

# Metagenomics 101

## Session 7: Databases for molecular functions

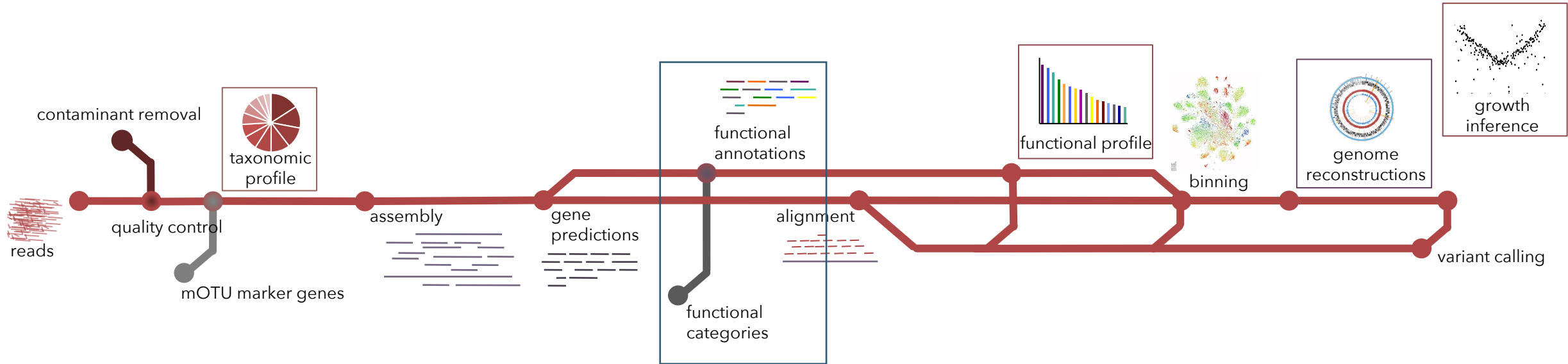
Anna Heintz-Buschart

April 2022



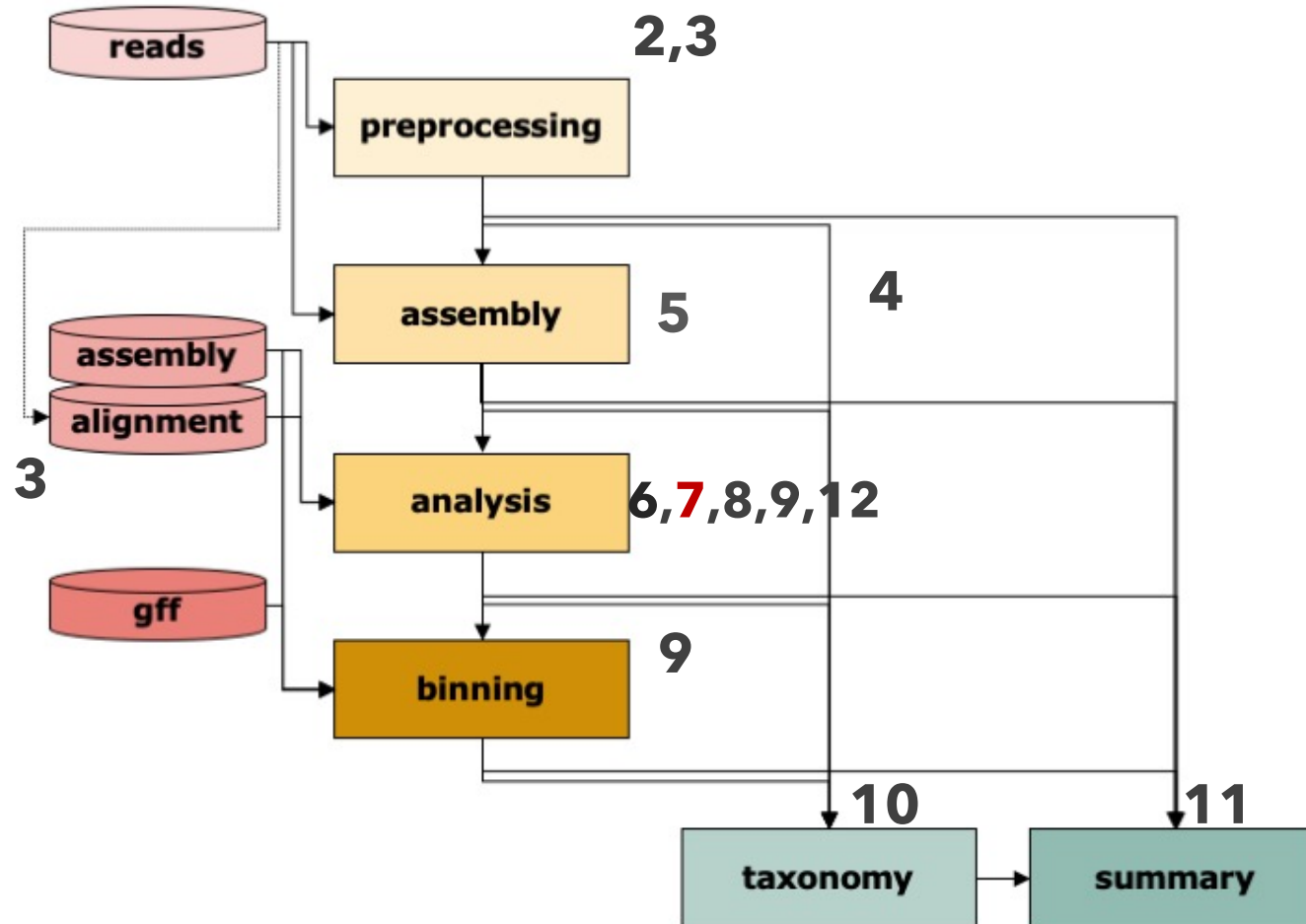


# Metagenomics (+ other omics) pipeline





# Metagenomics (+ other omics) pipeline





# Today

## Curated families/ontologies

- Pfam
- KEGG
- EggNOG

## Large collections

- UniProt
- NCBI

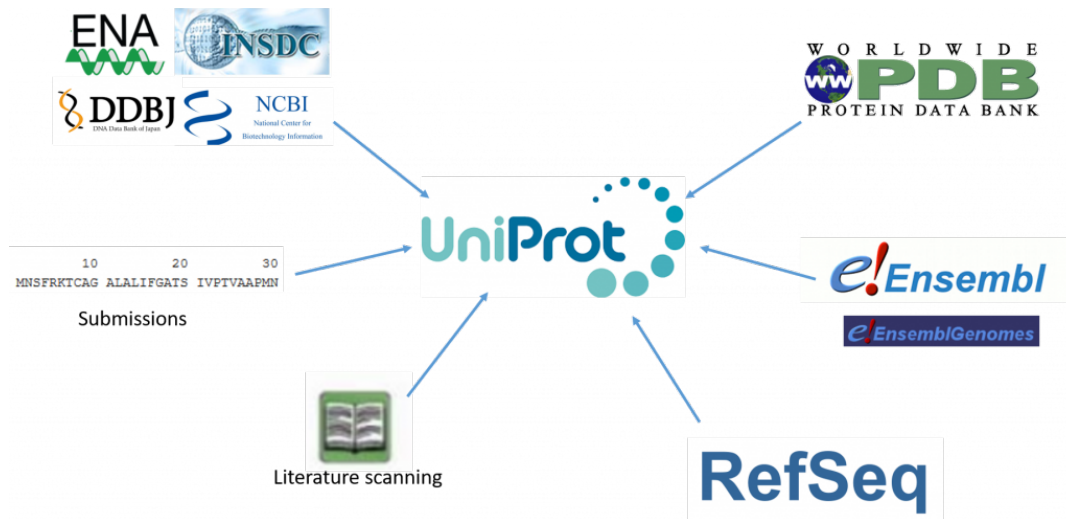
## Specialized databases

- antibiotics resistance: Resfams, CARD, ...
- specific metabolism: antiSMASH, CAZy, ...
- taxonomic/phylogenetic markers: BUSCO, CheckM, mOTUs, ...
- others: virulence, effectors, toxins, plasmids, phages, CRISPR...



# Large collections

- You will most likely not use these directly
- They form the basis of the more ordered and the more specific databases



**UniProtKB**  
UniProt Knowledgebase

Swiss-Prot (566,996)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (230,328,648)

Automatically annotated and not reviewed.

Records that await full manual annotation.

**UniProt** BLAST Align Peptide search SPARQL UniProtKB

**P08244 · PYRF\_ECOLI**

Orotidine 5'-phosphate decarboxylase · *Escherichia coli* (strain K12) · EC:4.1.1.23 · Gene: pyrF · 245 amino acids · Evidence at protein level · Annotation score: 100

Entry Feature viewer Publications External links History

BLAST Align Download Add Add a publication Entry feedback

