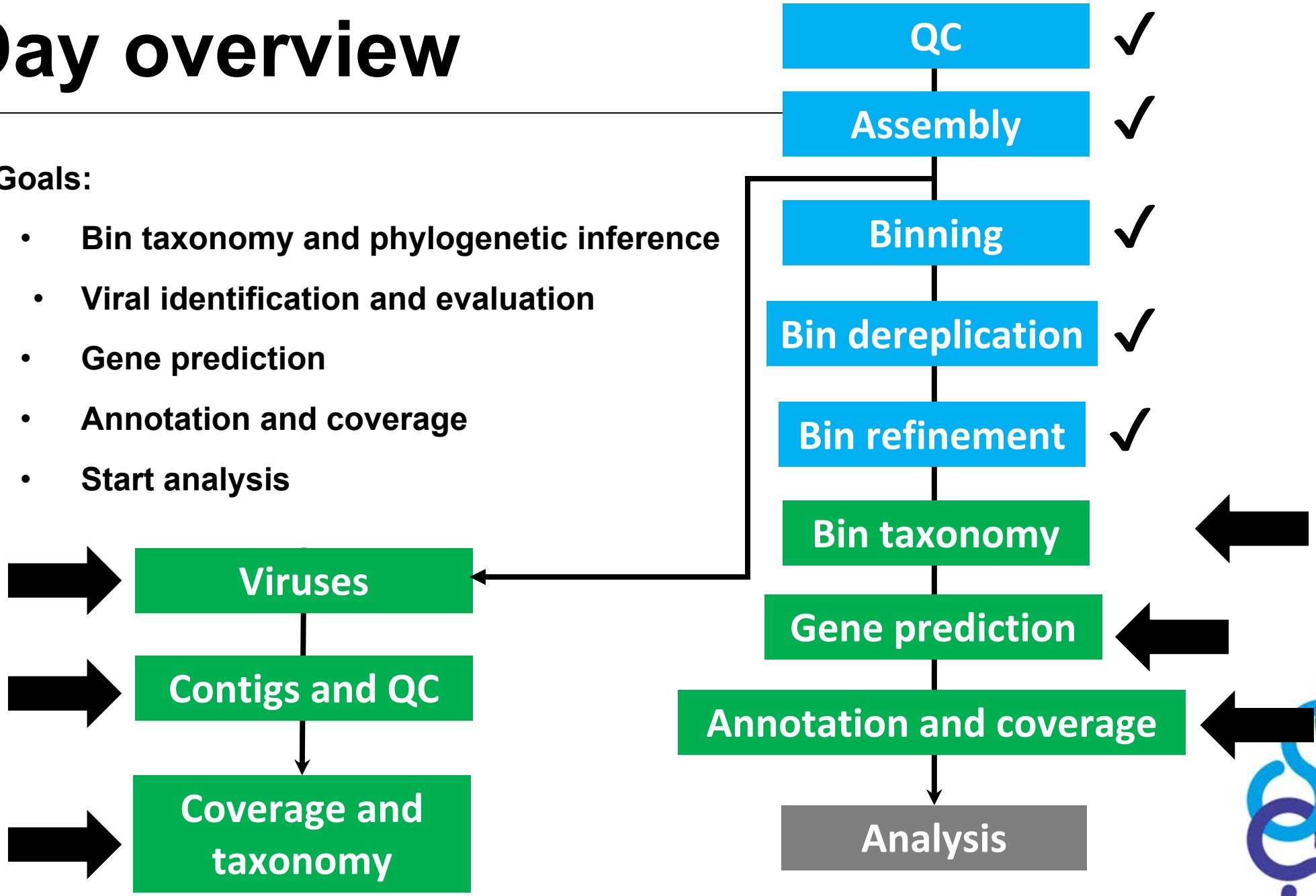
# Day 3

## Taxonomic classification & phylogenetic inference
## Viruses
## Gene prediction
## Gene annotation

# Day overview

- Goals:
  - Bin taxonomy and phylogenetic inference
  - Viral identification and evaluation
  - Gene prediction
  - Annotation and coverage
  - Start analysis



# Bin taxonomic classification

# Bin taxonomic classification

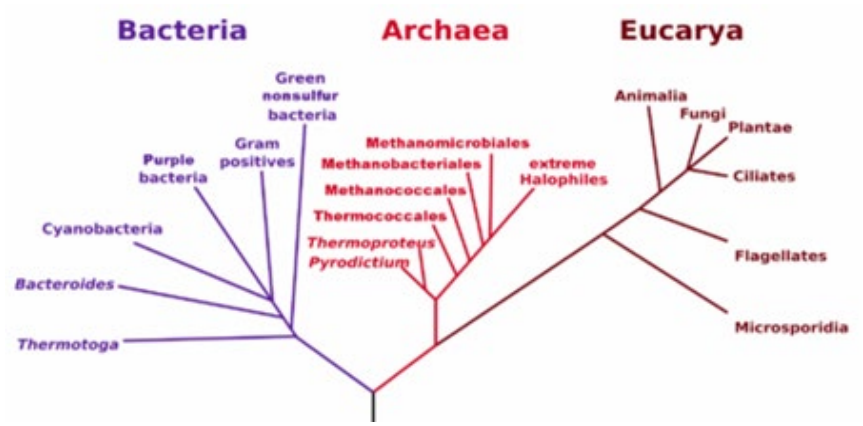
"…experience in many fields of science has demonstrated that naming is essential for accurate communication. Thus, astronomers have developed a system for naming stars, chemists a system for naming compounds, and biochemists a system for naming enzymes. All these fields share a common motivation, since it is simply not possible to provide precise descriptions of nature without clearly identifying the entities involved."

- William Whitman

# Bin taxonomic classification

- Taxonomy is a useful abstraction of the evolutionary process
  - Captures major routes of diversification
  - Not a perfect representation

- Fixed on a single molecular marker – 16S rRNA gene



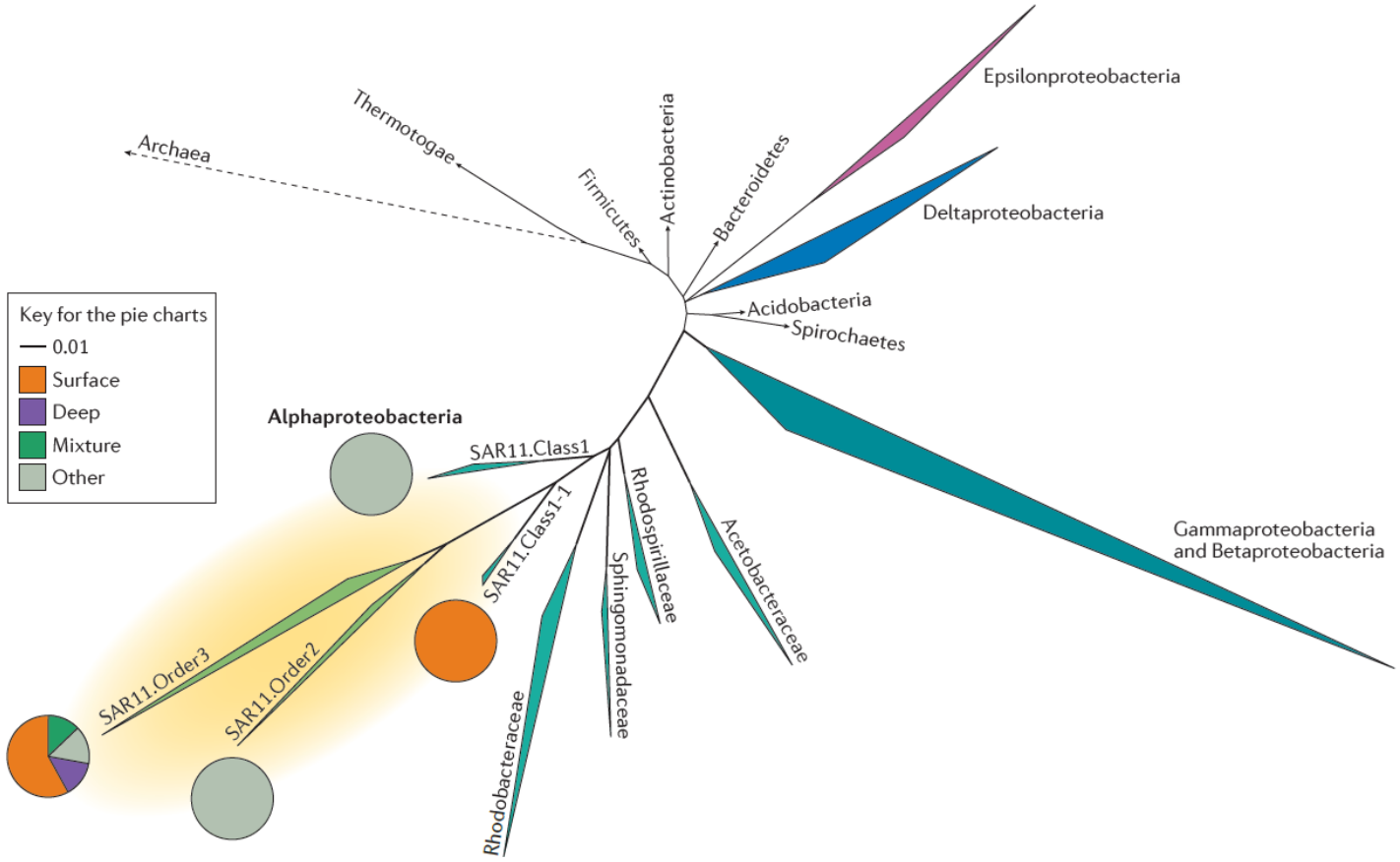

Woese *et al.* (1990) *Proc Natl Acad Sci USA 87: 4576*

# Bin taxonomic classification

- 16S rRNA commonly not recovered by *de novo* assembly

- Can recover 16S and 18S using EMIRGE
- Caveat: can be difficult to assign to genomes in complex communities with many similar taxa

# Bin taxonomic classification

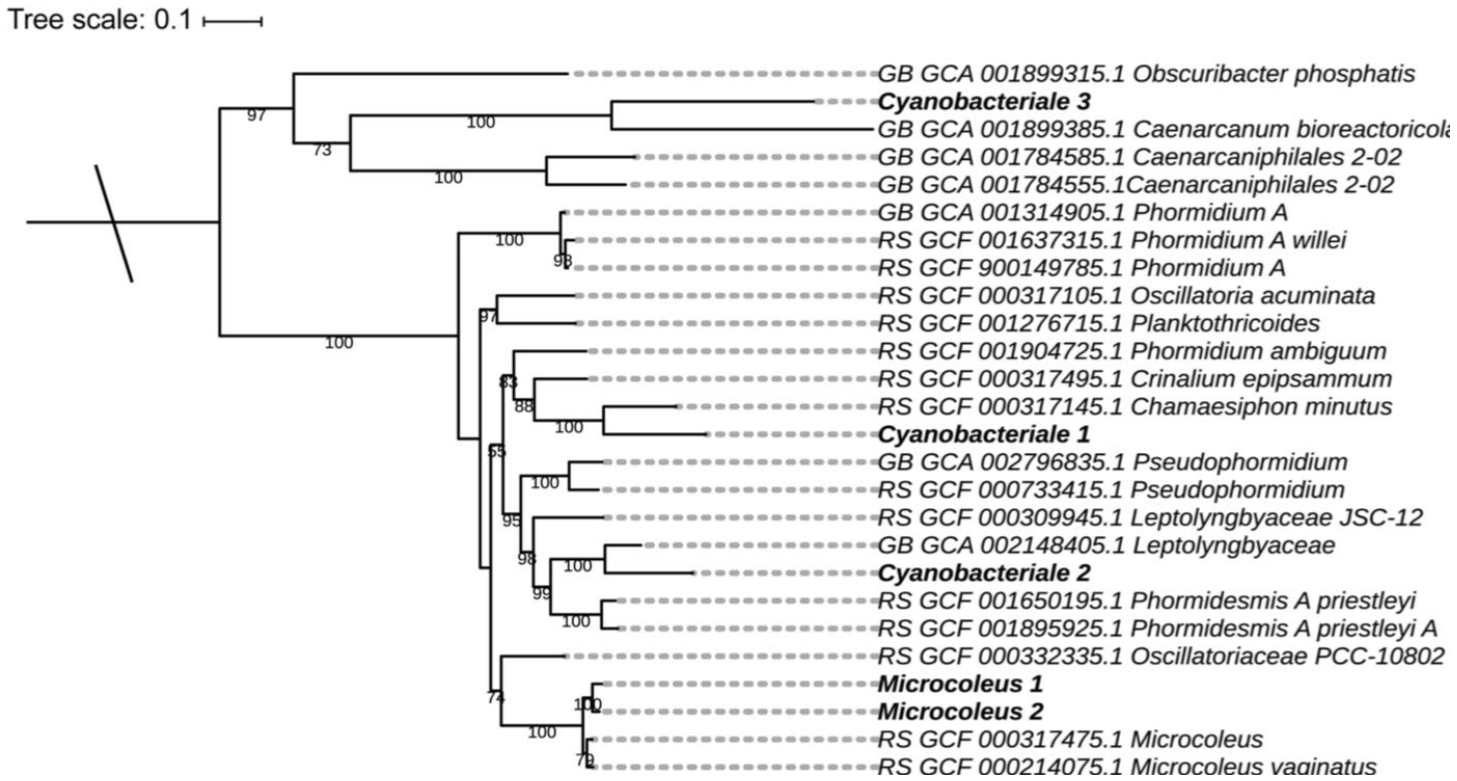



Yarza *et al.* (2014) *Nat Rev Microbiol 12, 635–645*

# Bin taxonomic classification

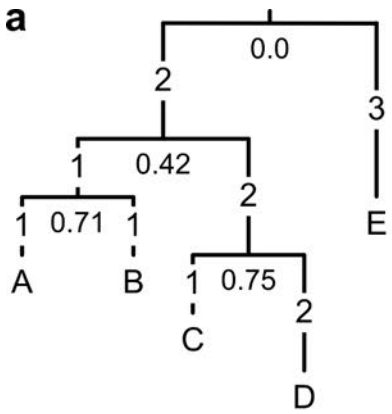
**Solution:**

- **Use one or more single copy core genes**
- **Concatenate protein sequences of multiple single copy core genes**

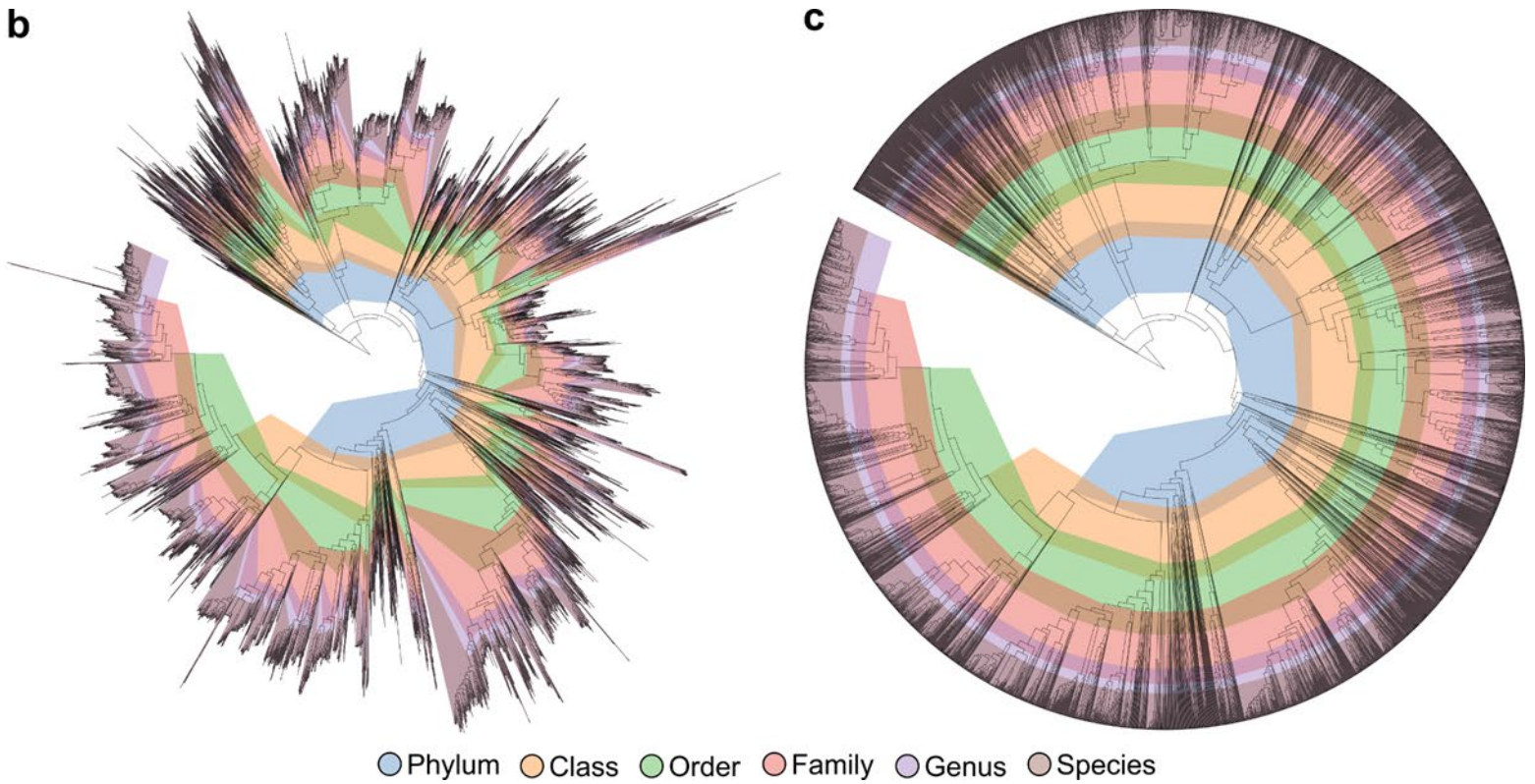



Concatenated protein sequence tree:
Phylogenetic placement of cyanobacterial genome bins (Wai-iti River, Nelson)

