Functional Annotation

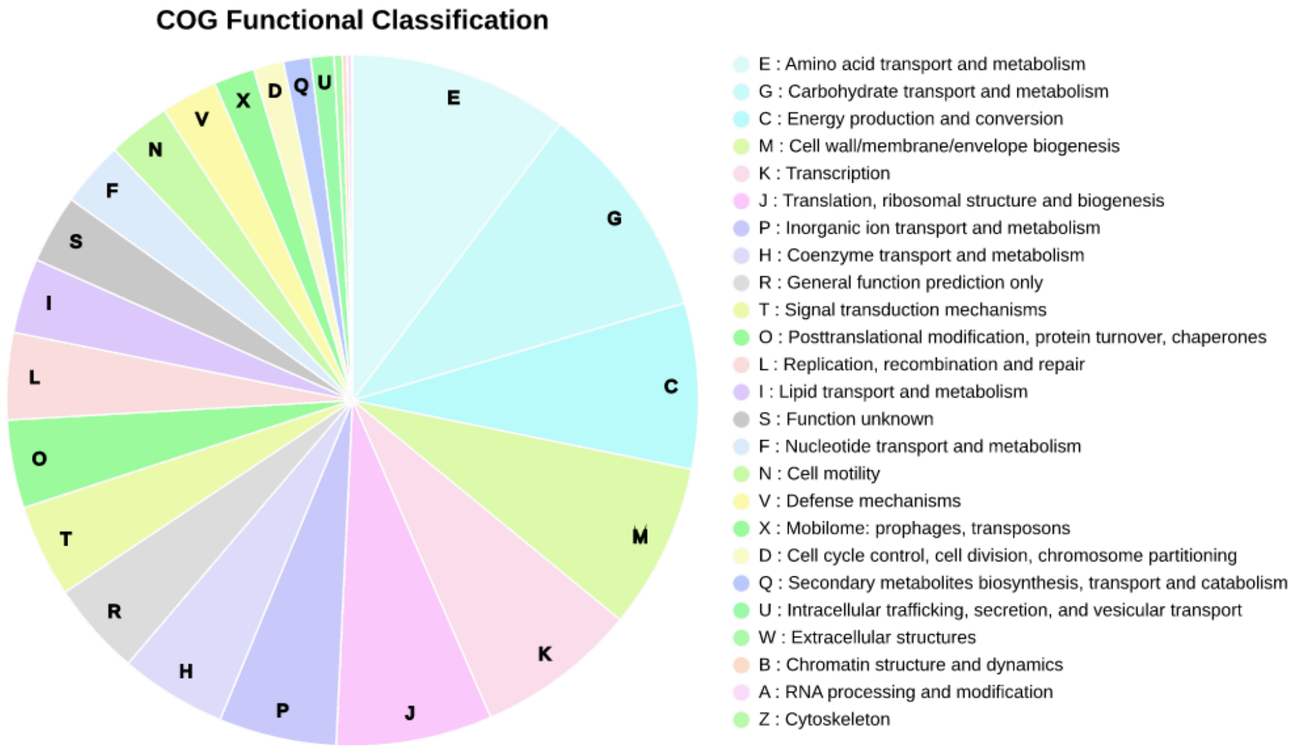
Jeferyd Yepes García

Swiss Institute of Bioinformatics

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

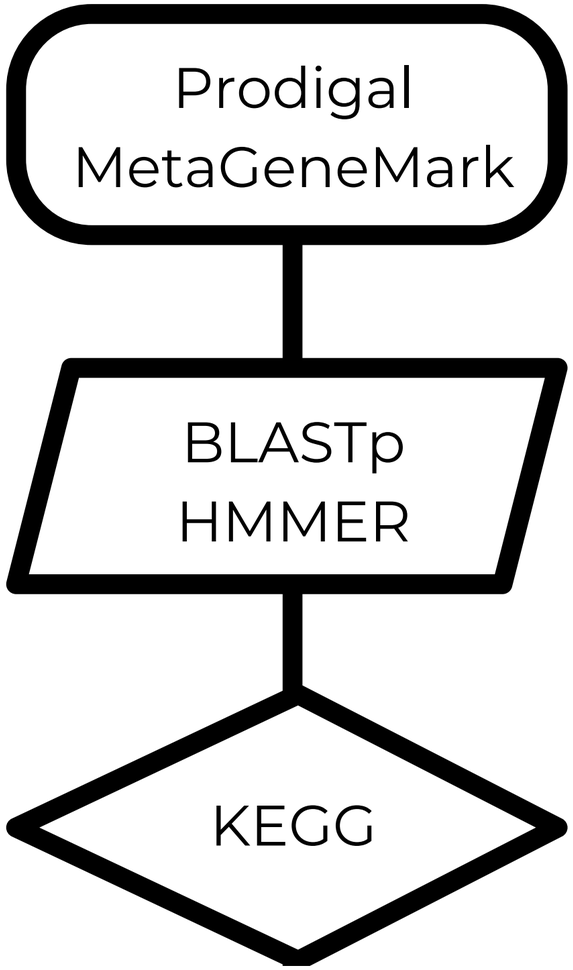
Importance

- Linking genetic content to ecological roles.
- Predicting metabolic pathways, environmental impacts, biotechnological potential.



Workflow

Gene Prediction



Characterization of genomic features.