**Function<sup>1</sup>**

Catalyzes the decarboxylation of orotidine 5'-monophosphate (OMP) to uridine 5'-monophosphate (UMP).

**Catalytic Activity**

H<sup>+</sup> + orotidine 5'-phosphate = CO<sub>2</sub> + UMP  
EC:4.1.1.23 (UniProtKB) | ENZYME C<sup>2</sup> | Rhea C<sup>2</sup> | Source: Rhea 11596 C<sup>2</sup>

**Pathway**

Pyrimidine metabolism; UMP biosynthesis via de novo pathway; UMP from orotate: step 2/2.

**Features**

Showing features for region<sup>1</sup>, active site<sup>1</sup>, binding site<sup>1</sup>.

4 20 40 60 80 100

[T A S S S R A V T N S P V V A L D Y H N R D A L A F V D K I D P R C R K V G K E M F T L E G P O F V R E L Q Q R G F D I F L D L K F H D]

TYPE -- Select -- ID POSITION(S)

Feature	Position(s)
Region	71-80
Active site	73
Binding site	22
Binding site	44
Binding site	131
Binding site	192
Binding site	201
Binding site	221
Binding site	222

**GO Annotations<sup>1</sup>**

Slimming set: agr

Cell color indicative of number of GO terms

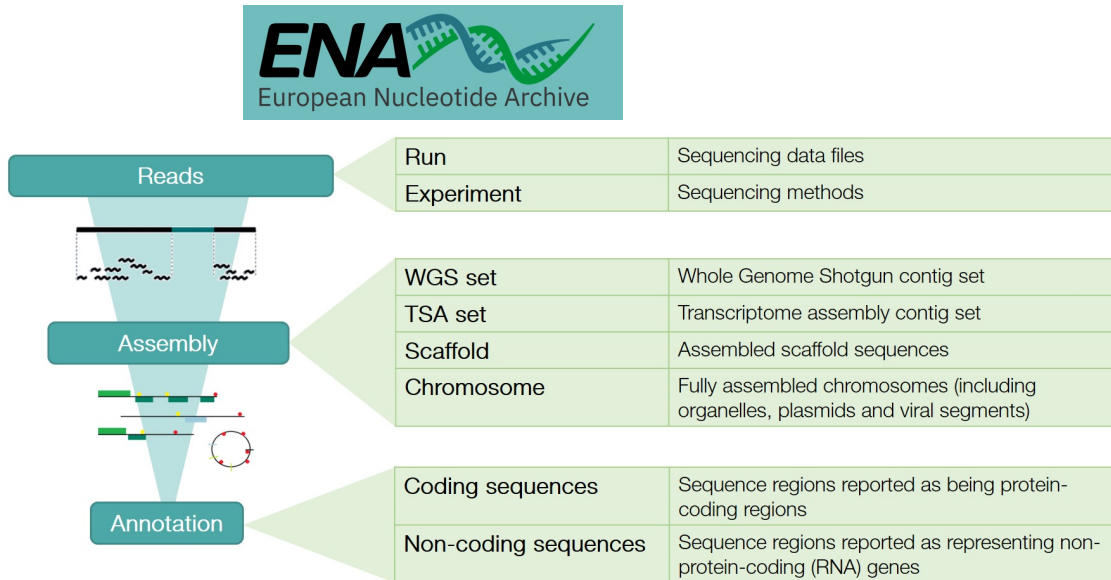
ASPECT	TERM
Cellular Component	cytoplasm
Cellular Component	cytosol
Molecular Function	carboxy-lyase activity
Molecular Function	orotidine-5'-phosphate decarboxylase activity
Biological Process	'de novo' pyrimidine nucleobase biosynthetic process
Biological Process	'de novo' UMP biosynthetic process
Biological Process	nucleobase-containing small molecule interconversion

