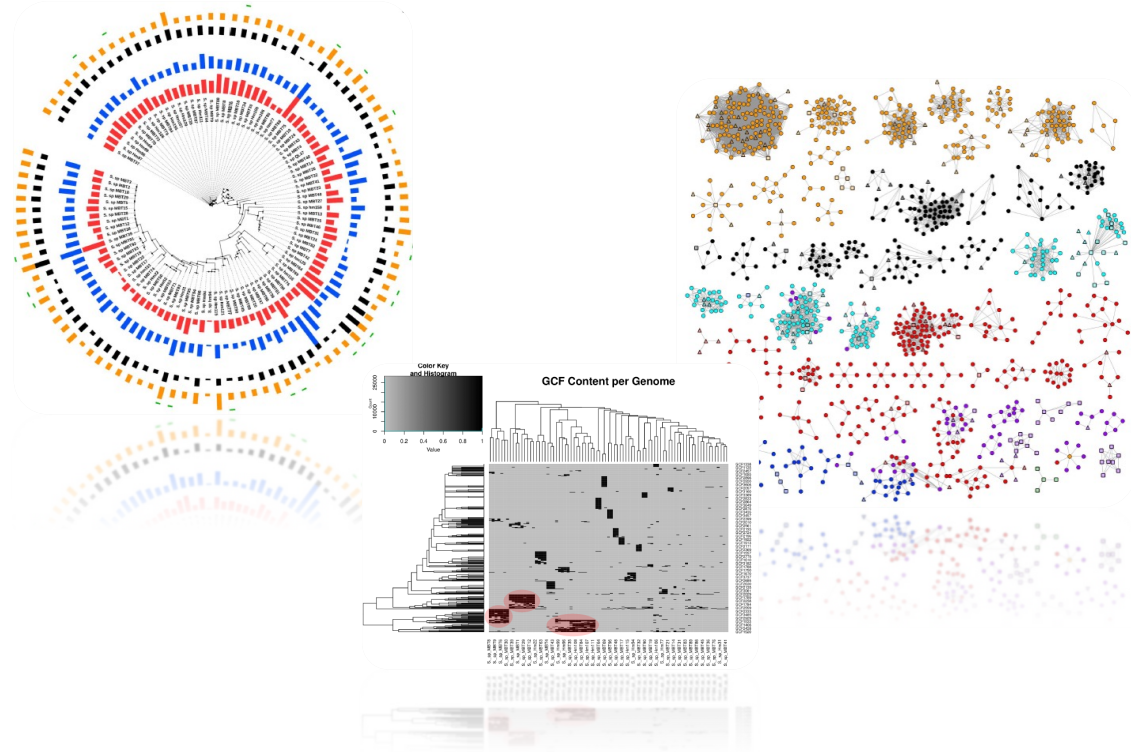
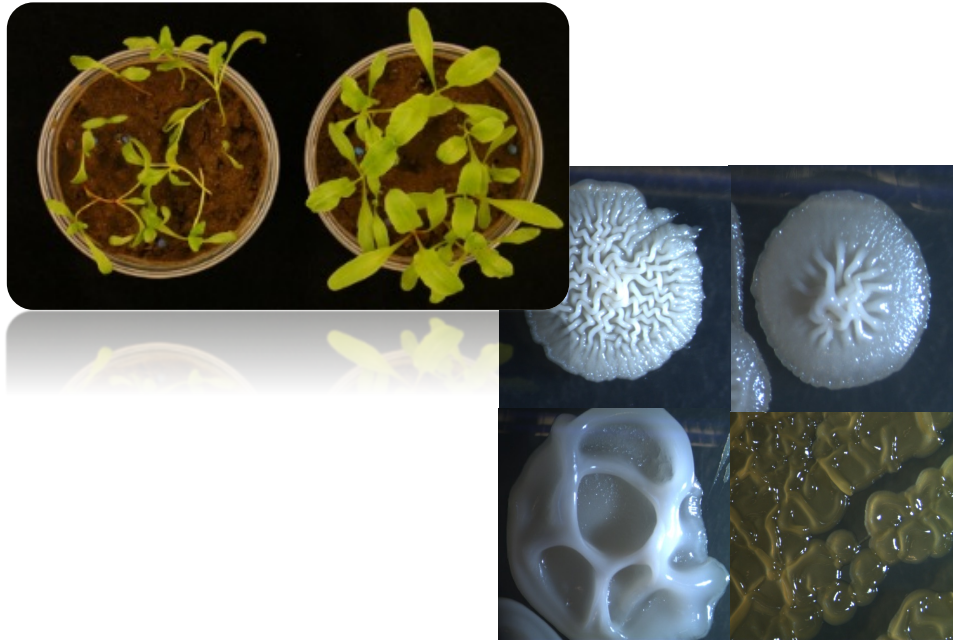


ANNOTATION TOOLS



Victor J Carrión

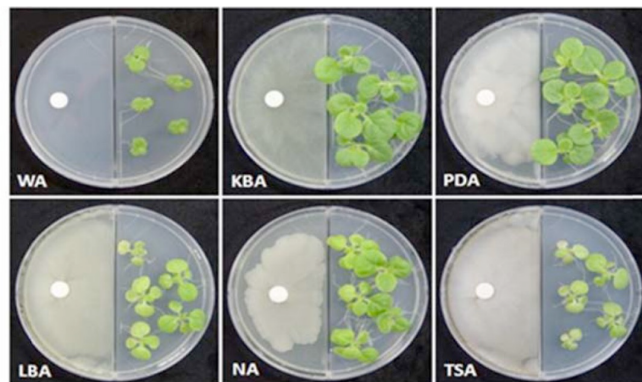
v.j.carrion.bravo@biology.leidenuniv.nl

 @VCarryOn1



Universiteit Leiden





DNA reads

Quality
control

FastQC

Filtered DNA reads

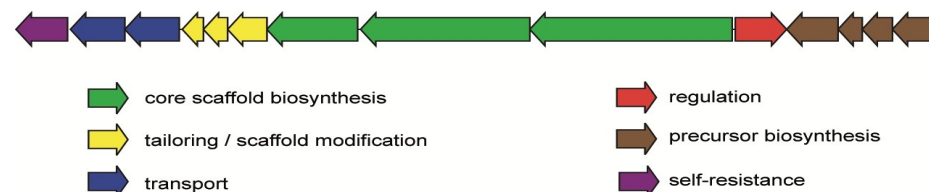
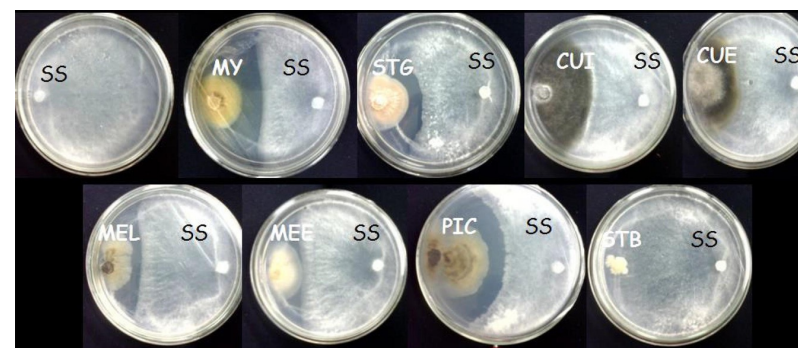
Genome
assembly

MEGAHIT

(meta)Genome

Annotation

Comp. Genomics



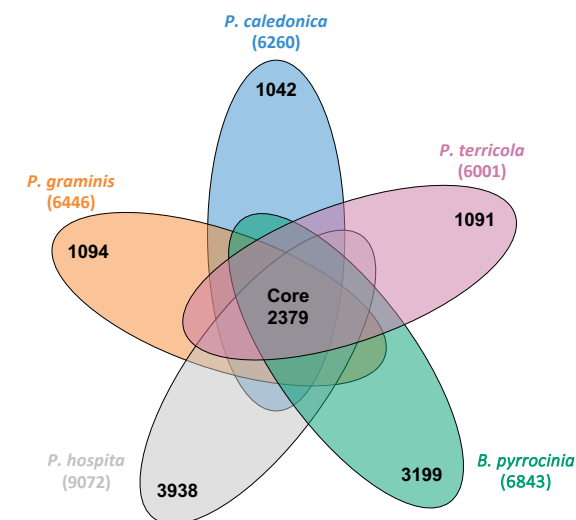
Velvet



MEGAHIT

Quast

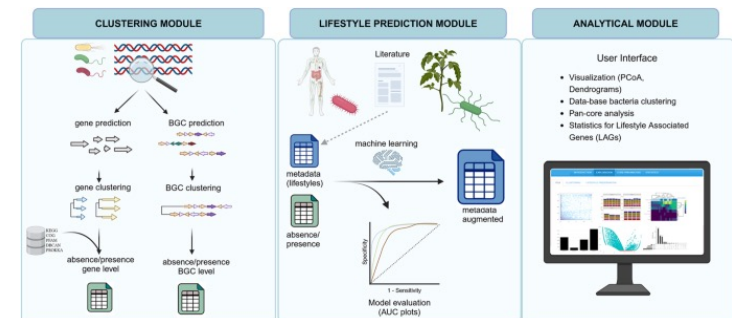
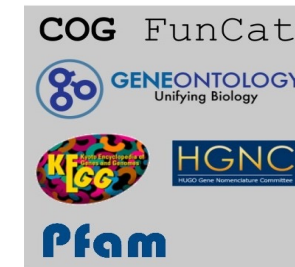
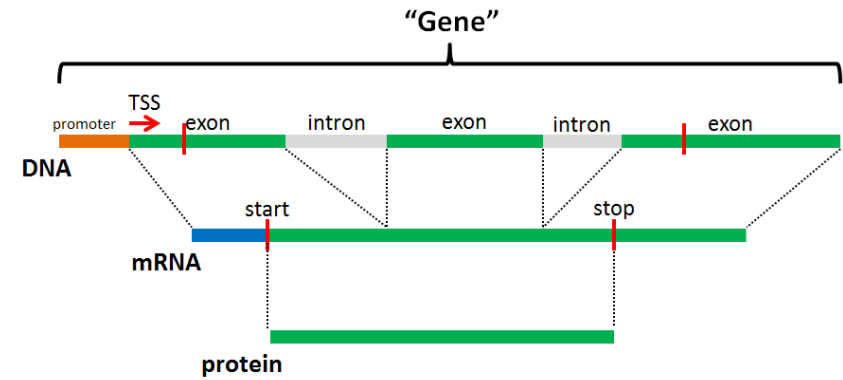
Quality Assessment Tool for Genome Assemblies



WHAT IS GENOME ANNOTATION?

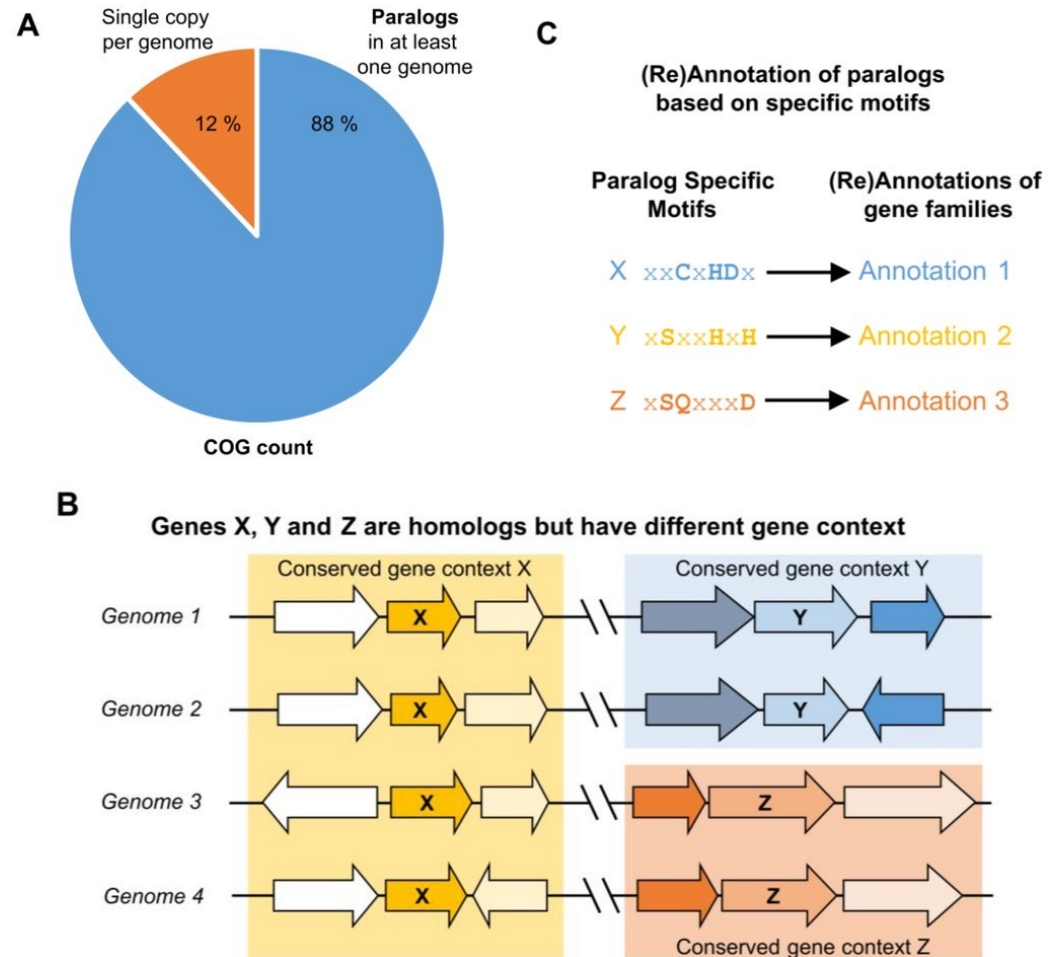
Genome annotation involves identifying genes, their structures, and functions in a genome.

