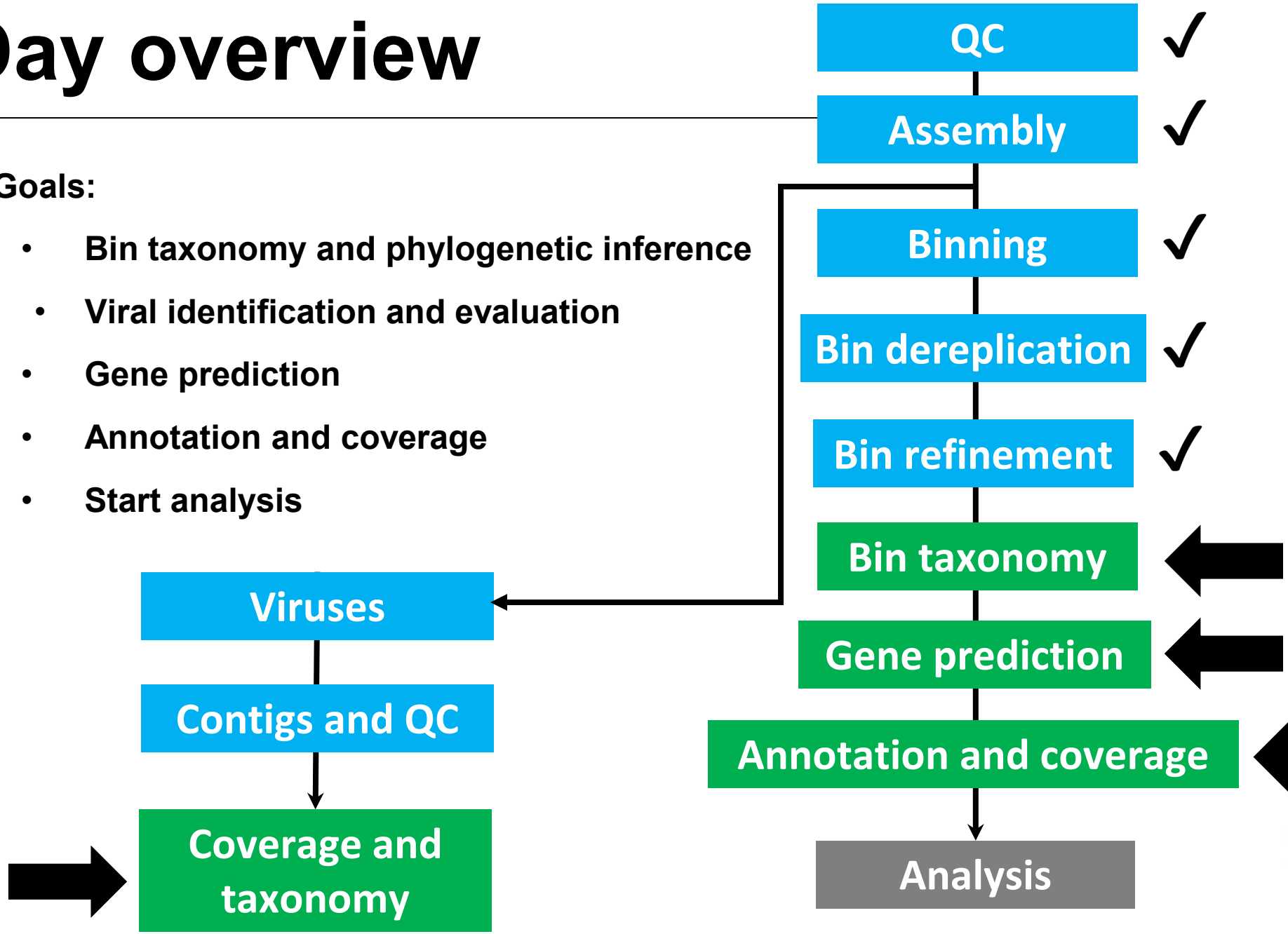
Day 3

Taxonomic classification
Phylogenetic inference
Viruses
Gene prediction & annotation

