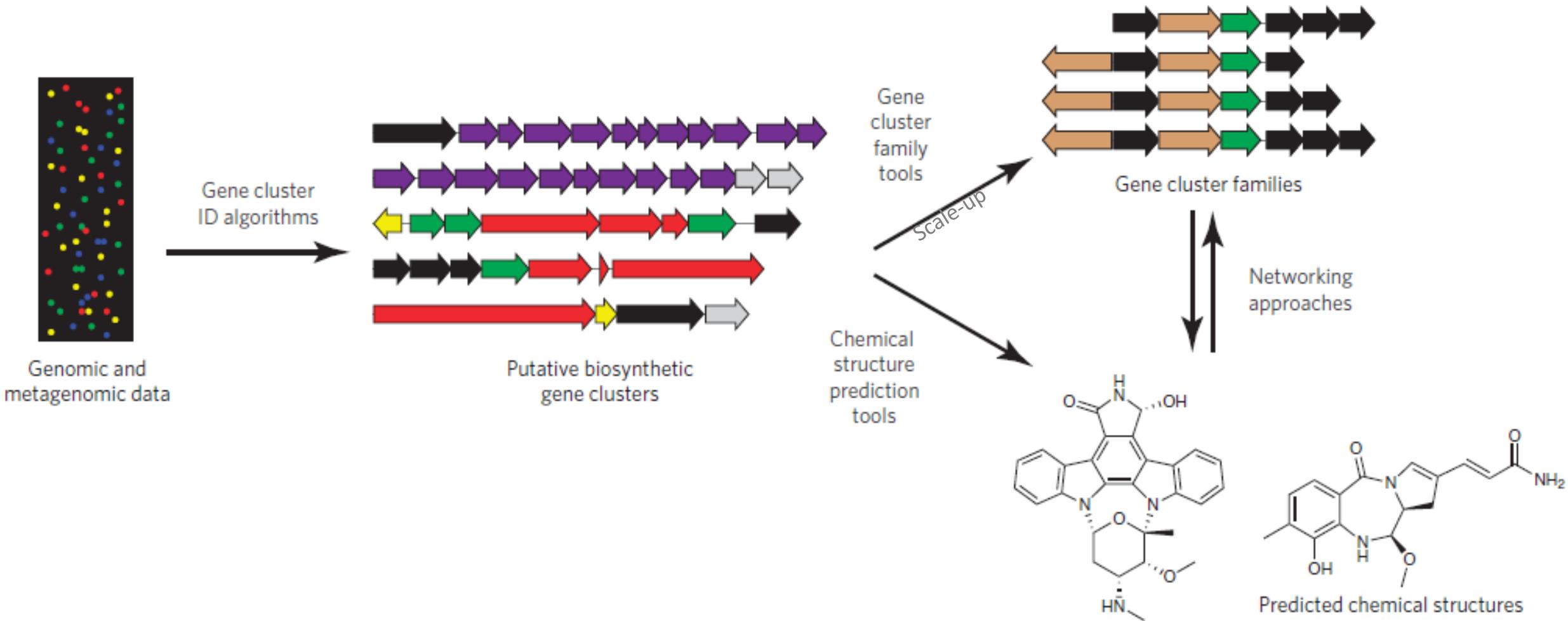
Through several different informatic algorithms and databases.



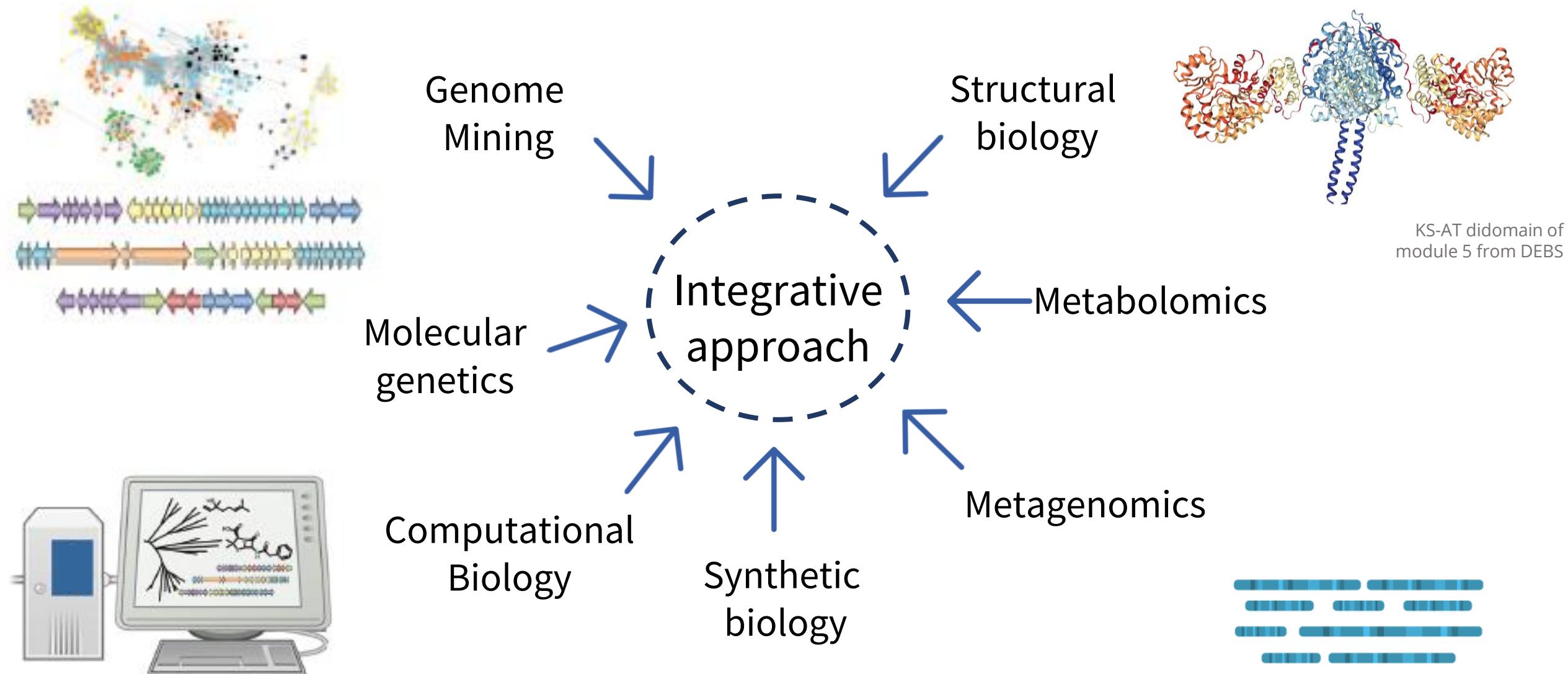
Discovery methods of new BGCs



Computational Approaches in Natural Products Discovery



The ultimate goal: an integrative approach



Case-study:

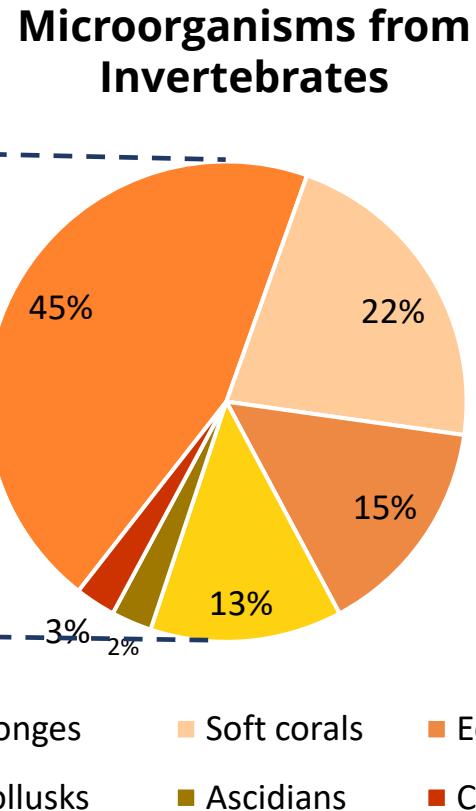
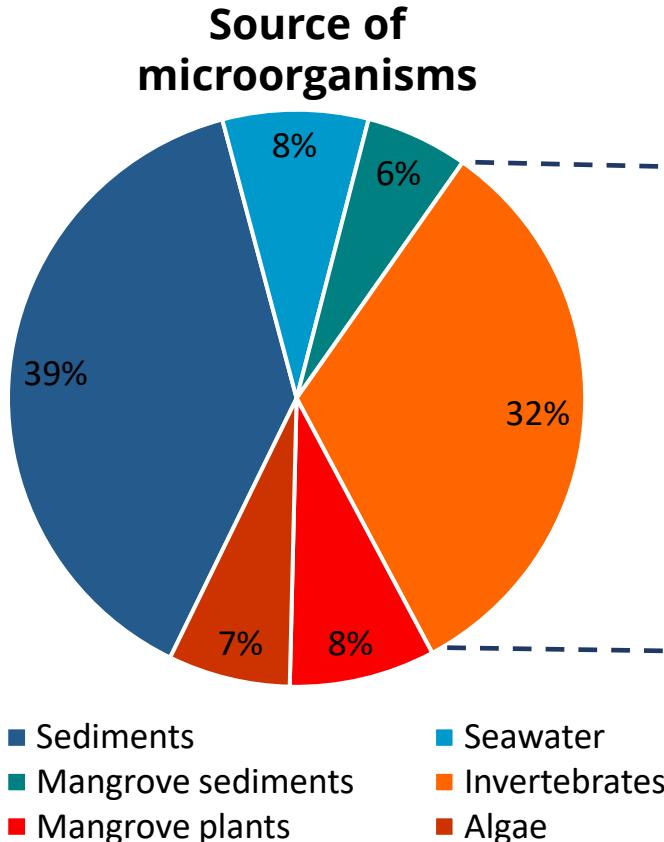
Secondary metabolite biosynthesis by *Aquimarina* species:
emerging bioactivities from the rare marine biosphere

The marine environment

**Is a prolific
source of novel
bioactive
natural products**



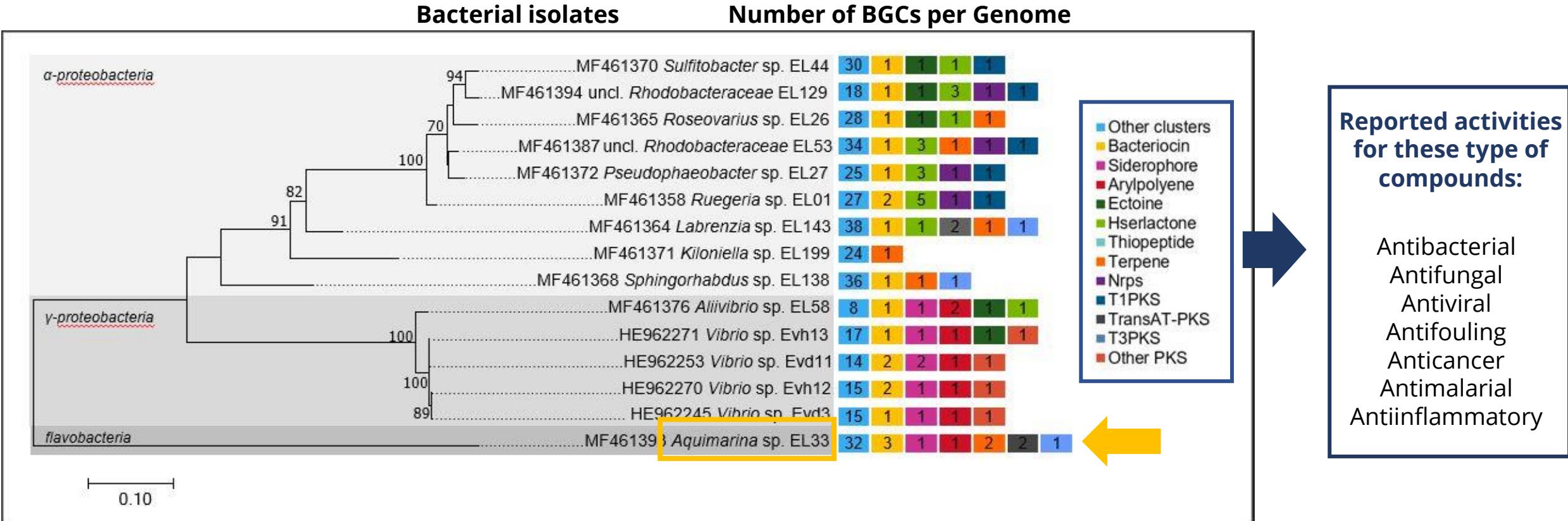
Novel Natural Products from the Seas



Sponges and soft corals stand out as hosts for microorganisms with potential for the biosynthesis of new compounds.

Year: 2013

Potential for Secondary Metabolite Synthesis in Soft Coral-Associated Bacteria



440 biosynthetic gene clusters (BGCs) on the genomes of 15 bacterial associates (12 genera) isolated from the soft corals *Eunicella labiata* and *Eunicella verrucosa*.

The *Aquimarina* genus



Phylum: Bacteroidetes
Family: Flavobacteriaceae
Genus: ***Aquimarina***

- Gram-negative bacteria;
- Strictly marine;
- Heterotrophic;
- Versatile carbon metabolism;
- Yellow or orange-pigmented.

The *Aquimarina* genus



Unknown
biotechnological
potential?

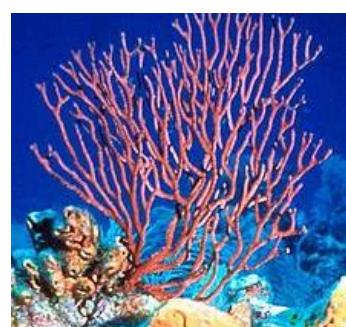
- Involved in the regulation of harmful microbial blooms through **mediation of carbon and nitrogen cycling**.
- Emerging evidence of **pathogenic behavior in some marine invertebrates**.
- **Distinct secondary metabolism** already observed for some isolates.

Comparative genomics reveals complex natural product biosynthesis capacities and carbon metabolism across host-associated and free-living *Aquimarina* (*Bacteroidetes*, *Flavobacteriaceae*) species

Sandra G. Silva ¹, Jochen Blom,² Tina Keller-Costa ¹
and Rodrigo Costa ^{1,3*}

Comparison of 26 *Aquimarina* genomes from several isolation sources

HOST-ASSOCIATED (HA)



Marine sponges

Gorgonian coral

FREE-LIVING (FL)



Marine sediments

Seawater

Methods

Analysis of all available *Aquimarina* genomes at NCBI (25/02/2019)