# A basis for robust phylogeny

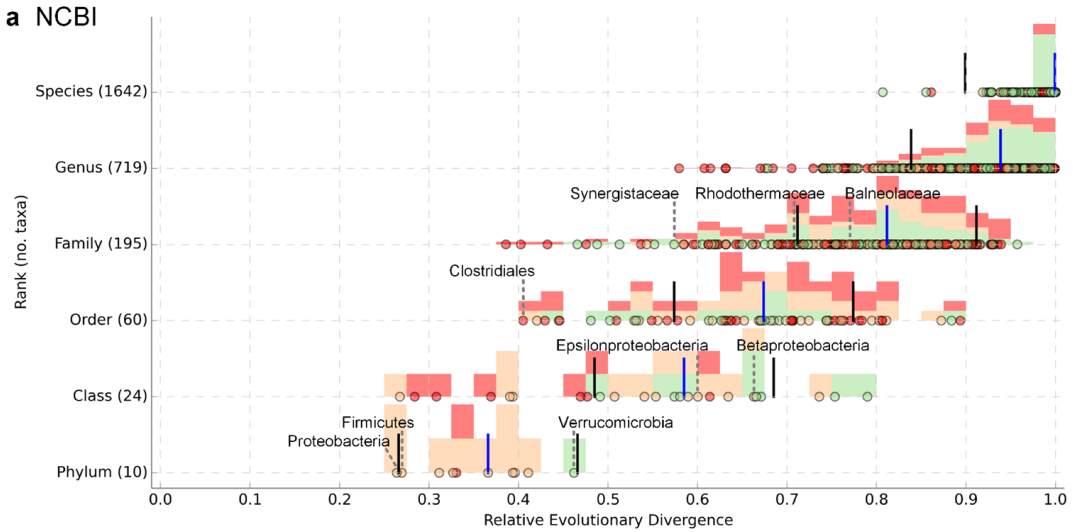


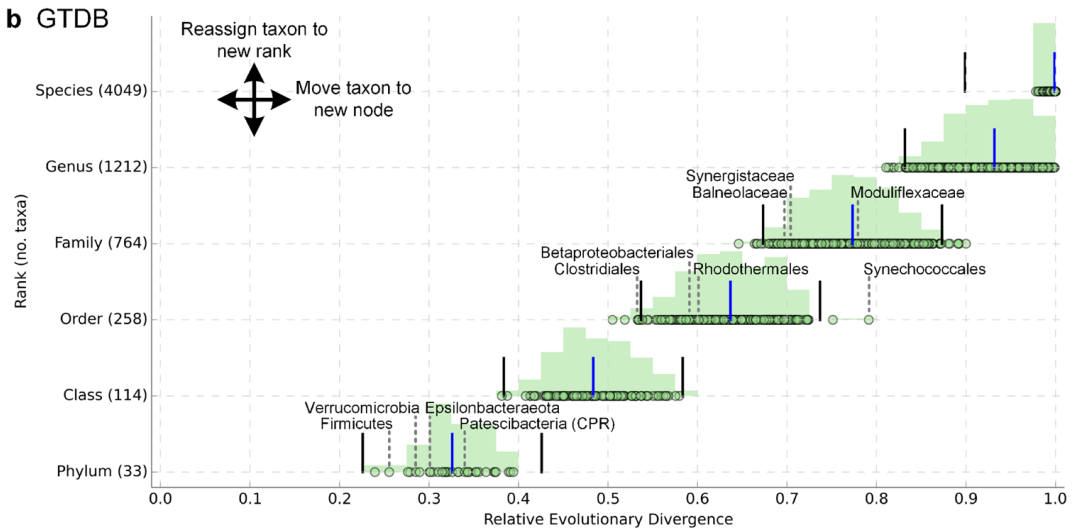

$$RED = \frac{d}{u}(1-p)$$

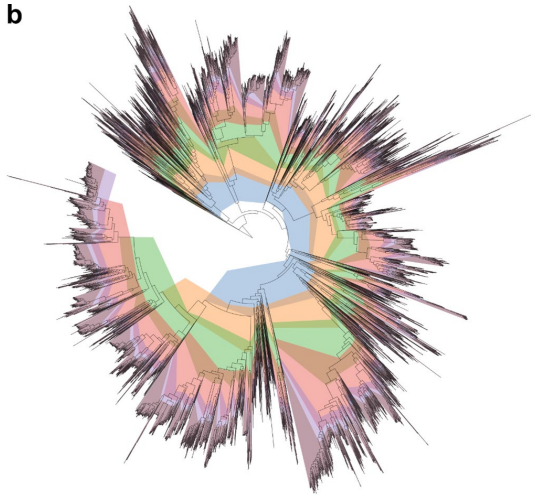


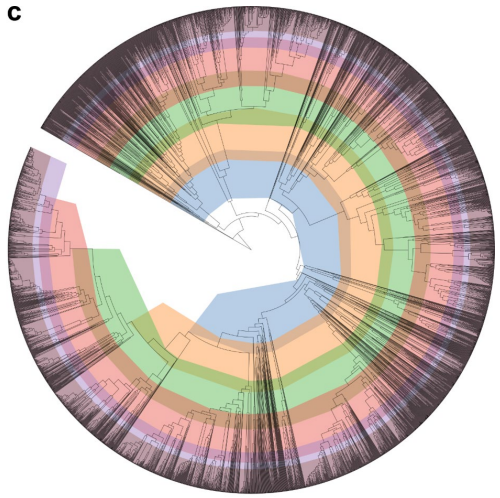

Parks *et al.* (2018) *Nat Biotechnol* 36: 996

# A basis for robust phylogeny



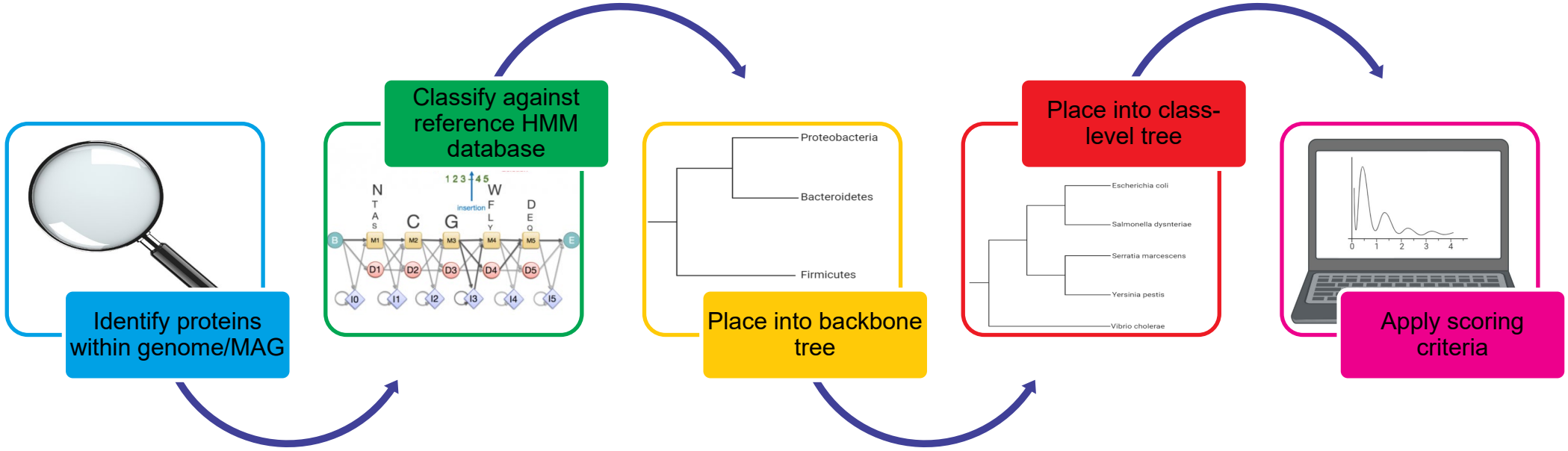




Parks *et al.* (2018) *Nat Biotechnol* 36: 996

# GTDB-TK version 2.1 – how it works



# Discriminate species

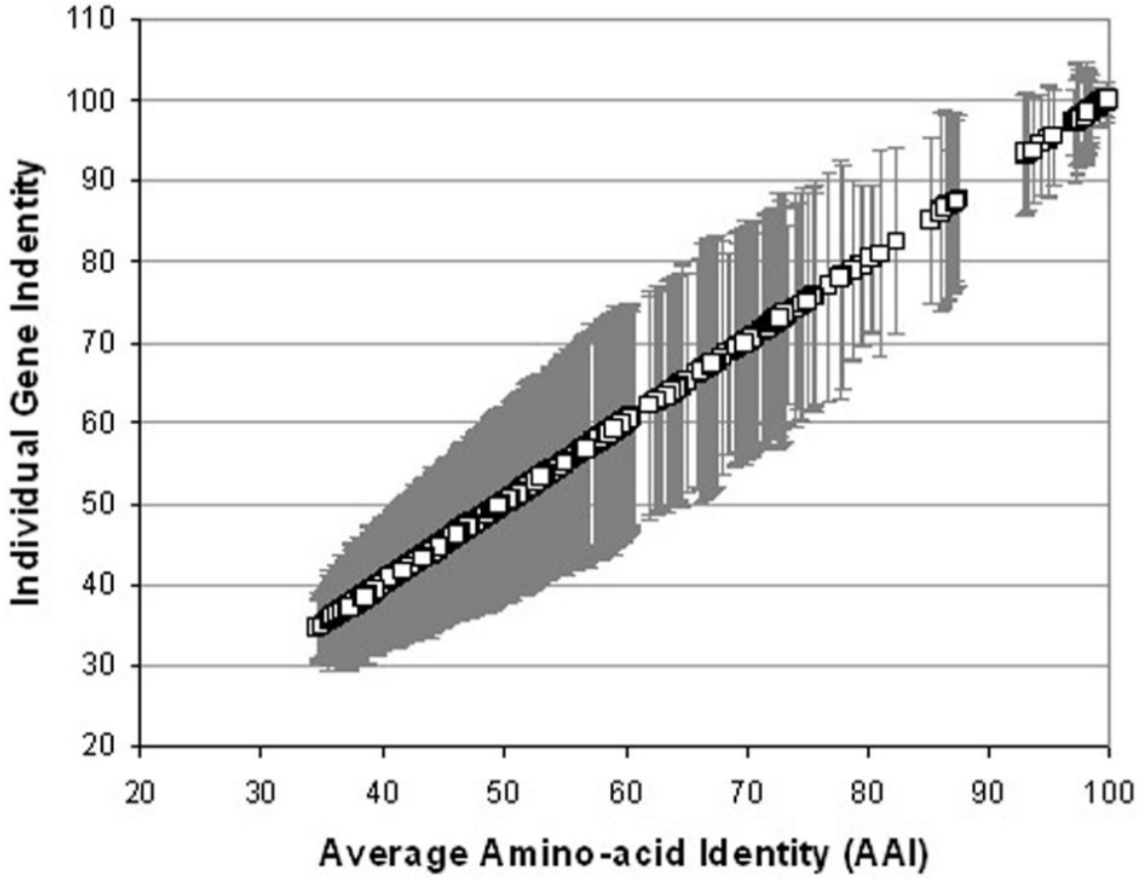
**Proxy for DNA-DNA hybridization**

**Pairwise genome comparisons:**

- **Average Nucleotide Identities (ANI)**
  - **gene comparisons**
- **Average Amino Acid Identities (AAI)**
  - **predicted protein comparisons**
- **Alignable Fraction (AF)**
  - **proportion of genes that align**

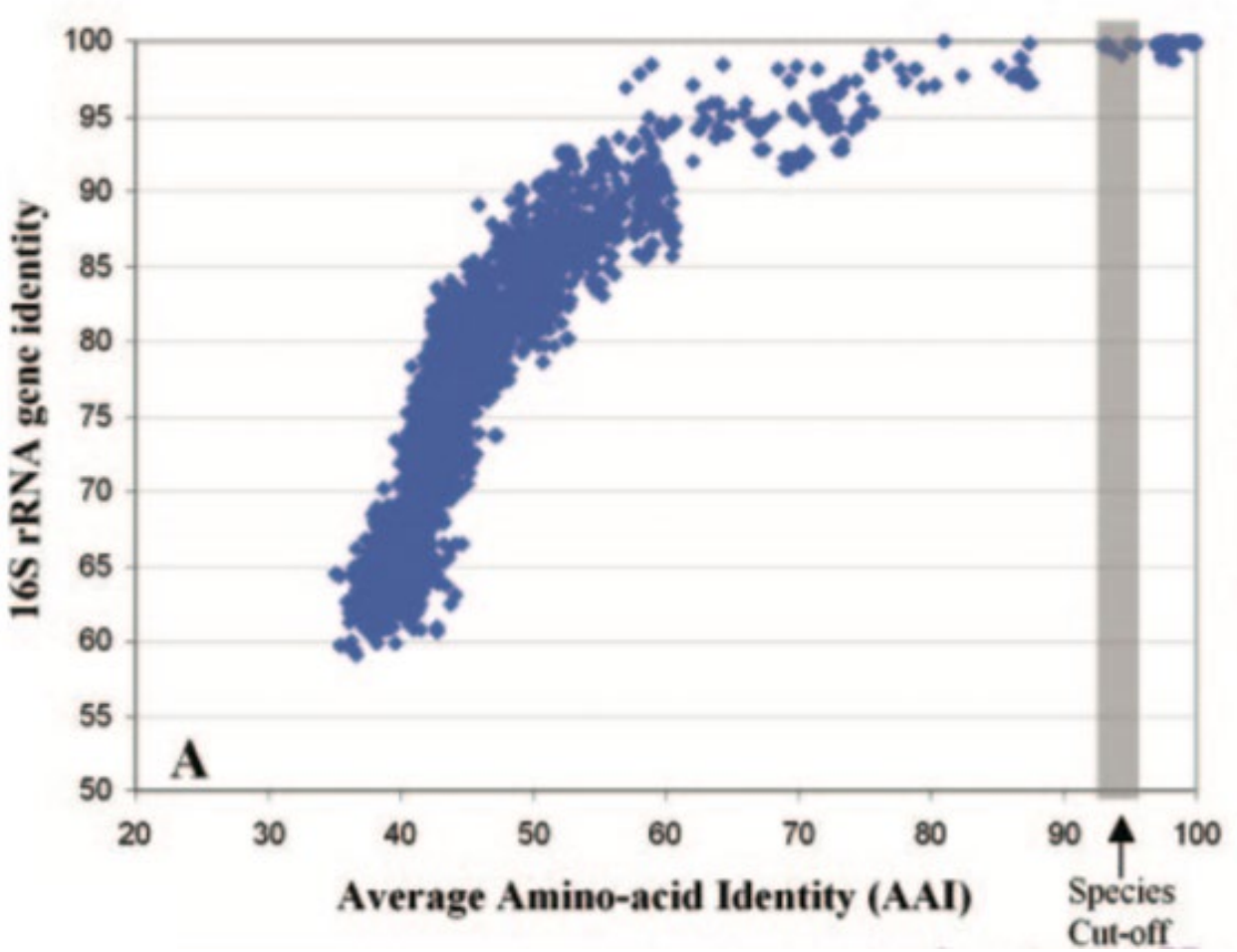
**Determine via: Pairwise BLAST-like search**

# Discriminate species




(Fig. 1, Konstantinidis and Tiedje, 2005, J Bacteriology)

# Discriminate species

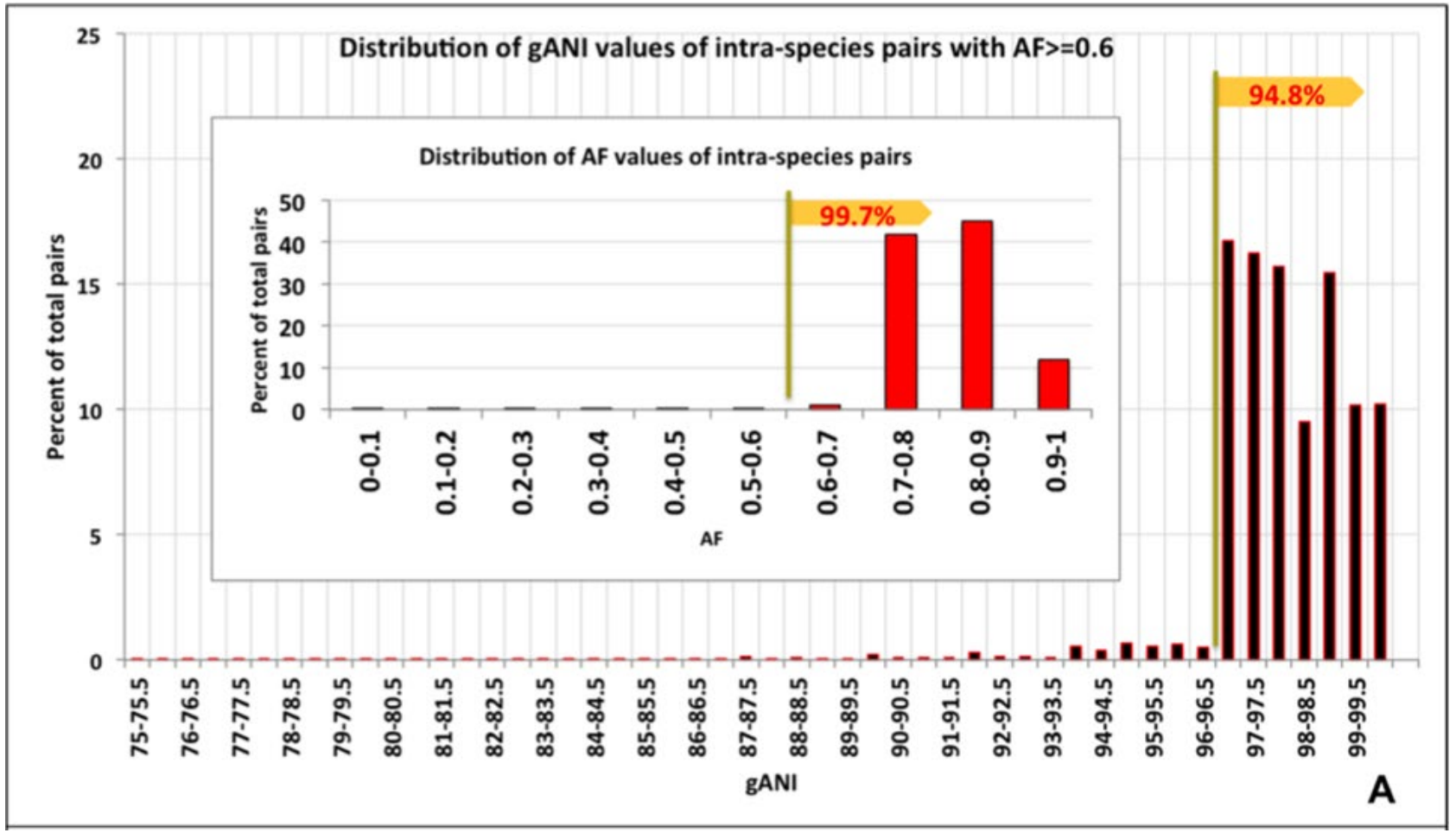



- AAI species cutoff ≈95-96% ID
- Equivalent to 70% DNA-DNA hybridization threshold for species

(Fig. 3, Konstantinidis and Tiedje, 2005, J Bacteriology)

# Discriminate species

10,998 IMG genomes -- 1,130,980 intra-species genome pairs




(Fig. 1A, Varghese et al., 2015, Nucleic Acids Research)