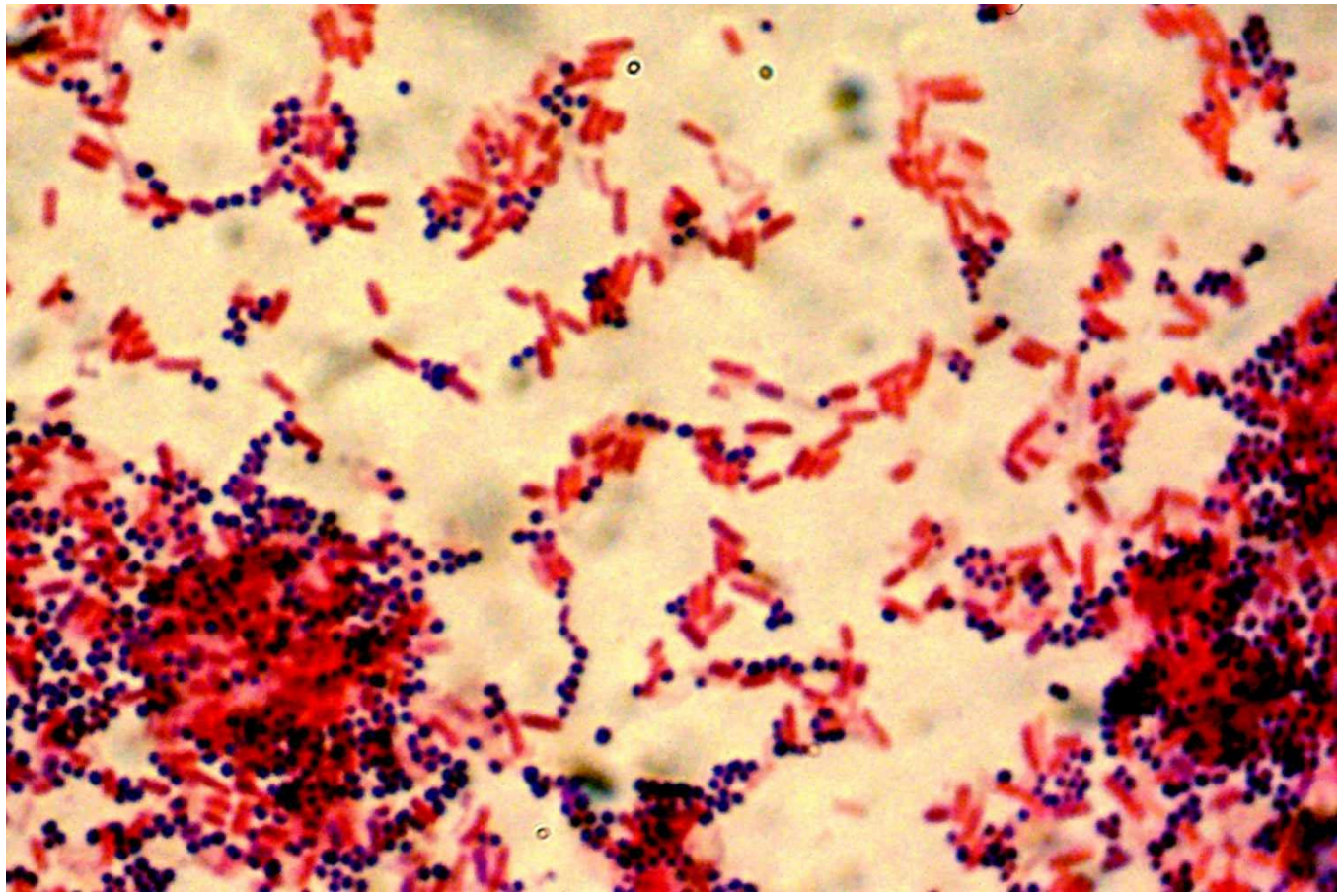
Day overview

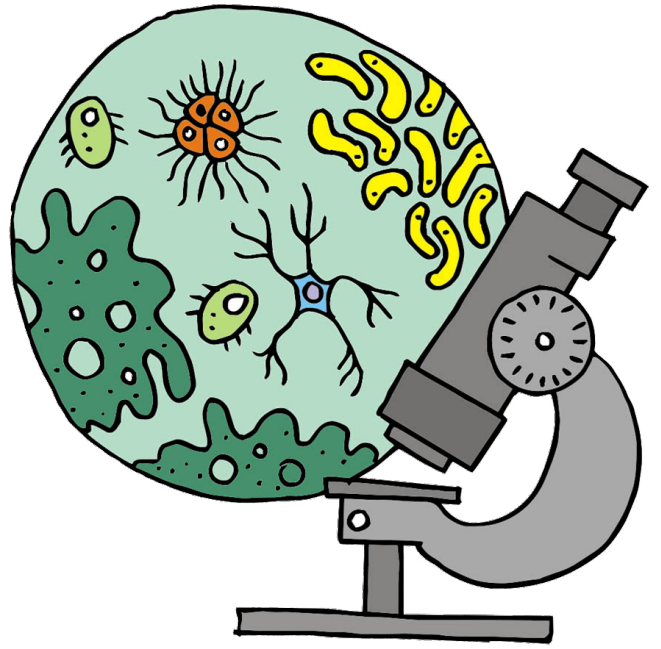
- Goals:
  - Bin taxonomy and phylogenetic inference
  - Viral identification and evaluation
  - Gene prediction
  - Annotation and coverage
  - Start analysis



Bin taxonomic classification

Bin taxonomic classification

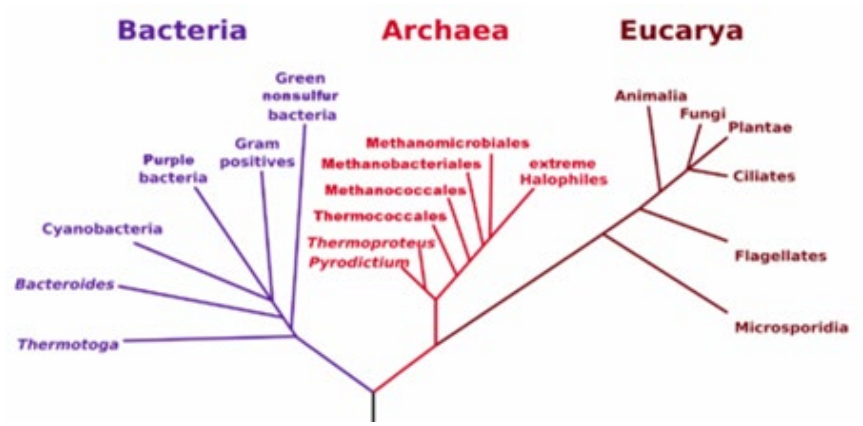




Bin taxonomic classification

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

- Traditionally used 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 PhyloFlash and/or EMIRGE
- Caveat: can be difficult to assign to genomes in complex communities with many similar taxa

How do we assign taxonomy to bins then?

