# Secondary metabolites: paths to discovery

**Sandra Godinho Silva**

sandragodinhosilva@tecnico.ulisboa.pt

**28 April 2021**

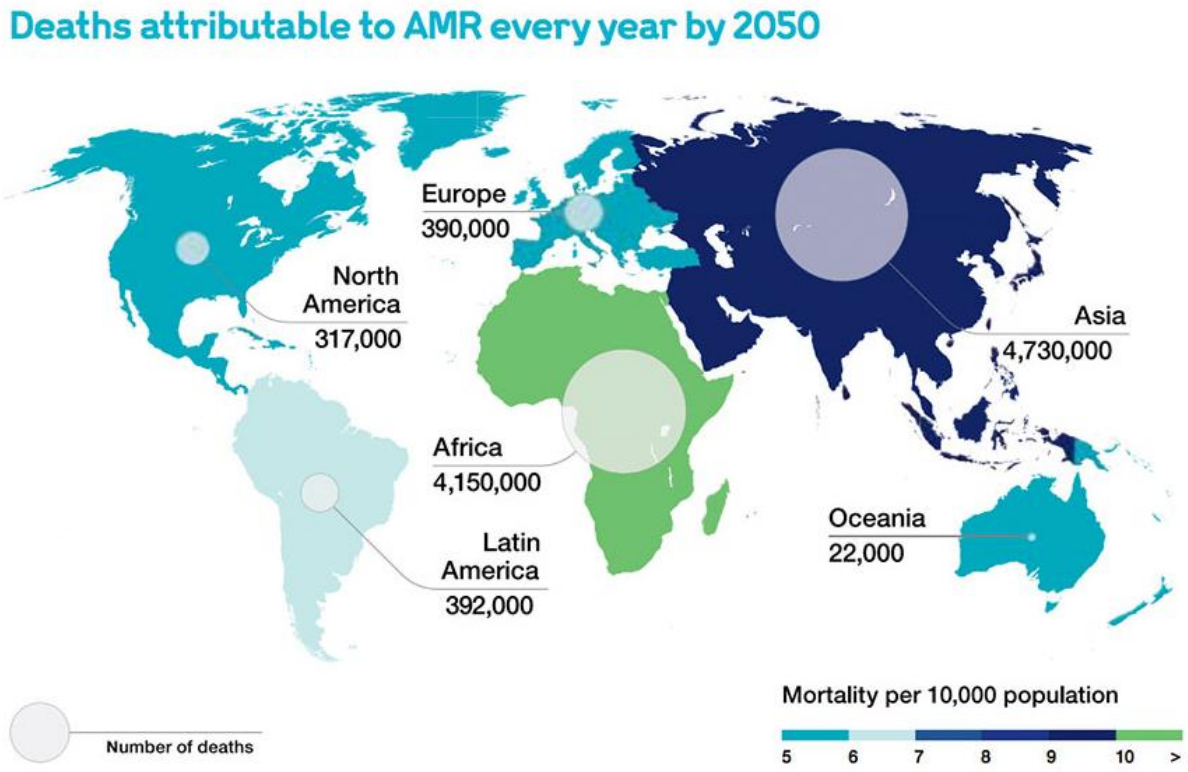
# Natural Products Discovery

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**

Harvey, A. L. 2008. “Natural products in drug discovery.” *Drug Discovery Today*, 13(19–20), 894–901.

# Natural Products Discovery




- **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.

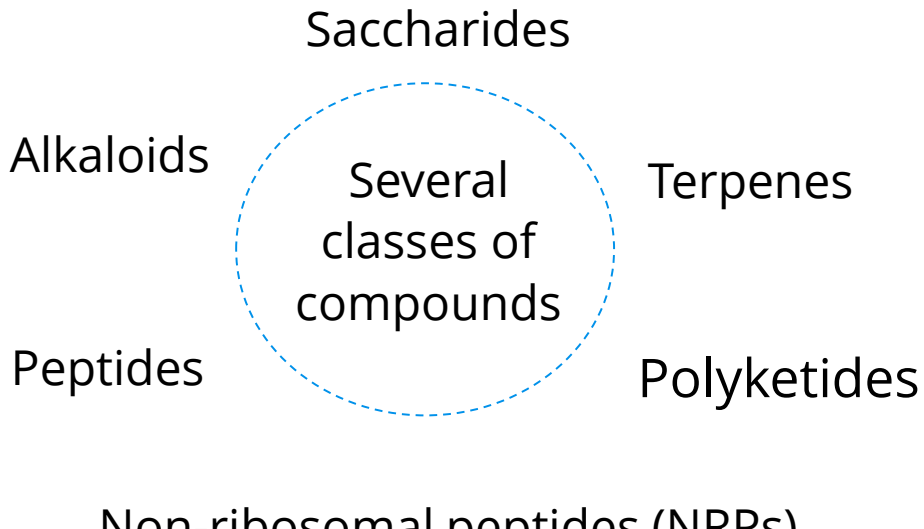
Source: Antimicrobial Resistance: *Tackling a Crisis for the Health and Wealth of Nations* (2014)

The need for new therapeutical drugs is real

# Natural Products Discovery

- Small organic molecules produced by living organisms;

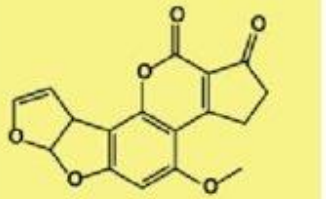
- Normally are secondary metabolites:
  - Not essential for growth and reproduction;
  - Provide **survival advantage.**




Harvey, A. L. 2008. “Natural products in drug discovery.” *Drug Discovery Today*, 13(19–20), 894–901.

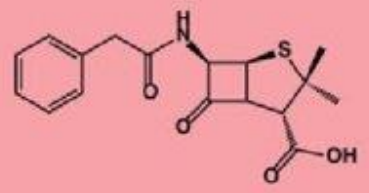
# Secondary metabolites – classes:

polyketide

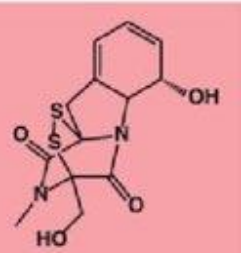


aflatoxin B1

non-ribosomal peptides

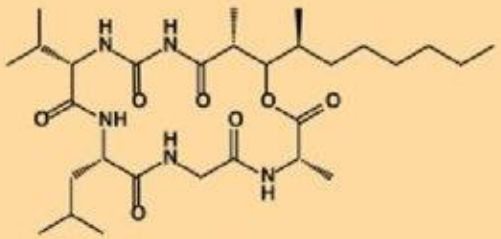


penicillin G

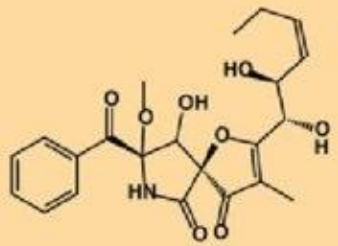


gliotoxin

polyketide/non-ribosomal peptide hybrids

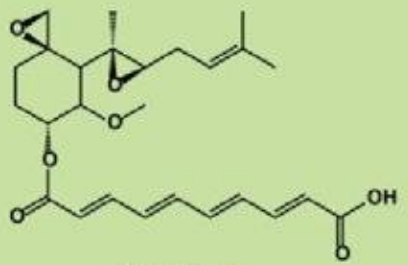


emericellamide A

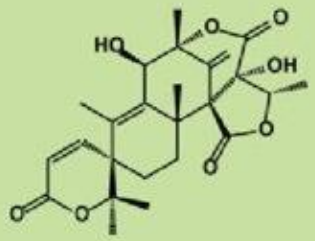


pseurotin A

meroterpenoids

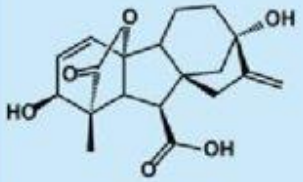


fumagillin



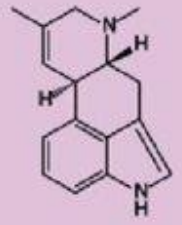
austinol

terpene



gibberellin A3

prenylated tryptophan derivative



agroclavine

non-canonical

norloline

Wieman & Keller 2013. ”Strategies for mining fungal natural products” *Journal of Industrial Microbiology* 41(2)

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:

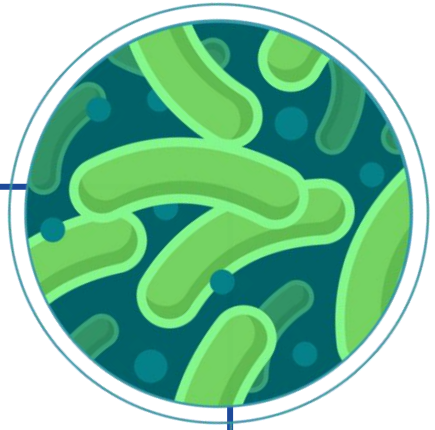
- Biosynthetic enzymes;
- Resistance enzymes;
- Enzymes to produce unusual building blocks;
- Regulatory machinery.

Cummings et al. 2014. “Steps towards the Synthetic Biology of Polyketide Biosynthesis.” *FEMS Microbiology Letters* 351(2): 116–25.

# 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

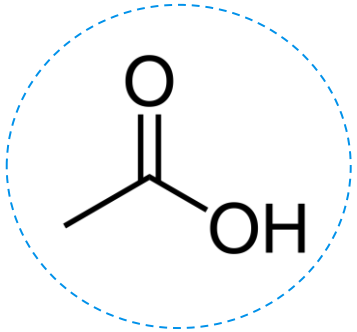
BGC diversification mechanisms

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

Jensen, Paul R. 2016. “Natural Products and the Gene Cluster Revolution.” *Trends in Microbiology* 24(12): 968–77

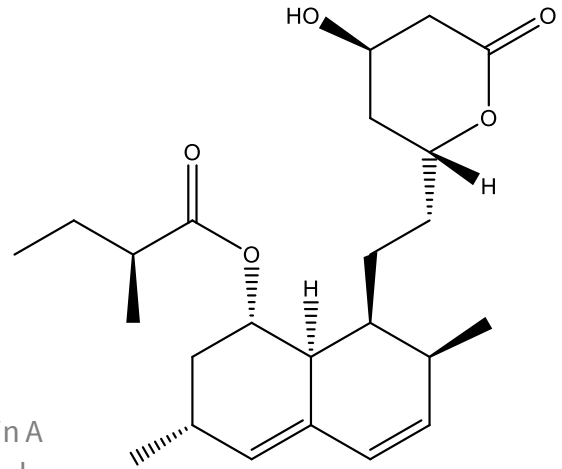
# Polyketides

- 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**

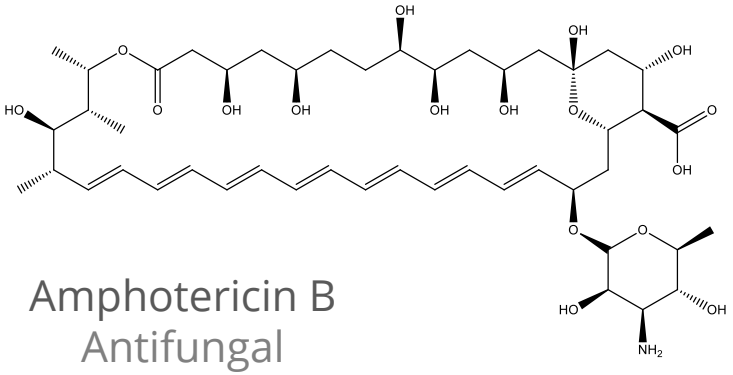
Most common:
- Malonyl-CoA
- Methylmanoyl-CoA