**UniProt**



# Large collections

- You will most likely not use these directly
- They form the basis of the more ordered and the more specific databases



## Index of /blast/db/FASTA

Name	Last modified	Size
<a href="#">Parent Directory</a>		-
<a href="#">nr.gz</a>	2022-04-16 19:29	124G
<a href="#">nr.gz.md5</a>	2022-04-16 20:16	40
<a href="#">nt.gz</a>	2022-04-18 13:33	172G
<a href="#">nt.gz.md5</a>	2022-04-18 14:58	40
<a href="#">pdbaa.gz</a>	2022-04-16 11:36	33M
<a href="#">pdbaa.gz.md5</a>	2022-04-16 11:36	43
<a href="#">swissprot.gz</a>	2022-04-16 11:36	135M
<a href="#">swissprot.gz.md5</a>	2022-04-16 11:36	47

### Announcements

March 11, 2022

**RefSeq Release 211 is available for FTP**

This release includes:

Proteins: 224,211,842

Transcripts: 43,956,061

Organisms: 117,030

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation: [Release Notes](#)

See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.

**NIH** National Library of Medicine  
National Center for Biotechnology Information

RefSeq


**RefSeq: NCBI Reference Sequence Database**


A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.



# Pfam-A

- families based on sequence homology
- represented by HMMs

EMBL-EBI  <https://pfam.xfam.org>

HOME | [SEARCH](#) | [BROWSE ABOUT](#) | [FTP](#) | [HELP](#) | 

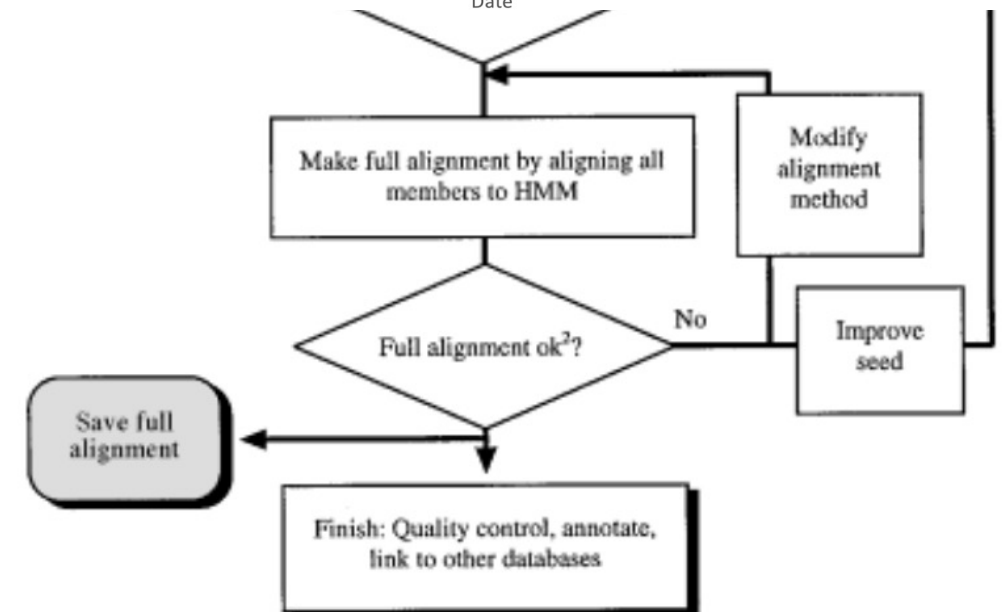
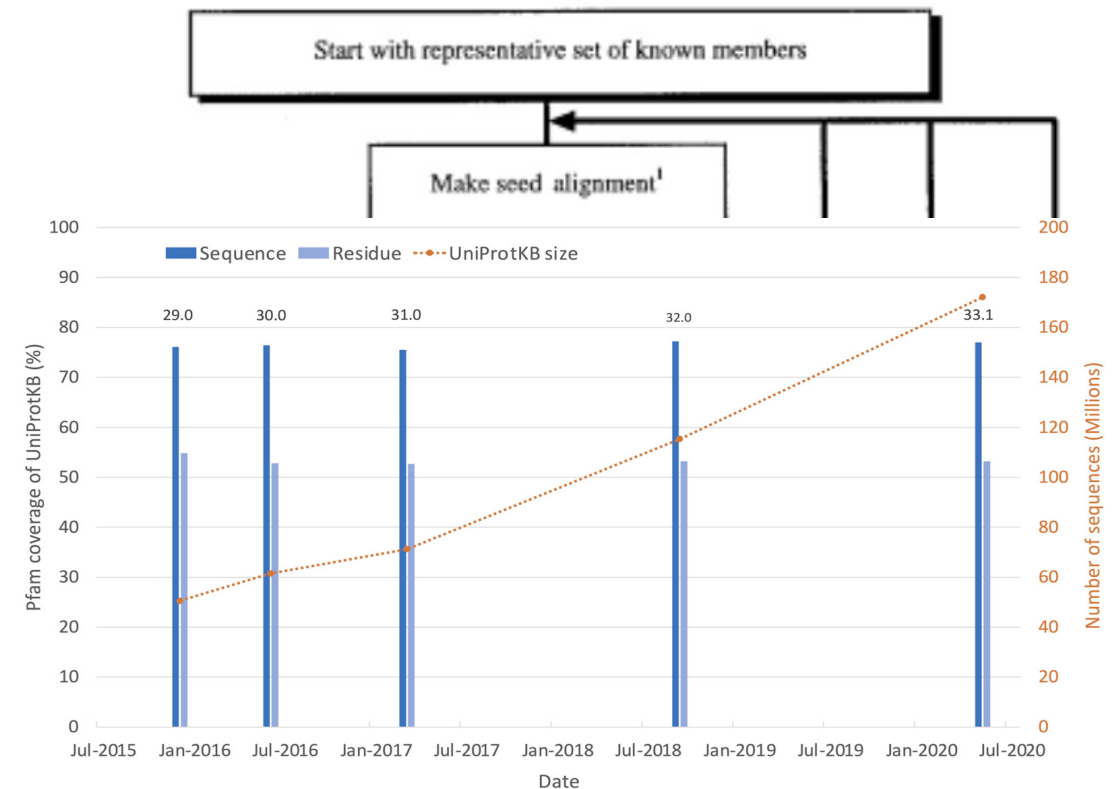
[ftp.ebi.ac.uk/pub/databases/Pfam/current\\_release/](ftp.ebi.ac.uk/pub/databases/Pfam/current_release/)