Bin taxonomic classification

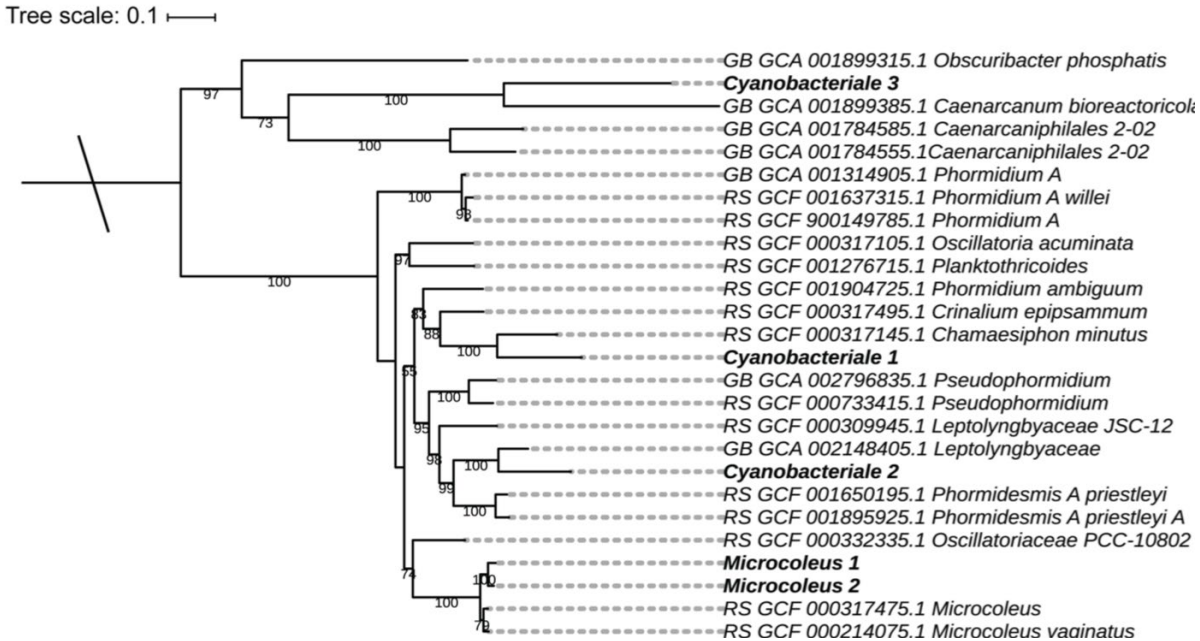
Solution:

- **Use one or more single copy core genes** (Wang et al., 2022, Scientific Reports)
- **Universal single-copy genes (USCGs) are marker genes that occur once and only once in almost every genome**

Bin taxonomic classification

Solution:

- **Use one or more single copy core genes** (Wang et al., 2022, Scientific Reports)
- **Universal single-copy genes (USCGs) are marker genes that occur once and only once in almost every genome**
- **Concatenate protein sequences of multiple single copy core genes**



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

Bin taxonomic classification

GTDB builds upon a number of existing public resources in order to provide a taxonomic resource that reflects recently proposed taxa, changes in taxonomic opinion, and the wealth of publicly available genomes.

(Parks et al., 2021, Nucleic Acids Research)

GTDB - how it works

Genomes are obtained from NCBI and must meet the following criteria to be included in the GTDB reference trees and database:

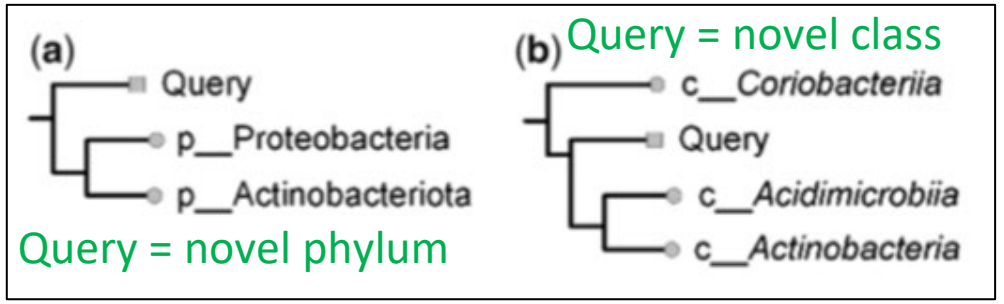
1. CheckM completeness estimate >50%
2. CheckM contamination estimate <10%
3. quality score, defined as completeness - 5*contamination, >50
4. contain >40% of the bac120 or arc53 marker genes
5. contain <1000 contigs
6. have an N50 >5kb
7. contain <100,000 ambiguous bases

Filtered genomes are manually inspected and exceptions are made for genomes of high nomenclatural or taxonomic significance

https://gtdb.ecogenomic.org/methods

Bin taxonomic classification

- **Genome Taxonomy Database Toolkit (GTDB-Tk)**
- **Use to classify bins against reference genome trees (GTDB)**
- **Uses set of 120 concatenated protein sequences (of single copy core genes)**