Functional Annotation

Biological information.

Pathway Reconstruction

Integrating gene functions into biological pathways.

Gene Prediction

Identified features:

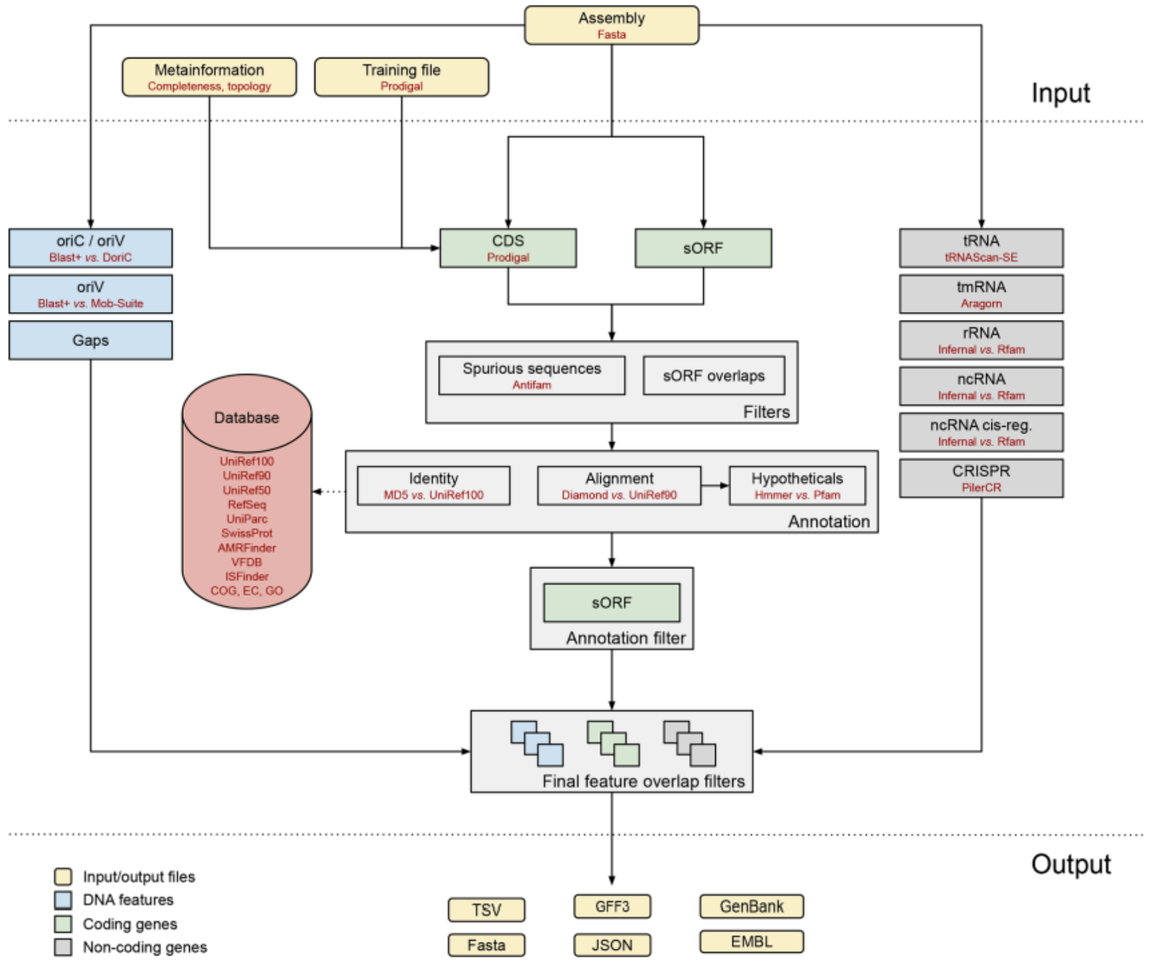
Prokka

https://github.com/tseemann/prokka

Rapid annotation tool of bacterial genomes
Numerous outputs

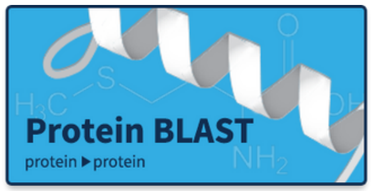
Tool (reference)	Features predicted
Prodigal (Hyatt 2010)	Coding sequence (CDS)
RNAmmer (Lagesen *et al.*, 2007)	Ribosomal RNA genes (rRNA)
Aragorn (Laslett and Canback, 2004)	Transfer RNA genes
SignalP (Petersen *et al.*, 2011)	Signal leader peptides
Infernal (Kolbe and Eddy, 2011)	Non-coding RNA