Download of
26 genomes



16S rRNA
phylogenetic analysis



Annotation:
COGs and Pfams

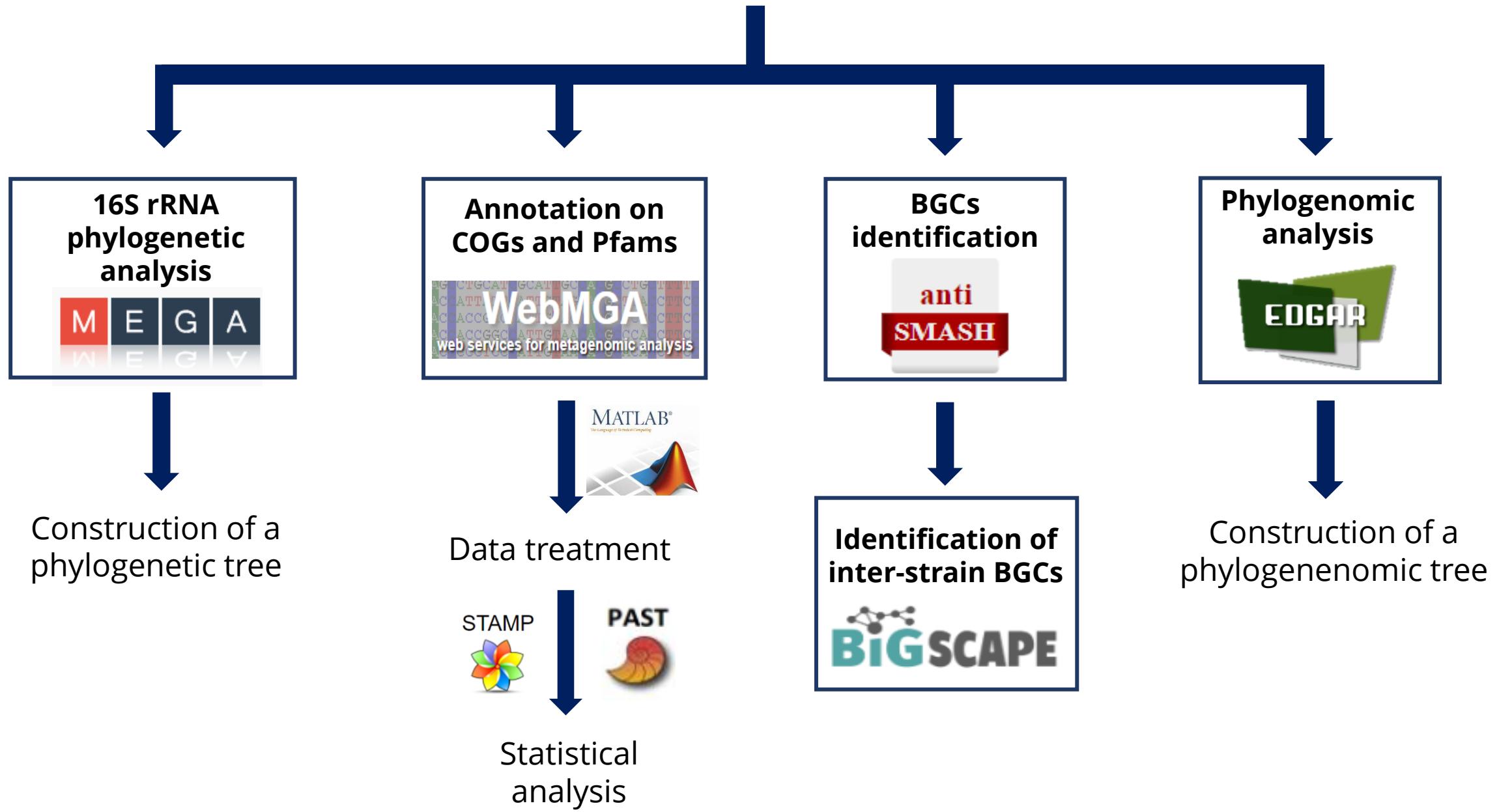


BGCs
identification



Phylogenomics
analysis

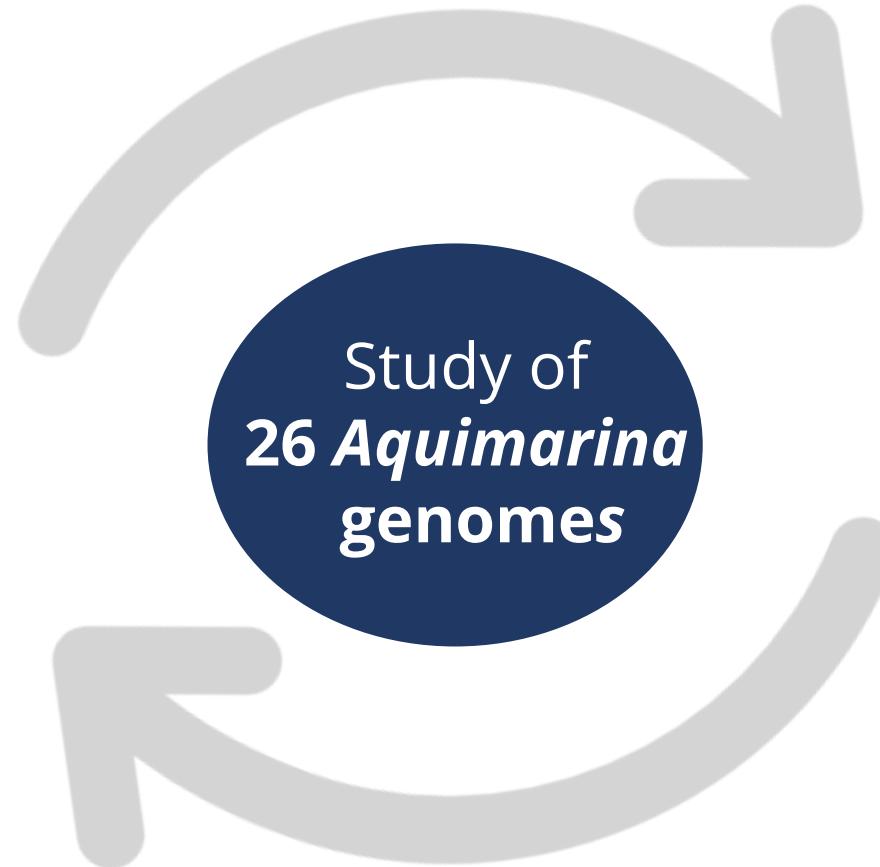




Goals

Search for biosynthetic gene clusters (BGCs).

Are *Aquimarina* species potential sources of novel bioactive natural products?



Describe general features of the genus.

Comparison between host-associated and free-living organisms.

Genomes Overview

- 26 genomes.
- Genome size range: from **4.07Mb** (*Aq. atlantica*) to **6.5 Mb** (*Aq. AU119*). Average: **5,6 Mb**.
- GC content range: from **31.4** (*Aq. muelleri*) to **35.9** (*Aq. spongiae*). Average: **32.72%**.
- Average number of coding sequences per genome: **5480 CDSs**.
- **Core genome:** 1226 CDSs.
- **Pan genome:** 21211 CDSs.



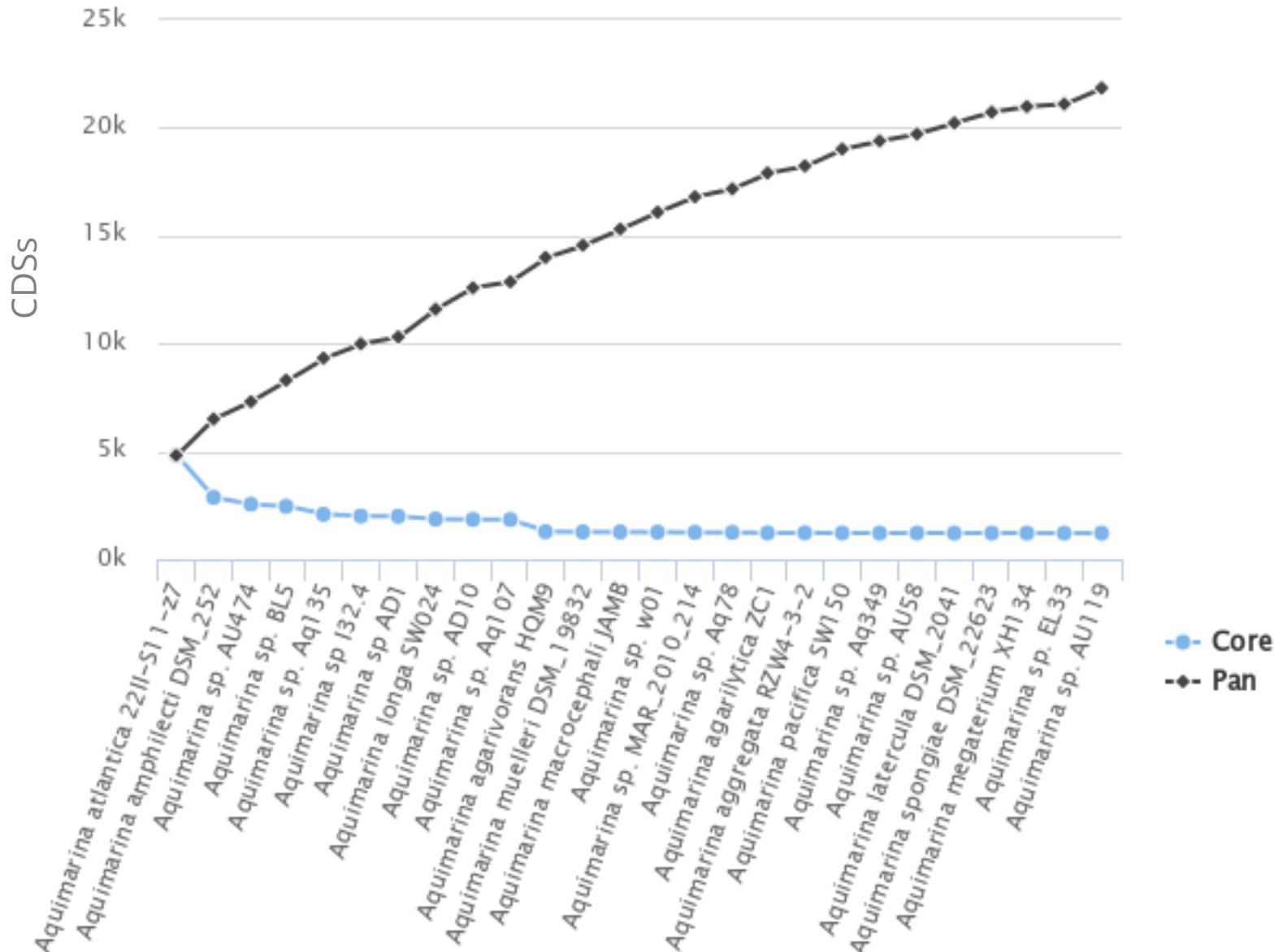
Whole-genome sequence alignment

These data suggests:
open pangenome.

Common in species living in a community.

Tendency to large genomes and high horizontal rate of genes transfer.

Core vs. Pan Plot



Functional Annotation

COG

Clusters of Orthologous Groups

44%

2320

1024

Number of different ORFs

Core

(Nr. ORFs present in all strains)

248

Unique

(Nr. of ORFs only present in one strain)

87646

Total number of ORFs

4187

1130

1716

27%

3371

Average of number of ORFs per strain

242234

9317

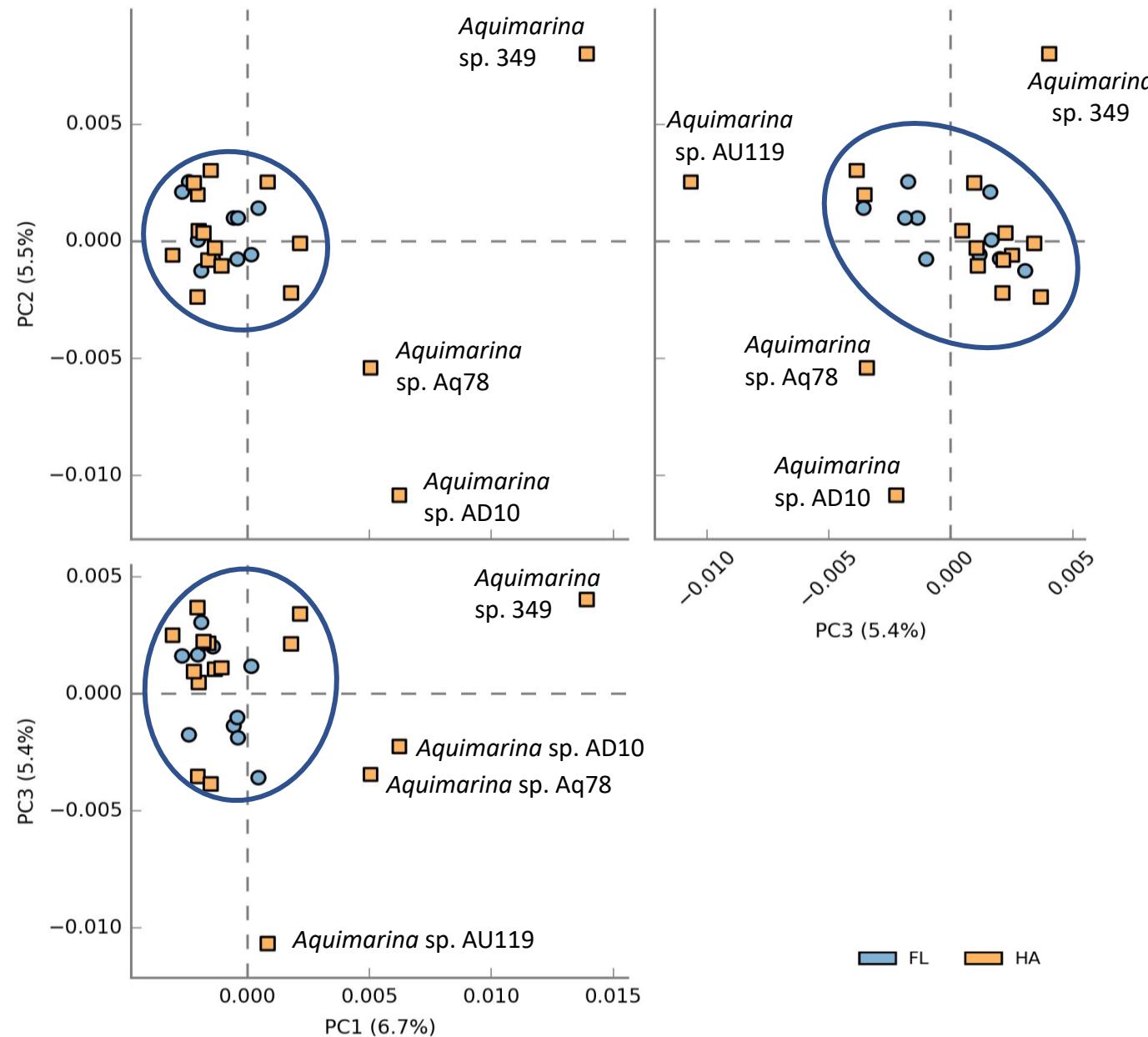
Pfam

Protein families' database

Principal Component Analysis (PCA)

Pfam
annotation

Absence of a statistical difference between annotated genomes

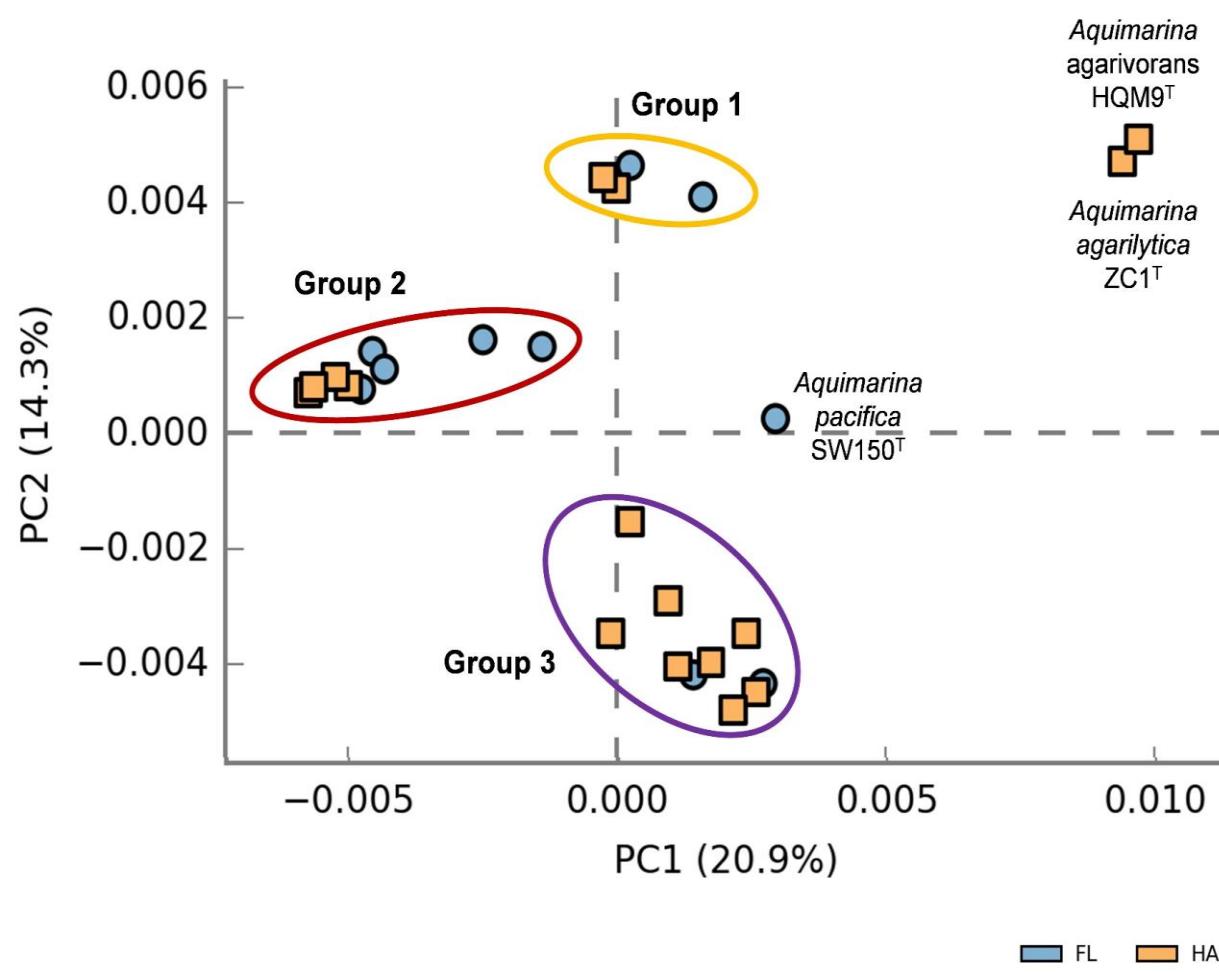


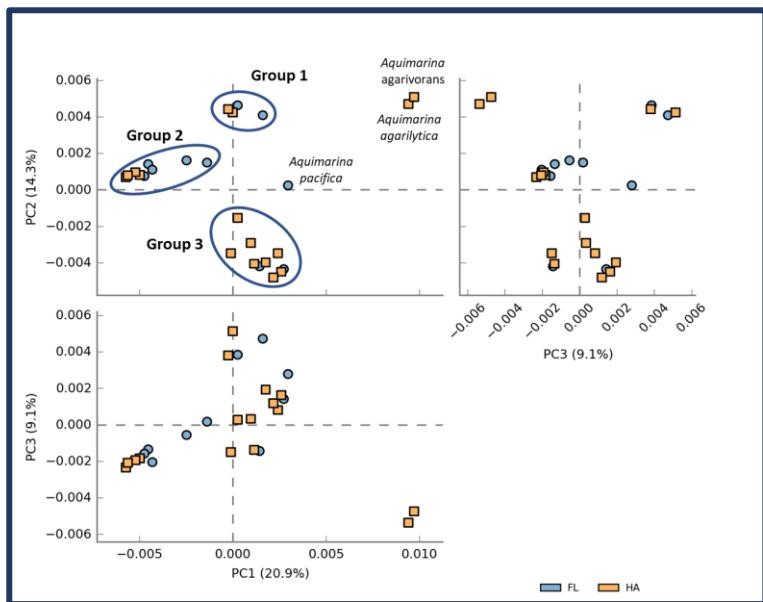
One single group:

- Aquimarina sp. I32.4
- Aquimarina longa
- Aquimarina muelleri
- Aquimarina sp. Aq135
- Aquimarina sp. w01
- Aquimarina sp. MAR
- Aquimarina atlantica
- Aquimarina macrocephali
- Aquimarina sp. AU58
- Aquimarina sp. EL33
- Aquimarina megaterium
- Aquimarina sp. AU474
- Aquimarina spongiae
- Aquimarina sp. Aq107
- Aquimarina aggregata
- Aquimarina latercula
- Aquimarina sp. BL5
- Aquimarina sp. AD10
- Aquimarina pacifica
- Aquimarina agarivorans
- Aquimarina agarilytica
- Aquimarina amphilecti

Outside of the group:

- Aquimarina sp. 349
- Aquimarina sp. 78
- Aquimarina sp. AD10
- Aquimarina sp. AU119





Are these groups
statistically significant?