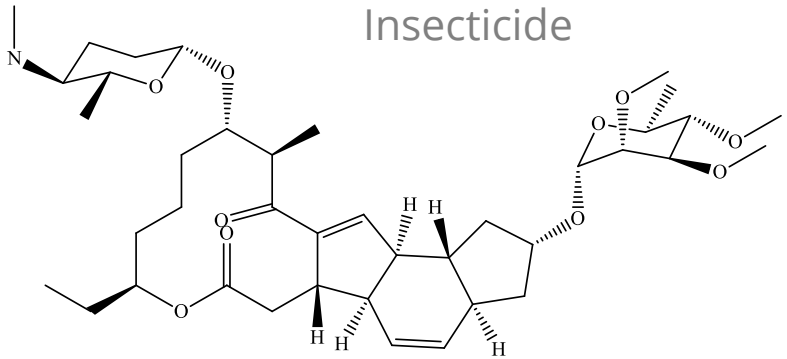

Lovastatin A
Cholesterol-lowering drug

# Famous Polyketides

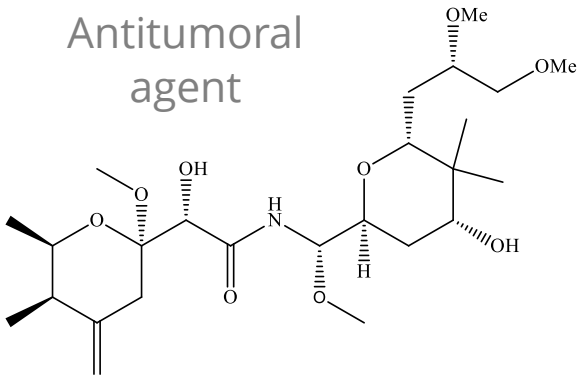
**Important therapeutic drugs**



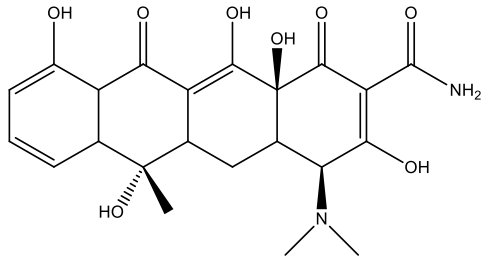
Amphotericin B
Antifungal



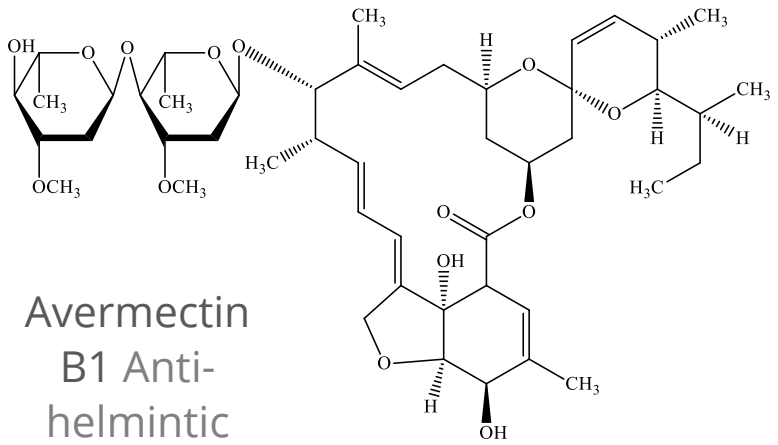
Spinosyn A
Insecticide



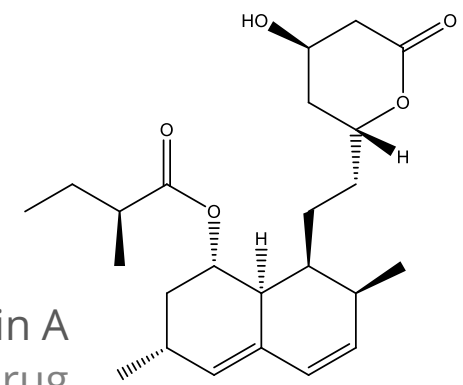
Pederin
Antitumoral agent



Tetracycline
Antibiotic

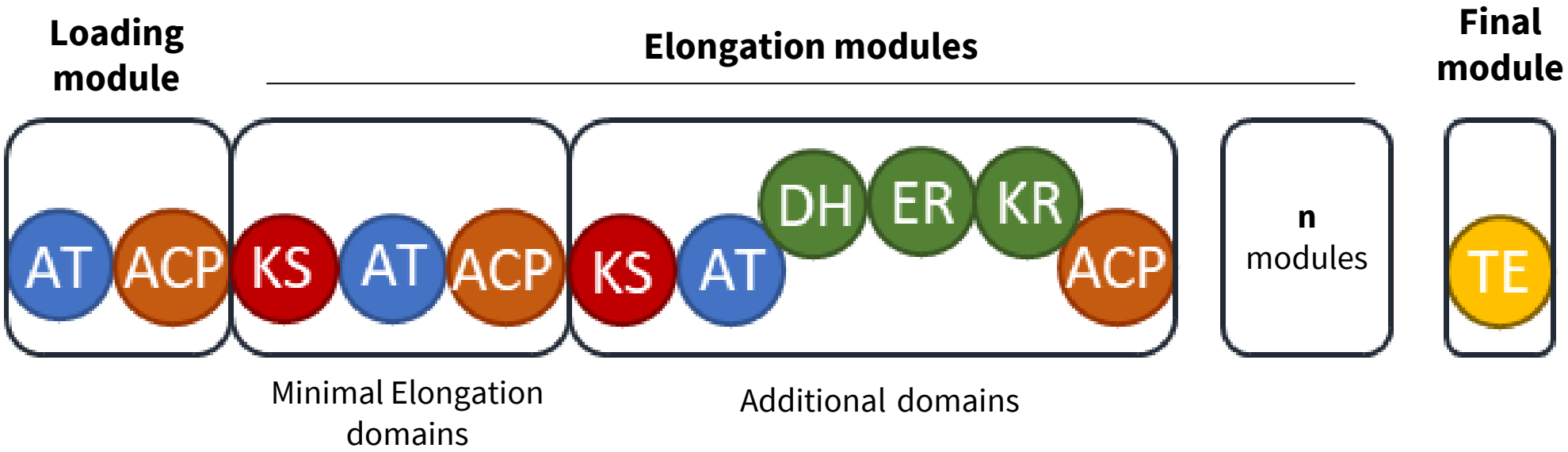


Avermectin B1 Anti-helmintic



Lovastatin A
Cholesterol-lowering drug

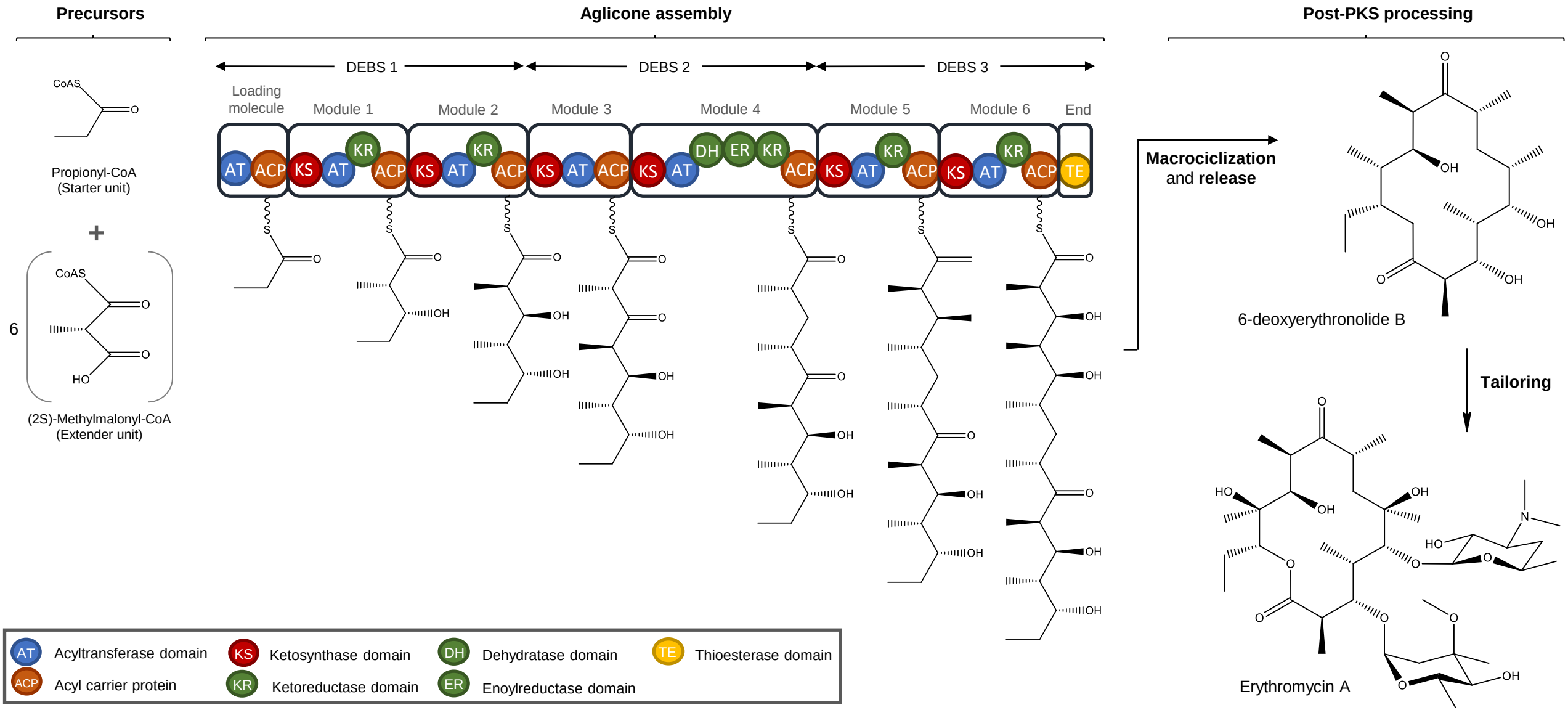
# General PKS composition





Weissman, Kira J., and Peter F. Leadlay. 2005. "Combinatorial Biosynthesis of Reduced Polyketides." *Nature Reviews Microbiology* 3(12): 925–36

# “DEBS”: the prototype of type I PKS



# PKSs classification

Based on
**enzyme architecture:**

Type I
Type II
Type III

PKSs

Based on
**domain organization**:

Iterative
Modular (only type I)

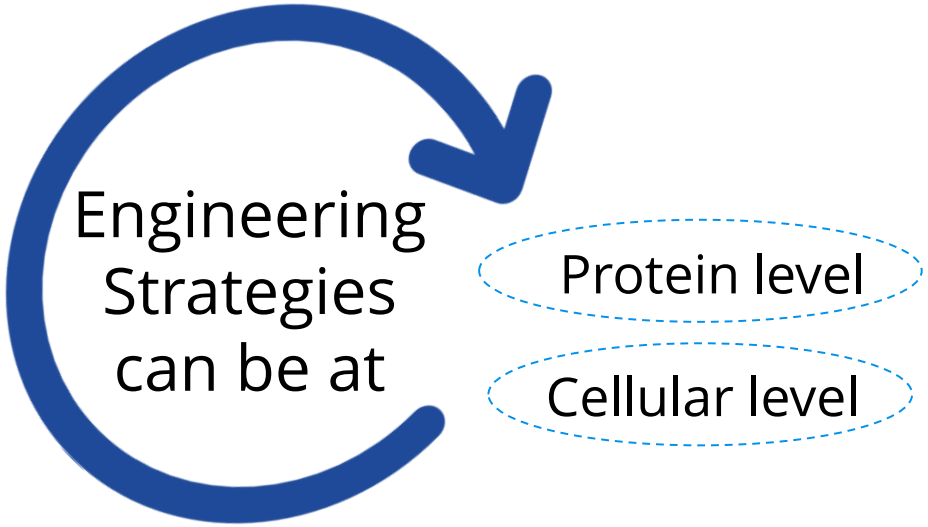
There are also diverse **PKS-NRPS hybrids** worth to mention.

# 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.

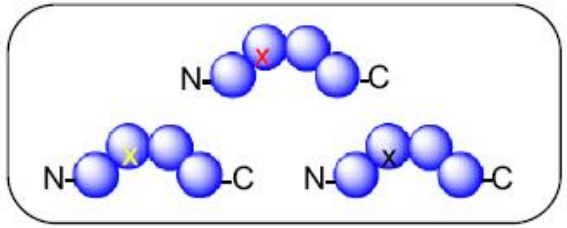



Barajas, Jesus F. et al. 2017. “Engineered Polyketides: Synergy between Protein and Host Level Engineering.” *Synthetic and Systems Biotechnology* 2(3): 147–66.

# PKS protein modifications

Engineering Strategies

Protein level




AT, KR mutagenesis

Reductive loop swaps

Possible strategies

Domain fusion

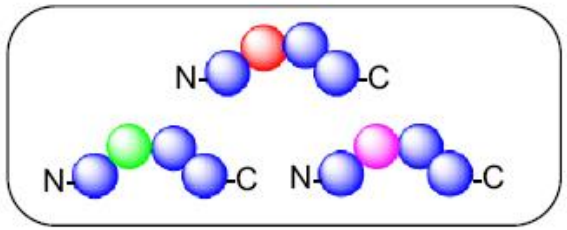
Modification of active sites

Different substrate specificity or higher substrate promiscuity.

Change of pool of building blocks.

Domain swapping

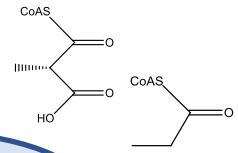
Entire modules or **single domains** (TE, AT).




Barajas, Jesus F. et al. 2017. "Engineered Polyketides: Synergy between Protein and Host Level Engineering." *Synthetic and Systems Biotechnology* 2(3): 147–66.

Sun, Huihua, Zihe Liu, Huimin Zhao, and Ee Lui Ang. 2015. "Recent Advances in Combinatorial Biosynthesis for Drug Discovery." *Drug Design*, Development and Therapy 9: 823–33.

# Polyketide diversification




Pool of starter units

Pool of extender units

**PKS**

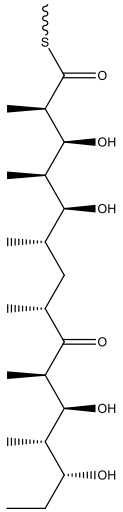
Polyketide backbone

Choice of building blocks
Chain length

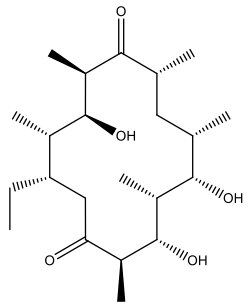
Domain composition

β-keto processing
Stereochemistry
Alkylation/branching

Primary core structure




1st tailoring



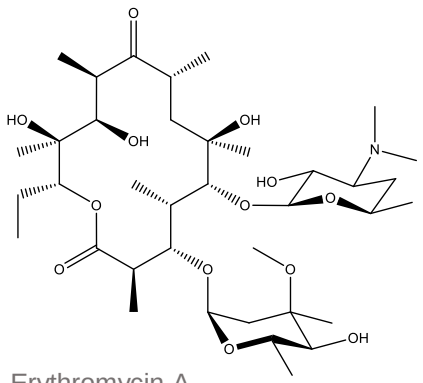

Secondary core structure

Rearrangements
2nd cyclization

2nd tailoring

Oxygenation
Halogenation
Alkylation
Glycosylation

Final polyketide metabolite




Erythromycin A

Combinatorial biosynthesis

Hertweck, Christian. 2009. "The Biosynthetic Logic of Polyketide Diversity." *Angewandte Chemie - International Edition* 48(26): 4688–4716.

# Heterologous expression of BGCs

Engineering Strategies

Cellular level

Strategy to produce compounds

→ Too complex to be chemically synthesized.

→ Produced by complex, slow-growing microorganisms.

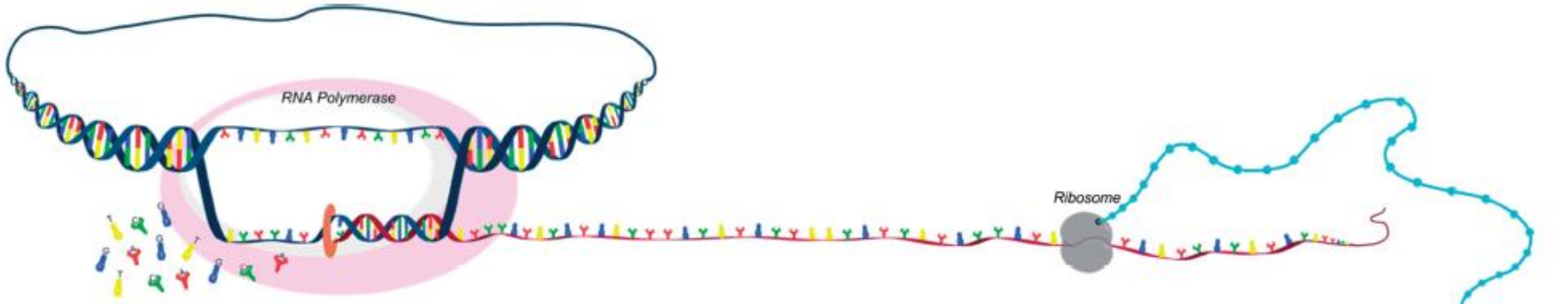
- Allow the expression of cryptic (silent) BGCs.
- Overproduction of target compounds.

Possible hosts:

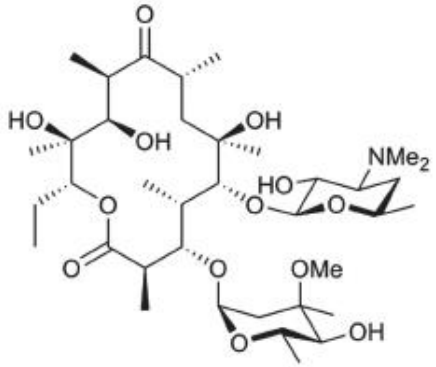
- *Streptomyces*
- Myxobacteria
- *Escherichia coli*
- *Saccharomyces cerevisiae*

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

# Challenges in Heterologous BGC expression

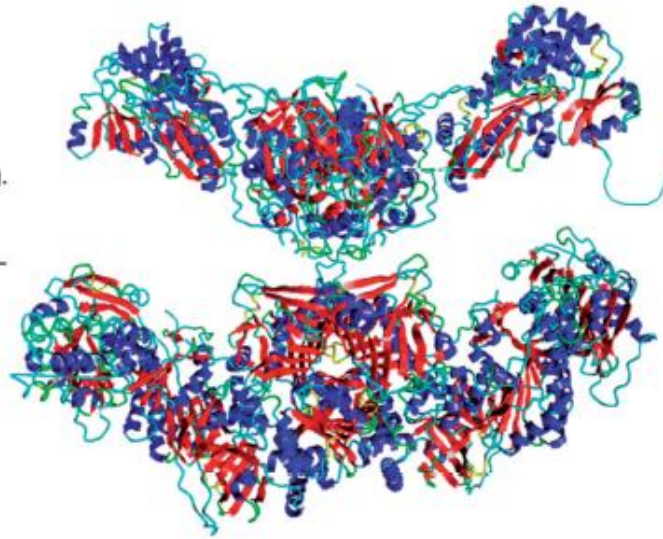



**Problem**: Transcription can be tightly regulated.
**Solutions**: **Promoter replacement**
**Transcription from native promoters**




Erythromycin A

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



**Problem**: Produced compounds can kill host.
**Solution**: Co-expression of resistance pathway/Sensitive host has not prevented mg/L of product formation.

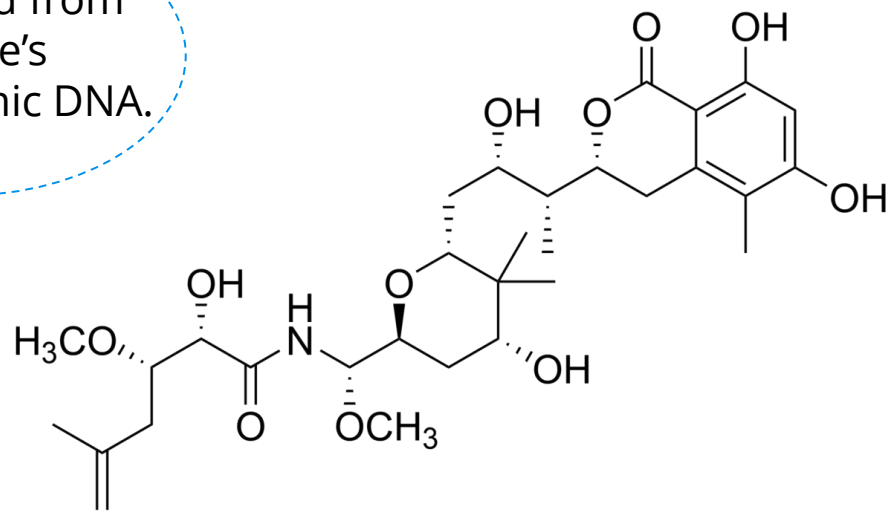
**Problem**: Proteins must fold and ACP must be phosphopantetheinylated.
**Solution**: Chaperones aid folding and phosphopantetheinyl transferases are added to hosts.

Stevens, D. Cole, Taylor P A Hari, and Christopher N. Boddy. 2013. "The Role of Transcription in Heterologous Expression of Polyketides in Bacterial Hosts." *Natural Product Reports* 30(11): 1391–1411.

# Discovery methods of new BGCs

Discovered from sponge's metagenomic DNA.

Psymberin




Metagenomics

Powerful tool for the discovery of novel PKSs, particularly from **uncultivable bacteria**.


Fisch, Katja M. et al. 2009. "Polyketide Assembly Lines of Uncultivated Sponge Symbionts from Structure-Based Gene Targeting." *Nature Chemical Biology* 5(7): 494–501

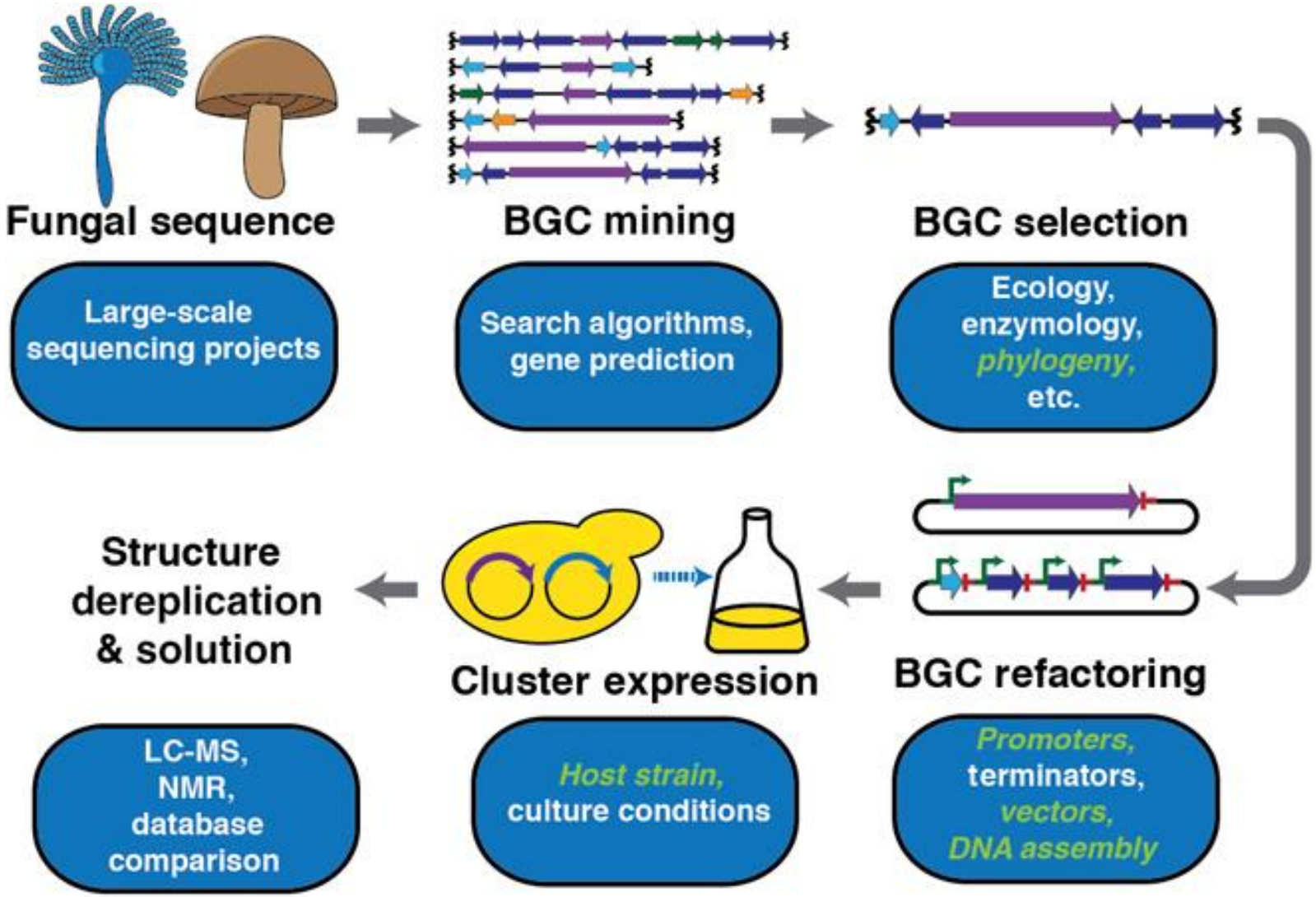
# Discovery methods of new BGCs

The Genomics Era → Has led to a shift in Natural Products Research.

Genome Mining → In silico discovery of novel BGCs.