- **Structural annotation:** Identifies coding regions and regulatory elements (Prodigal).
- **Functional annotation:** Assigns functions to genes based on homology and pathways.
- **Comparative annotation:** Analyzes evolutionary relationships and gene orthology.



LESSON 1: NEVER TRUST AN ANNOTATION

- Functional annotations provide a first hint ***but are often wrong***
- Often based on the protein with the best BLAST hit in some database, which may not be that closely related
- Errors are easily propagated
- To draw strong conclusions, manual validation is necessary, by looking at (conservation of) gene context, active-site motifs, etc.



GENERAL FUNCTIONAL ANNOTATION TOOLS

PFAM CAN BE USED TO RAPIDLY PROVIDE A HIGH-CONFIDENCE INITIAL ANNOTATION OF GENE FUNCTIONS



<http://www.ebi.ac.uk/Tools/hmmer/>

<http://pfam.xfam.org/>

wellcome trust sanger institute **Pfam**

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

keyword search Go

Family: *Piwi* (PF02171)

18 architectures 842 sequences 0 interactions 215 species 46 structures

Summary
Domain organisation
Clans
Alignments
HMM logo
Trees
Curation & models
Species
Interactions
Structures
Jump to...
enter ID/acc Go

Summary

Pfam includes annotations and additional family information from a range of different sources. These sources can be accessed via the tabs below.

Wikipedia: Piwi **Pfam** **Interpro**

The Pfam group coordinates the annotation of Pfam families in [Wikipedia](#). This family is described by a Wikipedia entry entitled "[Piwi](#)". [More...](#)

Piwi

[Edit Wikipedia article](#)

The **piwi** (sometimes also **PIWI**; originally **P-element induced wimpy testis** in *Drosophila*^[2]) class of genes was originally identified as encoding **regulatory proteins** responsible for maintaining incomplete **differentiation** in stem cells and maintaining the stability of cell division rates in germ line cells.^[3] Piwi proteins are highly conserved across evolutionary lineages and are present in both plants and animals.^[4] One of the major human homologs, whose upregulation is implicated in the formation of tumors such as seminomas, is called hiwi;^[5] other variants on the theme include the miwi protein in mice.^[6]

Contents [\[show\]](#)

- 1 Role in RNA interference
- 2 piRNAs and transposon silencing
- 3 References
- 4 External links

Role in RNA interference

The piwi domain^[7] is a protein domain homologous to piwi proteins and present in a large number of nucleic acid-binding proteins, especially those that bind and cleave RNA. The best-studied such family of proteins is the argonaute family; argonautes are RNase H-like enzymes that carry out the catalytic functions of the RNA-induced silencing complex (RISC). In the well-known cellular process of RNA interference, the argonaute protein in the RISC complex binds small interfering RNA (siRNA) generated from exogenous double-stranded RNA or microRNA (miRNA) generated from endogenous non-coding RNA, both the ribonuclease dicer. The RNA-RISC complex binds and cleaves complementary base pairing messenger RNA, destroying it and preventing its translation into a protein. Crystallized piwi domains have a conserved basic binding site for the 5' end of bound RNA; in the case of argonaute proteins binding siRNA strands, the last unpaired nucleotide base of the siRNA is also stabilized by base stacking interactions between the base and neighboring tyrosine residues.^[8]

Recent evidence suggests that the germ-line determination function of piwi proteins relies on their interaction with miRNAs, which are known to play a key role in early development and morphogenesis of *Drosophila melanogaster* embryos, in which germ-line maintenance has been extensively studied.^[9]

piRNAs and transposon silencing

Recently, a novel class of longer-than-average miRNAs known as Piwi-interacting RNAs (piRNAs) has been defined in mammalian cells, about 26-31 nucleotides long as compared to the more typical miRNA or siRNA of about 21 nucleotides. These piRNAs are expressed specifically in germline cells in the testes of

Piwi domain

Structure of *Pyrococcus furiosus* Argonaute.^[1]

Identifiers	
Symbol	Piwi
Pfam	PF02171 [E]
InterPro	IPR003165 [E]
PROSITE	PS50822 [E]

Available protein structures: [\[show\]](#)

PDB	RCSB PDB [E] ; PDBe [E]
PDBsum	structure summary [E]



HMMER

Biosequence analysis using profile hidden Markov Models

[Home](#)[Search](#)[Results](#)[Software](#)[Help](#)[About](#)[Contact](#)

Quick search

Paste in your sequence or use the [example](#) 

Enter your sequence

☒ Reference Proteomes ☐ UniProtKB ☐ SwissProt ☐ Pfam

[Submit](#)[Reset](#)[Clean](#)

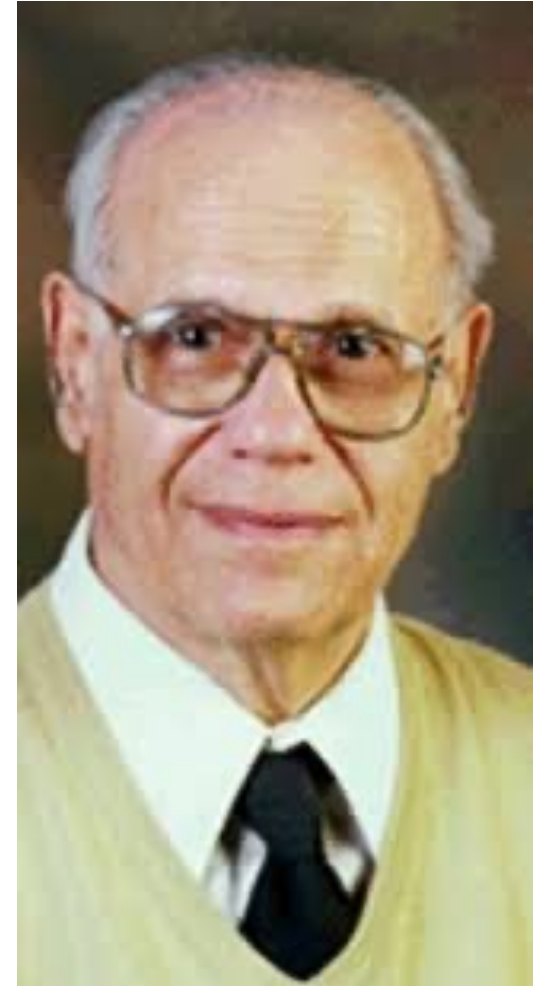
[Alternative search options](#)

The HMMER web server: fast and sensitive homology searches. This site has been designed to provide near **interactive searches** for most queries, coupled with **intuitive and interactive results** visualisations.

[Quickstart tutorial](#)[Online documentation](#)

Historia de los Hidden Markov Models (HMMs)

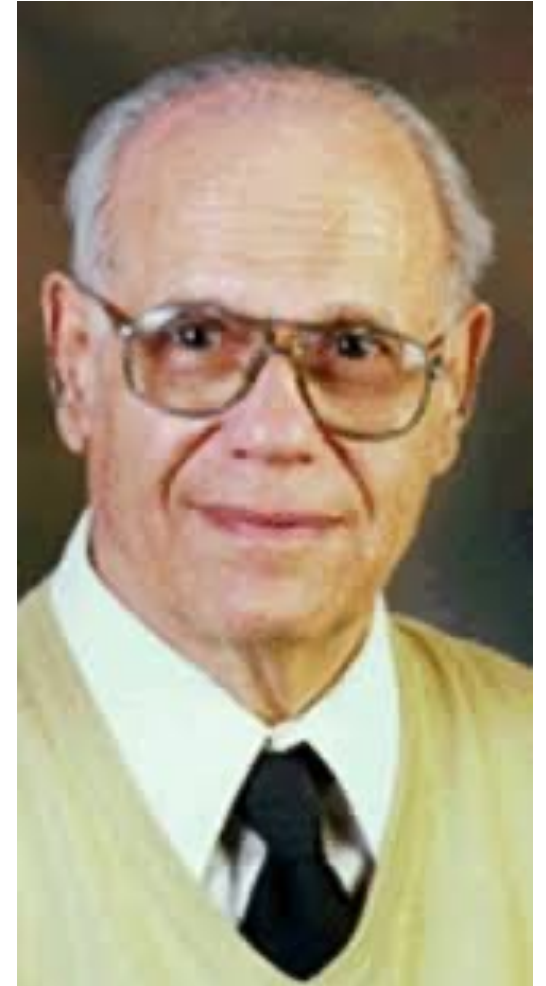
- Propuestos por Leonard E. Baum y sus colaboradores en la década de 1960.
- Inicialmente diseñados para **aplicaciones matemáticas y teóricas**.
- Permiten modelar sistemas complejos donde los estados internos no son directamente observables.
- Su capacidad de **inferencia probabilística** los hizo ideales para la biología computacional.
- Introducción en análisis de secuencias biológicas en los años 1990.
- Herramientas como HMMER utilizan HMMs para detectar familias de proteínas y regiones funcionales en ADN.



Historia de los Hidden Markov Models (HMMs)

Otras aplicaciones:

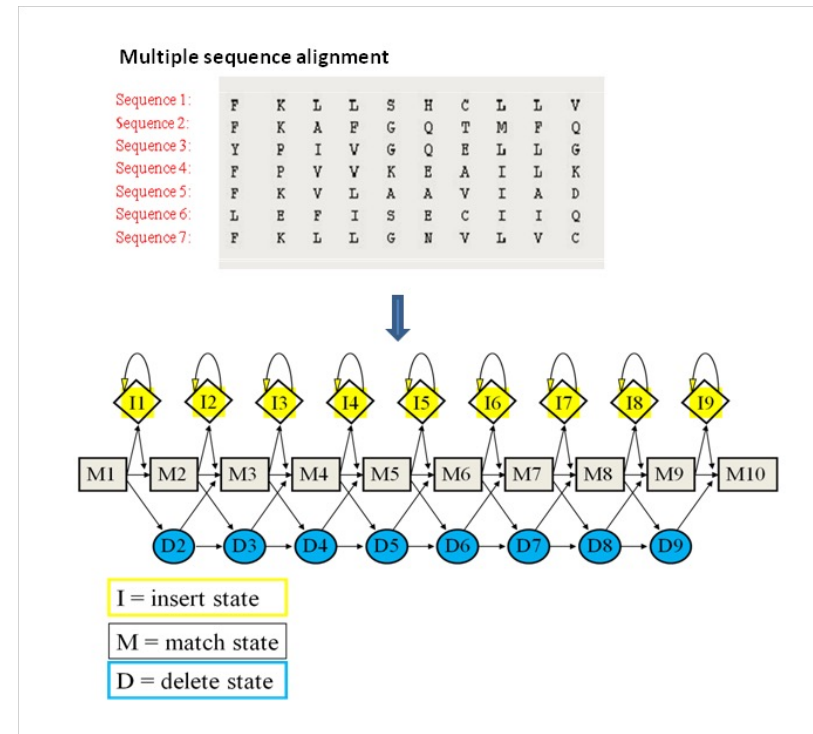
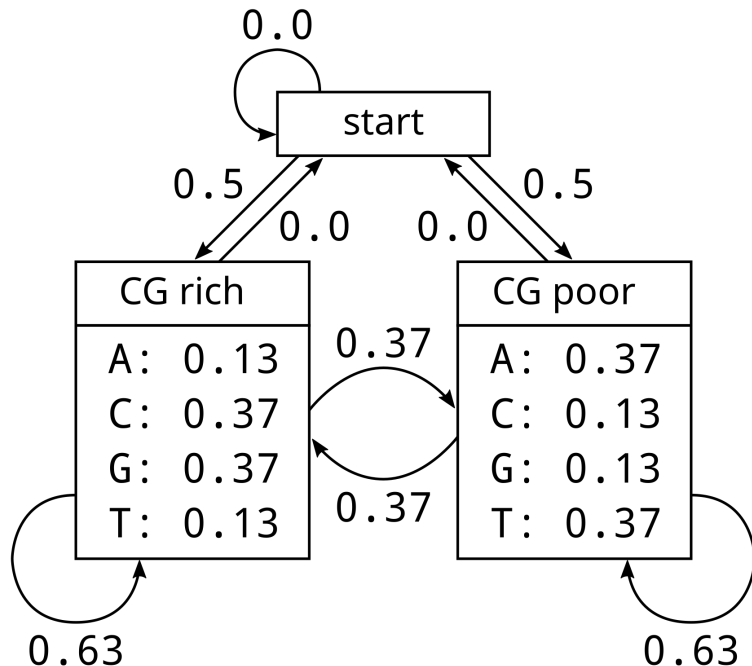
- **Reconocimiento de voz** 🎤 (los sonidos que escuchamos provienen de estados ocultos en las cuerdas vocales).
- **Finanzas** 📈 (inferir tendencias del mercado).
- **Análisis de ADN** 🧬 (detectar genes según patrones en la secuencia).



Modelos estadísticos que describen sistemas con estados ocultos.

En bioinformática: se usan para analizar secuencias de ADN.

- **Estados ocultos:** Ej. regiones codificantes (exones) y no codificantes (intrones).
- **Observaciones:** Bases nucleotídicas (A, T, C, G).
- **Probabilidades de transición:** Ej. paso de intrón a exón.
- **Probabilidades de emisión:** Probabilidad de observar una base en un estado específico.



Imagine you are a detective living in Santiago trying to infer the weather in a La Serena, but you can only see how people dress.

