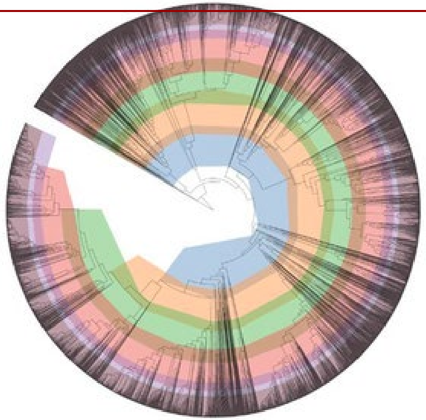
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

MAG functional annotation

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Samuel Aroney, Queensland University of Technology



NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)
As. Prof. Luis Pedro Coelho

You got MAGs!
You know that they are good
Now what?

```
ACTTCTCGGGCCGACGCCTGTTTCGATCGATGATTTGGTGCGTCTTGCGCAGACCACGGCGGCGATCGTGCGCGCTGCCCTGCTGGAACTCGAAATTGCCGGGCGGATTGAACG
CGATACCCCCAATCGTCATGTTCATCCGATTTCACCTGGCTGTCGCGGCTATACTGATCGTGCTGGTGGCGTCATCGCGGGGCCAGGAACAACCCGTTTCGGCTTTTCTCGCAG
CGCACGACGTTCGATACCGCCATGAAGGGATTCACCCCCGATCAACGGGTGATCACCGCGACCCGGCGTCAGCCGGAATACGGCAAACCCGTCGGCGACTACGTGAACGCCATG
AGGCCGCGAATGGGCAAAAACATTCGATGTCGTCGAAAAGAAGTTTCAGGTCGAACGCTGGGTCTTGCTCGCCTTGTGGGGCATGGAATCGGACTTCGGGTCAGAAAAAGATCG
ATGTGAAATTCCGCGACCCCTATTTTCGCAACGAGCTGATCGTGGCGATGCGCATCATGCAAGACAATCGCATCGCCCGCGAAAAGATGGTCAGCTCTTGGGCCGGCGCGATGG
TATGCGATCGATTTTTCCGGCGACGGACGGGCGGATATCTGGGGCAACGTACCGGATGTGCTTGCGTCGACCGCCAACTACCTGCGCAAATGGAAATGGAATCCGGCGCTTCCC
CTACATGCGCAGCCGCGCAAGTTTTGCGGAATGGCAAGCGCTCGGCGTGCGCCGCGCCGACGGCAAGGCGTTCCCCAACTCCGGACAGGGCATCCTCTTCTTTCCCAGCGGCGG
ATGTGCTCAAGGAATACAACAACTCCGATGCTTACGCGCTCGCGGTCGGGCACCTCGCCGACCGAATCCACGGCGGCGATCTGATCAAGACGCCCTGGCCTAAGGACGATCGCC
AGACTCGCGGCACTCGGCTACAAAGTGAACGAGTTTGAGGCCCACATCGATTTCGATTTGCGCGACAACATTCGCGTCGAGCAAAAGAAGCTGGGGATGGTCCCCGACGGCAAT
GCCTCGGCTCTAGGTCTCCAGCAGTACGCGGACCGCCTTGAAGCTGGCAATCCAGCCTTGATACAATTTACGGTTGTCGTCTCTTTCGTGCCCTATTTCGCTGGTCTGACACGG
TTCCCCATGTTCCGGGCCAAACATGCGGCCGATTCCGCCTTGATTCTCAACAGGCCTGGCGGCTAACCCATTGGATTTTCATGAATGTTGTCGTAGTCGAGTCGCCGTCCAAGG
GAGGTTCTGGCCTCATTTGGCCATATCCGGGACCTGCCCCCCAAGGATGGCTCCGTCGATCCCGACAATGATTTCCGCATGCTCTGGGAGGTCGACGCCAGGTCGAACCAGCGG
CAAGCTGATCCTCGCCACCGACCCGGATCGCGAGGGCGAAGCAATTTCCTGGCACGTGCTCGAGGTGCTGAAGGAAAAGAAGGCGCTCAAGGACCACAAGATCGAGCGCGTCGT