Yes

Confirmed by one-way
Permanova
(Permutational analysis of
variance)

Division of the 26
genomes into **3 clusters**



Which COGs are more contributive for
the formation of these 3 groups?



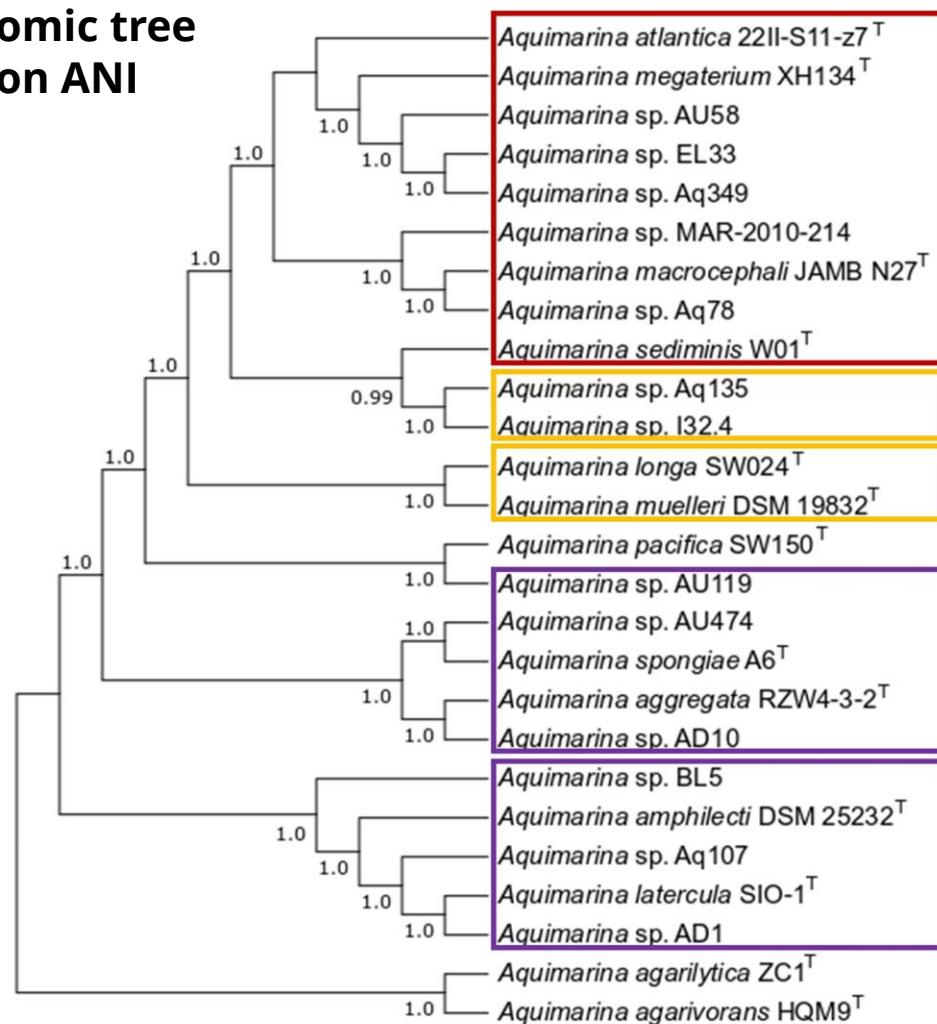
SIMPER analysis

**COGs related with
secondary metabolism**

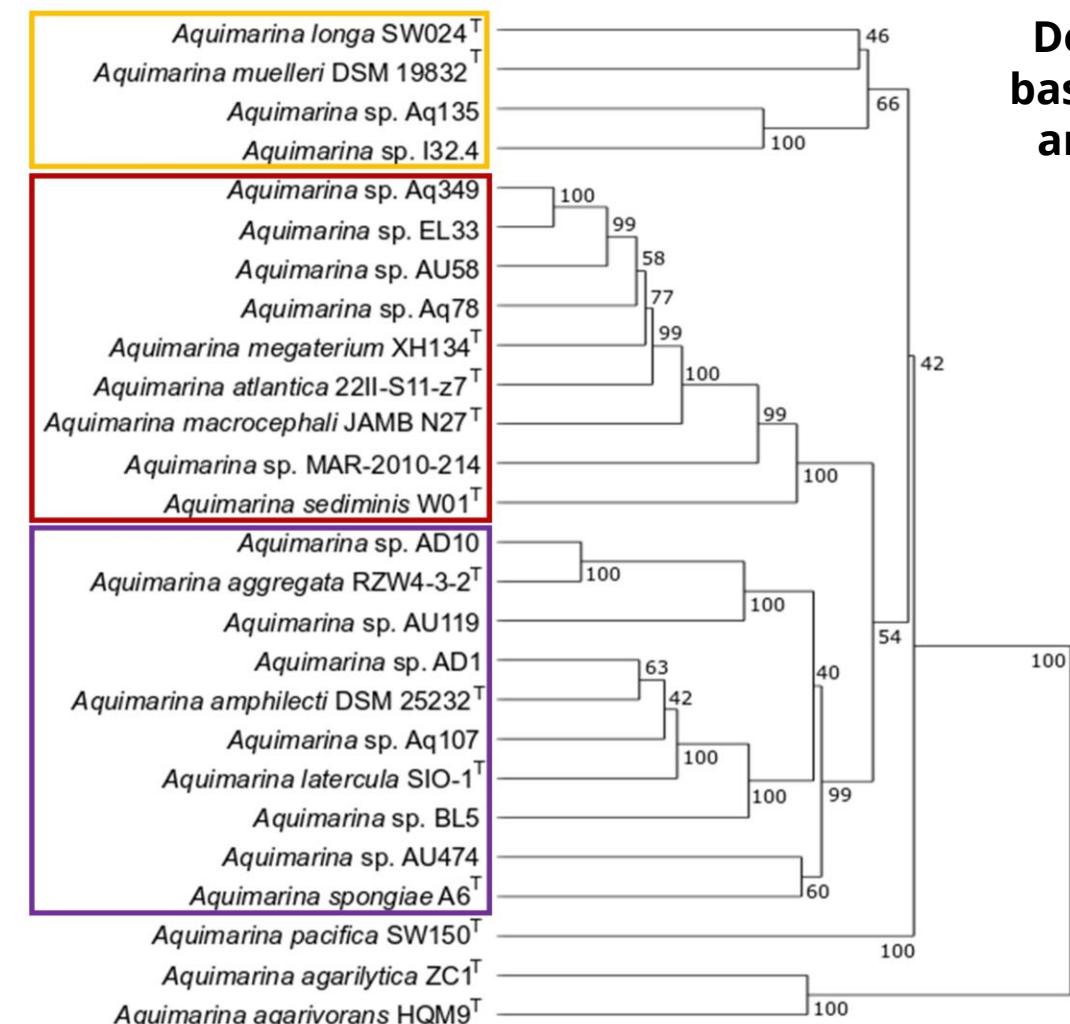
Taxon	Av. dissim	Contrib. %	Annotation
COG3321	0,05777	0,3766	Acyl transferase domain in polyketide synthase (PKS) enzymes
COG4886	0,04966	0,3238	Leucine-rich repeat (LRR) protein
COG2273	0,04762	0,3105	Beta-glucanase, GH16 family
COG2207	0,04713	0,3073	AraC-type DNA-binding domain and AraC-containing proteins
COG3275	0,04656	0,3036	Sensor histidine kinase, LytS/YehU family
COG1020	0,04599	0,2999	Non-ribosomal peptide synthetase component F
COG3279	0,04131	0,2693	DNA-binding response regulator, LytR/AlgR family
COG3979	0,03851	0,251	Chitodextrinase
COG3501	0,03683	0,2401	Uncharacterized conserved protein, implicated in type VI secretion and phage assembly
COG2335	0,03568	0,2327	Uncharacterized surface protein containing fasciclin (FAS1) repeats

Phylogeny primarily shapes the metabolism of *Aquimarina* species

**Phylogenomic tree
based on ANI**



■ Group 1 ■ Group 2 ■ Group 3



**Dendrogram
based on COG
annotation**

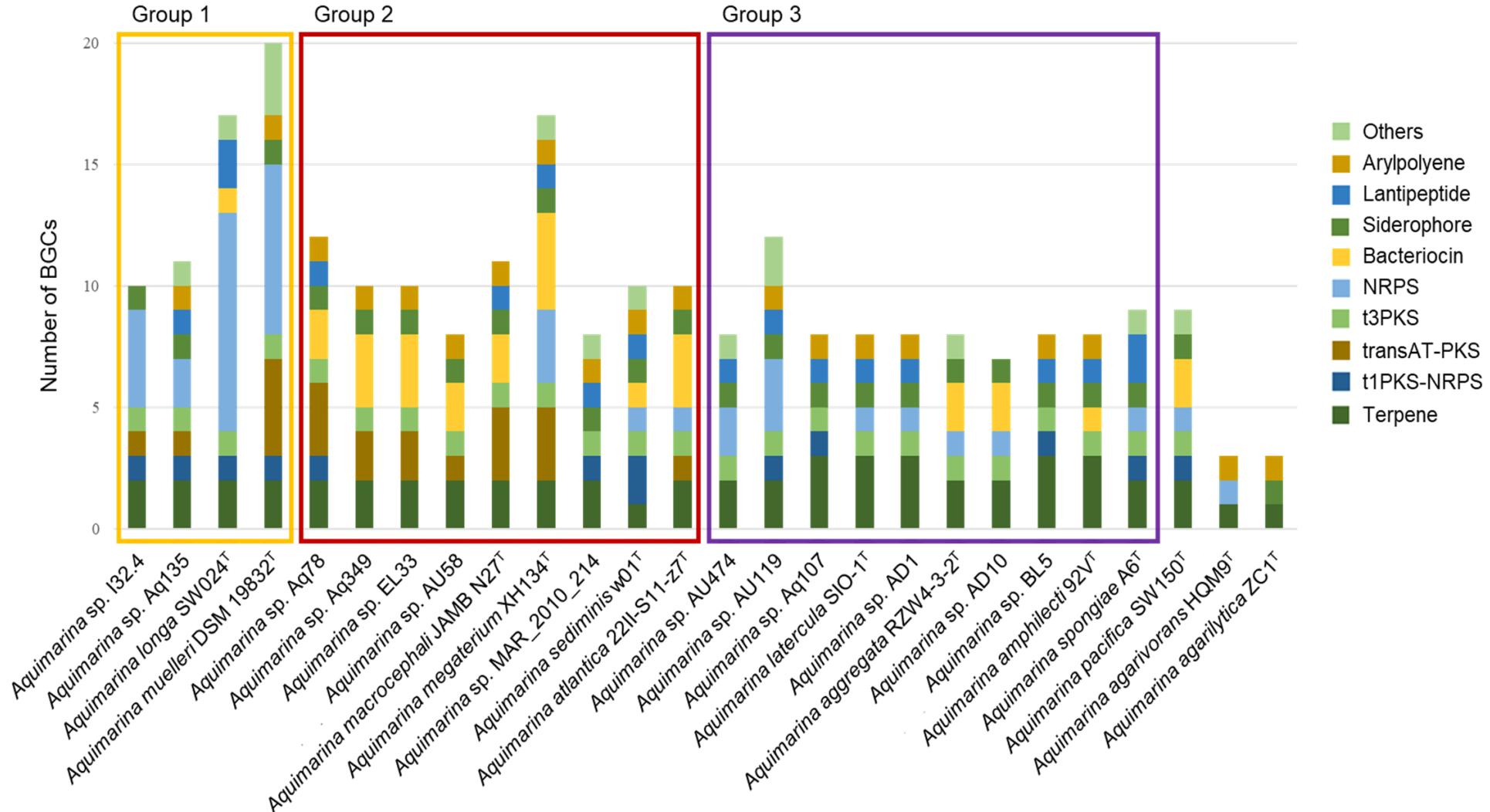
Identification of BGCs

Total count:
928 BGCs

anti
SMASH

54 terpenes
13 t1PKS-NRPS
21 transATPKS
24 t3pPKS
39 NRPS

High biosynthetic diversity



108 Saccharide GCFs

19 RiPPs GCFs

48 NRPS GCFs

13 T1PKS GCFs

209 Others GCFs

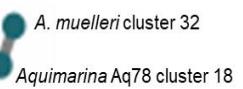
13 Terpenes GCFs

19 PKSother GCFs

10 PKS/NRPS GCFs

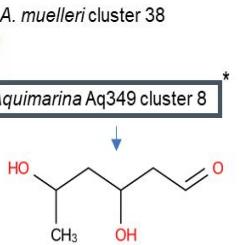
PKSother is composed by 8 different GCFs grouped into 5 clans.

PKSother D

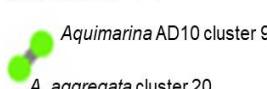


A. megaterium cluster 15
A. muelleri cluster 38

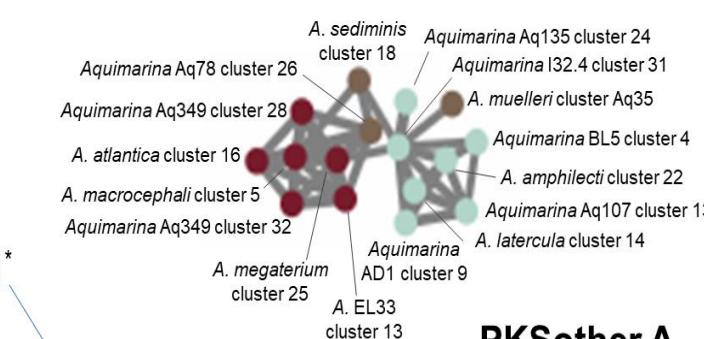
PKSother C



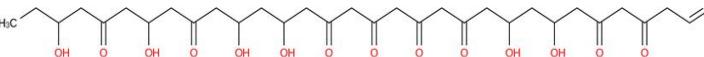
PKSother E



PKSother B



PKSother A

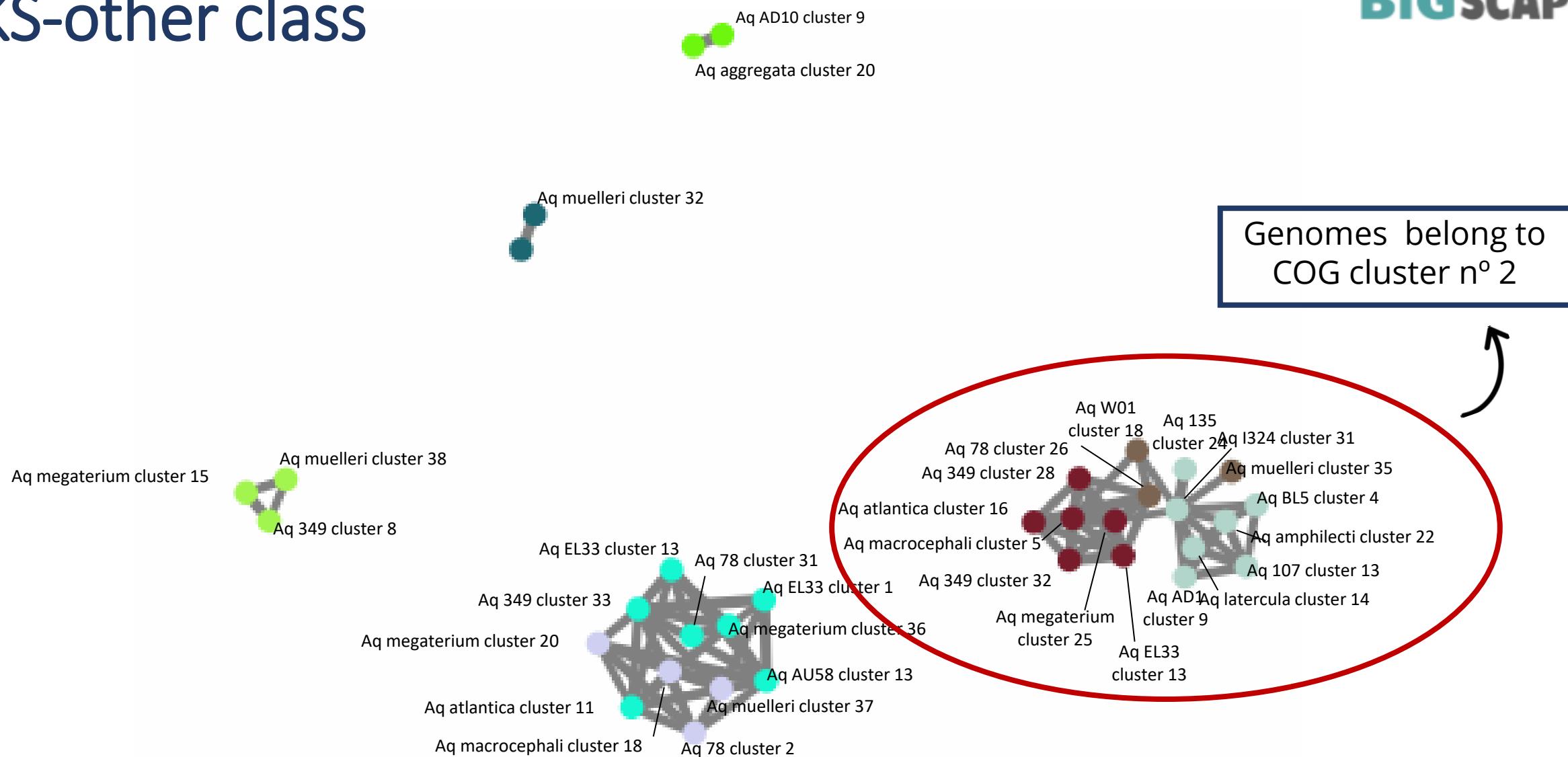


* Examples of clusters whose expected product was linked to structure predictions.