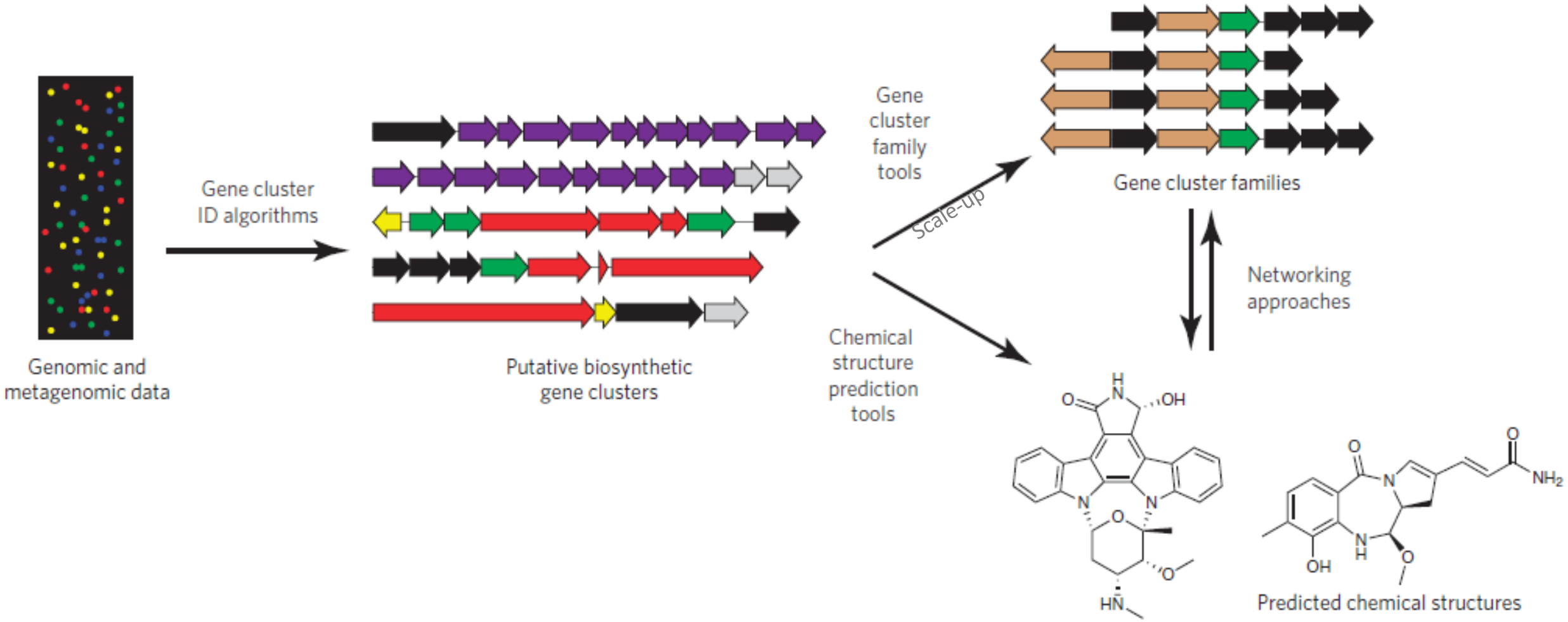
Genome Mining → Through several different informatic algorithms and databases.

# Discovery methods of new BGCs



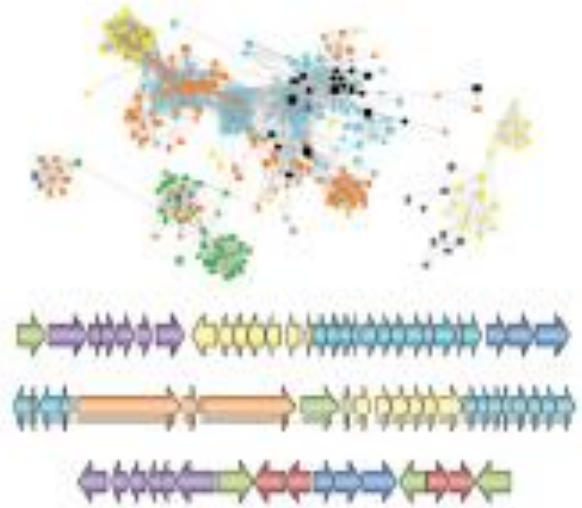

B Harvey CJ, Tang M, Schlecht U, Horecka J, Fischer CR, Lin H-C, et al. 2018 "HEx: A heterologous expression platform for the discovery of fungal natural products." *Sci Adv.*(4)

# Computational Approaches in Natural Products Discovery



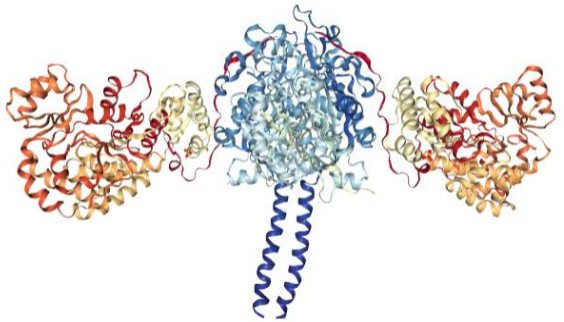

Medema, Marnix H., and Michael A. Fischbach. 2015. "Computational Approaches to Natural Product Discovery." *Nature Chemical Biology* 11(9): 639–48.

# The ultimate goal: an integrative approach



Genome Mining

Structural biology

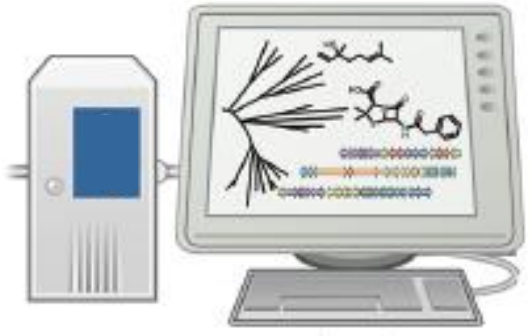


KS-AT didomain of module 5 from DEBS

Integrative approach

Molecular genetics

Metabolomics



Computational Biology

Synthetic biology

Metagenomics

Medema, Marnix H. 2018. “Computational Genomics of Specialized Metabolism: From Natural Product Discovery to Microbiome Ecology.” *mSystems* 3(2): e00182-17.

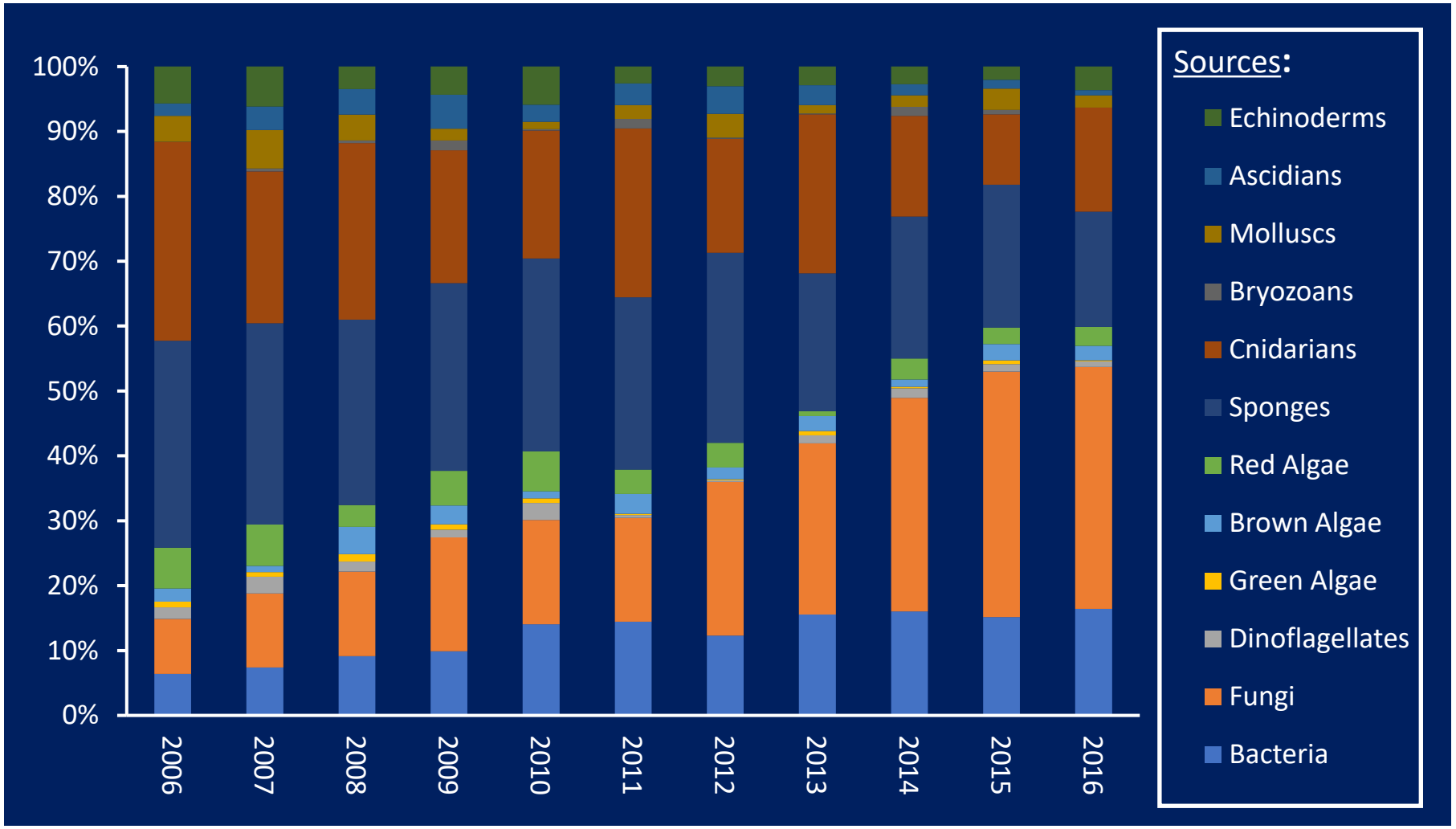
**Case-study:**

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

# The marine environment

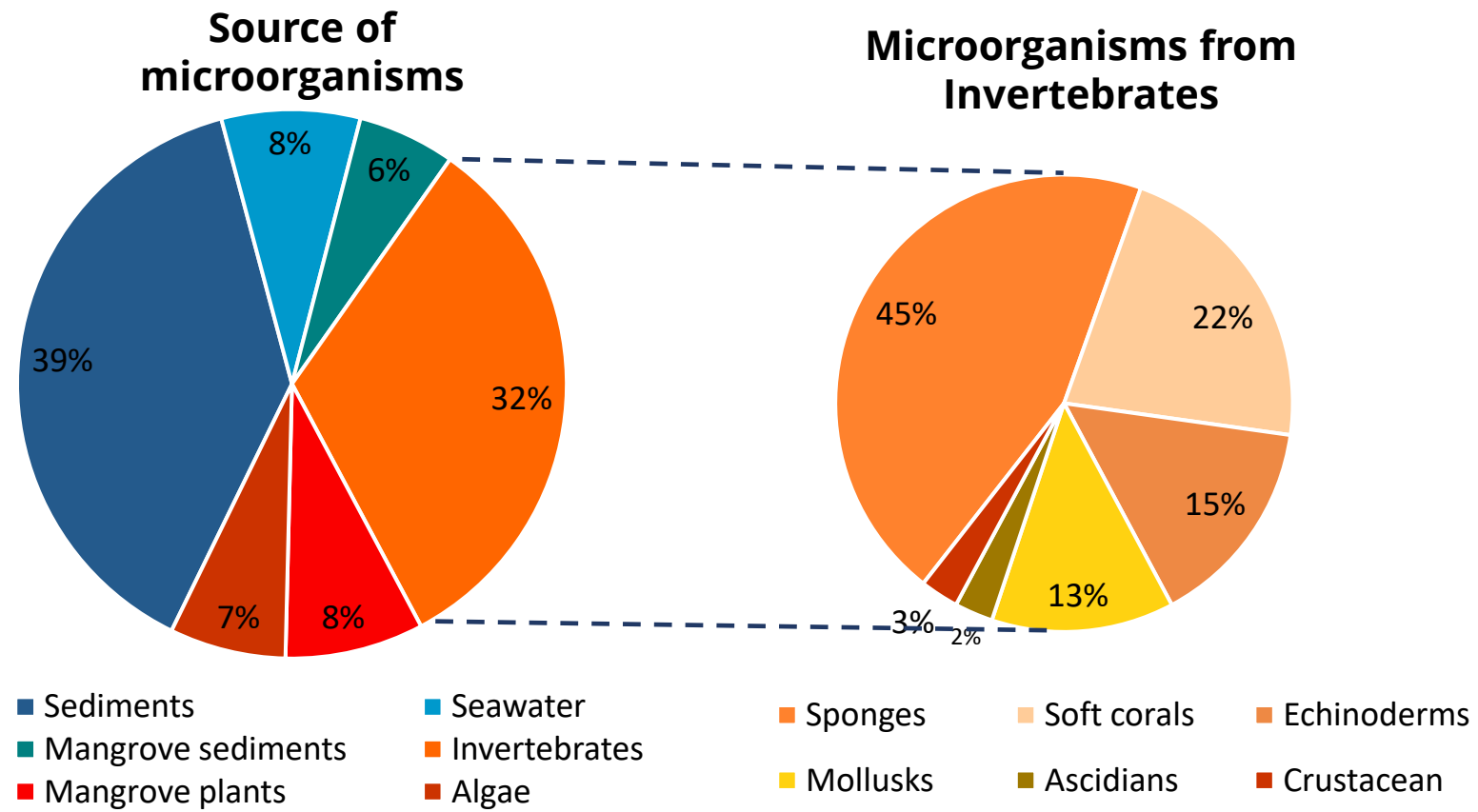
**Is a prolific source of novel bioactive natural products**

**Percentage of new marine drugs (2006 - 2016):**




Raimundo, I, **Silva, SG**., Costa, R & Keller-Costa, T 2018 *Marine Drugs*

# Novel Natural Products from the Seas

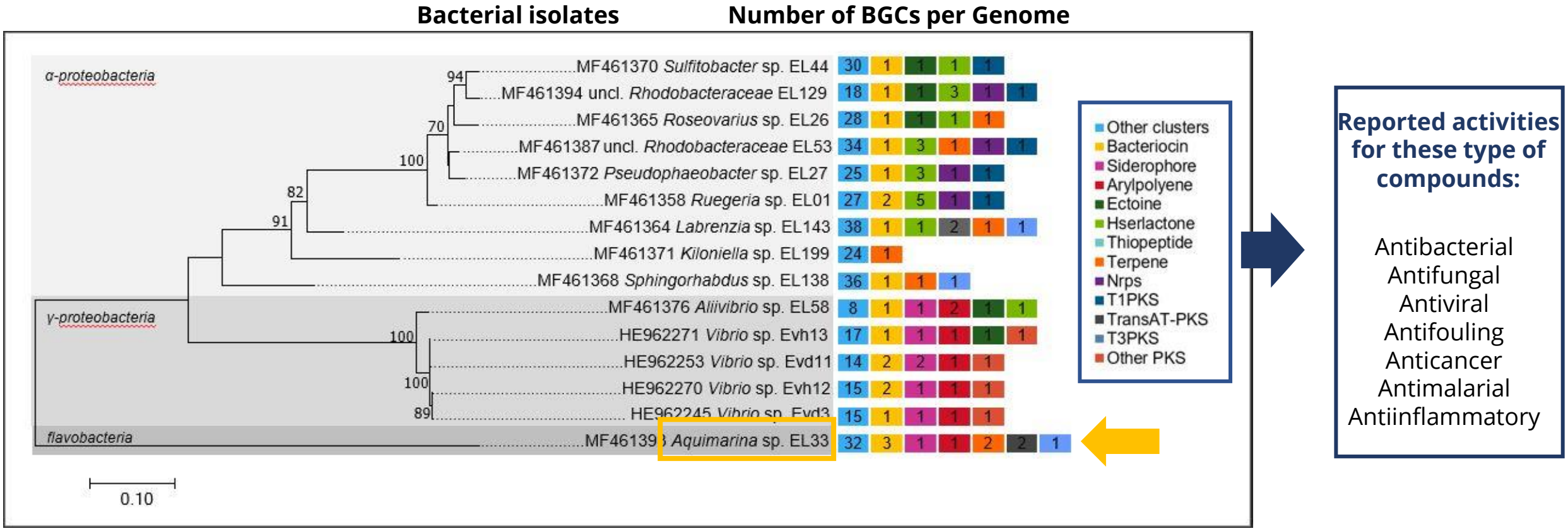



**Year: 2013**

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

Raimundo, I, **Silva, SG**., Costa, R & Keller-Costa, T 2018 *Marine Drugs*

# Potential for Secondary Metabolite Synthesis in Soft Coral-Associated Bacteria




**Reported activities for these type of compounds:**

Antibacterial
Antifungal
Antiviral
Antifouling
Anticancer
Antimalarial
Antiinflammatory

