Meta_Microbial Workshop

Metagenomic and bioinformatic insights into microbial communities

Untargeted Metagenomics - Taxa and Functional Annotation

@AdrianaRego10

adrianairego@gmail.com

ciimar

U.PORTO

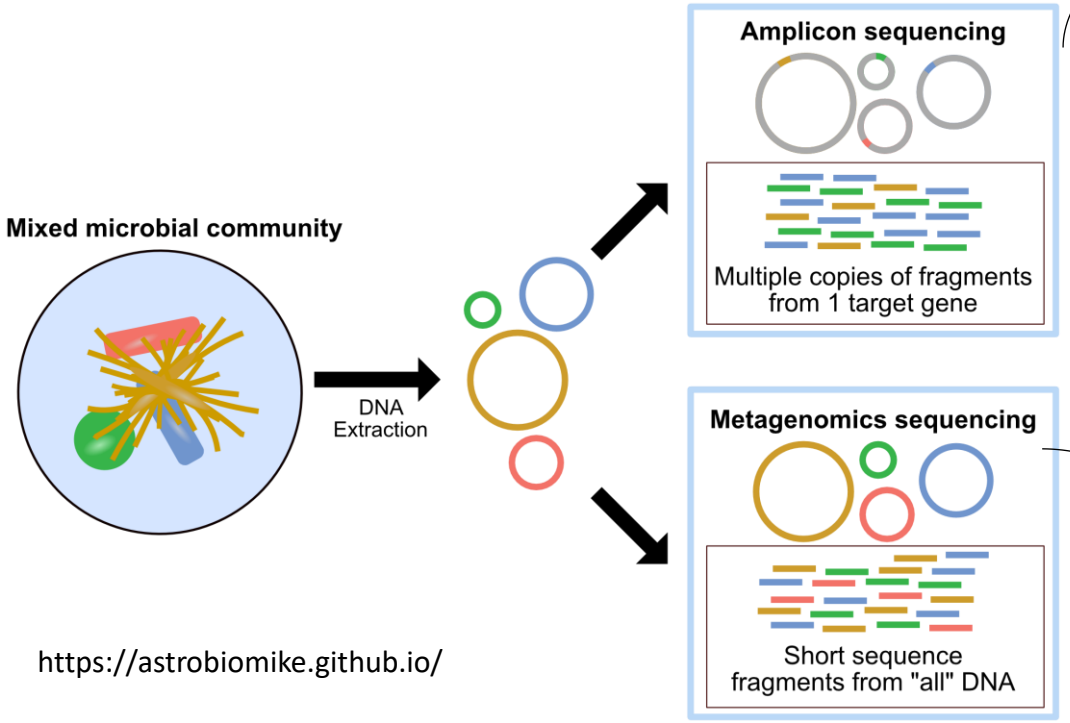
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Sponsorship

SPECO

BIOPORTUGAL S.A.
Químico, Farmacêutica

Amplicon vs shotgun metagenomics



https://astrobiomike.github.io/

PCR amplification of marker genes
Quick and cheap analysis
Applicable to samples contaminated with host DNA
Limitations
Only functional predictions (e.g. PICRUST)
Limited taxonomic resolution
PCR and primer bias

Sequencing of short/long random DNA fragments
Functional potential
Uncultured genomes
Taxonomic resolution (species/strain level)
Limitations
Expensive, time-consuming and computationally intense analysis
Host contamination

Amplicon vs shotgun metagenomics

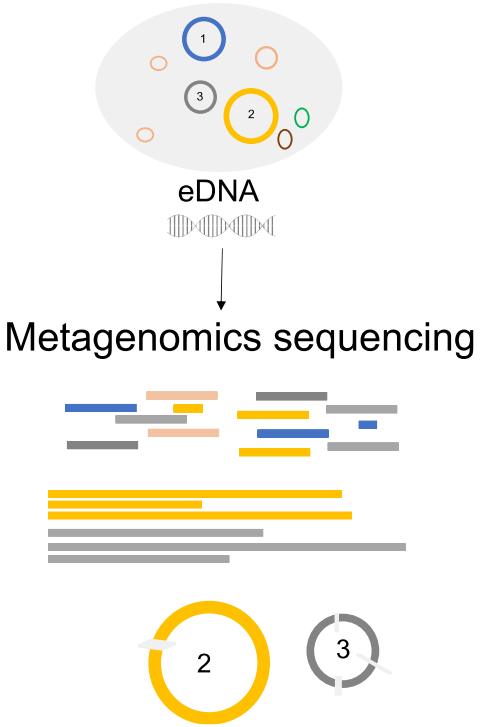
Shotgun metagenomics

Strengths

- High taxonomical resolution (species/strains)
- Functional metagenomics (genes)
- Comparative genomics

Challenges and limitations

- Completeness of databases and good references
- DNA extraction bias (cell-lysis efficiency)
- No information on expression of genes (transcriptomics instead)
- Complexity of ecosystems and sequencing depth
- Contamination



1 Introduction

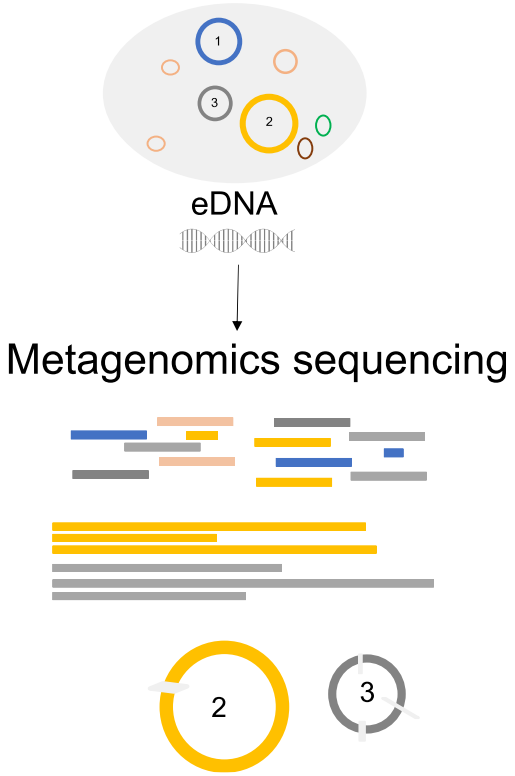
Shotgun sequencing metagenomics

Which questions can we answer using metagenomics data?

Who is there? - Taxonomical annotation

What can they do? - Functional annotation

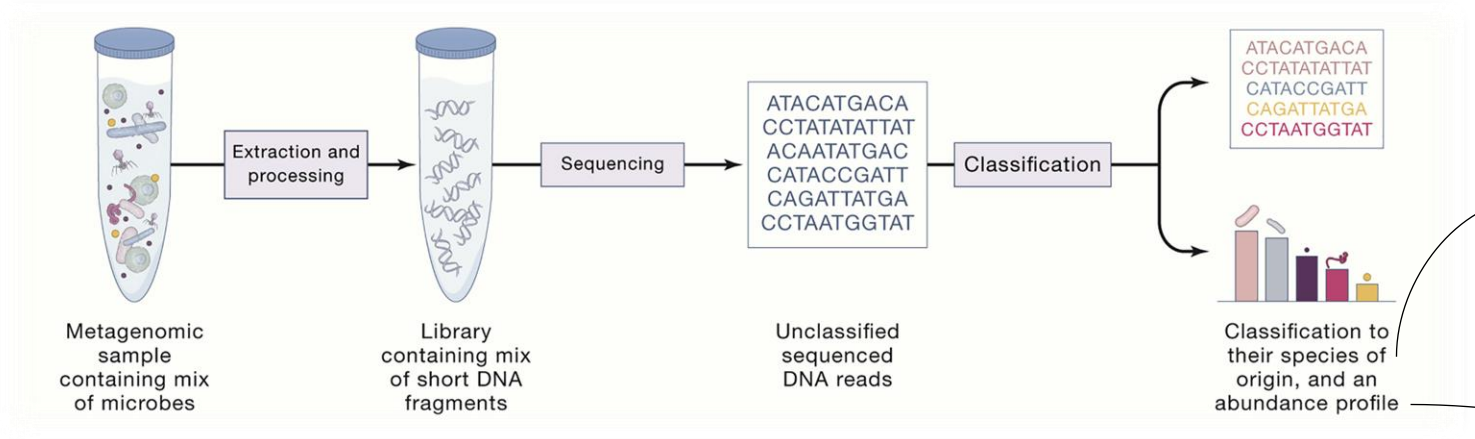
And not “What are they doing?” -> Transcriptomics



2 **Taxonomical annotation**

Who is there?

MAGs Taxonomic Classification and Profiling



Taxonomic classification/binning

Assigns taxon labels to individual sequences and results in taxon-specific sequence bins (and sequence abundance profiles).

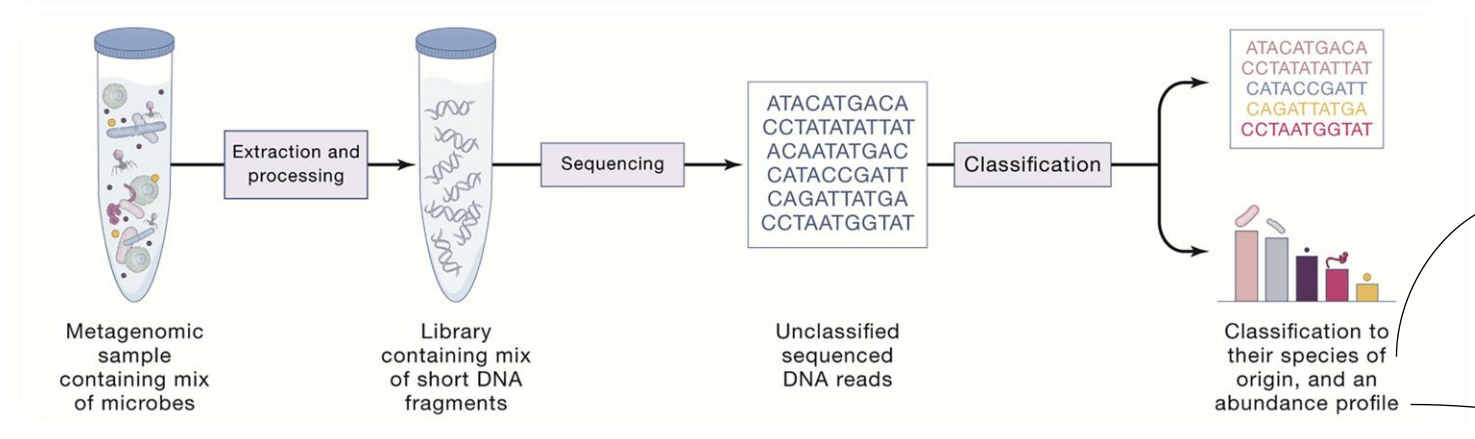
Taxonomic profiling

Quantify the presence and relative taxon abundances of microbial communities from metagenome samples.

2 **Taxonomical annotation**

Who is there?

MAGs Taxonomic Classification and Profiling



Taxonomic classification/binning

Assigns taxon labels to individual sequences and results in taxon-specific sequence bins (and sequence abundance profiles).

Taxonomic profiling

Quantify the presence and relative taxon abundances of microbial communities from metagenome samples.

2

Taxonomical annotation

Who is there?

Taxonomic classifiers

Traditional methods based on **16S rRNA** applied to metagenomics - limited resolution and 16S rRNAs can be poorly represented in MAGs.
Instead the use of several **single-copy marker genes** improves resolution.

Identification of marker genes in MAGs/metagenomes

MSA with marker genes from reference databases

Phylogenetic tree
ANI – average nucletide identity

2 **Taxonomical annotation**

Who is there?

Taxonomic classifiers

Traditional methods based on **16S rRNA** applied to metagenomics - limited resolution and 16S rRNAs can be poorly represented in MAGs.
Instead the use of several **single-copy marker genes** improves resolution.

Identification of marker genes in MAGs

MSA with marker genes from reference databases

Phylogenetic tree
ANI – average nucletide identity

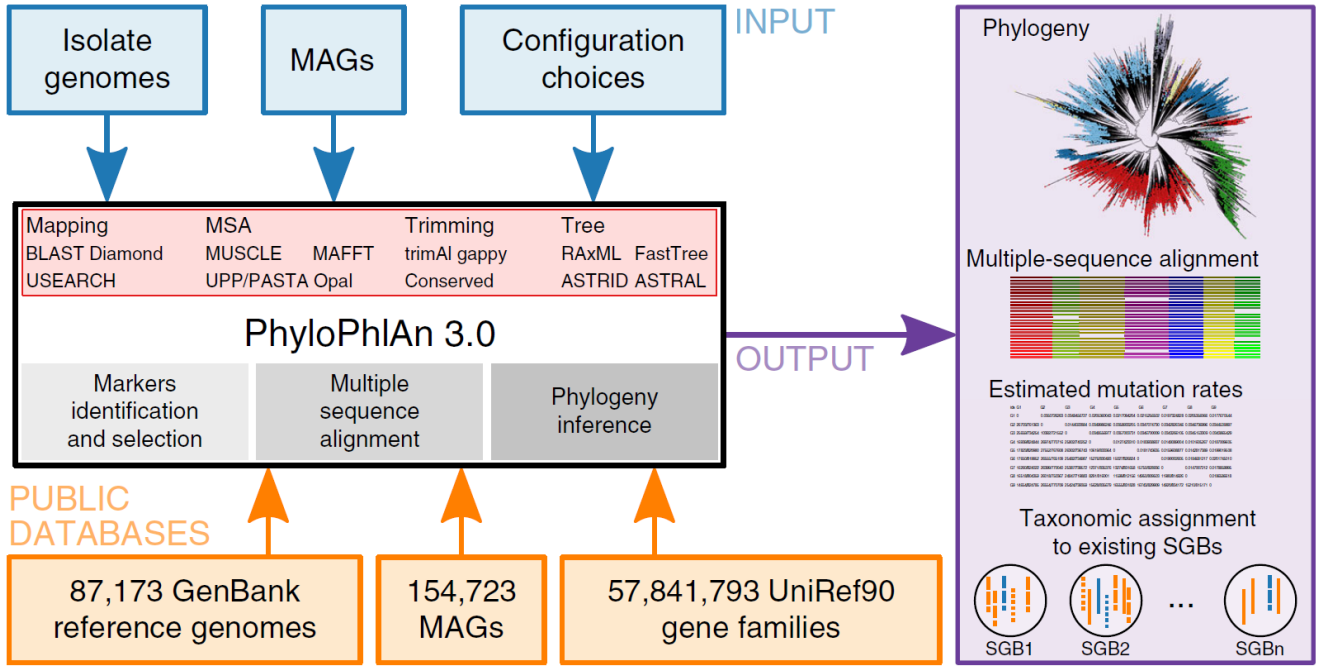
Calculation of average nucleotide identity (ANI) between genomes is used for **species identification**. ANI >96% - species
ANI of <93% - species differentiation.