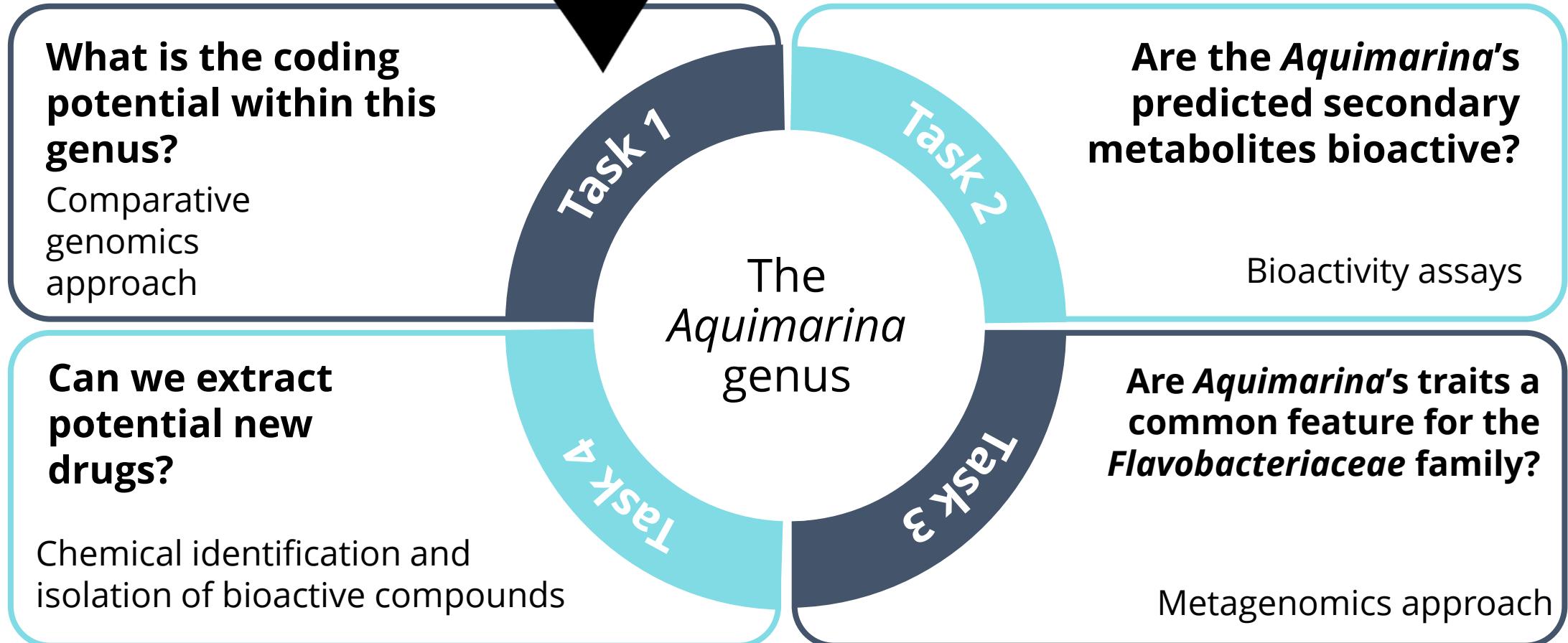
439

Gene Cluster Families

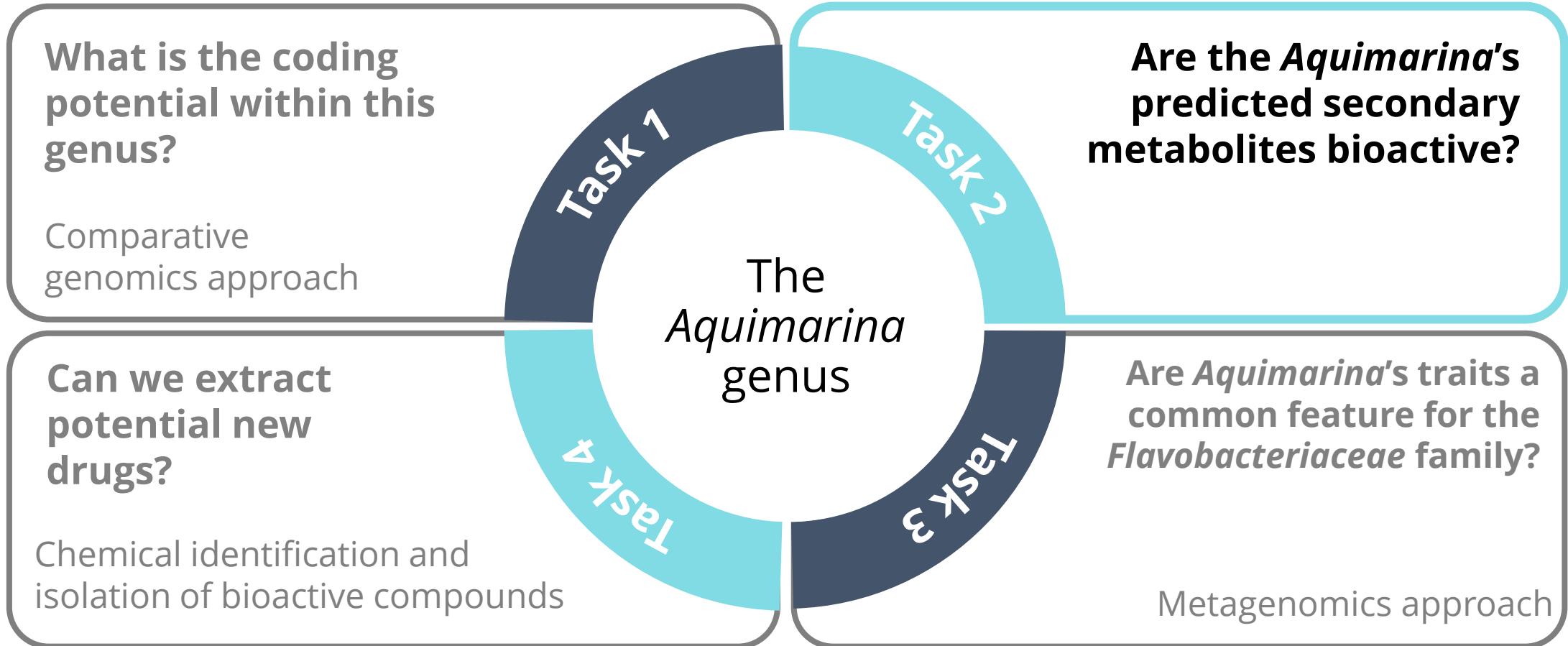
PKS-other class



Next steps

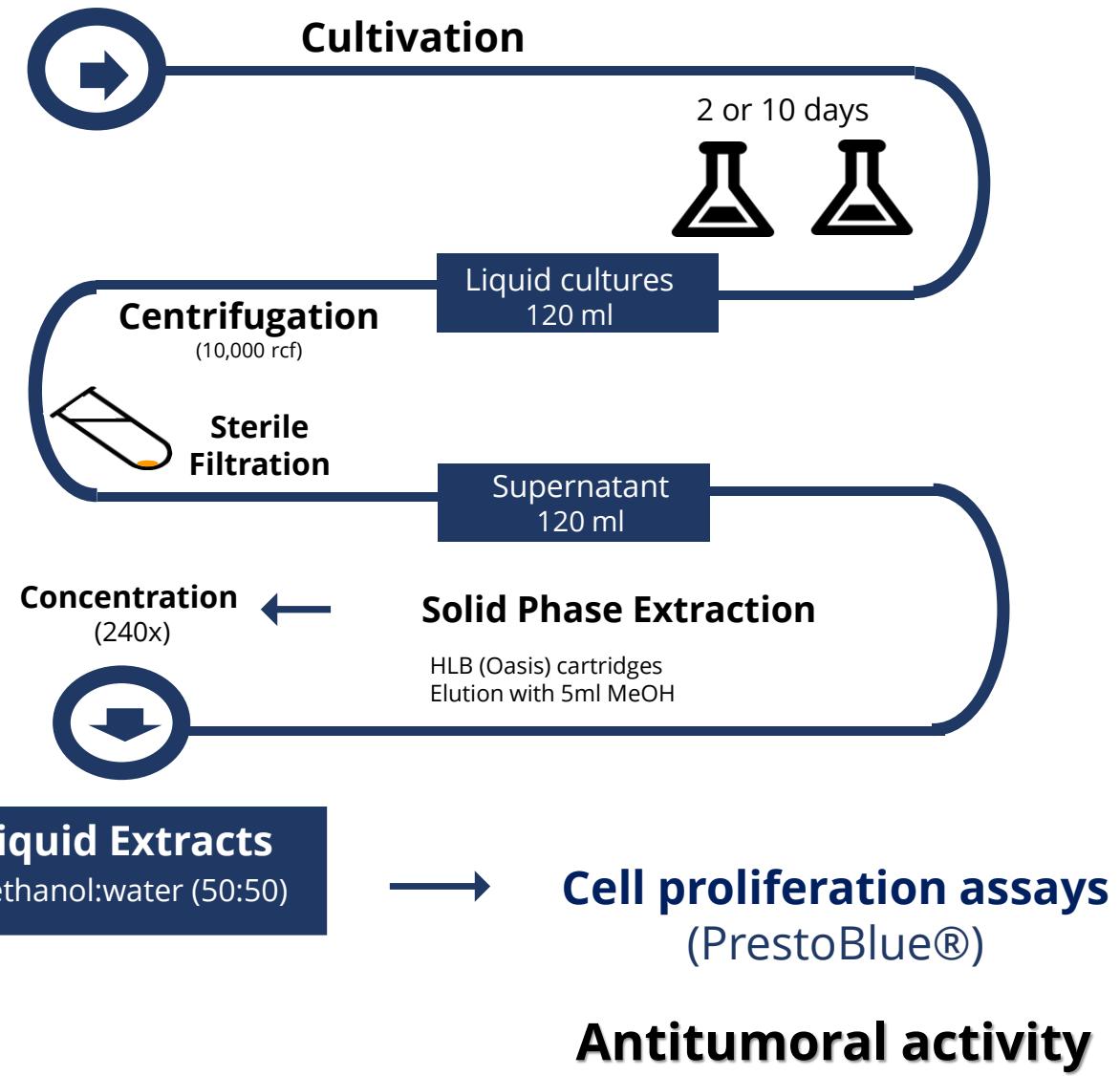


Next steps



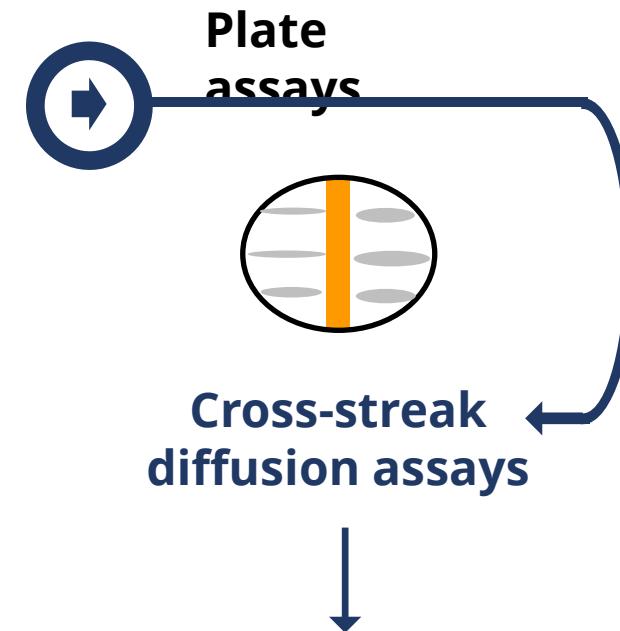
Are the *Aquimarina*'s predicted secondary metabolites bioactive?

9 *Aquimarina* strains



Are the *Aquimarina*'s predicted secondary metabolites bioactive?

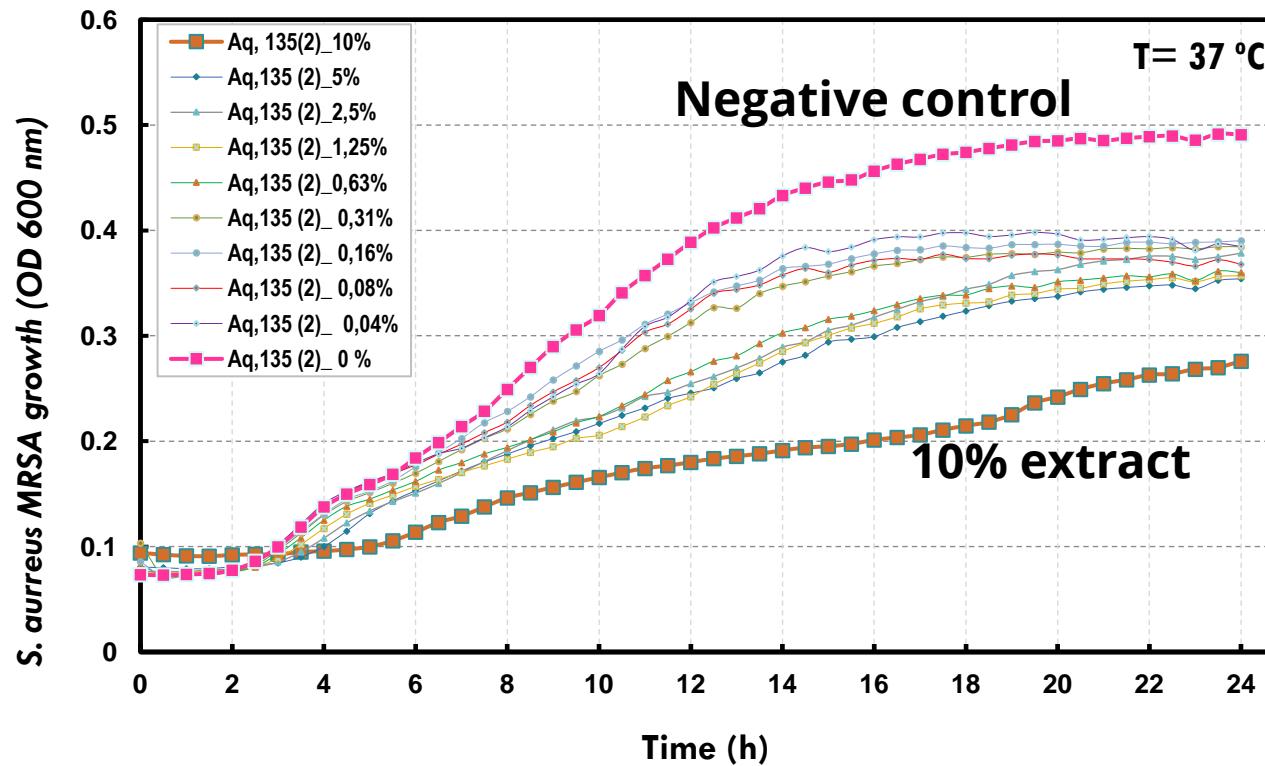
9 *Aquimarina* strains



Are the *Aquimarina*'s predicted secondary metabolites bioactive?

Aquimarina sp. strain Aq135 (day 2) extract versus *S. aureus* JE2 (MRSA)

Overnight
kinetic
assay



60% reduction of *S. aureus* MRSA growth
with the **10% Aq135 extract**

Are the *Aquimarina*'s predicted secondary metabolites bioactive?