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

Raimundo, I, **Silva, SG.**, Costa, R & Keller-Costa, T 2018 *Marine Drugs*

# 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.

environmental microbiology

Environmental Microbiology (2019) 21(11), 4002–4019 doi:10.1111/1462-2920.14747

# 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**

**Red algae**

**FREE-LIVING** (FL)

**Marine sediments**

**Seawater**

**Silva, SG**., Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*

# Methods

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

Download of 26 genomes

**Genome annotation on** RAST Rapid Annotation using Subsystem Technology version 2.0

**16S rRNA phylonetic analysis**

MEGA

**Annotation on COGs and Pfams**

WebMGA web services for metagenomic analysis

**BGCs identification**

antiSMASH

**Phylogenomics analysis**

EDGAR

**16S rRNA phylogenetic analysis**

MEGA

Construction of a phylogenetic tree

**Annotation on COGs and Pfams**

WebMGA
web services for metagenomic analysis

MATLAB

Data treatment

STAMP

PAST

Statistical analysis

**BGCs identification**

antiSMASH

**Identification of inter-strain BGCs**

BiGSCAPE

**Phylogenomic analysis**

EDGAR

Construction of a phylogenenomic tree

# Goals

**Search for biosynthetic gene clusters (BGCs).**

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


**Comparison between host-associated and free-living organisms.**

**Describe general features of the genus.**

# 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.**

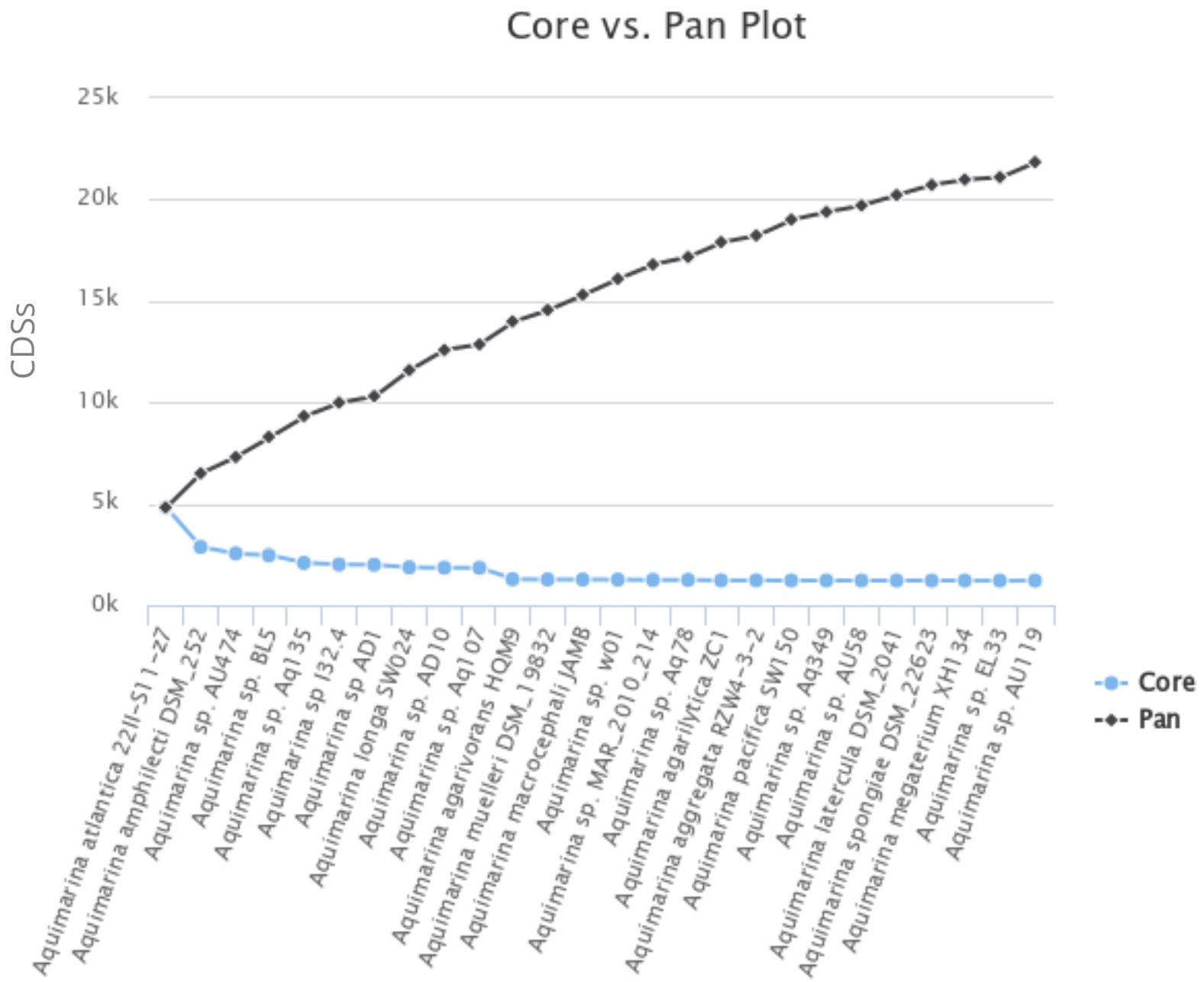
- **Core genome**: 1226 CDSs.
- **Pan genome**: 21211 CDSs.

# 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.




**Silva, SG**., Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*

# Functional Annotation

| **COG**<br>**Clusters of Orthologous Groups** | | **Pfam**<br>**Protein families' database** |
|---|---|---|
| 2320 | **Number of different ORFs** | 4187 |
| 1024 | **Core**<br>(Nr. ORFs present in all strains) | 1130 |
| 248 | **Unique**<br>(Nr. of ORFs only present in one strain) | 1716 |
| 87646 | **Total number of ORFs** | 242234 |
| 3371 | Average of number of ORFs per strain | 9317 |

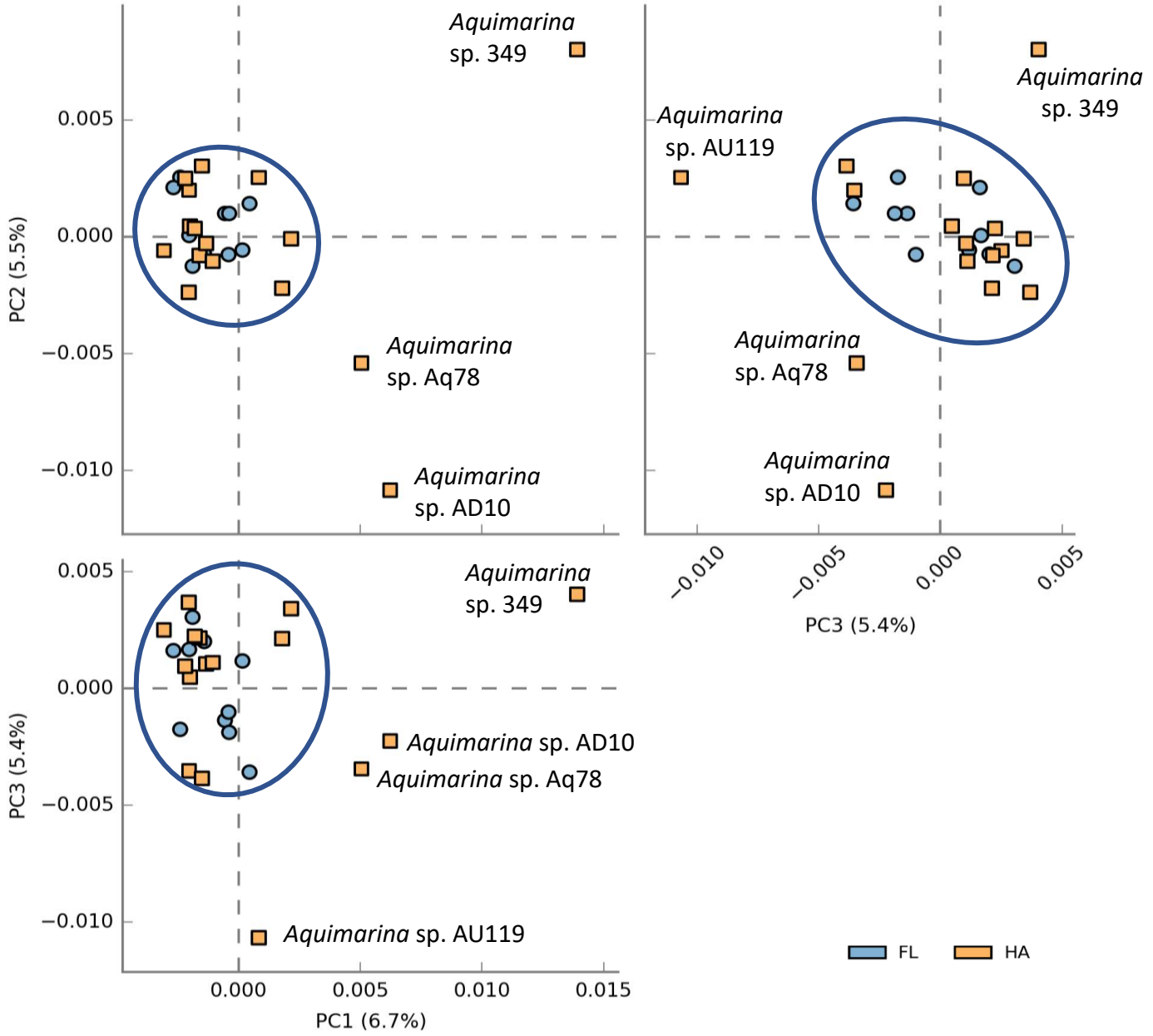
44%

27%

# Principal Component Analysis (PCA)

**Pfam annotation**

Absence of a statistical difference between annotated genomes




**One single group:**

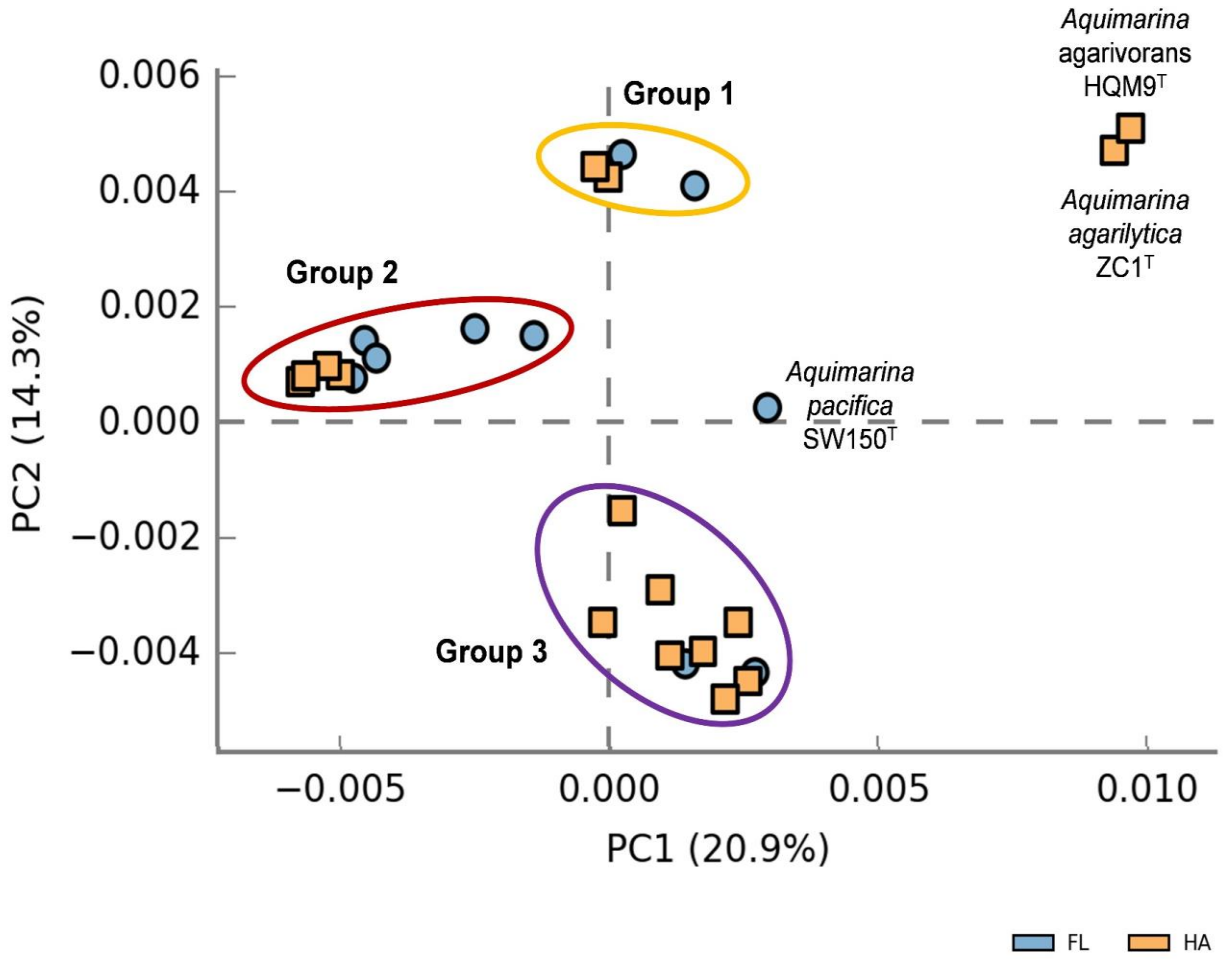
*Aquimarina* sp. I32.4
*Aquimarina longa*
*Aquimarina muelleri*
*Aquimarina* sp. Aq135
*Aquimarina* sp. w01
*Aquimarina* sp. MAR
*Aquimarina atlantica*
*Aquimarina macrocephali*
*Aquimarina* sp. AU58
*Aquimarina* sp. EL33
*Aquimarina megaterium*
*Aquimarina* sp. AU474
*Aquimarina spongiae*
*Aquimarina* 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

# COG
annotation

## Principal Component Analysis (PCA)



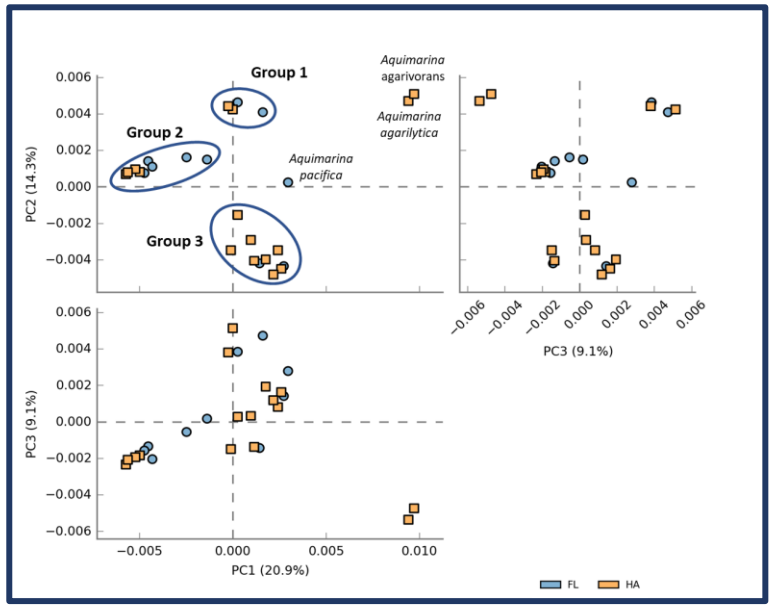

**Group 1**
*Aquimarina* sp. I32.4
*Aquimarina longa* SW024$^T$
*Aquimarina muelleri* DSM 19832$^T$
*Aquimarina* sp. Aq135

**Group 2**
*Aquimarina sediminis* w01$^T$
*Aquimarina* sp. MAR_2010_214
*Aquimarina atlantica* 22II-S11-z7$^T$
*Aquimarina macrocephali* JAMB N27$^T$
*Aquimarina* sp. Aq349
*Aquimarina* sp. AU58
*Aquimarina* sp. EL33
*Aquimarina* sp. Aq78
*Aquimarina megaterium* XH134$^T$

**Group 3**
*Aquimarina* sp. AU474
*Aquimarina* sp. AU119
*Aquimarina spongiae* A6$^T$
*Aquimarina* sp. Aq107
*Aquimarina aggregata* RZW4-3-2$^T$
*Aquimarina latercula* SIO-1$^T$
*Aquimarina* sp. BL5
*Aquimarina* sp. AD10
*Aquimarina* sp. AD1
*Aquimarina amphilecti* 92V$^T$

**Silva, SG.**, Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*

# COG
annotation




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

Are these groups statistically significant?