## Index of /pub/databases/Pfam/current\_release/

.. /	15-Nov-2021 15:48	-
<a href="#">database_files/</a>	15-Nov-2021 16:43	-
<a href="#">proteomes/</a>	15-Feb-2022 12:08	-
<a href="#">structure_models/</a>	15-Nov-2021 12:01	358156
<a href="#">Pfam-A.clans.tsv.gz</a>	15-Nov-2021 12:01	23577
<a href="#">Pfam-A.dead.gz</a>	15-Nov-2021 12:02	4509906931
<a href="#">Pfam-A.fasta.gz</a>	15-Nov-2021 12:06	15188156081
<a href="#">Pfam-A.full.gz</a>	15-Nov-2021 12:13	34204903419
<a href="#">Pfam-A.full.uniprot.gz</a>	15-Nov-2021 12:13	514890
<a href="#">Pfam-A.hmm.dat.gz</a>	15-Nov-2021 12:14	293000230
<a href="#">Pfam-A.hmm.gz</a>		

clan, UniProt sequence, PDB structure, etc.

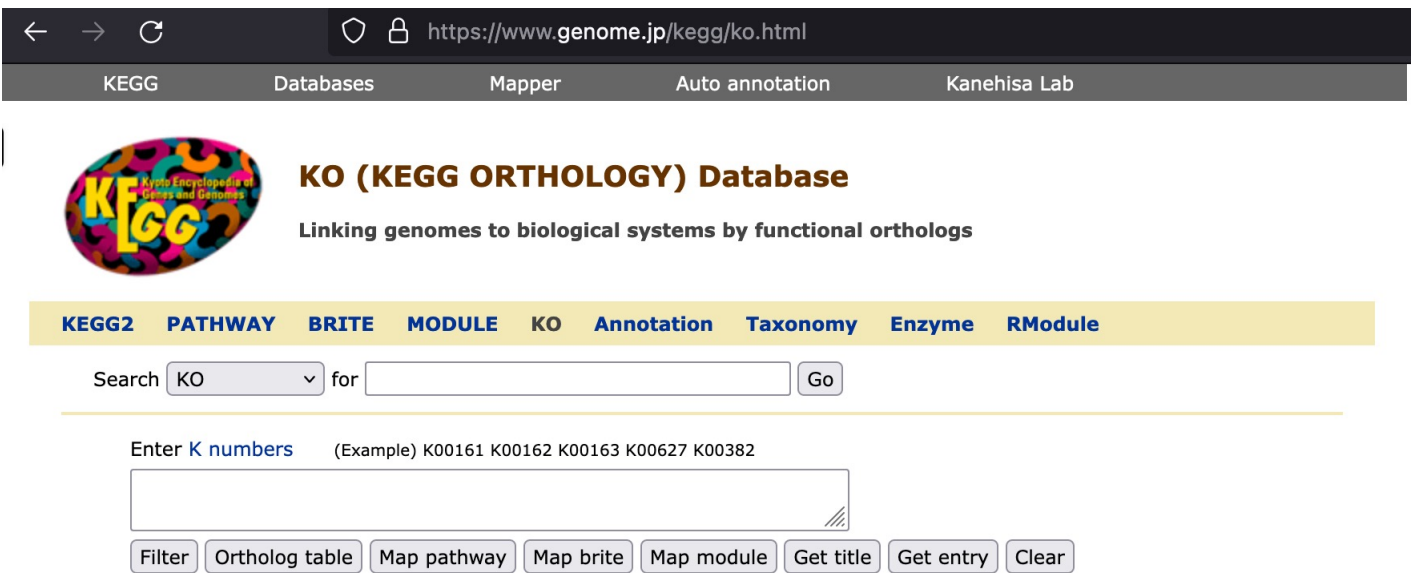
Or view the [help](#) pages for more information





# KEGG

- KOs
- pathways
- links to reactions, metabolites, genomes



The screenshot shows the KEGG KO (KEGG ORTHOLOGY) Database homepage. At the top, there is a browser address bar with the URL <https://www.genome.jp/kegg/ko.html>. Below the address bar is a navigation bar with links: KEGG, Databases, Mapper, Auto annotation, and Kanehisa Lab. The main heading is "KO (KEGG ORTHOLOGY) Database" with the tagline "Linking genomes to biological systems by functional orthologs". A secondary navigation bar contains links: KEGG2, PATHWAY, BRITE, MODULE, KO, Annotation, Taxonomy, Enzyme, and RModule. The search section includes a "Search" label, a dropdown menu set to "KO", a text input field, and a "Go" button. Below the search field, there is a prompt "Enter K numbers" with an example "(Example) K00161 K00162 K00163 K00627 K00382". A large text input field is provided for entering K numbers. At the bottom, there is a row of buttons: Filter, Ortholog table, Map pathway, Map brite, Map module, Get title, Get entry, and Clear.