Inhibition of marine bacteria growth in Cross-streak Diffusion Assays

	Vibro spp.								<i>Vibrio schiloi</i>	<i>Micrococcus</i> sp. MC110	<i>Pseudovibrio</i> sp. Pv125	<i>Labrenzia</i> sp. EL143
	EL22	EL36	EL38	EL41	EL49	EL62	EL67	EL112				
<i>A. muelleri</i> DSM19832	+++	++	++	+++	+++	++	+++	+++	-	-	+++	++
<i>A. latercula</i> DSM2041	-	-	-	+++	-	-	-	-	-	-	-	-
<i>A. spongiae</i> DSM22623	-	-	-	-	-	-	-	-	-	-	-	-
<i>Aquimarina</i> sp. Aq78	+	-	+	+++	-	+	++	+	-	-	+	-
<i>Aquimarina</i> sp. Aq107	+	-	+	+++	-	+	+	++	-	-	-	-
<i>Aquimarina</i> sp. Aq135	-	-	+	+++	+	-	++	+	-	-	++	++
<i>Aquimarina</i> sp. Aq349	++	-	+	+++	-	+	++	++	-	-	+	-
<i>Aquimarina</i> sp. EL33	++	-	++	+++	-	+	++	++	-	-	+	-
<i>Aquimarina</i> sp. EL43	++	-	+	+++	-	+	++	++	-	-	-	-

Growth inhibition	+++	++	+	+/-	-
	100%	75%	50%	25%	0%



Control +

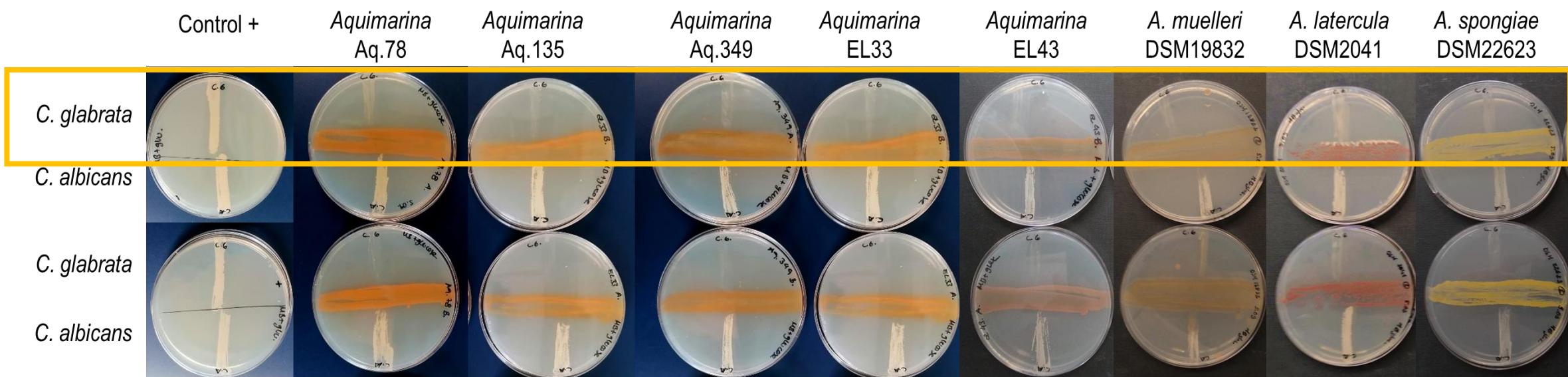


A. muelleri DSM19832

Are the *Aquimarina*'s predicted secondary metabolites bioactive?

Strong inhibitory activity of all *Aquimarina* strains against *C. glabrata* KChr606

Cross-streak Diffusion Assays

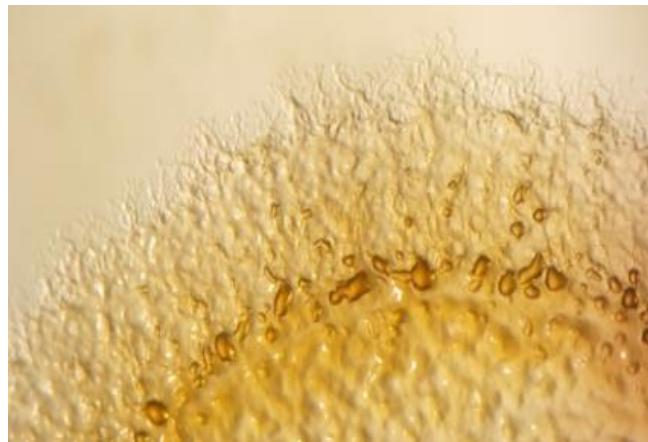


All cross-streak assays were performed at least in duplicates.

Are the *Aquimarina*'s predicted secondary metabolites bioactive?

Preliminary conclusions:

- *Aquimarina*'s extracts show anti-staphylococcal activity in "MIC"-type assays.
- In-plate activity was observed against several marine bacteria and *Candida glabrata*.

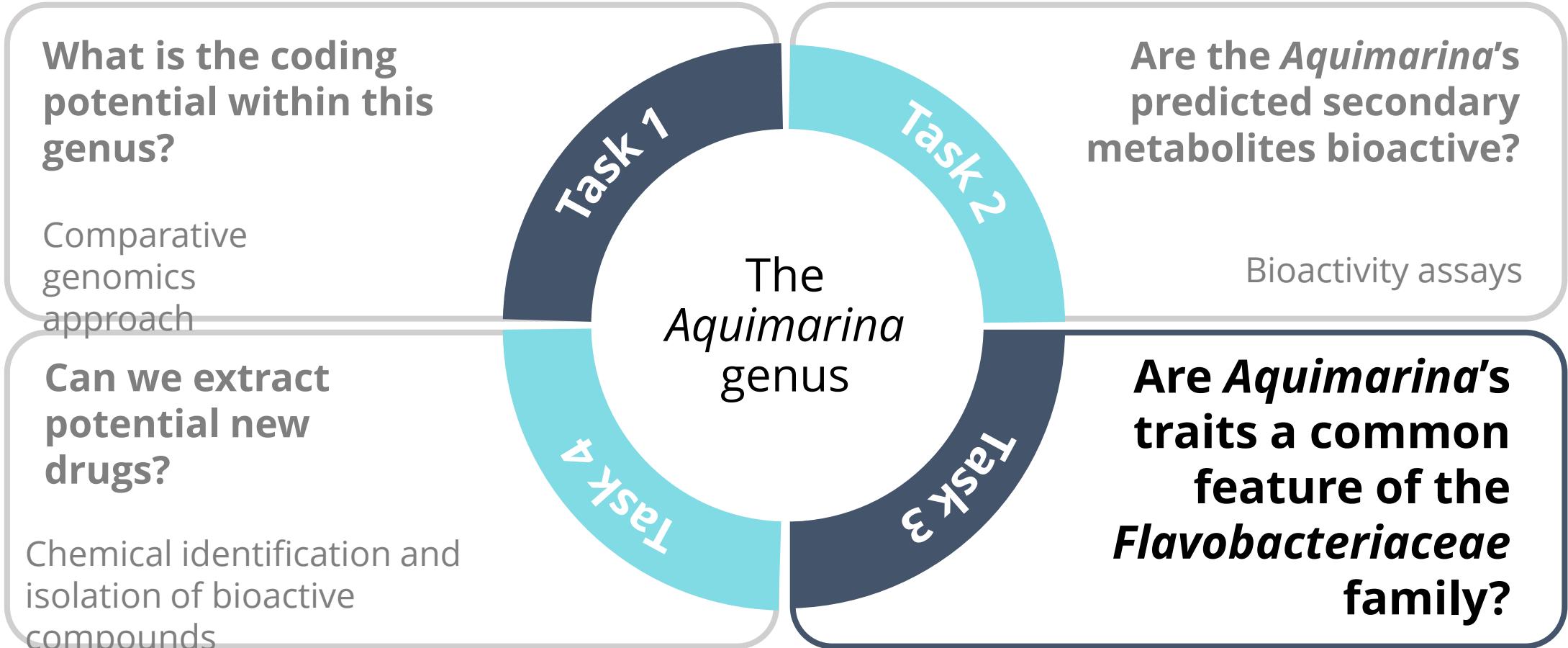


Aquimarina sp. EL33 (x25)

Future work:

- Conclusion of cell proliferation assays against human tumoral cells.
- LC-MS based "Metabolomic" profiling of the liquid extracts from *Aquimarina* strains.

Next steps

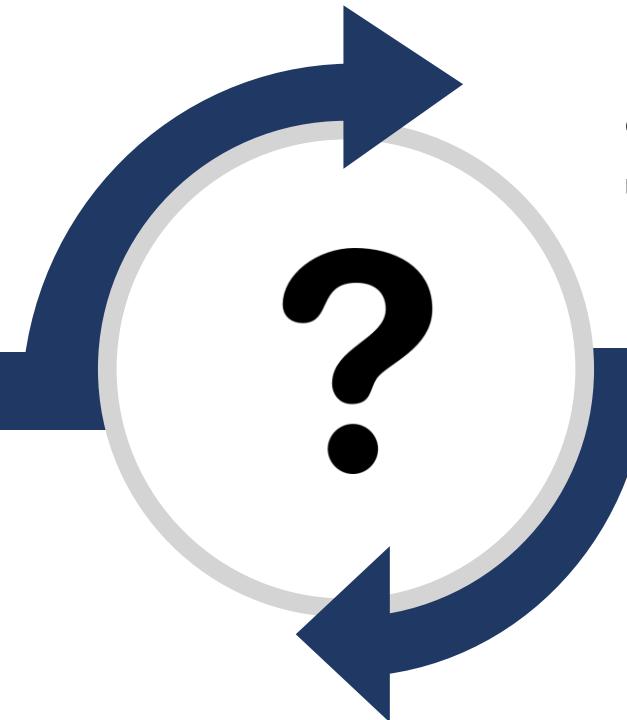


Are *Aquimarina*'s traits a common feature of the *Flavobacteriaceae* family?

The genomic identification of **BGCs** coupled with the observed **bioactivities** in the *Aquimarina* genus.

Raised an urgent question:
are all marine *Flavobacteriaceae* species an **underexplored biotechnological potential**?

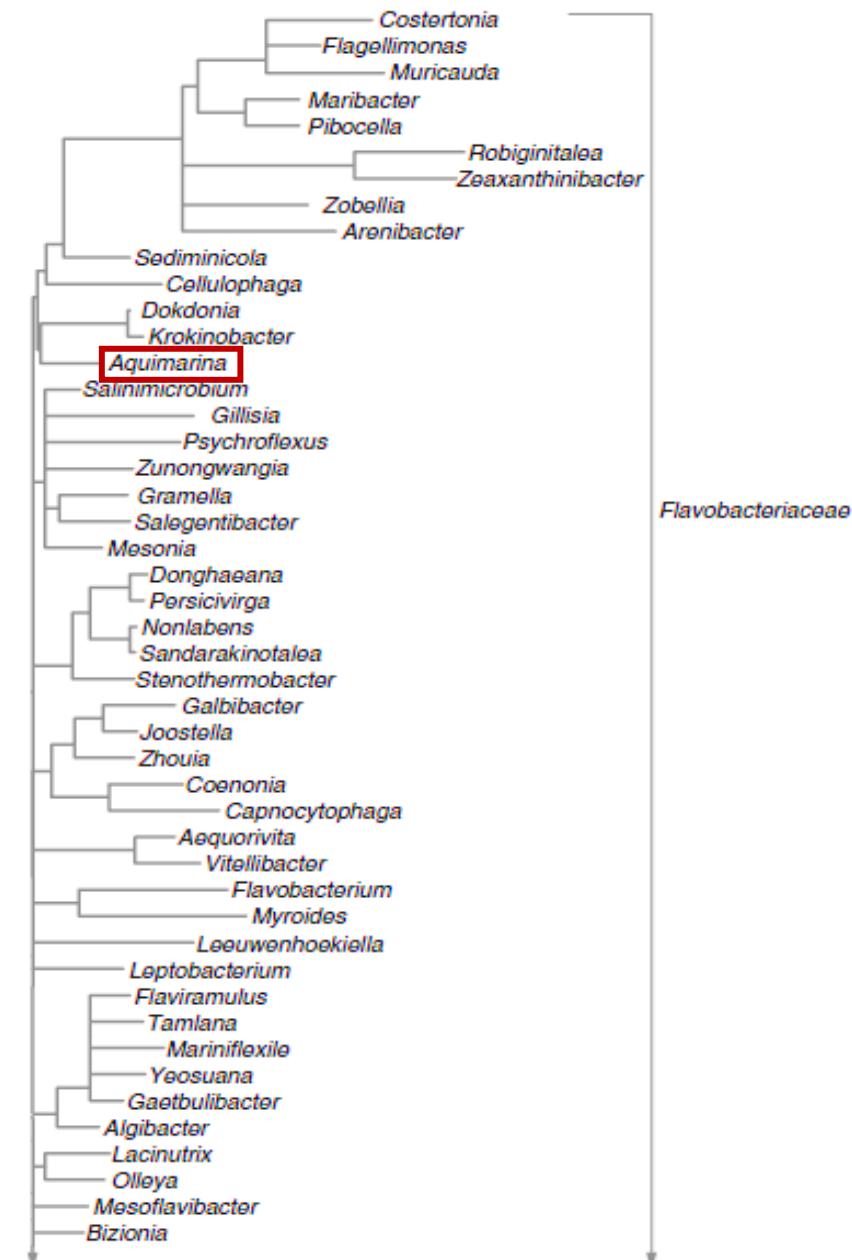
To answer this:
a metagenomic approach
was planned.



The Flavobacteriaceae family

Largest family in the phylum Bacteroidetes.

- Contains at least **90 genera** and hundreds of species.
- Wide variety of **marine, freshwater, and soil habitats**
Several known associations with animals and plants.



Are *Aquimarina*'s traits a common feature of the *Flavobacteriaceae* family?

The Flavobacteriaceae family

- Common specialization in the **degradation of high molecular weight (HMW) compounds.**

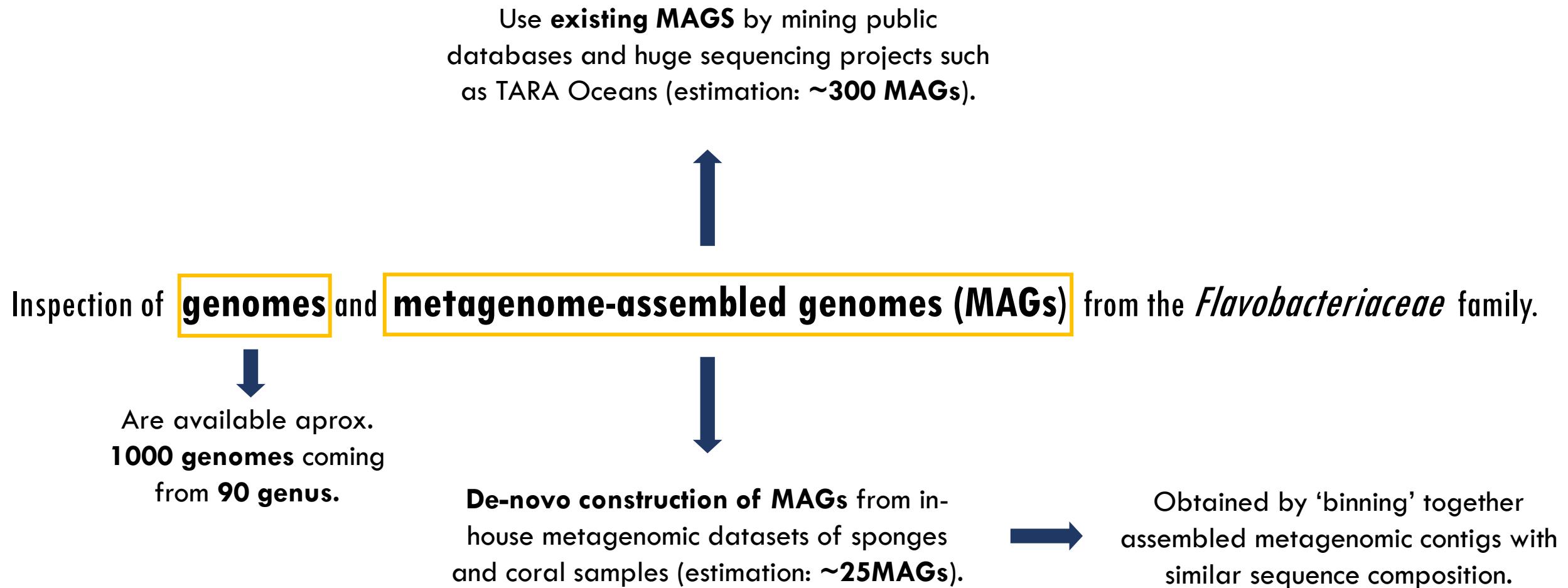


Found in high abundances during natural and induced **phytoplankton blooms.**

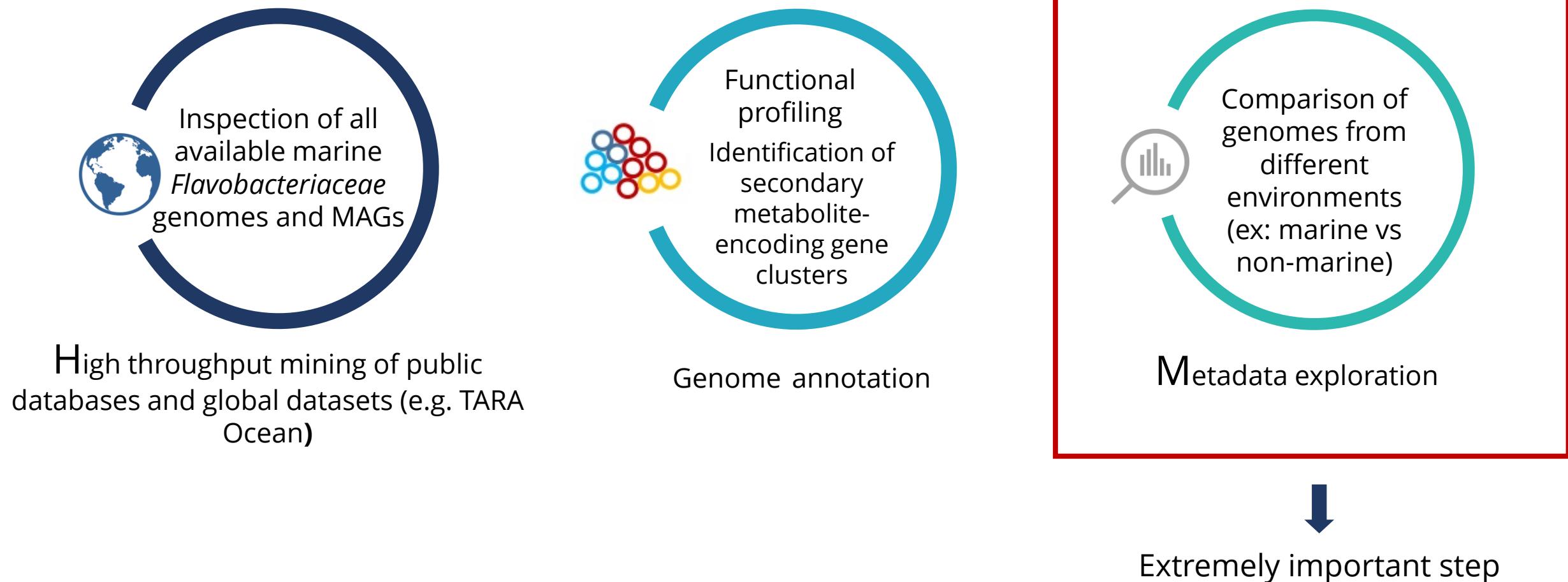
Phytoplankton Bloom
Iceland
June 24, 2010