(Chaumeil et al., 2019, Bioinformatics)

Rank assignment based on:

- Tree placement in GTDB ref tree
- Relative Evolutionary Divergence (value between 0=root and 1=tip)
- Species assignment:
  - ANI to reference genomes

https://github.com/Ecogenomics/GtdbTk

Discriminate species

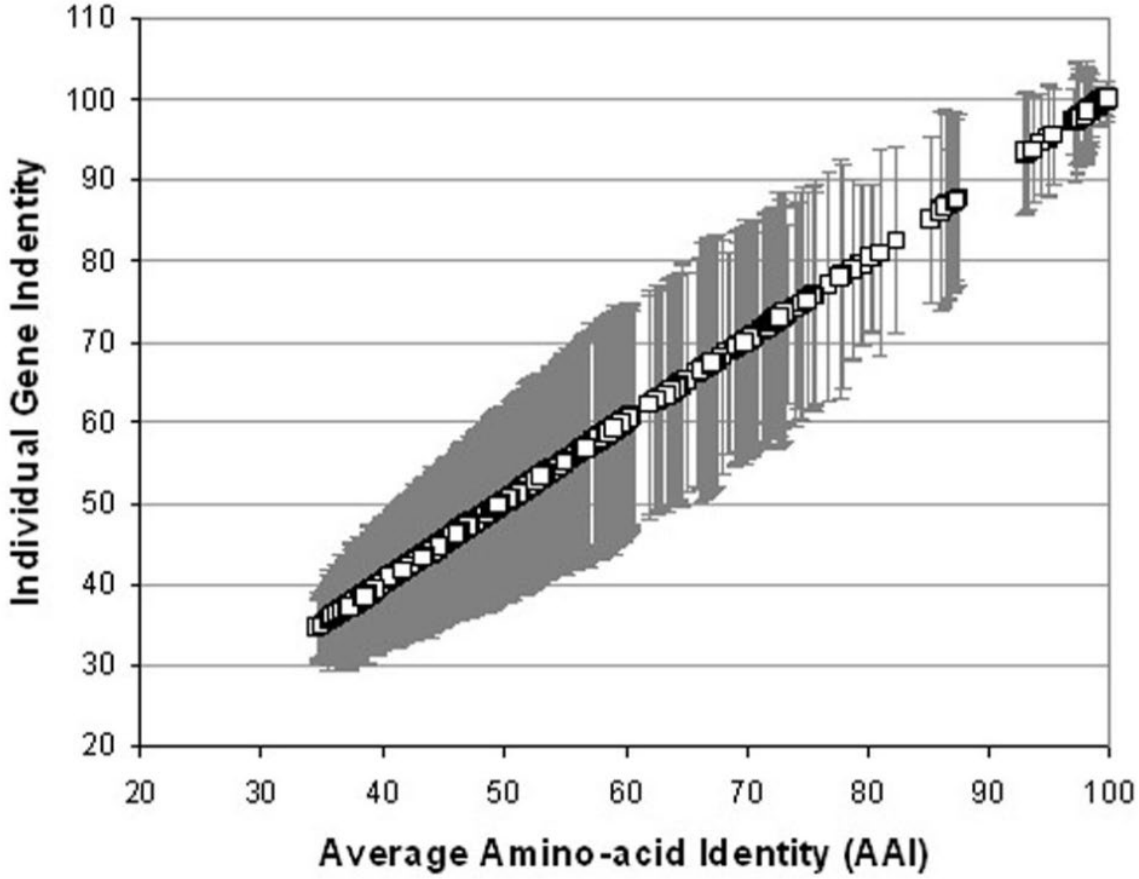
Proxy for DNA-DNA hybridization

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**

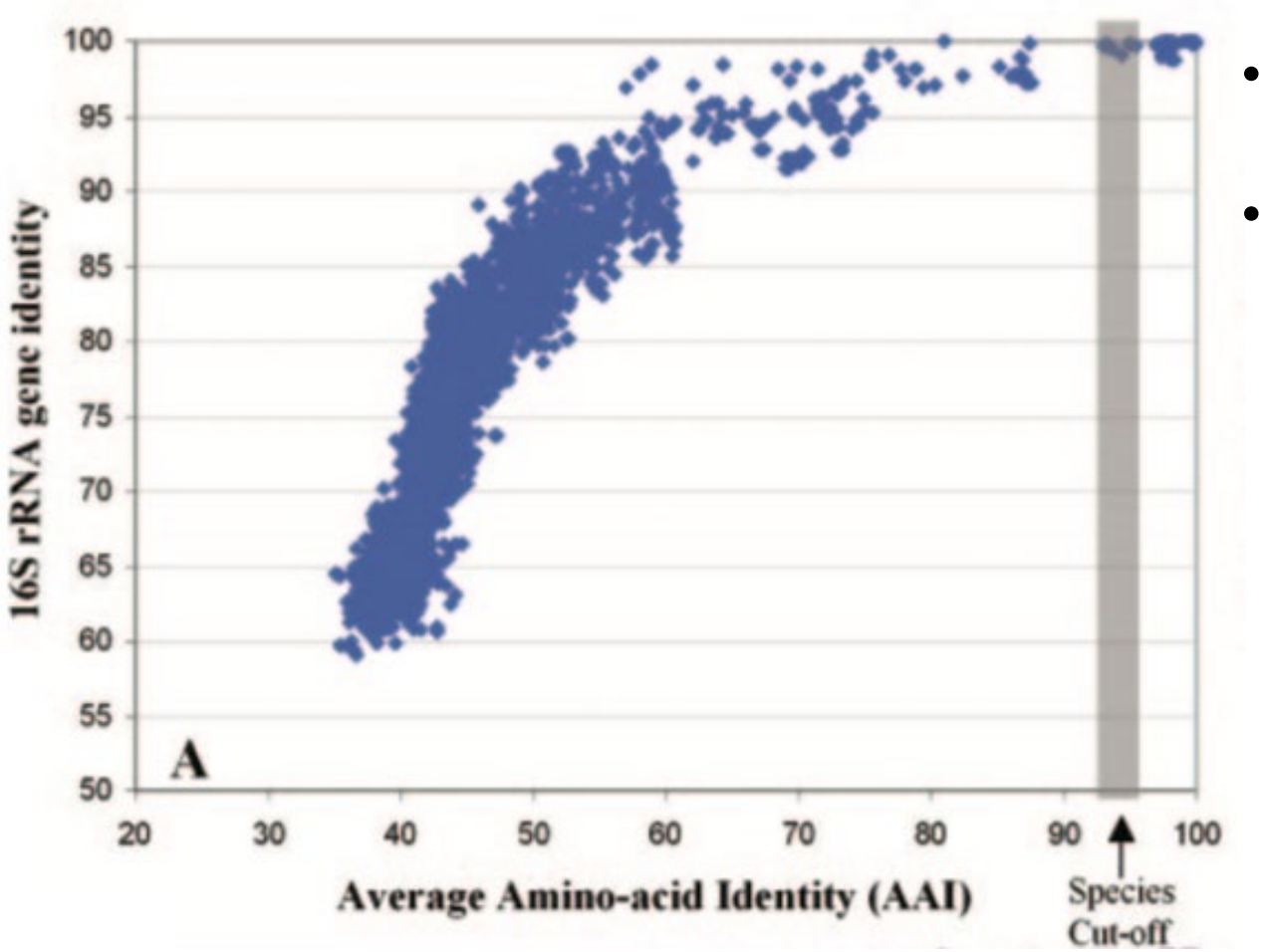
Determine via: Pairwise BLAST-like search