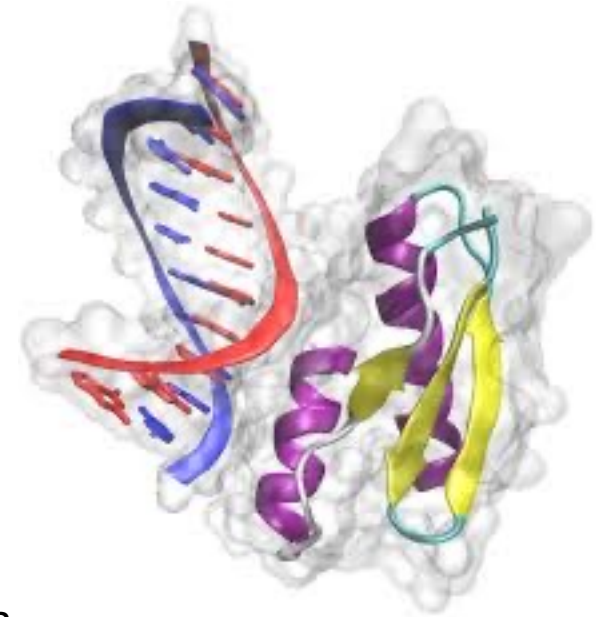
Hidden States: Weather (Sunny 🌞, Rainy 🌧️)

Observations: Clothing (T-shirt 👕, Coat 🧥, Umbrella 🌂)

Clasificación de Proteínas con HMMs: protein quinasas

Estado Dominio Quinasa:

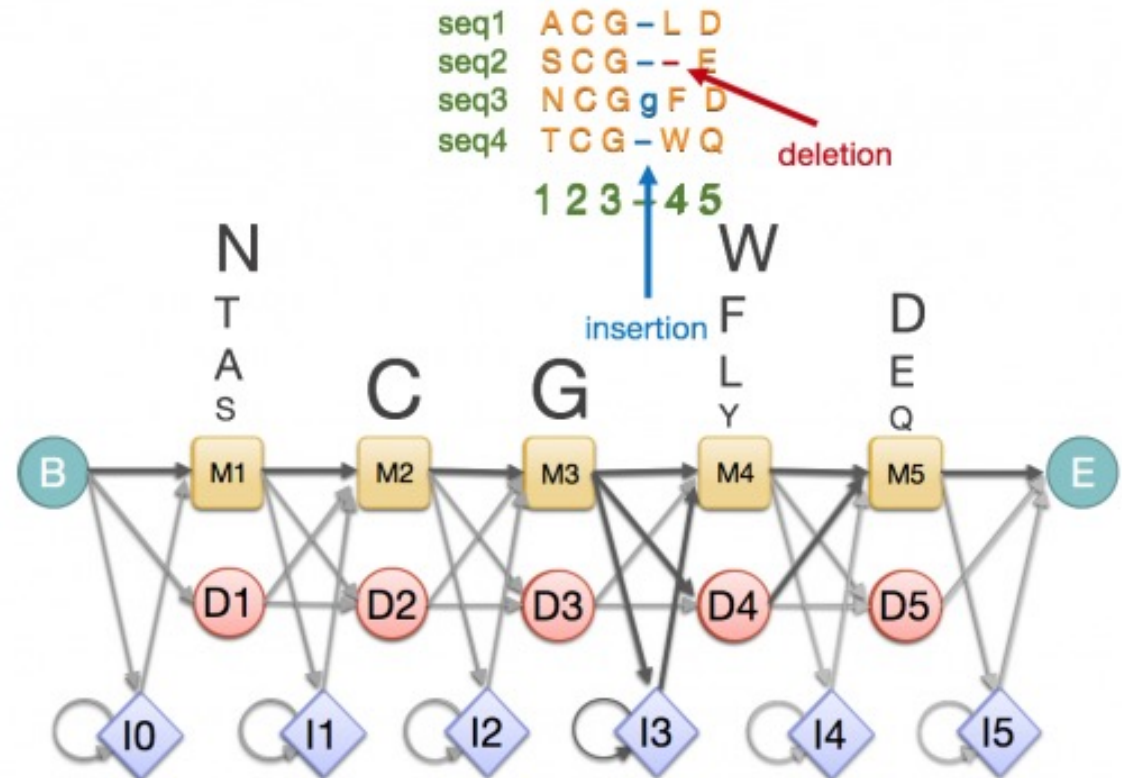
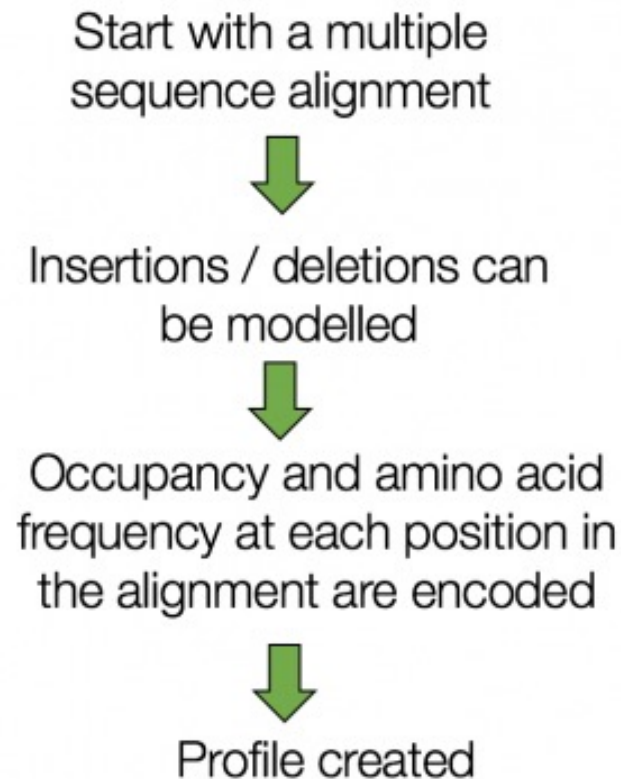
$P(\text{Ser}) = 0.25$, $P(\text{Thr}) = 0.20$, $P(\text{Lys}) = 0.15$, $P(\text{Asp}) = 0.15$, $P(\text{Otros}) = 0.25$



HMMs pueden ser entrenados con una base de datos de proteínas ya clasificadas y luego utilizados para analizar nuevas proteínas.

HIDDEN MARKOV MODELS (HMMs)

One of the computational algorithms used for predicting protein structure and function, identifies significant protein sequence similarities allowing the detection of homologs and consequently the transfer of information.

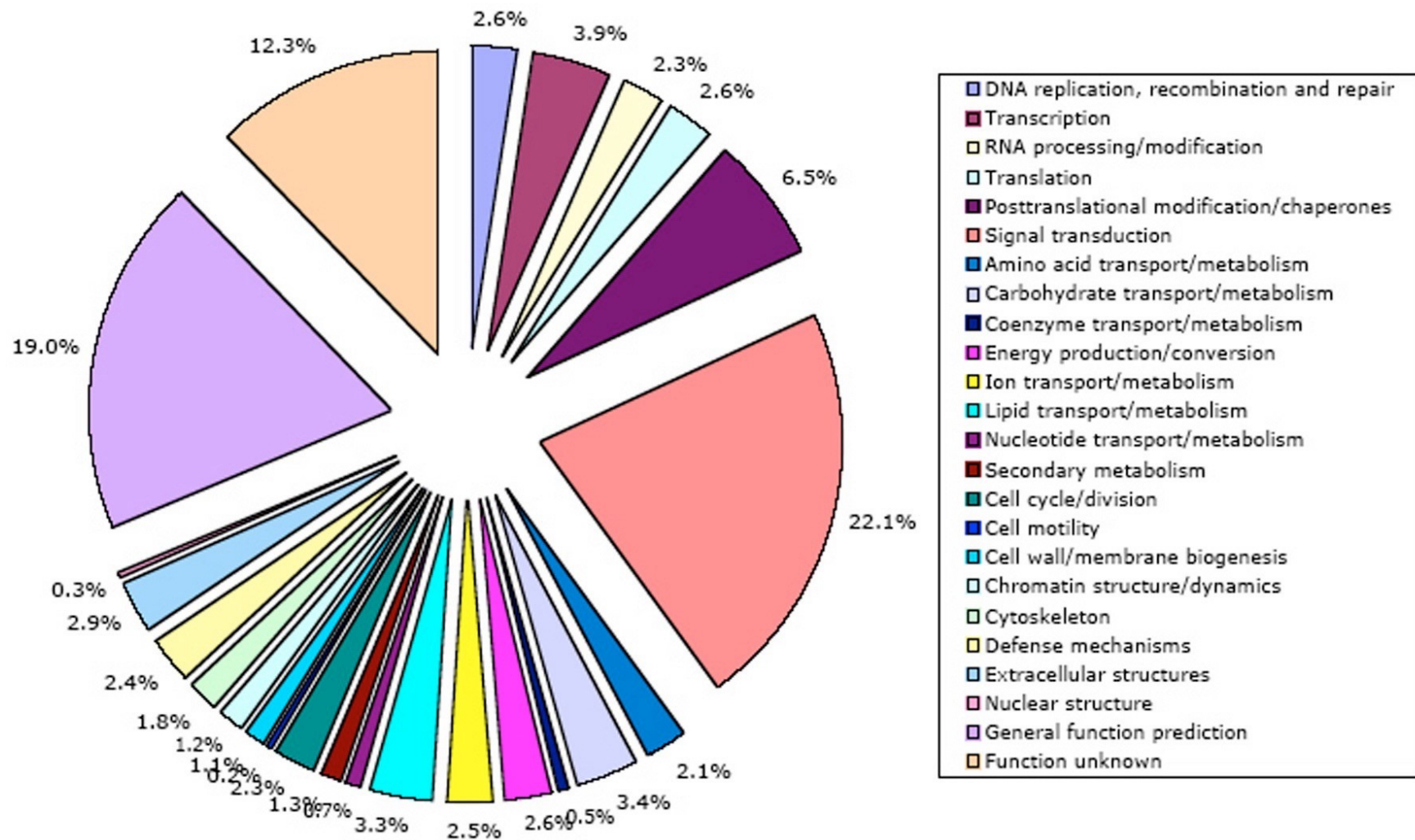


THE COG (CLUSTERS OF ORTHOLOGOUS GROUPS) DATABASE

- The protein database of Clusters of Orthologous Groups (COGs) is an attempt to phylogenetically classify the complete complement of proteins (both predicted and characterized) encoded by complete genomes.
- Each COG is a group of three or more proteins, i.e., they are direct evolutionary counterparts.

Proteínas ortólogas son aquellas que comparten una función y secuencia similar porque han evolucionado de un gen ancestral común mediante especiación

THE COG DATABASE CAN BE USED TO CLASSIFY LARGE SETS OF GENES INTO FUNCTIONAL CATEGORIES



KEGG: KYOTO ENCYCLOPEDIA OF GENES AND GENOMES

- KEGG is a collection of biological information compiled from published material curated database.
- Includes information on genes, proteins, metabolic pathways, molecular interactions, and biochemical reactions associated with specific organisms
- Provides a relationship (map) for how these components are organized in a cellular structure or reaction pathway.



KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions and relations

[KEGG2](#) [PATHWAY](#) [BRITE](#) [MODULE](#) [KO](#) [GENES](#) [COMPOUND](#) [NETWORK](#) [DISEASE](#) [DRUG](#)

Select prefix

map

Organism

Enter keywords

Go

[Help](#)

[\[New pathway maps | Update history \]](#)

Pathway Maps

KEGG PATHWAY is a collection of manually drawn [pathway maps](#) representing our knowledge of the molecular interaction, reaction and relation networks for:

1. Metabolism

[Global/overview](#) [Carbohydrate](#) [Energy](#) [Lipid](#) [Nucleotide](#) [Amino acid](#) [Other amino](#) [Glycan](#)
[Cofactor/vitamin](#) [Terpenoid/PK](#) [Other secondary metabolite](#) [Xenobiotics](#) [Chemical structure](#)

2. Genetic Information Processing

3. Environmental Information Processing

4. Cellular Processes

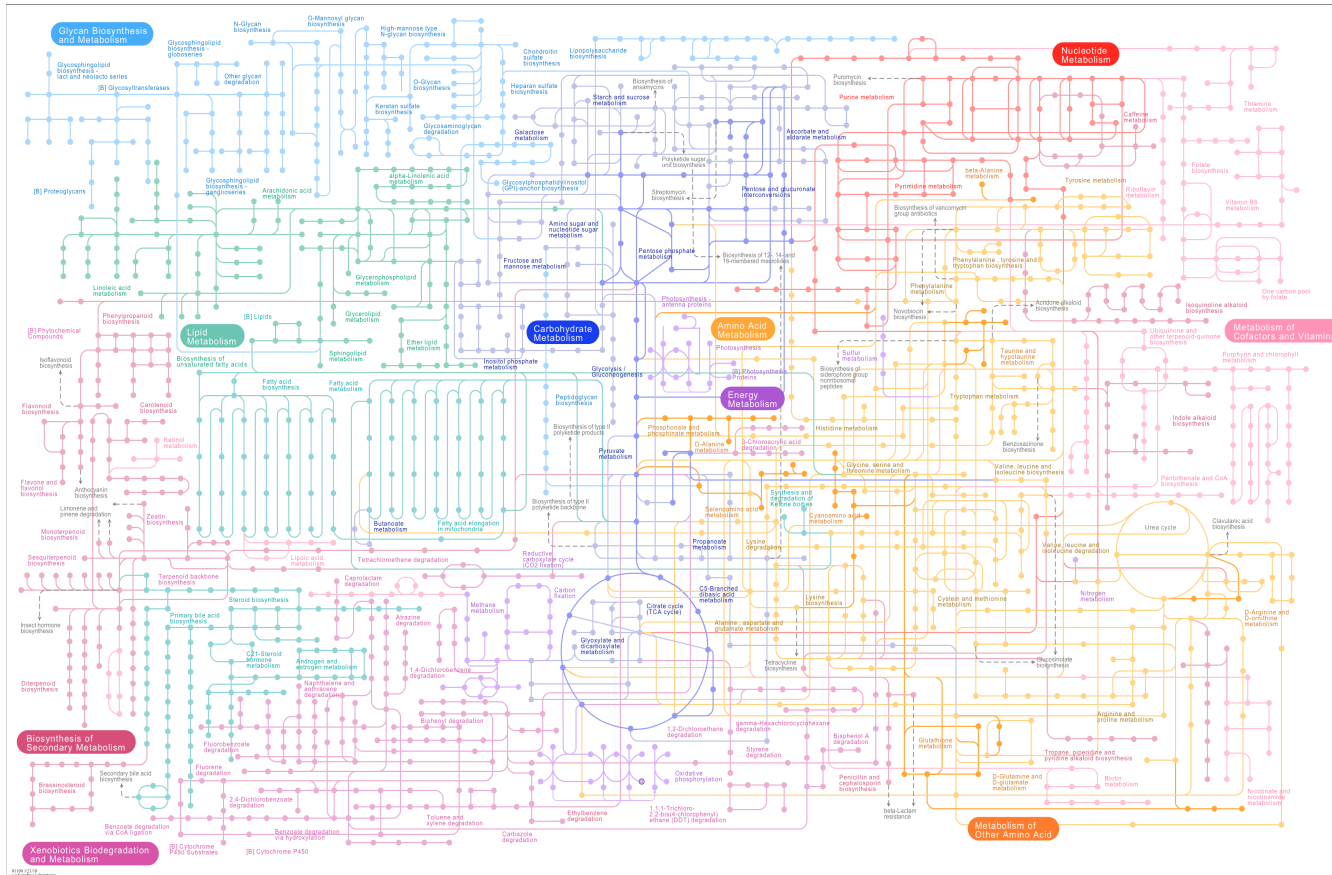
5. Organismal Systems

6. Human Diseases

7. Drug Development

The pathway map viewer linked from this page contains features of [KEGG mapping](#), especially for coloring map objects as described [here](#).