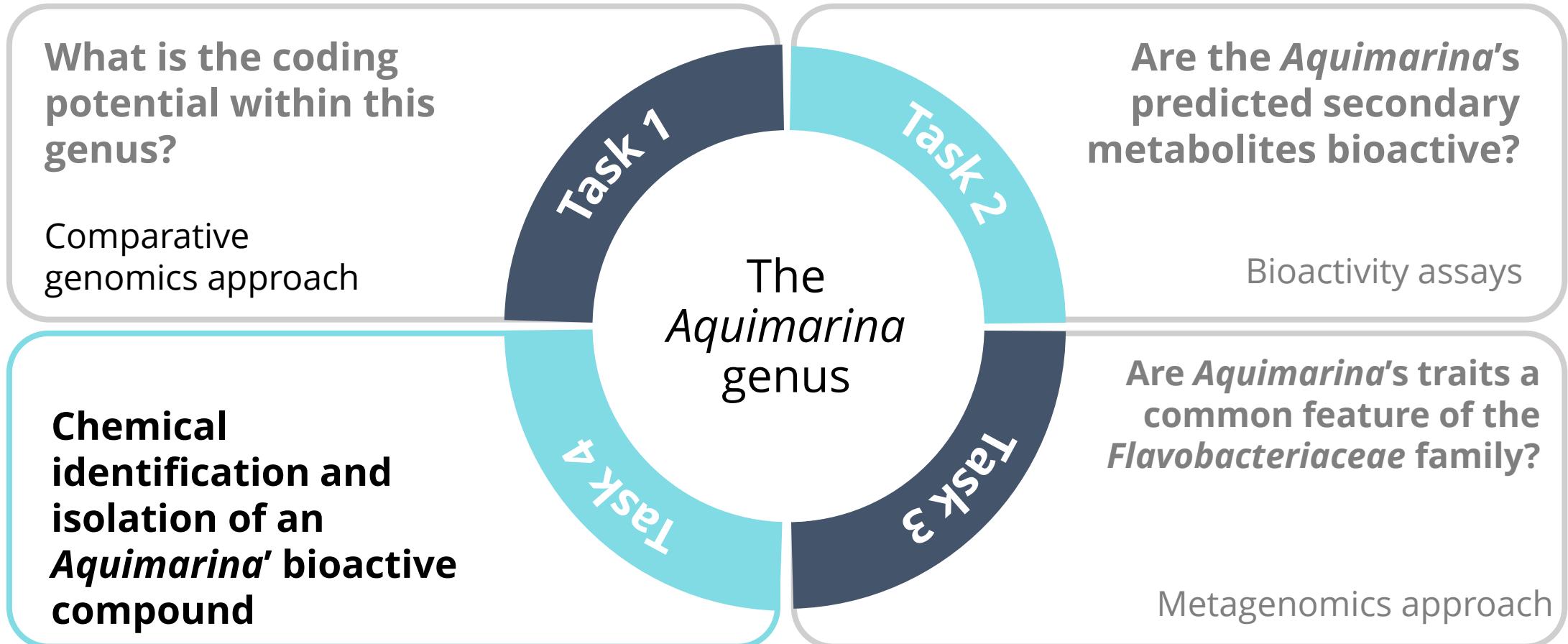
Are *Aquimarina*'s traits a common feature of the *Flavobacteriaceae* family?



Are *Aquimarina*'s traits a common feature of the *Flavobacteriaceae* family?



Next steps

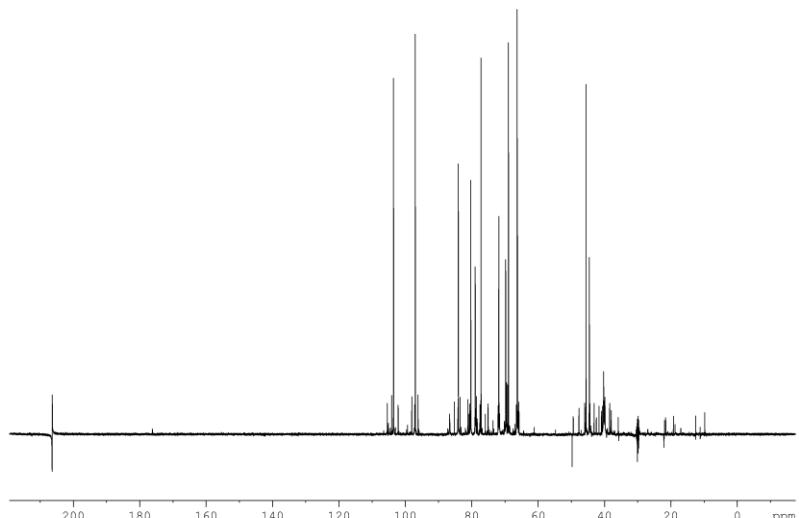


Chemical identification and isolation of an *Aquimarina*' bioactive compound

Among the in-house *Aquimarina* strains available, one will be chosen for further studies.



Goal: isolation and identification of a bioactive compound.



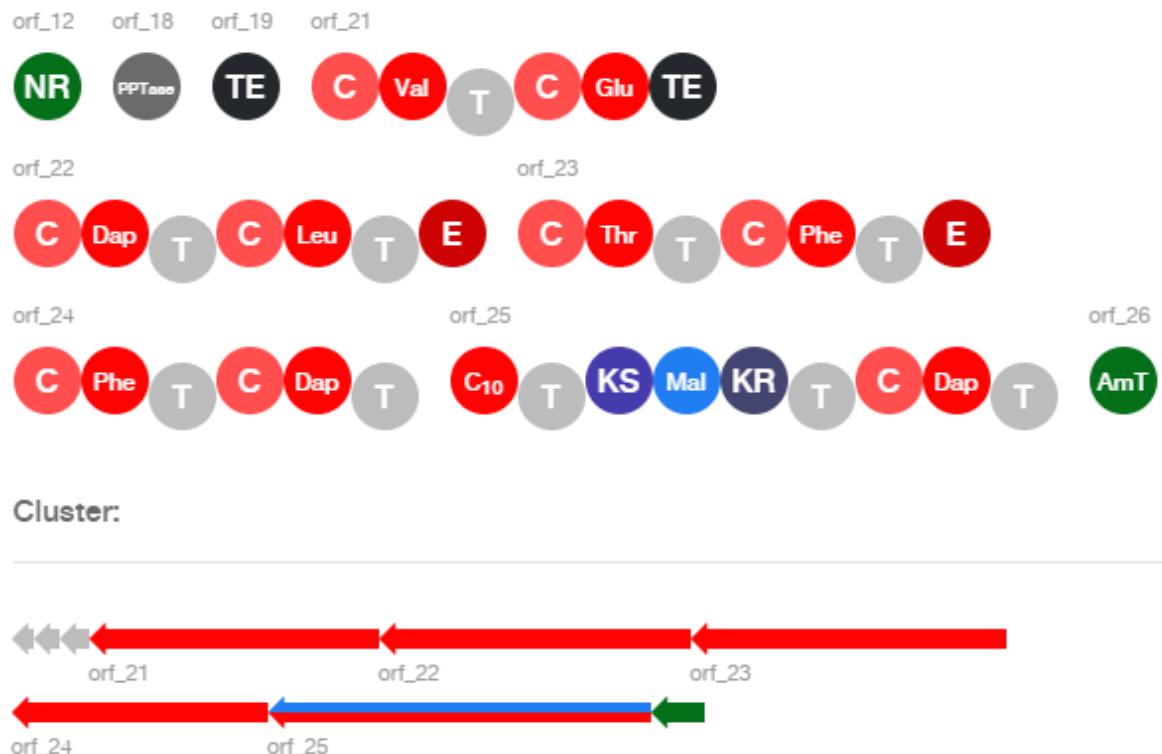
Erythromycin NMR spectra

- Secondary metabolite structure elucidation will be accomplished by:**
1. Bioassay-guided fractionation
 2. High-Resolution Mass Spectrometry

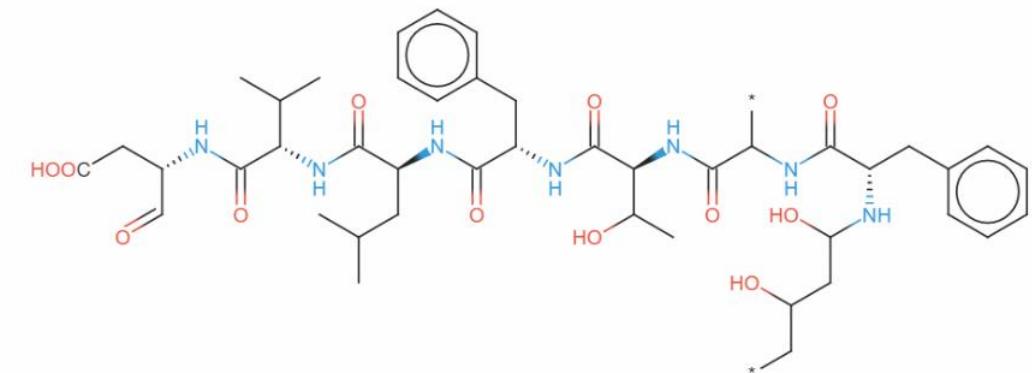
Chemical identification and isolation of an *Aquimarina*' bioactive compound

Genome and bioactivity guided approach

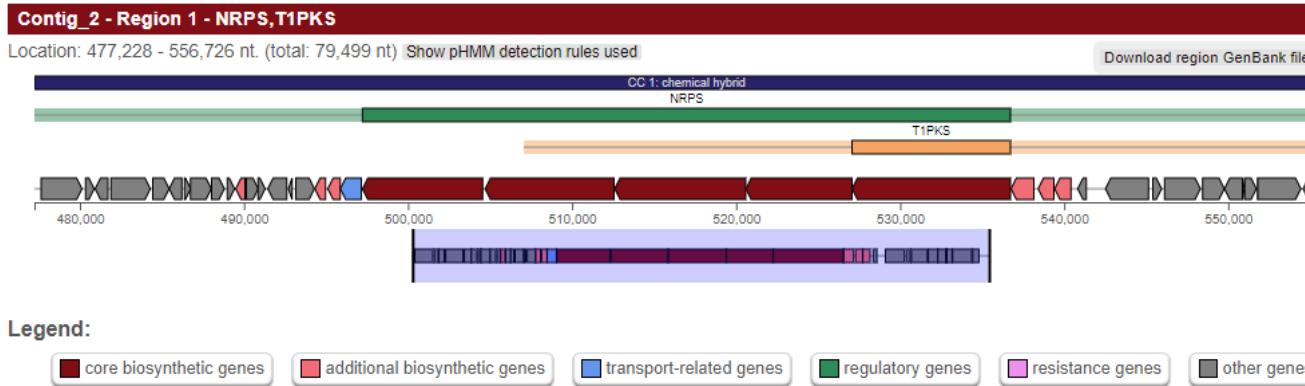
Example: **NRPS/PKS BGC** from *Aquimarina* sp. Aq78



Putative BGC



Chemical identification and isolation of an *Aquimarina*' bioactive compound

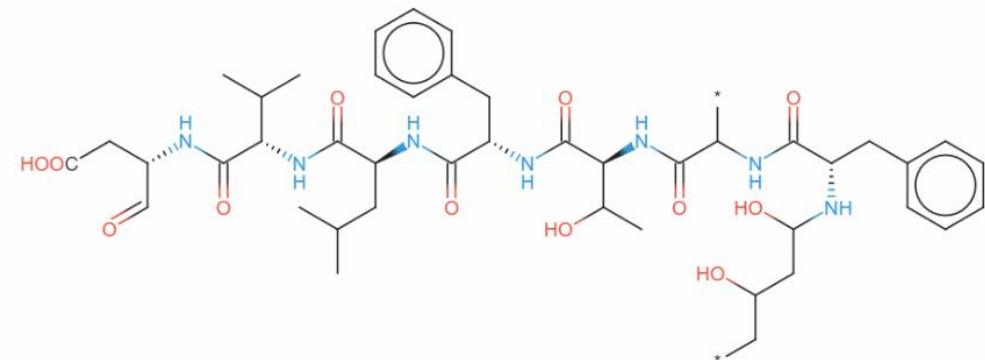


Cluster:



Legend:	
Aminocoumarin	Aminoglycoside
Beta-lactam	Bisindole
CDPS active site	Deoxysugar
Siderophore	Nucleoside
Regulatory	Resistance
Cleaved propeptide	Nonribosomal peptide
Hybrid	Polyketide
Other	Tailoring

Example: **NRPS/PKS BGC**
from *Aquimarina* sp. Aq78



Rough prediction of core scaffold based on assumed PKS/NRPS colinearity; tailoring reactions not taken into account

Polymer prediction:

(pk - ohmal - hydrophilic) + (phe - X) + (thr - phe) +
(hydrophilic - leu) + (val - asp)

Acknowledgments

Special thanks to:

- Prof. Dr. Rodrigo Costa
- Dr. Tina Keller-Costa
- Prof. Dr. Isabel Sá-Correia
- Prof. Dr. Margarida Casal

Collaborators:

- Patrícia Paula
- Matilde Marques
- Dr. Nuno Bernardes
- Dr. Dalila Mil-Homens
- Prof. Dr. Arsénio Fialho
- Prof. Dr. Miguel Teixeira
- Dr. Ulisses Nunes da Rocha (UFZ)
- Prof. Dr. Jörn Piel (ETH Zurich)



MicroEcoEvo team



[DP_AEM]
DOCTORAL PROGRAM IN
APPLIED AND ENVIRONMENTAL
MICROBIOLOGY



Lisb@20²⁰ PORTUGAL 2020



PD/BD/143029/2018
PTDC/MAR-BIO/1547/2014
PTDC/BIA-MIC/31996/2017
Project N.007317
UIDB/04565/2020

Hands-on 3:

Metagenome mining of secondary metabolite
biosynthetic gene clusters (SM-BGCs)



antiSMASH

<https://antismash.secondarymetabolites.org/>

Web-based tool that allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes.

It integrates and cross-links with many in silico secondary metabolite analysis tools and is powered by several open-source tools:

- NCBI BLAST+
- HMMer 3
- Muscle 3
- FastTree
- PySVG
- JQuery SVG.



Created in 2011
Current version: 5.0

antiSMASH - Job submission page

antiSMASH bacterial version 

Submit Bacterial Sequence  Submit Fungal Sequence  Submit Plant Sequence  Download  About  Help  Contact

Server status: **working**

Running jobs: **13**

Queued jobs: **0**

Jobs processed: **428044**

Nucleotide input Results for existing job

Search a genome sequence for secondary metabolite biosynthetic gene clusters

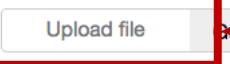
Load sample input Open example output

This is the antiSMASH 5 beta
While we feel it is pretty good already, this version might still be a bit rough at the edges. Until spring 2019, you can still run antiSMASH 4

Notification settings

your@email.com Email address (optional)

Data input

Upload file   

Extra features All off      

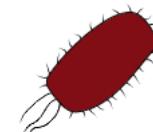
If you want all extra features.

Upload your sequence by using the “Upload file” button and selecting the sequence file (Fasta) to upload.

Enter your email address (optional, but highly recommended: you get an email when your results have been processed).

Submit

Please be considerate in your use of antiSMASH. Help us keep antiSMASH available for everybody by limiting yourself to 5 concurrent jobs. Need to run more? See the [antiSMASH install guide](#) for instructions for getting your own antiSMASH installation.



Now it's your turn!

Practical exercise: submit your metagenome sequences (fasta files) into antiSMASH.

<https://antismash.secondarymetabolites.org/>

anti
SMASH

antiSMASH – The output

The screenshot shows the antiSMASH version 5.0.0 interface. A red box labeled 1 highlights the 'antiSMASH version 5.0.0' text. A red box labeled 2 highlights the '1.13' button in the genomic region selector. A red box labeled 3 highlights the 'NC_003888.3 (Streptomyces coelicolor A3(2))' header. A red box labeled 4 highlights the '27' marker on the genomic map. A red box labeled 5 highlights the 'Region 6' row in the table.

antiSMASH version 5.0.0

Select genomic region:

Overview 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 1.10 1.11 1.12 1.13 1.14 1.15
1.16 1.17 1.18 1.19 1.20 1.21 1.22 1.23 1.24 1.25 1.26 1.27

Download About Help Contact

Identified secondary metabolite regions

NC_003888.3 (Streptomyces coelicolor A3(2))

Genomic map showing regions 1 through 27.