# Discriminate species

**Table S6. Pairwise average amino acid identities (AAI) shared between genome bins.**

| GSC | 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 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100 | 80 | 86 | 49 | 49 | 45 | 49 | 53 | 51 | 51 | 49 | 49 | 49 | 52 | 52 | 50 | 49 | 52 | 51 | 50 | 52 | 49 | 49 | 52 | 49 | 47 | 45 | 45 | 44 | 44 | 44 | 45 | 44 | 43 | 43 | 40 | 43 | 42 | 44 | 44 | 44 |
| 2 | 79 | 100 | 67 | 48 | 49 | 43 | 48 | 51 | 50 | 47 | 50 | 48 | 50 | 53 | 51 | 45 | 50 | 52 | 53 | 49 | 52 | 49 | 48 | 53 | 48 | 44 | 42 | 44 | 46 | 44 | 45 | 45 | 46 | 42 | 44 | 32 | 44 | 40 | 46 | 41 | 44 |
| 3 | 86 | 68 | 100 | 49 | 49 | 46 | 49 | 53 | 52 | 52 | 50 | 49 | 49 | 52 | 52 | 51 | 49 | 52 | 52 | 50 | 52 | 50 | 49 | 51 | 49 | 48 | 44 | 44 | 43 | 44 | 44 | 45 | 44 | 43 | 42 | 41 | 44 | 43 | 44 | 44 | 45 |
| 4 | 49 | 48 | 49 | 100 | 81 | 82 | 74 | 50 | 49 | 49 | 48 | 49 | 48 | 50 | 49 | 50 | 50 | 50 | 50 | 48 | 50 | 46 | 51 | 51 | 46 | 46 | 43 | 42 | 43 | 43 | 43 | 44 | 44 | 42 | 42 | 41 | 42 | 41 | 42 | 42 | 42 |
| 5 | 49 | 49 | 49 | 80 | 100 | 74 | 73 | 50 | 48 | 50 | 49 | 48 | 48 | 50 | 49 | 50 | 49 | 50 | 50 | 48 | 50 | 46 | 51 | 51 | 46 | 47 | 43 | 43 | 42 | 43 | 43 | 44 | 44 | 42 | 42 | 44 | 42 | 41 | 42 | 43 | 42 |
| 6 | 45 | 43 | 46 | 82 | 74 | 100 | 67 | 48 | 49 | 47 | 47 | 47 | 46 | 47 | 46 | 48 | 48 | 48 | 46 | 46 | 47 | 43 | 47 | 47 | 43 | 44 | 38 | 39 | 39 | 41 | 39 | 41 | 42 | 40 | 40 | 41 | 41 | 40 | 40 | 41 | 41 |
| 7 | 49 | 49 | 49 | 74 | 73 | 67 | 100 | 50 | 49 | 50 | 49 | 48 | 48 | 50 | 50 | 49 | 50 | 50 | 50 | 48 | 50 | 46 | 51 | 51 | 46 | 46 | 43 | 43 | 42 | 43 | 43 | 44 | 43 | 42 | 42 | 40 | 42 | 41 | 41 | 43 | 41 |
| 8 | 53 | 51 | 53 | 50 | 50 | 48 | 50 | 100 | 73 | 86 | 50 | 49 | 49 | 52 | 52 | 51 | 50 | 52 | 51 | 50 | 53 | 50 | 49 | 52 | 49 | 48 | 44 | 45 | 43 | 44 | 43 | 44 | 44 | 42 | 43 | 41 | 43 | 42 | 43 | 43 | 43 |
| 9 | 51 | 50 | 52 | 49 | 48 | 48 | 48 | 73 | 100 | 73 | 50 | 48 | 49 | 50 | 51 | 49 | 49 | 50 | 50 | 49 | 51 | 50 | 48 | 50 | 48 | 46 | 43 | 43 | 42 | 43 | 42 | 43 | 43 | 41 | 42 | 40 | 42 | 41 | 43 | 42 | 42 |
| 10 | 51 | 48 | 52 | 49 | 49 | 47 | 50 | 86 | 73 | 100 | 49 | 47 | 47 | 50 | 50 | 49 | 49 | 51 | 50 | 49 | 52 | 49 | 49 | 50 | 47 | 47 | 42 | 43 | 43 | 42 | 42 | 43 | 43 | 41 | 42 | 47 | 42 | 40 | 42 | 42 | 42 |
| 11 | 50 | 50 | 50 | 48 | 49 | 47 | 49 | 50 | 50 | 49 | 100 | 82 | 82 | 73 | 64 | 50 | 53 | 53 | 53 | 48 | 51 | 49 | 49 | 53 | 47 | 45 | 43 | 43 | 43 | 44 | 43 | 44 | 44 | 41 | 41 | 39 | 41 | 41 | 41 | 43 | 41 |
| 12 | 49 | 48 | 49 | 49 | 48 | 47 | 48 | 49 | 48 | 48 | 82 | 100 | 77 | 70 | 60 | 47 | 52 | 52 | 51 | 47 | 50 | 46 | 47 | 51 | 45 | 45 | 40 | 40 | 40 | 40 | 40 | 42 | 41 | 38 | 40 | 37 | 40 | 40 | 42 | 42 | 42 |
| 13 | 49 | 50 | 49 | 48 | 48 | 46 | 48 | 49 | 48 | 47 | 83 | 77 | 100 | 68 | 60 | 47 | 50 | 51 | 49 | 45 | 48 | 46 | 47 | 49 | 44 | 44 | 42 | 43 | 40 | 43 | 42 | 43 | 42 | 42 | 40 | 38 | 41 | 40 | 41 | 41 | 40 |
| 14 | 51 | 53 | 52 | 50 | 50 | 47 | 50 | 52 | 50 | 50 | 73 | 70 | 68 | 100 | 64 | 52 | 55 | 55 | 54 | 50 | 53 | 49 | 50 | 54 | 48 | 47 | 45 | 44 | 44 | 44 | 44 | 45 | 44 | 42 | 43 | 40 | 43 | 43 | 45 | 45 | 44 |
| 15 | 51 | 52 | 52 | 49 | 49 | 46 | 49 | 52 | 51 | 50 | 63 | 60 | 60 | 64 | 100 | 53 | 54 | 55 | 54 | 50 | 52 | 49 | 51 | 54 | 48 | 47 | 45 | 44 | 43 | 44 | 44 | 45 | 43 | 42 | 42 | 37 | 43 | 42 | 45 | 44 | 44 |
| 16 | 50 | 45 | 50 | 50 | 50 | 47 | 49 | 50 | 49 | 49 | 50 | 47 | 48 | 52 | 52 | 100 | 50 | 54 | 52 | 49 | 50 | 47 | 48 | 53 | 48 | 46 | 44 | 43 | 43 | 44 | 44 | 44 | 45 | 44 | 41 | 37 | 42 | 41 | 45 | 43 | 43 |
| 17 | 49 | 50 | 49 | 50 | 49 | 47 | 50 | 50 | 49 | 49 | 53 | 51 | 50 | 55 | 54 | 49 | 100 | 53 | 52 | 48 | 50 | 47 | 49 | 53 | 47 | 46 | 43 | 43 | 42 | 43 | 44 | 43 | 42 | 41 | 42 | 37 | 42 | 39 | 42 | 41 | 42 |
| 18 | 52 | 52 | 52 | 51 | 50 | 47 | 50 | 52 | 51 | 51 | 53 | 52 | 51 | 56 | 55 | 54 | 53 | 100 | 55 | 49 | 52 | 50 | 50 | 55 | 48 | 47 | 44 | 45 | 44 | 44 | 44 | 45 | 45 | 43 | 42 | 38 | 42 | 42 | 44 | 44 | 43 |
| 19 | 51 | 52 | 52 | 50 | 49 | 46 | 50 | 51 | 50 | 50 | 53 | 51 | 50 | 54 | 54 | 52 | 53 | 55 | 100 | 49 | 52 | 50 | 51 | 54 | 49 | 47 | 44 | 44 | 44 | 44 | 44 | 45 | 45 | 42 | 43 | 41 | 43 | 41 | 44 | 44 | 43 |
| 20 | 50 | 48 | 50 | 48 | 48 | 45 | 48 | 50 | 49 | 49 | 48 | 47 | 45 | 50 | 50 | 49 | 47 | 49 | 49 | 100 | 50 | 47 | 51 | 50 | 50 | 50 | 44 | 44 | 43 | 44 | 44 | 44 | 44 | 42 | 42 | 41 | 43 | 42 | 43 | 44 | 43 |
| 21 | 52 | 53 | 53 | 50 | 50 | 47 | 49 | 54 | 52 | 52 | 51 | 50 | 48 | 53 | 52 | 50 | 50 | 53 | 52 | 50 | 100 | 50 | 49 | 52 | 49 | 48 | 44 | 44 | 44 | 44 | 44 | 45 | 45 | 43 | 43 | 42 | 43 | 43 | 45 | 45 | 43 |
| 22 | 49 | 49 | 49 | 46 | 46 | 43 | 46 | 50 | 50 | 49 | 49 | 46 | 46 | 49 | 50 | 47 | 47 | 49 | 50 | 47 | 50 | 100 | 47 | 48 | 47 | 45 | 43 | 43 | 43 | 43 | 43 | 44 | 44 | 42 | 42 | 39 | 42 | 41 | 43 | 42 | 42 |
| 23 | 49 | 48 | 49 | 51 | 51 | 47 | 50 | 50 | 49 | 49 | 49 | 48 | 47 | 50 | 50 | 48 | 49 | 50 | 51 | 50 | 49 | 47 | 100 | 50 | 49 | 49 | 44 | 44 | 43 | 43 | 44 | 45 | 44 | 42 | 42 | 40 | 42 | 41 | 43 | 43 | 42 |
| 24 | 51 | 52 | 51 | 52 | 51 | 47 | 51 | 52 | 50 | 50 | 53 | 52 | 49 | 54 | 54 | 53 | 53 | 54 | 54 | 50 | 52 | 48 | 51 | 100 | 48 | 47 | 45 | 44 | 44 | 44 | 44 | 45 | 45 | 42 | 43 | 39 | 43 | 42 | 44 | 43 | 43 |
| 25 | 48 | 48 | 49 | 46 | 46 | 44 | 47 | 49 | 48 | 47 | 46 | 44 | 44 | 48 | 48 | 48 | 48 | 48 | 49 | 50 | 49 | 47 | 49 | 48 | 100 | 72 | 45 | 45 | 44 | 44 | 44 | 45 | 44 | 43 | 42 | 35 | 43 | 41 | 42 | 43 | 41 |
| 26 | 47 | 44 | 48 | 46 | 47 | 44 | 46 | 48 | 46 | 47 | 46 | 45 | 44 | 47 | 47 | 46 | 46 | 47 | 47 | 50 | 48 | 45 | 49 | 47 | 73 | 100 | 43 | 42 | 43 | 42 | 43 | 43 | 43 | 41 | 41 | 42 | 41 | 41 | 42 | 41 | 42 |
| 27 | 45 | 42 | 44 | 43 | 43 | 38 | 43 | 44 | 43 | 42 | 43 | 40 | 42 | 45 | 45 | 44 | 43 | 44 | 44 | 44 | 44 | 43 | 43 | 44 | 45 | 43 | 100 | 53 | 54 | 48 | 49 | 50 | 49 | 46 | 42 | 37 | 42 | 41 | 44 | 44 | 43 |
| 28 | 45 | 44 | 44 | 42 | 43 | 39 | 43 | 44 | 43 | 43 | 43 | 40 | 42 | 44 | 45 | 44 | 42 | 45 | 44 | 44 | 44 | 43 | 44 | 44 | 45 | 42 | 53 | 100 | 53 | 49 | 49 | 51 | 50 | 48 | 42 | 37 | 42 | 40 | 43 | 42 | 41 |
| 29 | 44 | 46 | 43 | 43 | 43 | 39 | 42 | 43 | 42 | 43 | 43 | 40 | 41 | 44 | 43 | 44 | 42 | 44 | 44 | 43 | 44 | 43 | 43 | 44 | 44 | 42 | 54 | 53 | 100 | 48 | 48 | 50 | 49 | 45 | 41 | 36 | 42 | 40 | 42 | 43 | 42 |
| 30 | 44 | 44 | 44 | 43 | 43 | 41 | 43 | 44 | 43 | 42 | 44 | 41 | 43 | 44 | 44 | 44 | 43 | 44 | 44 | 44 | 45 | 43 | 44 | 44 | 43 | 42 | 49 | 49 | 48 | 100 | 60 | 50 | 48 | 46 | 41 | 38 | 42 | 41 | 43 | 43 | 43 |
| 31 | 44 | 46 | 44 | 43 | 43 | 40 | 43 | 43 | 42 | 43 | 43 | 40 | 42 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 43 | 44 | 44 | 43 | 49 | 49 | 48 | 60 | 100 | 51 | 48 | 46 | 41 | 36 | 42 | 40 | 44 | 42 | 43 |
| 32 | 45 | 45 | 45 | 44 | 44 | 41 | 44 | 44 | 44 | 43 | 44 | 42 | 43 | 45 | 45 | 44 | 43 | 45 | 45 | 44 | 45 | 44 | 45 | 45 | 45 | 43 | 50 | 51 | 50 | 50 | 51 | 100 | 51 | 48 | 42 | 38 | 42 | 41 | 43 | 43 | 43 |
| 33 | 44 | 46 | 44 | 44 | 44 | 42 | 43 | 45 | 44 | 43 | 44 | 41 | 42 | 44 | 44 | 45 | 42 | 45 | 45 | 44 | 45 | 44 | 44 | 45 | 45 | 43 | 49 | 50 | 49 | 48 | 48 | 51 | 100 | 48 | 42 | 42 | 43 | 42 | 43 | 43 | 43 |
| 34 | 43 | 42 | 43 | 42 | 42 | 40 | 42 | 41 | 41 | 41 | 41 | 39 | 42 | 42 | 42 | 44 | 41 | 43 | 42 | 42 | 43 | 42 | 42 | 42 | 43 | 41 | 46 | 47 | 46 | 46 | 46 | 48 | 48 | 100 | 40 | 39 | 42 | 40 | 43 | 42 | 42 |
| 35 | 43 | 45 | 42 | 42 | 42 | 40 | 42 | 43 | 42 | 42 | 41 | 40 | 40 | 43 | 43 | 42 | 41 | 42 | 43 | 42 | 43 | 42 | 42 | 43 | 42 | 41 | 42 | 42 | 41 | 41 | 41 | 42 | 42 | 40 | 100 | 64 | 70 | 52 | 46 | 47 | 47 |
| 36 | 40 | 32 | 42 | 41 | 43 | 41 | 40 | 41 | 40 | 47 | 39 | 37 | 37 | 40 | 37 | 37 | 37 | 38 | 41 | 41 | 41 | 39 | 41 | 39 | 34 | 41 | 38 | 38 | 36 | 38 | 35 | 38 | 40 | 39 | 64 | 100 | 59 | 44 | 46 | 42 | 46 |
| 37 | 43 | 44 | 44 | 42 | 42 | 42 | 42 | 43 | 42 | 42 | 42 | 40 | 41 | 43 | 43 | 42 | 42 | 43 | 43 | 43 | 43 | 42 | 42 | 43 | 43 | 41 | 42 | 42 | 42 | 42 | 42 | 42 | 43 | 42 | 70 | 59 | 100 | 52 | 47 | 49 | 47 |
| 38 | 42 | 40 | 42 | 41 | 41 | 40 | 41 | 42 | 41 | 40 | 41 | 40 | 40 | 43 | 42 | 41 | 39 | 42 | 41 | 42 | 42 | 41 | 42 | 42 | 42 | 41 | 41 | 40 | 40 | 41 | 40 | 41 | 42 | 40 | 51 | 44 | 52 | 100 | 46 | 47 | 45 |
| 39 | 44 | 46 | 44 | 42 | 43 | 41 | 42 | 43 | 43 | 42 | 41 | 43 | 41 | 46 | 45 | 45 | 42 | 45 | 44 | 43 | 44 | 44 | 43 | 44 | 43 | 42 | 44 | 43 | 42 | 43 | 44 | 43 | 43 | 43 | 46 | 46 | 47 | 46 | 100 | 56 | 57 |
| 40 | 44 | 41 | 44 | 42 | 43 | 41 | 43 | 43 | 42 | 42 | 43 | 42 | 41 | 45 | 44 | 43 | 42 | 44 | 44 | 44 | 45 | 42 | 43 | 44 | 43 | 41 | 44 | 42 | 43 | 43 | 42 | 43 | 43 | 42 | 47 | 43 | 49 | 47 | 56 | 100 | 59 |
| 41 | 44 | 44 | 45 | 42 | 42 | 40 | 42 | 43 | 42 | 42 | 41 | 42 | 40 | 44 | 44 | 43 | 42 | 43 | 44 | 43 | 43 | 42 | 42 | 43 | 41 | 42 | 43 | 41 | 42 | 43 | 42 | 43 | 43 | 42 | 47 | 49 | 47 | 45 | 57 | 58 | 100 |

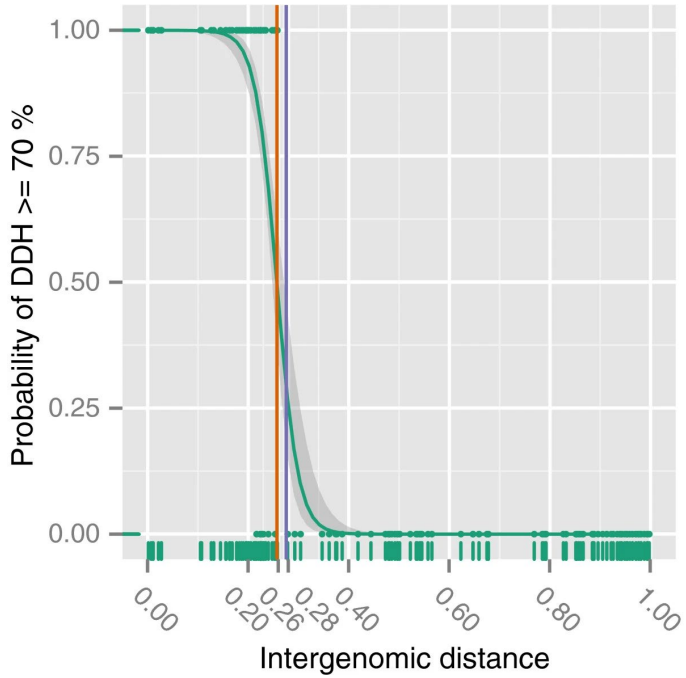
(Handley et al., 2017, ISME J)

- Pairwise AAI comparisons between genomic bins from the Gulf of Mexico seafloor
- All unique species (i.e. <95-96% AAI)
- Figures shows clusters of similar genomes
- Red = more similar
- Blue = dissimilar

# Discriminate species

**Alternative to gANI or AAI:**

- GGDC (Genome Blast Distance Phylogeny) https://ggdc.dsmz.de/
  - Makes use of generalised linear models

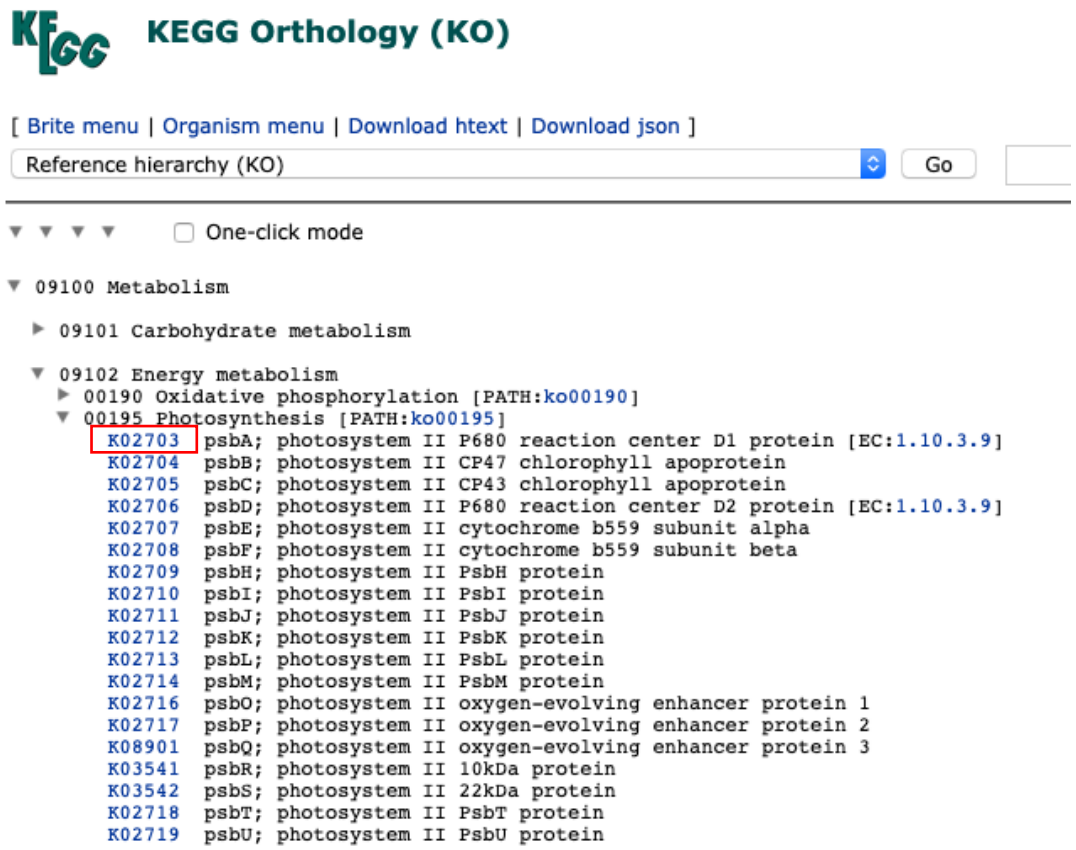


# Phylogenetic trait distributions

- Interactive phylogenetic and trait based tree
- Annotree (http://annotree.uwaterloo.ca/annotree/)
- Trait searches by:
  - Taxonomic hierarchy
  - KEGG (KO number)
  - Pfam
  - TIGRFAM

# Phylogenetic trait distributions

## Get KEGG KO number from the KEGG website or your annotations

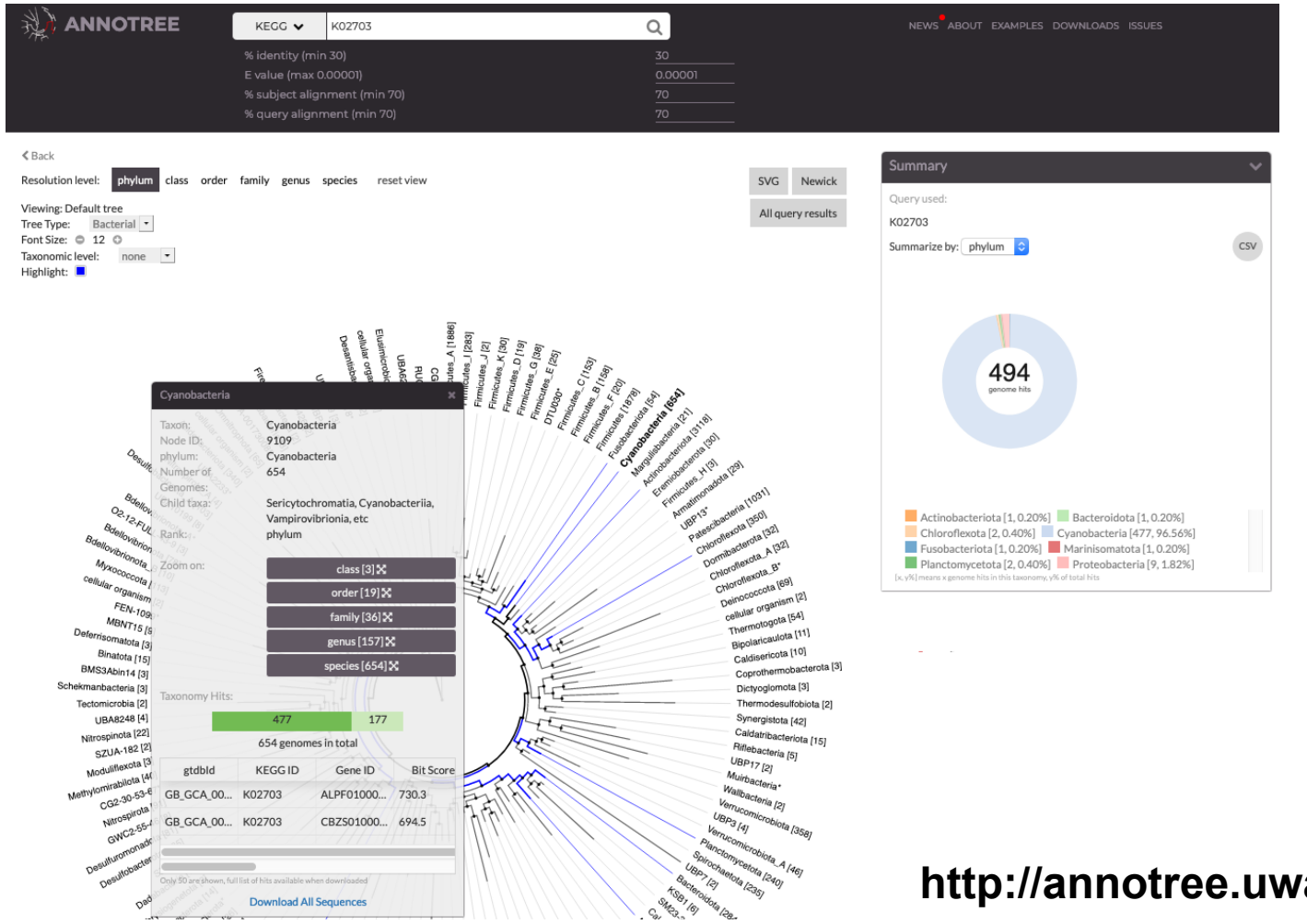



K02703

https://www.genome.jp/kegg-bin/get_htext#C17

# Phylogenetic trait distributions

## Add to ANNOTREE search box and select hierarchy



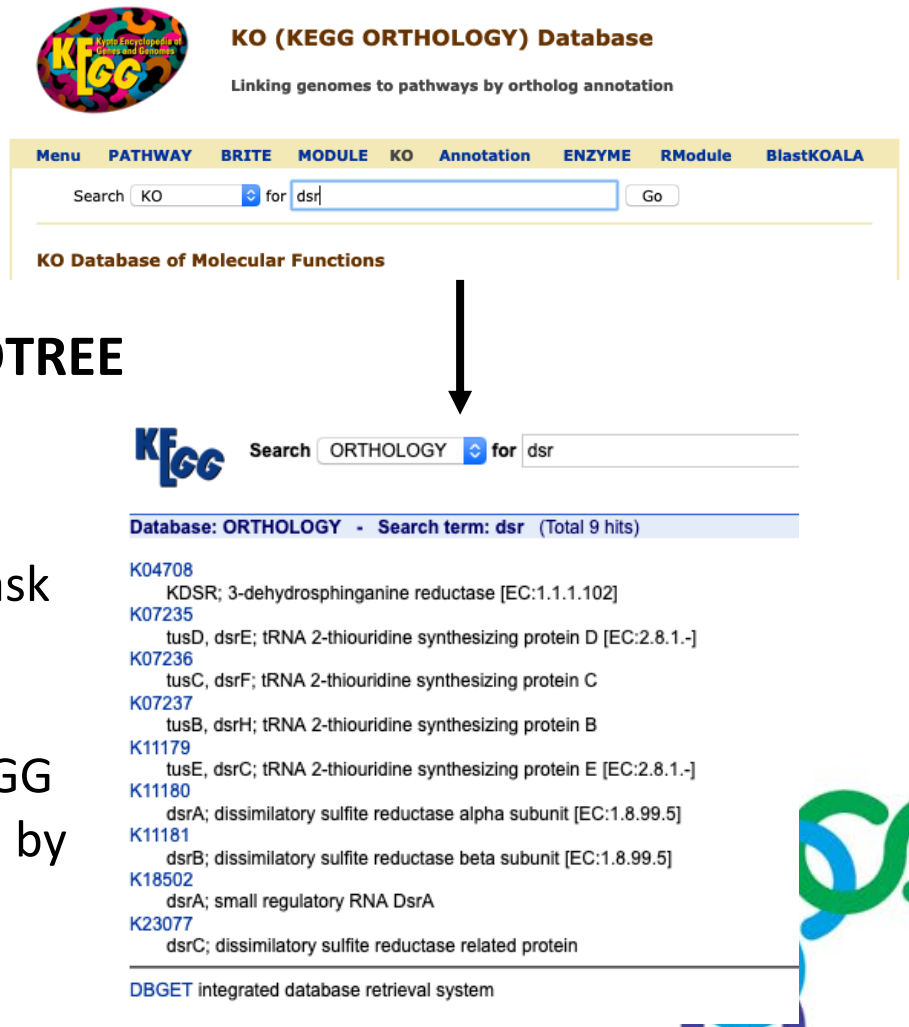

http://annotree.uwaterloo.ca/annotree/

# Task: Genome taxonomic classification

## Go to Github MGSS webpage

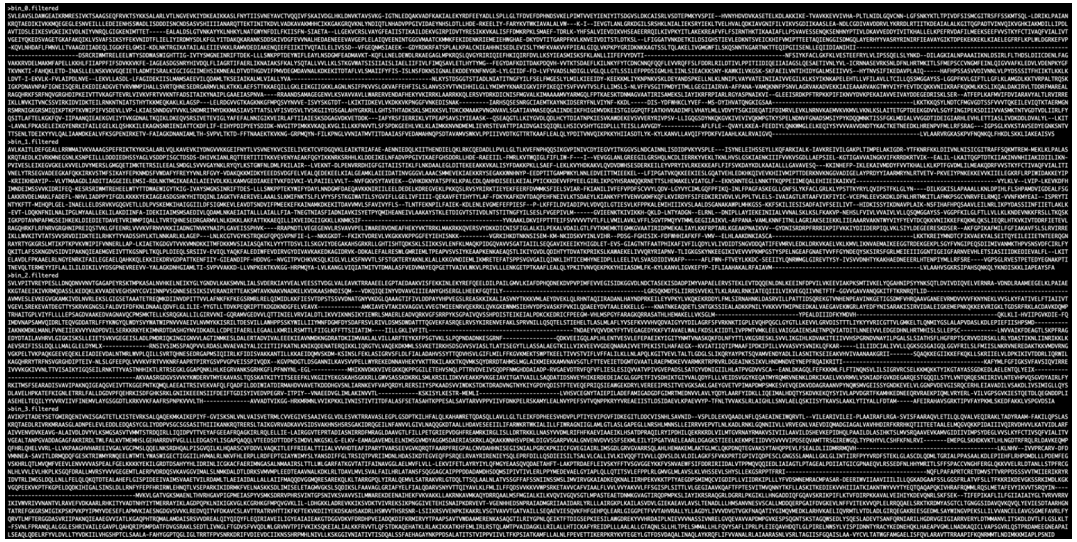
**Tasks:**

- **Taxonomic classification using GTDB-Tk**
- **View phylogenetic trait distribution using ANNOTREE**
  - Use ANNOTREE to explore the phylogenetic distribution of functions
  - Try using attribute annotations for your group task
  - You can use your KEGG Orthology (KO) numbers
  - **Note:** You can also get KO numbers from the KEGG website (https://www.genome.jp/kegg/ko.html) by searching for gene names



# Phylogenetics

GTDB-Tk output file: Multiple Sequence Alignment (MSA) of 120 marker genes (protein sequences)



Bin_0

Bin_1

Bin_2

Bin_3


Phylogenetic tree software takes the MSA as input and builds an evolutionary tree

# Phylogenetics

THE ESTIMATION OF LINKAGE FROM THE OFFSPRING OF SELFED HETEROZYGOTES.