CGATGAAGCATCCACGGATGATCGACGCCGCATTGGTCGATGCCTACCTCGCGCGCCGCGCGCTCGACTATCTCGTCGGCTTCACCCTTTCACCGGTGCTGTGGCGAAAGCTGG
GCGCTTCGGCTTGTGTGCGATCGCGAACTCGAAATCGAGAAGTTTGTTGCGAAGGAGTATTGGTCGATTCTCGCCAGGCTCGCGACGCCGCGCAACGAAGTGTTCGAAGCGCGT
CGACATAGGTTCGGGCGCCGAAGCGGAAGCTTTCACCCGAGACCTCGAGAATGCGACCTTCAAGGTGACGTCGGTCGAGGCAAAGCCCGCACGGCGCAATCCGCCGCCGCCCTT
AGCTCGGCTTTGCACCGGCGGTCGCCATGCGTCTCGCCCAGCGGCTCTACGAAGGCGTGGAAATCGACGGCGAAGCGACCGGCTTGATTACGTATATGCGTACTGACGGCATCG
ATGTTGGGCCGCAACTACGGCAAGGAGTACGTCCCTGCGTCGCCGCGCGAGTACCACAACAAATCCAAGAACGCGCAGGAGGCGCACGAAGCGGTGCGCCCGACCAGCGCGAAC
CGACCAGGCGCGGCTCTACGAGTTGATCTGGAACCGCGCGGTCGCGAGCCAAATGGAATCCGCCGAGCTCGAGCGCACCACGGTCGACATTGTTGCGAAGGCGGGCTCACGCAA
TCGACGGCTTTCTGACGCTCTATCAGGAAGGCCAGGACGAGACGCCGGACGATGACGAGTCGCGGCGTCTGCCCGCGATGTCGGAAGGCGAAACGCTCAGCAAGCAGGCGATCG
TTCTCGGAAGCGGCGCTGGTCAAGCGGATGGAAGAGCTCGGCATCGGCCGGCCCTCCACTTATGCCTCCGTCCTCCAGGTGCTGCAGGATCGCGGCTATGTGCGGATCGACAAG
CGTCGCCTTCCTTGAAAGCTTCTTCGCGCGCTACGTCGAATACGACTTCACAGCCAGCCTCGAGGAAAAACTCGACGAGATTTCGGCGGGCAATATCGACTGGCGGGCCGTGCT
ACGACATCAAGGAAGTACGCAATCGCGTCGTGCTCGATGCGCTCAACGACCTACTTGAGCCGCATATTTTTCCCGAGCGCACTGACGGCAAACCGCGCCGCCAATGCCCGCAGT
TTCGGGGCTTTCGTCGGTTGCTCGAATTATCCGGAATGCAATTTCACGCGGCAAATTACTCCGAGCGCCGATGGCATTCAGGGCAAAAGGGTCCTGGGTGAAGATCCGGCCACA
GCCCTATCTGCAGCTTGGCGAGCAGATCAGGCCGCCCAAGCCGAAGAAAGGCGAGAAAAAACCGGAAGTCGAAAAGCCCAAGCGCGCCGGAATTCCAAAAGGTGTGTCGCCCGA
CGCTGCCGCGTGAAATTGGATTGTCGCCGGAAGACGGCGAGCCGATAGTCGCCGGCATCGGACGCTTCGGCTCATACGTGAAGCACGGCAAGATCTACGCCAACCTGGAAGAAG
GTCACATTGATCGCCGAGAAGAAGGCGAACCCGAAGAAGGGCCGGCGGTTCGGCGCCGATCCAGGCAAGGTCTTGGGCGAACATCCCGATCAGGGCGGCCCGGTTGTCGTCAAG
CATCAACGCGACGATAACCGGCGACAGAACGCCCGAGACCATCACGCTGACTGAAGCGGTCGTGCTGCTTGACGCGCGCGCCGATCAACTGAGTTCTCAGCCGCGGCGCGCAGG
CGAAGGGGCGCAAGAAAAAGACGGCGGAACCGGCAGACAAGGCCGAAAAATCAGCCAAGCCGCGCAAGAAGCCGGCAAAAAAGGCCGCAAAATTCACCGCCGCCGAATAGCTGG
AGCGCCGTTTTGGCGCAGGACTGTCTATTTTGAATAAGAAATCCAAACGAACACCCTCCCTCCCTTCCAAAGCCGACATACTCGCCTTCATCGGTAATCAGCCCGGAAAAGTCG
```