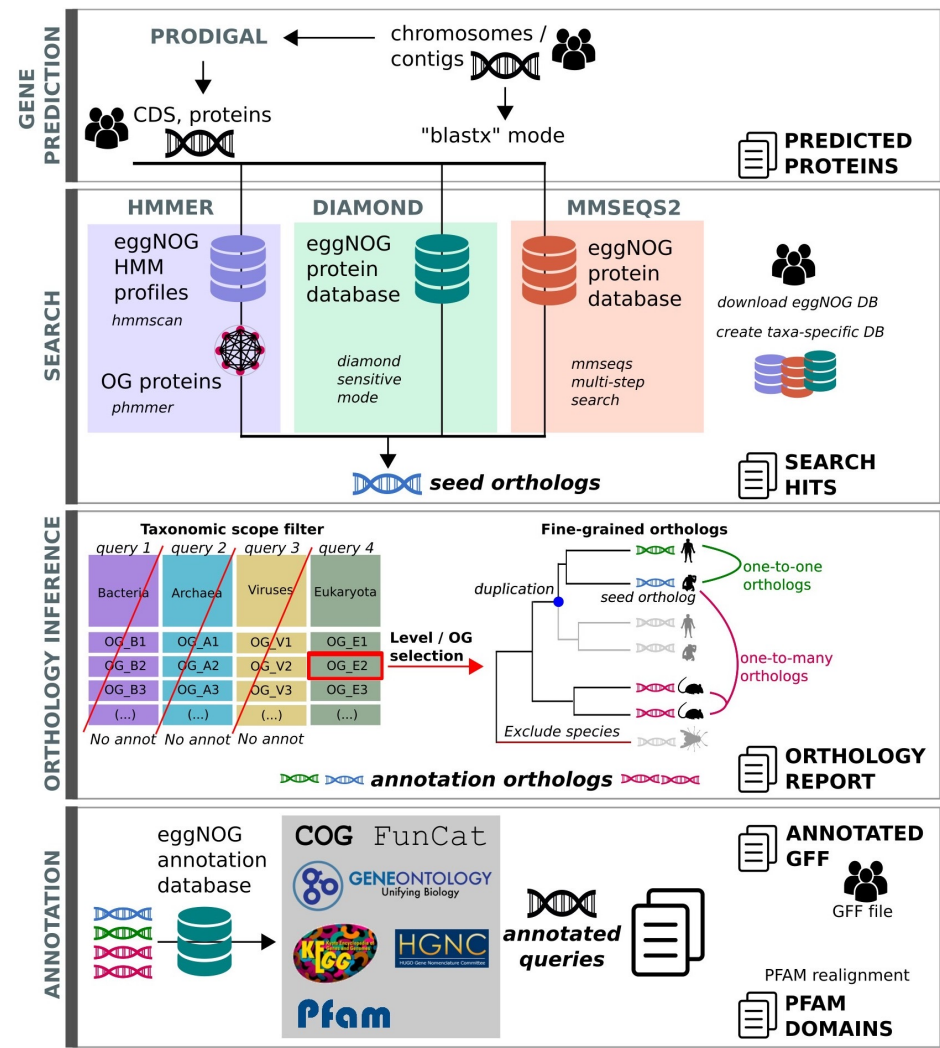
KEGG: A BROAD OVERVIEW OF KNOWN METABOLIC PATHWAYS



Very good for common (primary metabolic) pathways, often incomplete for specialized primary and secondary metabolism

<https://www.genome.jp/kegg/>

EGGNOG: ASSIGNMENT OF PFAM, COG, KEEG AND GO ANNOTATIONS



Annotate a file

What kind of data?

☐ Proteins ☐ CDS ☐ Genomic ☒ Metagenomic

Up to 1,000 contigs in FASTA format (max total nucleotides: 10,000,000).

Gene prediction method

☒ Prodigal ☐ Blastx-like

Proteins predicted by Prodigal will be used for searching.

Upload sequences

Files may be compressed in gzip format (file name must end in '.gz')

Bladeren... Geen bestand geselecteerd.

Email address (Required for job scheduling and notifications)

Enter email

Advanced Options

Search filters

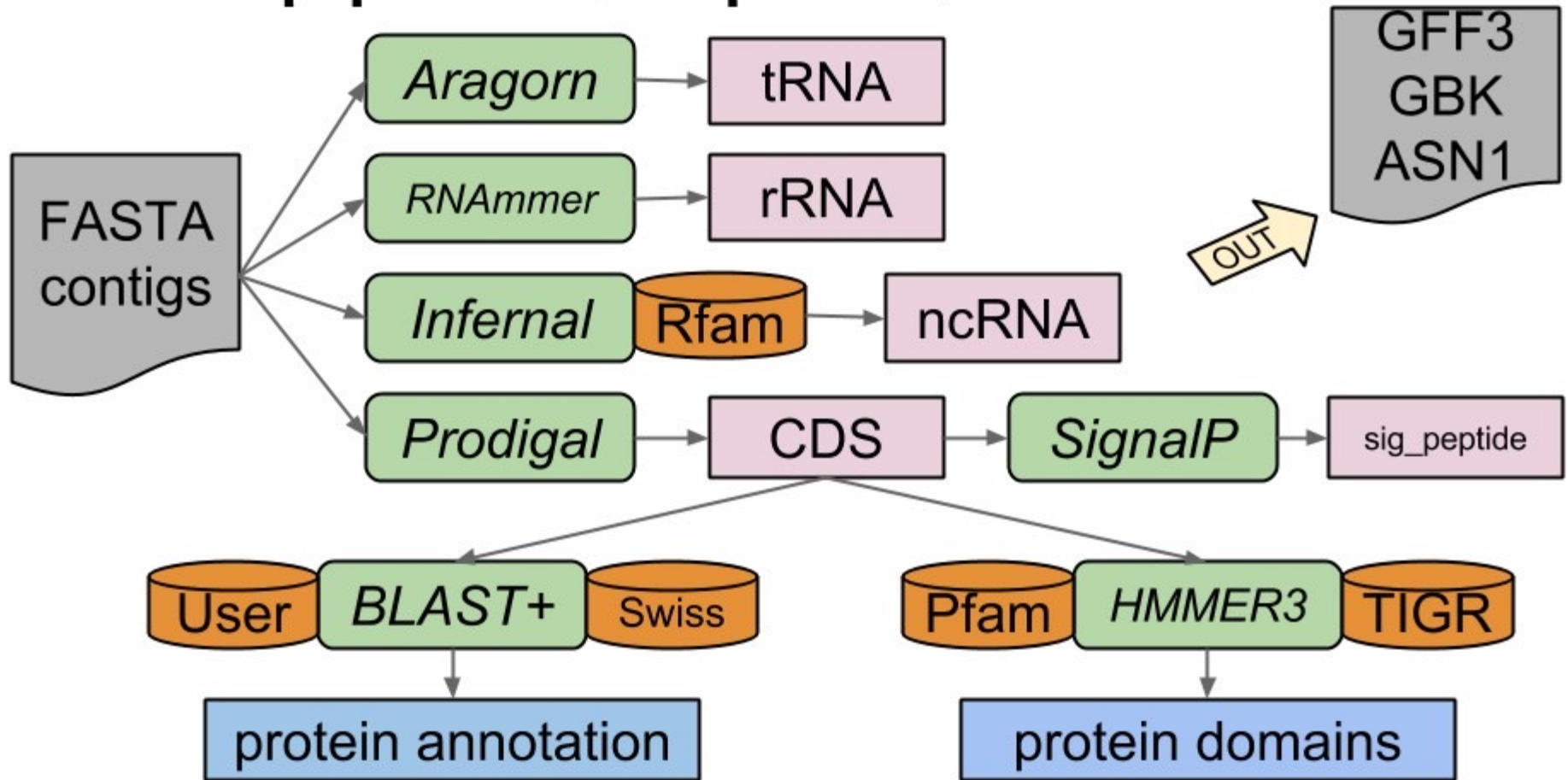
Annotation options

Database

<http://eggno5.embl.de>
<http://eggno-mapper.embl.de/>

PROKKA: AUTOMATED PIPELINE FOR GENOMES ANNOTATION

Prokka pipeline (simplified)



SPECIFIC FUNCTIONAL ANNOTATION TOOLS

THE CAZY DATABASE CLASSIFIES CARBOHYDRATE-ACTING ENZYMES, DBCAN PROVIDES LIBRARIES OF HMMs TO DETECT THEM

Bacteroides thetaiotaomicron DSM 2079

Taxonomy ID : [818](#)

Lineage: cellular organisms; Bacteria; FCB group; Bacteroidetes/Chlorobi group; Bacteroidetes; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides

Glycoside Hydrolase Family	2	3	13	16	18	20	23	25	27	28	29	30	31	32	33	35	36
Number of sequences	31	10	7	3	12	14	3	1	5	9	9	3	6	4	2	3	4

GlycosylTransferase Family	1	2	3	4	5	8	9	14	19	25	28	30	32	35	51	101	NC
Number of sequences	1	40	1	26	1	1	1	3	1	1	1	1	3	2	4	2	5

Polysaccharide Lyase Family	1	8	9	10	11	12	13	15	26	27	29	33	40	42
Number of sequences	5	3	2	1	1	2	1	1	1	1	1	2	1	1

Carbohydrate Esterase Family	2	4	6	7	8	9	11	12	19	20	NC
Number of sequences	1	1	1	1	2	3	1	4	1	9	1

Carbohydrate-Binding Module Family	6	20	32	35	50	51	57	58	91	93	NC
Number of sequences	1	2	13	2	7	1	1	1	5	1	5



List Of Proteins

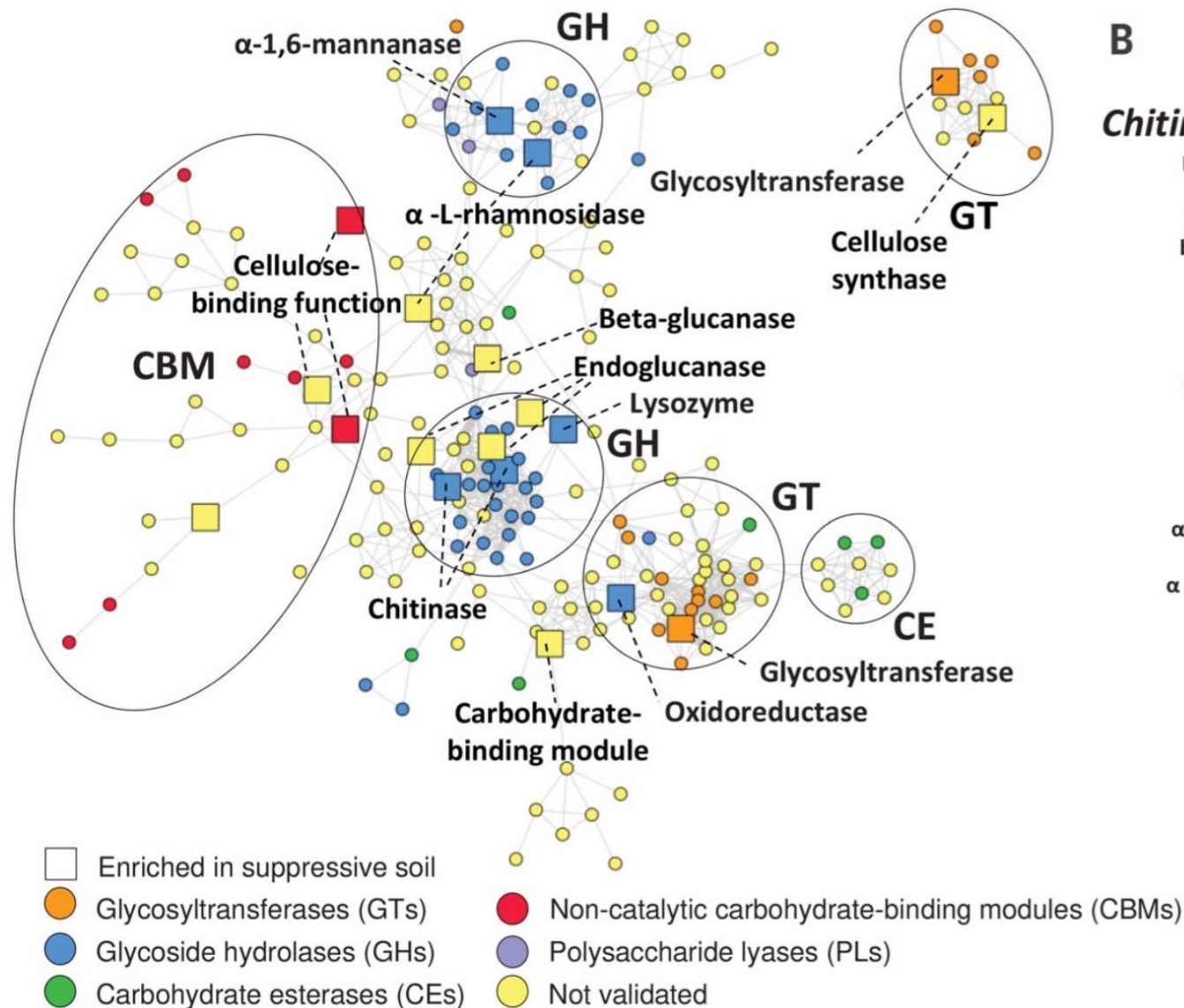
Protein Name	Family	Reference Accession
FE838_00230	CBM20,CBM20,GH77	QMW84610.1
FE838_00240	GT4	QMW84612.1
FE838_00255	GT2	QMW84615.1
FE838_00455	GH29	QMW84652.1
FE838_00465	CBM32	QMW89084.1
FE838_00480	GH2	QMW84658.1