By R. A. FISHER, Sc.D., and BHAI BALMUKAND, M.A., M.Sc. (*Punjab University*).

(With Two Text-figures.)

https://link.springer.com/article/10.1007/BF02983317

Ronald Fisher

In 1928, Fisher was the first to use equations to calculate the distribution of allele frequencies and the estimation of genetic linkage by **maximum likelihood methods** among populations.



In the absence of linkage their deviations will be independent, but if linkage is present the mean value of $pq$ may be found to be

$$-3n\frac{1-4x}{3},$$

or, the correlation between $p$ and $q$ is

$$\rho = -\frac{1-4x}{3}.$$

The simultaneous deviation of $p$ and $q$ from zero will therefore be measured by

$$Q^2 = \frac{1}{3n}\left\{\frac{1}{1-\rho^2}(p^2 - 2\rho pq + q^2)\right\}$$
$$= \frac{3}{8(1-x)(1+2x)n}\left\{p^2 + q^2 + \frac{2}{3}(1-4x)\,pq\right\}.$$

This expression, which of course depends upon $x$, is a quadratic function of the frequencies; in this it resembles $\chi^2$, and on comparing term by term the two expressions it appears that

$$\chi^2 = Q^2 + \frac{1}{I}\left\{\frac{a}{2+x} - \frac{b+c}{1-x} + \frac{d}{x}\right\}^2,$$

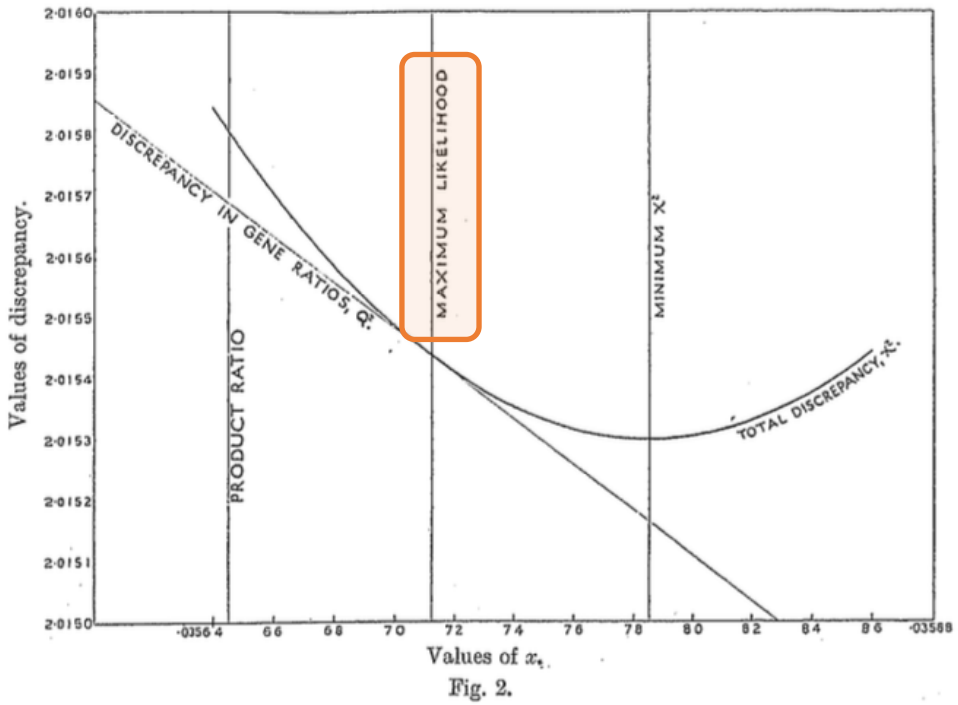



Fig. 2.

**Maximum Likelihood Phylogenetic Inference** estimates the most likely phylogenetic tree that represents the branching patterns and relationships among a set of biological sequences, usually DNA or protein sequences.

# Phylogenetics

Phylogenetic inference using maximum likelihood:

FastTree 2

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

## FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments

Morgan N. Price, Paramvir S. Dehal, Adam P. Arkin

Published: March 10, 2010 • https://doi.org/10.1371/journal.pone.0009490

| | |
|---|---|
| 4,589 Save | 9,606 Citation |
| 70,863 View | 4 Share |

MOLECULAR BIOLOGY AND EVOLUTION

JOURNAL ARTICLE

## IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, Robert Lanfear Author Notes

*Molecular Biology and Evolution*, Volume 37, Issue 5, May 2020, Pages 1530–1534, https://doi.org/10.1093/molbev/msaa015

**Published:** 03 February 2020

| CITATIONS | VIEWS | ALTMETRIC |
|---|---|---|
| 4.5k | 51,184 | 138 |

# Phylogenetics

**10 bins / MAGs**

CPU: 02
RAM: 64MB
Time: 4 seconds

**Pros:** Speed, memory efficiency, easy to use

**Cons:** It may not always produce the most accurate phylogenetic trees compared to more computationally intensive methods.

Limited customization and flexibility

**10 bins / MAGs**

CPU: 22
RAM: 80MB
Time: 03:56:34

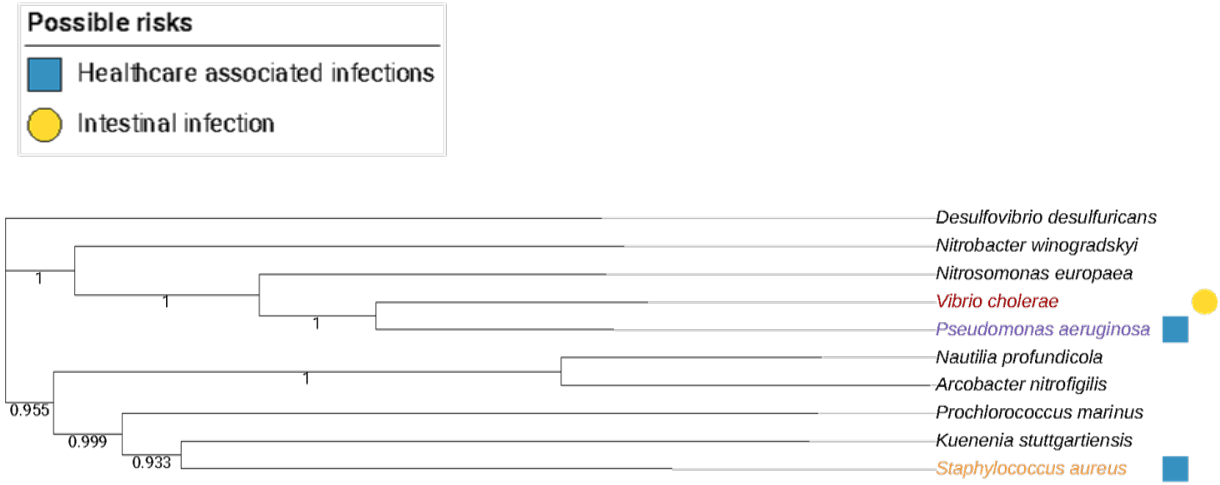
**Pros:** Provides better accuracy due to its advanced optimization methods (**Maximum Likelihood Optimization)**

More customizations and flexibilities

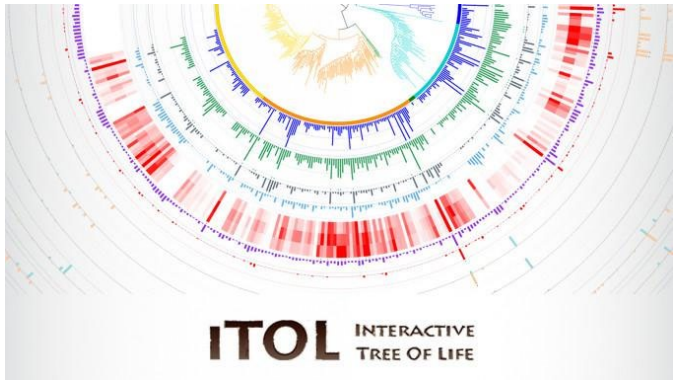
**Cons:** Require more computational resources and time

# Phylogenetics

Tree file: `(((bin_3.filtered:0.4163372898,(bin_5.filtered:0.2927075220,bin_9.filtered:0.3087018139):0.1195543871)`







# Viruses

# Viral assembly alternatives

- **metaSPAdes**
  - We ran this in day 1
- **metaviralSPAdes**
  - DNA viruses only
  - Based on variation in viral and bacterial sequence coverage
- **rnaSPAdes**
  - Transcriptome assembly
  - Not ideal for RNA viruses
- **rnaviralSPAdes**
  - *de novo* assembler tailored for RNA viral datasets
  - transcriptome, metatranscriptome, metavirome

# Viruses

Identify viral sequences

- VirSorter2
  - Database-based multi-classifier
  - https://github.com/jiarong/VirSorter2
- VIBRANT
  - Neural network/machine learning
  - https://github.com/AnantharamanLab/VIBRANT
- DeepVirFinder
  - Machine learning approach based on k-mer frequencies
  - https://github.com/jessieren/DeepVirFinder

**VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses**

Jiarong Guo, Ben Bolduc, Ahmed A. Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O. Delmont, Akbar Adjie Pratama, M. Consuelo Gazitúa, Dean Vik, Matthew B. Sullivan & Simon Roux

*Microbiome* **9**, Article number: 37 (2021) | Cite this article

**16k** Accesses | **112** Citations | **73** Altmetric | Metrics

Methodology | Open Access | Published: 10 June 2020

**VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences**

Kristopher Kieft, Zhichao Zhou & Karthik Anantharaman

*Microbiome* **8**, Article number: 90 (2020) | Cite this article

# Viruses

Identify viral sequences

## Caveats

- Filtering out host reads will prevent identification of endogenous viral elements in the host genome
- Results might include cryptic phage

# Viruses

## Inspect viral sequence quality

## CheckV

- Removes host regions based on hits to multiple databases and GC content
- Estimates completeness and quality by comparing against known and complete virus sequences
- Identifies closed genomes (direct terminal repeats)

Article | Open Access | Published: 21 December 2020

**CheckV assesses the quality and completeness of metagenome-assembled viral genomes**

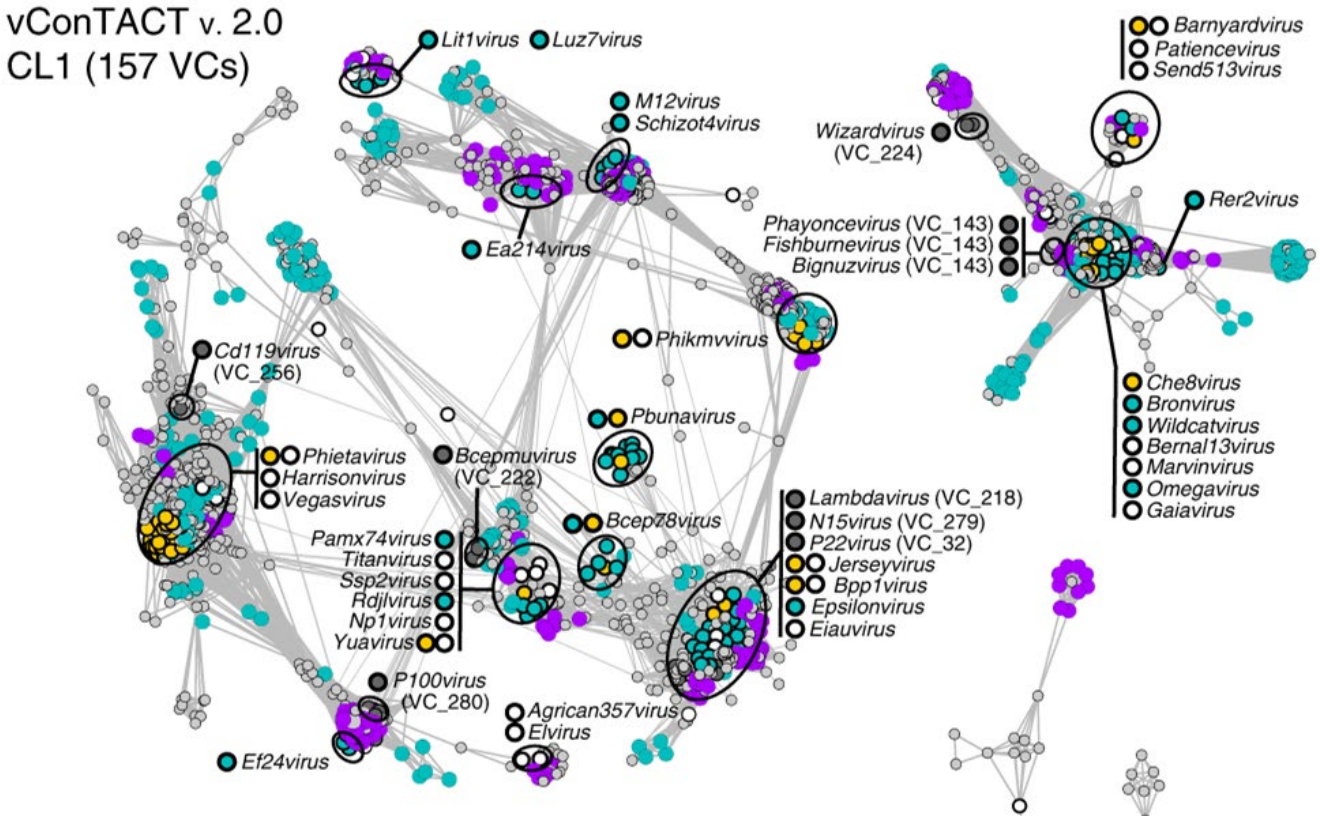
Stephen Nayfach, Antonio Pedro Camargo, Frederik Schulz, Emiley Eloe-Fadrosh, Simon Roux & Nikos C. Kyrpides

*Nature Biotechnology* **39**, 578–585 (2021) | Cite this article

# Viral taxonomic prediction

**vConTACT v.2.0** (https://bitbucket.org/MAVERICLab/vcontact2/src/master/)

- Clustering-based: guilt-by-contig-association taxonomic prediction
- Reference database (Viral RefSeq) + identified viral contigs



# Viruses - workshop workflow

- Viral identification: **VirSorter2**
- (OPTIONAL: Dereplication across multiple assemblies)
- Viral QC: **CheckV**
- (OPTIONAL: Viral taxonomy and gene-sharing network: **vConTACT2** and **Cytoscape**)
- Viral gene prediction and annotation: **DRAM-v**

# Task: QC and taxonomic classification

Go to Github MGSS webpage

Tasks:

- ✓ Identifying viral contigs using VirSorter2
- ✓ QC of viral contigs using CheckV
- Examine viral output files from VirSorter2 and CheckV
- Taxonomic classification of viruses using vContact2

# Gene prediction

# Gene prediction

- Genome annotation is the process of attaching biological information to sequences
- It consists of three main steps:
  - Gene prediction
  - Prediction of protein sequences
  - Functional annotation: Attaching biological information to these elements

# Gene prediction

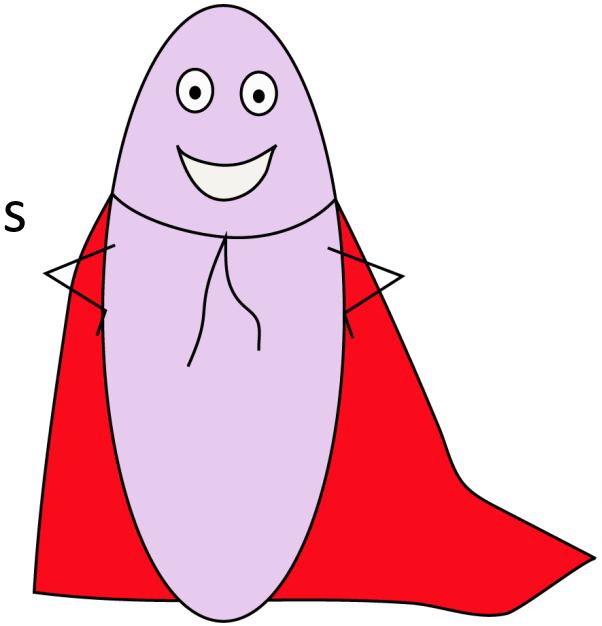
Aim:

- To identify regions of genomic DNA that encode putative genes present in high quality genomes

About 1/1000th of a human genome in size, but with only 1/10$^{th}$ less coding DNA sequence
 100 x more power packed!!!