**Yes**

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

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

**SIMPER analysis**

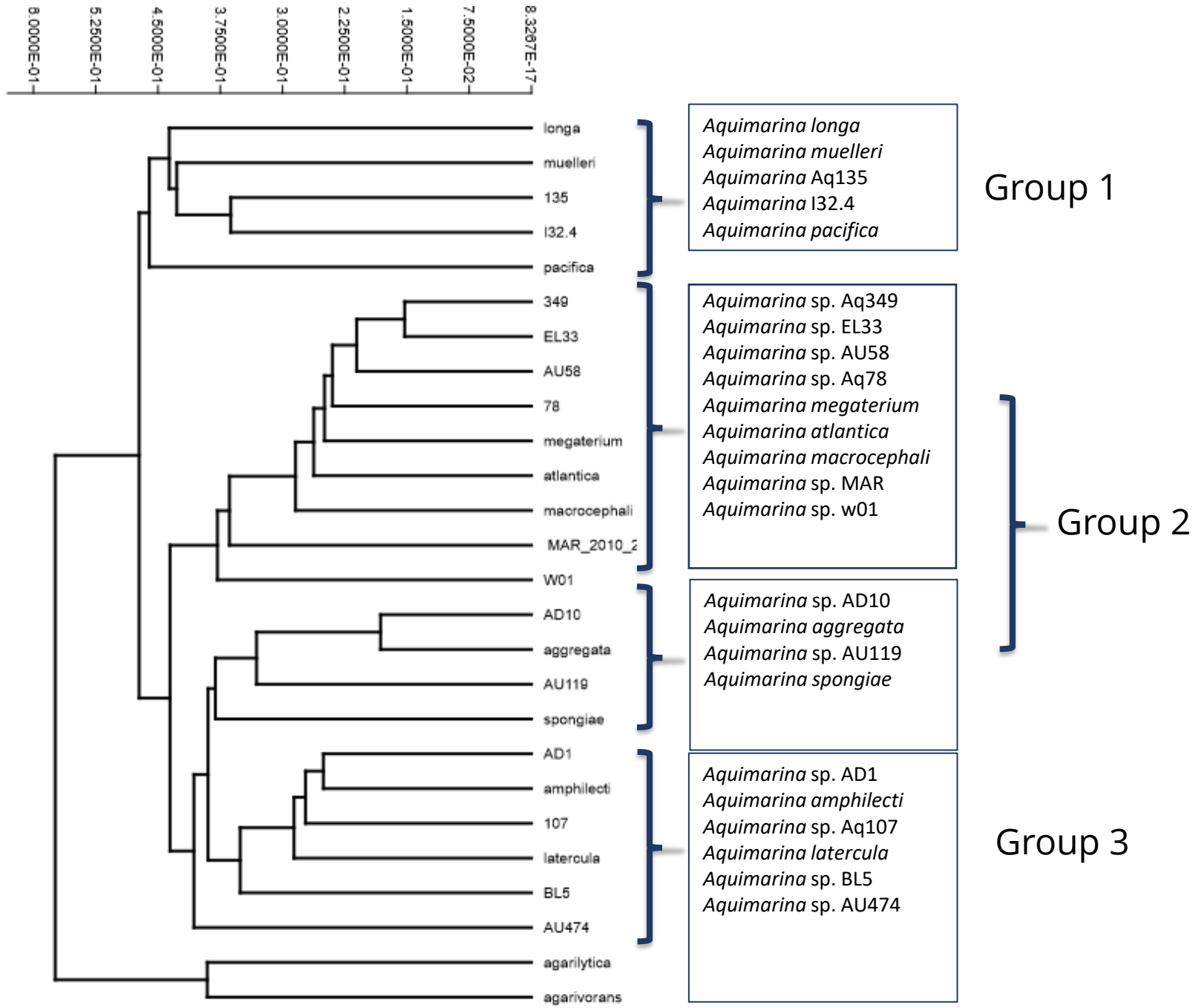
COG
annotation

→ 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 |

# COG
## annotation

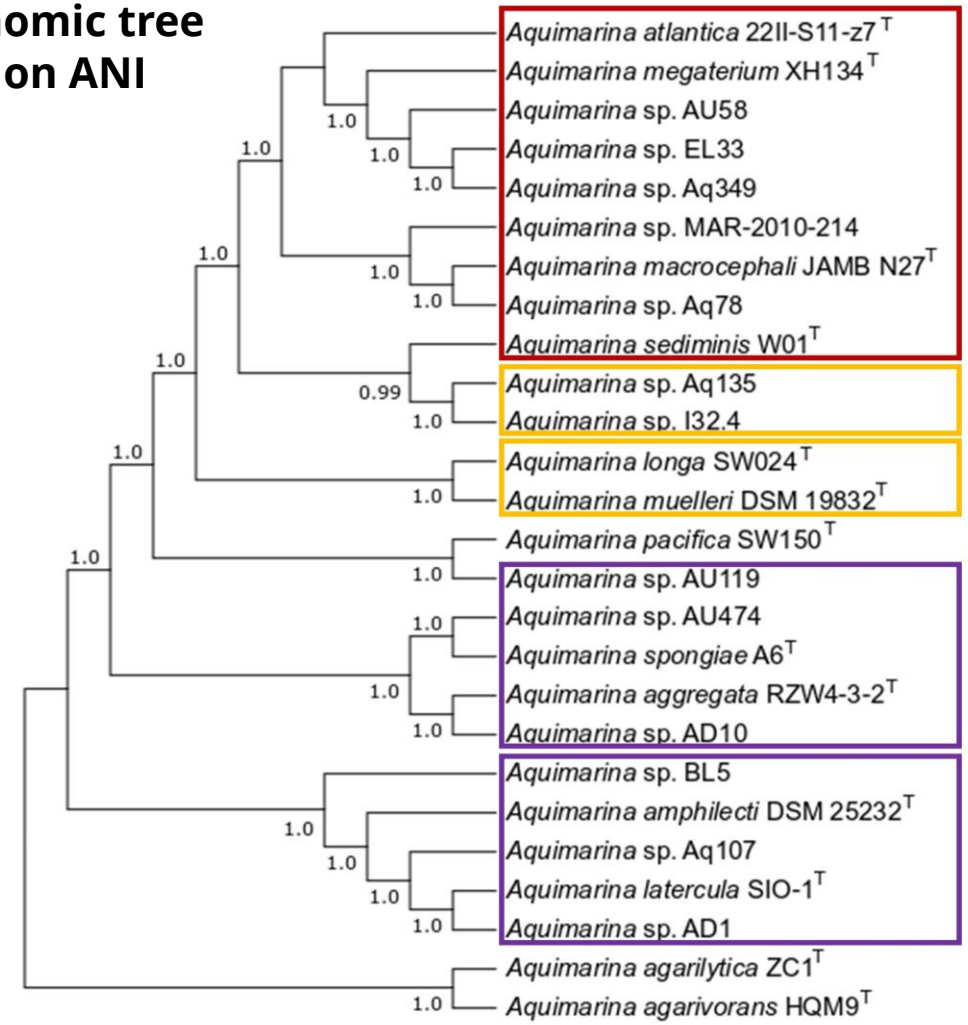



Clustering with the paired group (UPGMA).

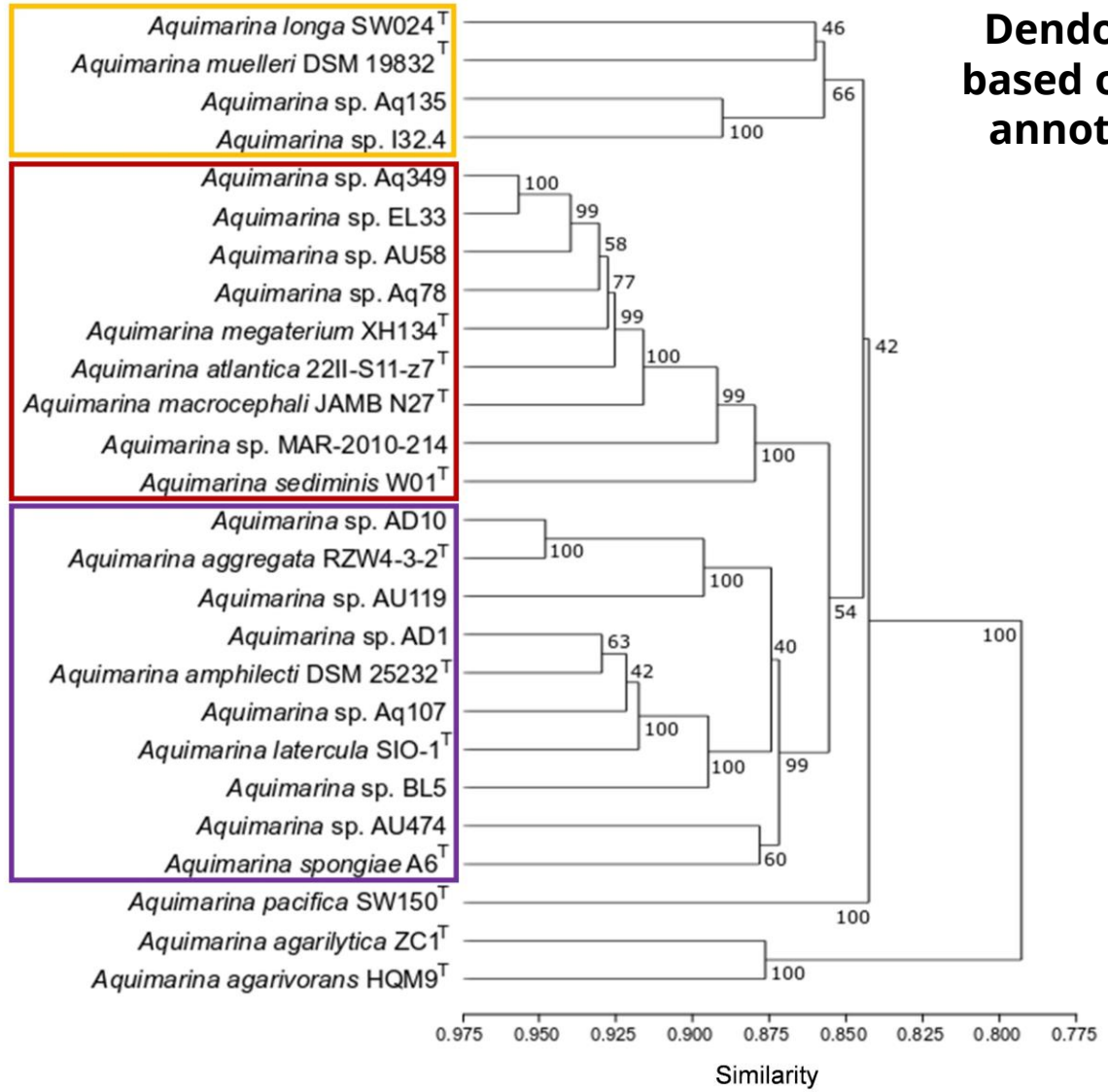
5000 bootstraps.
Euclidean similarity index

