# Introduction to Bioinformatics in Natural Products Discovery: an historical overview

**Adriana Rego**

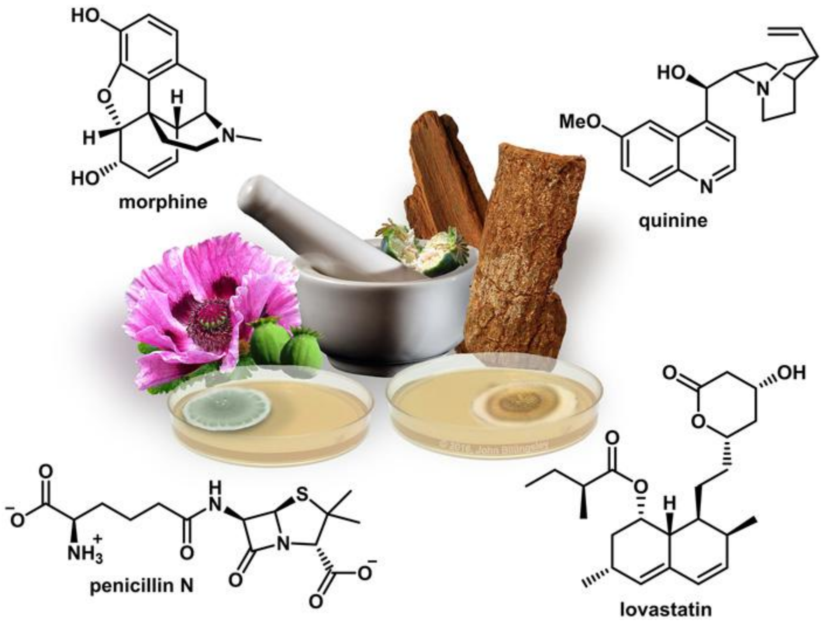
**Short Course in Bioinformatics – Decoding (Meta)Genomes for Natural Products Discovery**

**17th – 19th March 2025**

# Microorganims as a prolific source of natural products: an historical overview

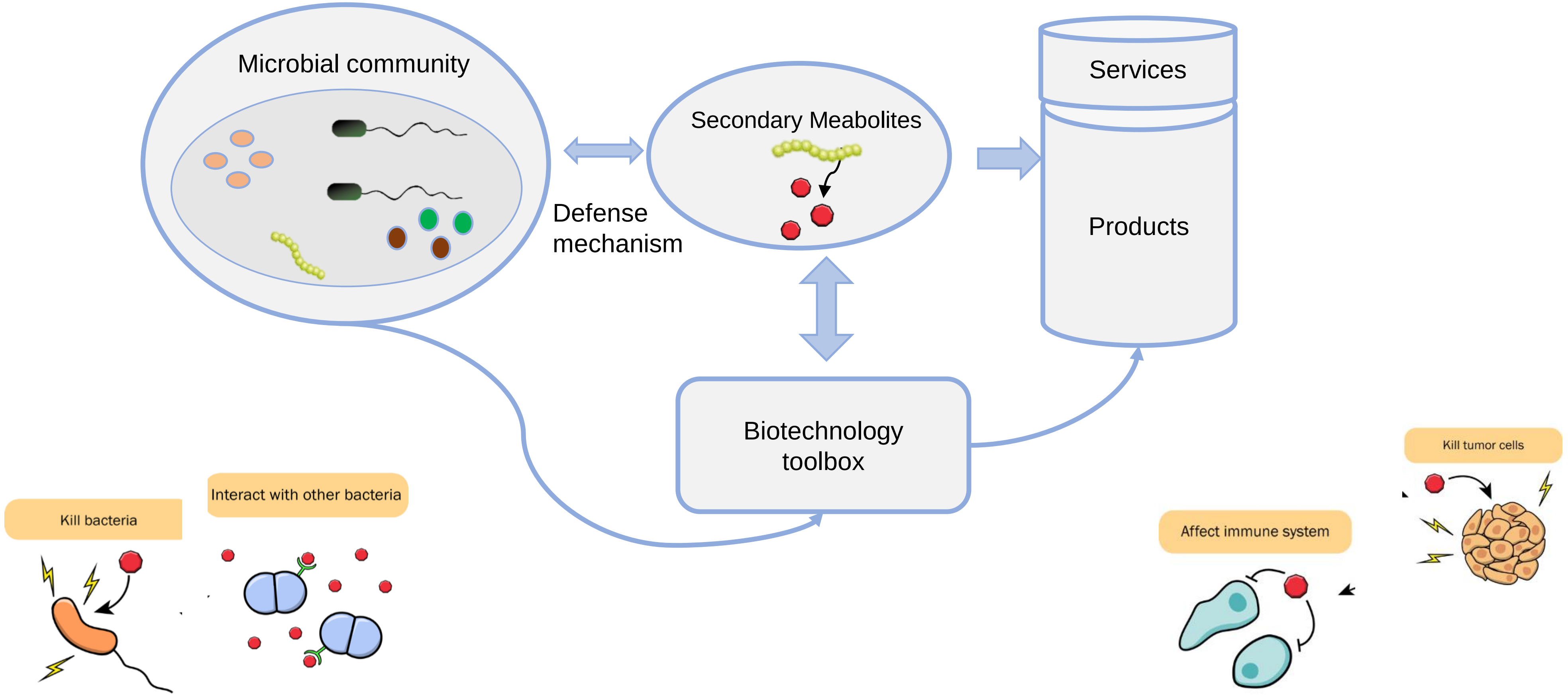
Natural products, also called secondary or specialized metabolites could be defined broadly as any **molecules found in nature**.

More traditionally in organic and medicinal chemistry communities, natural products **are defined as small organic molecules** (MW < 1500 daltons) generated **from secondary metabolic pathways.**



*Nat. Prod. Biosynth*. **2022**, RSC Publishing

# Microorganims as a prolific source of natural products: an historical overview



Microbial community

Defense mechanism

Secondary Meabolites

Services

Products

Biotechnology toolbox

Kill bacteria

Interact with other bacteria

Affect immune system

Kill tumor cells

Confer some form of advantage or protection to the producers

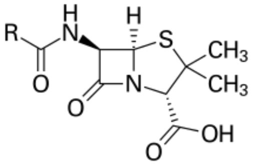
Pharmacologic activities

(Adapted from Senft 2020)

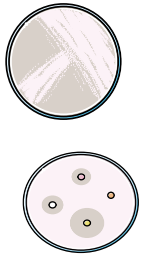
# Microorganims as a prolific source of natural products: an historical overview

**1929**

Penicillin

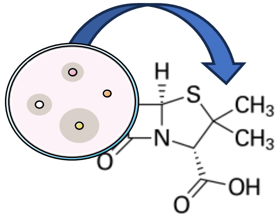


Easily cultivable bacteria



23 000 new NPs

Re-discovery



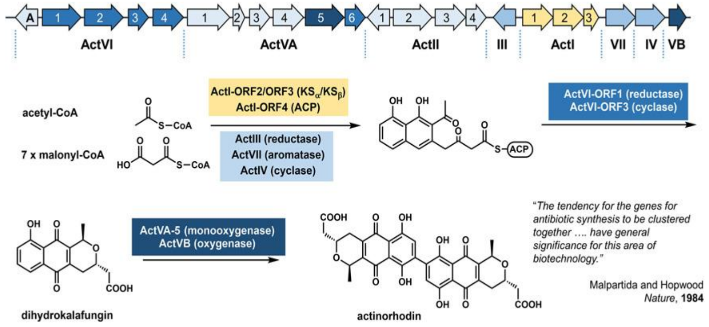
Beginning of an age of natural product-driven medicine

Potential of cultured microbes had been exhausted

**bioactivity-guided fractionation**

# Microorganims as a prolific source of natural products: an historical overview

## 1984



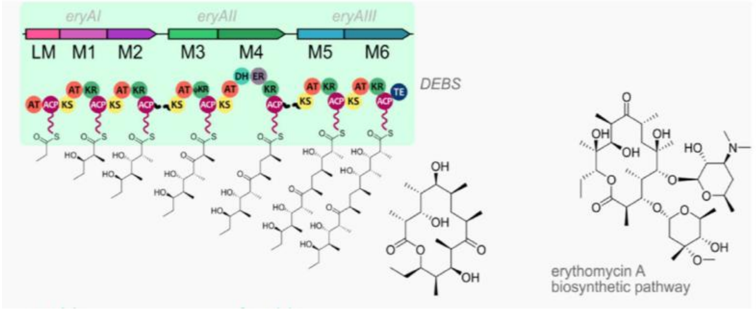
*"The tendency for the genes for antibiotic synthesis to be clustered together …. have general significance for this area of biotechnology."*