Functional annotation

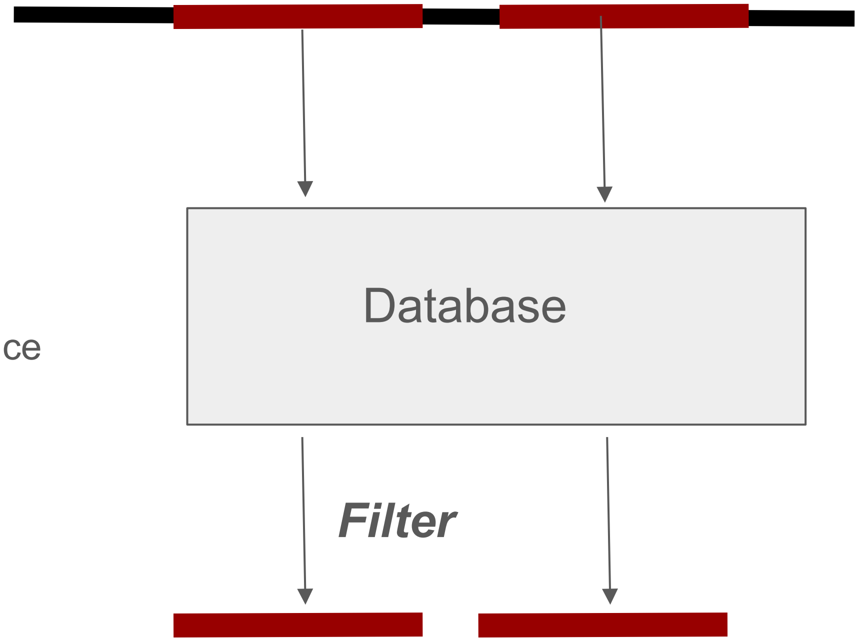
- Genes
- *eggNOG*
- *RGI*
- *Other gene annotations databases*
- *…*

Congrats!
You got big FASTA files!