2 **Taxonomical annotation**

Who is there?

Taxonomic classifiers

PhyloPhlAn - https://huttenhower.sph.harvard.edu/phylophlan/ is an integrated pipeline for large-scale phylogenetic profiling of genomes and metagenomes. PhyloPhlAn is an accurate, rapid, and easy-to-use method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution. PhyloPhlAn can assign both genomes and metagenome-assembled genomes (MAGs) to species-level genome bins (SGBs).



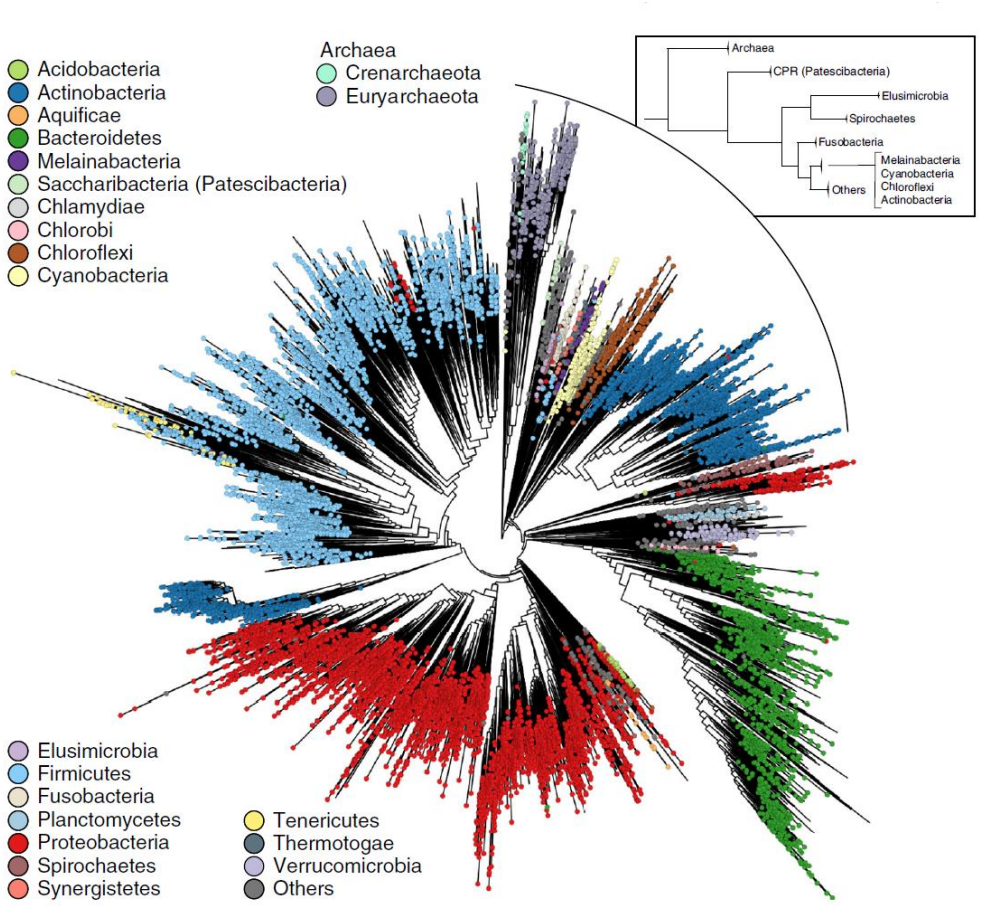
(Adapted from Asnicar et al. 2020)

2 **Taxonomical annotation**

Who is there?

Taxonomic classifiers

PhyloPhlAn - https://huttenhower.sph.harvard.edu/phylophlan/ is an integrated pipeline for large-scale phylogenetic profiling of genomes and metagenomes. PhyloPhlAn is an accurate, rapid, and easy-to-use method for large-scale microbial genome characterization and phylogenetic analysis at multiple levels of resolution. PhyloPhlAn can assign both genomes and metagenome-assembled genomes (MAGs) to species-level genome bins (SGBs).

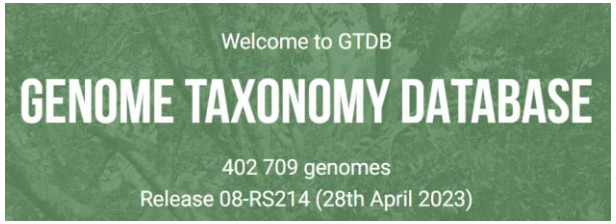


PhyloPhlAn 3.0 microbial tree-of-life with 17,672 species-representative genomes from 51 known and 84 candidate phyla.

(Adapted from Asnicar et al. 2020)

2 **Taxonomical annotation**

Who is there?



Taxonomic classifiers

GTDB-Tk https://github.com/Ecogenomics/GTDBTk

- is a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes based on the Genome Database Taxonomy.
- command line version

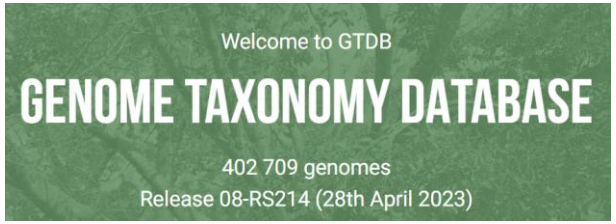
GTDB R214 spans 402,709 genomes organized into 85,205 species clusters.

	Bacteria	Archaea	Total
Phylum	161	20	181
Class	488	60	548
Order	1,624	148	1,772
Family	4,262	508	4,772
Genus	19,153	1,586	20,739
Species	80,789	4,416	85,205

```
gtdbtk classify_wf --genome_dir All_bins/ --out_dir
All_bins_gtdbk_output -x fa --cpus 50
```

2 **Taxonomical annotation**

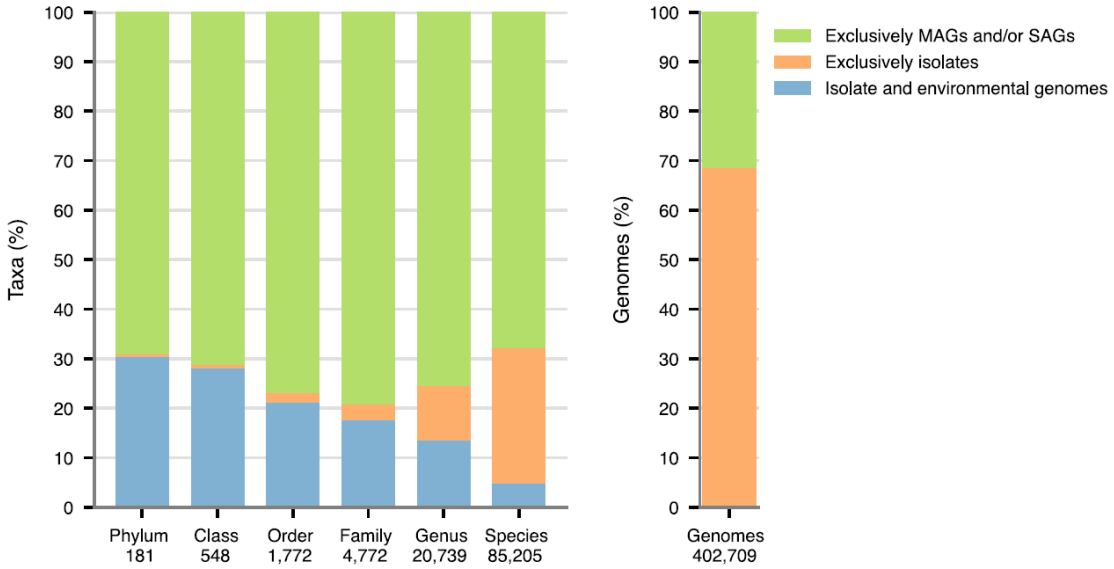
Who is there?



Taxonomic classifiers

GTDB-Tk https://github.com/Ecogenomics/GTDBTk

- is a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes based on the Genome Database Taxonomy.
- command line version



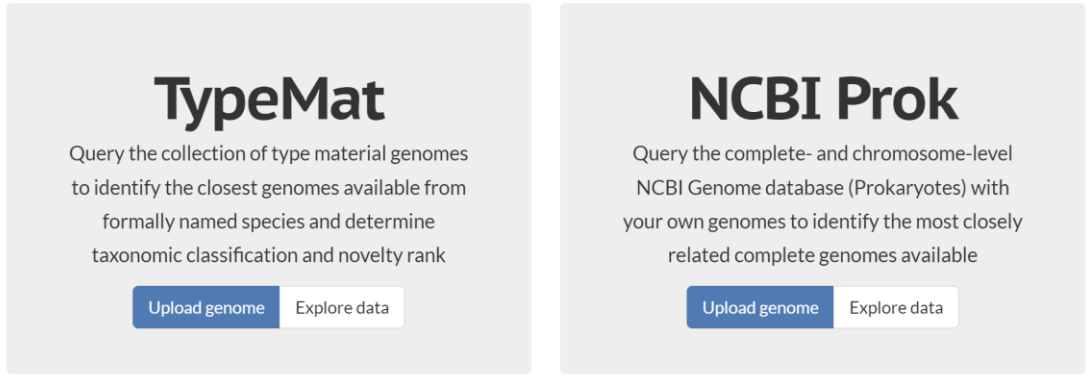

```
gtdbtk classify_wf --genome_dir All_bins/ --out_dir
All_bins_gtdbk_output -x fa --cpus 50
```

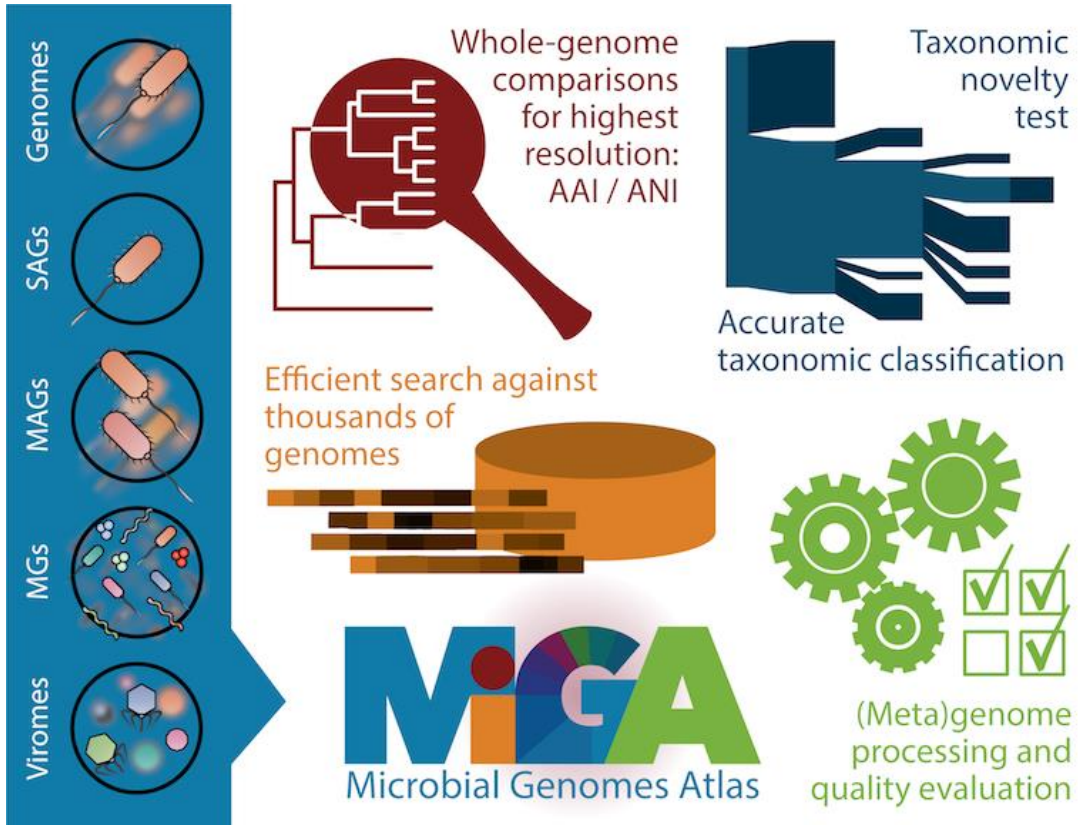
2 **Taxonomical annotation**

Who is there?