Malpartida and Hopwood
*Nature*, **1984**

*Nature* **1984,** 309, 462–464

This study established that the necessary **biosynthetic genes are clustered on a contiguous stretch of DNA.**

## 1990, 1991



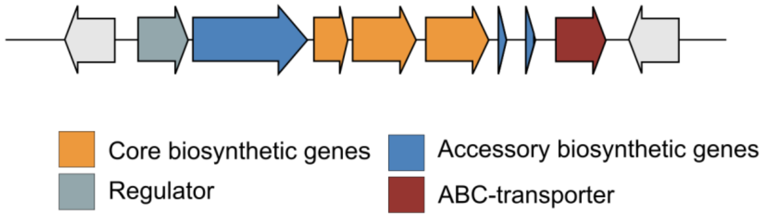
*Nature* **1990**, *348*, 176; *Science***1991**, *252*, 671

The genes required to assemble the 14-member macrolactone core of erythromycin are **contiguous** in *Saccharopolyspora Erythraea.*

# Microorganims as a prolific source of natural products: an historical overview

**A Biosynthetic Gene Cluster (BGC)** can be defined as a physically clustered group of two or more genes in a particular genome that together encode a biosynthetic pathway for the production of a specialized metabolite (including its chemical variants).

*Nature Chemical Biology* **2015**, 11, 625–631

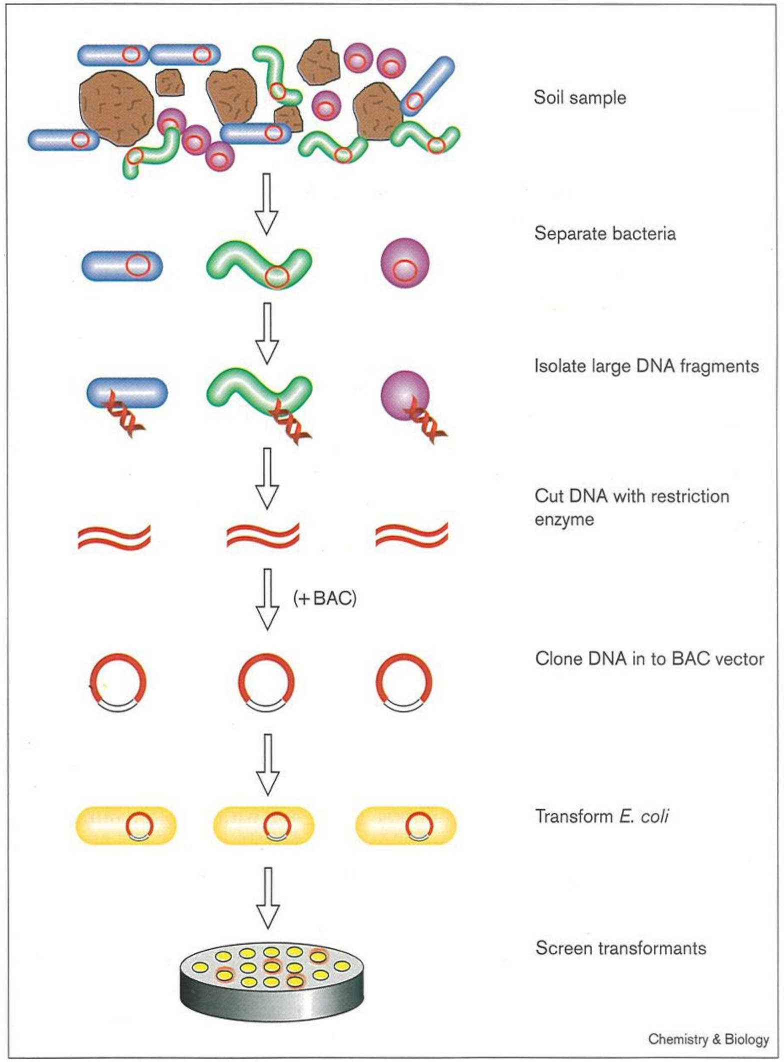


# Microorganims as a prolific source of natural products: an historical overview

## 1998

‘A new frontier of science is emerging that unites biology and chemistry - the **exploration of natural products from previously uncultured soil microorganisms.**
The methodology has been made possible by advances in molecular biology and eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we **term the metagenome of the soil**.’

*Chemistry & Biology* **1998,** 5, 245-249

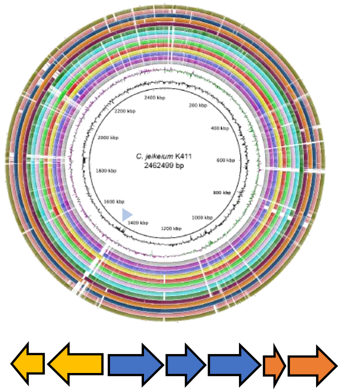


# Bioinformatic analysis of natural product biosynthetic capacity

## The emergence of a new core discipline - the genome mining

**2000**

Genome sequencing and mining



Recovery of complete BGCs

Renaissance and Genome-guided discovery of new molecules

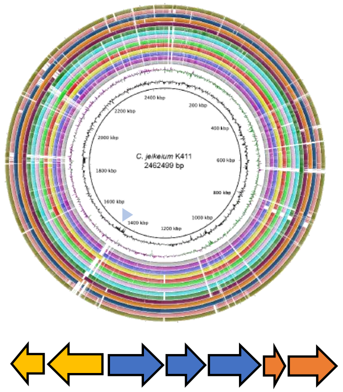
In the early 2000s, the sequencing of the first *Streptomyces* bacterial genomes revealed that the vast majority of small molecules produced by microbes had yet to be discovered, thus opening the door for future discovery efforts and for the emergence of a new core discipline – the **genome mining.**

## Bioinformatic analysis of natural product biosynthetic capacity

### The emergence of a new core discipline - the genome mining

**2000**

Genome sequencing and mining



Recovery of complete BGCs

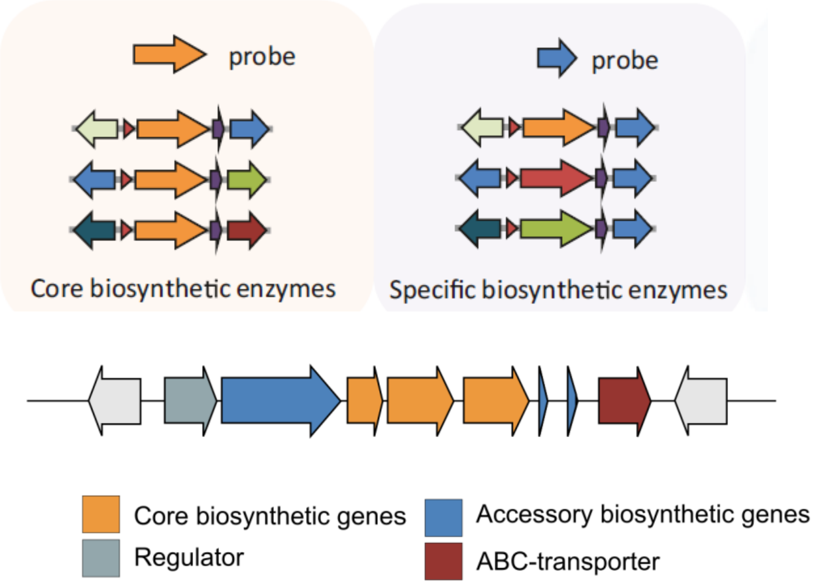
Renaissance and Genome-guided discovery of new molecules

**Genome mining** describes the targeted bioinformatic analysis of (meta-)genomes to identify gene clusters involved in the biosynthesis of NPs.