Physalia
Courses

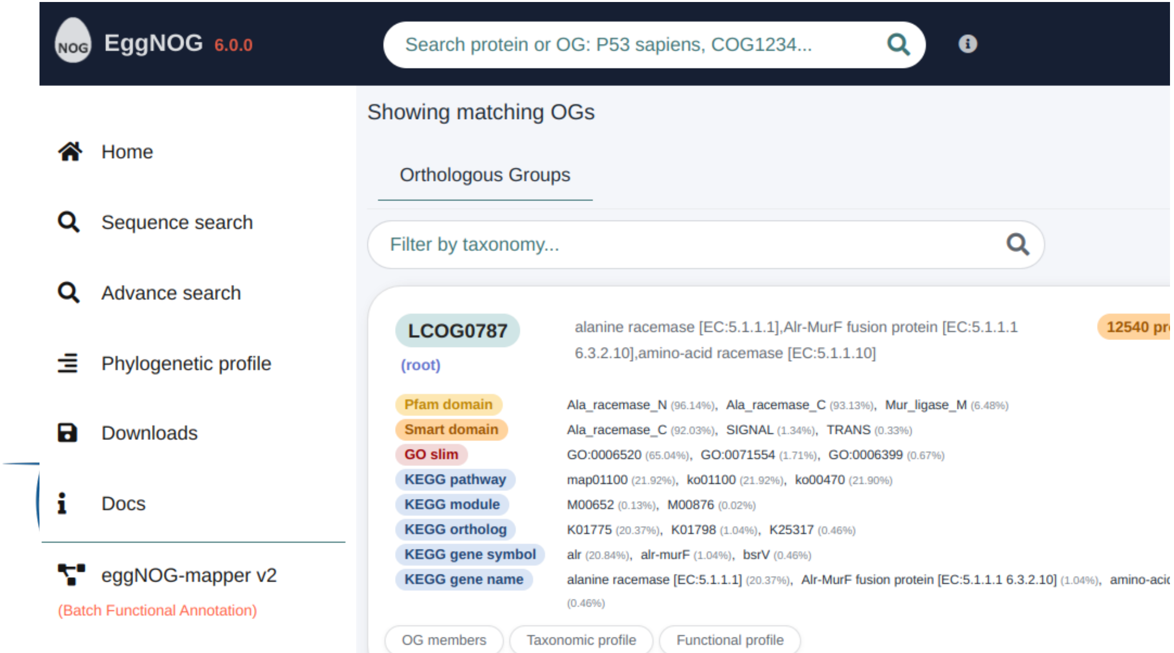
Functional annotation

1. Predict genes
   a. Prodigal
   b. Pyrodigal
2. Predict function of genes
   a. Eggnog-mapper (generic)
   b. RGI (specialized to antibiotic resistance genes)
   c. …
3. Use informed human judgement!!



Gene annotation with eggnog-mapper

Basic concept: *eggNOG orthologous group*



EggNOG 6.0.0

Search protein or OG: P53 sapiens, COG1234...

Home

Sequence search

Advance search

Phylogenetic profile

Downloads

Docs

eggNOG-mapper v2

(Batch Functional Annotation)

Showing matching OGs

Orthologous Groups

Filter by taxonomy...

LCOG0787 (root)

alanine racemase [EC:5.1.1.1],Alr-MurF fusion protein [EC:5.1.1.1 6.3.2.10],amino-acid racemase [EC:5.1.1.10]