<http://www.cazy.org>

<https://bcb.unl.edu/dbCAN2>

THE dBCAN DATABASE (USING HMMs): EXAMPLE USE CASE



B

Chitin

Be

Ca

bir

(F

En

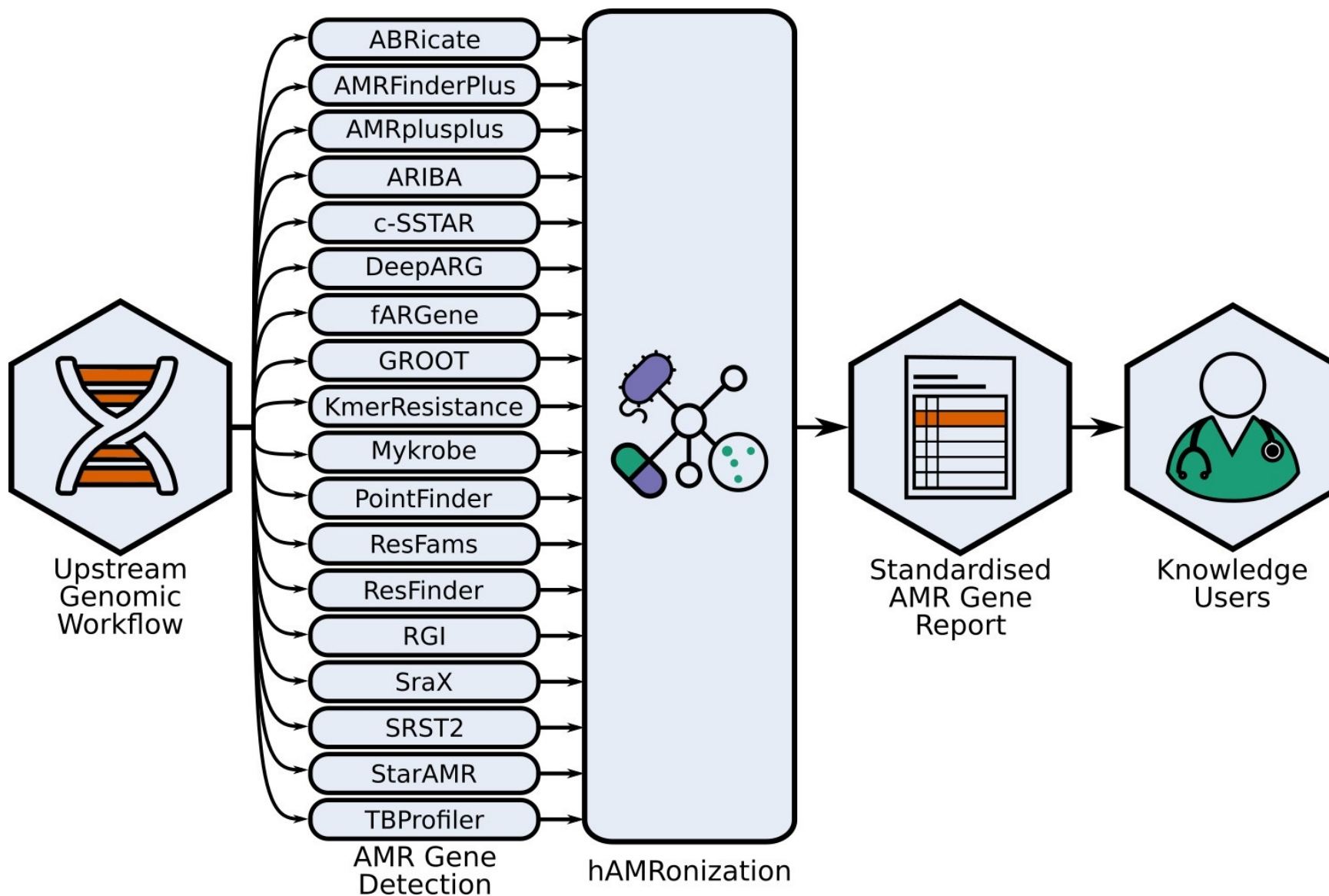
(F

α-1

α-L

(F

IDENTIFYING ANTIMICROBIAL RESISTANCE



BUT ... WHY ... SO MANY ANNOTATIONS????????

COG							PFAM		KEGG		
Gen_id	COG.descr	COG.typ	COG.ID	pfam_dom	pfam_descr	kegg	kegg_d				
Gen 1	Ribosomal_protein_S18_acetylase_RimI_and_related_acety		COG0456	PF00583	Acetyltransferase_(GN K03395	aac3-l;_ar					
Gen 2	Ribosomal_protein_S18_acetylase_RimI_and_related_acety		COG0456	PF00583	Acetyltransferase_(GN K03395	aac3-l;_ar					
Gen 3	Ribosomal_protein_S18_acetylase_RimI_and_related_acety		COG0456	PF00583	Acetyltransferase_(GN K03395	aac3-l;_ar					
Gen 4	Protein_N-acetyltransferase,_RimJ/RimL_family	JO	COG1670	PF13523	Acetyltransferase_(GN K00663	aacA;_ami					
Gen 5	Aminoglycoside_N3'-acetyltransferase	V	COG2746	PF02522	Aminoglycoside_3-N-a K00662	aacC;_ami					
Gen 6	Aminoglycoside_N3'-acetyltransferase	V	COG2746	PF02522	Aminoglycoside_3-N-a K00662	aacC;_ami					
Gen 7	Aminoglycoside_N3'-acetyltransferase	V	COG2746	PF02522	Aminoglycoside_3-N-a K00662	aacC;_ami					
Gen 8	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 9	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 10	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 11	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 12	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 13	Acyl-coenzyme_A_synthetase/AMP-(fatty)_acid_ligase	I	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs					
Gen 14	A1	S1	S2	S3	SR1	SR2	SR3	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs
Gen 15	A1	68	13	37	28	27	34	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs
Gen 16	A2	79	47	55	41	37	26	COG0365	PF16177;PF00501;PF1	Acetyl-coenzyme_A_s K01907	AACS,_acs
Gen 17	A3	23	33	11	24	24	16	COG0365	PF00501;PF13193	AMP-binding_enzyme K01907	AACS,_acs
Gen 18	A3	23	33	11	24	24	16	COG0365	PF00501;PF13193	AMP-binding_enzyme K01907	AACS,_acs
Gen 19	A4	24	18	17	14	17	9	COG3321	PF00975	Thioesterase_domain K01907	AACS,_acs
Gen 20	F5	33	9	26	105	68	20	COG1708	PF13427;PF01909	Domain_of_unknown K00984	aadA;_stre
Gen 21	F6	16	14	52	24	26	20	COG1708	PF13427;PF01909	Domain_of_unknown K00984	aadA;_stre
Gen 22	F7	4	19	8	18	9	6	COG1708	PF13427;PF01909	Domain_of_unknown K00984	aadA;_stre
Gen 23	F5	190	165	147	101	160	50	COG1708	PF01909	Nucleotidyltransferase K00984	aadA;_stre
Gen 24	F5	190	165	147	101	160	50	COG0657	PF07859	alpha/beta_hydrolase K13616	AADAC;_a

COUNT/ABUNDANCE
TABLE

10	27	67	13	18	16	29
11	86	45	38	94	164	15
12	724	527	475	635	560	372
13	154	463	507	542	508	70
16	112	58	45	37	45	44
17	161	326	221	562	470	68
18	21	235	68	41	92	35

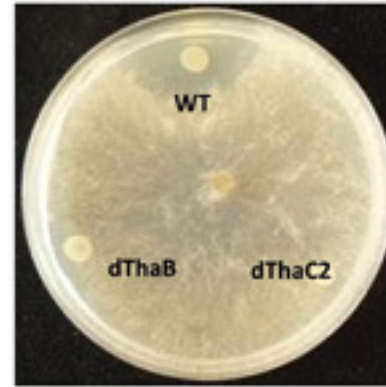
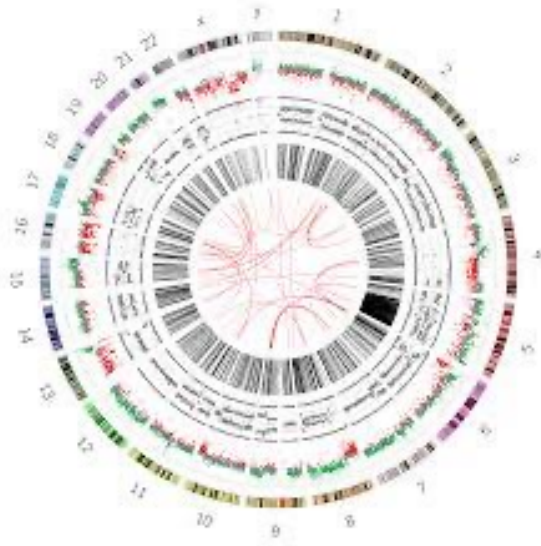
SampleID	Treatment	Description	Fungi	Status
C1	Endo_Conductive	Endosphere	No_Rhizo	Conductive
C2	Endo_Conductive	Endosphere	No_Rhizo	Conductive
C3	Endo_Conductive	Endosphere	No_Rhizo	Conductive
C4	Endo_Conductive	Endosphere	No_Rhizo	Conductive
S1	Endo_Suppressive	Endosphere	No_Rhizo	Suppressive
S2	Endo_Suppressive	Endosphere	No_Rhizo	Suppressive
S3	Endo_Suppressive	Endosphere	No_Rhizo	Suppressive
SR1	Endo_SuppressiveR	Endosphere	Rhizo	Suppressive
SR2	Endo_SuppressiveR	Endosphere	Rhizo	Suppressive
SR3	Endo_SuppressiveR	Endosphere	Rhizo	Suppressive

METADATA

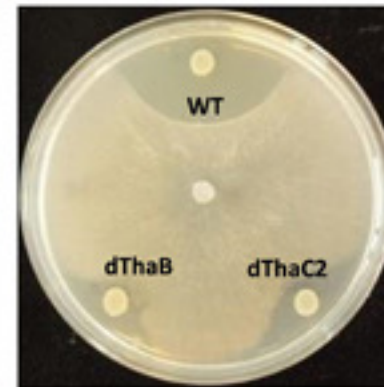


S. Rico
IBFG

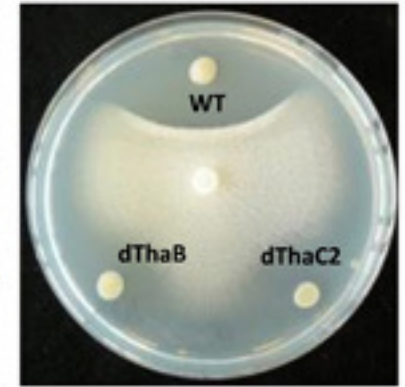
SECONDARY METABOLISM



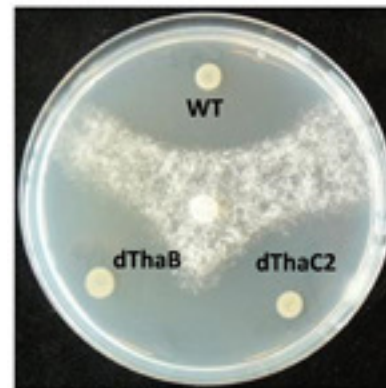
R. solani



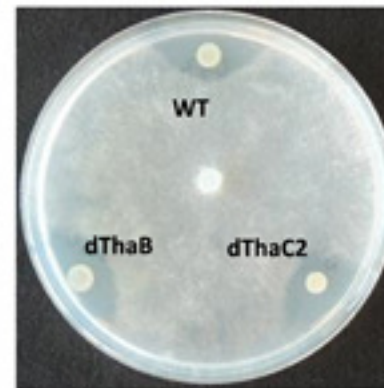
B. cinerea



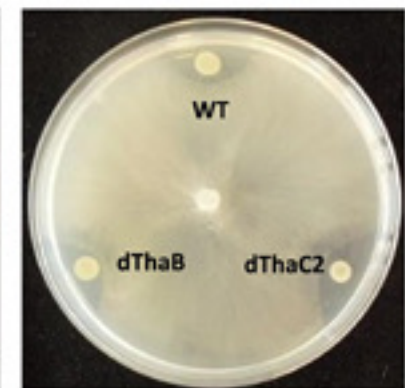
Geotrichum sp.