Prokaryote genomes:

- High gene density
- Genes = continuous stretches of coding DNA
- Absence of introns in the protein coding regions



# Gene prediction

**Gene finding algorithms for prokaryotes**

- Homology:
  - Search by sequence similarity to homologous sequences
  - Based on the assumption that functional regions are more conserved evolutionarily than non-functional regions

- *Ab initio*:
  - Search by content: find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA
  - Search by signals/sites, e.g. promoters, start and stop codons

# Gene prediction

**Homology: Sequence similarity searches**

- Finding similarity in gene sequences between expressed sequence tags (ESTs), proteins, or other genomes to the input genome
- Local alignment:
  - BLAST family tools: https://blast.ncbi.nlm.nih.gov/Blast.cgi
  - Global alignment
  - GeneWise: https://www.ebi.ac.uk/Tools/psa/genewise/

# Gene prediction

***Ab initio* search by content:** Markov Model Based Algorithms

- Most widespread algorithms for gene finding in prokaryotes are based on Markov Models
- Aim is to capture compositional differences among coding regions, “shadow” coding regions (coding on the opposite DNA strand) and non-coding DNA

# Gene prediction

**Markov Model Based Algorithms:** Glimmer

- http://ccb.jhu.edu/software/glimmer/index.shtml
- Interpolated Markov model (IMM) DNA discriminator
- Log-likelihood that a given interval on a DNA sequence was generated by a model of coding versus non-coding DNA

# Gene prediction

**Markov Model Based Algorithms:** GeneMark/GeneMarkHMM/MetaGeneMark

- http://exon.gatech.edu/GeneMark/
- GeneMark is a family of gene prediction tools
- Genomic sequences can be analysed either by the self-training program GeneMarkS (sequences >50 kb) or using Heuristic Models by GeneMarkHMM
- Pre-trained model parameters are available for many species
- Metagenomics sequences can be analysed with MetaGeneMark

# Gene prediction

**Prodigal** (PROkaryotic Dynamic Programming Genefinding ALgorithm)

- https://github.com/hyattpd/Prodigal
- Based on Dynamic Programming, not Markov Models
- Gene-finding algorithm for prokaryote genomes developed to predict translation initiation sites more accurately.
- High accuracy in high GC content genomes
- Tends to predict longer genes rather than more genes (minimising number of false positives)

# Gene prediction

**Prodigal for metagenomics:**

- Use anon (meta) mode with metagenomic data (or short sequence data)
  - Copes with diverse genomes
  - Unlike normal mode, it does not attempt to study the input sequence, and predict based on these assumptions
  - Uses pre-calculated training files, and predicts genes based on the best results
- Alternatively, use normal mode on each individual genome bin

# Gene prediction

**Prodigal for metagenomics:**

- Caveat: unusual genetic codes
  - First uses genetic code 11 (stop codons TAA, TGA, TAG)
  - If genes are too short, uses alternative code 4 (TGA not a stop codon)
  - Will not try code 25, but will issue warning if genes are short
  - Must manually select code 25

# Gene prediction

**Prodigal for metagenomics:**

- Important note:
  - Prodigal predicts coding DNA sequence ONLY
  - Provides nucleic acid (.fna) and amino acid (.faa) files
  - DOES NOT identify other features (e.g. rRNA, tRNA)
  - Combine with other prediction tools

# Gene prediction

Predicting RNA features and non-coding regions:

- MeTaxa2: predicts ribosomal RNA sequences in a genome
- Aragorn: predicts tRNA and tmRNA sequences

# Gene prediction

Predicting protein coding sequences in unassembled (short) reads

- FragGeneScan:
    - Tuning parameters for short sequences (and hence incomplete genes)
    - Model sequence error

# Task: Gene prediction

Go to Github MGSS webpage

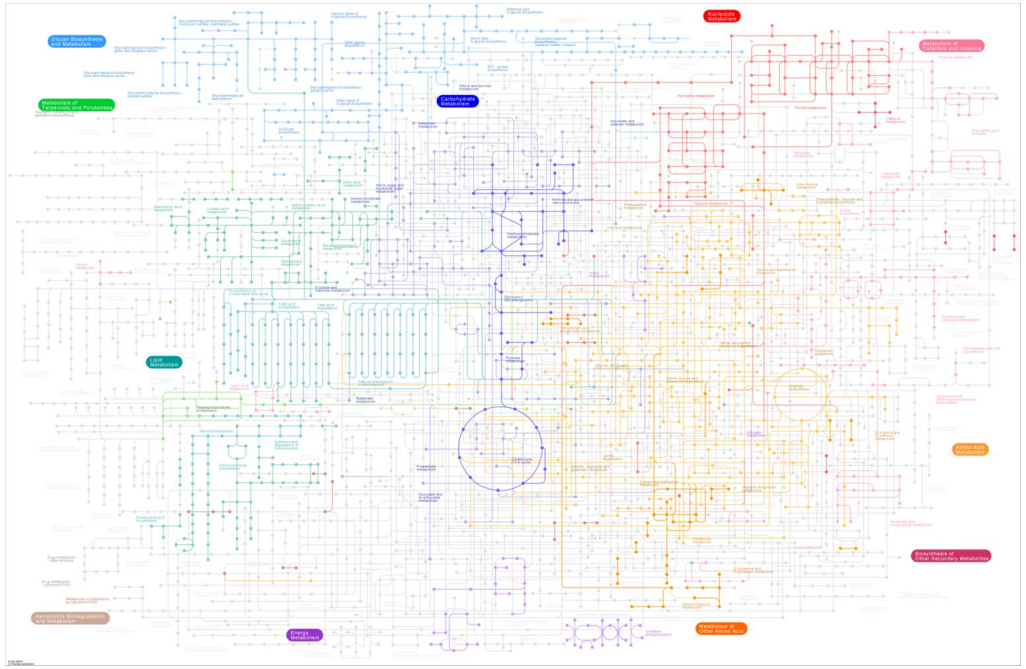
**Tasks:**

- **Predict open reading frames and protein sequences**

# Gene annotation (part 1)

# Gene annotation

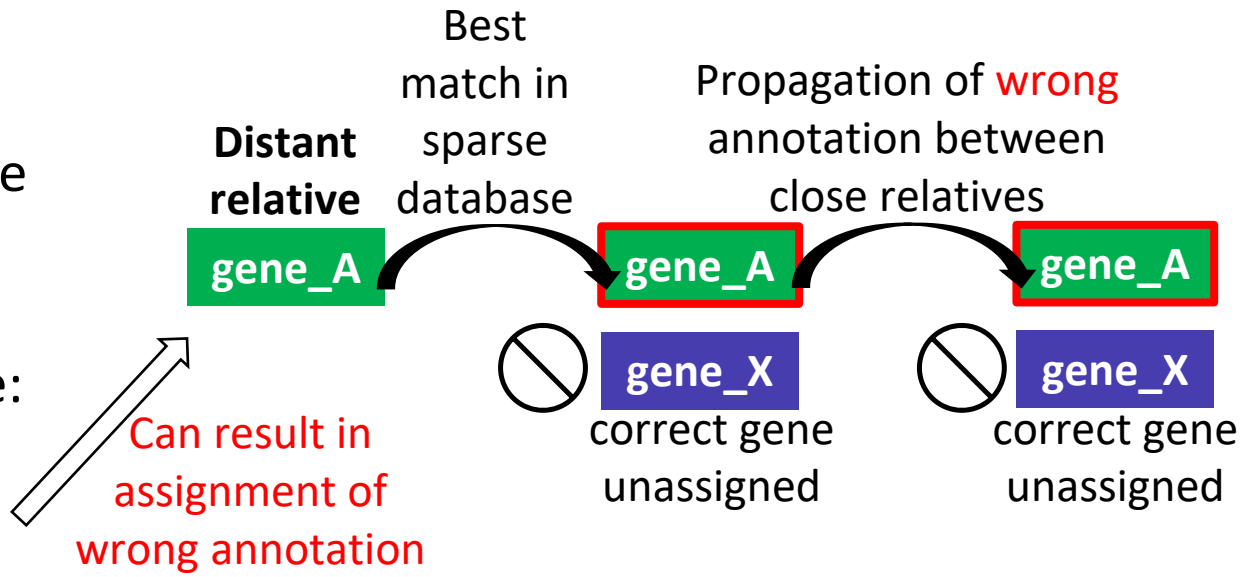
- Genome annotation attempts to predict gene function
- Predicted genes or protein sequences are compared against a curated set of reference sequences for which function is known, or is strongly suspected



# Gene annotation

Caveat:

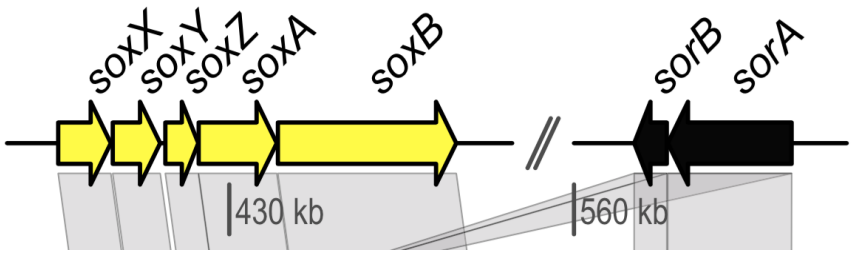
- Annotations are dependent on the reference database
- Environmental genomes can have:
  - Genes with distant homology matches to unrelated taxa
  - Large numbers of "hypothetical" gene annotations (= genes of unknown function)



# Gene annotation

Caveat:

- Annotations are “advice”
- Automated annotations often need to be manually curated
- Interrogate if: expected functional gene is missing from annotations
- Gene synteny is a useful for missing gene discovery, e.g.:
  - check genes co-located in operons for putative functions
  - check for operon truncation (due to contig break)




(Handley et al., 2014, Environmental Microbiology)

# Gene annotation

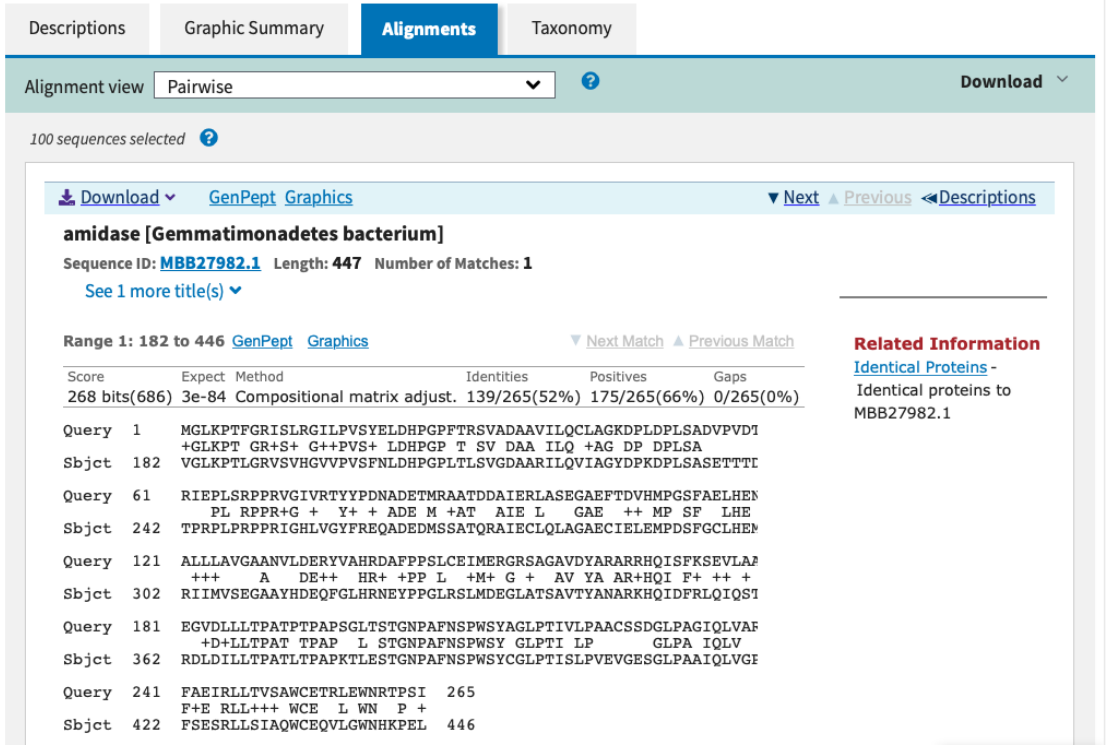
There are two main ways to perform gene annotation with protein sequences:

- BLAST-like gene annotation
- Domain annotation

# Gene annotation

## BLAST-like gene annotation

- Pairwise local alignment between the gene of interest (query sequence) and the sequences in the database (target sequence)
- Tools:
  - BLAST: web-based and stand alone (usually too slow for metagenomics)
  - USEARCH (64-bit): fast (subscription needed)
  - Diamond: fast



# Gene annotation

## HMM-profiling of domains:

- Considers the query sequences as a collection of independently functioning protein folding domains
- Uses database of Hidden Markov models built from a collection of proteins that share a common domain
- Profiles build from statistical map of the
  - amino acid transitions (from position to position),
  - variations (differences at a position),
  - insertions/deletions between positions
- Tools: HMMer software (http://hmmer.org/)

# Gene annotation

**Common functional databases**

- KEGG (Kyoto Encyclopedia of Genes and Genomes) (https://www.kegg.jp)

  Very popular, each entry is well annotated, and often linked into “Modules” or “Pathways” (Full access now requires a license fee)

- COGs (Clusters of Orthologous Groups of proteins) (https://www.ncbi.nlm.nih.gov/COG/)

  Classify proteins from completely sequenced genomes on the basis of the orthology concept

- PFAM (now integrated with InterPro: https://www.ebi.ac.uk/interpro/)

  Focused more on protein domains based on hidden Markov models

- TIGRfam (now integrated into NCBI: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/tigrfams/)

  Database of protein family definitions based on hidden Markov models

# Gene annotation

**Common functional databases** (continued)

- The PANTHER (**P**rotein **AN**alysis **TH**rough **E**volutionary **R**elationships) Classification System (http://pantherdb.org)

  Proteins are classified according to Family and subfamily, molecular function, biological process and pathway

- UniRef (UniProt Reference Clusters ) (https://www.uniprot.org/)

  Protein clustering at different levels (e.g. UniRef100, UniRef90, UniRef50)

- BioCyc Database Collection (https://biocyc.org)

  14735 Pathway/Genome Databases (PGDBs), plus software tools

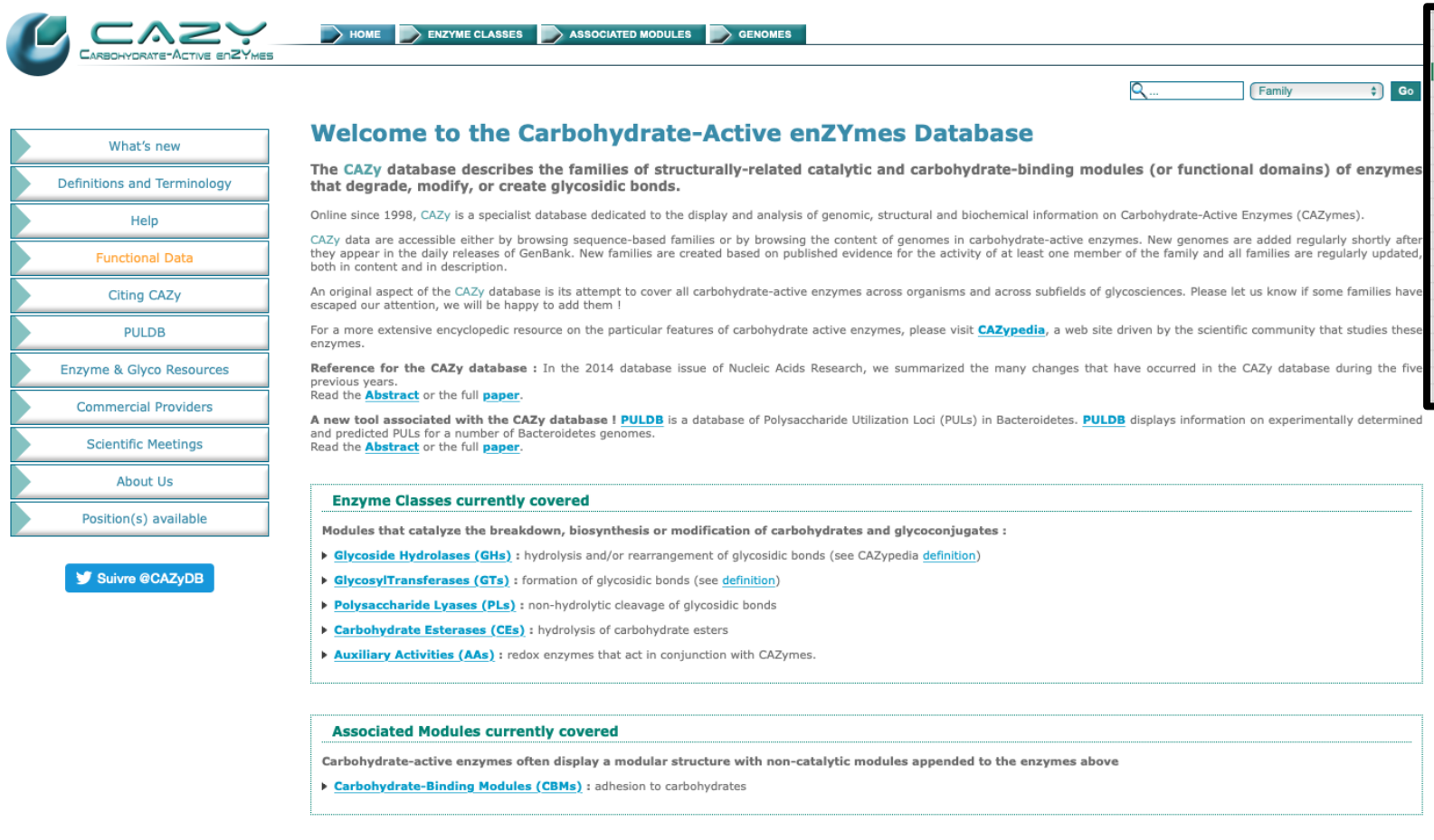
  Subscriptions are required to access most of BioCyc

- MetaCyc Metabolic Pathway Database (https://metacyc.org)

  2722 pathways from 3009 different organisms

# Gene annotation

Identification of Carbohydrate-Active enZYmes - **CAZY Database**



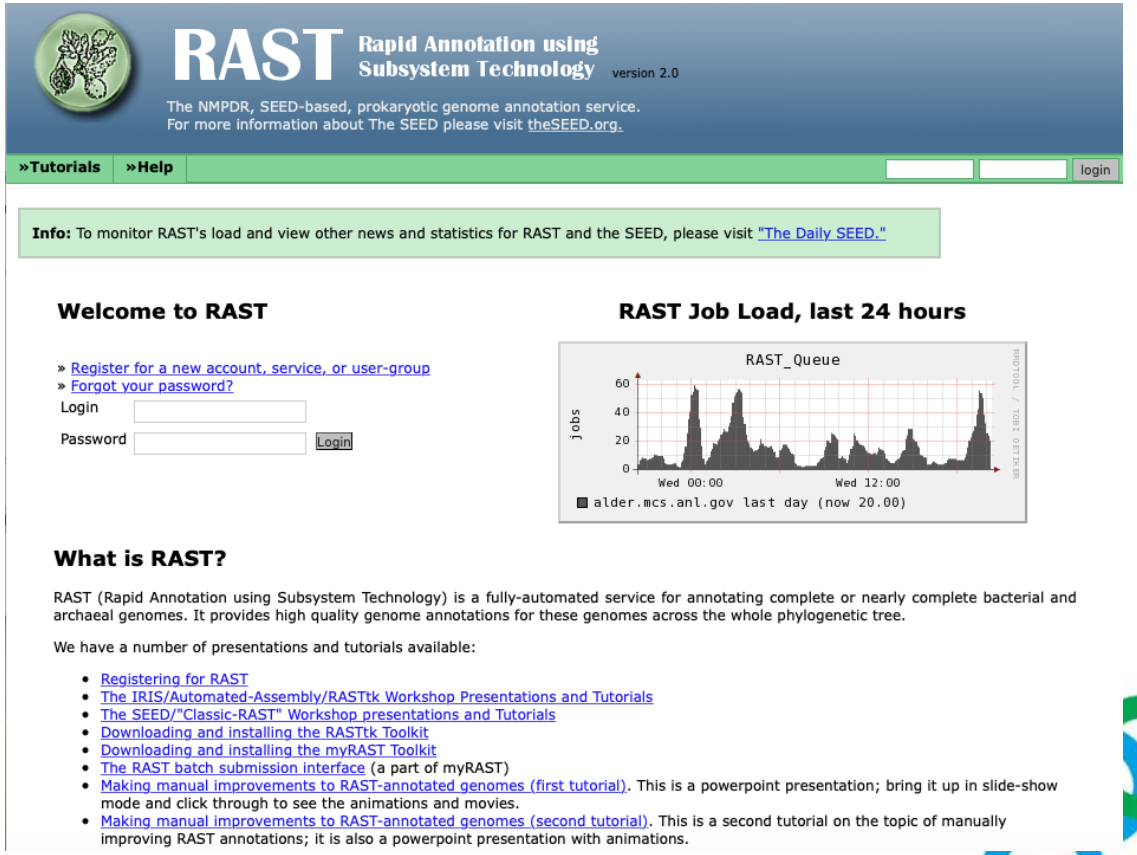

| Query | Subject | E-value | Subject start | Subject end | Query start | Query end | Covered fraction |
|---|---|---|---|---|---|---|---|
| fig\|6666666.197029.peg.1003 | GH109.hmm | 6.90E-12 | 1 | 117 | 2 | 112 | 0.920634921 |
| fig\|6666666.197029.peg.1015 | GT4.hmm | 3.00E-35 | 8 | 157 | 193 | 340 | 0.93125 |
| fig\|6666666.197029.peg.1034 | PL17.hmm | 0.00017 | 2 | 84 | 577 | 648 | 0.589928058 |
| fig\|6666666.197029.peg.1083 | GH38.hmm | 3.20E-28 | 25 | 182 | 33 | 183 | 0.583643123 |
| fig\|6666666.197029.peg.1124 | GH117.hmm | 2.60E-06 | 64 | 161 | 39 | 125 | 0.45971564 |
| fig\|6666666.197029.peg.1151 | CE14.hmm | 1.10E-16 | 1 | 124 | 9 | 127 | 0.991935484 |
| fig\|6666666.197029.peg.1183 | GH109.hmm | 2.70E-09 | 63 | 121 | 23 | 84 | 0.46031746 |
| fig\|6666666.197029.peg.1208 | GH109.hmm | 2.30E-12 | 3 | 112 | 6 | 106 | 0.865079365 |
| fig\|6666666.197029.peg.1232 | GH109.hmm | 1.70E-07 | 2 | 121 | 25 | 145 | 0.944444444 |
| fig\|6666666.197029.peg.1233 | GH74.hmm | 2.70E-11 | 30 | 117 | 164 | 254 | 0.373390558 |
| fig\|6666666.197029.peg.1233 | GH74.hmm | 4.70E-08 | 43 | 118 | 228 | 301 | 0.321888412 |
| fig\|6666666.197029.peg.1247 | PL12.hmm | 1.70E-27 | 1 | 138 | 388 | 516 | 0.992753623 |
| fig\|6666666.197029.peg.127 | GH109.hmm | 1.60E-12 | 2 | 122 | 5 | 119 | 0.952380952 |
| fig\|6666666.197029.peg.1289 | GH109.hmm | 1.60E-07 | 1 | 122 | 10 | 133 | 0.96031746 |
| fig\|6666666.197029.peg.1297 | GH109.hmm | 2.20E-08 | 4 | 121 | 5 | 111 | 0.928571429 |
| fig\|6666666.197029.peg.130 | GH32.hmm | 1.80E-08 | 43 | 171 | 93 | 232 | 0.436860068 |
| fig\|6666666.197029.peg.1325 | GH109.hmm | 1.20E-12 | 4 | 105 | 9 | 105 | 0.801587302 |
| fig\|6666666.197029.peg.1326 | GH109.hmm | 1.10E-09 | 4 | 118 | 10 | 115 | 0.904761905 |
| fig\|6666666.197029.peg.1327 | GH109.hmm | 2.30E-08 | 3 | 122 | 6 | 116 | 0.944444444 |
| fig\|6666666.197029.peg.1334 | CE10.hmm | 9.80E-16 | 99 | 316 | 3 | 221 | 0.636363636 |
| fig\|6666666.197029.peg.1340 | GH28.hmm | 7.60E-07 | 59 | 217 | 148 | 307 | 0.486153846 |
| fig\|6666666.197029.peg.1359 | CBM9.hmm | 1.60E-32 | 1 | 181 | 39 | 240 | 0.989010989 |

http://www.cazy.org

# Gene annotation

RAST Annotation Server (Rapid **Annotation** using Subsystem Technology):