# Phylogeny primarily shapes the metabolism of *Aquimarina* species

**Phylogenomic tree based on ANI**




**Dendogram based on COG annotation**



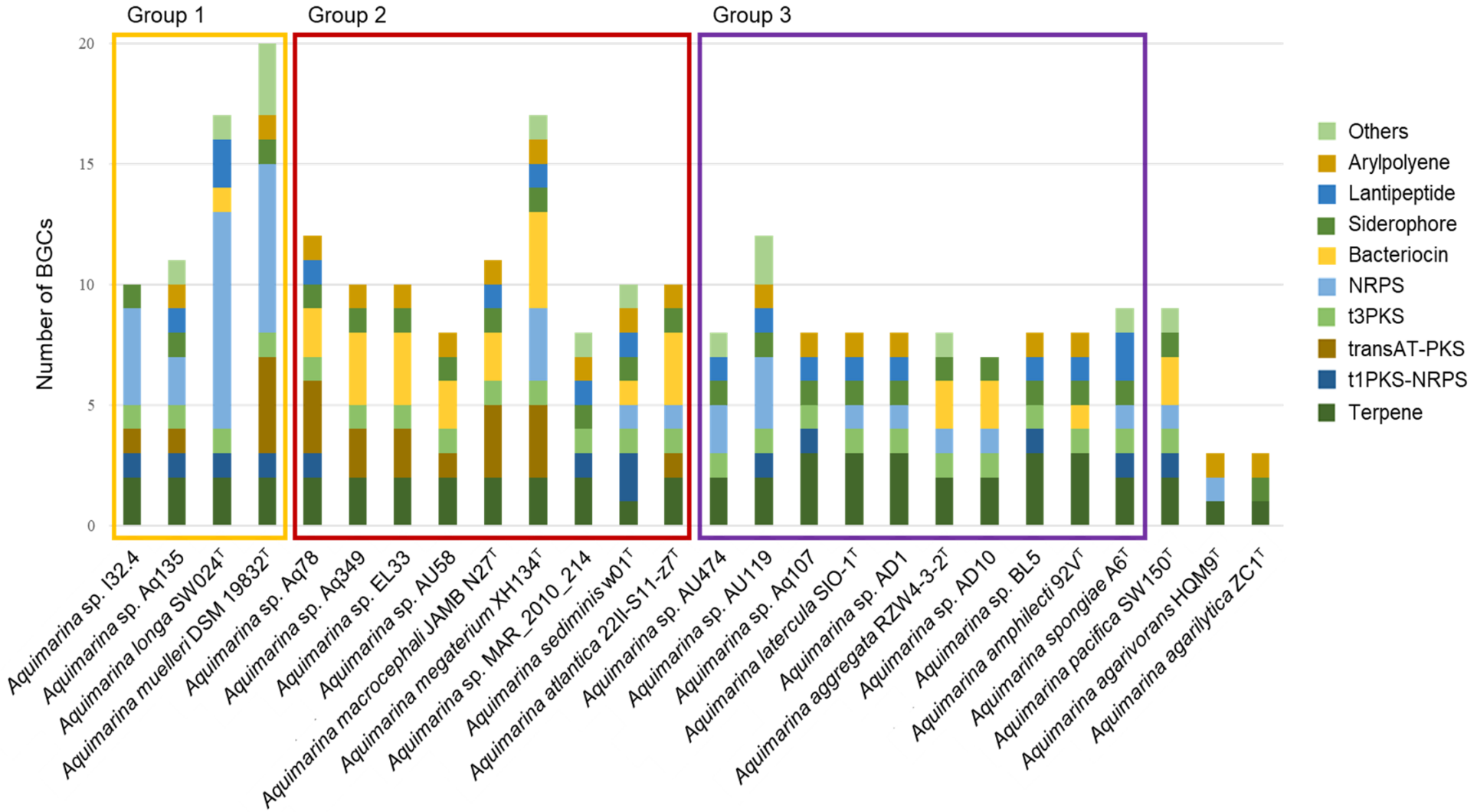

**Silva, SG.**, Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*

# Identification of BGCs

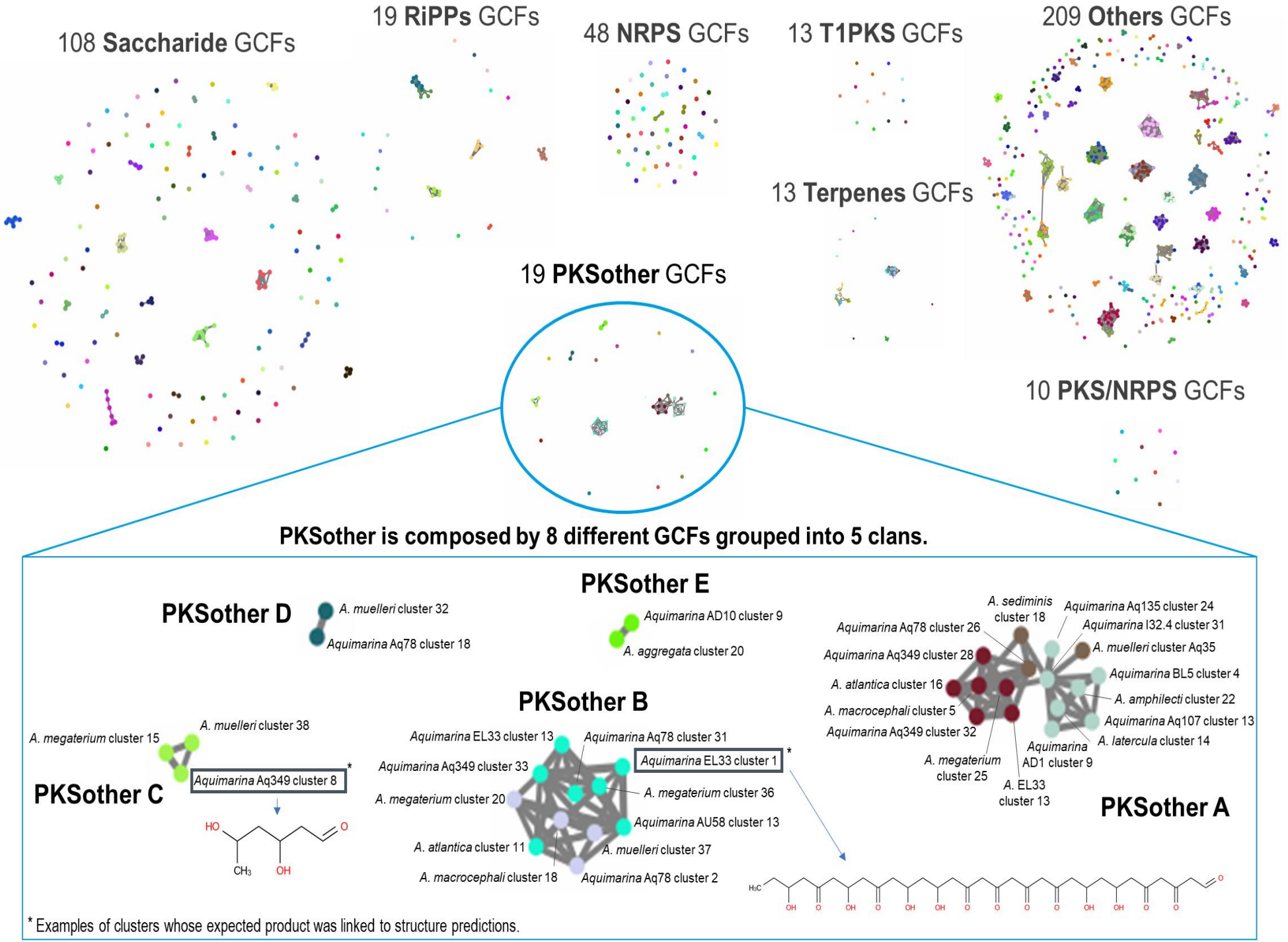
Total count:
928 BGCs

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

High biosynthetic diversity

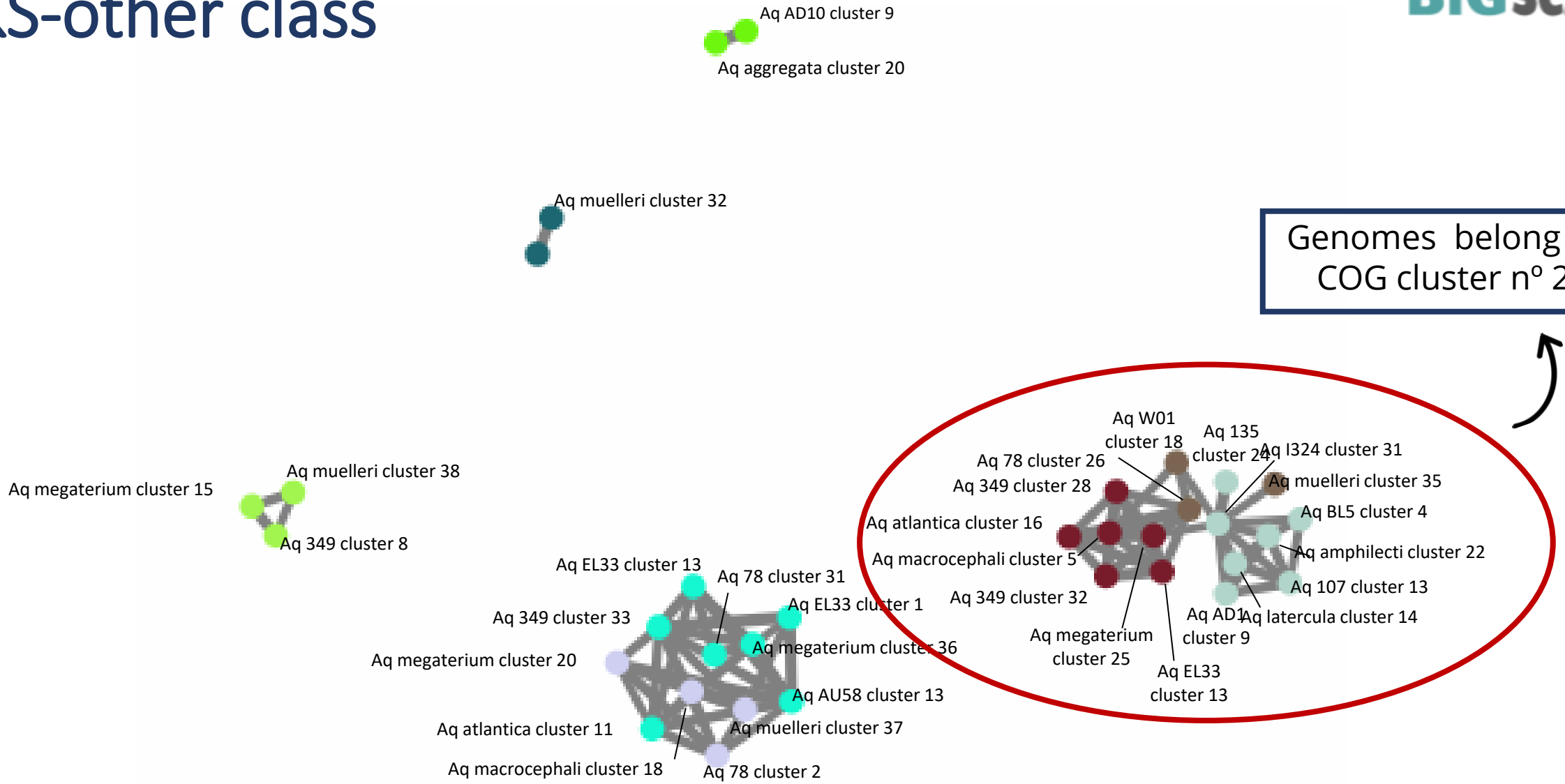



**Silva, SG.**, Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*




439
**Gene Cluster Families**

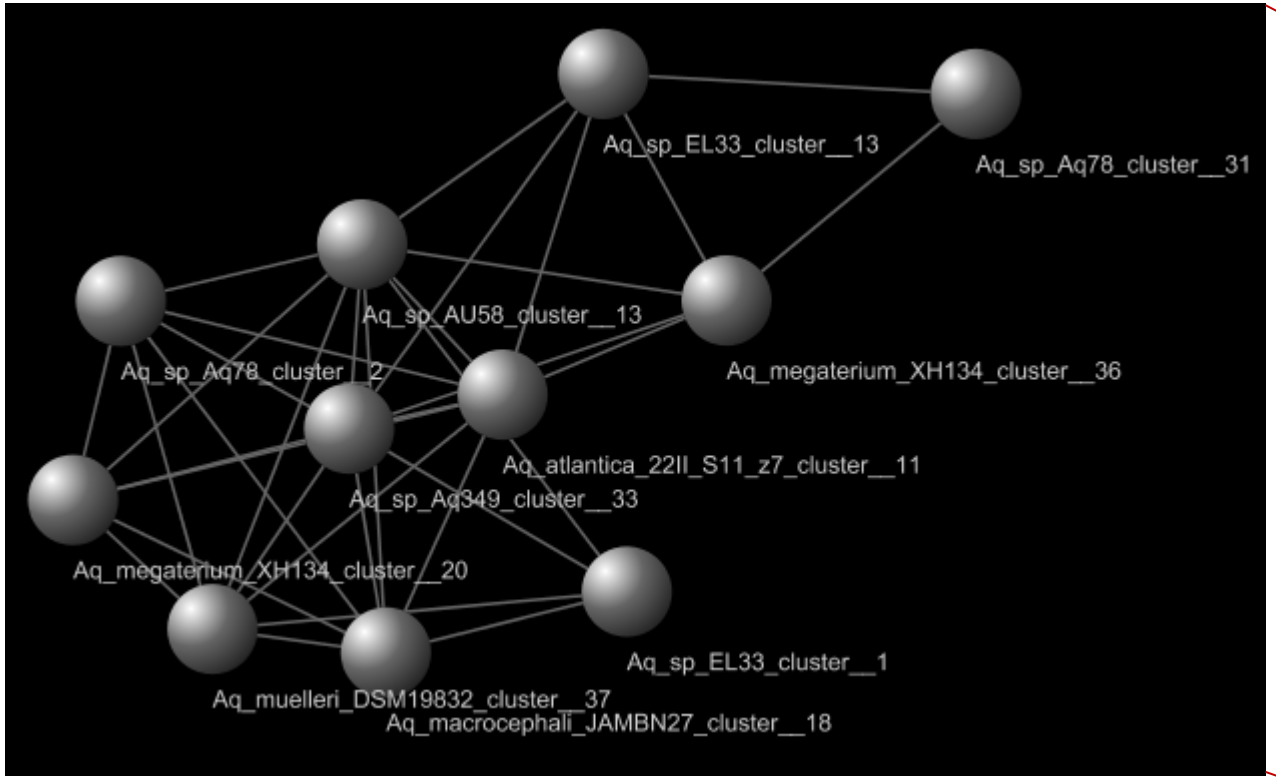
# PKS-other class

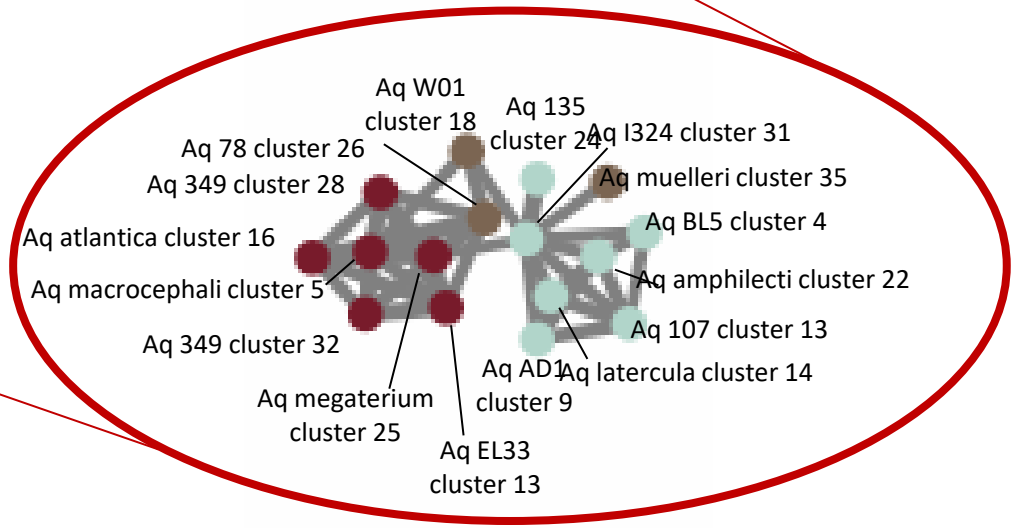


BiGSCAPE

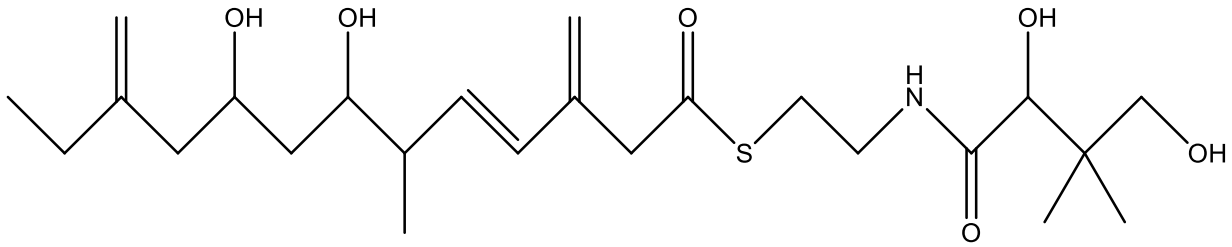
# Cuniculene BGC

*trans*-AT PKS

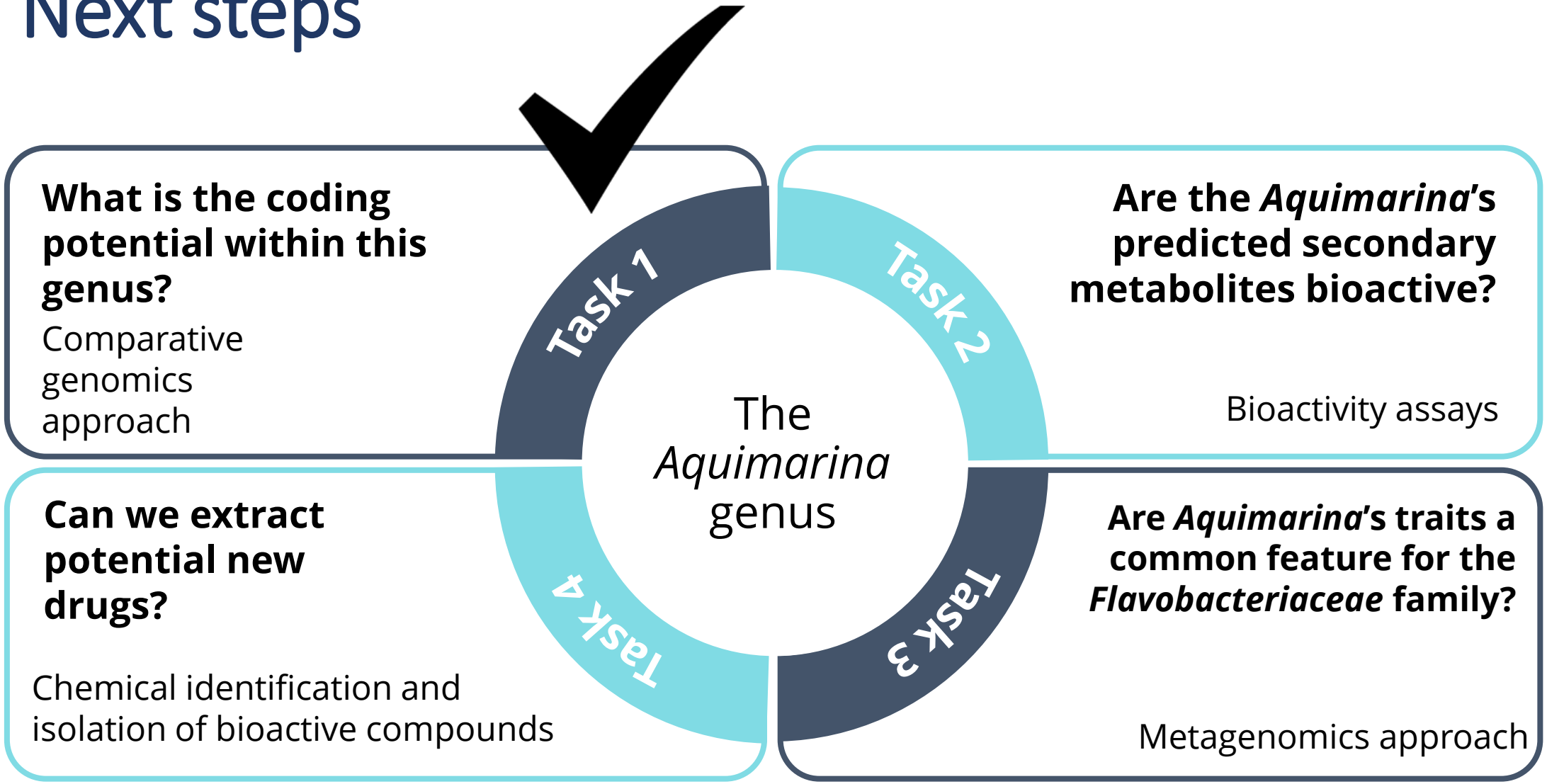



Isolated biosynthetic intermediate



# Next steps


What is the coding potential within this genus?
Comparative genomics approach
Task 1
Task 2
Are the *Aquimarina*'s predicted secondary metabolites bioactive?
Bioactivity assays
The *Aquimarina* genus
Task 4
Task 3
Can we extract potential new drugs?
Chemical identification and isolation of bioactive compounds
Are *Aquimarina*'s traits a common feature for the *Flavobacteriaceae* family?
Metagenomics approach

# 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


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

**Hands-on:**

# Bioinformatic tools for the identification of 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.

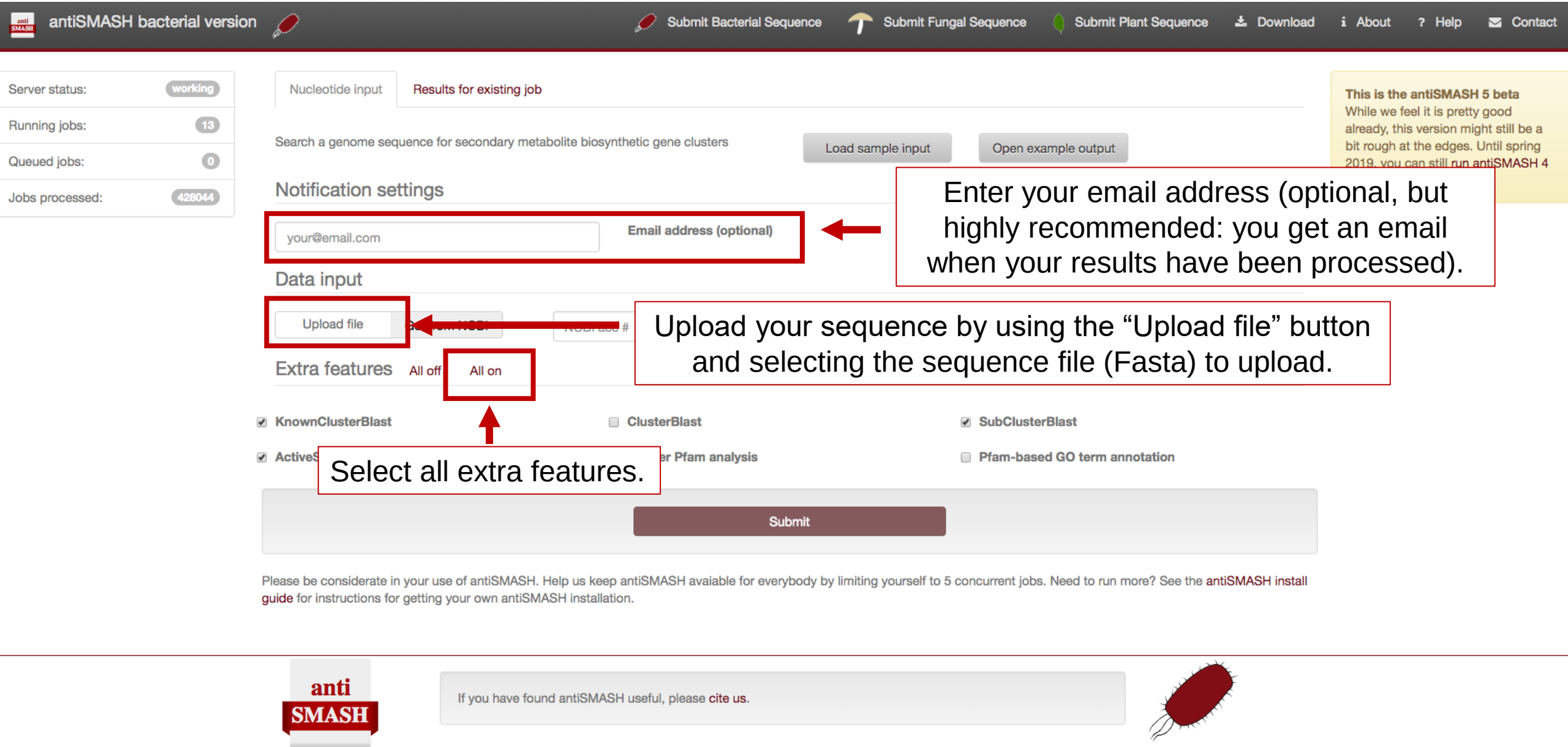
It integrates and cross-links with many in silico secondary metabolite analysis tools and is powered by several open-source tools:

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


Created in 2011
Current version: 5.0

antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Kai Blin et al (2019) doi: 10.1093/nar/gkz310.

# antiSMASH - Job submission page

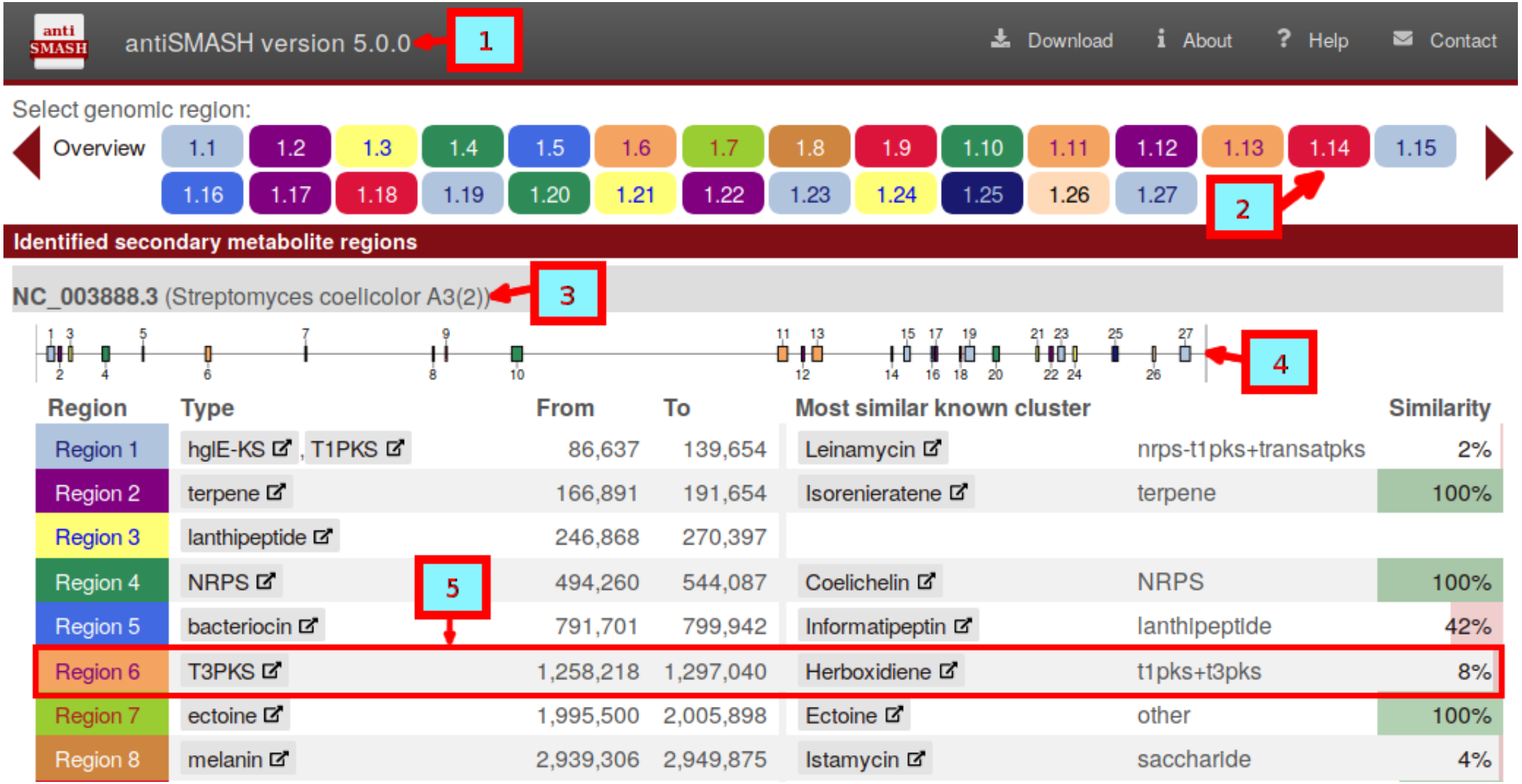


# Now it’s your turn!

Practical exercise: submit your metagenome sequences into antiSMASH.