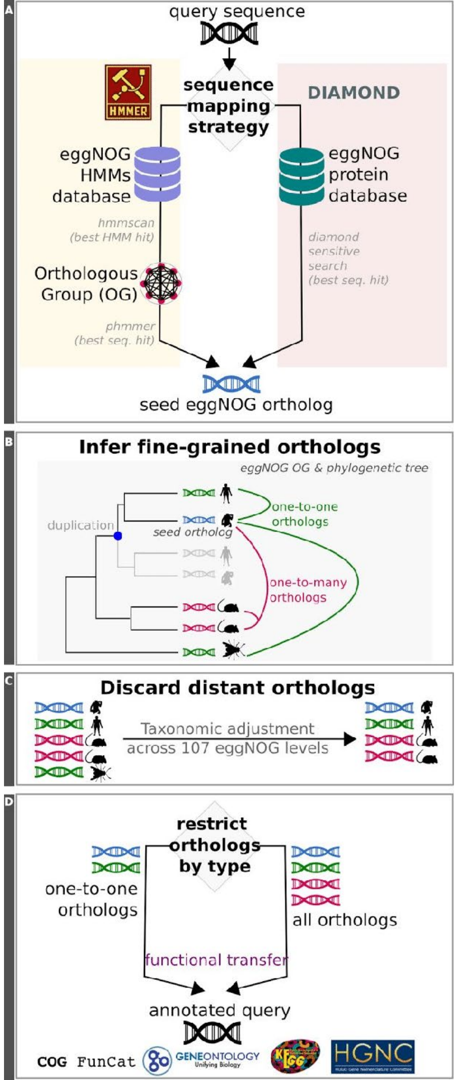
12540 pr

Pfam domain	Ala_racemase_N (96.14%), Ala_racemase_C (93.13%), Mur_ligase_M (6.48%)
Smart domain	Ala_racemase_C (92.03%), SIGNAL (1.34%), TRANS (0.33%)
GO slim	GO:0006520 (65.04%), GO:0071554 (1.71%), GO:0006399 (0.67%)
KEGG pathway	map01100 (21.92%), ko01100 (21.92%), ko00470 (21.90%)
KEGG module	M00652 (0.13%), M00876 (0.02%)
KEGG ortholog	K01775 (20.37%), K01798 (1.04%), K25317 (0.46%)
KEGG gene symbol	alr (20.84%), alr-murF (1.04%), bsrV (0.46%)
KEGG gene name	alanine racemase [EC:5.1.1.1] (20.37%), Alr-MurF fusion protein [EC:5.1.1.1 6.3.2.10] (1.04%), amino-acic (0.46%)

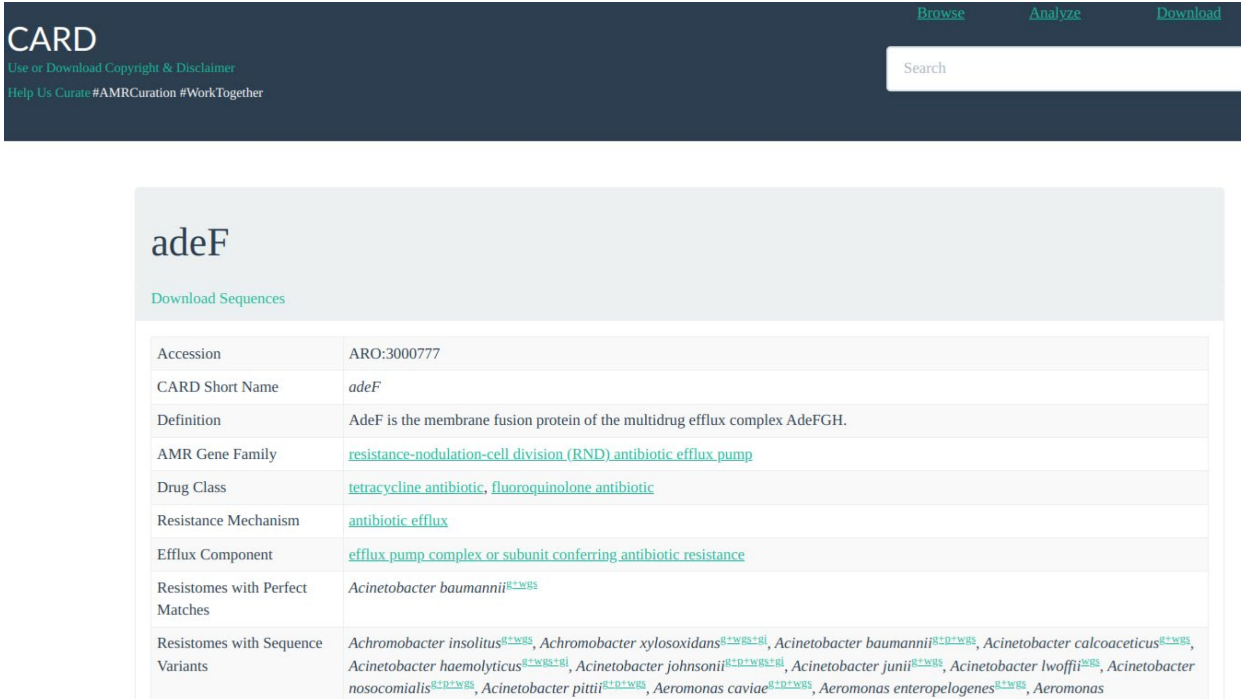
OG members

Taxonomic profile

Functional profile



Gene annotation with RGI (CARD)