[ [Pathway menu](#) | [Pathway entry](#) | [Image file](#) | [Help](#) ]

Change pathway type

Scale:  60%

## ▼ Search

Go

## ▼ ID search

Go

▼ **Color**

▼ **Module**

## □ Pathway modules

☐ Amino acid metabolism

☐ Aromatic amino acid metabo

☐ M00042 Catecholamine bi

☐ M00043 Thyroid hormone

☐ M00044 Tyrosine degrada

☐ M00533 Homoprotocatech

## ▼ Network

**nt06422 Dopamine metabolism**

☐ N01036 L-DOPA generation

☐ N01038 DOPAL generation

nt06322 TRH-TSH-TH signaling





# KEGG

- KOs
- pathways
- links to reactions, metabolites, genomes



## KofamKOALA - KEGG Orthology Search

K number assignment based on KO-dependent scoring criteria

BlastKOALA

GhostKOALA

KofamKOALA

KOALA job status 2022/04/20 13:43:47 (GMT+9)

	Blast	Ghost	Kofam
Number of jobs in the queue	6	1	0
Submission of last completed job	2022/04/20 10:45:49	2022/04/20 12:28:55	2022/04/20 13:17:39

KofamKOALA assigns K numbers to the user's sequence data by HMMER/HMMSEARCH against Kofam (a customized HMM database of KEGG Orthologs (KOs)). K number assignments with scores above the predefined thresholds for individual KOs are more reliable than other proposed assignments. Such high score assignments are highlighted with asterisks '\*' in the output. The K number assignments facilitate the interpretation of the annotation results by linking the user's sequence data to the KEGG pathways and EC numbers.

Enter FASTA Sequences

or upload a sequence file

No file selected.

## Index of /ftp/db/kofam

<a href="#">Name</a>	<a href="#">Last modified</a>	<a href="#">Size</a>	<a href="#">Description</a>
----------------------	-------------------------------	----------------------	-----------------------------

<a href="#">Parent Directory</a>		-	
<a href="#">archives/</a>	30-Mar-2022 13:42	-	
<a href="#">ko_list.gz</a>	28-Mar-2022 21:12	781K	
<a href="#">profiles.tar.gz</a>	30-Mar-2022 13:48	1.3G	



## KO (KEGG ORTHOLOGY) Database

Linking genomes to biological systems by functional orthologs

[KEGG2](#) [PATHWAY](#) [BRITE](#) [MODULE](#) [KO](#) [Annotation](#) [Taxonomy](#) [Enzyme](#) [RModule](#)

Search  for

Enter K numbers (Example) K00161 K00162 K00163 K00627 K00382



# EggNOG

- orthologous groups
- bacteria, archaea, viruses
- links to and parsing of phylogeny

eggnog5.embl.de/#/app/home

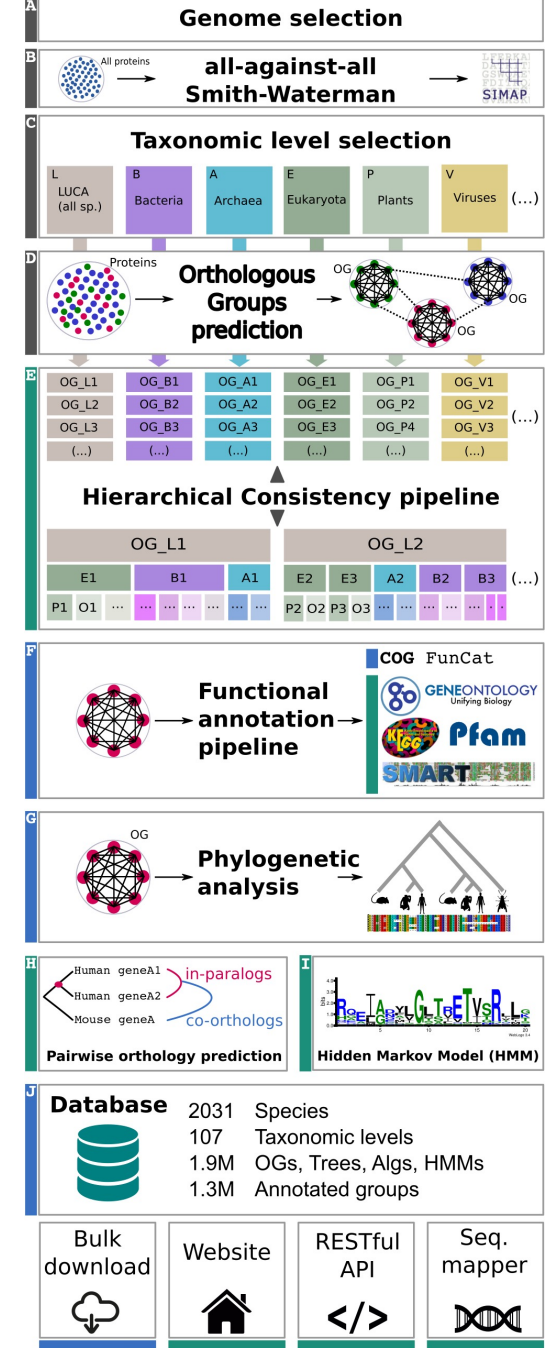
Search protein or OG

## EggNOG v5.0

A database of orthology relationships, functional annotation, and gene evolutionary histories.

Organisms	Viruses	Orthologous Groups	Tree & Algs
5,090	2,502	4.4M	4.4M

Search





# Specific databases: antibiotics resistance

- Resfams
- CARD

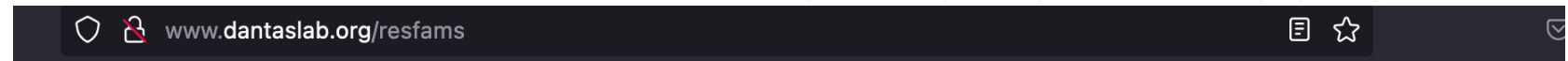
The ISME Journal (2014), 1–10  
© 2014 International Society for Microbial Ecology All rights reserved 1751-7362/14  
www.nature.com/ismej



## ORIGINAL ARTICLE

### Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology

Molly K Gibson<sup>1</sup>, Kevin J Forsberg<sup>1</sup> and Gautam Dantas<sup>1,2,3</sup>



#### DOWNLOAD RESFAMS

- [Resfams HMM Database \(Core\)](#) - v1.2, updated 2015-01-27  
Database version for annotation of microbial proteins in the absence of any functional confirmation for antibiotic resistance.
- [Resfams HMM Database \(Full\)](#) - v1.2, updated 2015-01-27  
Database version for annotation of microbial proteins when functional confirmation for antibiotic resistance is available (such as functional metagenomic selections).