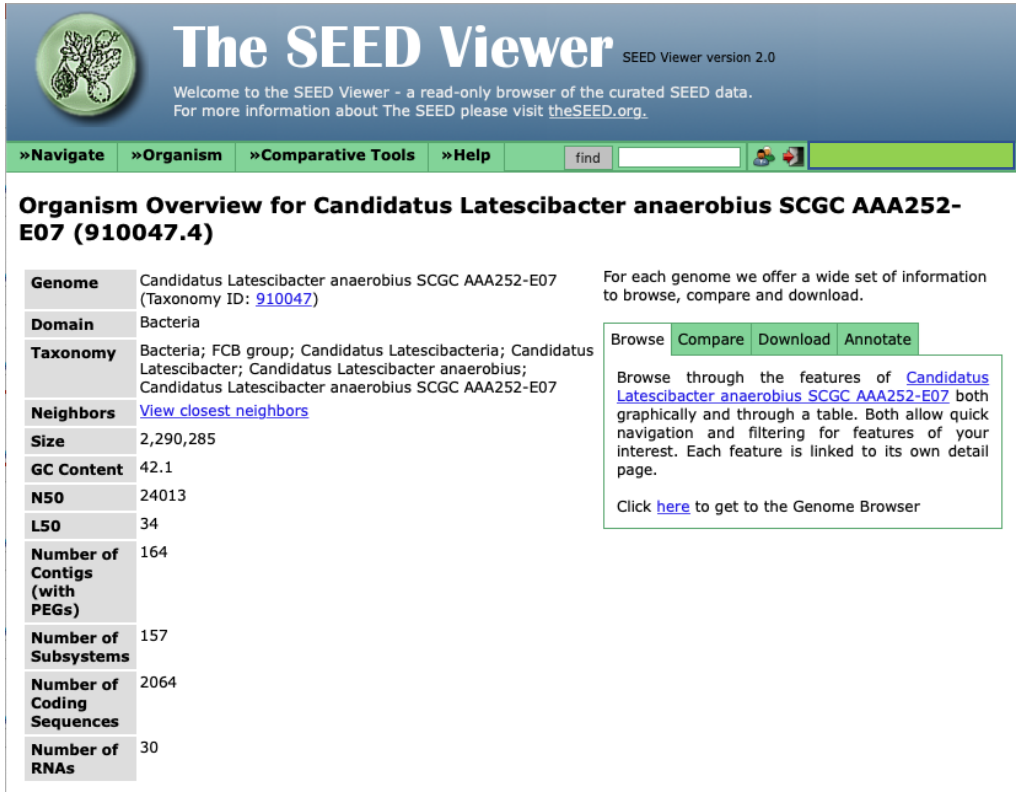
- Fast annotation (~1 genome/day)
- Can use for individual genome bins
- It works well for genomes similar to large groups of reference genomes
- As usual: requires manual curation after initial annotation

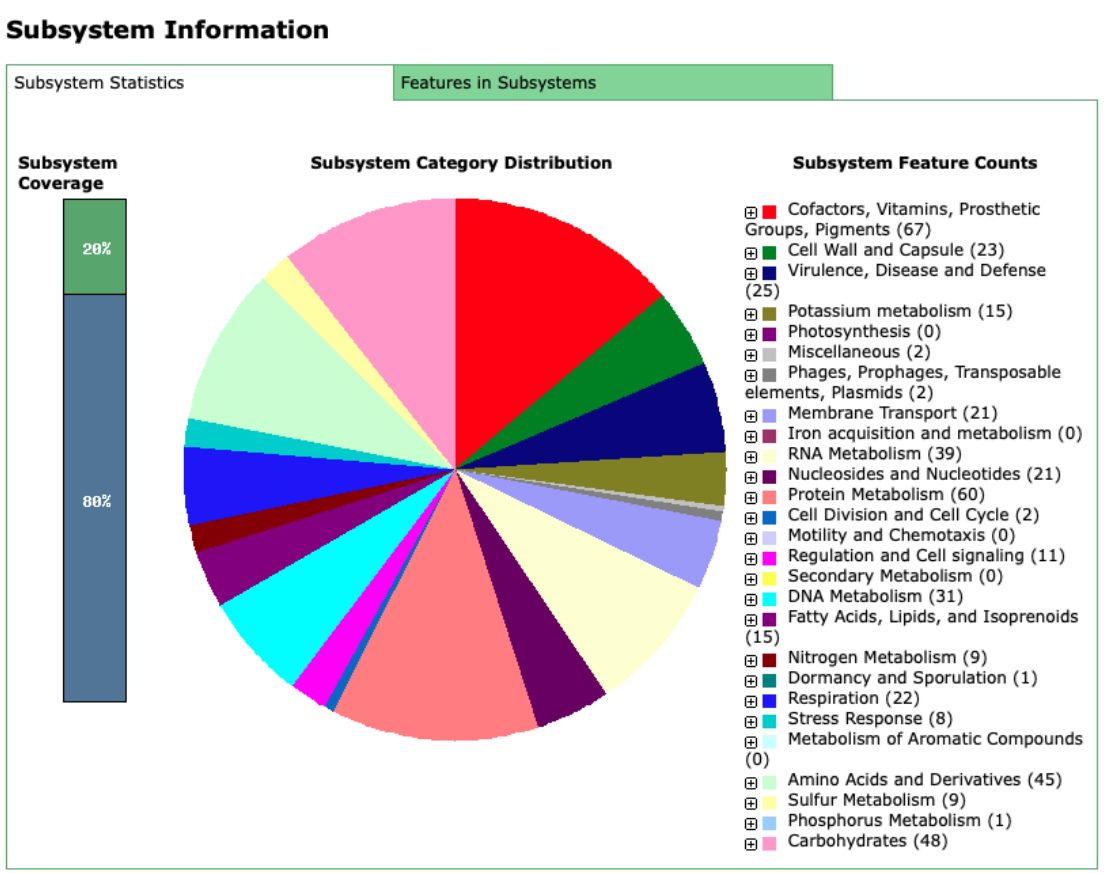



http://rast.theseed.org/FIG/rast.cgi

# Gene annotation

RAST Annotation Server (Rapid **Annotation** using Subsystem Technology)



# Gene annotation

**Distilling and Refining Annotations of Metabolism (DRAM;** Shaffer et al. 2020. Nucleic Acids Research 48(16)**)**

- Tool for gene prediction and gene annotation of MAGs (DRAM-v for viruses)
  - Functional annotation using KEGG (if provided), UniRef 90, MEROPS, Pfam, dbCAN and VOGDB. tRNAs and rRNAs also detected
  - Genome annotations to metabolic functions in three levels:

**1. RAW**
Each gene nucleotide and amino acid sequence with annotations

**2. DISTILLATE**
Taxonomy (GTDB-tk), quality statistics (checkM), and key metabolisms summarized by genome

**3. LIQUOR**
Genome metabolisms classified by key functional gene, with gene FASTAs output

- **Data compiler**: checkM and GTDB-tk taxonomy summary

# Task: Gene annotation

Go to Github MGSS webpage

**Tasks:**

- **Gene annotation using DIAMOND and HMMER3**

# Online resources and data analysis

# Gene annotation

**Some web-based annotation tools:**

- Web BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
- RAST/MG-RAST (Rapid Annotation using Subsystem Technology) Annotation Server
- KEGG Automatic annotation and KEGG mapping service
  - BLAST-Koala: BLAST search (https://www.kegg.jp/blastkoala/)
  - GHOST-Koala: GHOSTX search (https://www.kegg.jp/ghostkoala/)
  - KofamKOALA: HMM profile search (https://www.genome.jp/tools/kofamkoala/)
- IMG/M (The Integrated Microbial Genomes and Microbiomes) (https://img.jgi.doe.gov)

# Gene annotation

KEGG: https://www.genome.jp/kegg/pathway.html

**KEGG PATHWAY Database**

**Wiring diagrams of molecular interactions, reactions and relations**

Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

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 on 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**

KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

**Pathway Identifiers**

Each pathway map is identified by the combination of 2-4 letter prefix code and 5 digit number (see KEGG Identifier). The prefix has the following meaning:

| | |
|---|---|
| map | manually drawn reference pathway |
| ko | reference pathway highlighting KOs |
| ec | reference metabolic pathway highlighting EC numbers |
| rn | reference metabolic pathway highlighting reactions |
| <org> | organism-specific pathway generated by converting KOs to gene identifiers |

and the numbers starting with the following:

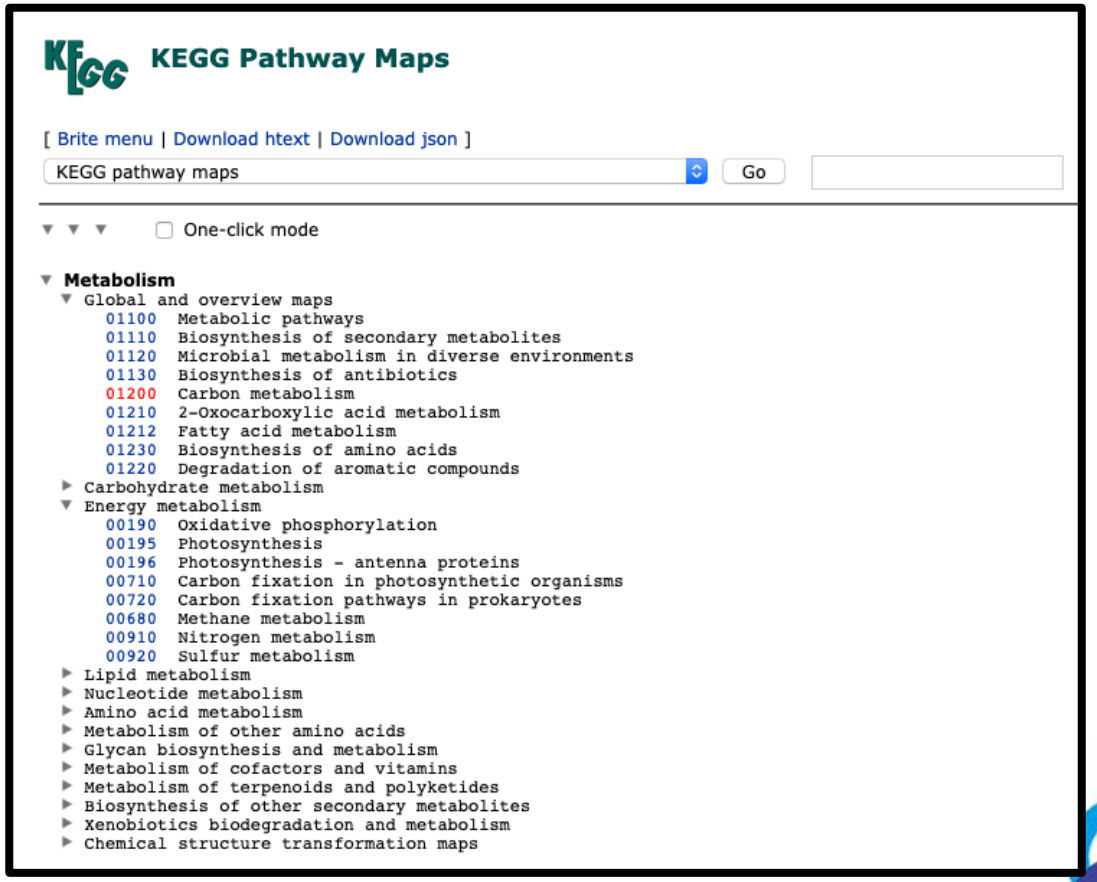
| | |
|---|---|
| 011 | global map (lines linked to KOs) |
| 012 | overview map (lines linked to KOs) |
| 010 | chemical structure map (no KO expansion) |
| 07 | drug structure map (no KO expansion) |
| other | regular map (boxes linked to KOs) |

are used for different types of maps.

**1. Metabolism**

**1.0 Global and overview maps**

01100 Metabolic pathways
01110 Biosynthesis of secondary metabolites
01120 Microbial metabolism in diverse environments
01130 Biosynthesis of antibiotics
01200 Carbon metabolism



**KEGG Pathway Maps**

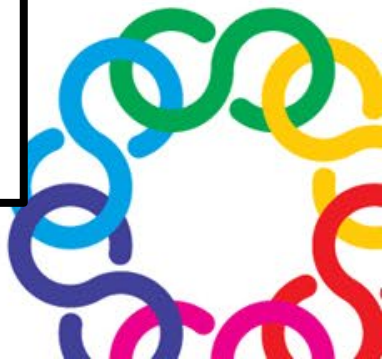
[ Brite menu | Download htext | Download json ]

KEGG pathway maps | Go

One-click mode

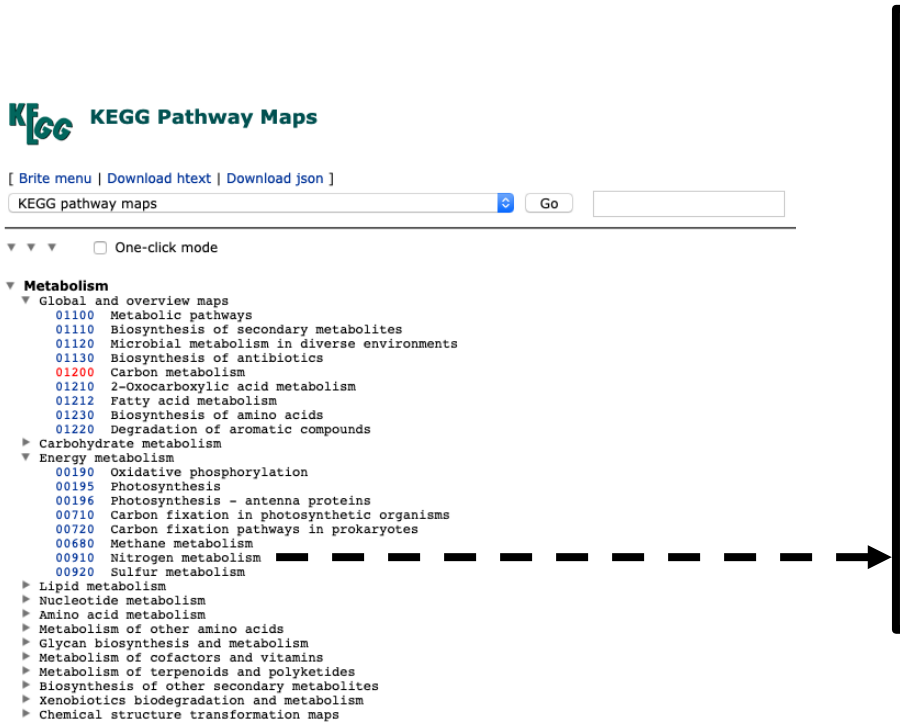
**Metabolism**

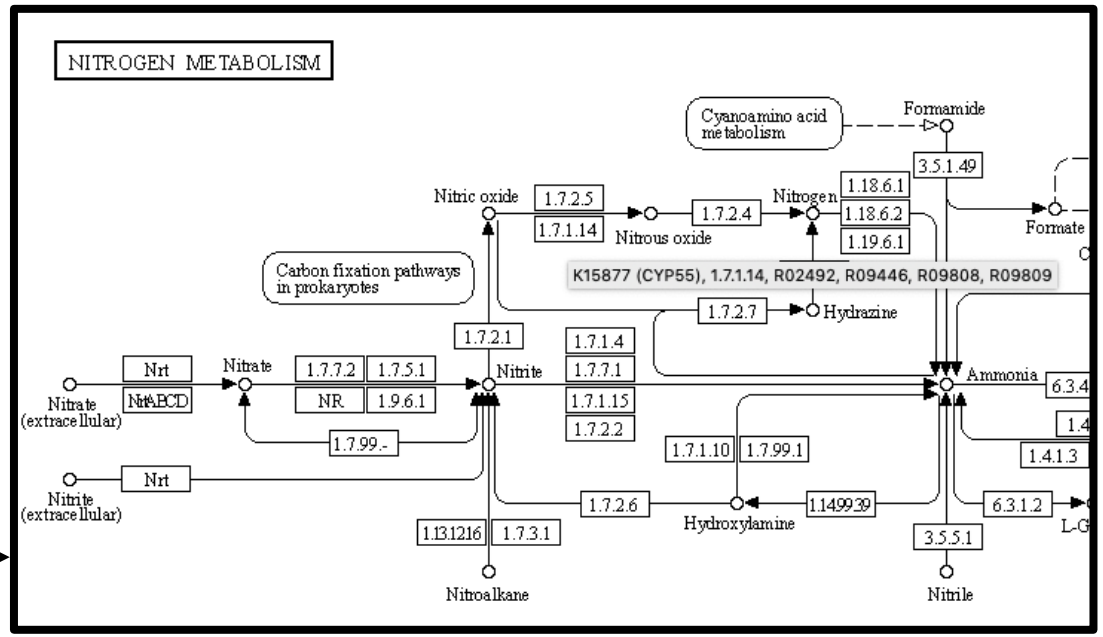
- Global and overview maps
  - 01100 Metabolic pathways
  - 01110 Biosynthesis of secondary metabolites
  - 01120 Microbial metabolism in diverse environments
  - 01130 Biosynthesis of antibiotics
  - 01200 Carbon metabolism
  - 01210 2-Oxocarboxylic acid metabolism
  - 01212 Fatty acid metabolism
  - 01230 Biosynthesis of amino acids
  - 01220 Degradation of aromatic compounds
- Carbohydrate metabolism
- Energy metabolism
  - 00190 Oxidative phosphorylation
  - 00195 Photosynthesis
  - 00196 Photosynthesis - antenna proteins
  - 00710 Carbon fixation in photosynthetic organisms
  - 00720 Carbon fixation pathways in prokaryotes
  - 00680 Methane metabolism
  - 00910 Nitrogen metabolism
  - 00920 Sulfur metabolism
- Lipid metabolism
- Nucleotide metabolism
- Amino acid metabolism
- Metabolism of other amino acids
- Glycan biosynthesis and metabolism
- Metabolism of cofactors and vitamins
- Metabolism of terpenoids and polyketides
- Biosynthesis of other secondary metabolites
- Xenobiotics biodegradation and metabolism
- Chemical structure transformation maps



# Gene annotation

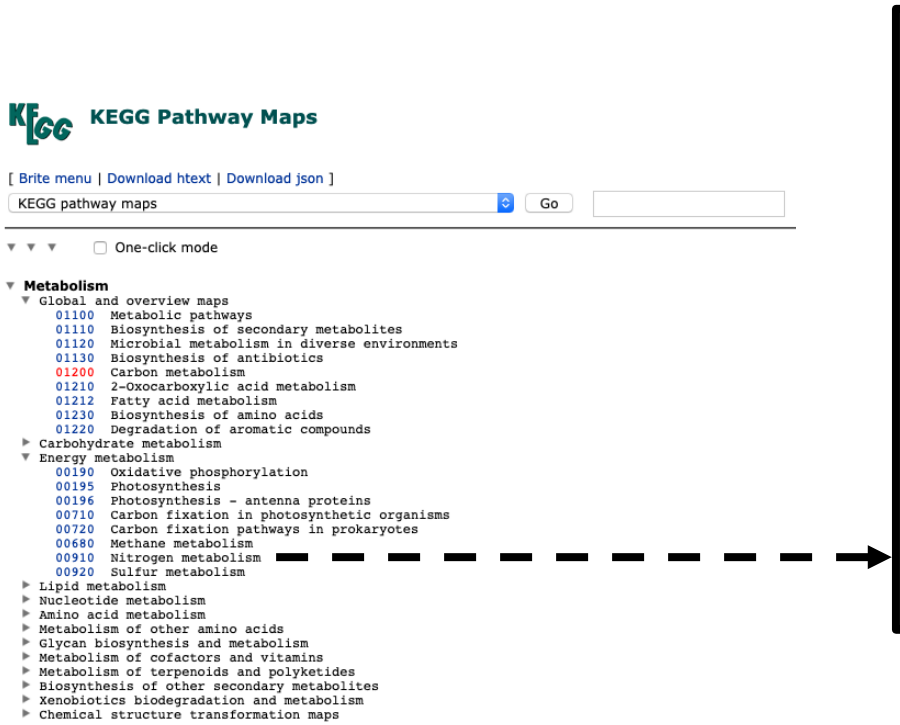
KEGG: https://www.genome.jp/kegg/pathway.html