CARD

Use or Download Copyright & Disclaimer

Help Us Curate #AMRCuration #WorkTogether

adeF

Download Sequences

Accession	ARO:3000777
CARD Short Name	*adeF*
Definition	AdeF is the membrane fusion protein of the multidrug efflux complex AdeFGH.
AMR Gene Family	resistance-nodulation-cell division (RND) antibiotic efflux pump
Drug Class	tetracycline antibiotic, fluoroquinolone antibiotic
Resistance Mechanism	antibiotic efflux
Efflux Component	efflux pump complex or subunit conferring antibiotic resistance
Resistomes with Perfect Matches	*Acinetobacter baumannii*$^{g+wgs}$
Resistomes with Sequence Variants	*Achromobacter insolitus*$^{g+wgs}$, *Achromobacter xylosoxidans*$^{g+wgs+gi}$, *Acinetobacter baumannii*$^{g+p+wgs}$, *Acinetobacter calcoaceticus*$^{g+wgs}$, *Acinetobacter haemolyticus*$^{g+wgs+gi}$, *Acinetobacter johnsonii*$^{g+p+wgs+gi}$, *Acinetobacter junii*$^{g+wgs}$, *Acinetobacter lwoffii*^{wgs}, *Acinetobacter nosocomialis*$^{g+p+wgs}$, *Acinetobacter pittii*$^{g+p+wgs}$, *Aeromonas caviae*$^{g+p+wgs}$, *Aeromonas enteropelogenes*$^{g+wgs}$, *Aeromonas*

Similar principle: annotate to existing database of functions

Gene annotation with RGI (CARD): Annotation foibles

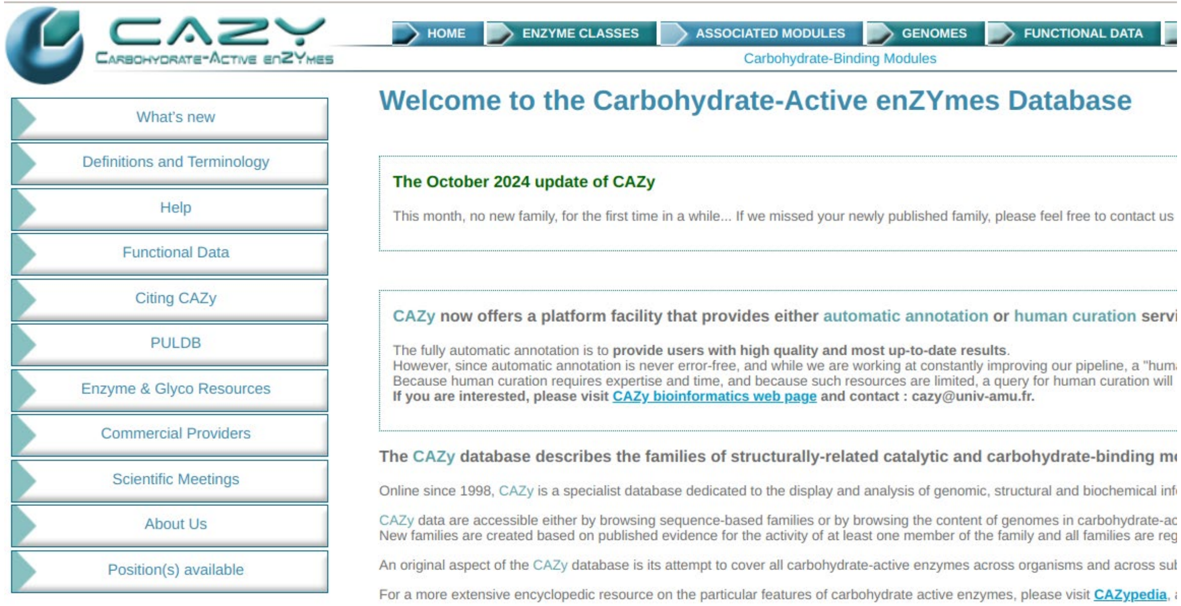
Strict: a match above threshold
Loose: a match below threshold
Nudged: a match below threshold, but high identity over a fragment

Beware the false positives!