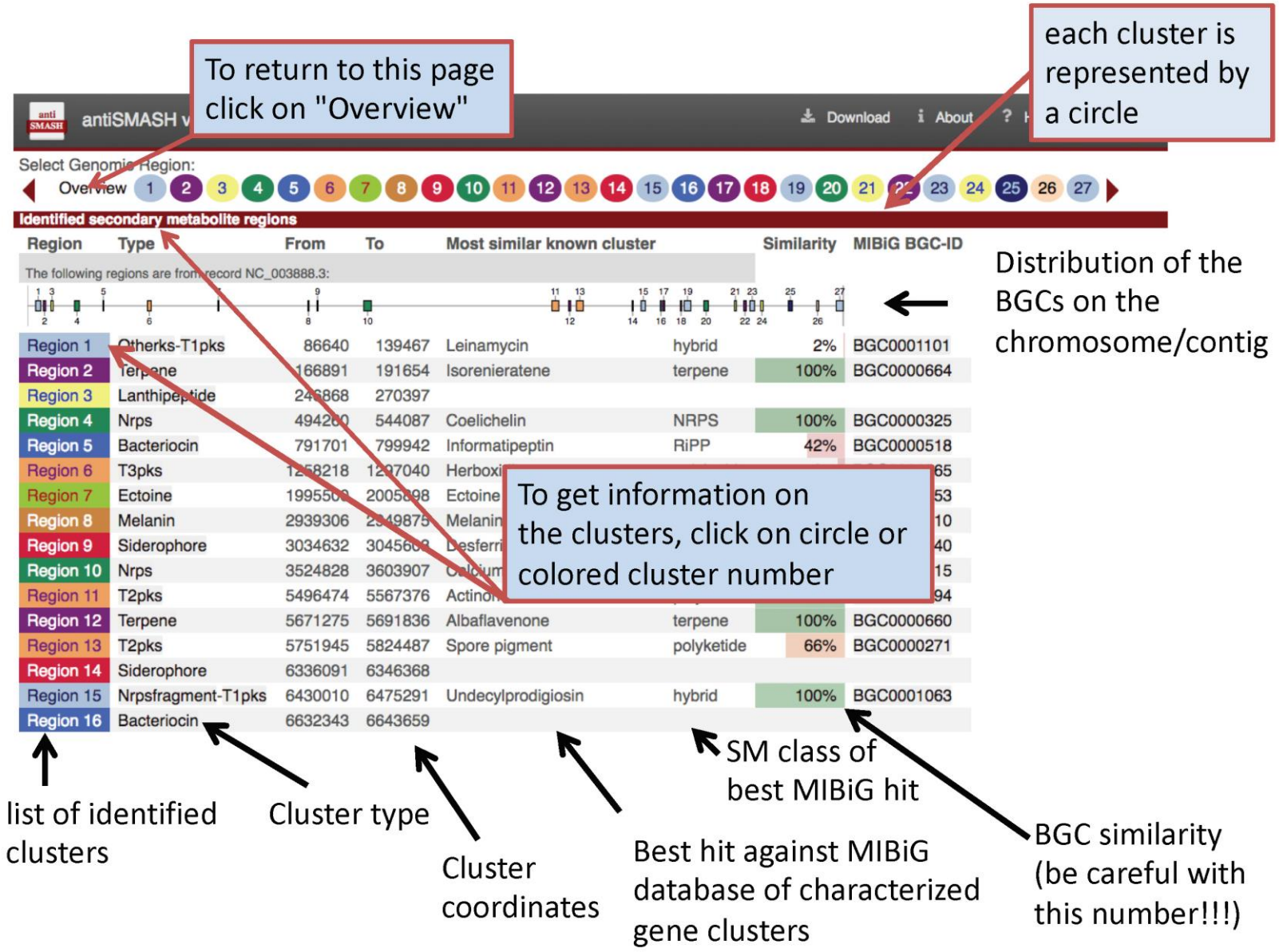
https://antismash.secondarymetabolites.org/

# antiSMASH – The output




| Region | Type | From | To | Most similar known cluster | | Similarity |
|---|---|---|---|---|---|---|
| Region 1 | hglE-KS, T1PKS | 86,637 | 139,654 | Leinamycin | nrps-t1pks+transatpks | 2% |
| Region 2 | terpene | 166,891 | 191,654 | Isorenieratene | terpene | 100% |
| Region 3 | lanthipeptide | 246,868 | 270,397 | | | |
| Region 4 | NRPS | 494,260 | 544,087 | Coelichelin | NRPS | 100% |
| Region 5 | bacteriocin | 791,701 | 799,942 | Informatipeptin | lanthipeptide | 42% |
| Region 6 | T3PKS | 1,258,218 | 1,297,040 | Herboxidiene | t1pks+t3pks | 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% |

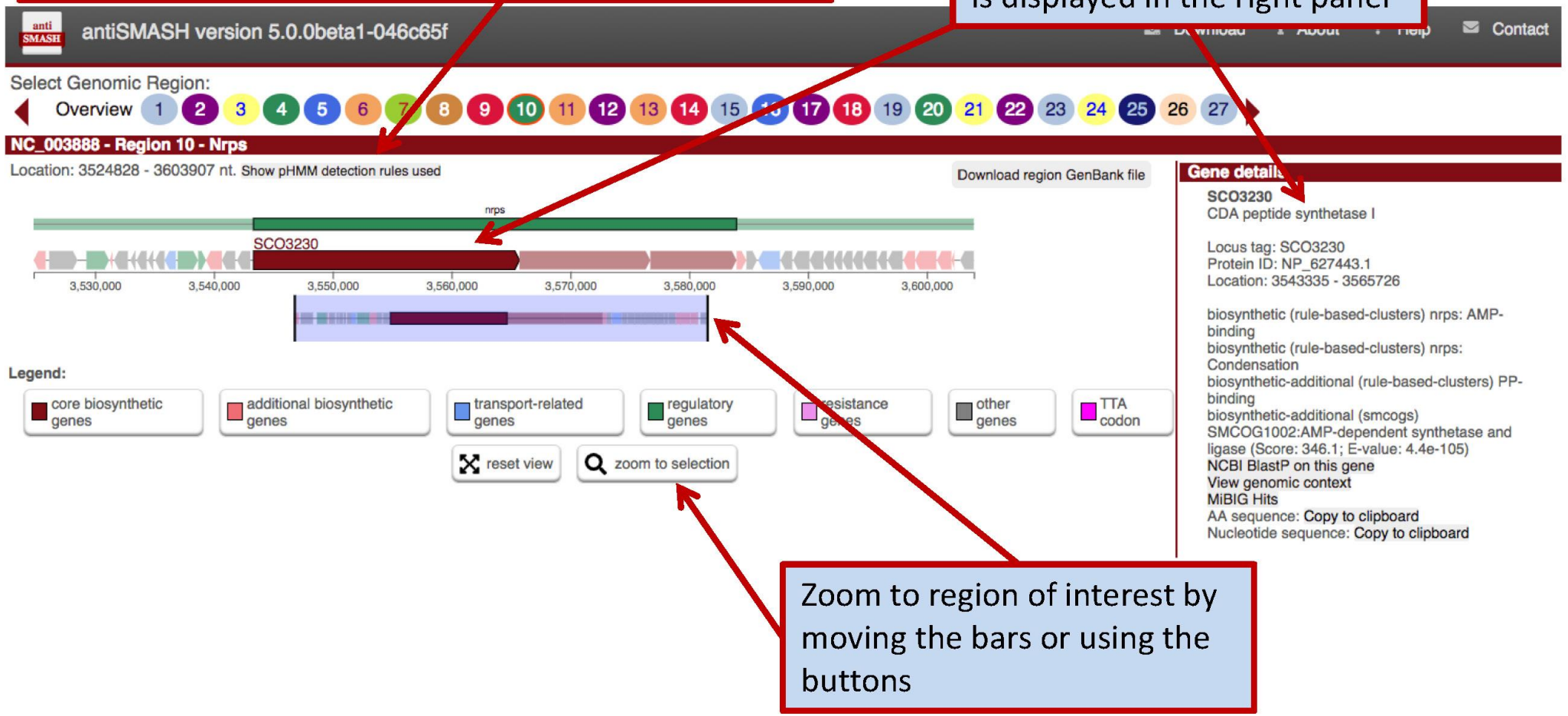
# antiSMASH – The output



# antiSMASH – The output

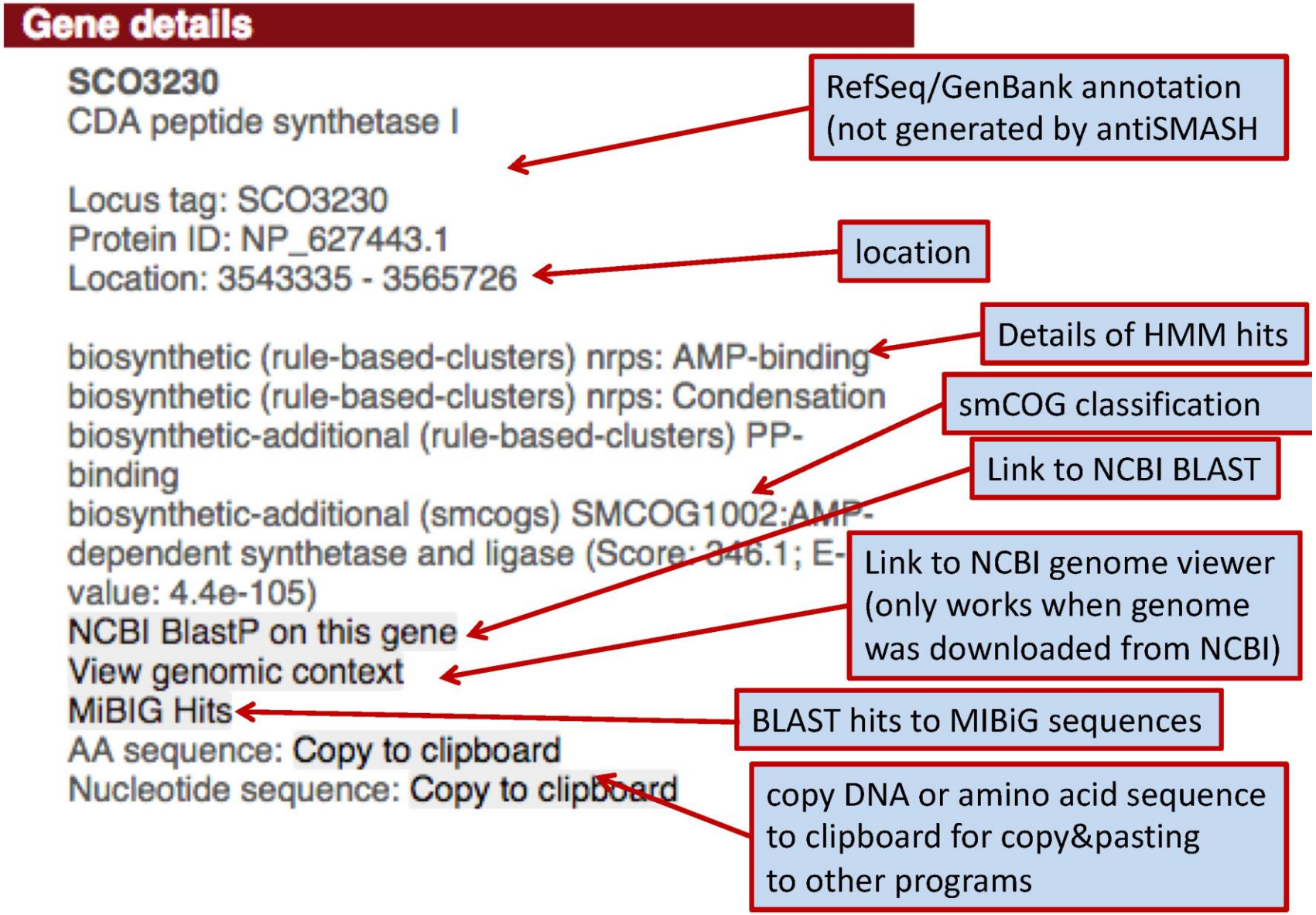
To get information on
the rule that antiSMASH used to identify the genetic region as a secondary metabolite biosynthetic gene cluster, click here