Taxonomic classifiers

Microbial Genome Atlas
http://microbial-genomes.org/





2 **Taxonomical annotation**

Who is there?

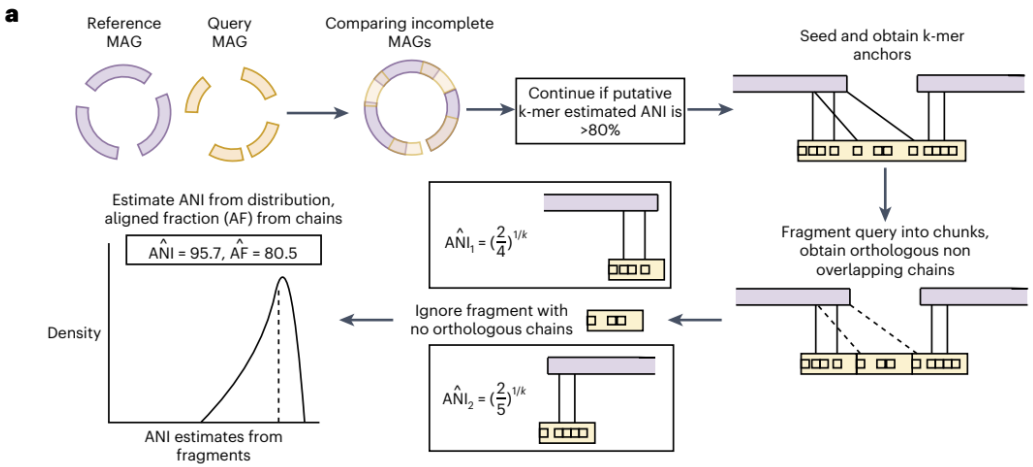
Taxonomic classifiers

ANI tools

Comparison of ANI against a specific (even a created) database

FastANI - https://github.com/ParBLiSS/FastANI

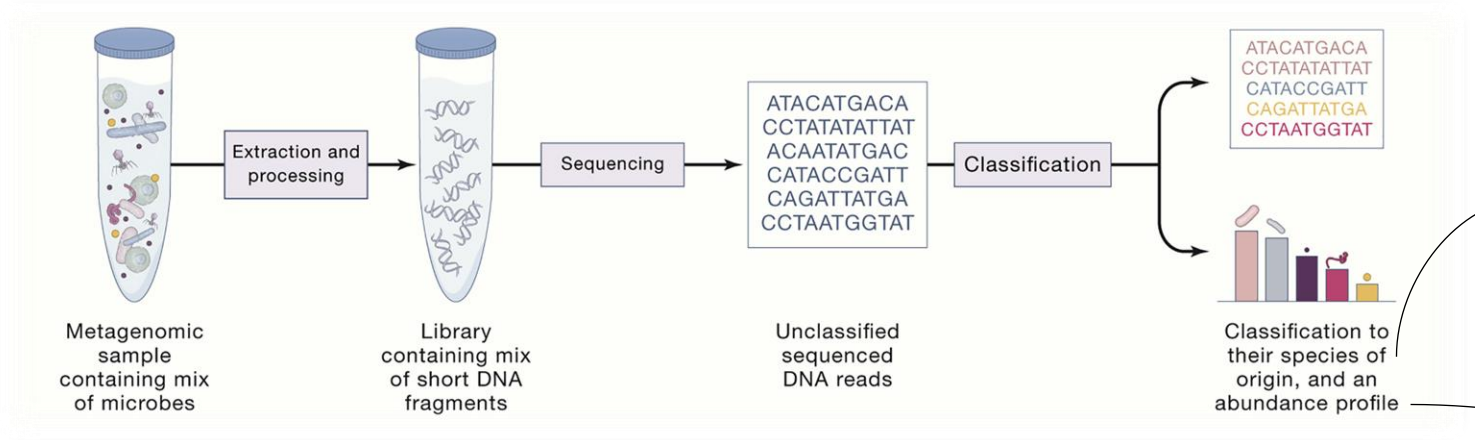
Skani – https://github.com/bluenote-1577/skani
performs better for incompete MAGs (e.g. for metagenor environmental studies)



2 **Taxonomical annotation**

Who is there?

MAGs Taxonomic Classification and Profiling



Taxonomic classification/binning

Assigns taxon labels to individual sequences and results in taxon-specific sequence bins (and sequence abundance profiles).

Taxonomic profiling

Quantify the presence and relative taxon abundances of microbial communities from metagenome samples.

2 **Taxonomical annotation**

Who is there?

Metagenomic profilers

Metagenomic profilers can be categorized based on their **reference database type**.

Approaches for Taxonomic classification and profiling

Reference database	DNA to DNA	DNA to protein	Marker gene based
	DNA-to-DNA Complete genomes as reference	DNA-to-protein Coding DNA as reference	DNA-to-marker Single copy markers as reference
Software	Kraken 2	Diamond Kaiju	MetaPhlan mOTUs

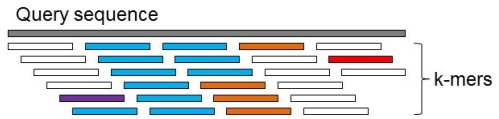
2 **Taxonomical annotation**

Who is there?

Metagenomic profilers

The **comparison** of reads to database sequences can be performed by different strategies.

Approaches for Taxonomic classification and profiling

	K-mer based	Marker gene based	Genome based
	Very fast Lack sensitivity Big fraction unclassified	Estimation of species abundance Low detection accuracy and that the unclassified percentage is unknown	High detection accuracy, unclassified percentage is known High-resolution genomic comparisons are possible
Software	 **Kraken 2**	**MetaPhlan**	**GTDB-Tk**

2 **Taxonomical annotation**

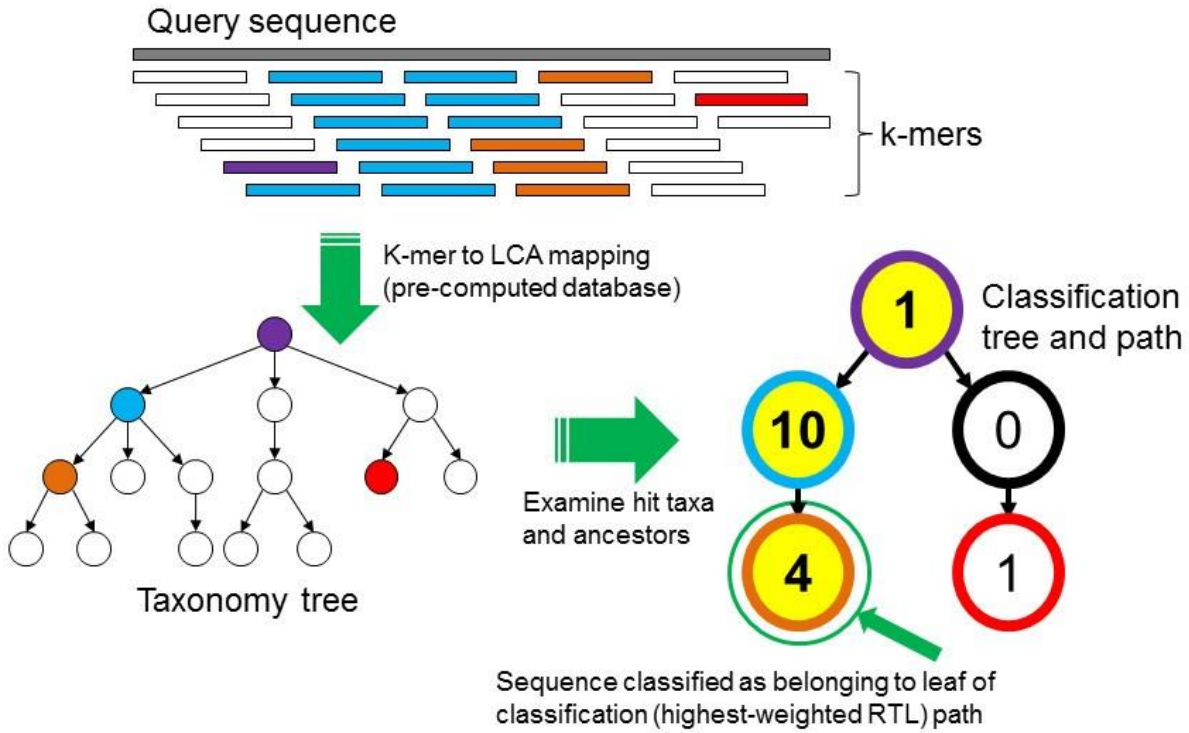
Who is there?

Kraken 2 - https://ccb.jhu.edu/software/kraken2/

Kraken is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies.

Reference database – DNA to DNA
Software – kmer-based

***MiniKraken**: To allow users with low-memory computing environments



Classification

To classify a set of sequences (reads), use the `kraken` command:

```
kraken --db $DBNAME seqs.fa
```

2 **Taxonomical annotation**

Who is there?

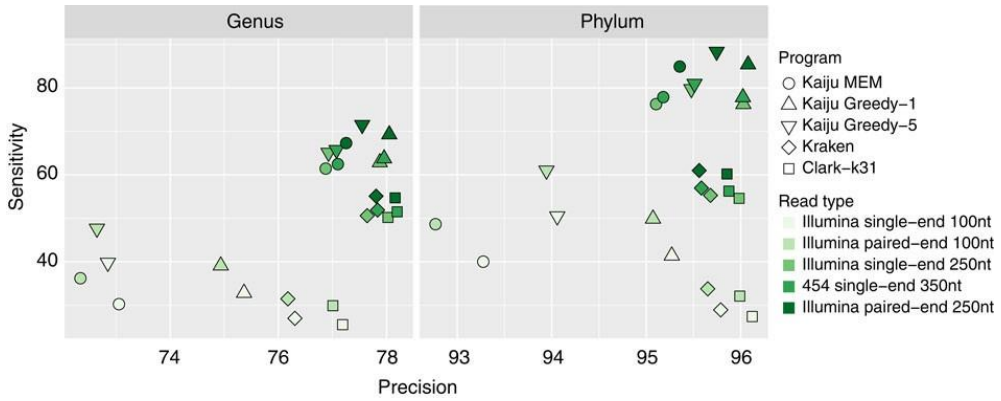
Kaiju - https://bioinformatics-centre.github.io/kaiju/

Kaiju is a program for fast and sensitive taxonomic classification of high-throughput sequencing reads from metagenomic, whole genome sequencing or metatranscriptomics experiments.

Reference database – DNA to protein

KAIJU

Fast and sensitive taxonomic classification for metagenomics



(Adapted from Menzel et al. 2016)

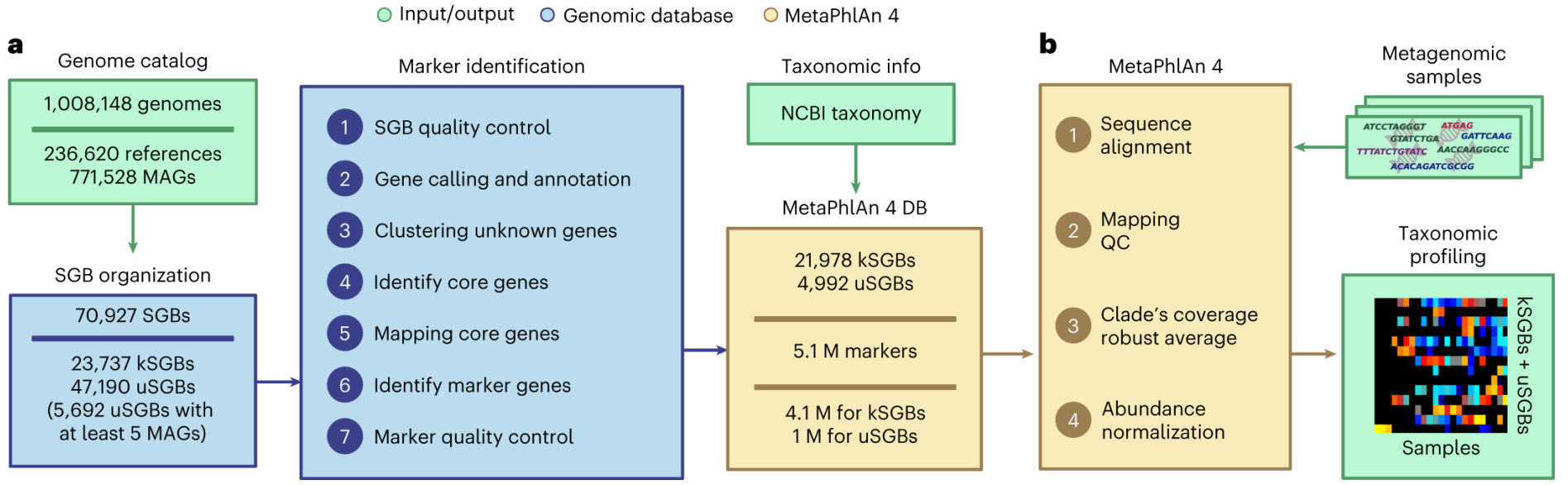
2 **Taxonomical annotation**

Who is there?

Software – marker gene based

MetaPhlAn: Metagenomic Phylogenetic Analysis

- is a computational tool for profiling the composition of microbial communities with **species-level**. With **StrainPhlAn**, it is possible to perform accurate strain-level microbial profiling.



MetaPhlAn 4 relies on ~**5.1M** unique clade-specific marker genes identified from ~**1M** microbial genomes (~236,600 references and 771,500 metagenomic assembled genomes) spanning 26,970 species-level genome bins (SGBs).

2 **Taxonomical annotation**

Who is there?