Discriminate species



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

Discriminate species

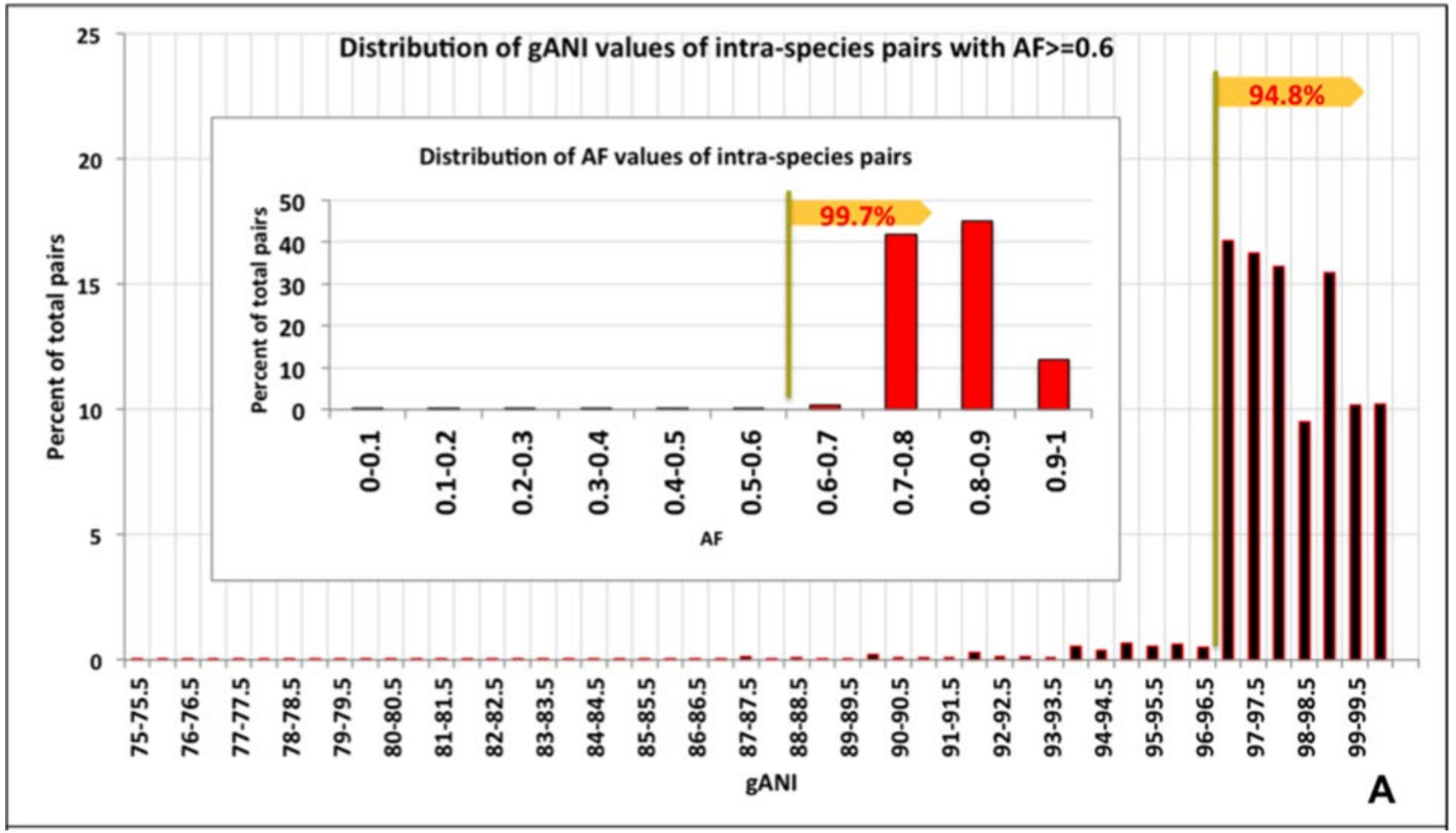


- AAI species cutoff ≈95-96% ID
- Equivalent to 70% DNA-DNA hybridization threshold for species

(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	45	44	43	45	45	44	45	44	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