P. infestans



P. ultimum



P. capsici

ANTISMASH: A WEB SERVER FOR THE DETECTION AND ANALYSIS OF BIOSYNTHETIC GENE CLUSTERS



Web server and a stand-alone software to identify, annotate, and compare gene clusters that encode the biosynthesis of secondary metabolites in bacterial and fungal genomes (Medema et al. 2011)

Table 1 Examples of secondary metabolite types

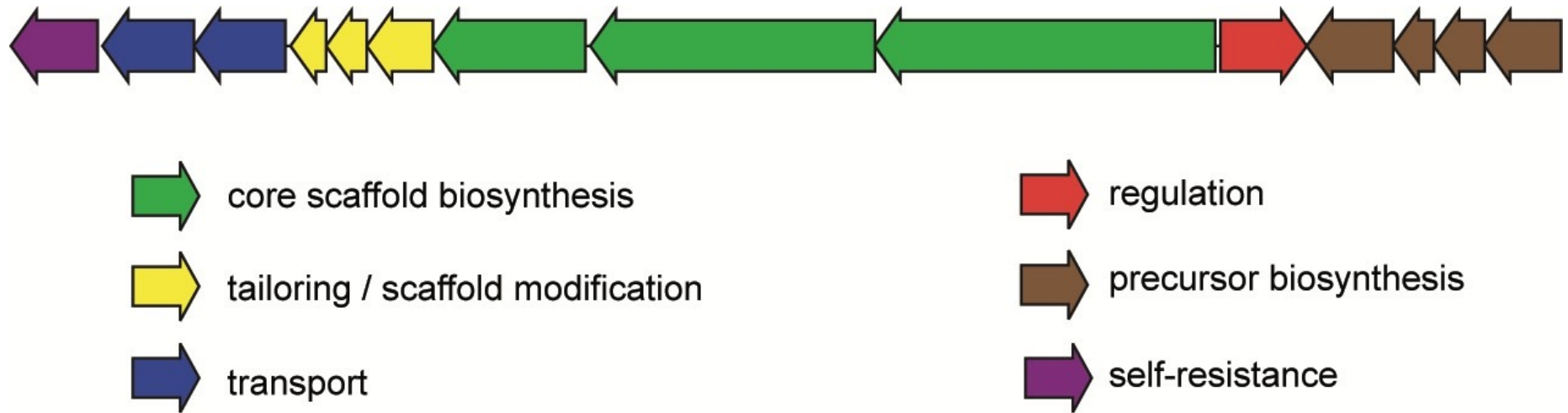
Secondary metabolite type	No. records*	Example	Reference
PK	799	Erythromycin	Rawlings (2001)
NRP	605	Penicillin	Fierro <i>et al.</i> (2006)
RiPP	258	Bottromycin	Shimamura <i>et al.</i> (2009)
Saccharide	173	Streptomycin	Ohnishi <i>et al.</i> (2008)
Terpene	130	Mycophenolic acid	Regueira <i>et al.</i> (2011)
Alkaloid	49	Violacein	Hoshino (2011)
Other	253	Clavulanic acid	Aidoo <i>et al.</i> (1994)

PK, polyketide; NRP, nonribosomal peptide; RiPPS, ribosomally synthesized and post-translationally modified peptide.

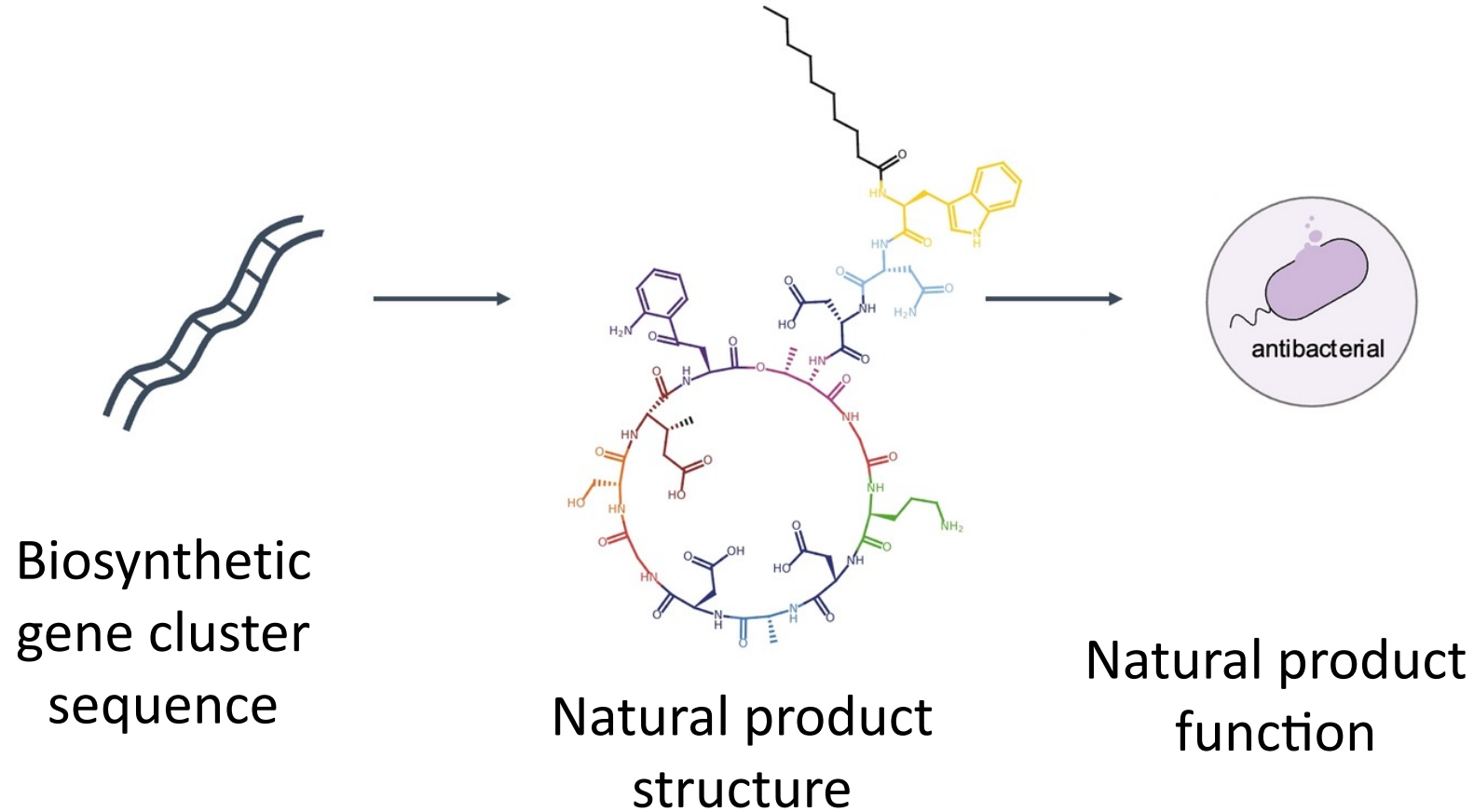
*Number of gene clusters related to each type of secondary metabolite in the MIBIG platform (<https://mibig.secondarymetabolites.org/stats>).

BIOSYNTHETIC GENE CLUSTERS (BGCs) COMPRISE SEVERAL TYPES OF GENE FUNCTIONS

A 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).

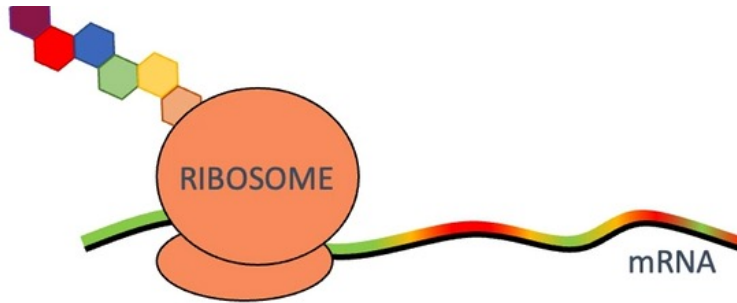


THE “CENTRAL DOGMA” OF SPECIALIZED METABOLISM



RIBOSOMAL PEPTIDE SYNTHESIS

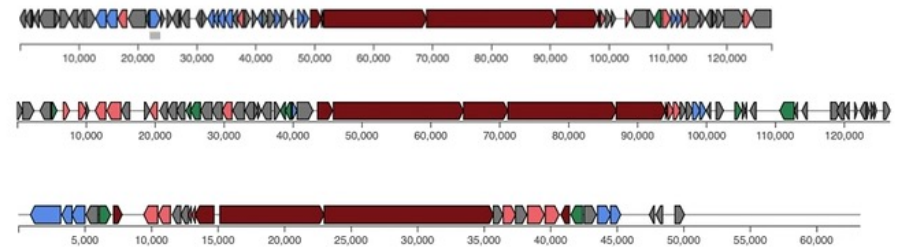
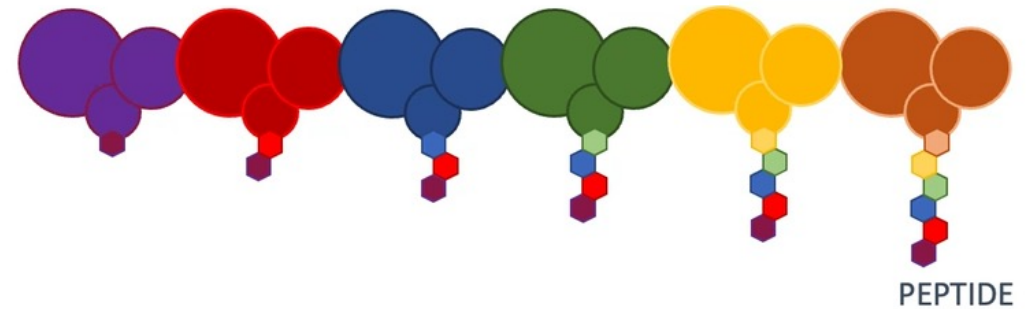
PEPTIDE



Peptide sequence
depends on template

NON-RIBOSOMAL PEPTIDE SYNTHESIS

NON-RIBOSOMAL PEPTIDE
SYNTHETASE



Peptide sequence depends on
machinery



Server status:	working
Running jobs:	20
Queued jobs:	0
Jobs processed:	1429532

Nucleotide input

Results for existing job

Search a genome sequence for secondary metabolite biosynthetic gene clusters

Load sample input

Open example output

Notification settings

your@email.com

Email address (optional)

antiSMASH beta features

☒ Enable antiSMASH beta

Data input

Upload file

Get from NCBI

NCBI acc #

NCBI accession number of desired sequence

Extra features

All off

All on

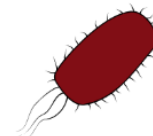
☒ KnownClusterBlast☐ ClusterBlast☒ SubClusterBlast☐ MIBiG cluster comparison☒ ActiveSiteFinder☒ RREFinder☐ Cluster Pfam analysis☐ Pfam-based GO term annotation☐ TIGRFam analysis☒ TFBS analysis

Submit

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



If you have found antiSMASH useful, please [cite us](#).

**antiSMASH 7 beta is now available**

Dear antiSMASH users, we have just made antiSMASH 7 available as a beta version. We're very happy with how it works, but if you still want to run the old 6.1.1 release, just uncheck the "enable antiSMASH beta" checkbox.

ANTISMASH: GENOME

antiSMASH version 6.1.0-800225e Download About Help Contact

Select genomic region:

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

NC_003888 - Region 10 - NRPS ? **Gene details** ?

Location: 3,524,828 - 3,603,907 nt. (total: 79,080 nt) Show pHMM detection rules used Download region SVG Download region GenBank file

Legend:

- core biosynthetic genes
- additional biosynthetic genes
- transport-related genes
- regulatory genes
- other genes
- resistance
- TTA codons

reset view zoom to selection

NRPS/PKS domains **MIBiG comparison** **ClusterBlast** **KnownClusterBlast** **SubClusterBlast** **NRPS/PKS modules** **Pfam domains** **TIGRFAM domains**

Detailed domain annotation ? Download

Selected features only ☐ **Show module domains** ☒

SCO3227 **AmT**

SCO3230 C A C A C A E C A C A E

SCO3231 C A C A C A E

NRPS/PKS products **NRPS/PKS monomers** **NRPS/PKS monomer predictions** ?