# Bioinformatic analysis of natural product biosynthetic capacity

## What makes BGCs being easily detected by Genome Mining?

While NPs are chemically diverse, their biosynthetic machineries **are often highly conserved**. Core biosynthetic enzymes are characterized by **high amino-acid sequence similarity**, which allows screening of genomic data for the presence of **specific biosynthetic genes that encode the required enzymatic activity**.


probe
probe
Core biosynthetic enzymes
Specific biosynthetic enzymes
Core biosynthetic genes
Accessory biosynthetic genes
Regulator
ABC-transporter

*Nat Commun* **2021**, 12, 3864

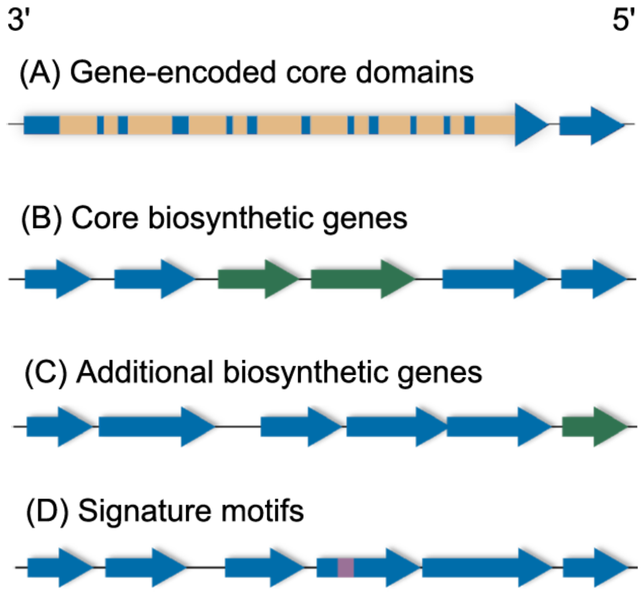
# Bioinformatic analysis of natural product biosynthetic capacity

## What makes BGCs being easily detected by Genome Mining?

Bacterial biosynthetic genes are **clustered on a contiguous stretch of DNA (co-localized).**

**Existence of core biosynthetic genes –** highly conserved across different organisms (e.g. KS and A domains)

Similar **genetic pattern organization** – core genes, regulatory genes and transporters.



3'

5'

(A) Gene-encoded core domains

(B) Core biosynthetic genes

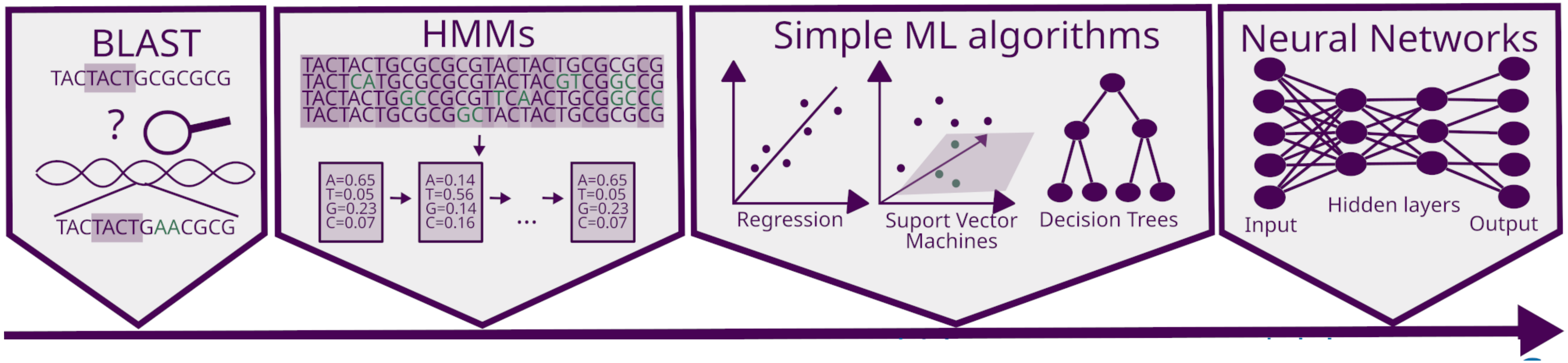
(C) Additional biosynthetic genes

(D) Signature motifs

*J. Org. Chem.* **2022,** *18,* 1656–1671.

# Bioinformatic analysis of natural product biosynthetic capacity

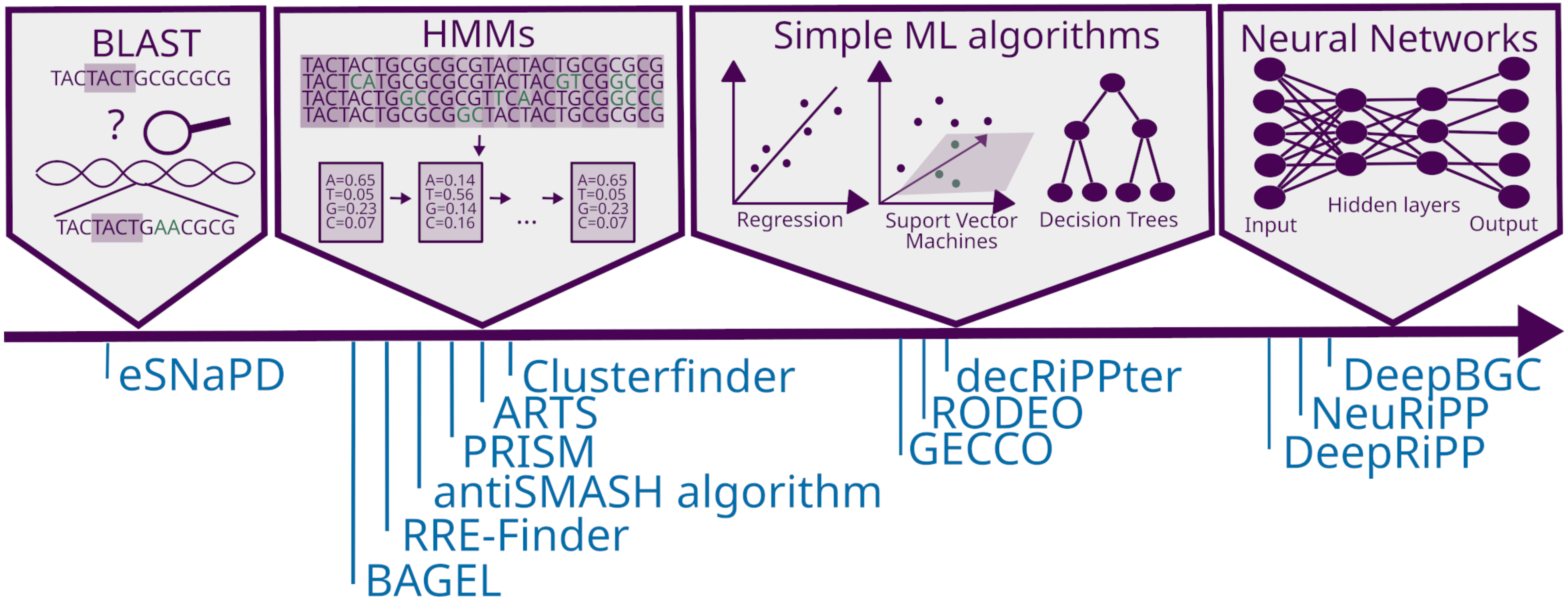
## Homology search tools/Algorithms for BGCs prediction



*J. Org. Chem.* **2022,** *18,* 1656–1671.

# Bioinformatic analysis of natural product biosynthetic capacity

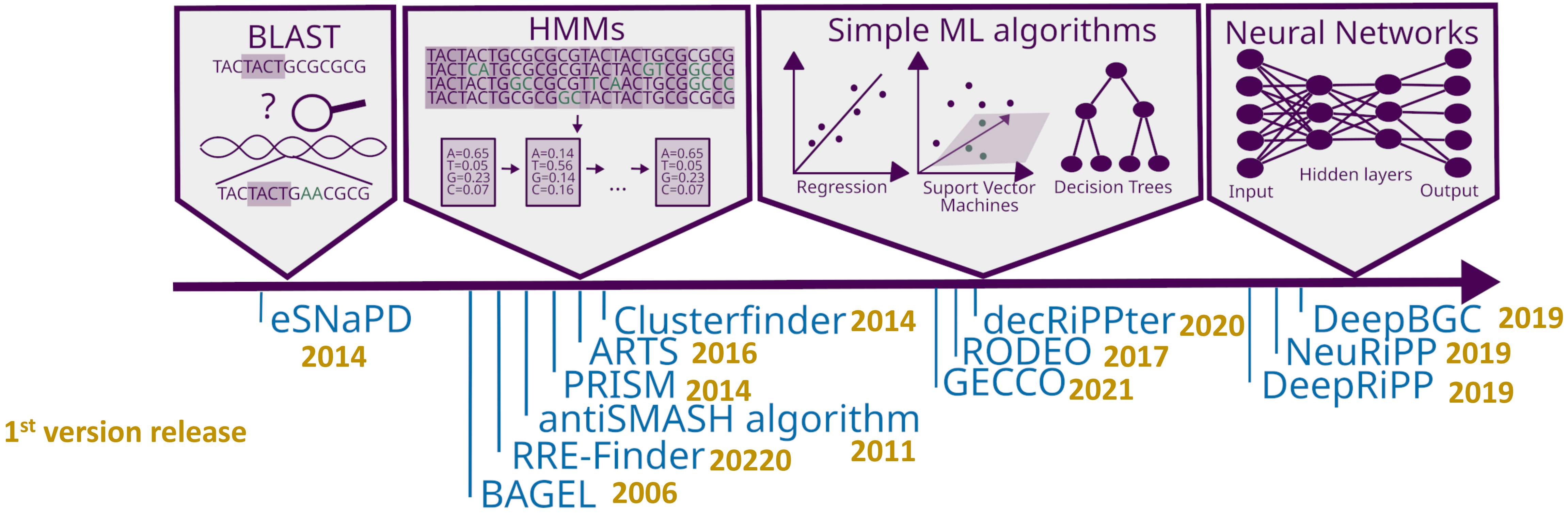
## Homology search tools/Algorithms for BGCs prediction



*J. Org. Chem.* **2022,** *18,* 1656–1671.