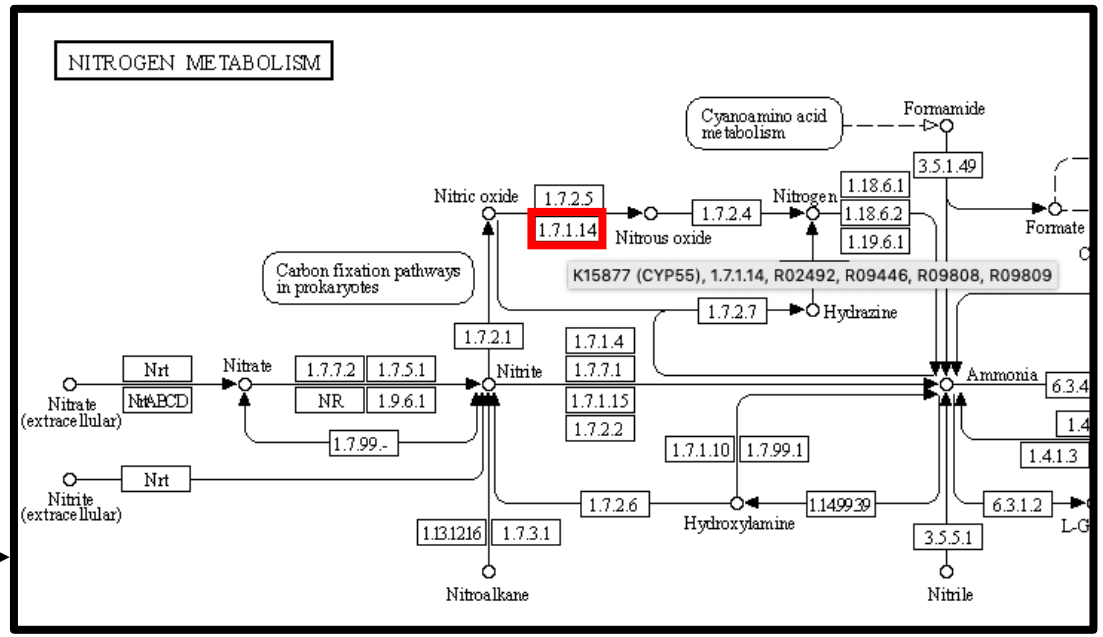


# Gene annotation

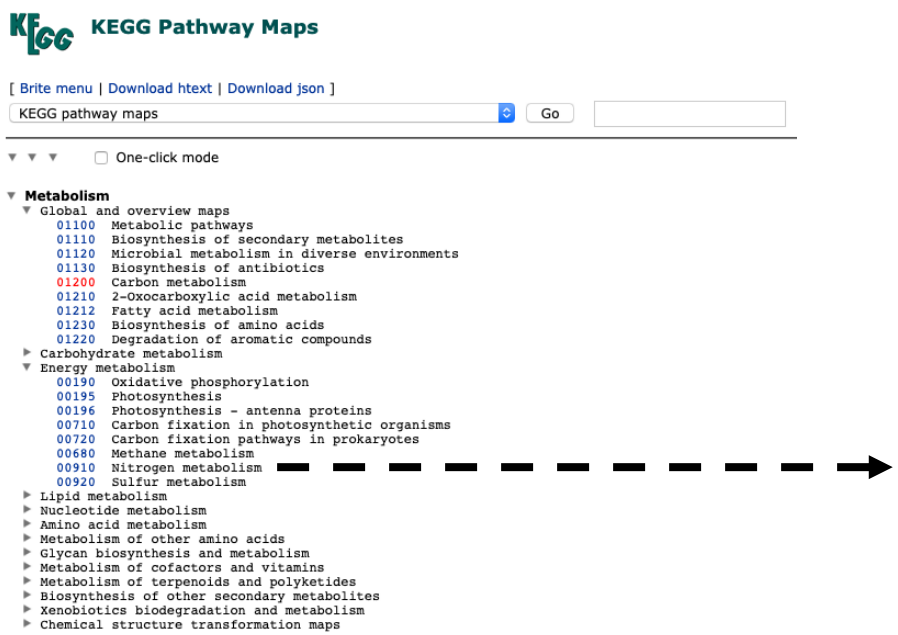
KEGG: https://www.genome.jp/kegg/pathway.html

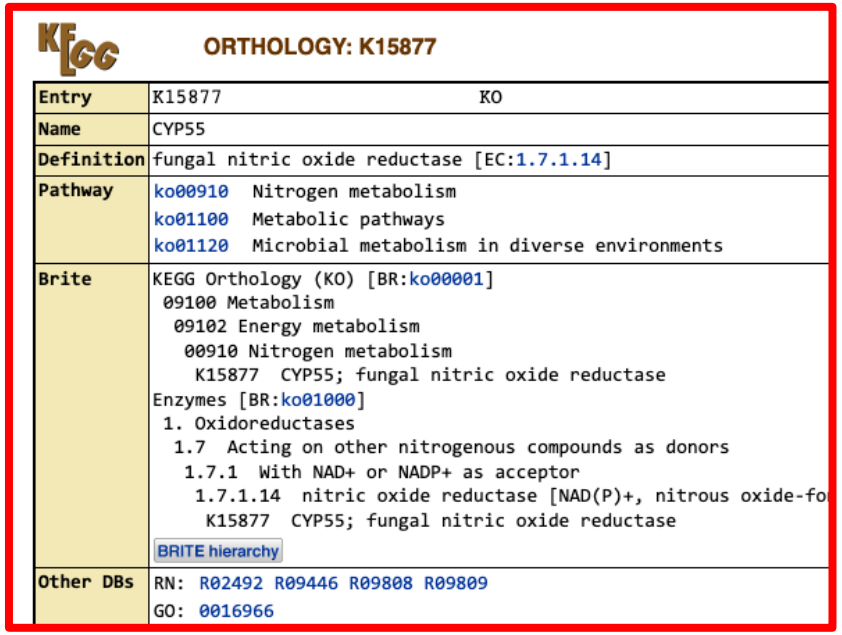


# Gene annotation

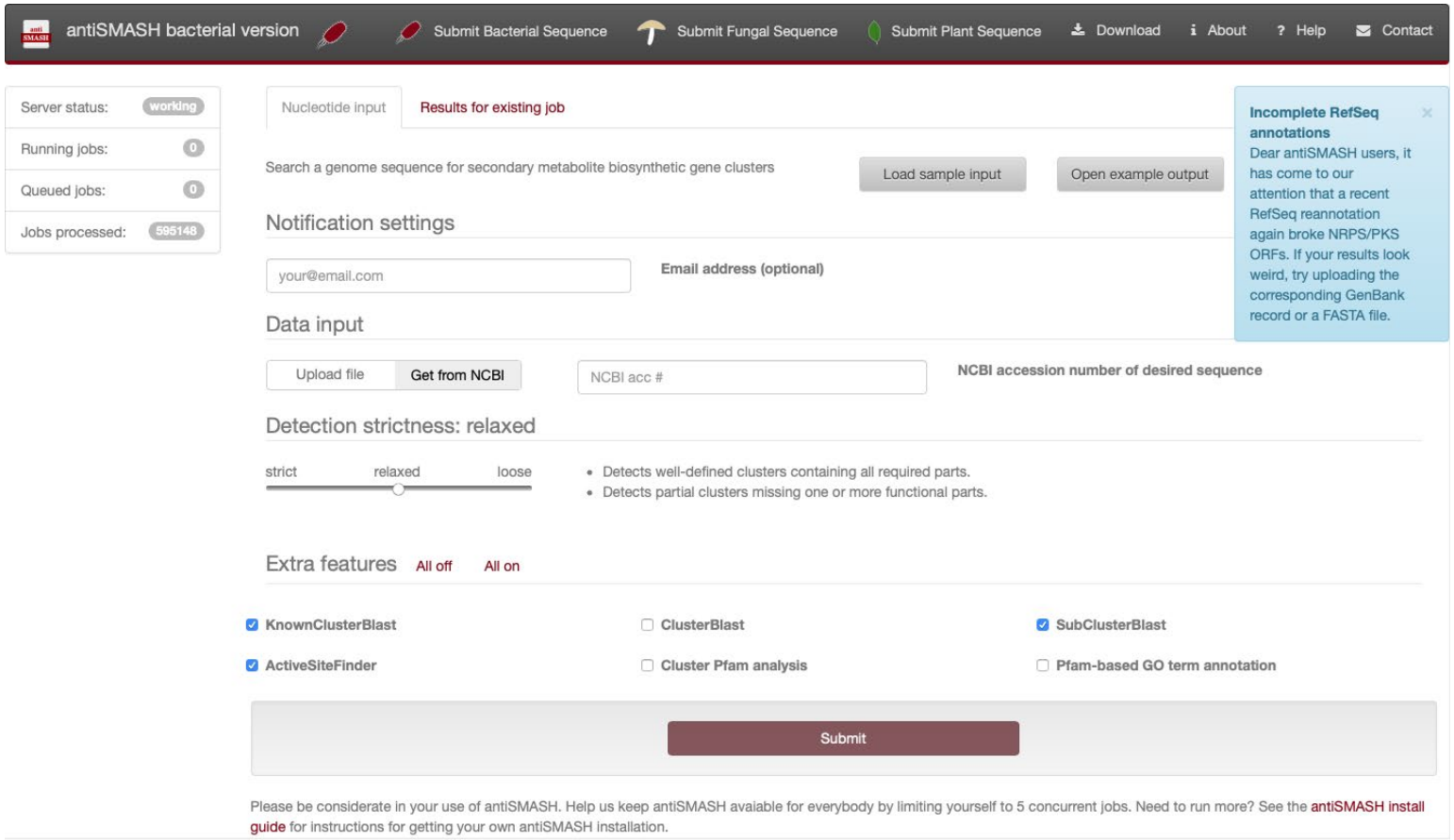
KEGG: https://www.genome.jp/kegg/pathway.html

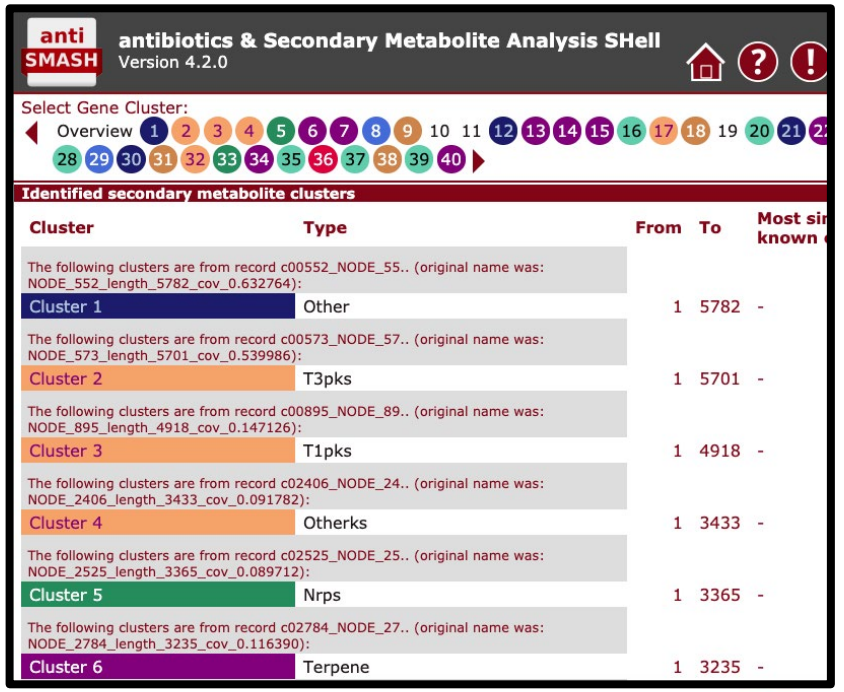


# Gene annotation

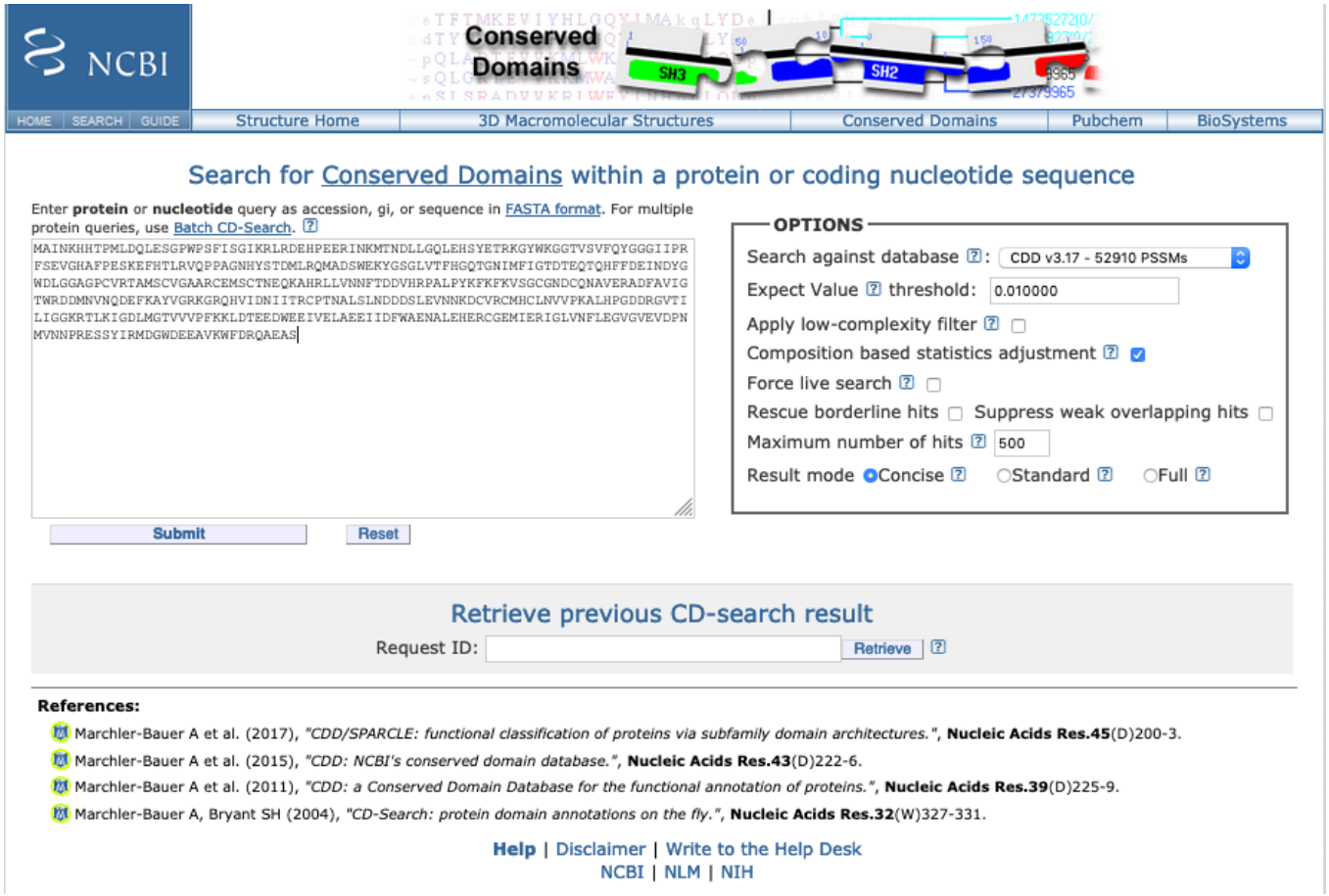
Identification of Biosynthetic Gene Clusters with **antiSMASH**



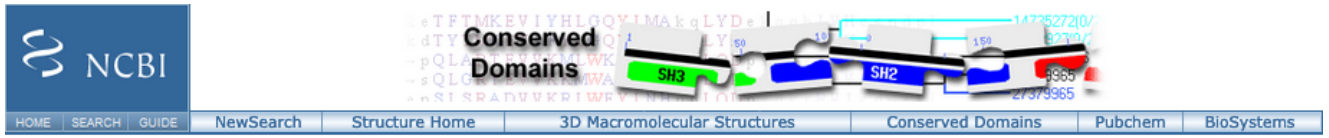

https://antismash.secondarymetabolites.org/

# NCBI Conserved Domain Search




Search nucleotide/protein sequence(s) for conserved domains

Individual search: https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Batch: https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi

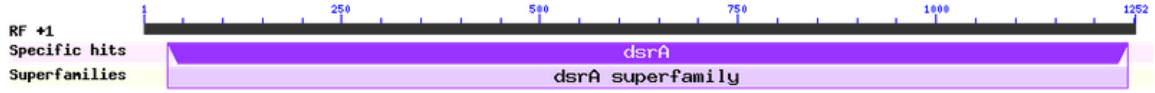



## Conserved domains on [lcl|1]

View: Concise Results

NZ_JRAA01000001.1:c722683-721433 Solemya velum gill symbiont strain WH SV_sym_Scaffold_1, whole genome shotgun sequence

**Graphical summary** | Zoom to residue level | show extra options »




Search for similar domain architectures | Refine search

**List of domain hits**

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| dsrA | TIGR02064 | sulfite reductase, dissimilatory-type alpha subunit; Dissimilatory sulfite reductase catalyzes ... | 31-1242 | 0e+00 |

sulfite reductase, dissimilatory-type alpha subunit; Dissimilatory sulfite reductase catalyzes the six-electron reduction of sulfite to sulfide, as the terminal reaction in dissimilatory sulfate reduction. It remains unclear however, whether trithionate and thiosulfate serve as intermediate compounds to sulfide, or as end products of sulfite reduction. Sulfite reductase is a multisubunit enzyme composed of dimers of either alpha/beta or alpha/beta/gamma subunits, each containing a siroheme and iron sulfur cluster prosthetic center. Found in sulfate-reducing bacteria, these genes are commonly located in an unidirectional gene cluster. This model describes the alpha subunit of sulfite reductase. [Central intermediary metabolism, Sulfur metabolism]

:

**Pssm-ID:** 273948 [Multi-domain] **Cd Length:** 402 **Bit Score:** 667.31 **E-value:** 0e+00

```
                    10        20        30        40        50        60        70        80
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1              11 LDQLESGPWPSFISGIKRLRDEHPEERINKMTNDLLGQLEHSYETRKGYWKGGTVSVFQYGGGIIPRFSEVGHAFPESKE 90
Cdd:TIGR02064   1 LDQLEKGPWPSFVSEIKKTAAYRADYQVPVDPEDLLGVLELSYDERKTHWKGGIVSVFGYGGGVIGRYSDQGEKFPGVAE 80

                    90        100       110       120       130       140       150       160
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1              91 FHTLRVQPPAGNHYSTDMLRQMADSWEKYGSGLVTFHGQTGNIMFIGTDTEQTQHFFDEINDYGWDLGGAGPCVRTAMSC 170
Cdd:TIGR02064  81 FHTVRVAQPSGKFYSTDYLRQLCDVWEKYGSGLTNFHGQTGDIVFLGTQTPQLQEIFEELTNLGTDLGGSGSNLRTPESC 160

                    170       180       190       200       210       220       230       240
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1             171 VGAARCEMSCTNEQKAHRLLVNNFTDDVHRPALPYKFKFKVSGCGNDCQNAVERADFAVIGTWRDDMNVNQDEFKAYVGR 250
Cdd:TIGR02064 161 VGPARCEFACYDTLKACYELTMEYQDELHRPAFPYKFKFKFSGCPNDCVAAIARSDFAVIGTWKDDIKVDQEAVKAYIAG 240

                    250       260       270       280       290       300       310       320
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1             251 KGRQHVIDNIITRCPTNALSLNDDDSLEVNNKDCVRCMHCLNVVPKALHPGDDRGVTILIGGKRTLKIGDLMGTVVVPFK 330
Cdd:TIGR02064 241 WGKFDIENEVVNRCPTKAISWDGSKELSIDNRECVRCMHCINKMPKALHPGDERGVTILIGGKAPILDGAQMGWVVVPFV 320

                    330       340       350       360       370       380       390       400
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1             331 klDTEEDWEEIVELAEEIIDFWAENALEHERCGEMIERIGLVNFLEGVGVEVDPNMVNNPRESSYIRMDGWDEEAVKWFD 410
Cdd:TIGR02064 321 --EAEEPYDEIKELVEKIIDWWDEEGKNRERIGETIKRLGLQKFLEVIGIEPDPQMVKEPRTNPYIFFKVEDEVPGGWDA 398

                    ....
1             411 RQAE 414
Cdd:TIGR02064 399 DIAE 402
```

**References:**

Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.

**Conserved Domain Search results:**

*dsrA* gene of *Solemya velum* gill symbiont strain WH


Search for similar domain architectures | Refine search

## List of domain hits

| Name | Accession | Description | Interval | E-value |
|---|---|---|---|---|
| dsrA | TIGR02064 | sulfite reductase, dissimilatory-type alpha subunit; Dissimilatory sulfite reductase catalyzes ... | 31-1242 | 0e+00 |

sulfite reductase, dissimilatory-type alpha subunit; Dissimilatory sulfite reductase catalyzes the six-electron reduction of sulfite to sulfide, as the terminal reaction in dissimilatory sulfate reduction. It remains unclear however, whether trithionate and thiosulfate serve as intermediate compounds to sulfide, or as end products of sulfite reduction. Sulfite reductase is a multisubunit enzyme composed of dimers of either alpha/beta or alpha/beta/gamma subunits, each containing a siroheme and iron sulfur cluster prosthetic center. Found in sulfate-reducing bacteria, these genes are commonly located in an unidirectional gene cluster. This model describes the alpha subunit of sulfite reductase. [Central intermediary metabolism, Sulfur metabolism]

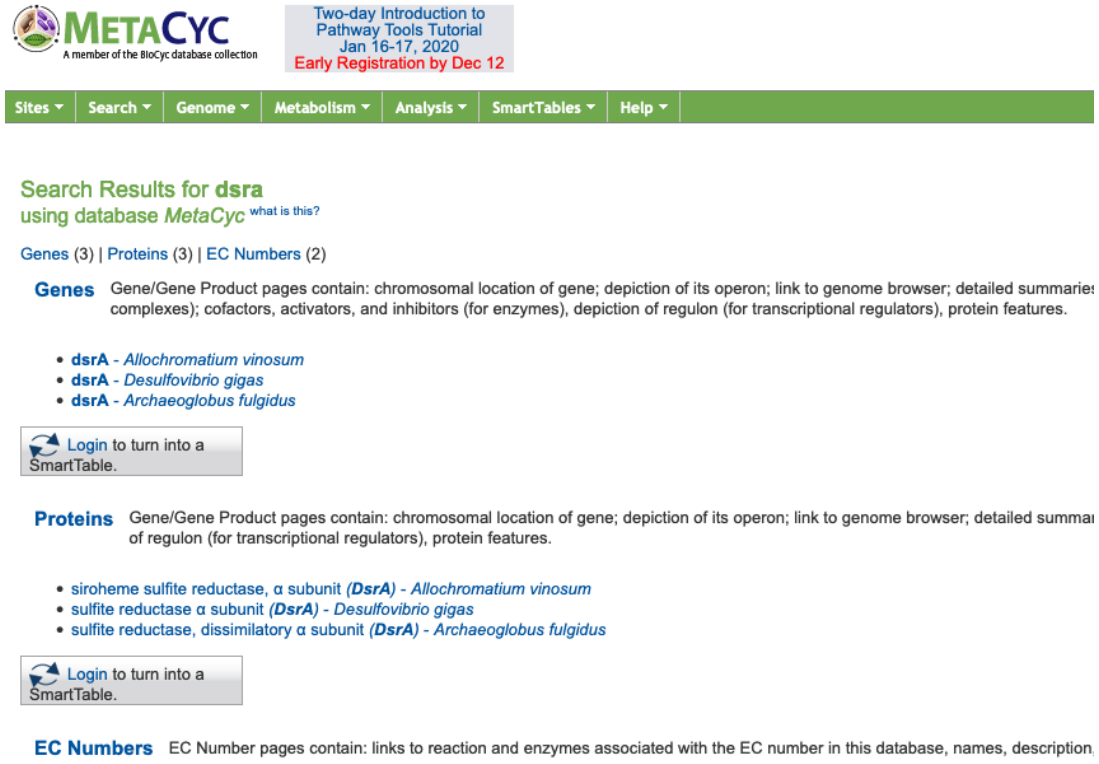
```
                    170       180       190       200       210       220       230       240
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1           171 VGAARCEMSCTNEQKAHRLLVNNFTDDVHRPALPYKFKFKVSGCGNDCQNAVERADFAVIGTWRDDMNVNQDEFKAYVGR 250
Cdd:TIGR02064 161 VGPARCEFACYDTLKACYELTMEYQDELHRPAFPYKFKFKFSGCPNDCVAAIARSDFAVIGTWKDDIKVDQEAVKAYIAG 240

                    250       260       270       280       290       300       310       320
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1           251 KGRQHVIDNIITRCPTNALSLNDDDSLEVNNKDCVRCMHCLNVVPKALHPGDDRGVTILIGGKRTLKIGDLMGTVVVPFK 330
Cdd:TIGR02064 241 WGKFDIENEVVNRCPTKAISWDGSKELSIDNRECVRCMHCINKMPKALHPGDERGVTILIGGKAPILDGAQMGWVVVPFV 320

                    330       340       350       360       370       380       390       400
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1           331 klDTEEDWEEIVELAEEIIDFWAENALEHERCGEMIERIGLVNFLEGVGVEVDPNMVNNPRESSYIRMDGWDEEAVKWFD 410
Cdd:TIGR02064 321 --EAEEPYDEIKELVEKIIDWWDEEGKNRERIGETIKRLGLQKFLEVIGIEPDPQMVKEPRTNPYIFFKVEDEVPGGWDA 398

                    ....
1           411 RQAE 414
Cdd:TIGR02064 399 DIAE 402
```

**References:**

Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.