SCO3230: ser - thr - trp - asp - asp - hpg +

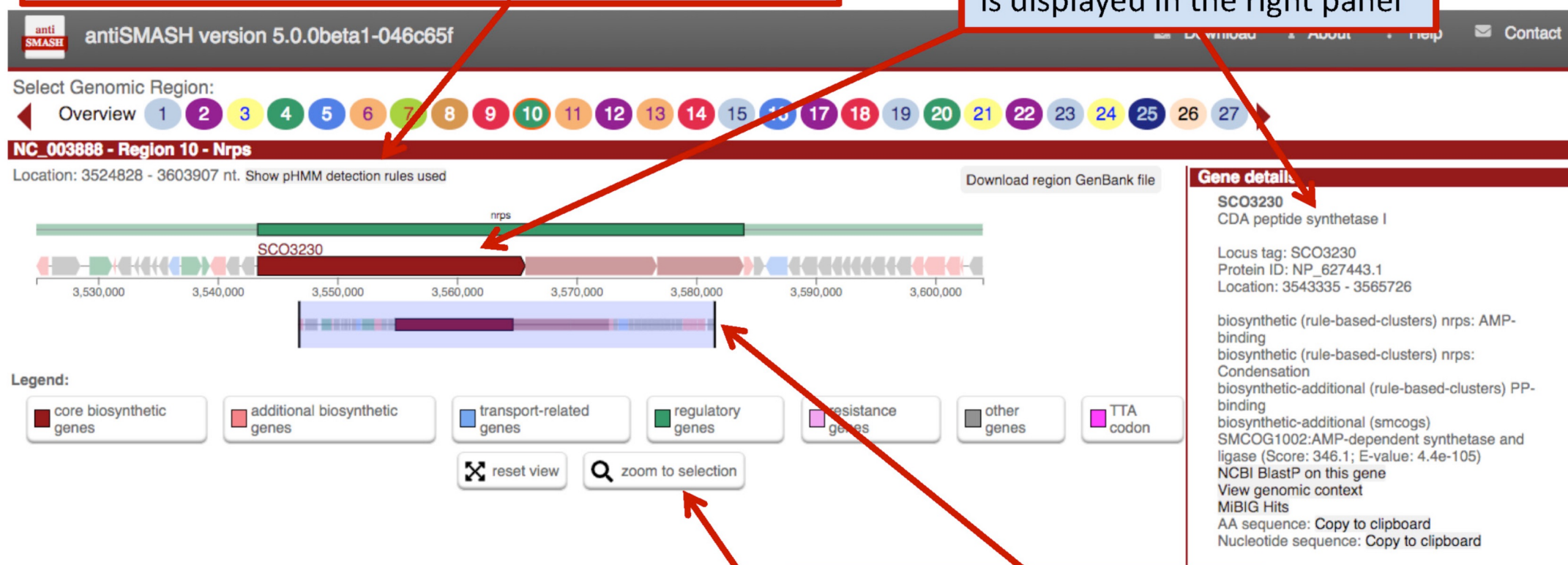
SCO3231: asp - gly - asn +

SCO3232: 3-me-glu - trp +

Medema et al. (2011) *Nucl. Acids Res.* 39: W339-W346.
 Blin, Medema et al. (2013) *Nucl. Acids Res.* 41: W204-W212.
 Weber et al. (2015) *Nucl. Acids Res.* 43: W237-W243.
 Blin et al. (2017) *Nucl. Acids Res.* 45: W36-W41.
 Blin et al. (2019) *Nucl. Acids Res.* 47: W81-W87.
 Blin et al. (2021) *Nucl. Acids Res.* 49: W29-W35.

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



SCO3227

SCO3230

SCO3231

SCO3232

SCO3248

KS

Domain type

Location of domain (with respect to full length protein)

Link to NCBI BlastP

A-domain specificity predictions

copy amino acid /DNA sequence of selected domain to clipboard for copy&pasting to other programs

AMP-binding
Location: 1674-2097 AA
Run BlastP on this domain
Substrate predictions:
-consensus: thr
AA sequence: Copy to clipboard
Nucleotide sequence: Copy to clipboard

Homologous gene clusters

All hits

Download graphic

Query sequence



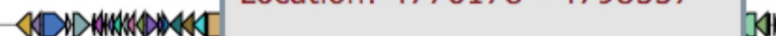
AL939115_c1: *Streptomyces coelicolor* A3(2) complete genome; segment 12/29. (100% of genes show similarity)



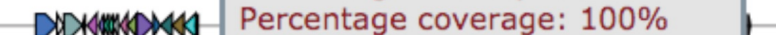
CP009124_c16: *Streptomyces lividans* TK24, complete genome. (94% of genes show similarity)



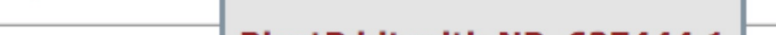
GG657756_c15: *Streptomyces lividans* TK24 ger... (102% of genes show similarity)



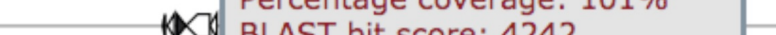
CM001889_c12: *Streptomyces lividans* 1326 chr... (105% of genes show similarity)



JXWX01000029_c1: *Streptomyces* sp. WM6391 c... (47% of genes show similarity)



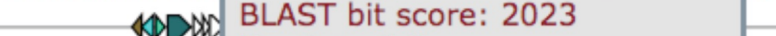
CP006259_c17: *Streptomyces collinus* Tu 365, c... how similarity)



CP012382_c14: *Streptomyces ambofaciens* ATCC... of genes show similarity)



LJCV01000261_c1: Actinobacteria bacterium OK... (23% of genes show similarity)



CP000249_c15: *Frankia* sp. CcI3, complete genome. (28% of genes show similarity)



CP007155_c32: *Kutzneria albida* DSM 43870, complete genome. (23% of genes show similarity)



CDA_peptide_synthetase_I
Location: 4776178 - 4798557

BlastP hit with NP_627443.1
Percentage identity: 98%
Percentage coverage: 100%
BLAST bit score: 14443
E-value: 0.0

BlastP hit with NP_627444.1
Percentage identity: 63%
Percentage coverage: 101%
BLAST bit score: 4242
E-value: 0.0

BlastP hit with NP_733597.1
Percentage identity: 52%
Percentage coverage: 92%
BLAST bit score: 2023
E-value: 0.0



antiSMASH version 5.0.0beta1-046c65f



Download



About



Help



Contact

Select Genomic Region:

Overview 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

NC_003888 - Region 10 - Nrps

Location: 3524828 - 3603907 nt. Show pHMM detection rules used

nrps

Download all results

Download GenBank summary file

Download log file

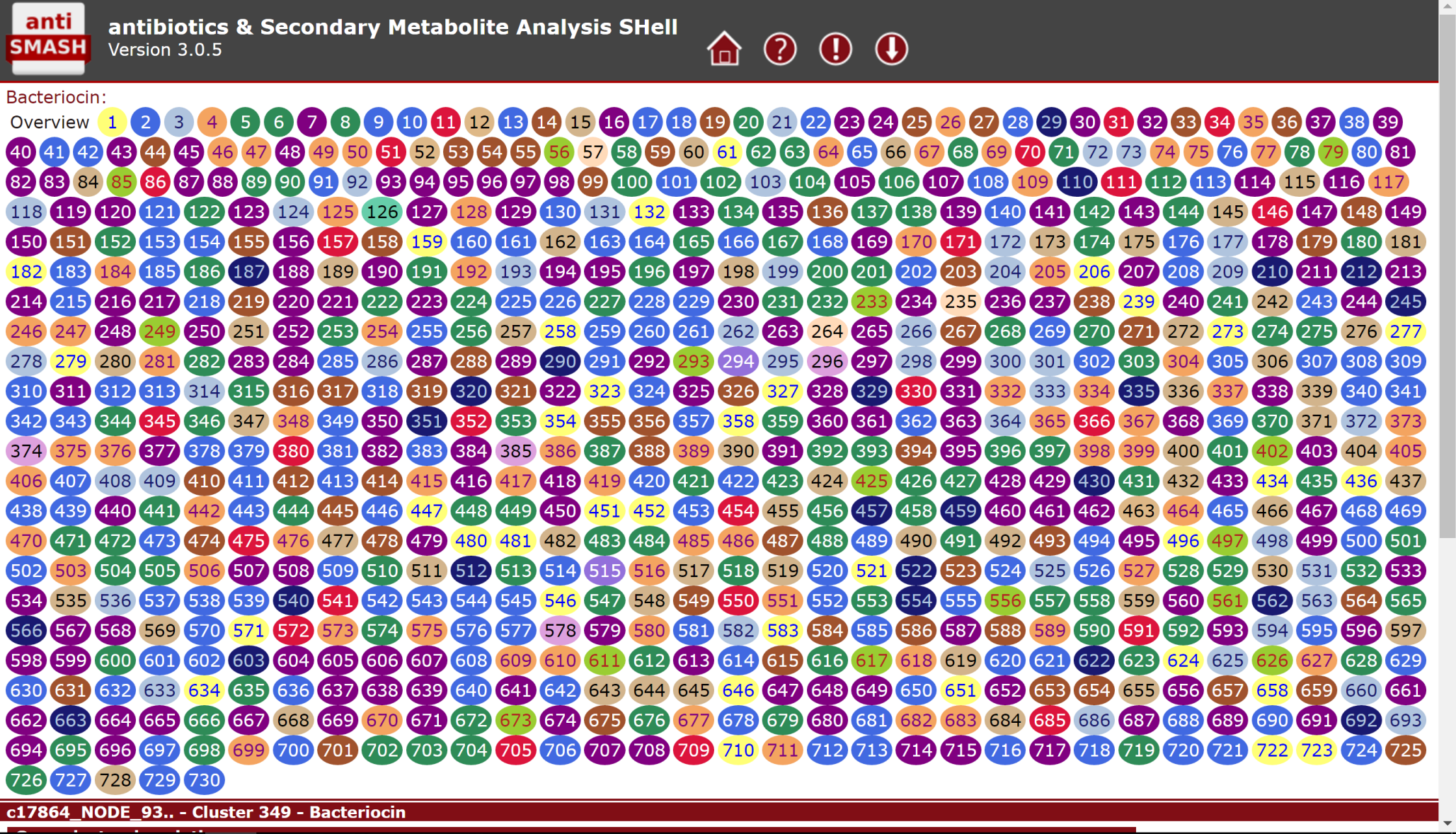
Download

SC03230

CDA peptide synthetase I

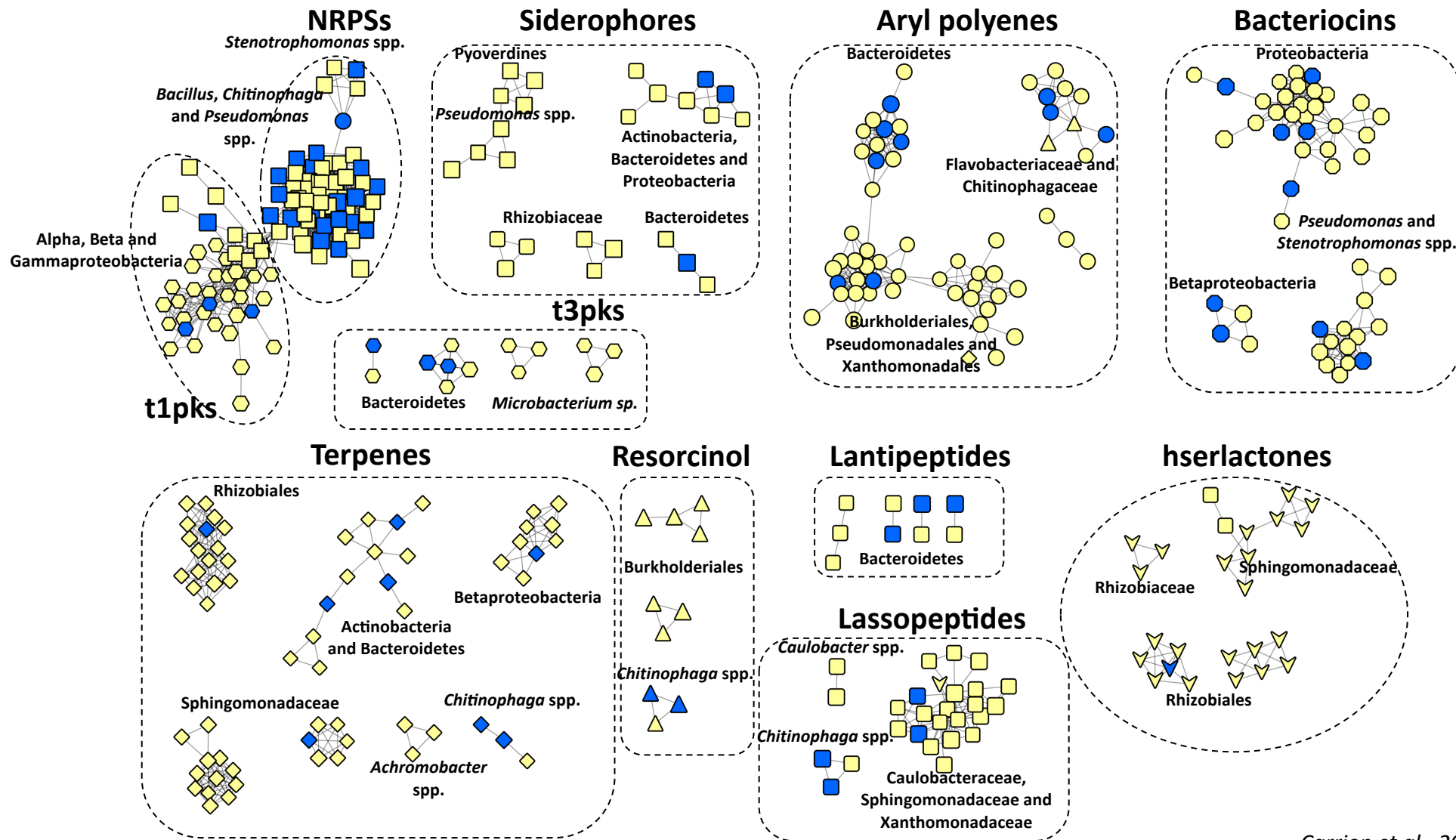
Locus tag: SC03230

ANTISMASH: METAGENOME



Enriched functions: secondary metabolism

(network based on HMM domains of BGCs)





Contact

[View antiSMASH-generated output](#)

Nucleotide sequence: [Copy to clipboard](#)

General information about the BGC

 [Enterobactin: an archetype for microbial iron transport.](#) Raymond KN et al., Proc Natl Acad Sci U S A (2003) PMID:12655062
 [Enterobactin: An archetype for microbial iron transport.](#) Raymond, K et al., Proceedings of the National Academy of Sciences (2003)
 DOI:10.1073/pnas.0630018100

A RICH WEB INTERFACE ALLOWS EFFECTIVELY EXPLORING BIG-SCAPE OUTPUTS



Biosynthetic Genes Similarity Clustering and Prospecting Engine
Version 0.0.0r

Networks: **Overview** PKSother Terpene RPPs NRPS Others PKSI PKS-NRP_Hybrids Saccharides