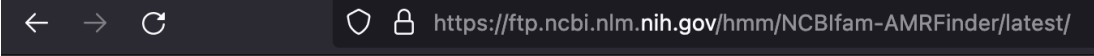
#### SUPPORTING DATAFILES

- [Resfams profile HMM Metadata](#) - v1.2.2, updated 2018-02-21
- DEPRECATED: [Resfams profile HMM Metadata](#) - v1.2.1, updated 2017-03-25  
Metadata on profile HMMs in resfams including description, ARO identifiers, HMM database source, and mechanism classification. Updated to reflect latest CARD (v1.1.5)
- [Resfams AR Proteins](#) - v1.2, updated 2015-01-27  
Proteins used to build Resfams base HMM database



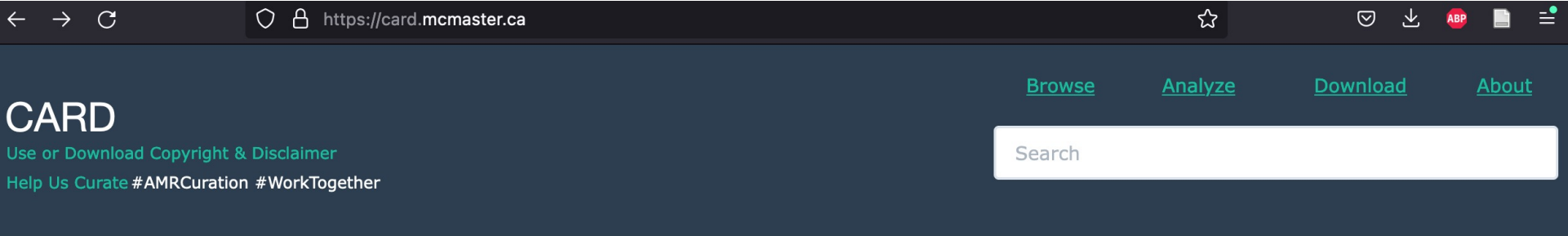
# Specific databases: antibiotics resistance

- Resfams
- CARD



Index of /hmm/NCBIfam-AMRFinder/latest

Name	Last modified	Size
<a href="#">Parent Directory</a>		-
<a href="#">NCBIfam-AMRFinder.HMM.tar.gz</a>	2022-04-07 11:30	6.4M
<a href="#">NCBIfam-AMRFinder.LIB</a>	2022-04-07 11:30	87M
<a href="#">NCBIfam-AMRFinder.SEED.tar.gz</a>	2022-04-07 11:30	685K
<a href="#">NCBIfam-AMRFinder.changelog.txt</a>	2022-04-07 11:30	66K
<a href="#">NCBIfam-AMRFinder.tsv</a>	2022-04-07 11:30	111K



CARD

[Use or Download Copyright & Disclaimer](#)  
[Help Us Curate](#) #AMRCuration #WorkTogether

[Browse](#) [Analyze](#) [Download](#) [About](#)

Search

## The Comprehensive Antibiotic Resistance Database

A bioinformatic database of resistance genes, their products and associated phenotypes.

6501 Ontology Terms, 4970 Reference Sequences, 1922 SNPs, 2913 Publications, 5016 AMR Detection Models

Resistome predictions: 263 pathogens, 16719 chromosomes, 2675 genomic islands, 33860 plasmids, 136704 WGS assemblies, 285146 alleles

[CARD Bait Capture Platform 1.0.0](#) | [State of the CARD 2021 Presentations & Demonstrations](#)



# Specific databases: antibiotics resistance

Arango-Argoty *et al. Microbiome* (2018) 6:23  
DOI 10.1186/s40168-018-0401-z

Microbiome

SOFTWARE

Open Access

DeepARG: a deep learning approach for  
predicting antibiotic resistance genes from  
metagenomic data



Gustavo Arango-Argoty<sup>1</sup>, Emily Garner<sup>2</sup>, Amy Pruden<sup>2</sup>, Lenwood S. Heath<sup>1</sup>, Peter Vikesland<sup>2</sup> and Liqing Zhang<sup>1\*</sup>