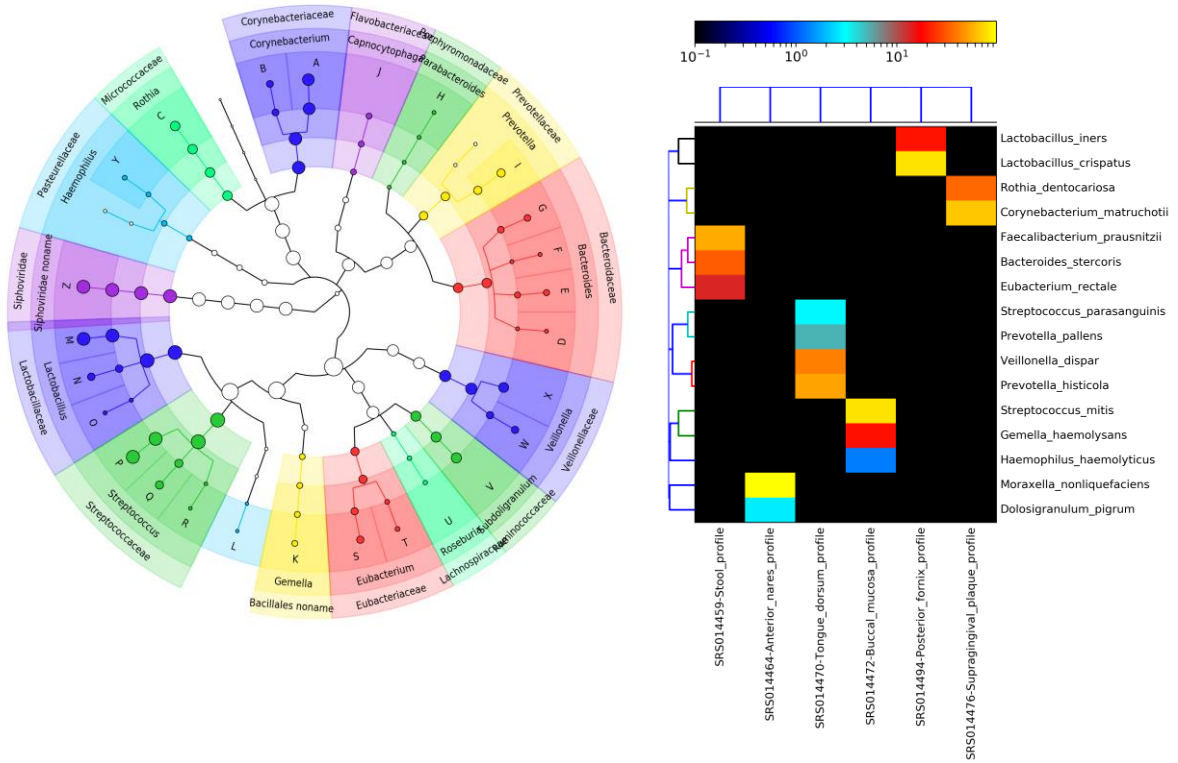
Software – marker gene based

MetaPhlAn: Metagenomic Phylogenetic Analysis

Input
- **Metagenomic shotgun sequencing reads** (FASTQ file) from a biological sample

Output
- A **taxonomic profile** of the sample (i.e. a list of microbial taxa and their relative abundances in %)

Visualization
- Sample-by-taxon **heatmap** of microbe abundances
- Tree (**cladogram**) summary of abundances



Examples visualization (https://github.com/biobakery/biobakery/wiki/metaphlan4)

2 **Taxonomical annotation**

Who is there?

Benchmarking of existent tools for taxonomic classification

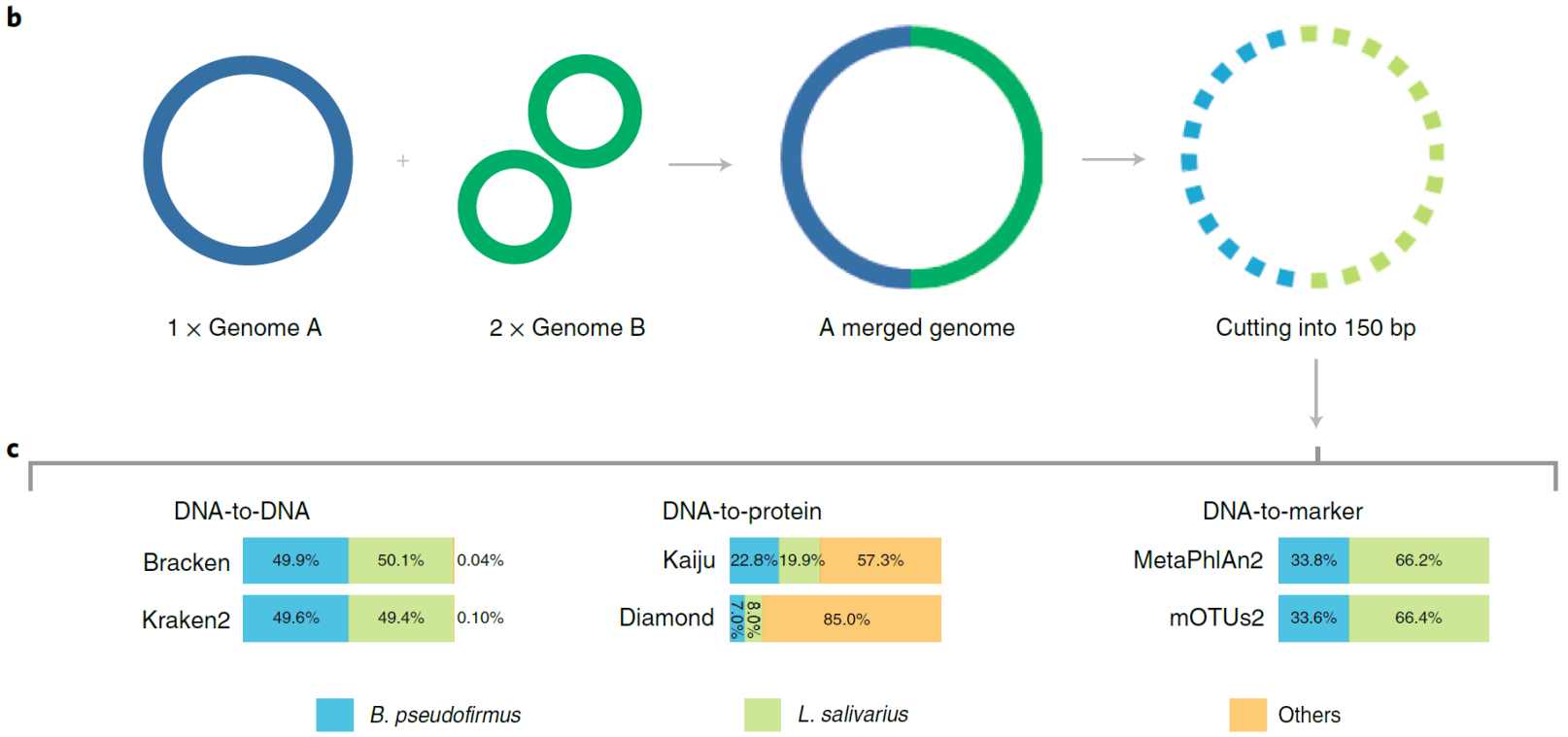
How to choose the best tool?

- **Question** - do I want to classify a specific MAG or to have an overview of the taxonomical diversity of a dataset? Classify a potential new genus, species?
- **Sample type** –completeness of reference databases
- **Availability** of computational resources and time

2 **Taxonomical annotation**

Who is there?

Benchmarking of existent tools for taxonomic classification

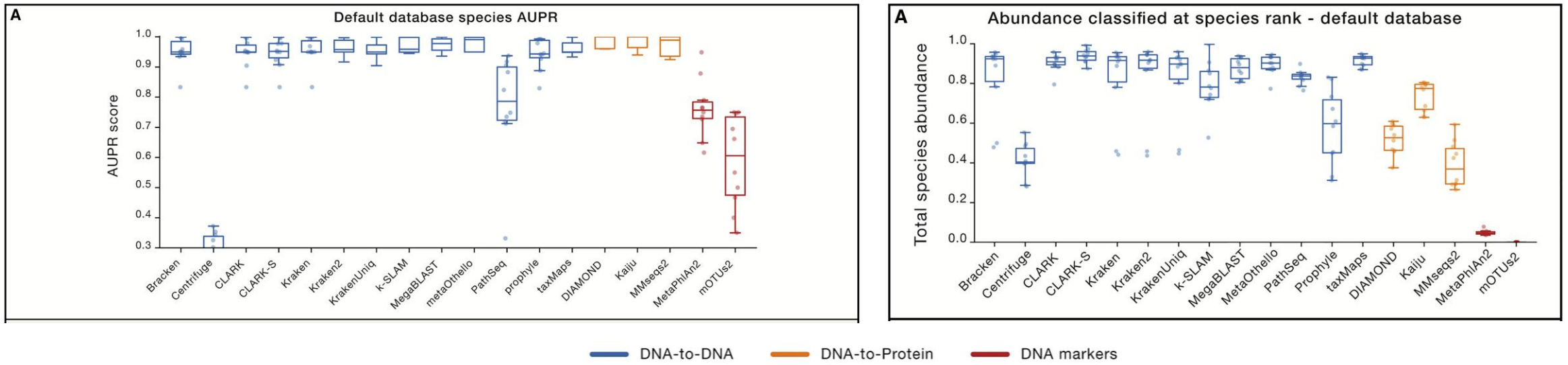


(Adapted from Sun et al. 2021)

2 **Taxonomical annotation**

Who is there?

Benchmarking of existent tools for taxonomic classification

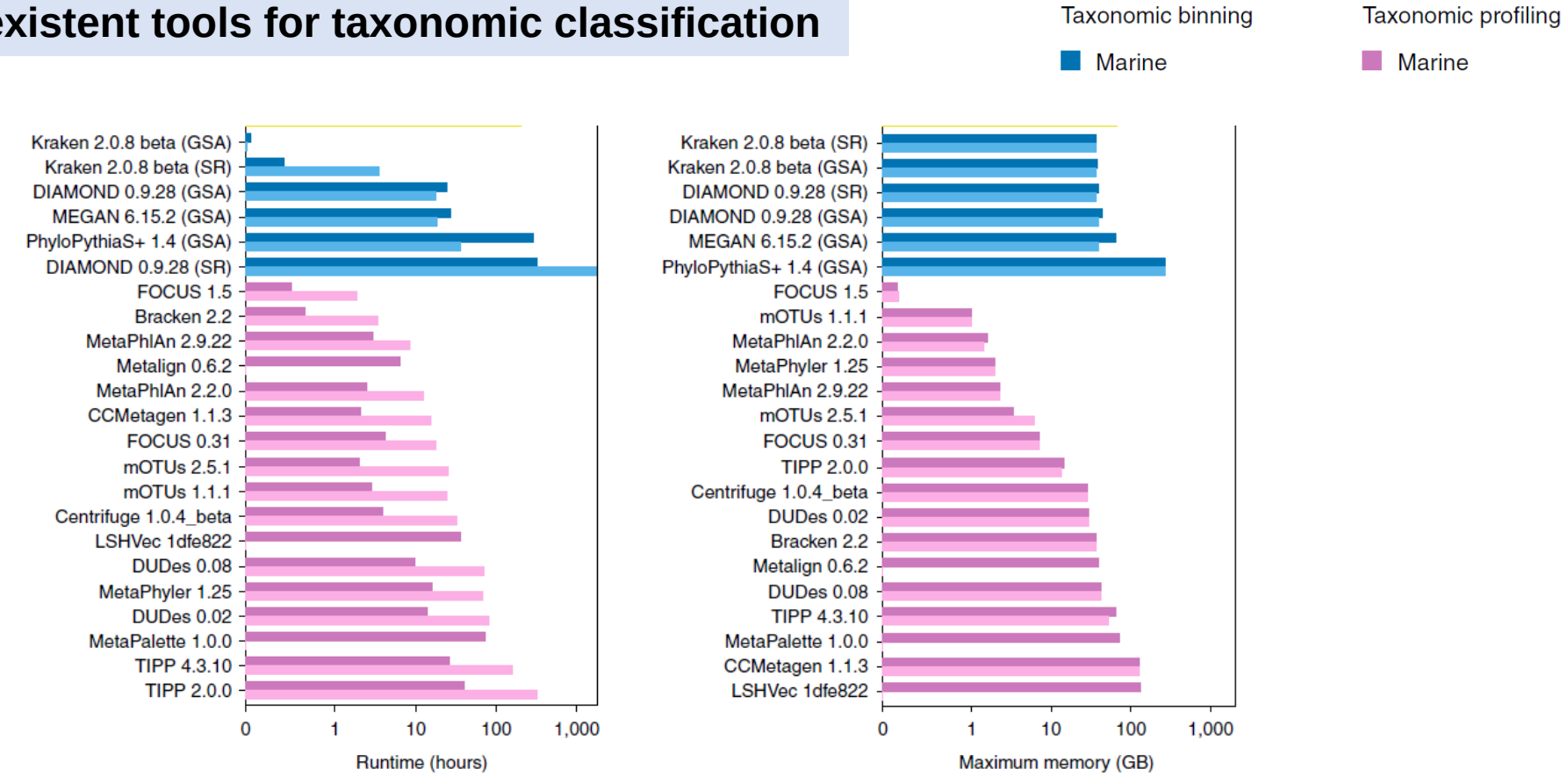


(Adapted from Ye et al. 2019)

2 **Taxonomical annotation**

Who is there?