## Bioinformatic analysis of natural product biosynthetic capacity

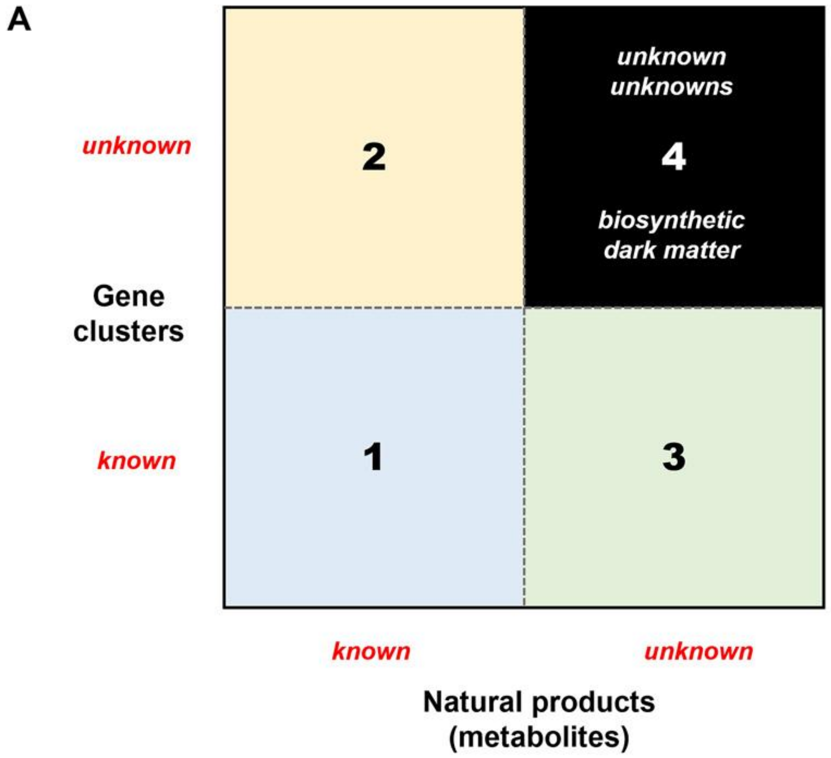
### Homology search tools/Algorithms for BGCs prediction



*J. Org. Chem.* **2022,** *18,* 1656–1671.

# Bioinformatic analysis of natural product biosynthetic capacity

## Homology search tools/Algorithms for BGCs prediction



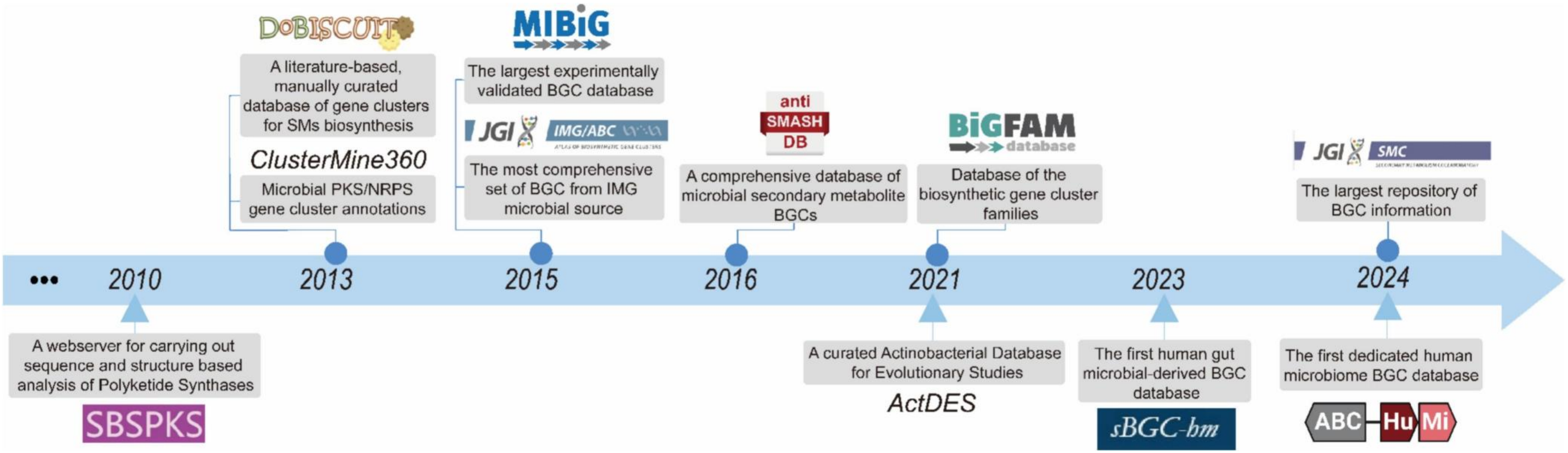
*Nat. Prod. Biosynth.* **2022**, RSC Publishing

Genome mining tools rely on existing datasets for accurate predictions.

Large databases are essential to train algorithms effectively.

# Bioinformatic analysis of natural product biosynthetic capacity

## Biosynthetic gene clusters (BGCs) databases



A webserver for carrying out sequence and structure based analysis of Polyketide Synthases

A literature-based, manually curated database of gene clusters for SMs biosynthesis

Microbial PKS/NRPS gene cluster annotations

The largest experimentally validated BGC database

The most comprehensive set of BGC from IMG microbial source

A comprehensive database of microbial secondary metabolite BGCs

Database of the biosynthetic gene cluster families

A curated Actinobacterial Database for Evolutionary Studies

The first human gut microbial-derived BGC database

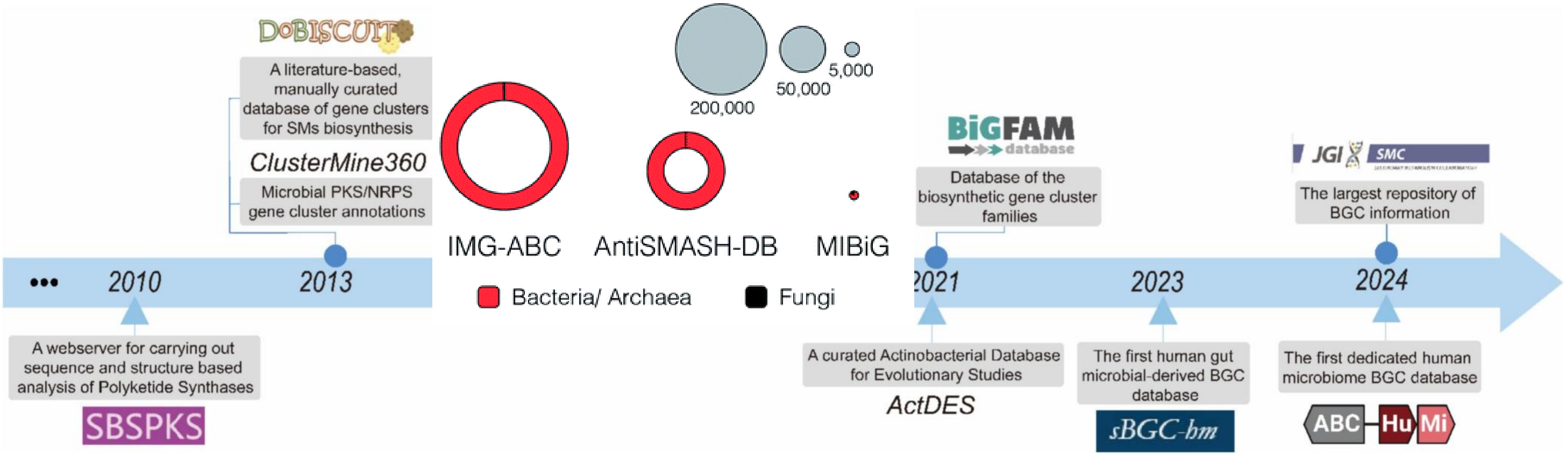
The largest repository of BGC information

The first dedicated human microbiome BGC database

*Biotechnol Adv*. **2025**, 79:108532.