Runs: 2018-05-04_18-11-47_global_c0.5

Run Information

Analysis Started: 04/05/2018 18:11:47
Parameters:
--pfam_dir ./pfam/ --inputdir
/mnt/scratch/roete009/antismash_output_minimal/
--output
/home/xnava009/public_html/streptomyces_out/
--mibig --hybrids -c 40 --mode lcs -l c0.5 --
cutoffs 0.5 --clan_cutoff 0.5 0.7
Analysis Completed: 04/05/2018 18:23:02 (0h11m15s)

Input Data

Total Number of Genomes: 96
Total BGCs: 3138

Network Overview

PKSother	Terpene	RPPs	NRPS	Others	PKSI	PKS- NRP_Hybrids
Saccharides						
Number of families:						117
Average number of BGCs per family:						2
Max number of BGCs in a family:						28
Families with MIBiG Reference BGCs:						19

GCF absence/presence heatmap

Cluster GCF based on:

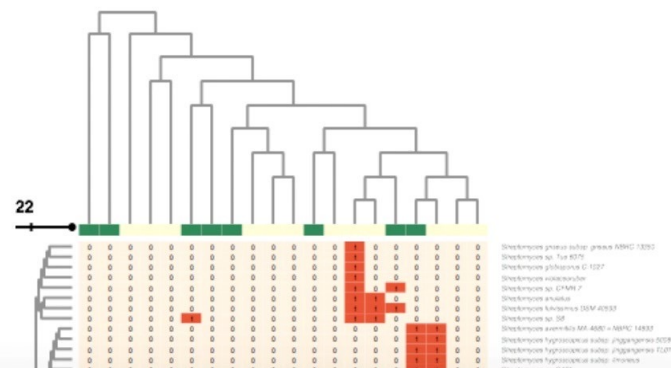
Genomes Absence/Presence

Cluster Genomes based on:

Family Absence/Presence

Show:

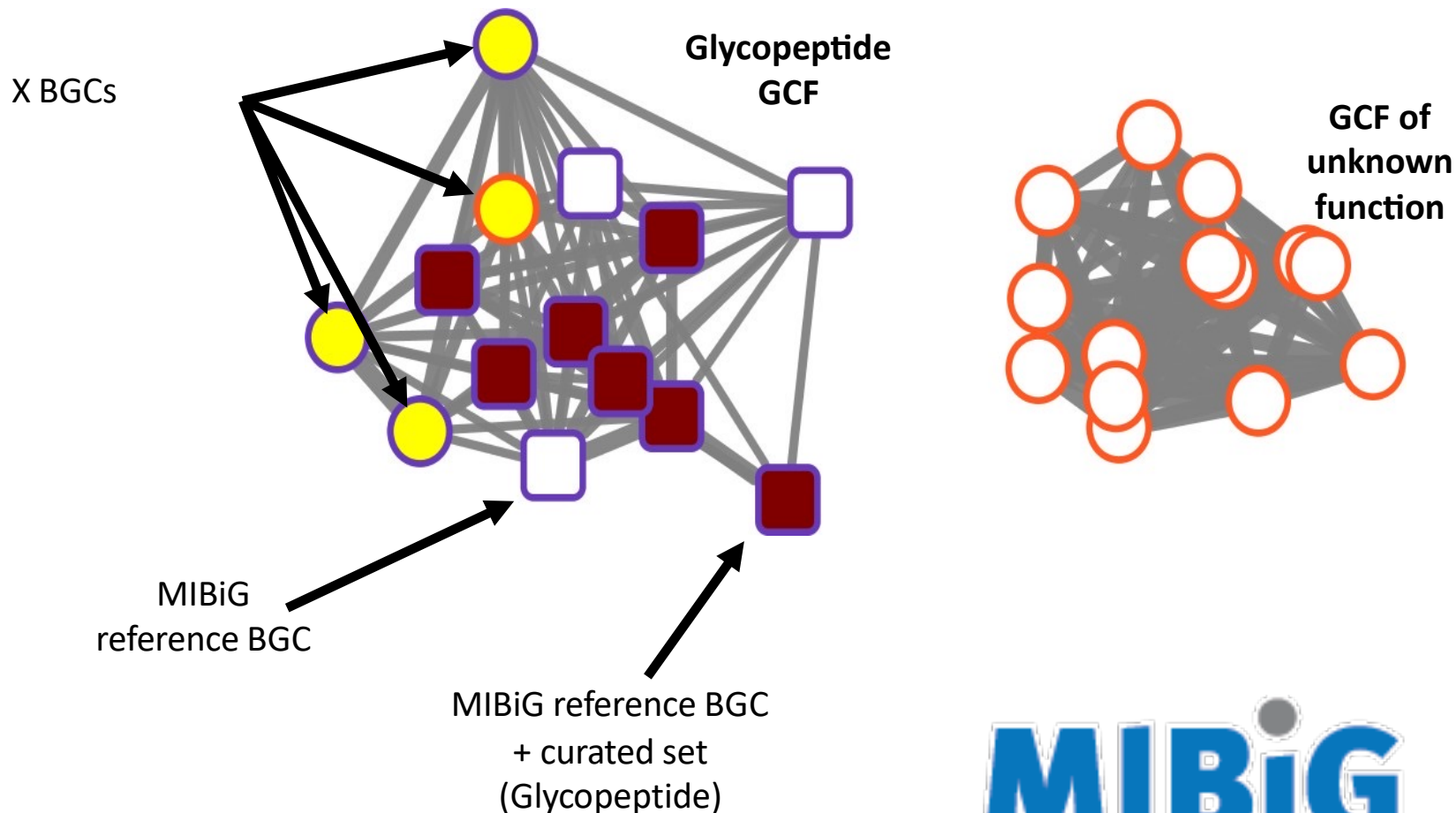
20 largest GCFs



<http://bigscape-corason.secondarymetabolites.org>

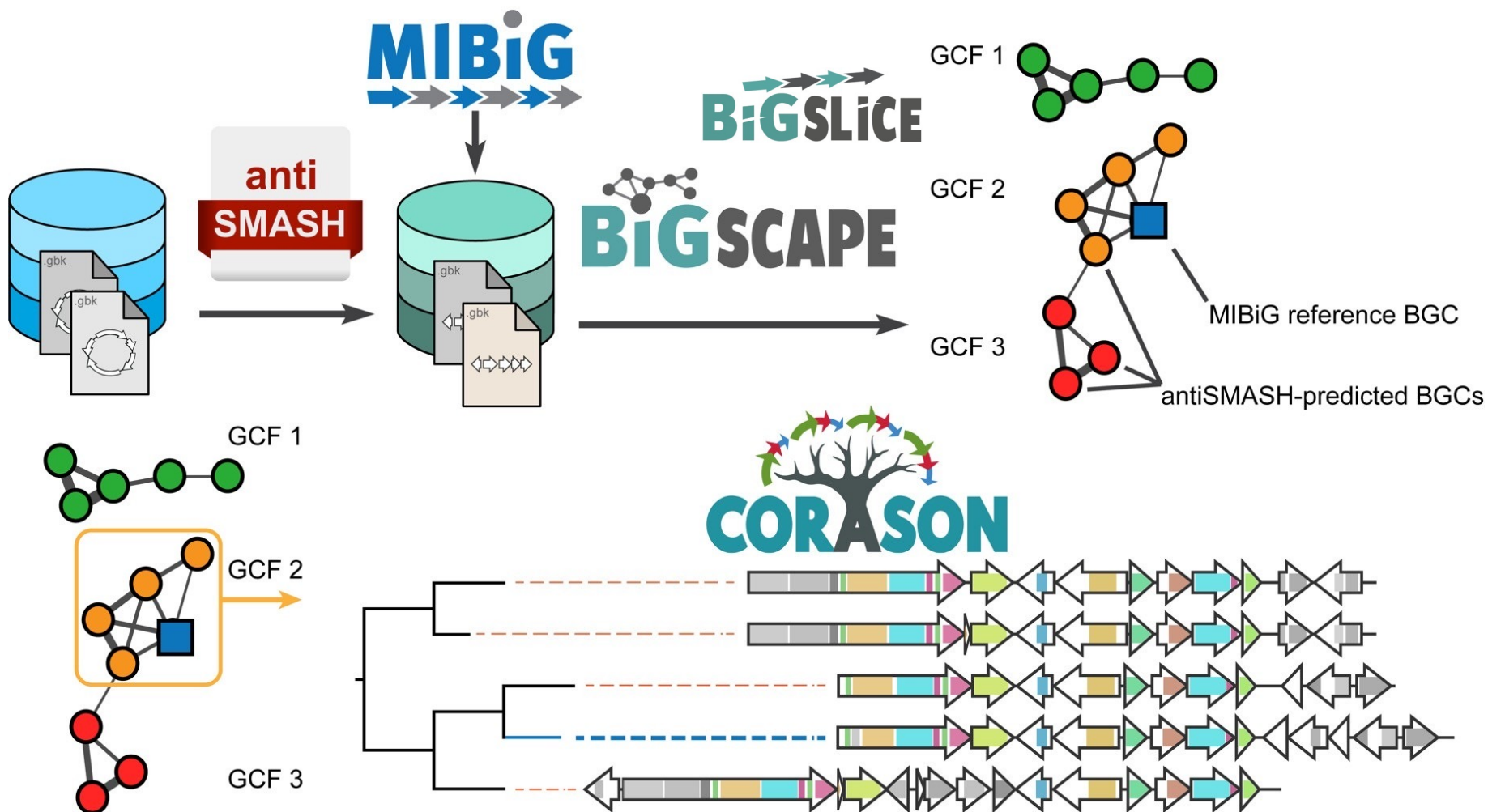
<https://git.wageningenur.nl/medema-group/BiG-SCAPE>

MIBiG REFERENCE DATA ALLOWS RAPID ANNOTATION PROPAGATION AND NETWORK ANALYSIS



<https://mibig.secondarymetabolites.org>

AN EXTENSIVE SOFTWARE ECOSYSTEM FACILITATES COMPUTATIONAL GENOMIC ANALYSIS OF LARGE-SCALE BIOSYNTHETIC DIVERSITY

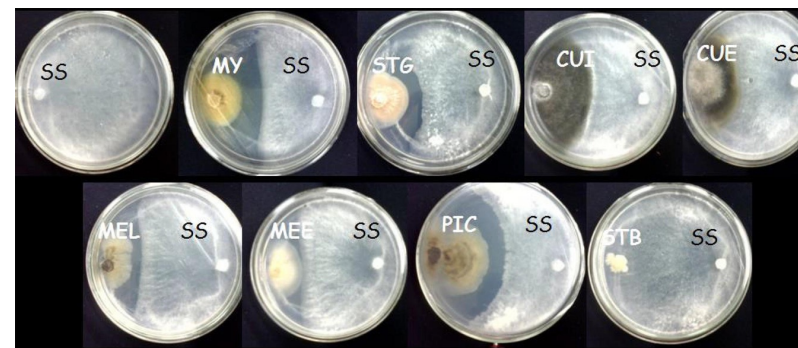
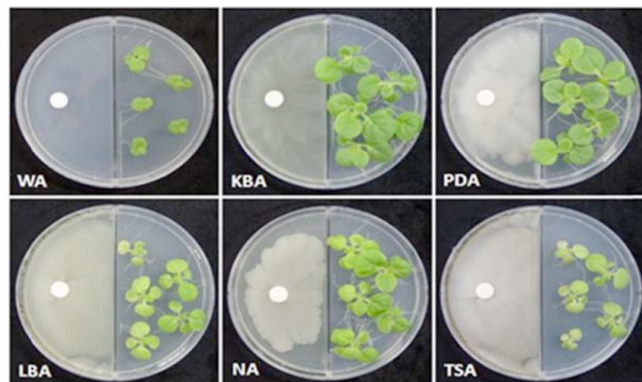


<https://antismash.secondarymetabolites.org>

<https://mibig.secondarymetabolites.org>

<http://bigscape-corason.secondarymetabolites.org>

Navarro-Muñoz, Selem-Mojica, Mullowney et al. (2020) *Nature Chemical Biology* 16(1):60-68.



DNA reads

Quality
control

FastQC

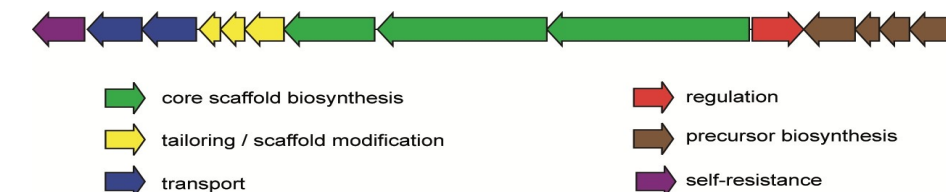
Filtered DNA reads

Genome
assembly

MEGAHIT

(meta)Genome

Annotation



Velvet



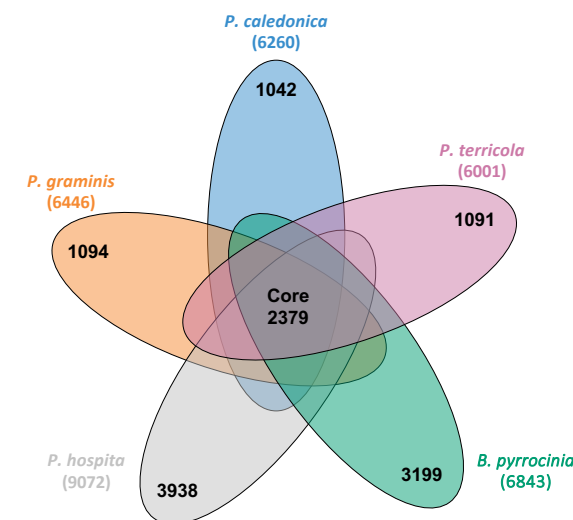
MEGAHIT

Quast

Quality Assessment Tool for Genome Assemblies



Comp. Genomics



WHAT HAVE WE LEARN?

1. Annotation tools (general and specific)
2. Secondary metabolism and more ...