Benchmarking of existent tools for taxonomic classification

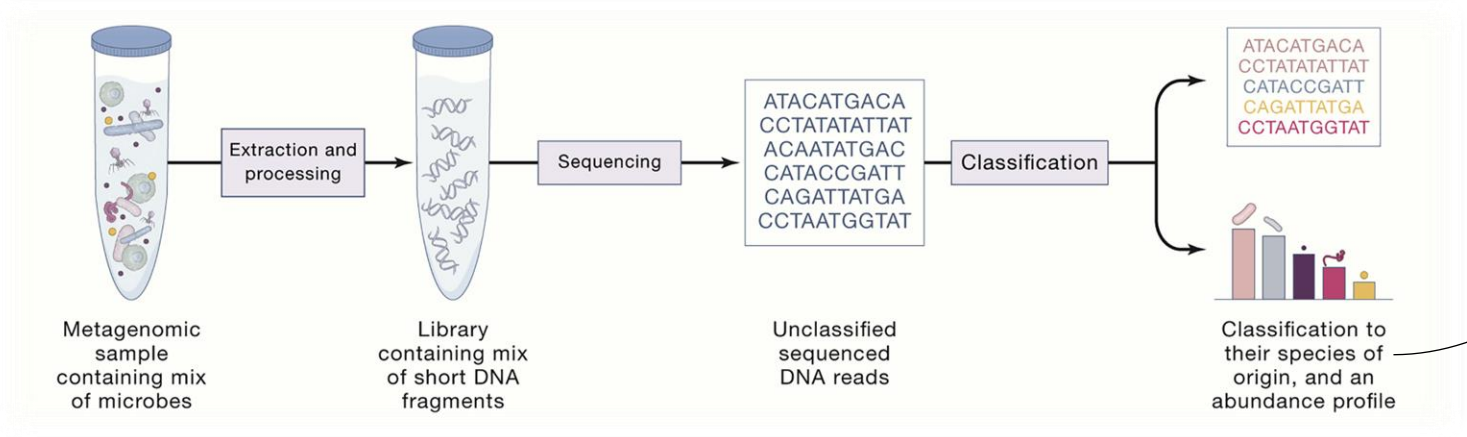


(Adapted from Meyer et al. 2022)

3

Functional annotation

What can they do?



Functional annotation

What can the identified microbial community (potentially) do?

→ **function prediction** for protein coding gene

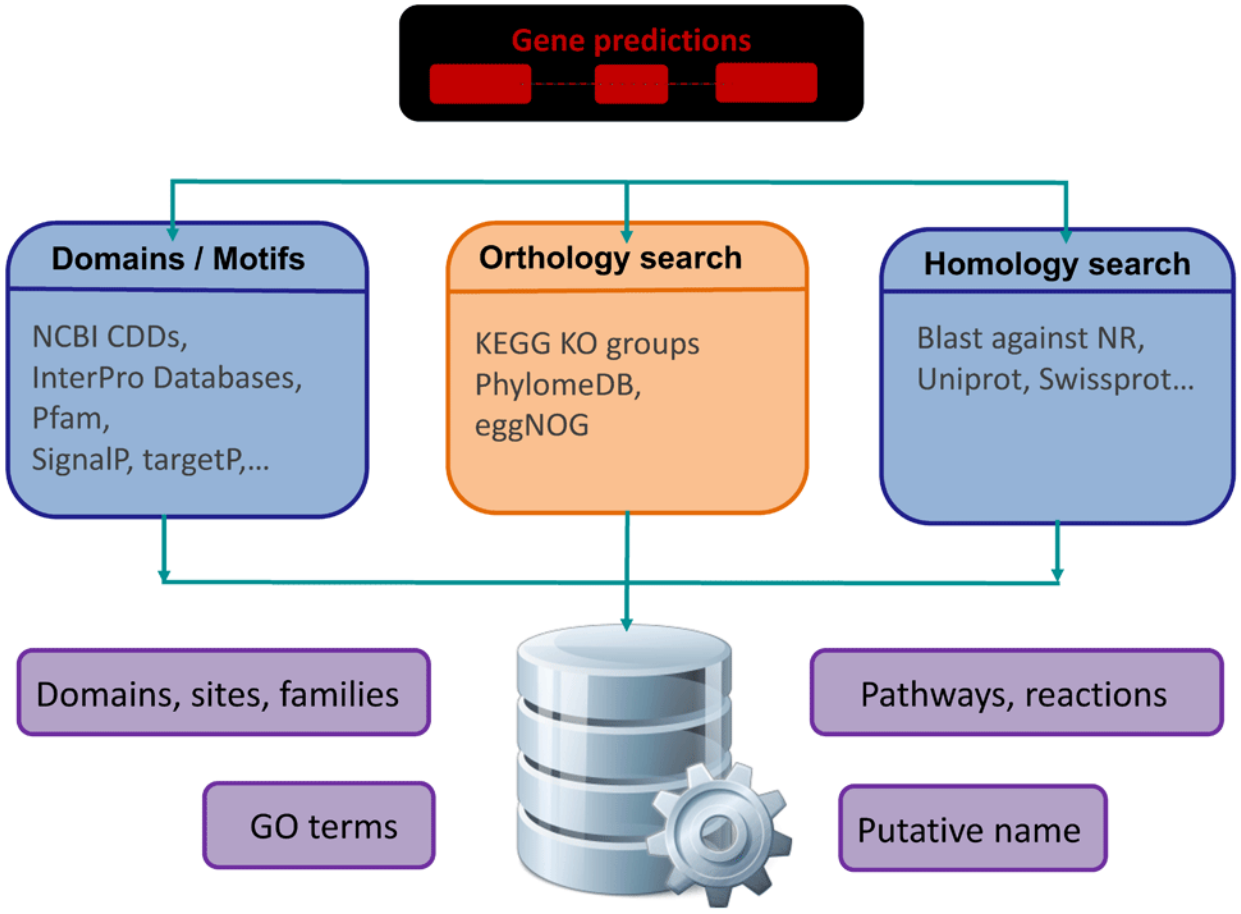
3 **Functional annotation**

What can they do?

Gene functional annotation tools can be classified into two categories:

1) tools with broad scopes to evaluate **full functional potential**

2) tools with narrow scopes focusing on one or a few **specific biological processes**.



(Adapted from Angel et al. 2018)

3 **Functional annotation**

What can they do?

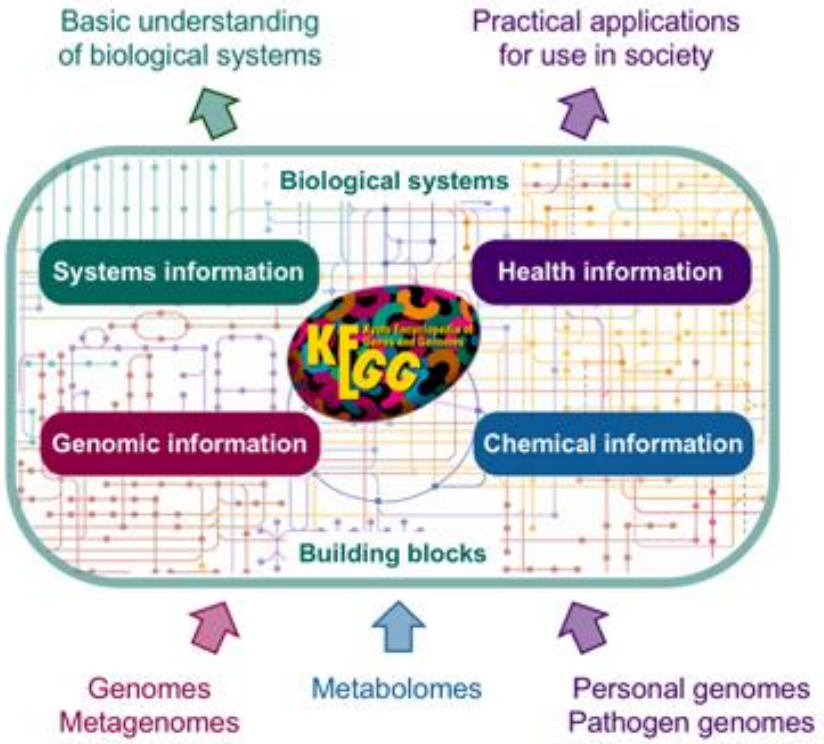
KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

Databases for Functional Annotation

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system.
- computer representation of the biological system, consisting of molecular building blocks of genes and proteins (**genomic information**) and chemical substances (**chemical information**) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information). It also contains disease and drug information (**health information)** as perturbations to the biological system.



3 Functional annotation

What can they do?

Databases for Functional Annotation

COG - Database of Clusters of Orthologous Genes (COGs)

https://www.ncbi.nlm.nih.gov/research/cog

COGs	Genomic loci	Taxonomic Categories	Organims	Protein IDs	COG symbols
4,877	3,456,089	37	1,309	3,213,255	3,828

3 **Functional annotation**

What can they do?

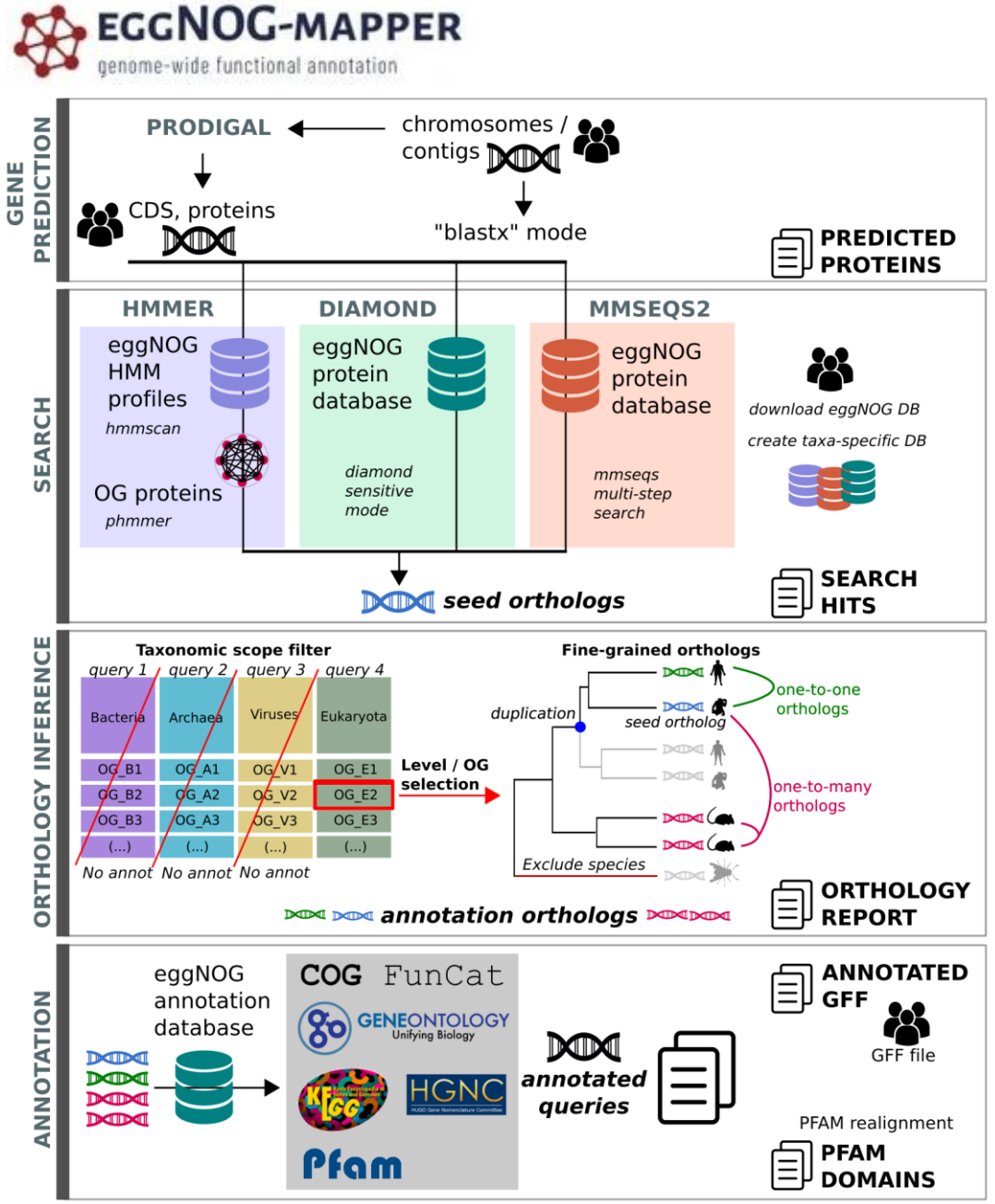
Databases for Functional Annotation

EggNOG http://eggnog5.embl.de/#/app/home

A database of orthology relationships, functional annotation and gene evolutionary histories, based on 5090 organisms and 2502 viruses.

Software for Functional Annotation

EggNOG-mapper - tool for fast functional annotation of novel sequences. It uses precomputed Orthologous Groups (OGs) and phylogenies from the EggNOG database.