Seemann, 2014



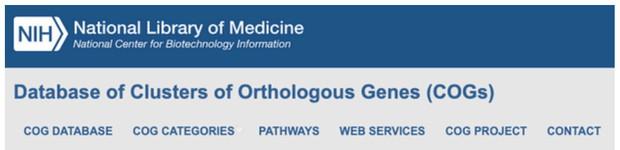
Functional Annotation

1. Homology-based methods:



2. Domain and Motif Detection:

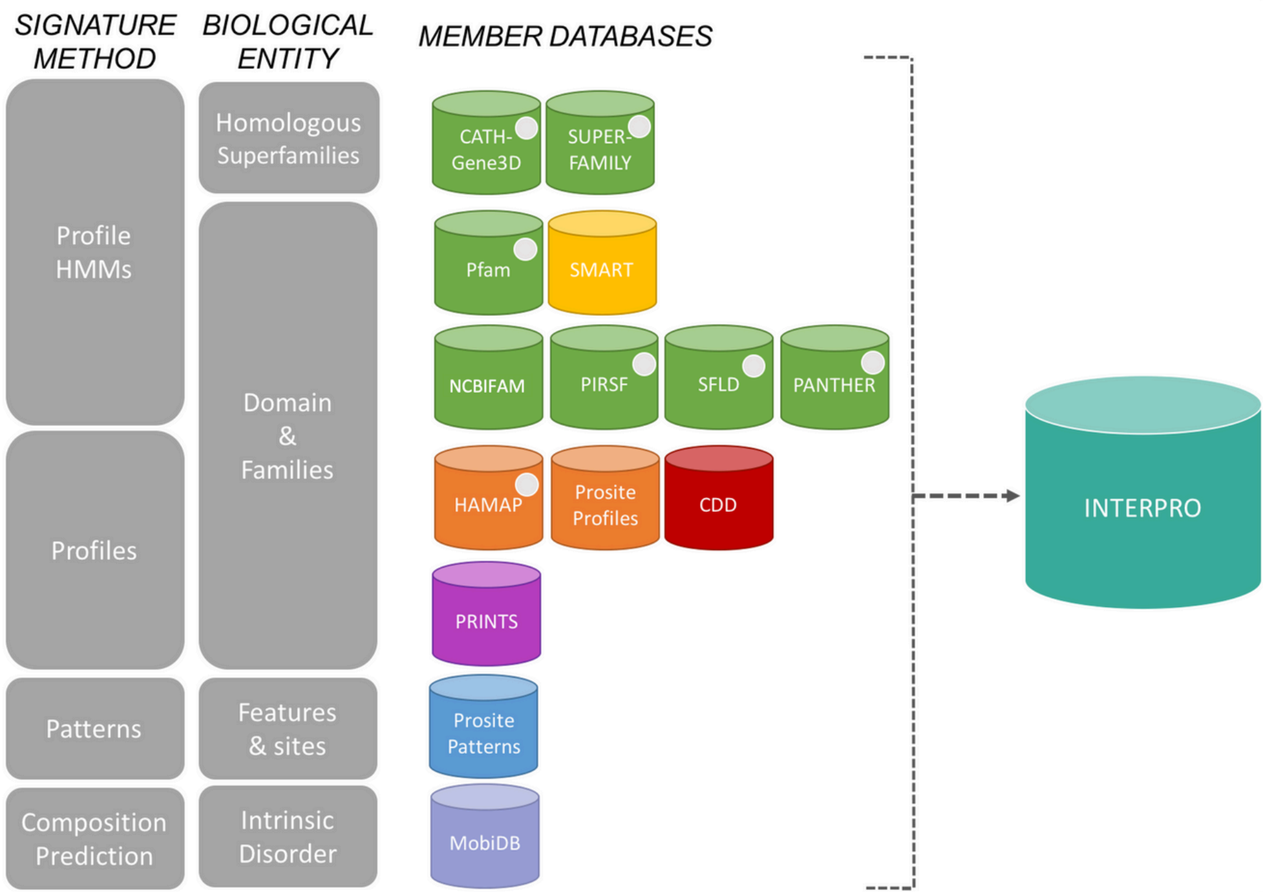
Databases



TIGRFAMs

Functional Annotation

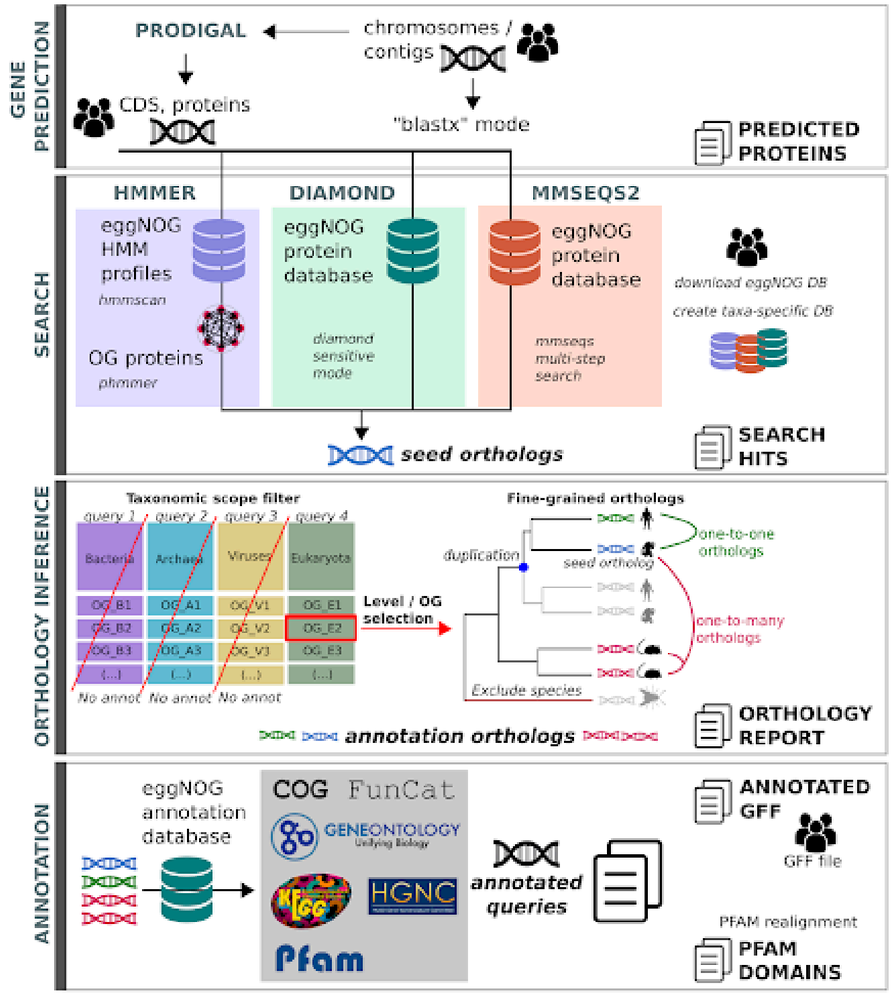
3. Automated pipelines:



Functional Annotation

3. Automated pipelines:

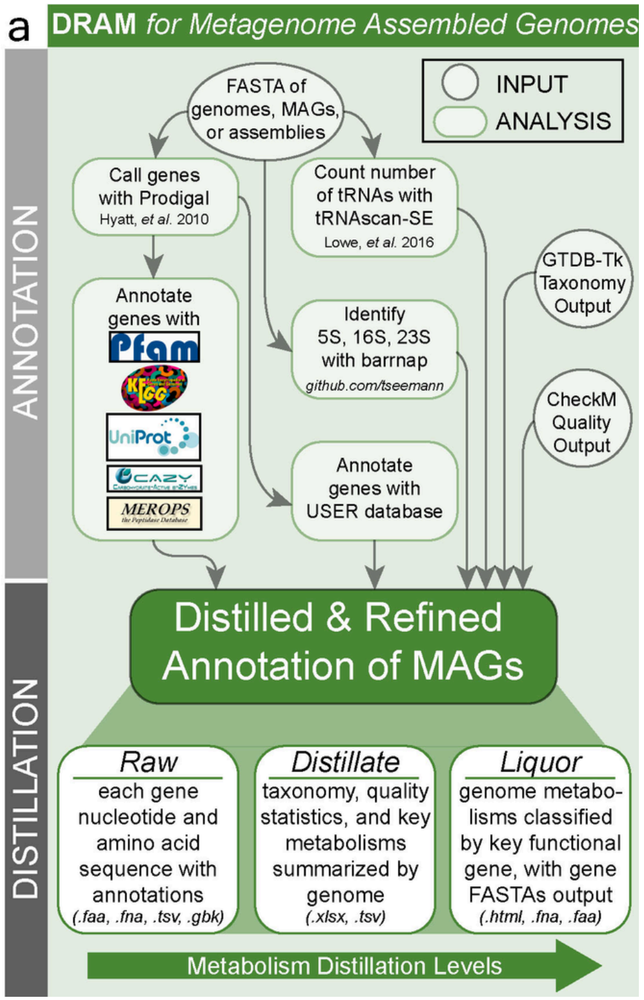
EggNOG mapper



Functional Annotation

3. Automated pipelines:

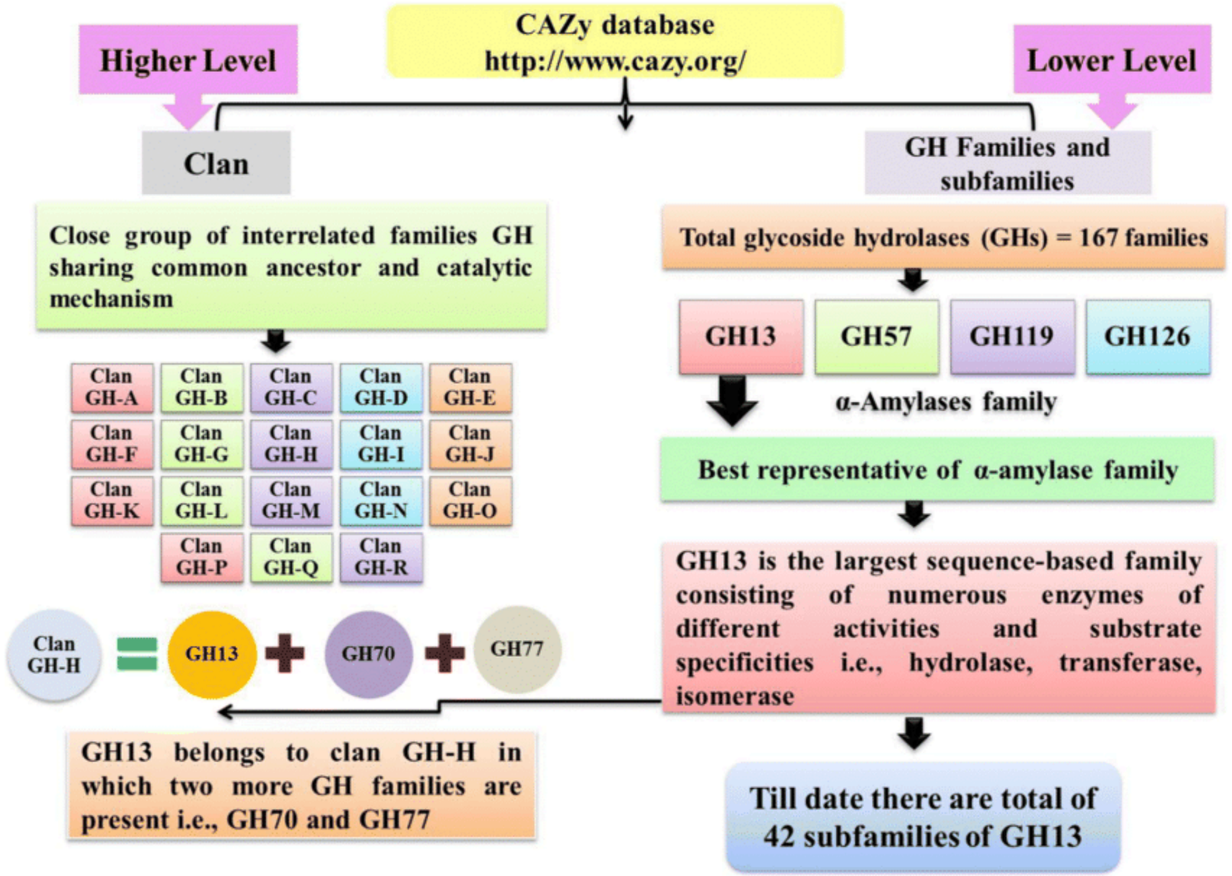