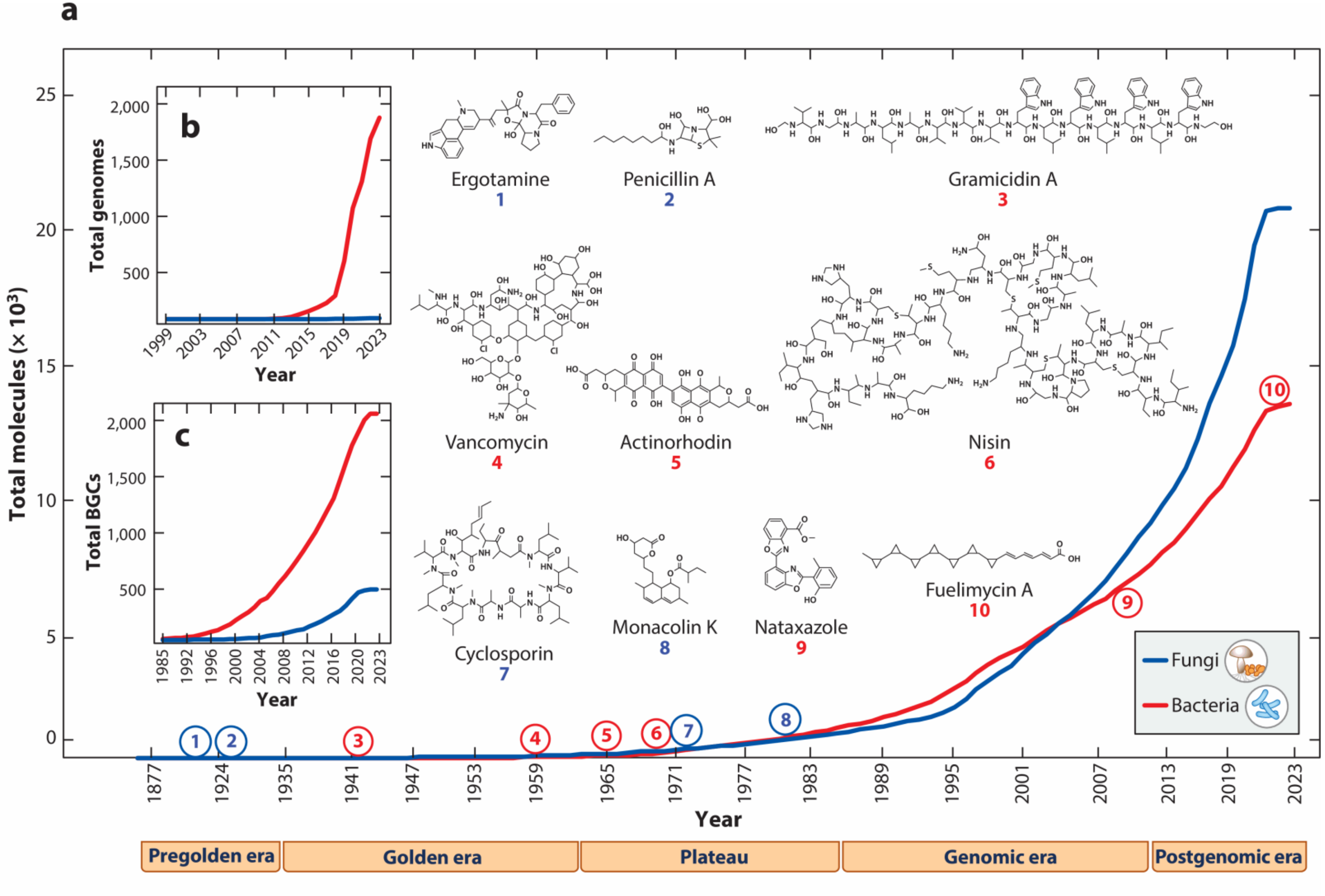
# Bioinformatic analysis of natural product biosynthetic capacity

## Biosynthetic gene clusters (BGCs) databases


DoBISCUIT
A literature-based, manually curated database of gene clusters for SMs biosynthesis
ClusterMine360
Microbial PKS/NRPS gene cluster annotations
200,000
50,000
5,000
IMG-ABC
AntiSMASH-DB
MIBiG
🟥 Bacteria/ Archaea
⬛ Fungi
BiGFAM database
Database of the biosynthetic gene cluster families
JGI SMC
The largest repository of BGC information
••• 2010
2013
2021
2023
2024
A webserver for carrying out sequence and structure based analysis of Polyketide Synthases
SBSPKS
A curated Actinobacterial Database for Evolutionary Studies
ActDES
The first human gut microbial-derived BGC database
sBGC-hm
The first dedicated human microbiome BGC database
ABC-HuMi

*Biotechnol Adv*. **2025**, 79:108532.

# Mapping the biosynthetic gene clusters diversity



*Annu. Rev. Biochem.* **2024**. 93:411–45

# Mapping the biosynthetic gene clusters diversity

**BiG-SCAPE** - Biosynthetic Gene Similarity Clustering and Prospecting Engine

**BiG-SLICE** - Biosynthetic Gene clusters - Super Linear Clustering Engine
Is a powerful tool for analyzing BGCs at scale, enabling clustering of millions of BGCs based on sequence similarity.

**BIG-FAM database**
Is an online repository for "homologous" groups of biosynthetic gene clusters (BGCs) putatively encoding the production of similar specialized metabolites. BiG-FAM facilitates querying putative BGCs to rapidly **find their position on the diversity map and gain a better understanding of their novelty or (probable) functions**, based on relationships with other known and predicted BGCs from publicly available data.

# Mapping the biosynthetic gene clusters diversity

**BiG-SCAPE -** Biosynthetic Gene Similarity Clustering and Prospecting Engine

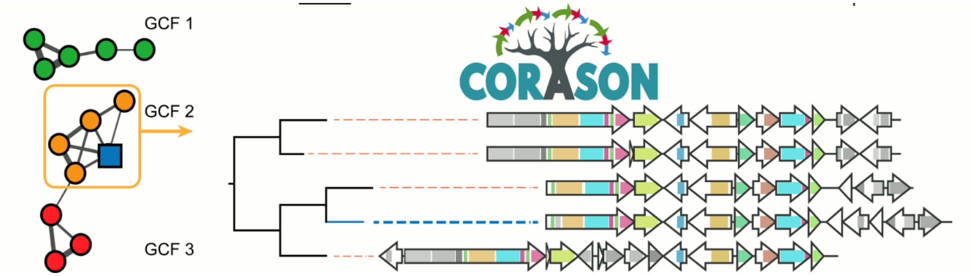
| | |
|---|---|
| Introduction to antiSMASH and BiG-SCAPE workflows | Catarina Loureiro |
| | |
| Hands- on session antiSMASH and BiG-SCAPE | Catarina Loureiro |