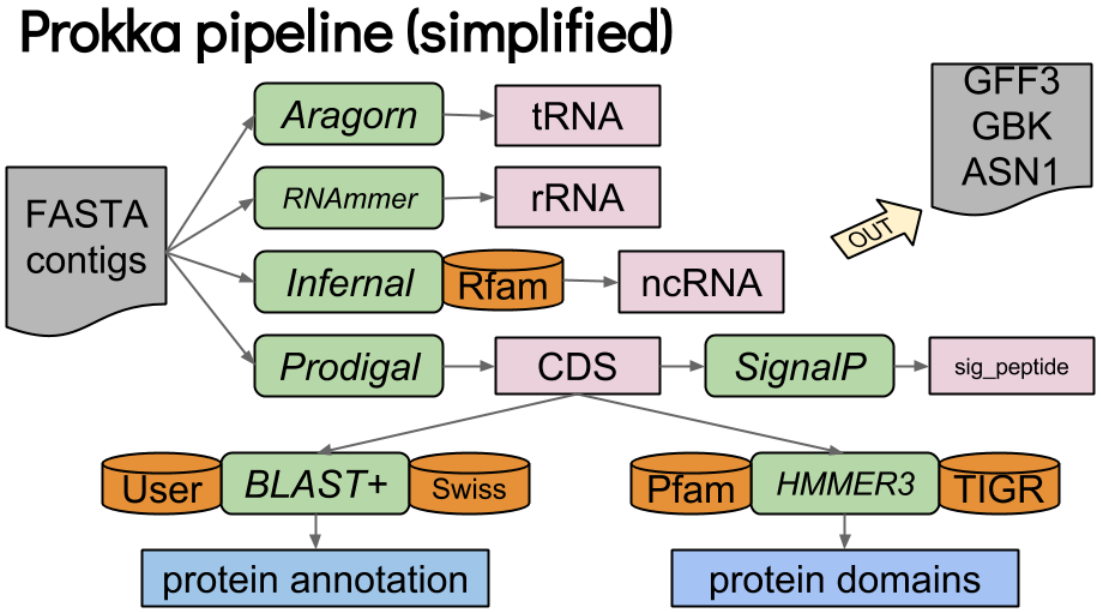
3 **Functional annotation**

What can they do?

Software for Functional Annotation

Prokka: rapid prokaryotic genome annotation

https://github.com/tseemann/prokka
is a pipeline that runs several other tools to annotate prokaryotic genomes.




```
# Choose the names of the output files
% prokka --outdir mydir --prefix mygenome contigs.fa
```

3 **Functional annotation**

What can they do?

Software for Functional Annotation

Prokka: rapid prokaryotic genome annotation

https://github.com/tseemann/prokka
is a pipeline that runs several other tools to annotate prokaryotic genomes.

locus_tag	ftype	length_bp	gene	EC_number	COG	product
OCAFAAEJ_00	CDS	1620				hypothetical protein
OCAFAAEJ_00	CDS	597				hypothetical protein
OCAFAAEJ_00	CDS	1059				hypothetical protein
OCAFAAEJ_00	CDS	474				hypothetical protein
OCAFAAEJ_00	CDS	456				hypothetical protein
OCAFAAEJ_00	CDS	786				hypothetical protein
OCAFAAEJ_00	CDS	888				hypothetical protein
OCAFAAEJ_00	CDS	969	dusB	1.3.1.-	COG0042	tRNA-dihydrouridine synthase B
OCAFAAEJ_00	CDS	549	yvqK	2.5.1.17	COG2096	Corrinoid adenosyltransferase
OCAFAAEJ_00	CDS	366				hypothetical protein
OCAFAAEJ_00	CDS	939				hypothetical protein
OCAFAAEJ_00	CDS	1830	btuB_1			Vitamin B12 transporter BtuB
OCAFAAEJ_00	CDS	195	def_1	3.5.1.88		Peptide deformylase
OCAFAAEJ_00	CDS	96				hypothetical protein
OCAFAAEJ_00	CDS	2166				hypothetical protein
OCAFAAEJ_00	CDS	171				hypothetical protein

```
# Choose the names of the output files
% prokka --outdir mydir --prefix mygenome contigs.fa
```

3 **Functional annotation**

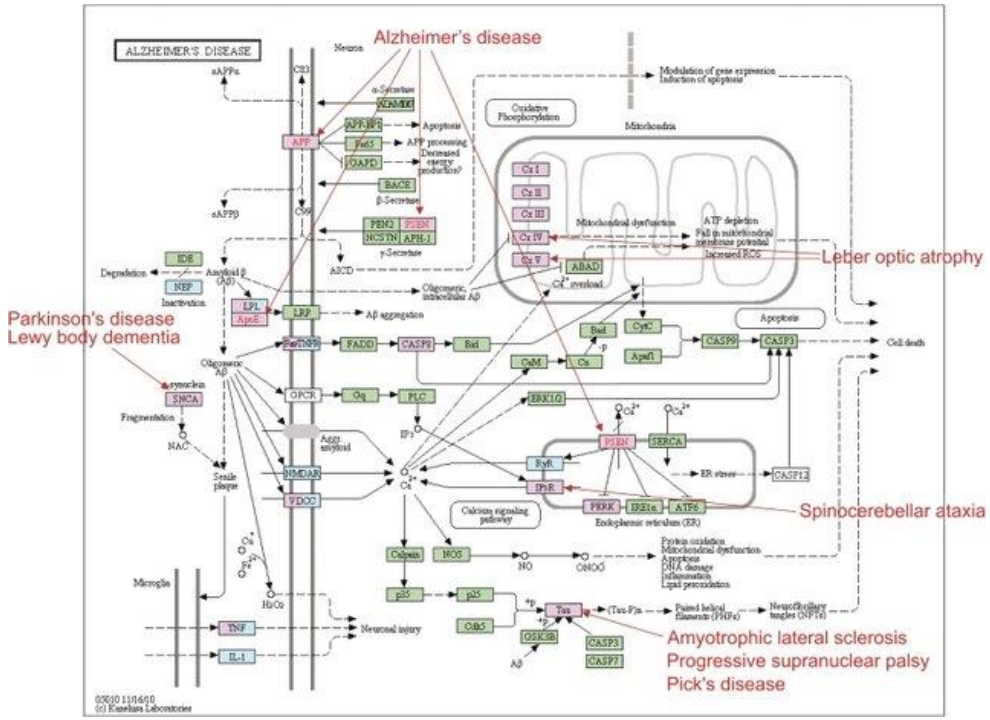
What can they do?

Visualization of Functional Annotation

KEGG Mapper

The Reconstruct tool "reconstructs" **KEGG pathway maps** and other network entities from a set of K numbers (KO identifiers).

KEGG Mapper – Reconstruct



3 **Functional annotation**

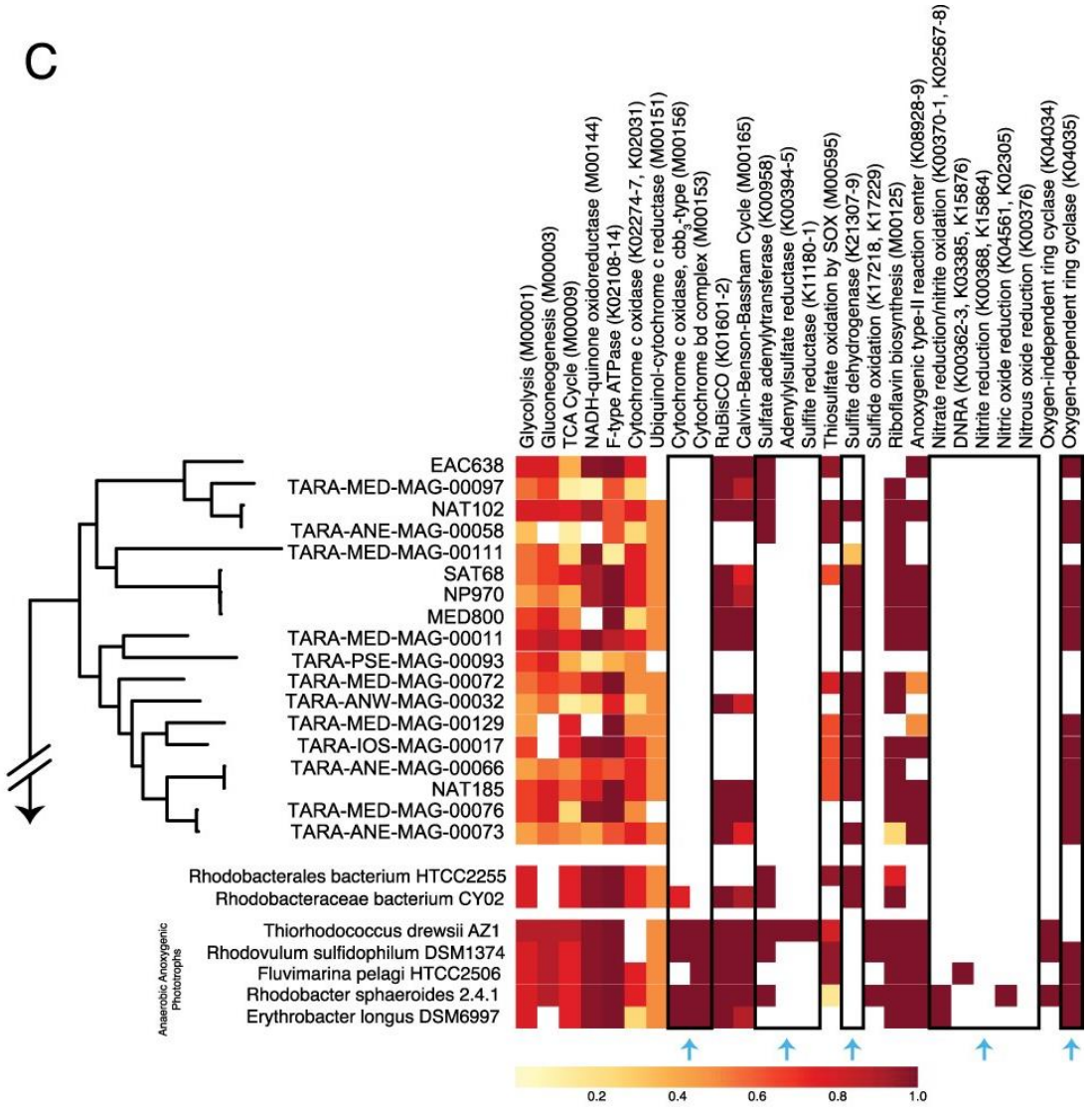
What can they do?

Visualization of Functional Annotation

KEGG decoder

https://github.com/bjtully/BioData/blob/master/KEGGDecoder/README.md

Designed to parse through a KEGG-Koala outputs (including blastKOALA, ghostKOALA, KOFAMSCAN) to determine the **completeness** of various metabolic pathways.



(Adapted from Graham et al. 2018)

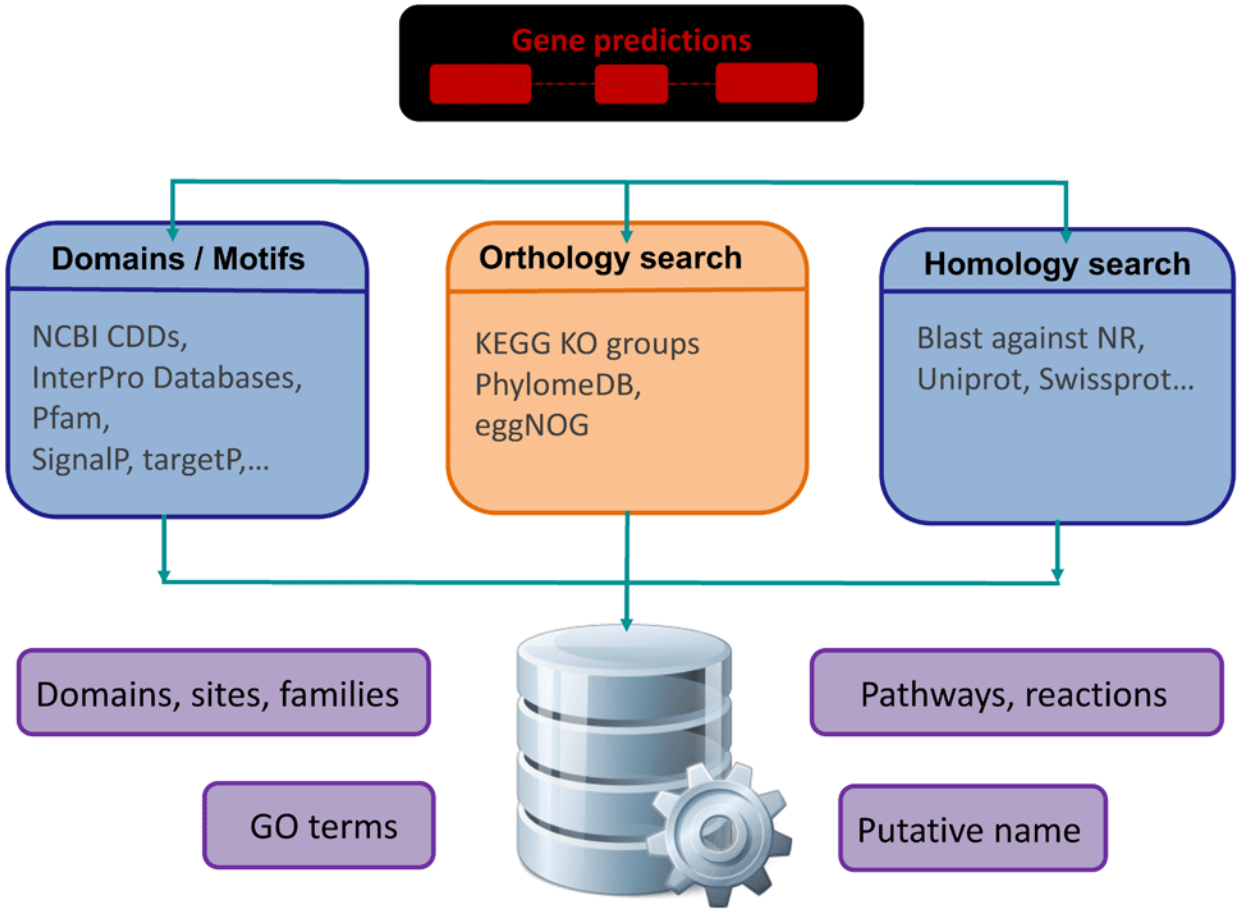
3 **Functional annotation**

What can they do?