# Gene annotation

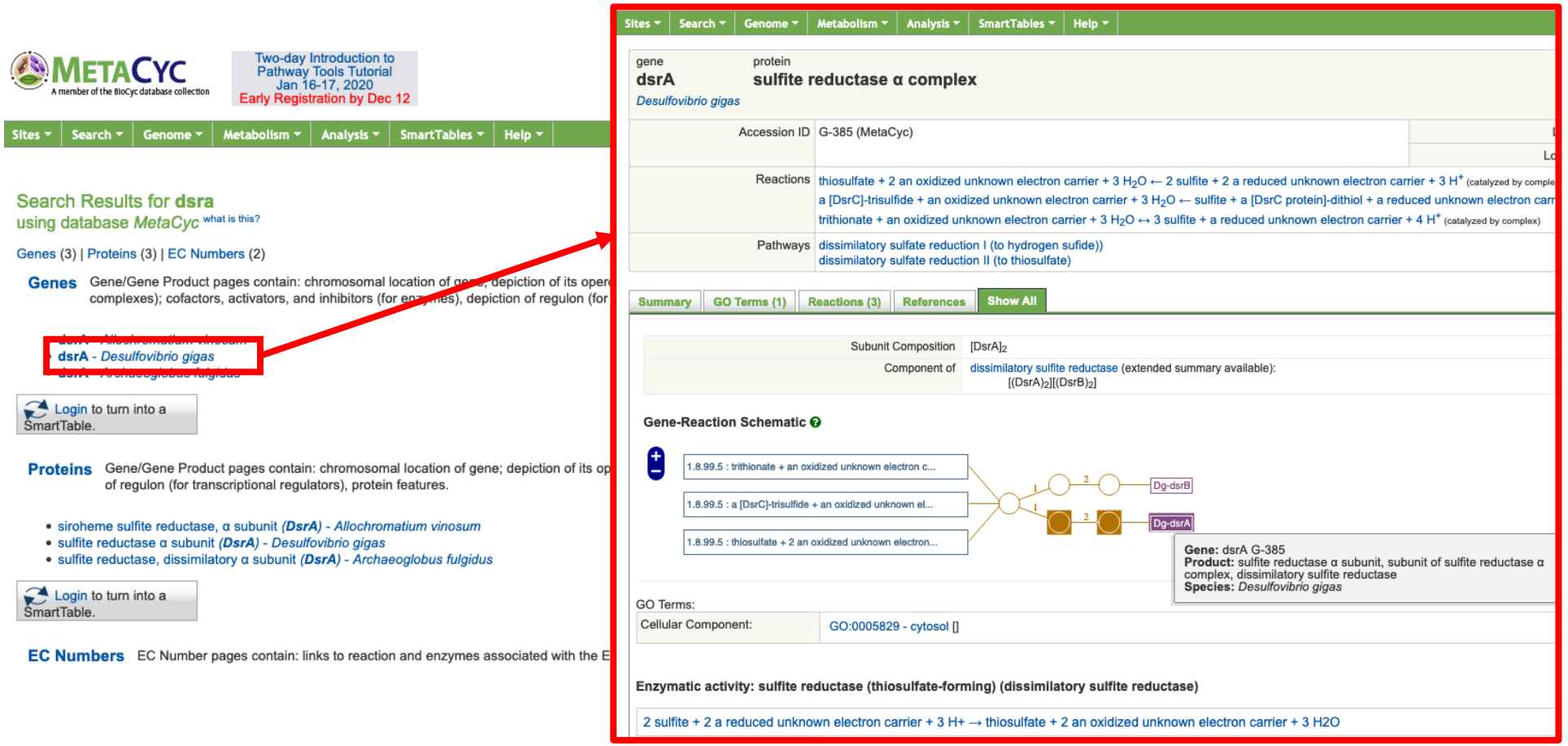
Metacyc: experimentally curated metabolic pathways




https://metacyc.org/

# Gene annotation

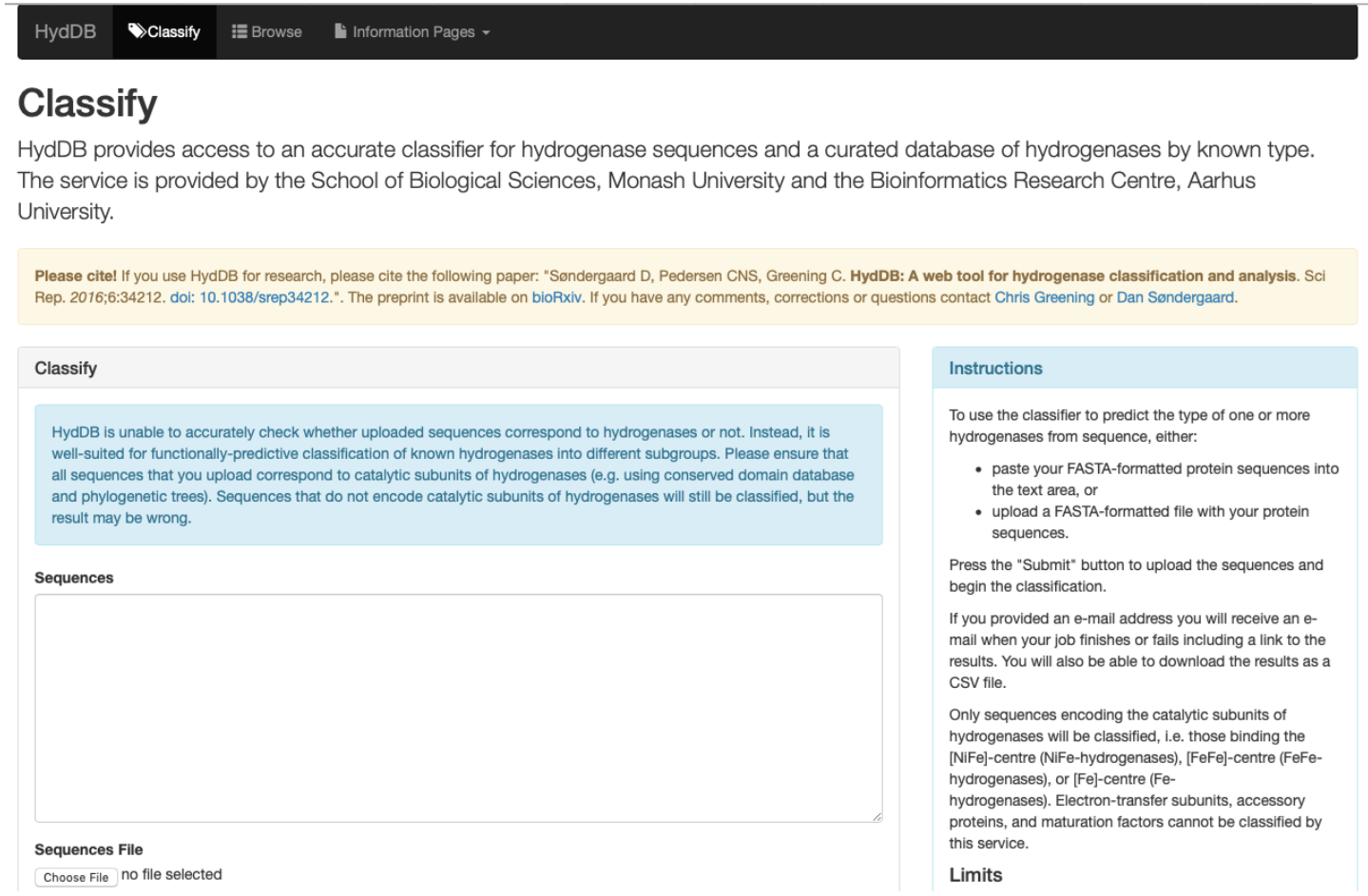
Metacyc: experimentally curated metabolic pathways




https://metacyc.org/

# Gene annotation

## Accurate classifier for hydrogenase sequences - **HydDB**




https://services.birc.au.dk/hyddb/

# Gene annotation

- The **PSORT** family - prediction of protein localization sites in cells.
- Useful for making cell schematics!

psortb

Updates | Documentation | Resources | Contact

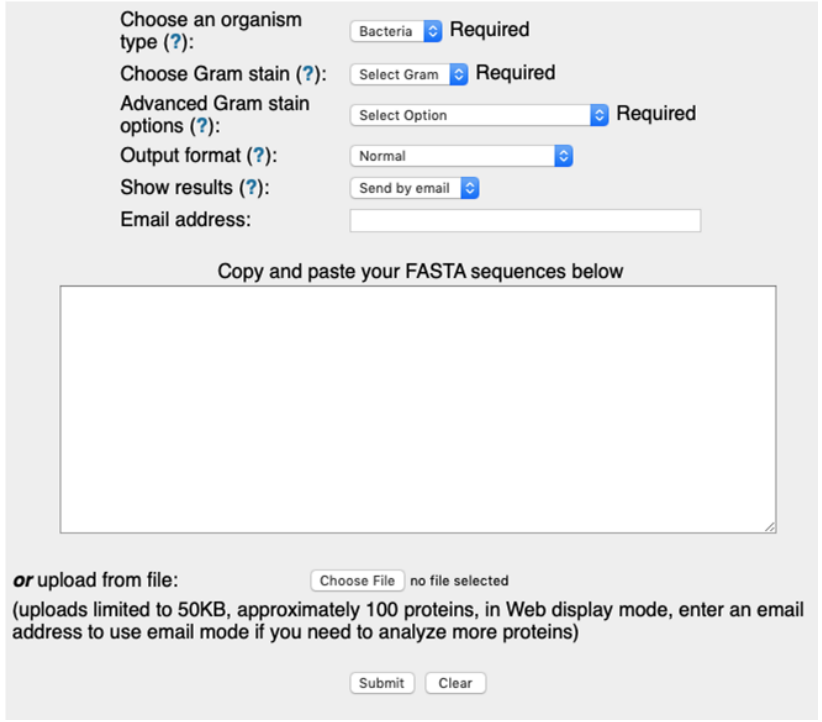
**Submit a Sequence to PSORTb version 3.0.2**

Based on a study last performed in 2010, PSORTb v3.0.2 is the most precise bacterial localization prediction tool available. PSORTb v3.0.2 has a number of **improvements** over PSORTb v2.0.4. Version 2 of PSORTb is maintained **here**.

You can currently submit one or more Gram-positive or Gram-negative bacterial sequences or archaeal sequences in FASTA format (**?**). Copy and paste your FASTA-formatted sequences into the textbox below or select a file containing your sequences to upload from your computer. Web display mode is limited to the analysis of approximately 100 proteins. For larger analyses, either enter your email address in the form below (results of up to 5000 per submission returned by email) or for even larger analyses we can help you or you can download the standalone version.

See also:

- **Updates**
- **Precomputed genome results**
- **Limitations of PSORTb v.3.0**
- **PSORTb User's Guide**
- **Docker PSORTb web service** (what is **docker**?)
- **Download standalone PSORTb**
- **Docker standalone PSORTb** (what is **docker**?)




http://psort.org/

# Gene annotation

## SignalP 6.0 (https://services.healthtech.dtu.dk/service.php?SignalP)

- Predicts signal peptides on amino acid sequences across kingdoms of life.
- Uses Protein Language Models

Submission | Instructions | Data | Article abstract | FAQ | Version history | Portable | Downloads

### Submit data

**Sequence submission: paste the sequence(s) *and/or* upload a local file**

*Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 5000.*
*The long output format might timeout for more than 100 entries.*

Mirror Use SignalP 6.0 on BioLib if this server is heavily loaded.

Enter protein sequence(s) in fasta format...

For example proteins Click here
Format directly from your local disk:
Browse... No file selected.

**Organism**
- Eukarya
- Other

"Eukarya" only predicts Sec/SPI SPs.

**Output format:**
- Long output
- Short output (no figures)

**Model mode:**
- Fast
- Slow

The slow mode takes 6x longer to compute. Use when accurate region borders are needed.

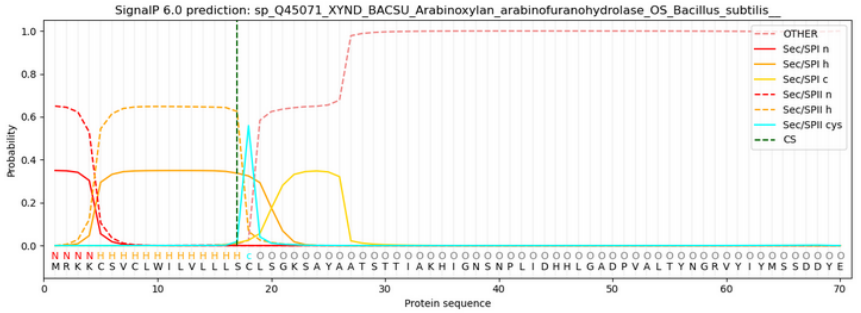
Submit Clear fields

**sp_Q45071_XYND_BACSU_Arabinoxylan_arabinofuranohydrolase_OS_Bacillus_subtilis__**
**Prediction:** Lipoprotein signal peptide (Sec/SPII)

Cleavage site between pos. 17 and 18. Probability 0.625165

| Protein type | Other | Signal Peptide (Sec/SPI) | Lipoprotein signal peptide (Sec/SPII) | TAT signal peptide (Tat/SPI) | TAT Lipoprotein signal peptide (Tat/SPII) | Pilin-like signal peptide (Sec/SPIII) |
|---|---|---|---|---|---|---|
| **Likelihood** | 0.0012 | 0.3394 | 0.6584 | 0.0004 | 0.0003 | 0.0003 |

**Download:** PNG / EPS / Tabular



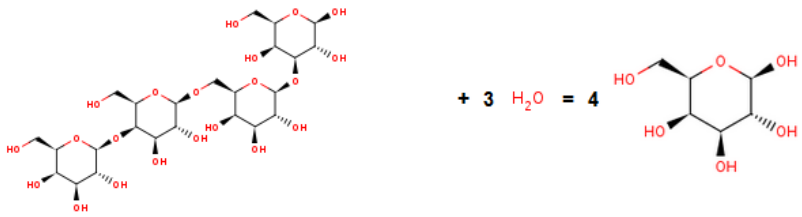
# Gene annotation

BRENDA: https://www.brenda-enzymes.org/

- Enzyme database
- Has reaction and pathway schematics
  - In-house pathways
  - KEGG pathways

**Reaction Schemes** hide

beta-D-galactopyranosyl-(1-4)-beta-D-galactopyranosyl-(1-6)-beta-D-galactopyranosyl-(1-3)-beta-D-galactopyranose + 3 H2O = 4 beta-D-galactopyranose




## Information on EC 3.2.1.23 - beta-galactosidase

for references in articles please use BRENDA:EC3.2.1.23

show all entries

**EC Tree**

- 3 Hydrolases
  - 3.2 Glycosylases
    - 3.2.1 Glycosidases, i.e. enzymes that hydrolyse *O*- and *S*-glycosyl compounds
      - 3.2.1.23 beta-galactosidase

**IUBMB Comments**

Some enzymes in this group hydrolyse alpha-L-arabinosides; some animal enzymes also hydrolyse beta-D-fucosides and beta-D-glucosides; cf. EC 3.2.1.108 lactase.

top print hide 7 entries Go to Pathway Search

| PATHWAY SOURCE ▲▼ | PATHWAYS ▲▼ |
|---|---|
| BRENDA | metabolism of disaccharids |
| KEGG | Galactose metabolism, Glycosaminoglycan degradation, Glycosphingolipid biosynthesis - ganglio series, Other glycan degradation, Sphingolipid metabolism |
| - | Sphingolipid metabolism |

# Summary of online resources

**Resources to help interpret your data:**

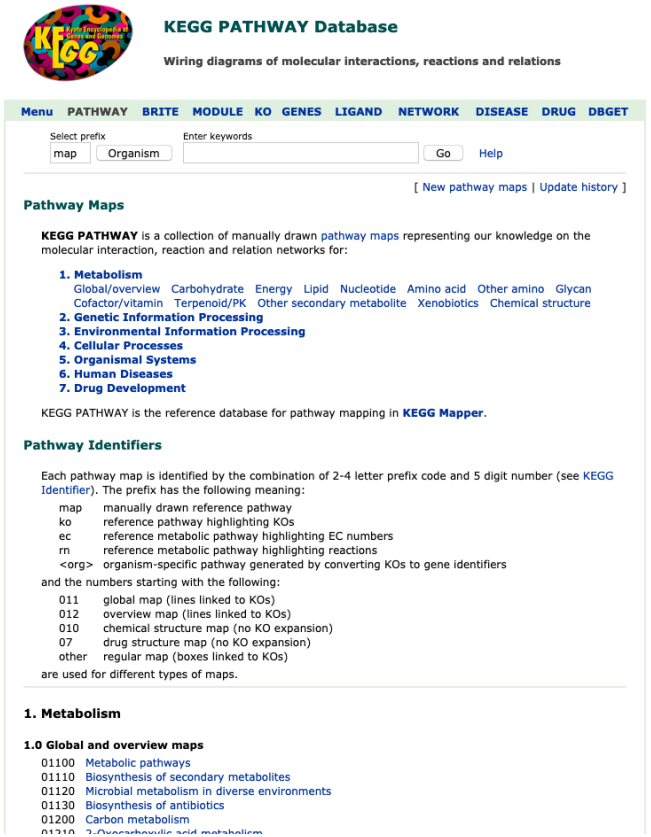
- **KEGG:** https://www.genome.jp/kegg/pathway.html
- **BioCyc:** https://biocyc.org/
- **MetaCyc:** https://metacyc.org/
- **HydDB:** https://services.birc.au.dk/hyddb/
- **PSORT:** https://psort.hgc.jp/
- **SignalP:** https://services.healthtech.dtu.dk/service.php?SignalP
- **BRENDA:** https://www.brenda-enzymes.org/

# Task: Gene annotation

**Tasks:**

- **View KEGG annotation in KEGG website**

KEGG: https://www.genome.jp/kegg/pathway.html



**KEGG PATHWAY Database**

**Wiring diagrams of molecular interactions, reactions and relations**

Menu PATHWAY BRITE MODULE KO GENES LIGAND NETWORK DISEASE DRUG DBGET

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 on 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**

KEGG PATHWAY is the reference database for pathway mapping in **KEGG Mapper**.

**Pathway Identifiers**

Each pathway map is identified by the combination of 2-4 letter prefix code and 5 digit number (see KEGG Identifier). The prefix has the following meaning:

| | |
|---|---|
| map | manually drawn reference pathway |
| ko | reference pathway highlighting KOs |
| ec | reference metabolic pathway highlighting EC numbers |
| rn | reference metabolic pathway highlighting reactions |
| <org> | organism-specific pathway generated by converting KOs to gene identifiers |

and the numbers starting with the following:

| | |
|---|---|
| 011 | global map (lines linked to KOs) |
| 012 | overview map (lines linked to KOs) |
| 010 | chemical structure map (no KO expansion) |
| 07 | drug structure map (no KO expansion) |
| other | regular map (boxes linked to KOs) |

are used for different types of maps.

**1. Metabolism**

**1.0 Global and overview maps**

01100 Metabolic pathways
01110 Biosynthesis of secondary metabolites
01120 Microbial metabolism in diverse environments
01130 Biosynthesis of antibiotics
01200 Carbon metabolism

# Recorded Talk:
# Acidithiobacillus and thermophilic adaptations
# (Chanenath Sriaporn)

# Task: DRAM

Go to Github MGSS webpage

**Tasks:**

- **MAG annotation with DRAM**

# Task: Coverage calculation

Go to Github MGSS webpage

**Tasks:**

- **Coverage calculation using Bowtie2**

# Task: Pick group challenge!

**Tasks:**

- **Introduce group project goals**
- **Dividing into working groups / get a group name**
- **Select a goal from your project**

# Task: Pick group challenge!

**Determine which genome(s) have the following attributes, and the genetic mechanisms used for these attributes:**

1. **Denitrification** (Nitrate or nitrite to nitrogen)
2. **Ammonia oxidation** (Ammonia to nitrite or nitrate)
3. **Anammox** (Ammonia and nitrite to nitrogen)
4. **Sulfur oxidation** (SOX pathway, thiosulfate to sulfate)
5. **Sulfur reduction** (DSR pathway, sulfate to sulfide)
6. **Photosynthetic carbon fixation**
7. **Non-photosynthetic carbon fixation** (Reverse TCA or Wood-Ljungdahl)
8. **Non-polar flagella expression due to a chromosomal deletion**
9. **Plasmid-encoded antibiotic resistance**
10. **Aerobic (versus anaerobic) metabolism**