# Mapping the biosynthetic gene clusters diversity

**CORASON** - CORe Analysis of Syntenic Orthologs to prioritize Natural Product-Biosynthetic Gene Cluster

**cblaster** - a remote search tool for rapid identification and visualization of homologous gene clusters

**clinker** - a pipeline for easily generating publication-quality gene cluster comparison figures.

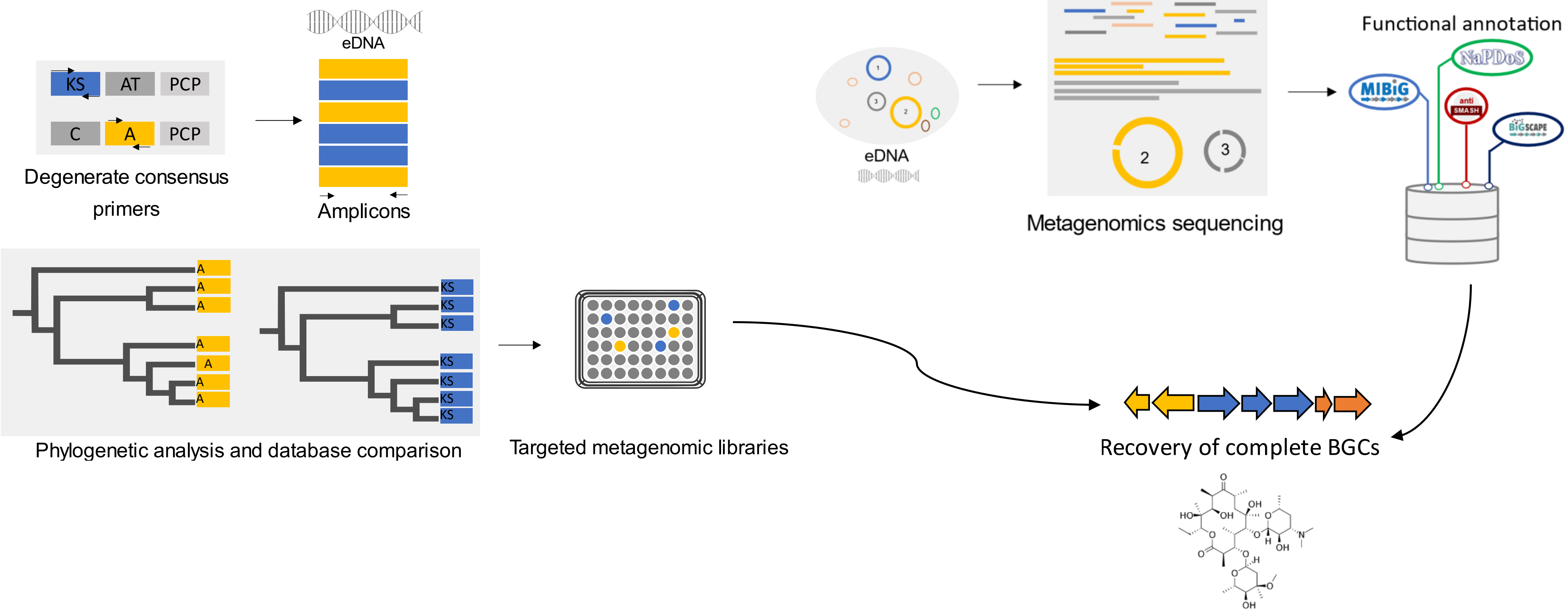


| | | |
|---|---|---|
| 11h00 30'+45' | Genomic context of target biosynthetic genes | Adriana Rego |
| | (short) Hands-on session genomic context | Adriana Rego and Catarina Loureiro |