DRAM



Functional Annotation

Specialized tools and databases:

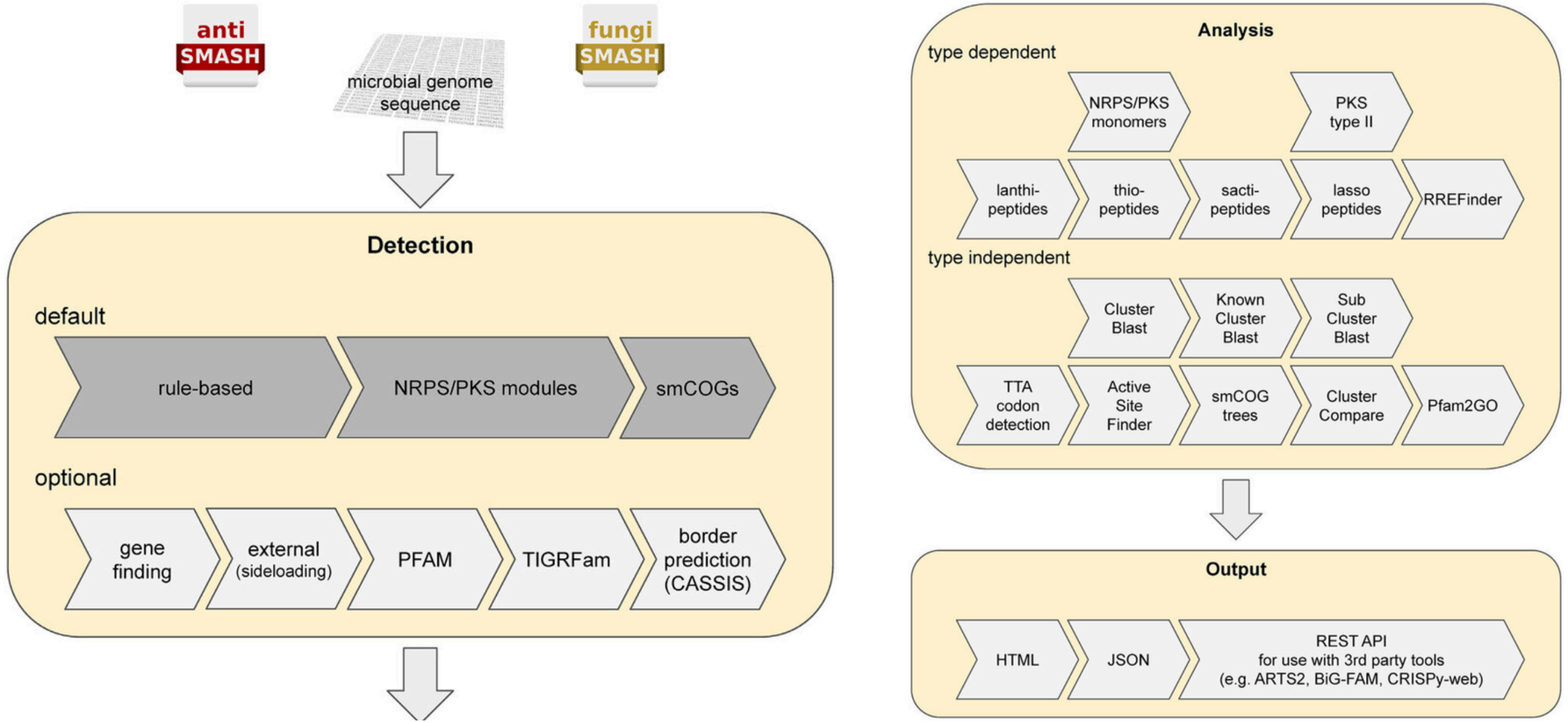
CAZy and dbCAN3



Functional Annotation

Specialized tools and databases:

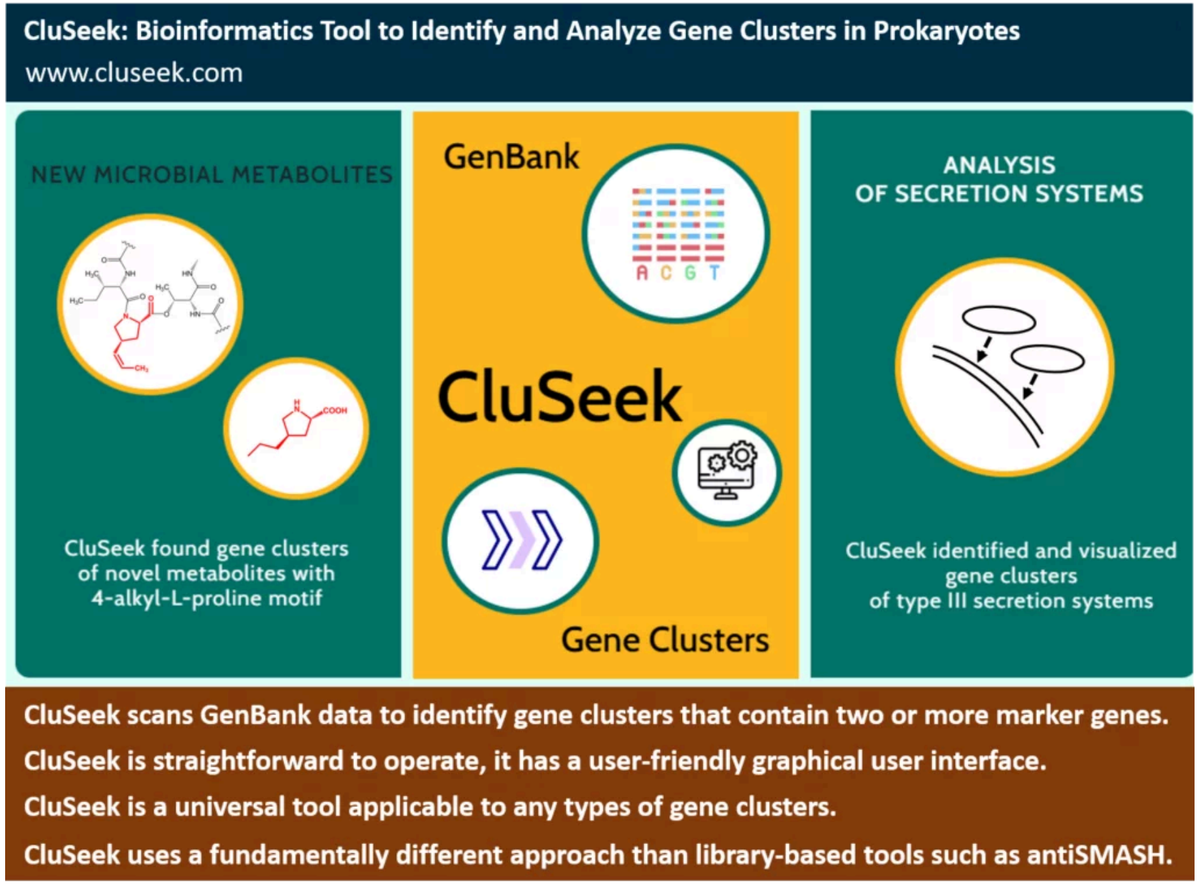
antiSMASH



Functional Annotation

Specialized tools and databases:

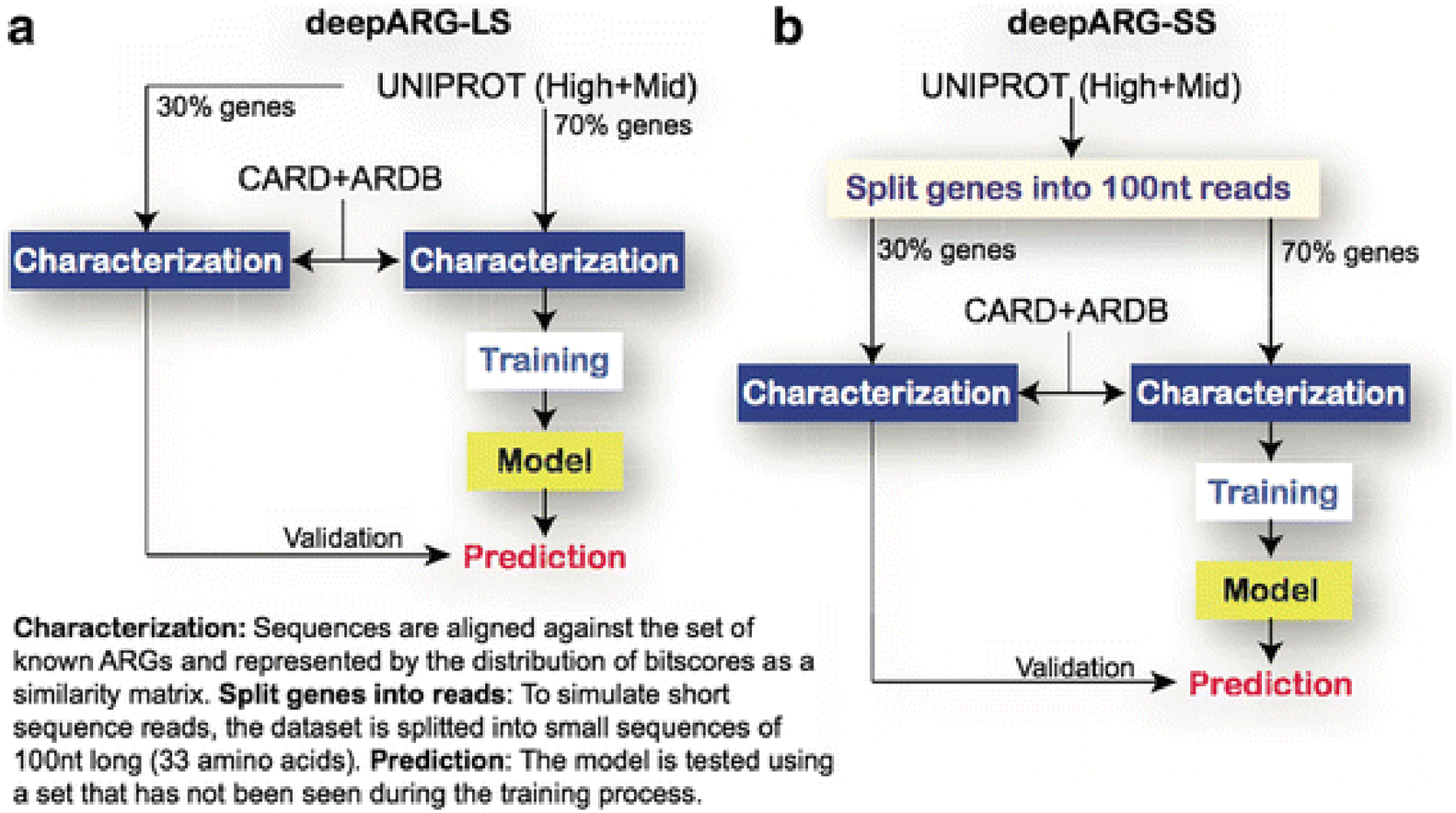
CluSeek



Functional Annotation

Specialized tools and databases:

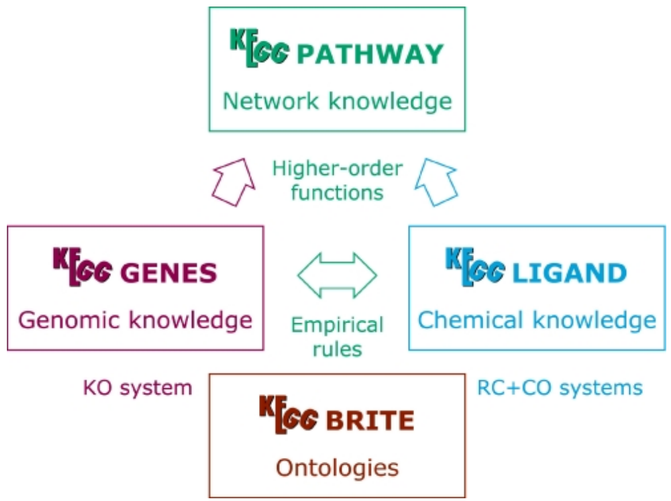
deepARG

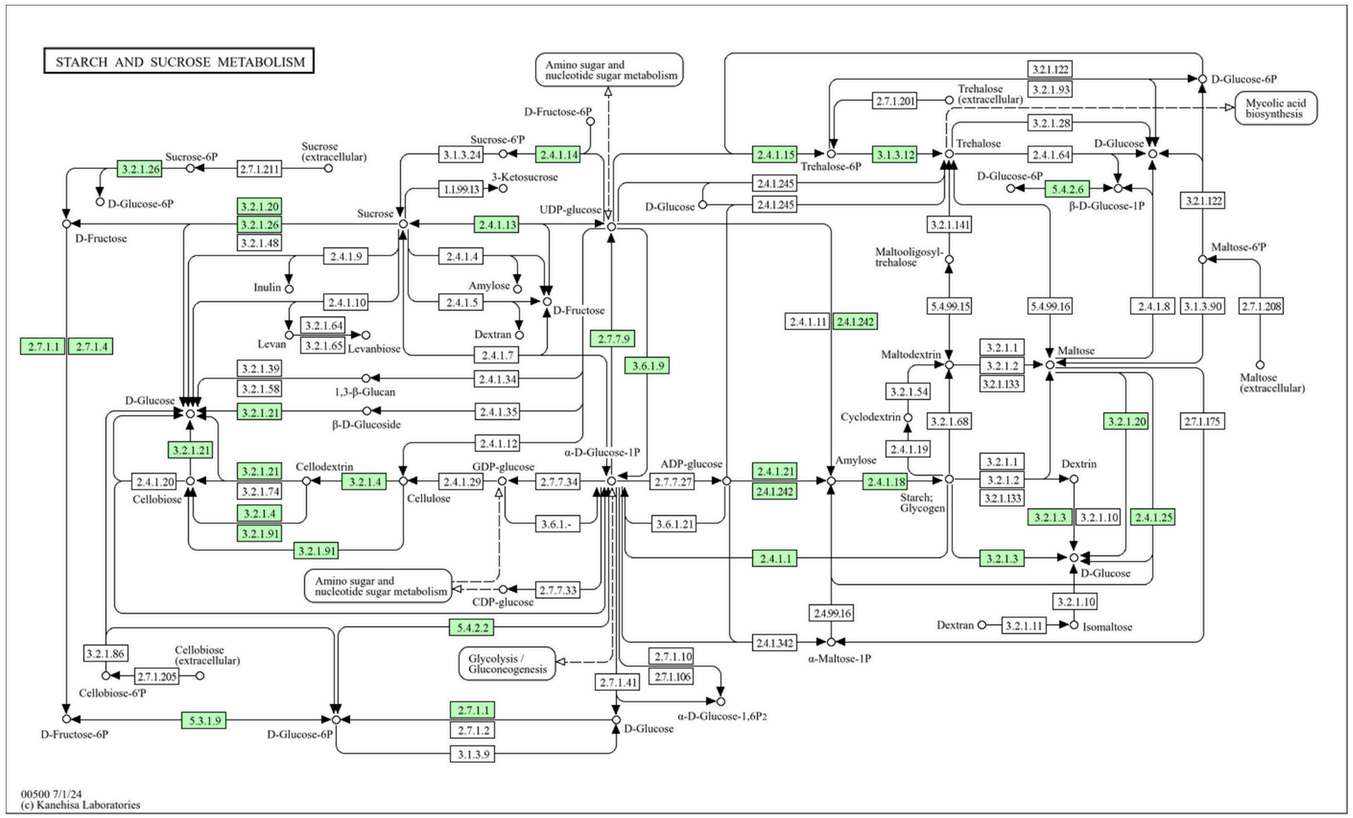


Characterization: Sequences are aligned against the set of known ARGs and represented by the distribution of bitscores as a similarity matrix. **Split genes into reads**: To simulate short sequence reads, the dataset is splitted into small sequences of 100nt long (33 amino acids). **Prediction**: The model is tested using a set that has not been seen during the training process.

Pathway Reconstruction

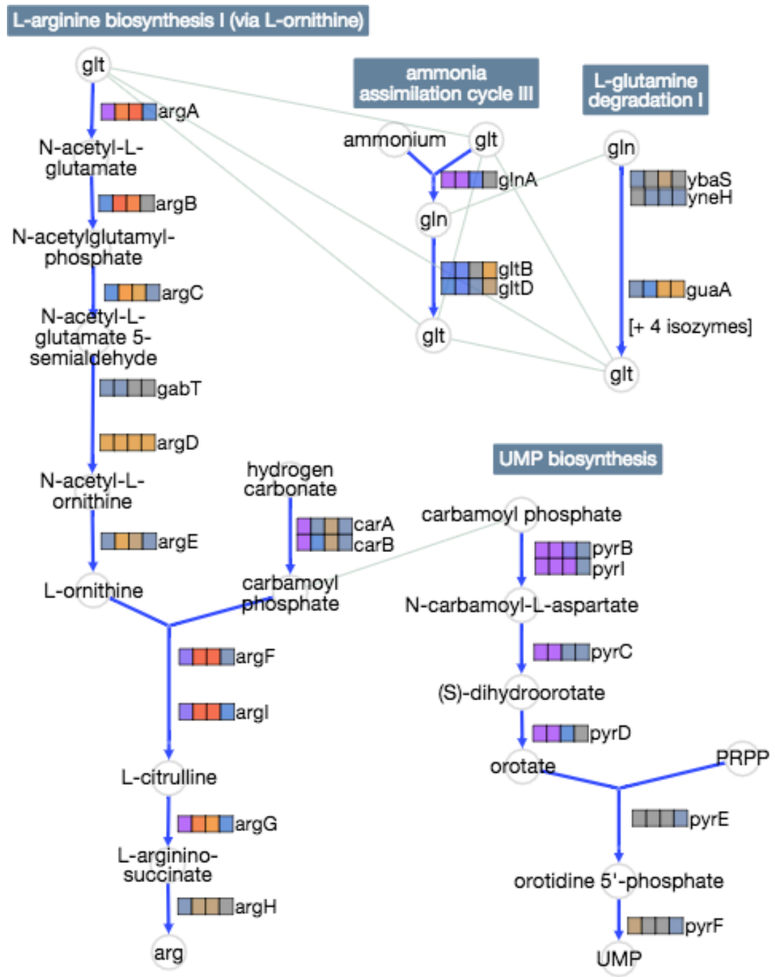
KEGG Pathway Mapper:





Pathway Reconstruction

Pathway Tools (BioCyc):



Challenges

- Incomplete genomes (fragmented, missing genes).
- Contamination and misassembly issues.
- Lack of close reference genomes (novelty of many MAGs).
- Biases in existing functional databases.

Take-home messages

- MAG functional annotation is a **multi-step process** that requires several tools and databases.

- Combination of tools with **different scope and functionalities** can expand the knowledge about ecological functions a MAG can perform.

- Automated pipelines are **interesting approches**, although they may be slow, web-server dependent or time-consuming during debugging.

Let’s have fun!

- Complete the practical tutorial as suggested with the different tools. Afterwards, choose one of the provided MAGs and show interesting annotations relying in Proksee visualization.