# Specific databases: specific metabolism

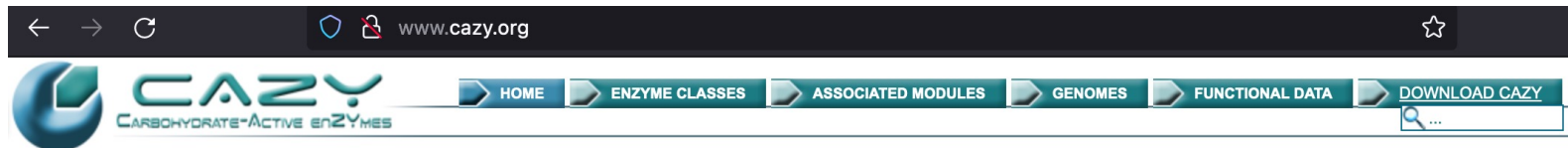
- antiSMASH
- CAZy

Published online 12 May 2021

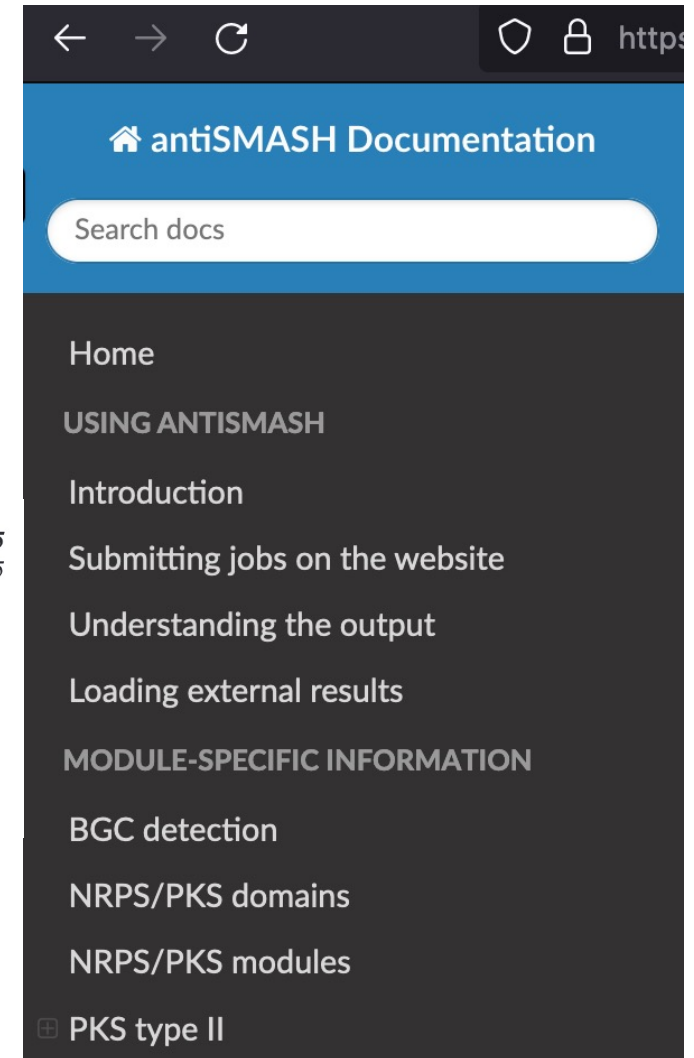
*Nucleic Acids Research*, 2021, Vol. 49, Web Server issue W29–W35  
<https://doi.org/10.1093/nar/gkab335>

## antiSMASH 6.0: improving cluster detection and comparison capabilities

Kai Blin<sup>1,\*</sup>, Simon Shaw<sup>1</sup>, Alexander M. Kloosterman<sup>2</sup>, Zach Charlop-Powers<sup>3</sup>, Gilles P. van Wezel<sup>2,4</sup>, Marnix H. Medema<sup>2,5,\*</sup> and Tilmann Weber<sup>1,\*</sup>



Welcome to the Carbohydrate-Active enZymes Database





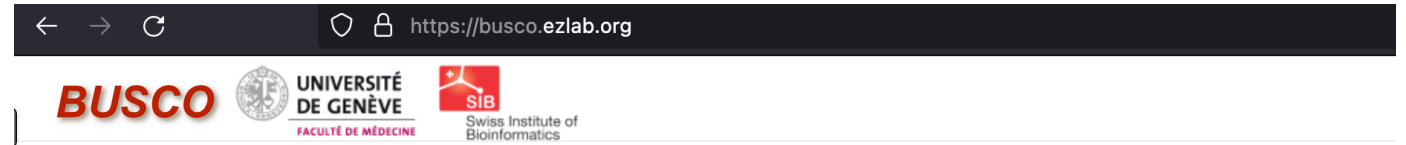
# Specific databases: marker genes

- BUSCO
- CheckM
- fetchMG-markers

## Method

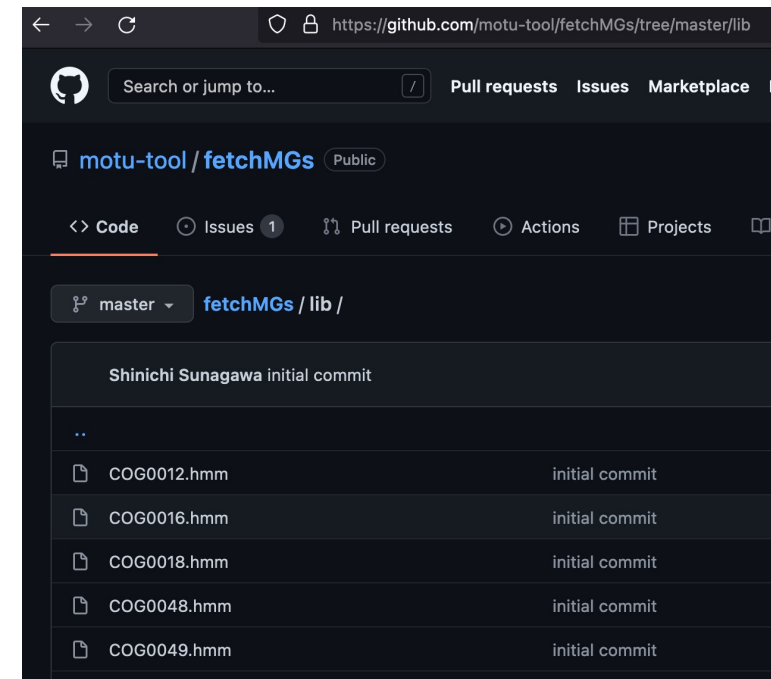
### CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes

Donovan H. Parks,<sup>1</sup> Michael Imelfort,<sup>1</sup> Connor T. Skennerton,<sup>1</sup> Philip Hugenholtz,<sup>1,2</sup> and Gene W. Tyson<sup>1,3</sup>