- Efflux pumps!
- vanR (the regulator) without the regulatees
    - probably not ARG
- …

There are other specialized databases

- CAZy is very popular
- https://www.cazy.org/



Welcome to the Carbohydrate-Active enZYmes Database

The October 2024 update of CAZy

This month, no new family, for the first time in a while... If we missed your newly published family, please feel free to contact us

CAZy now offers a platform facility that provides either automatic annotation or human curation servi

The fully automatic annotation is to **provide users with high quality and most up-to-date results**.
However, since automatic annotation is never error-free, and while we are working at constantly improving our pipeline, a "hum
Because human curation requires expertise and time, and because such resources are limited, a query for human curation will
If you are interested, please visit CAZy bioinformatics web page and contact : cazy@univ-amu.fr.

The CAZy database describes the families of structurally-related catalytic and carbohydrate-binding m

Online since 1998, CAZy is a specialist database dedicated to the display and analysis of genomic, structural and biochemical inf

CAZy data are accessible either by browsing sequence-based families or by browsing the content of genomes in carbohydrate-ac
New families are created based on published evidence for the activity of at least one member of the family and all families are reg

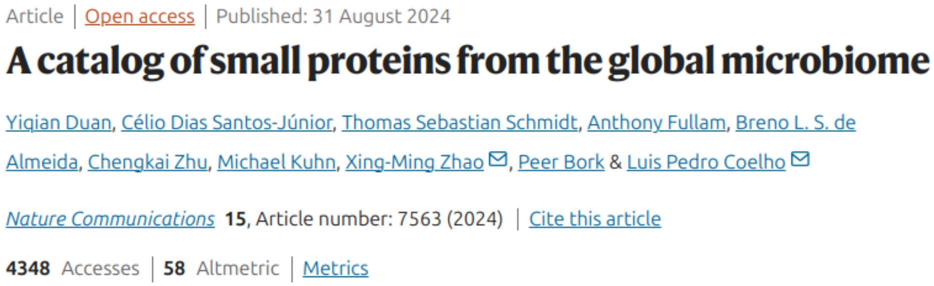
An original aspect of the CAZy database is its attempt to cover all carbohydrate-active enzymes across organisms and across sul

For a more extensive encyclopedic resource on the particular features of carbohydrate active enzymes, please visit CAZypedia, a

What about small genes?

Traditionally ignored

Some groups have started to look into them (and others).



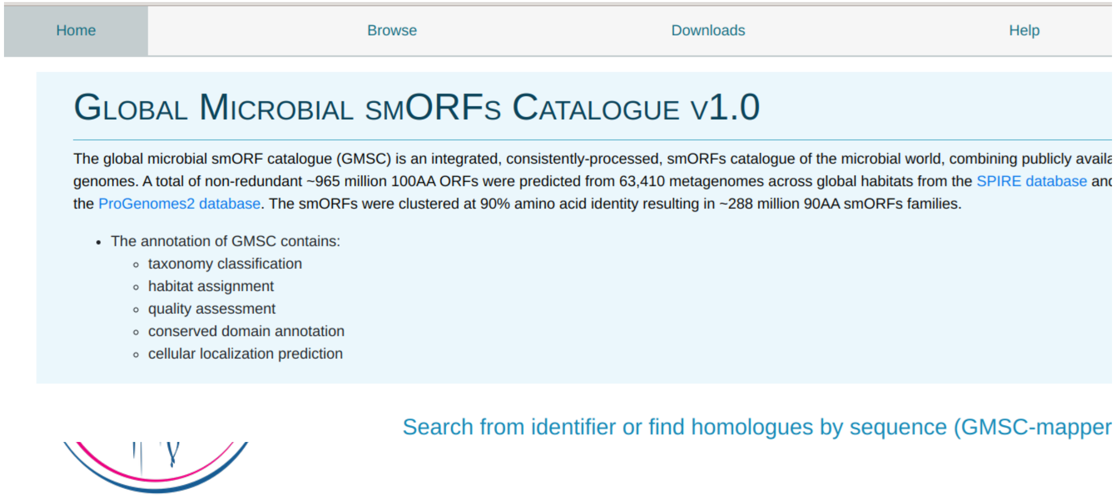
Article | Open access | Published: 31 August 2024