# A new era of metagenomics-driven NPs discovery

## Targeted

## Untargeted



Degenerate consensus primers

Amplicons

Phylogenetic analysis and database comparison

Targeted metagenomic libraries

Metagenomics sequencing

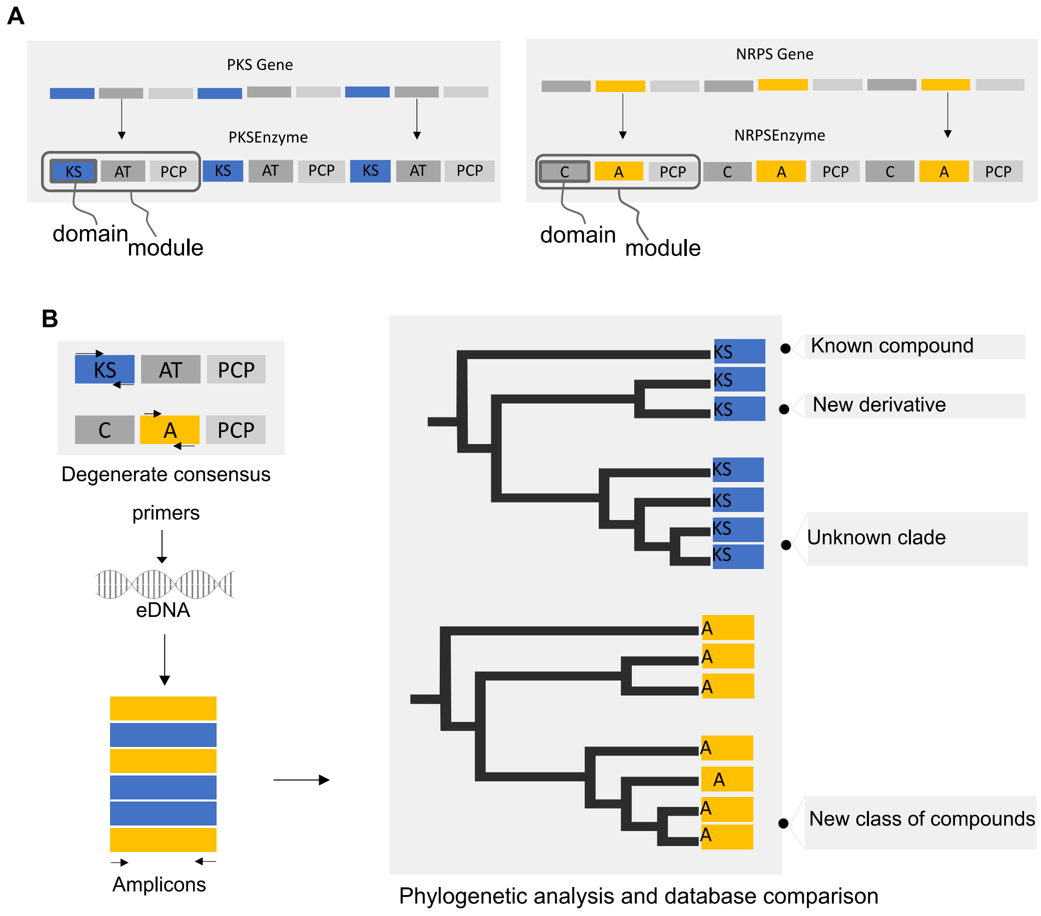
Functional annotation

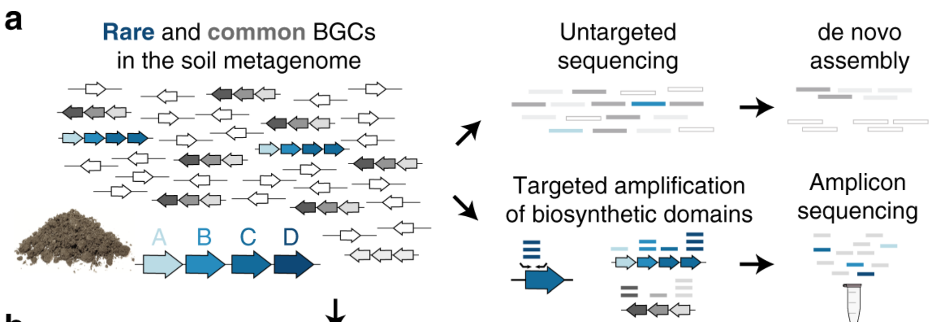
Recovery of complete BGCs

Expression, isolation and characterization of NPs

# A new era of metagenomics-driven NPs discovery

## Amplicon-based or targeted approaches





**Main applications:**

Poorly assembled genomes

Amplicons

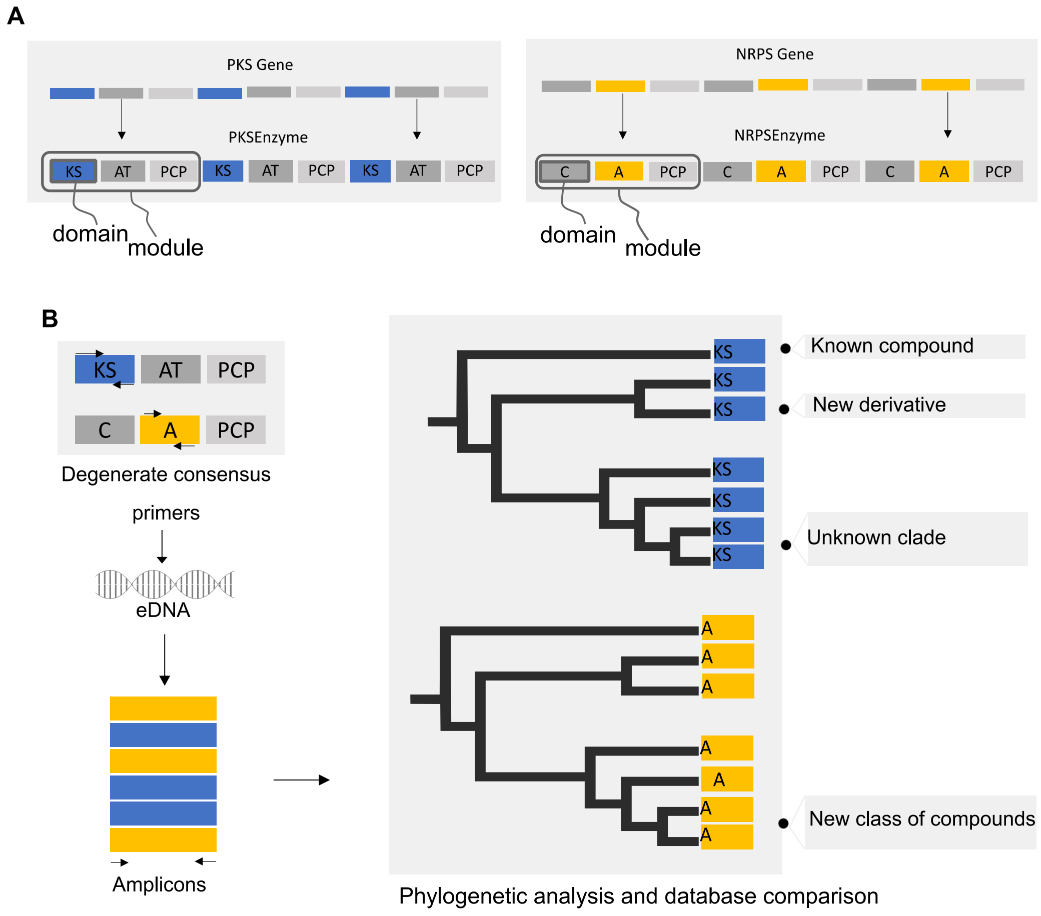
Metagenomes

**Advantage** - quick estimate biosynthetic potential

New clades = new functionality

# A new era of metagenomics-driven NPs discovery

## Amplicon-based or targeted approaches

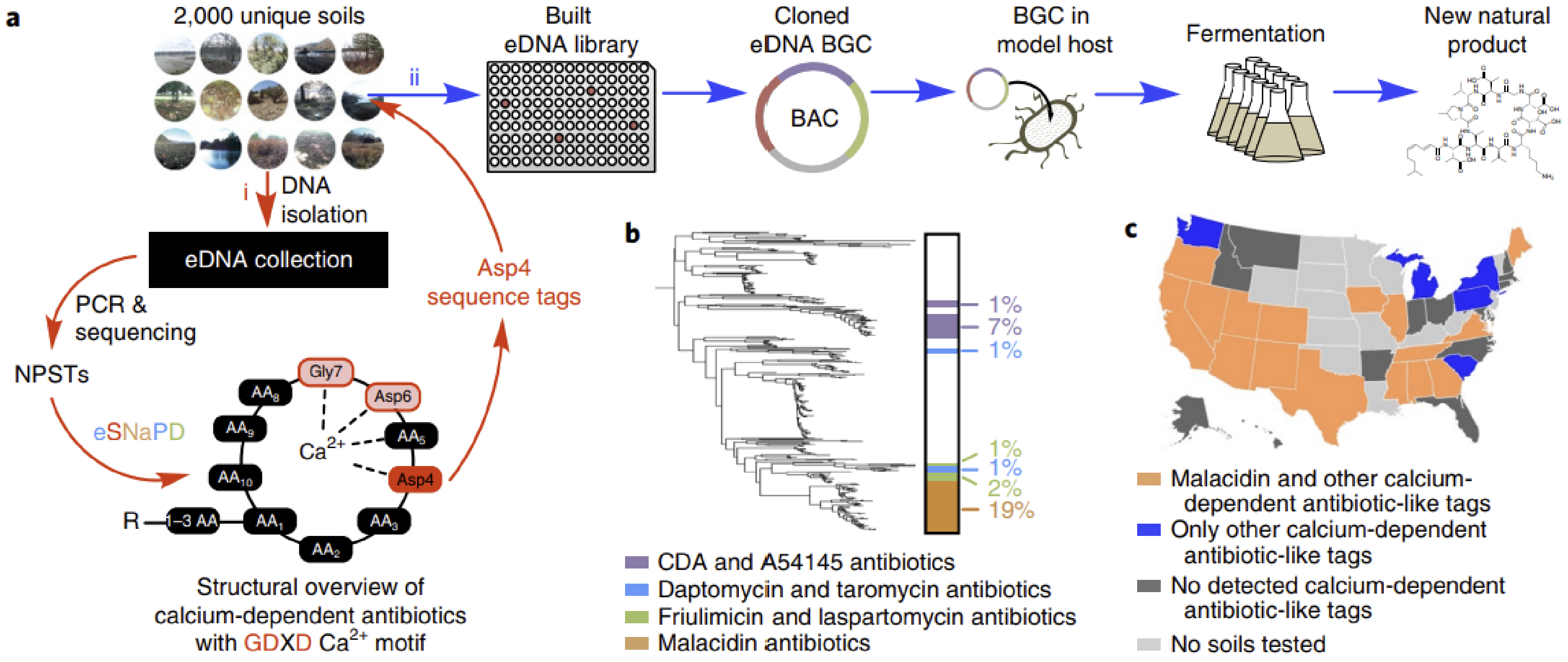


**Hans Singh**
(University of Hawai'i at Mānoa, USA)

| Time | Duration | Topic |
|---|---|---|
| 9h00 | 60'+30' | Phylogenetic approaches to natural product discovery<br>(short) Hands-on session NaPDos2 |

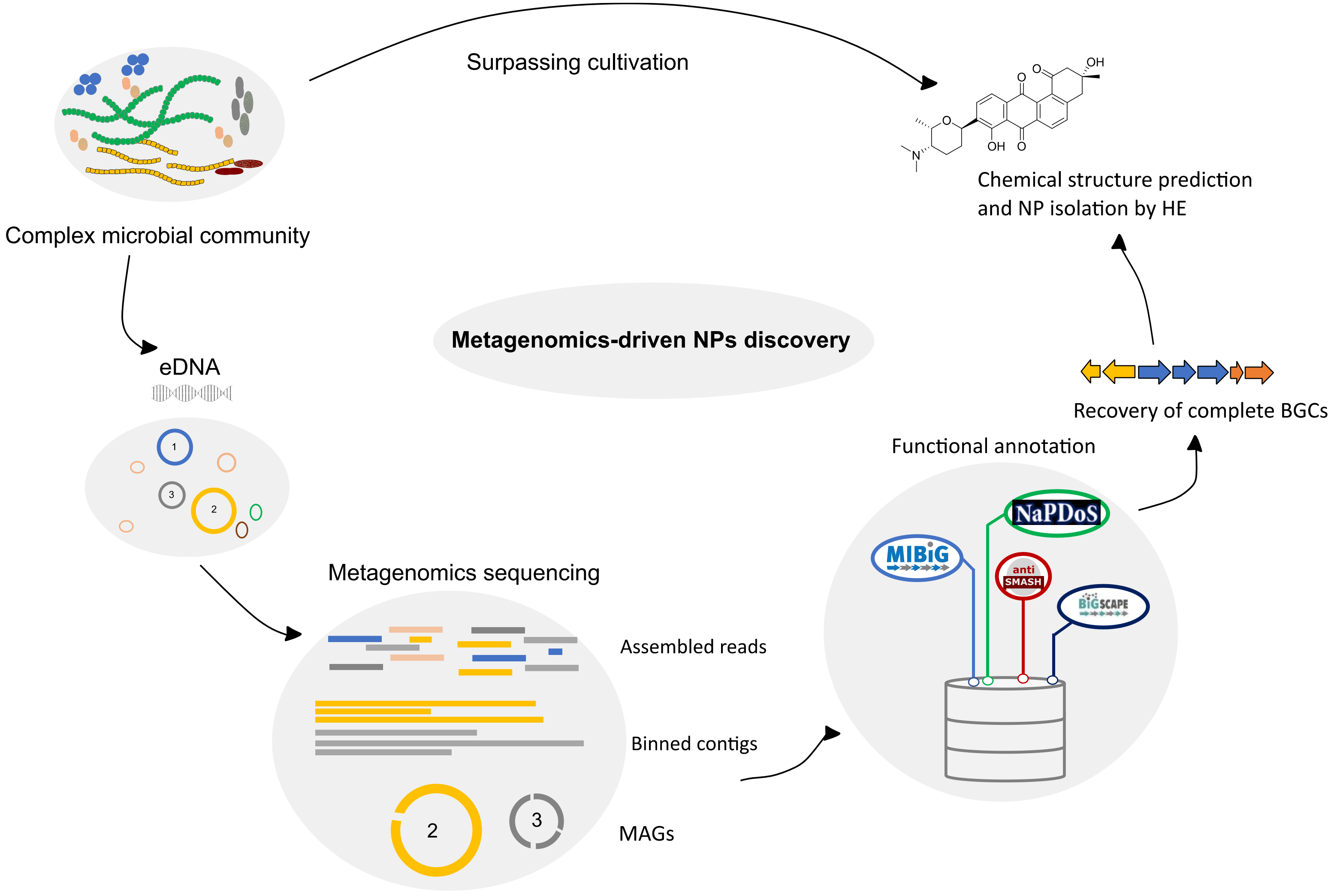
# Bioinformatic analysis of natural product biosynthetic capacity