# BUSCO

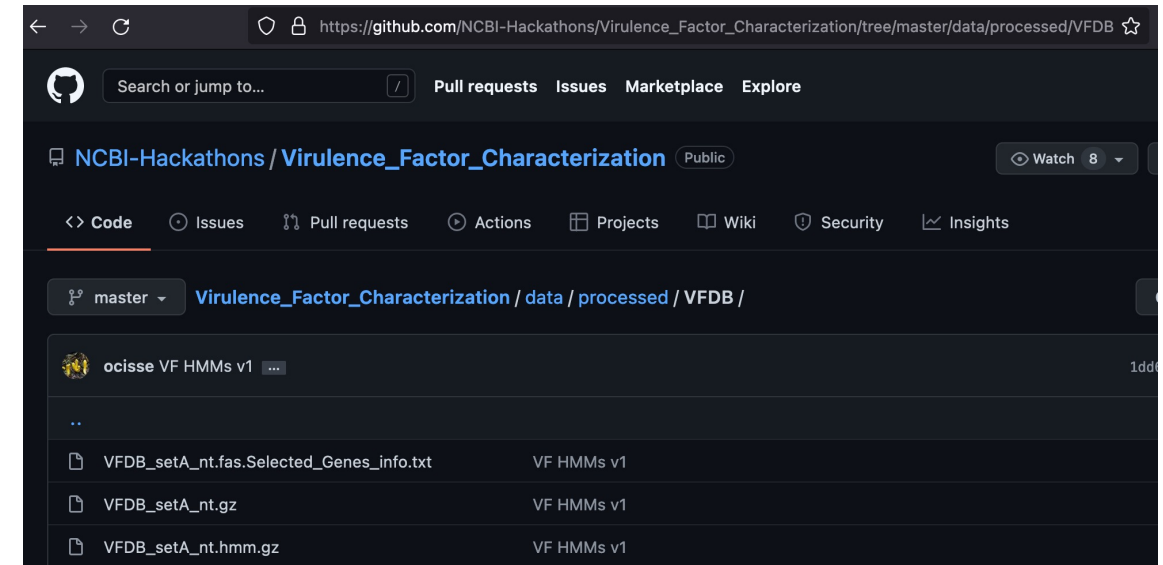
from QC to gene prediction and phylogenomics





# Specific databases: others

- virulence
- effectors



Nies et al. *Microbiome* (2021) 9:49  
<https://doi.org/10.1186/s40168-020-00993-9>

Microbiome

<https://cran.r-project.org/web/packages/effectR/vignettes/effectR.html>

## effectR: An R package to call oomycete effectors

Javier F. Tabima


2018-09-30

The `effectR` package is an R package designed to call oomycete RxLR and CRN effectors by searching for the motifs of interest using regular expression searches and hidden markov models (HMM).

## SOFTWARE ARTICLE

Open Access

## PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data

Laura de Nies<sup>1</sup>, Sara Lopes<sup>1</sup>, Susheel Bhanu Busi<sup>1</sup>, Valentina Galata<sup>1</sup>, Anna Heintz-Buschart<sup>1,2,3</sup>, Cedric Christian Laczny<sup>1</sup>, Patrick May<sup>4</sup> and Paul Wilmes<sup>1\*</sup> 





# Specific databases: others

- phages
- plasmids



Database, 2019, 1–8  
doi: 10.1093/database/baz093  
Database tool



Database tool

**CasPDB: an integrated and annotated database  
for Cas proteins from bacteria and archaea**

Zhongjie Tang<sup>1,†</sup>, ShaoQi Chen<sup>1,†</sup>, Ang Chen<sup>1</sup>, Bifang He<sup>1,2</sup>,  
Yuwei Zhou<sup>1</sup>, Guoshi Chai<sup>1</sup>, FengBiao Guo<sup>1,\*</sup> and Jian Huang<sup>1,\*</sup>

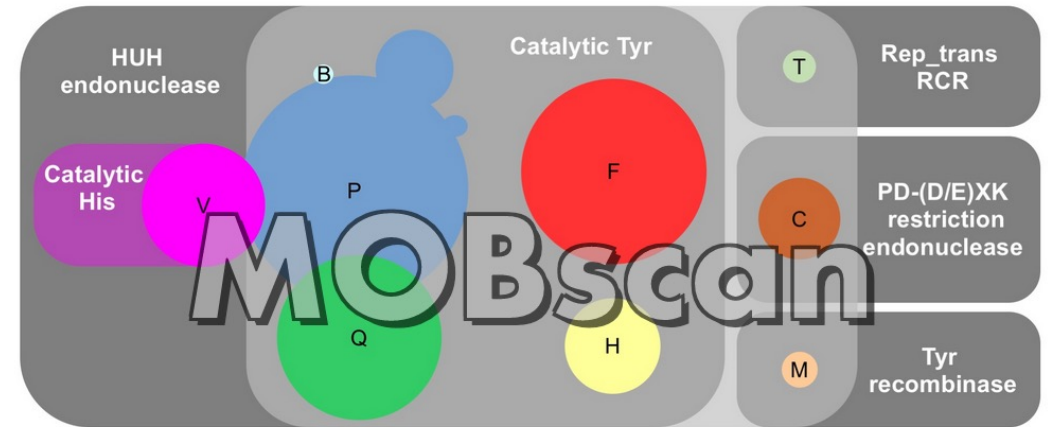


**viruses**

Article

## Classifying the Unclassified: A Phage Classification Method

Cynthia Maria Chibani, Anton Farr, Sandra Klama, Sascha Dietrich and Heiko Liesegang \*<sup>id</sup>



MOBscan is a web application for identifying relaxase MOB families. It uses the hmmscan function of the HMMER3 software suite ([Eddy, 2011](#)) to search against [MOBfamDB](#), a curated relaxase profile HMM database. If you find it useful for your work, please cite it as:

Garcillán-Barcia M.P., Redondo-Salvo S., Vielva L., de la Cruz F. (2020) "MOBscan: Automated Annotation of MOB Relaxases". In: de la Cruz F. (eds) *Horizontal Gene Transfer. Methods in Molecular Biology*, vol 2075. Humana, New York, NY



# And what about ...?

- genome databases and gene catalogues:  
we will look at these on the 8th June





# Thanks for your attention!



[a.u.s.heintzbuschart@uva.nl](mailto:a.u.s.heintzbuschart@uva.nl)

SP C2.205



[github.com/a-h-b](https://github.com/a-h-b)



[twitter.com/\\_a\\_h\\_b\\_](https://twitter.com/_a_h_b_)