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

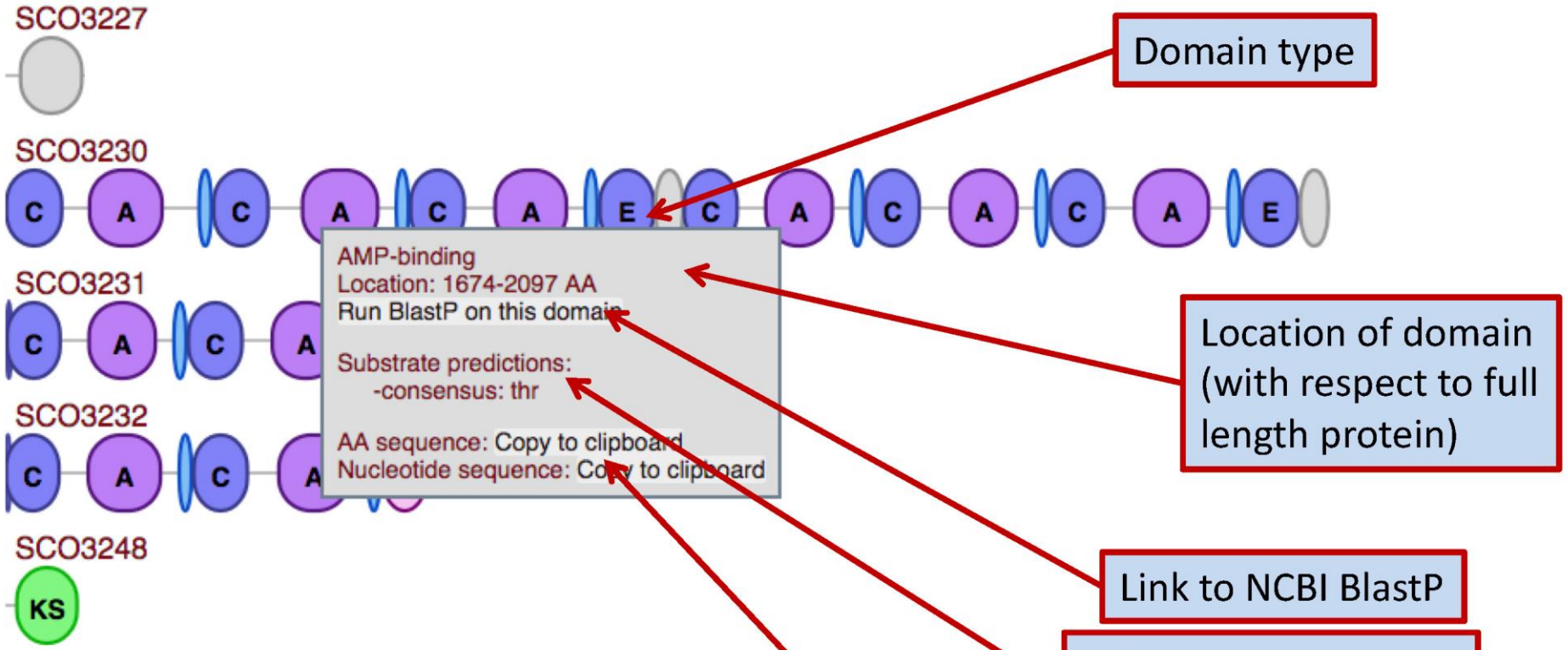



Zoom to region of interest by moving the bars or using the buttons

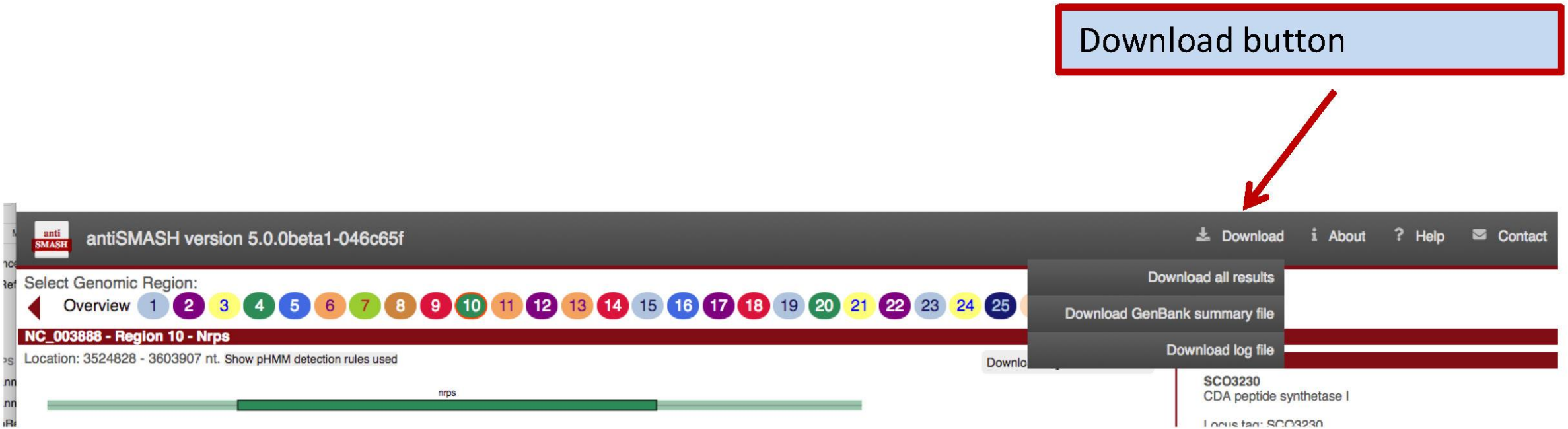
# antiSMASH – The output



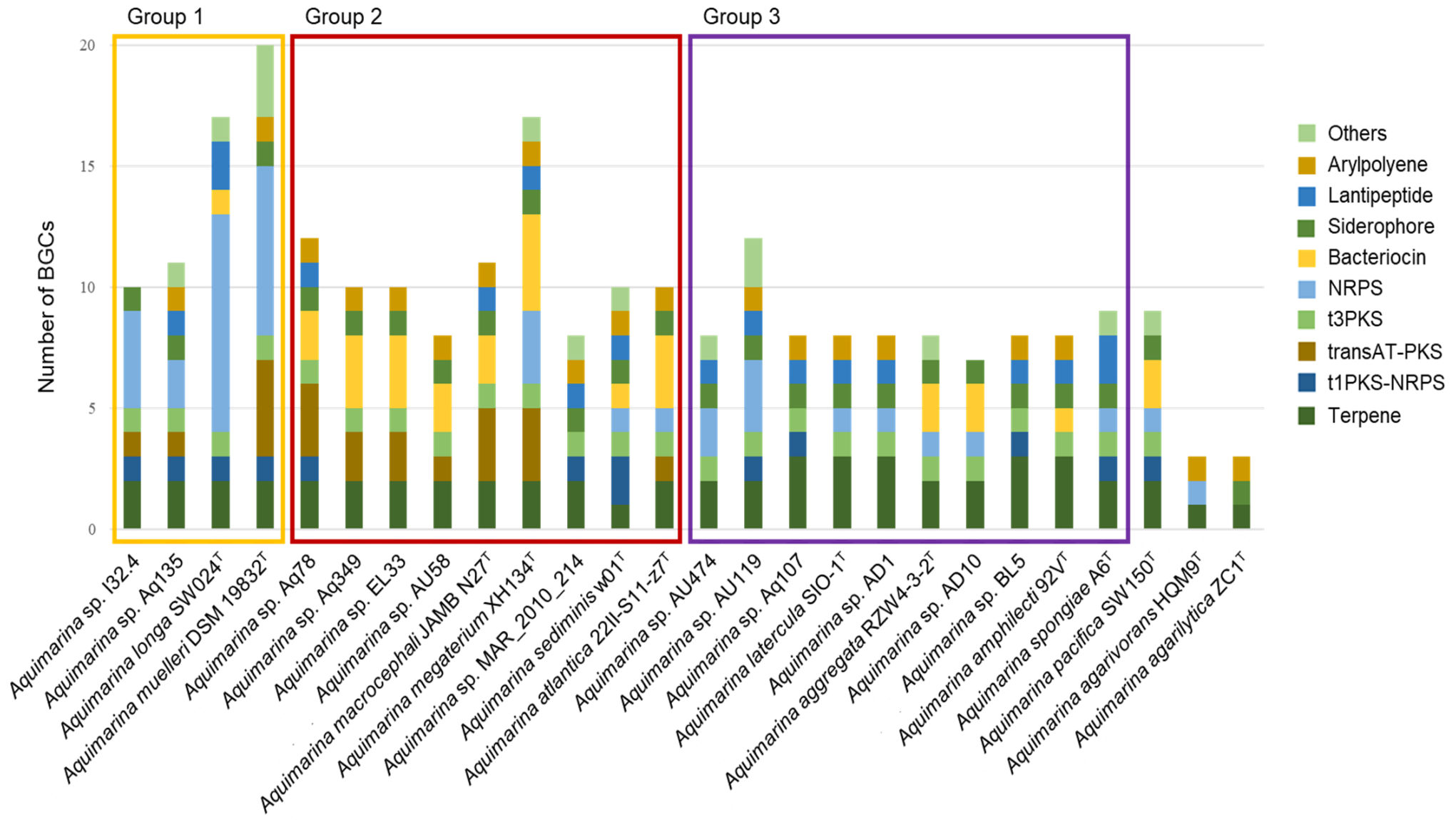
# antiSMASH – The output



# antiSMASH – The output



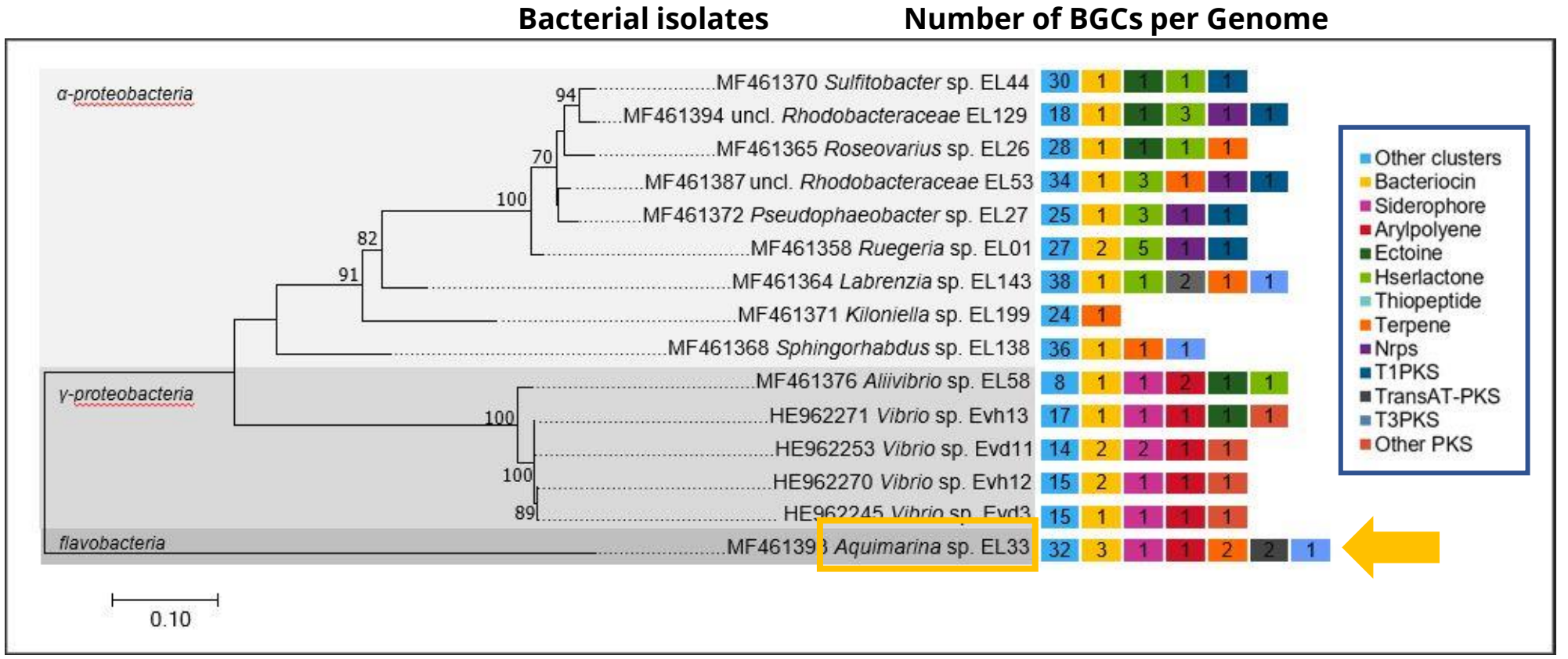
# Result obtained for the *Aquimarina study:*




**Silva, SG.**, Blom, J., Keller-Costa, T, Costa, R. 2019 *Environmental Microbiology*

**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*.

Raimundo, I, **Silva, SG.**, Costa, R & Keller-Costa, T 2018 *Marine Drugs*

# 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.

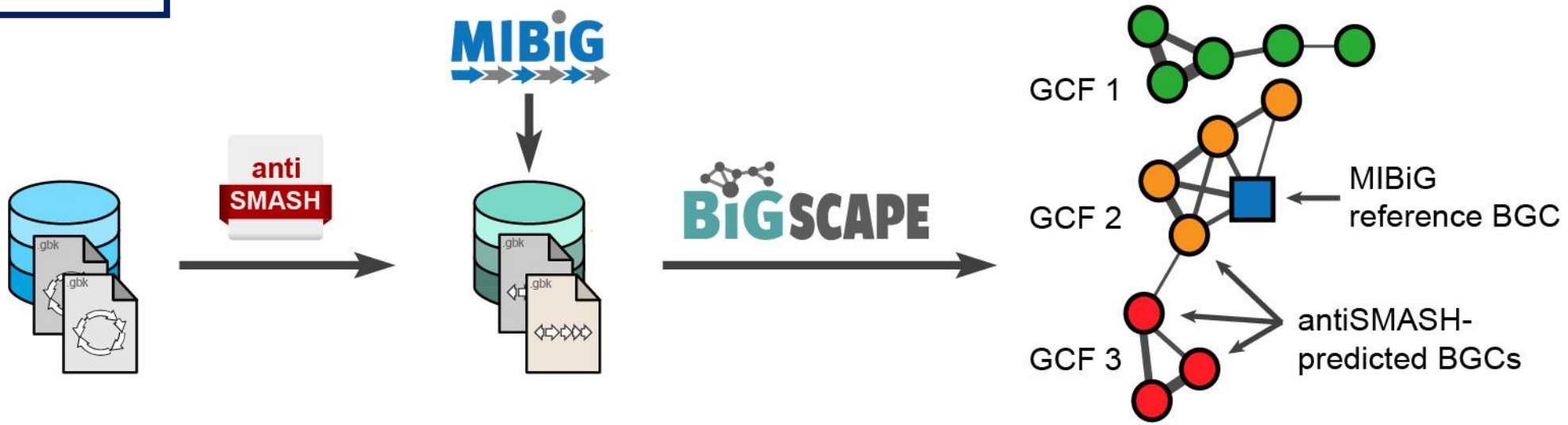
BiG-SCAPE's interactive visualizations of these similarity networks allows effective exploration of the diversity of BGCs, linking them to knowledge from reference data within the **MIBiG repository**

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

Navarro-Muñoz, J.C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* **16,** 60–68 (2020).

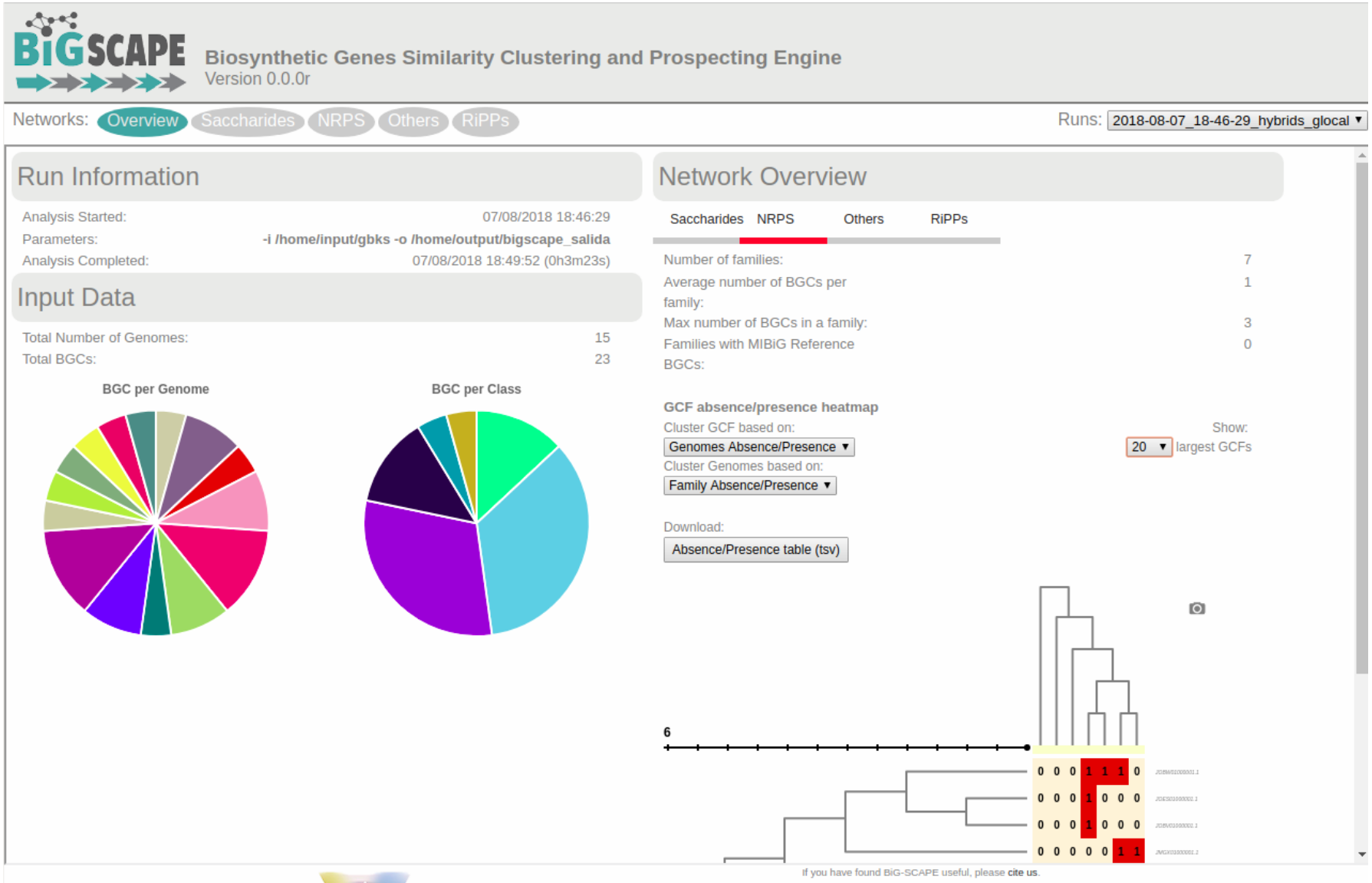
# The BiG-SCAPE workflow uses sequence similarity networking to group biosynthetic gene clusters into families

**Input:** results from antiSMASH




**Output:** graphical visualization of similarity networks

# BiG-SCAPE – The output

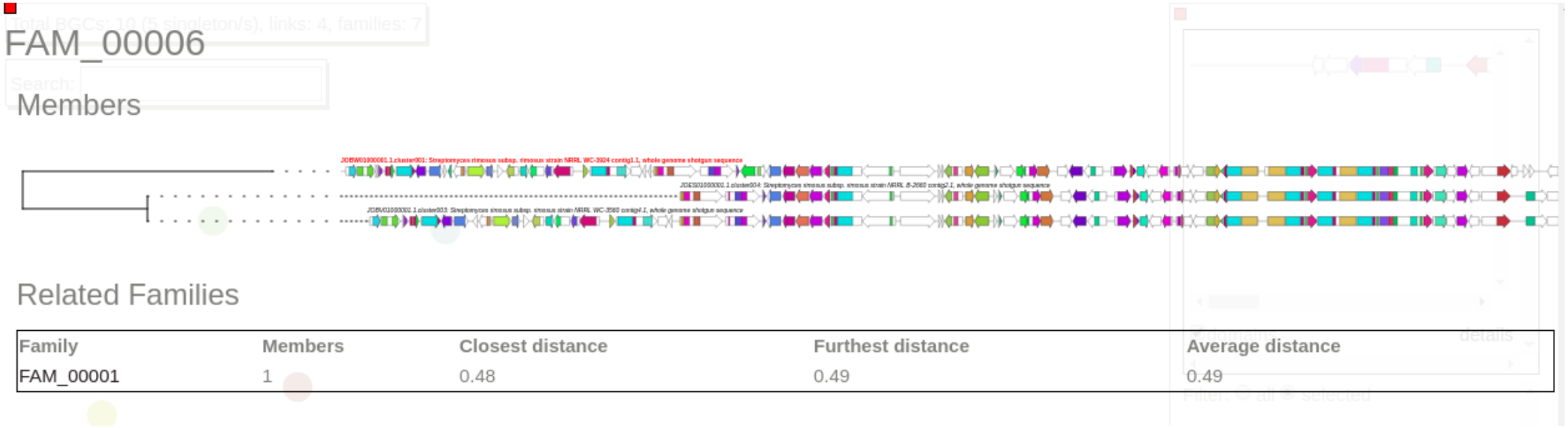


# BiG-SCAPE – The output


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

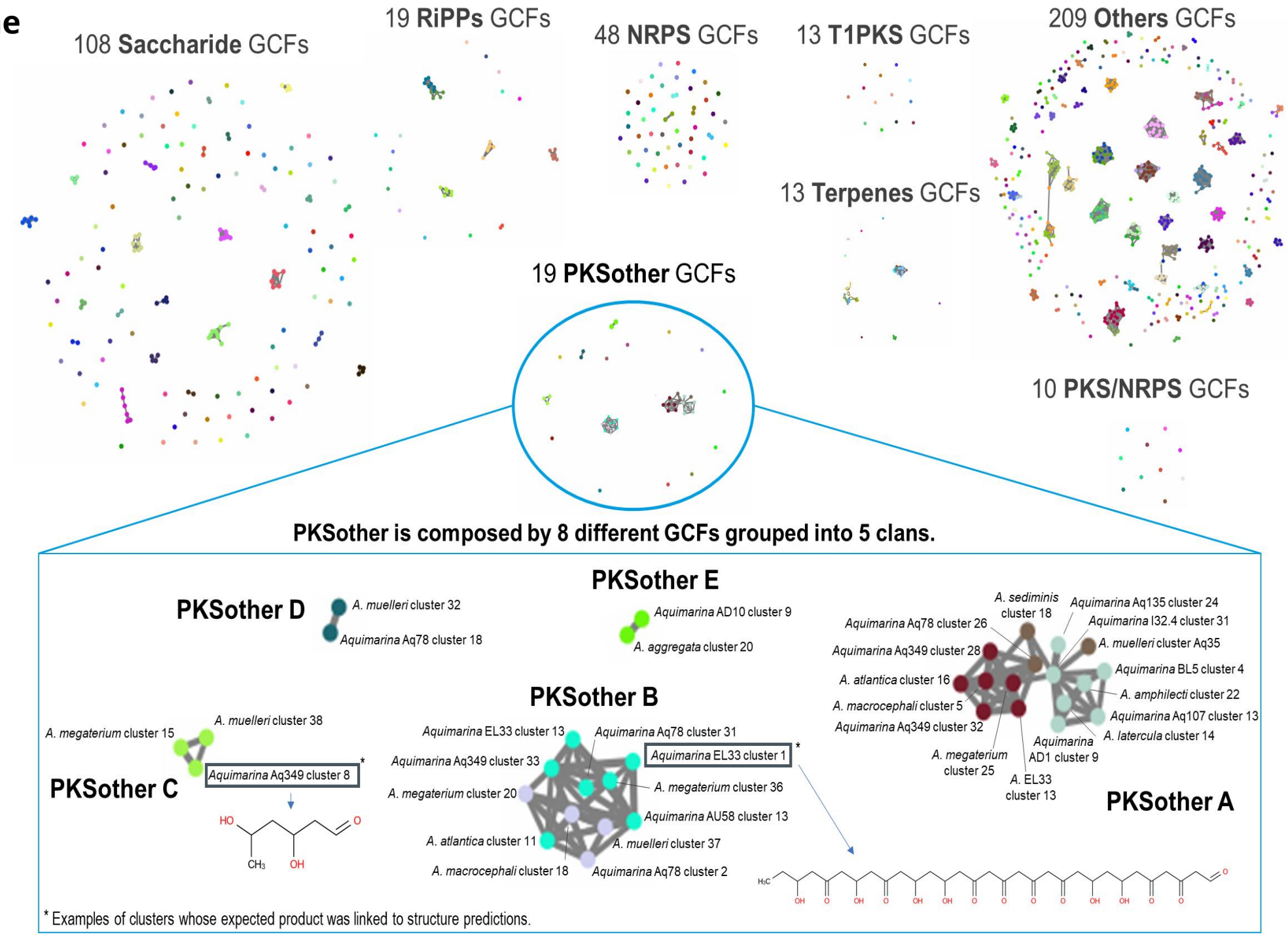
# BiG-SCAPE – The output

**Gene Cluster Family (GCF) example:**



# Result obtained for the *Aquimarina* study:

439
**Gene Cluster Families**




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

**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@gmail.com
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.

TÉCNICO LISBOA

iBB Institute for Bioengineering and Biosciences

biological sciences research group

PD F PROGRAMAS DE DOUTORAMENTO FCT

[DP_AEM] DOCTORAL PROGRAM IN APPLIED AND ENVIRONMENTAL MICROBIOLOGY

# Thank you for your attention.

**Sandra Godinho Silva**

sandragodinhosilva@tecnico.ulisboa.pt