A catalog of small proteins from the global microbiome

Yiqian Duan, Célio Dias Santos-Júnior, Thomas Sebastian Schmidt, Anthony Fullam, Breno L. S. de Almeida, Chengkai Zhu, Michael Kuhn, Xing-Ming Zhao ✉, Peer Bork & Luis Pedro Coelho ✉

Nature Communications **15**, Article number: 7563 (2024) | Cite this article

4348 Accesses | **58** Altmetric | Metrics



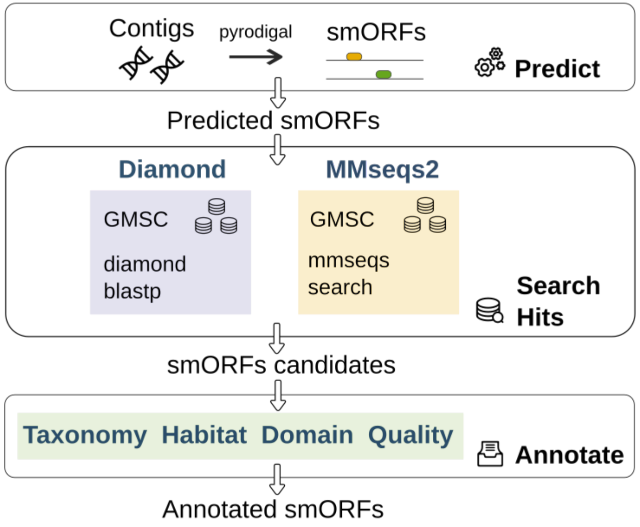
Home | Browse | Downloads | Help

Global Microbial smORFs Catalogue v1.0

The global microbial smORF catalogue (GMSC) is an integrated, consistently-processed, smORFs catalogue of the microbial world, combining publicly availa genomes. A total of non-redundant ~965 million 100AA ORFs were predicted from 63,410 metagenomes across global habitats from the SPIRE database anc the ProGenomes2 database. The smORFs were clustered at 90% amino acid identity resulting in ~288 million 90AA smORFs families.

- The annotation of GMSC contains:
  - taxonomy classification
  - habitat assignment
  - quality assessment
  - conserved domain annotation
  - cellular localization prediction

Search from identifier or find homologues by sequence (GMSC-mapper



Contigs → pyrodigal → smORFs

Predict

Predicted smORFs

Diamond: GMSC, diamond blastp

MMseqs2: GMSC, mmseqs search

Search Hits

smORFs candidates

Taxonomy Habitat Domain Quality

Annotate

Annotated smORFs

Public databases of MAGs

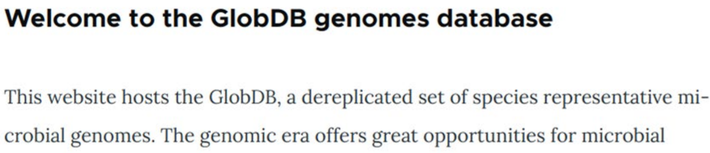
Large-scale genome recovery not submitted to NCBI (so not in GTDB):

- SPIRE (50k new species)
- mOTUs (50k new species)
- SMAG (13k new species)
- GEM (4k new species)
- OceanDNA (2k new species)

Public databases of MAGs – GlobDB

Globdb includes all of GTDB and SPIRE, mOTUs, etc.

- 306,260 species (~2x in GTDB)
- Standardised naming, QC, gene calling, annotation
- Complete taxonomy (including new clades)
- SingleM metapackage (taxonomic profiling)
- 82 million protein clusters + ProtT5 protein language model embeddings



Welcome to the GlobDB genomes database

This website hosts the GlobDB, a dereplicated set of species representative microbial genomes. The genomic era offers great opportunities for microbial

What can you do with these databases?

1. Download their MAGs
2. Compare your MAGs to theirs
   - Dereplication
   - Pangenome
   - Taxonomy (novel GlobDB clades coming soon to Sandpiper!)
3. Check geographic distributions

Physalia
Courses