## Amplicon-based or targeted approaches


a
2,000 unique soils
ii
Built eDNA library
Cloned eDNA BGC
BAC
BGC in model host
Fermentation
New natural product
i
DNA isolation
eDNA collection
PCR & sequencing
NPSTs
eSNaPD
Asp4 sequence tags
Gly7
Asp6
AA8
AA9
AA5
Ca2+
AA10
Asp4
R
1–3 AA
AA1
AA2
AA3
Structural overview of calcium-dependent antibiotics with GDXD Ca2+ motif
b
1%
7%
1%
1%
1%
2%
19%
CDA and A54145 antibiotics
Daptomycin and taromycin antibiotics
Friulimicin and laspartomycin antibiotics
Malacidin antibiotics
c
Malacidin and other calcium-dependent antibiotic-like tags
Only other calcium-dependent antibiotic-like tags
No detected calcium-dependent antibiotic-like tags
No soils tested

*Nat Microbiol* **2018**, **3**, 415–422

# A new era of metagenomics-driven NPs discovery



Metagenomics-driven NPs discovery

Complex microbial community

Surpassing cultivation

Chemical structure prediction and NP isolation by HE

eDNA

Metagenomics sequencing

Assembled reads

Binned contigs

MAGs

Functional annotation

Recovery of complete BGCs

# A new era of metagenomics-driven NPs discovery

©Helena Klein

Article

## Biosynthetic potential of the global ocean microbiome

*Nature* **2022** 607, 111–118

Article

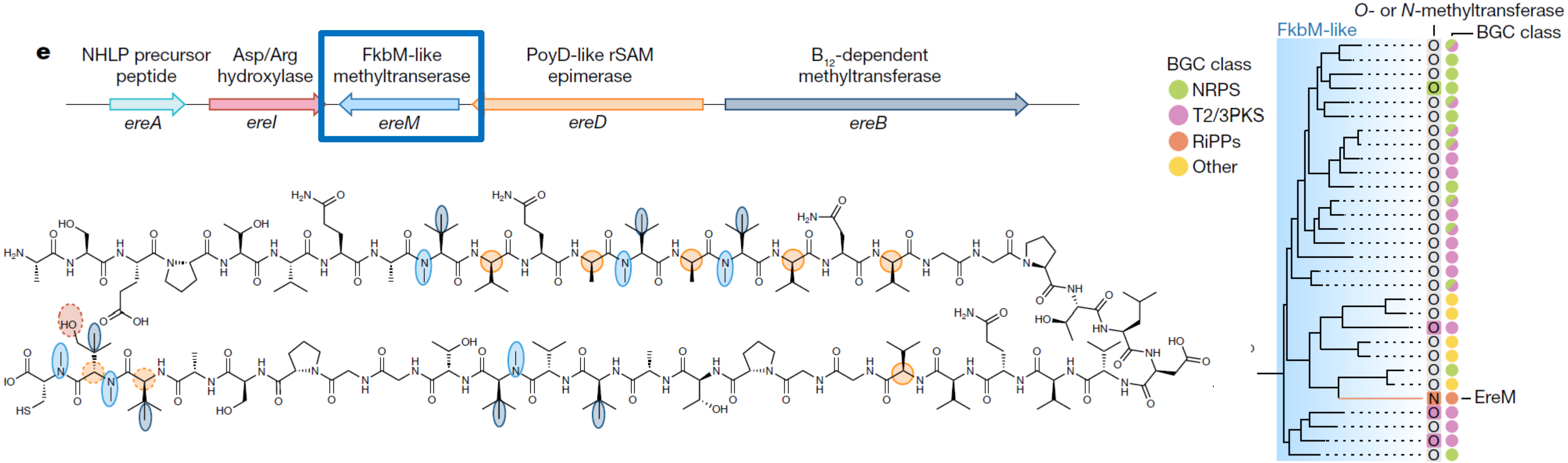
## Global marine microbial diversity and its potential in bioprospecting

*Nature* **2024** 633, 371–379

# A new era of metagenomics-driven NPs discovery

**Are existent computational approaches sufficiently powerful to predict new enzymology and natural products ?**

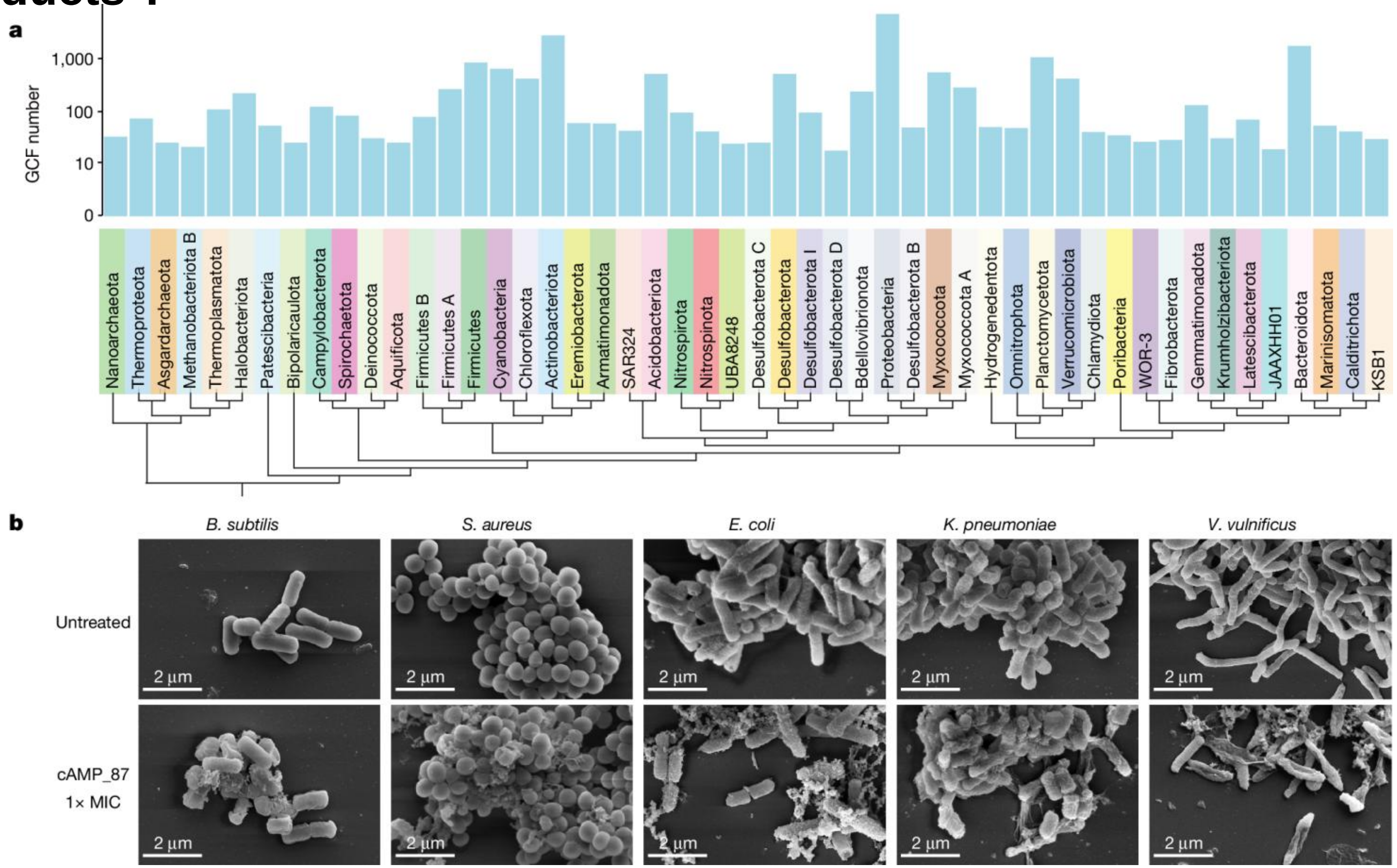
unique trans amide-N-methylation



*Nature* **2022** 607, 111–118

# A new era of metagenomics-driven NPs discovery

**Are existent computational approaches sufficiently powerful to predict new enzymology and natural products ?**



*Nature* **2024** 633, 371–379

# Outlook

We still only grasp a small diversity of the existent **biosynthetic potential** and chemical diversity, including the understanding of their **ecological functions and applications**.

Genome mining is a **continuously evolving field,** although current sequence homology- and ML-based computational approaches are capable of identifying new biosynthetic pathways and enzymes, that share even low levels of similarity with known BGCs.

Future bioprospection is **environmental and metagenomics-driven**.