Gene functional annotation tools can be classified into two categories:

1) tools with broad scopes to evaluate **full functional potential**

2) tools with narrow scopes focusing on one or a few **specific biological processes**.



(Adapted from Angel et al. 2018)

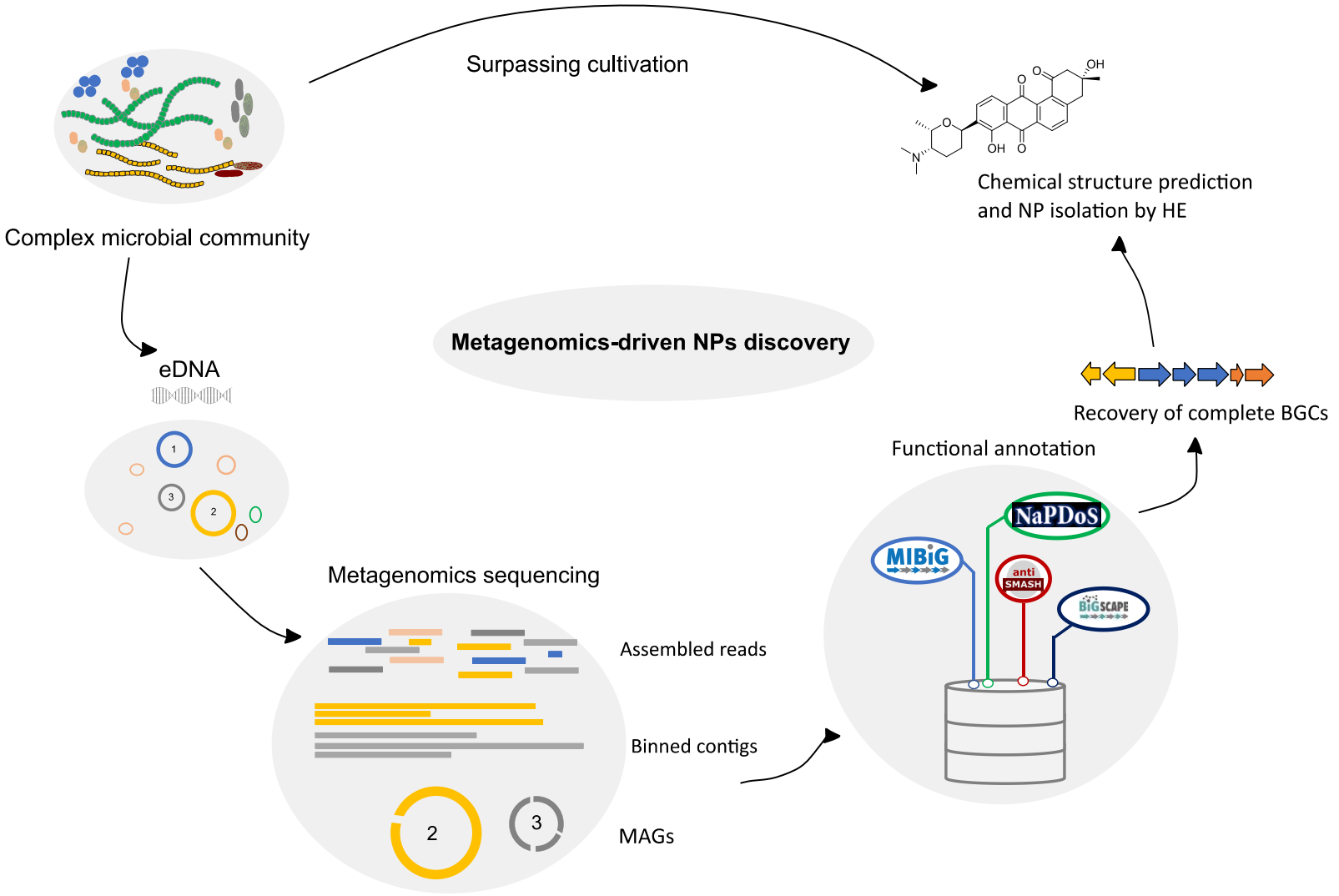
4

Functional metagenomics

What can they do?

Specialized metabolites

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). (Medema et al. 2015)



What can they do?

Biosynthetic gene clusters- identification

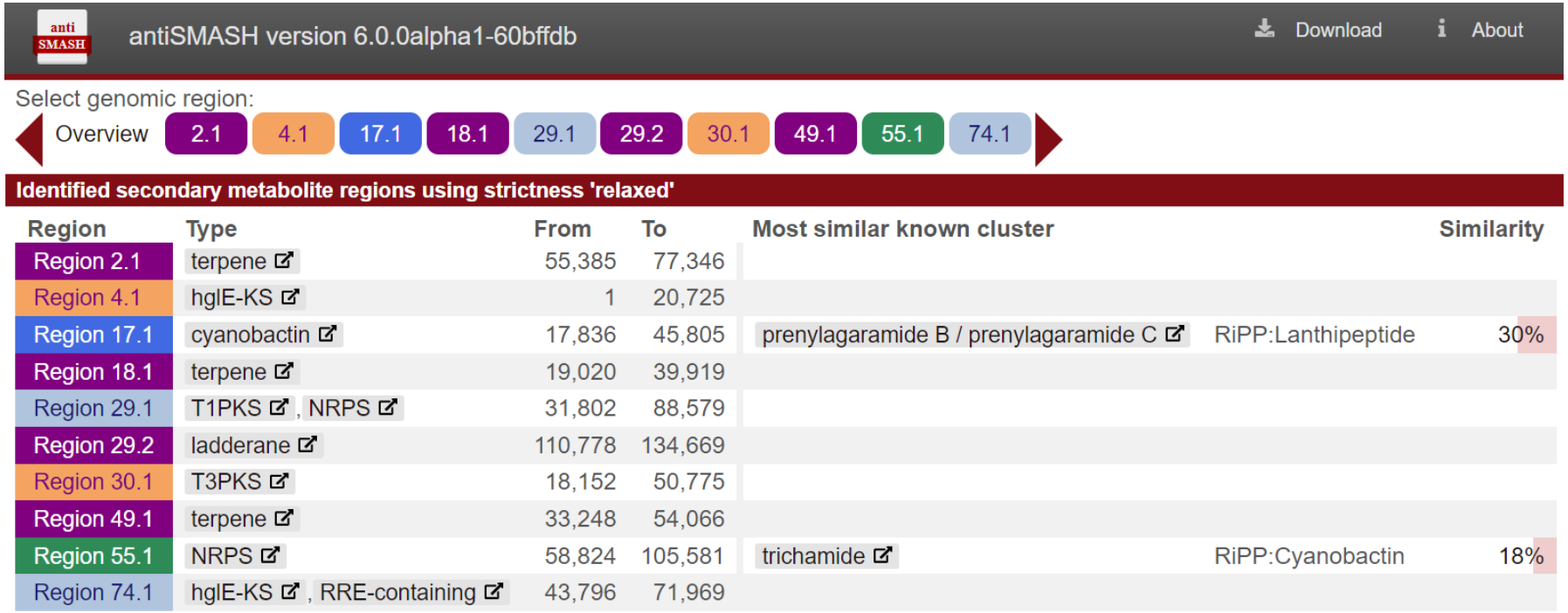
https://antismash.secondarymetabolites.org/#!/start

- allows the rapid genome-wide identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes.
- Web version and command line

```
antismash  MAG_example.fa --cb-general  --cb-subclusters --cb-knownclusters --genefinding-tool prodigal --output-dir MAG_example_antismash_output
```

4 **Functional metagenomics**

What can they do?



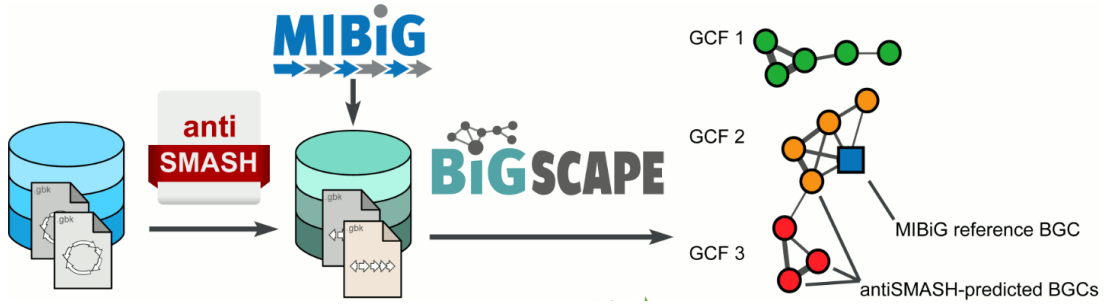
Region	Type	From	To	Most similar known cluster		Similarity
Region 2.1	terpene	55,385	77,346			
Region 4.1	hglE-KS	1	20,725			
Region 17.1	cyanobactin	17,836	45,805	prenylagaramide B / prenylagaramide C	RiPP:Lanthipeptide	30%
Region 18.1	terpene	19,020	39,919			
Region 29.1	T1PKS, NRPS	31,802	88,579			
Region 29.2	ladderane	110,778	134,669			
Region 30.1	T3PKS	18,152	50,775			
Region 49.1	terpene	33,248	54,066			
Region 55.1	NRPS	58,824	105,581	trichamide	RiPP:Cyanobactin	18%
Region 74.1	hglE-KS, RRE-containing	43,796	71,969			

antismash **MAG_example.fa** --cb-general --cb-subclusters --cb-knownclusters --genefinding-tool prodigal --output-dir MAG_example_antismash_output

4 **Functional metagenomics**

What can they do?

Biosynthetic gene clusters- identification



BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine)
- https://bigscape-corason.secondarymetabolites.org/

Is a software package that constructs **sequence similarity networks** of Biosynthetic Gene Clusters (BGCs) and groups them into **Gene Cluster Families** (GCFs).

```
python bigscape.py -i  antismash_gbk_files  --mix --mibig --include_singletons --cutoffs 0.3, 0.7 -o bigscape_output
```

4 Functional metagenomics

What can they do?

Article

Biosynthetic potential of the global ocean microbiome

https://doi.org/10.1038/s41586-022-04862-3

Received: 21 May 2021

Accepted: 12 May 2022

Published online: 22 June 2022

Open access

Check for updates

Lucas Paoli[1], Hans-Joachim Ruscheweyh[1,18], Clarissa C. Forneris[2,18], Florian Hubrich[2,18], Satria Kautsar[3], Agneya Bhushan[2], Alessandro Lotti[2], Quentin Clayssen[1], Guillem Salazar[1], Alessio Milanese[1], Charlotte I. Carlström[1], Chrysa Papadopoulou[1], Daniel Gehrig[1], Mikhail Karasikov[4,5,6], Harun Mustafa[4,5,6], Martin Larralde[7], Laura M. Carroll[7], Pablo Sánchez[8], Ahmed A. Zayed[9], Dylan R. Cronin[9], Silvia G. Acinas[8], Peer Bork[7,10,11], Chris Bowler[12,13], Tom O. Delmont[13,14], Josep M. Gasol[8], Alvar D. Gossert[15], André Kahles[4,5,6], Matthew B. Sullivan[8,16], Patrick Wincker[13,14], Georg Zeller[7], Serina L. Robinson[2,17]✉, Jörn Piel[2]✉ & Shinichi Sunagawa[1]✉

MICROBIAL GENOMICS

RESEARCH ARTICLE

Rego *et al.*, *Microbial Genomics* 2021;7:000731

DOI 10.1099/mgen.0.000731

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes

Adriana Rego[1,2], Antonio Fernandez-Guerra[3], Pedro Duarte[4], Philipp Assmy[4], Pedro N. Leão[1,*] and Catarina Magalhães[1,5,*]

4 **Functional metagenomics**

What can they do?