Region	Type	From	To	Most similar known cluster	Similarity
Region 1	hglE-KS ↗, T1PKS ↗	86,637	139,654	Leinamycin ↗	nrps-t1pkstransatpk 2%
Region 2	terpene ↗	166,891	191,654	Isorenieratene ↗	terpene 100%
Region 3	lantipeptide ↗	246,868	270,397		
Region 4	NRPS ↗	494,260	544,087	Coelichelin ↗	NRPS 100%
Region 5	bacteriocin ↗	791,701	799,942	Informatipeptin ↗	lantipeptide 42%
Region 6	T3PKS ↗	1,258,218	1,297,040	Herboxidiene ↗	t1pkst3pk 8%
Region 7	ectoine ↗	1,995,500	2,005,898	Ectoine ↗	other 100%
Region 8	melanin ↗	2,939,306	2,949,875	Istamycin ↗	saccharide 4%

antiSMASH – The output

To return to this page click on "Overview"

each cluster is represented by a circle

Distribution of the BGCs on the chromosome/contig

To get information on the clusters, click on circle or colored cluster number

list of identified clusters

Cluster type

Cluster coordinates

Best hit against MIBiG database of characterized gene clusters

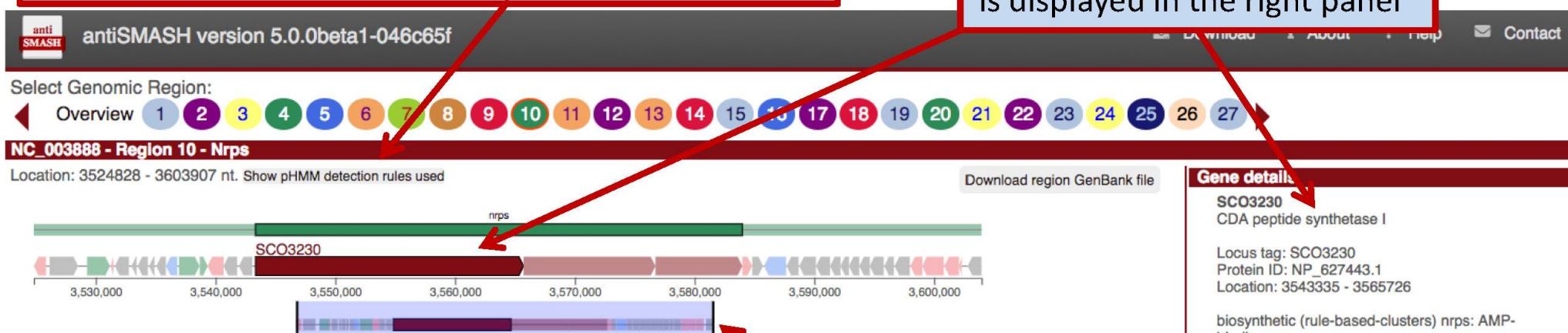
SM class of best MIBiG hit

BGC similarity (be careful with this number!!!)

Region	Type	From	To	Most similar known cluster	Similarity	MIBiG BGC-ID
Region 1	Otherks-T1pk	86640	139467	Leinamycin	hybrid	2% BGC0001101
Region 2	Terpene	166891	191654	Isorenieratene	terpene	100% BGC0000664
Region 3	Lanthipeptide	246868	270397			
Region 4	Nrps	494260	544087	Coelichelin	NRPS	100% BGC0000325
Region 5	Bacteriocin	791701	799942	Informatipeptin	RiPP	42% BGC0000518
Region 6	T3pk	1258218	1297040	Herboxin		65
Region 7	Ectoine	1995500	2005398	Ectoine		53
Region 8	Melanin	2939306	2949875	Melanin		10
Region 9	Siderophore	3034632	3045603	Desferrin		40
Region 10	Nrps	3524828	3603907	Calcum		15
Region 11	T2pk	5496474	5567376	Actinomycetin		94
Region 12	Terpene	5671275	5691836	Albaflavenone	terpene	100% BGC0000660
Region 13	T2pk	5751945	5824487	Spore pigment	polyketide	66% BGC0000271
Region 14	Siderophore	6336091	6346368			
Region 15	Nrpsfragment-T1pk	6430010	6475291	Undecylprodigiosin	hybrid	100% BGC0001063
Region 16	Bacteriocin	6632343	6643659			

antiSMASH – The output

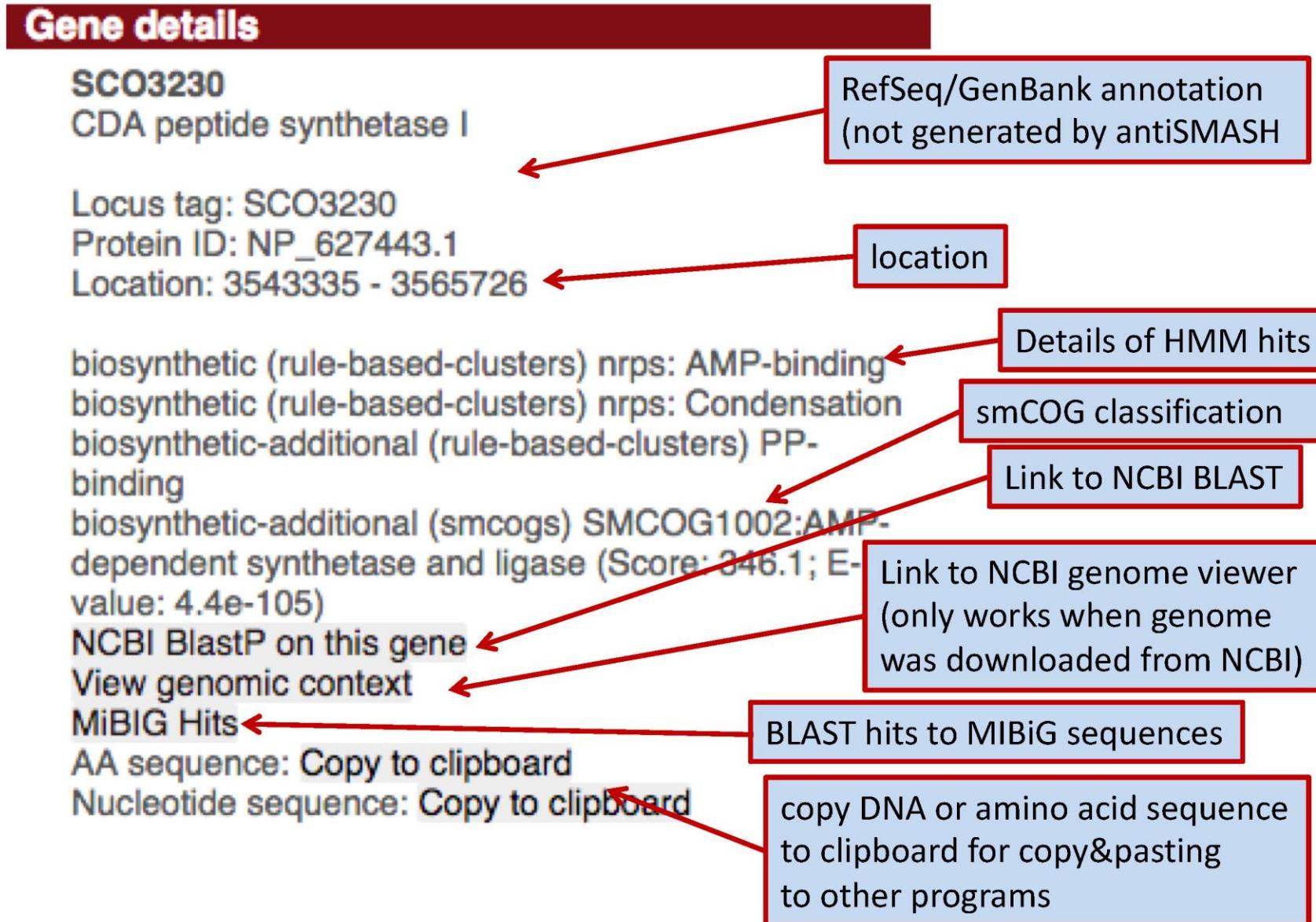
To get information on the rule that antiSMASH used to identify the genetic region as a secondary metabolite biosynthetic gene cluster, click here



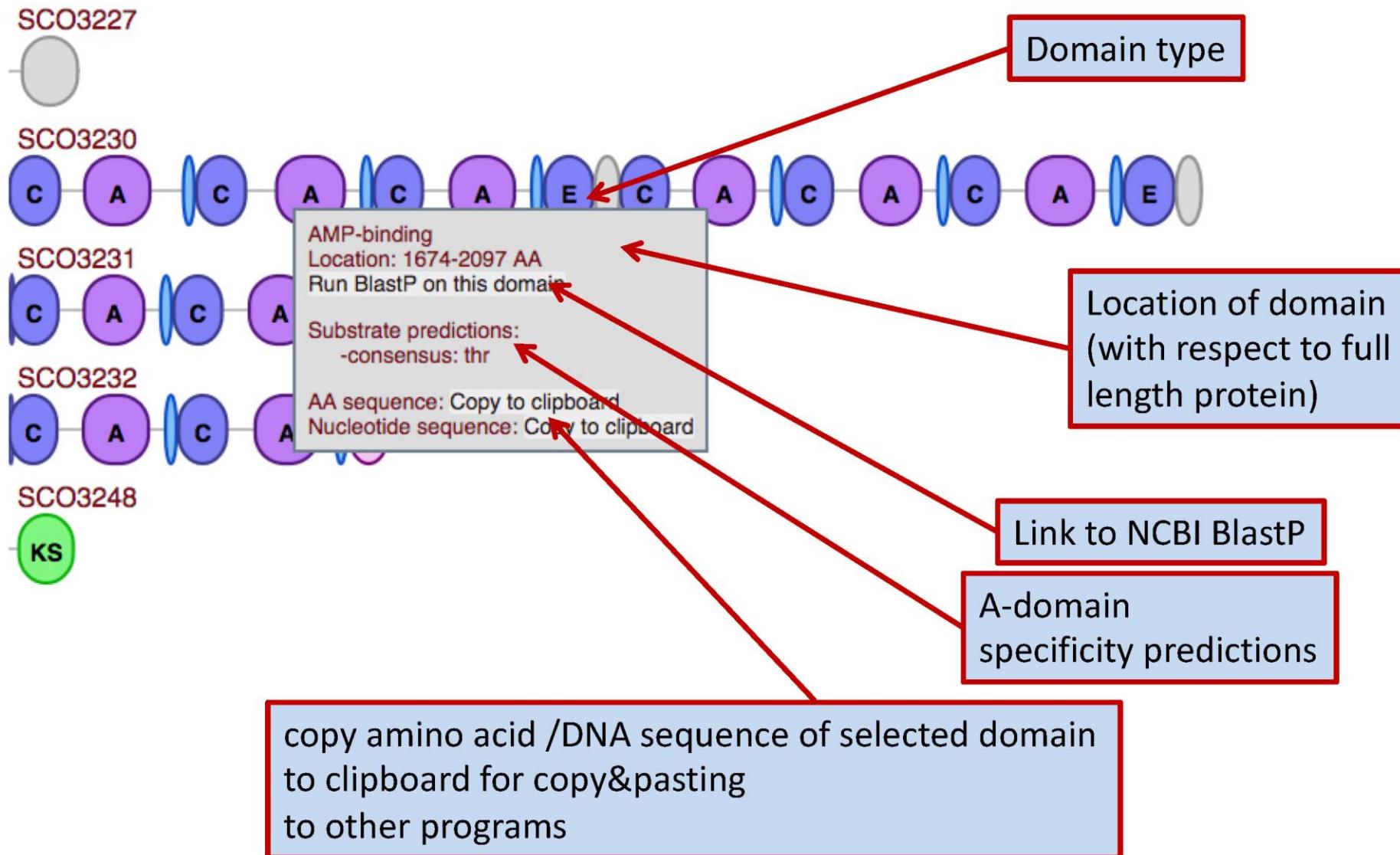
To get information on a specific gene of the cluster, click on the gene arrows; info is displayed in the right panel

Zoom to region of interest by moving the bars or using the buttons

antiSMASH – The output



antiSMASH – The output

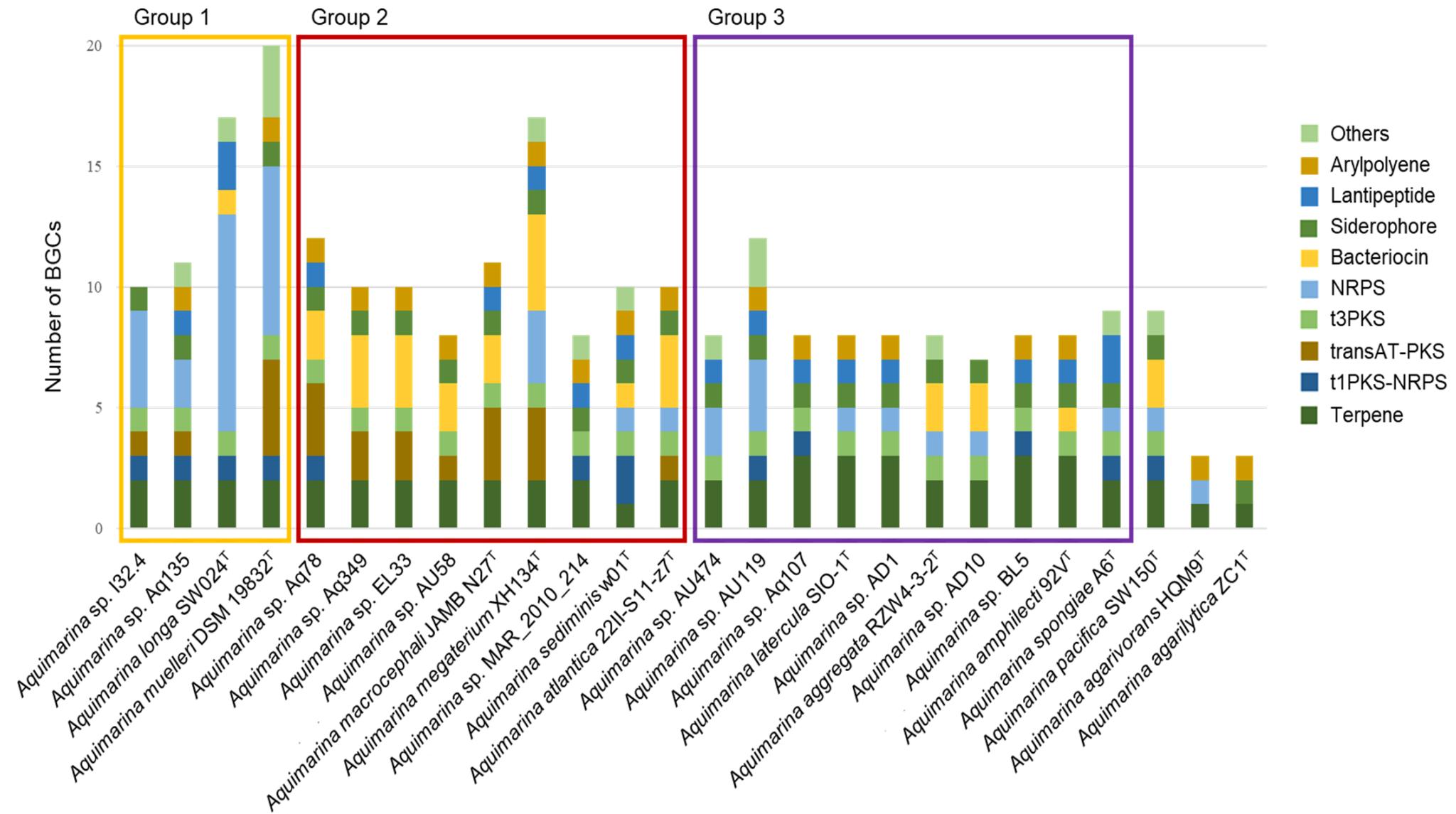


antiSMASH – The output

The screenshot shows the antiSMASH web interface. At the top, it displays "antiSMASH version 5.0.0beta1-046c65f". Below this, a navigation bar includes links for "Overview", "About", "Help", and "Contact". A "Select Genomic Region" section features a circular menu with numbered options from 1 to 25, where option 10 is highlighted in purple. The main content area is titled "NC_003888 - Region 10 - Nrps" and shows a genomic map with a green bar labeled "nrps". On the right side, a "Download" button is expanded to show three download options: "Download all results", "Download GenBank summary file", and "Download log file". A red arrow points to the "Download" button. Below the download menu, specific details for locus tag SCO3230 are listed: "SCO3230 CDA peptide synthetase I" and "Locus tag: SCO3230".