Article

Biosynthetic potential of the global ocean microbiome

https://doi.org/10.1038/s41586-022-04862-3

Received: 21 May 2021

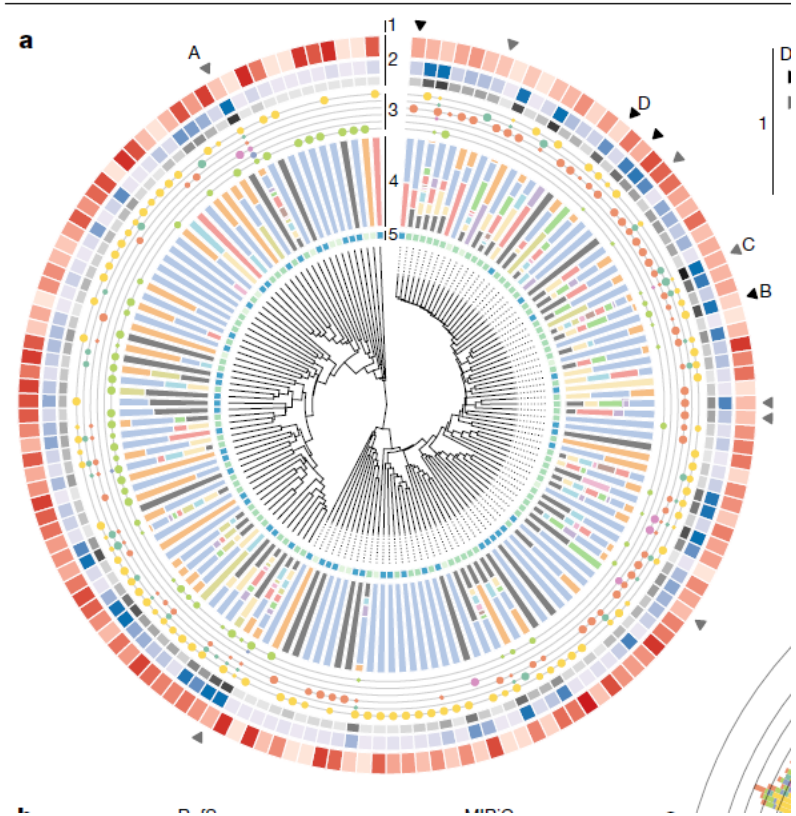
Accepted: 12 May 2022

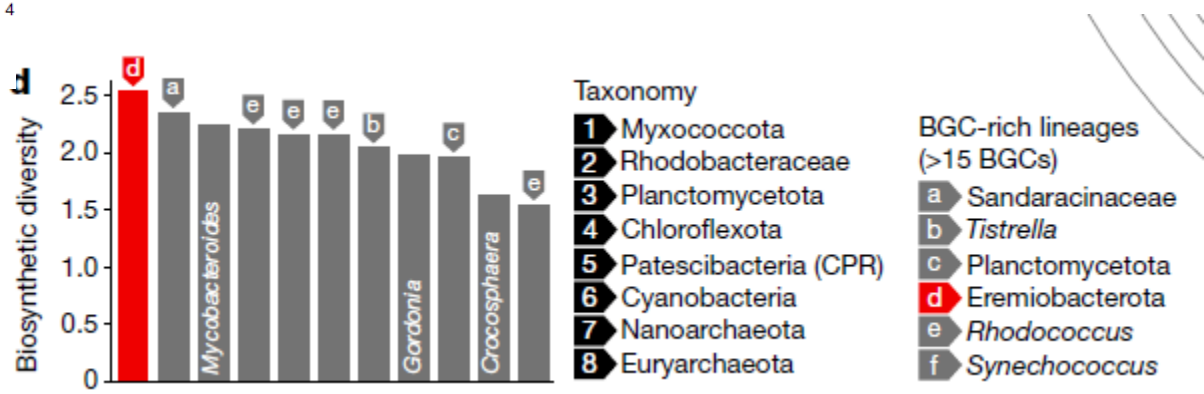
Published online: 22 June 2022

Open access

Check for updates

Lucas Paoli[1], Hans-Joachim Ruscheweyh[1,18], Clarissa C. Forneris[2,18], Florian Hubrich[2,18], Satria Kautsar[3], Agneya Bhushan[2], Alessandro Lotti[2], Quentin Clayssen[1], Guillem Salazar[1], Alessio Milanese[1], Charlotte I. Carlström[1], Chrysa Papadopoulou[1], Daniel Gehrig[1], Mikhail Karasikov[4,5,6], Harun Mustafa[4,5,6], Martin Larralde[7], Laura M. Carroll[7], Pablo Sánchez[8], Ahmed A. Zayed[9], Dylan R. Cronin[9], Silvia G. Acinas[8], Peer Bork[7,10,11], Chris Bowler[12,13], Tom O. Delmont[13,14], Josep M. Gasol[8], Alvar D. Gossert[15], André Kahles[4,5,6], Matthew B. Sullivan[8,16], Patrick Wincker[13,14], Georg Zeller[7], Serina L. Robinson[2,17]✉, Jörn Piel[2]✉ & Shinichi Sunagawa[1]✉





4

Functional metagenomics

What can they do?

Article

Biosynthetic potential of the global ocean microbiome

https://doi.org/10.1038/s41586-022-04862-3

Received: 21 May 2021

Accepted: 12 May 2022

Published online: 22 June 2022

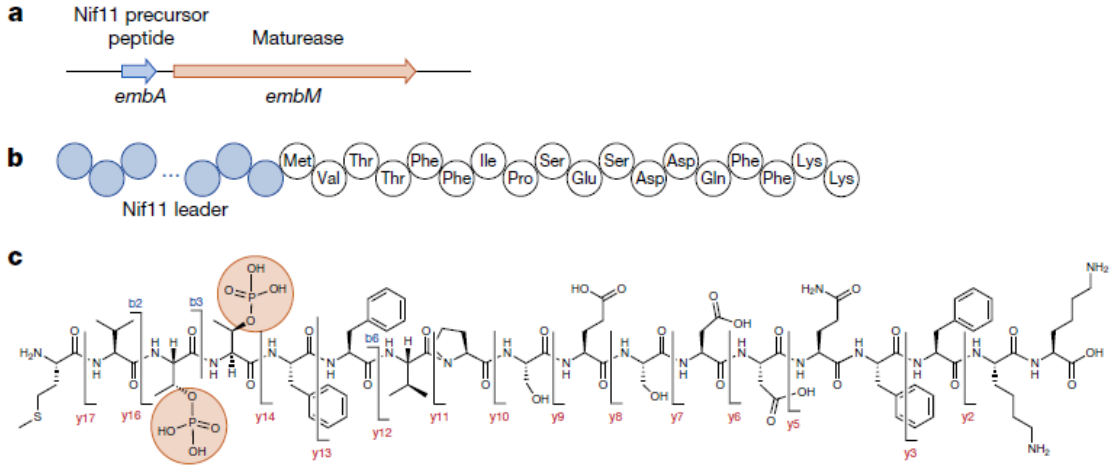
Open access

Check for updates

Lucas Paoli[1], Hans-Joachim Ruscheweyh[1,18], Clarissa C. Forneris[2,18], Florian Hubrich[2,18], Satria Kautsar[3], Agneya Bhushan[2], Alessandro Lotti[2], Quentin Clayssen[1], Guillem Salazar[1], Alessio Milanese[1], Charlotte I. Carlström[1], Chrysa Papadopoulou[1], Daniel Gehrig[1], Mikhail Karasikov[4,5,6], Harun Mustafa[4,5,6], Martin Larralde[7], Laura M. Carroll[7], Pablo Sánchez[8], Ahmed A. Zayed[9], Dylan R. Cronin[9], Silvia G. Acinas[8], Peer Bork[7,10,11], Chris Bowler[12,13], Tom O. Delmont[13,14], Josep M. Gasol[8], Alvar D. Gossert[15], André Kahles[4,5,6], Matthew B. Sullivan[8,16], Patrick Wincker[13,14], Georg Zeller[7], Serina L. Robinson[2,17]✉, Jörn Piel[2]✉ & Shinichi Sunagawa[1]✉

New enzymes and natural products

We finally sought to experimentally validate the promising prospects of our microbiomics-driven work for the discovery of new pathways, enzymes and natural products. Among the different BGC classes, RiPP



4

Functional metagenomics

What can they do?

MICROBIAL GENOMICS

RESEARCH ARTICLE

Rego *et al.*, *Microbial Genomics* 2021;7:000731

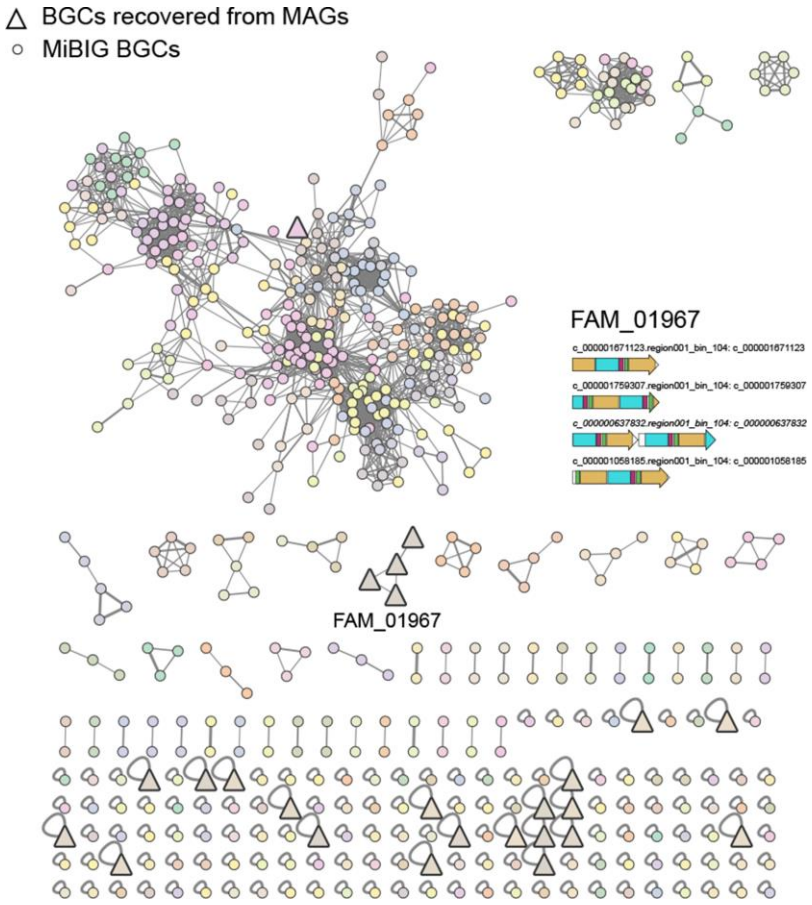
DOI 10.1099/mgen.0.000731

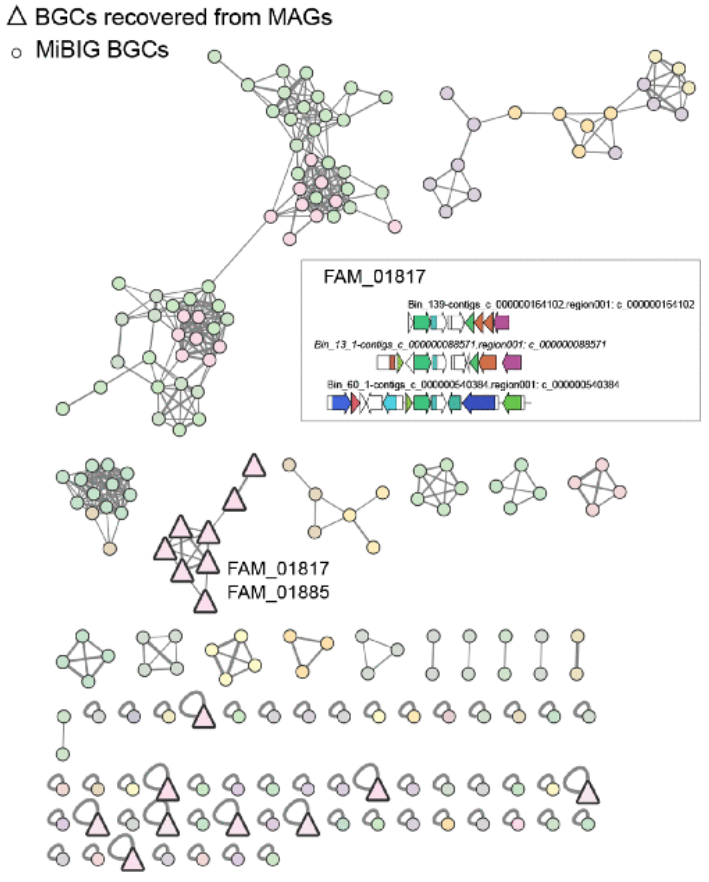
MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes

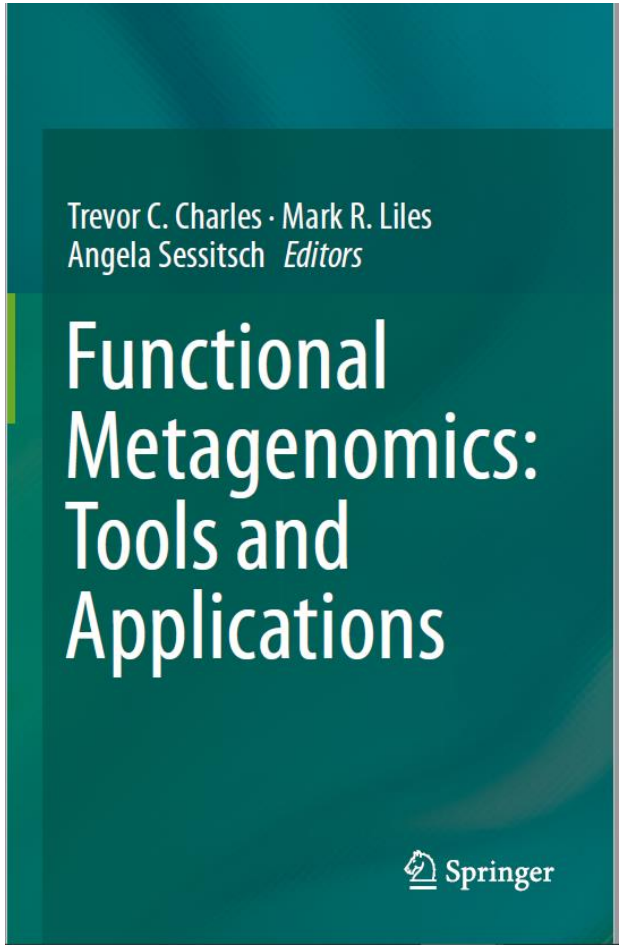
Adriana Rego[1,2], Antonio Fernandez-Guerra[3], Pedro Duarte[4], Philipp Assmy[4], Pedro N. Leão[1,*] and Catarina Magalhães[1,5,*]





8 Resources

Bibliography



Questions?

ciimar

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Sponsorship

SPECO

BIOPORTUGAL S.A.
Químico, Farmacêutica