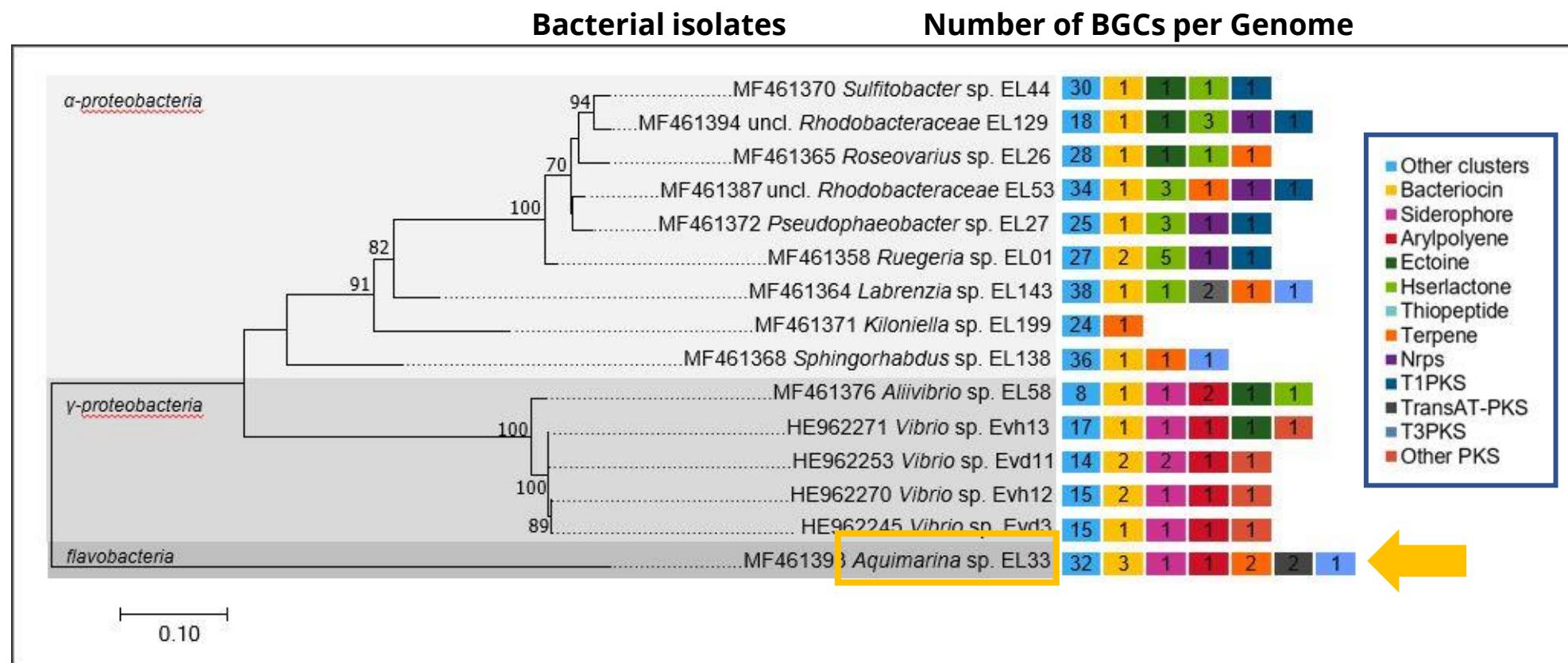
Download button

Result obtained for the *Aquimarina* study:



Another example:

Potential for Secondary Metabolite Synthesis in Soft Coral-Associated Bacteria



440 biosynthetic gene clusters (BGCs) on the genomes of 15 bacterial associates (12 genera) isolated from the soft corals *Eunicella labiata* and *Eunicella verrucosa*.

BiG-SCAPE

Biosynthetic Gene Similarity Clustering and Prospecting Engine

BiG-SCAPE is a tool that **calculates distances between BGCs** in order to map the BGC diversity onto sequence similarity networks, which are then processed for automated reconstruction of **Gene Cluster Families**



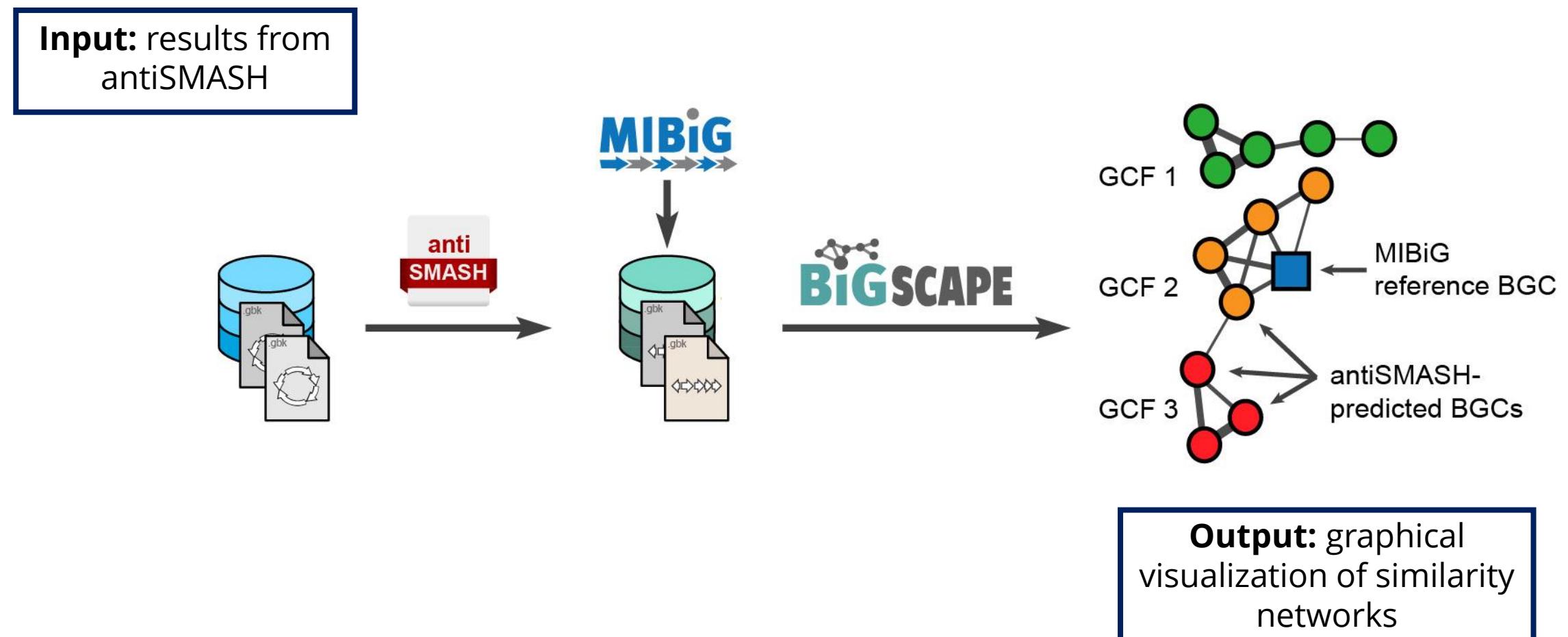
Groups of gene clusters that encode biosynthesis of highly similar or identical molecules.

BiG-SCAPE's interactive visualizations of these similarity networks allows effective exploration of the diversity of BGCs, linking them to knowledge from reference data within the **MiBIG repository**



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

The BiG-SCAPE workflow uses sequence similarity networking to group biosynthetic gene clusters into families



BiG-SCAPE – The output

BiG-SCAPE Biosynthetic Genes Similarity Clustering and Prospecting Engine Version 0.0.0r

Networks: [Overview](#) [Saccharides](#) [NRPS](#) [Others](#) [RiPPs](#)

Runs: 2018-08-07_18-46-29_hybrids_glocal

Run Information

Analysis Started: 07/08/2018 18:46:29
Parameters: -i /home/input/gbks -o /home/output/bigscape_salida
Analysis Completed: 07/08/2018 18:49:52 (0h3m23s)

Input Data

Total Number of Genomes: 15
Total BGCs: 23

BGC per Genome

BGC per Class

Network Overview

Saccharides	NRPS	Others	RiPPs
Number of families:	7		
Average number of BGCs per family:	1		
Max number of BGCs in a family:	3		
Families with MiBiG Reference BGCs:	0		

GCF absence/presence heatmap

Cluster GCF based on: [Genomes Absence/Presence](#) ▾
Cluster Genomes based on: [Family Absence/Presence](#) ▾

Show: 20 ▾ largest GCFs

Download: [Absence/Presence table \(tsv\)](#)

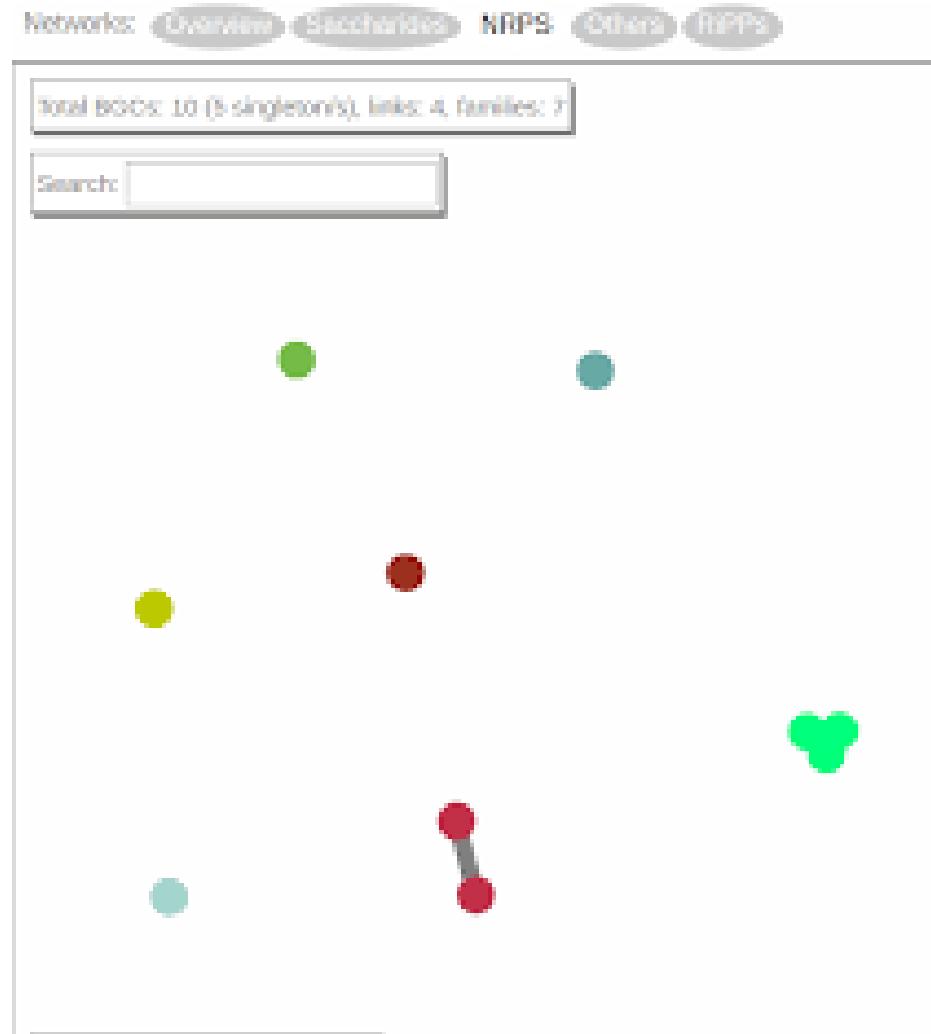
If you have found BiG-SCAPE useful, please cite us.

WAGENINGEN UR
For quality of life

WARWICK

DTU
crb
The Novo Nordisk Foundation
Center for Biosustainability

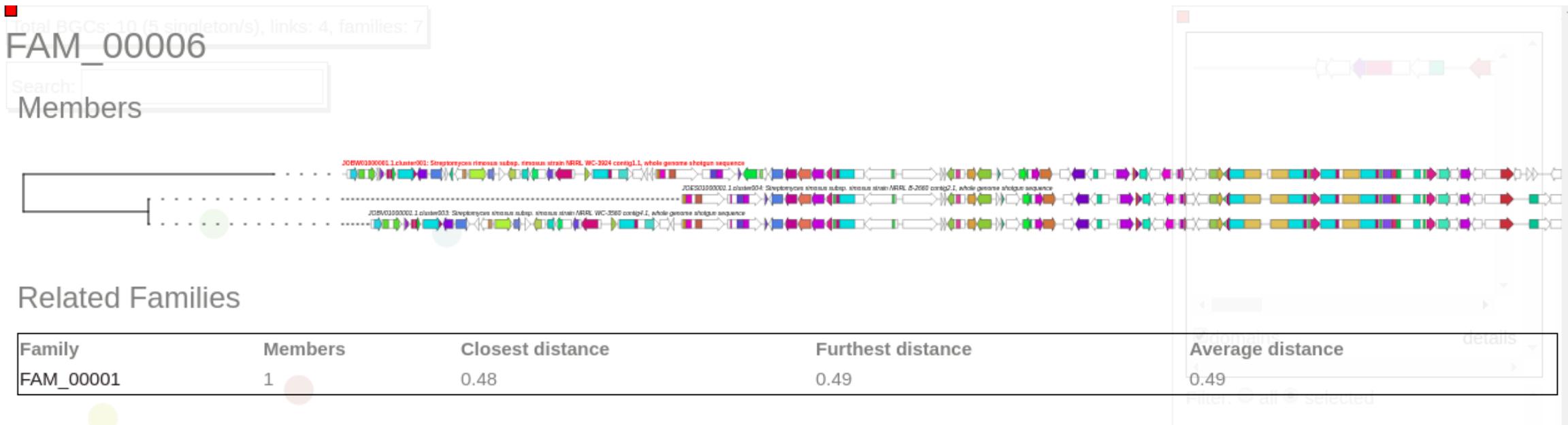
BiG-SCAPE – The output



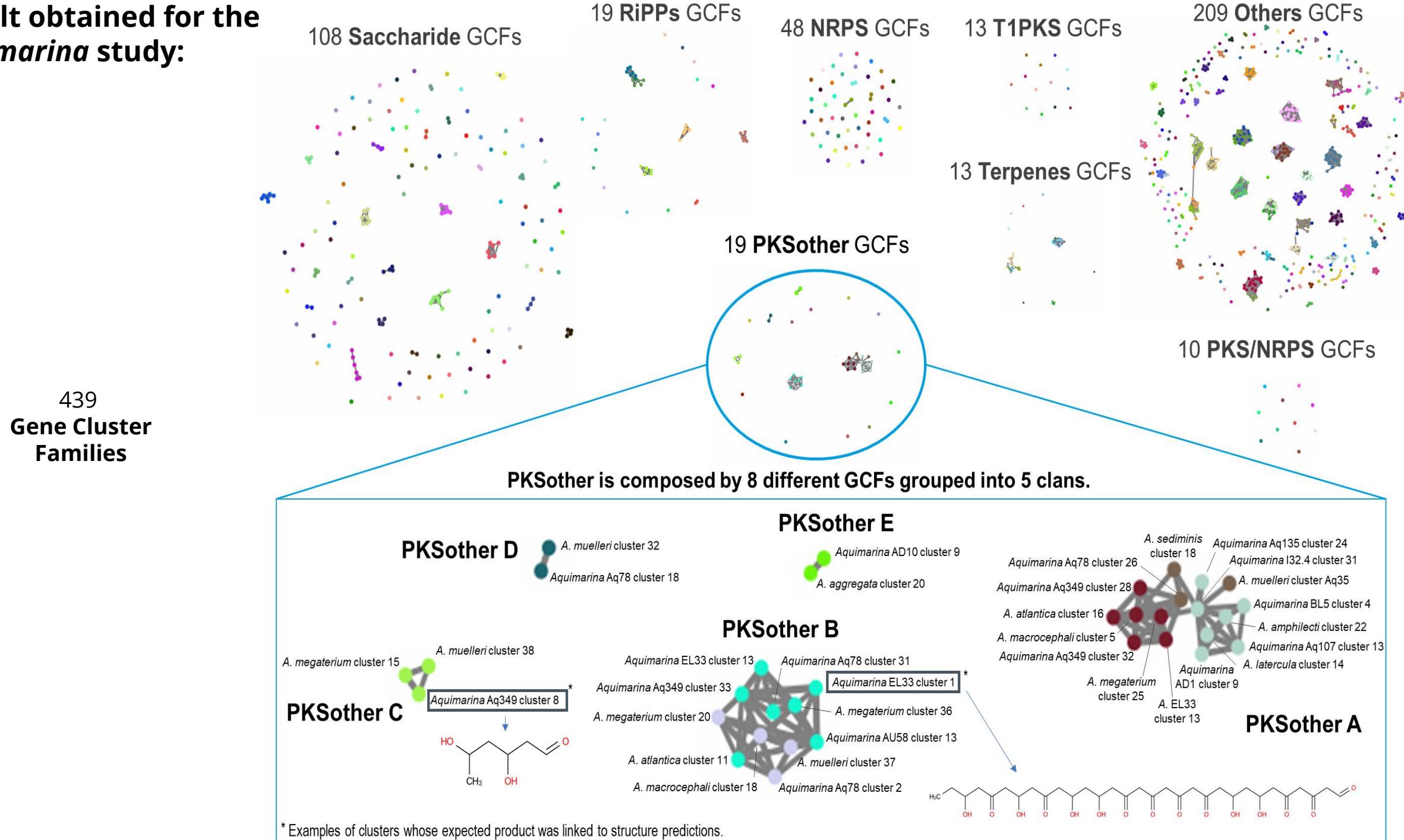
The distances for each cutoff value will be used to automatically define '**Gene Cluster Families**' (**GCFs**) for each compound class.

BiG-SCAPE – The output

Gene Cluster Family (GCF) example:



Result obtained for the *Aquimaria* study:



What you'll need to apply the BiG-SCAPE workflow to your antiSMASH results:

- 1)** Once the antiSMASH run is finished, you'll receive an email. Open the link provided.
- 2)** On the top of "overview" page", you will find the option "Download all results". Click on it. A zip folder will be downloaded. Unzip it.
- 3)** On each antiSMASH results folder you will find several different files in different formats. For the BiG-SCAPE workflow you will need the GenBank (.gbk) files corresponding to each BGC identified.

For later identification, the filename of each gbk file must be renamed so that you later know from which genome each BGC came from.

Example of a suitable filename : **Aq_Aq78_contig_1.region001.gbk**
Make sure that you don't have any spaces in the filename.

- 4)** Move all the .gbk files into a single folder. Zip the folder to make the file transfer easier.
- 5)** Send this zipped folder to: sandragodinhosilva@tecnico.ulisboa.pt
I'll run the BiG-SCAPE pipeline and return the results to you as soon as possible.



If you want to run BiG-SCAPE on your own:

- 1)** Unfortunately, this workflow needs to be run on a Linux operating system. As most of us have a Windows operating system on our computers, this might be the major difficulty.

- 2)** If that isn't a problem for you, you can try to install BiG-SCAPE. All the instructions to do so are available in the following link:
<https://git.wageningenur.nl/medema-group/BiG-SCAPE/-/wikis/installation>

- 3)** Please talk with me and I'll be happy to help you on this process.



Thank you for your attention.

Sandra Godinho Silva

sandragodinhosilva@tecnico.ulisboa.pt