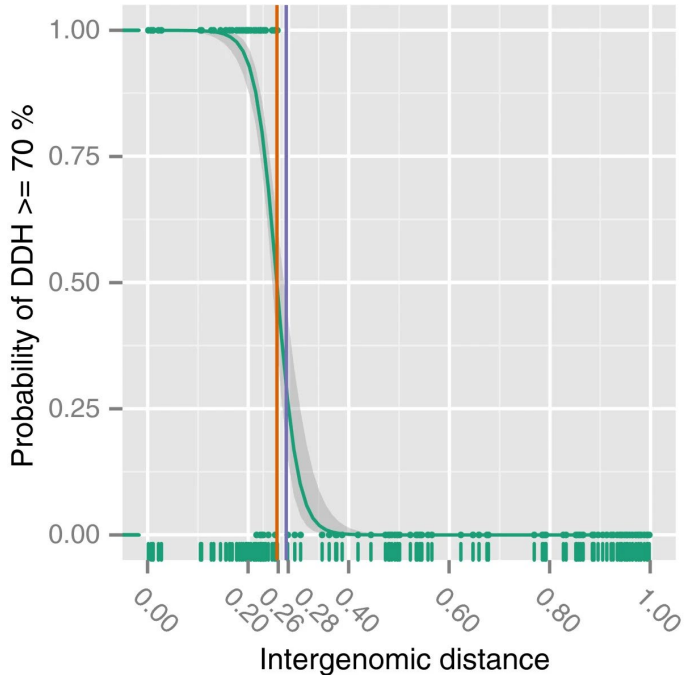
(Handley et al., 2017, ISME J)

- 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



Task: Genome taxonomic classification

Go to Github MGSS webpage

Tasks:

- **Taxonomic classification of bins using GTDB-Tk**

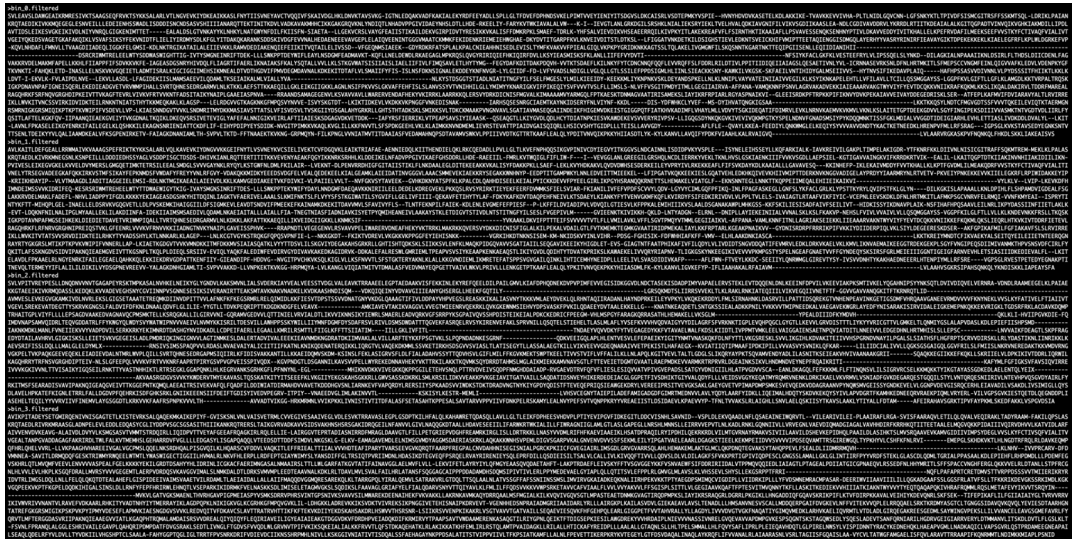
Phylogenetics

Hot take

“There’s no such thing as a genome tree”

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

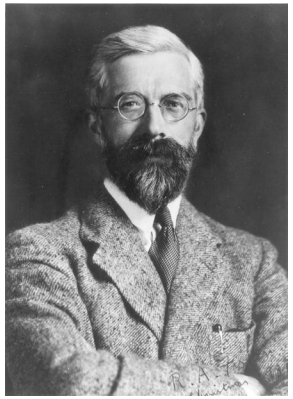
Phylogenetics

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 χ^2, and on comparing term by term the two expressions it appears that

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

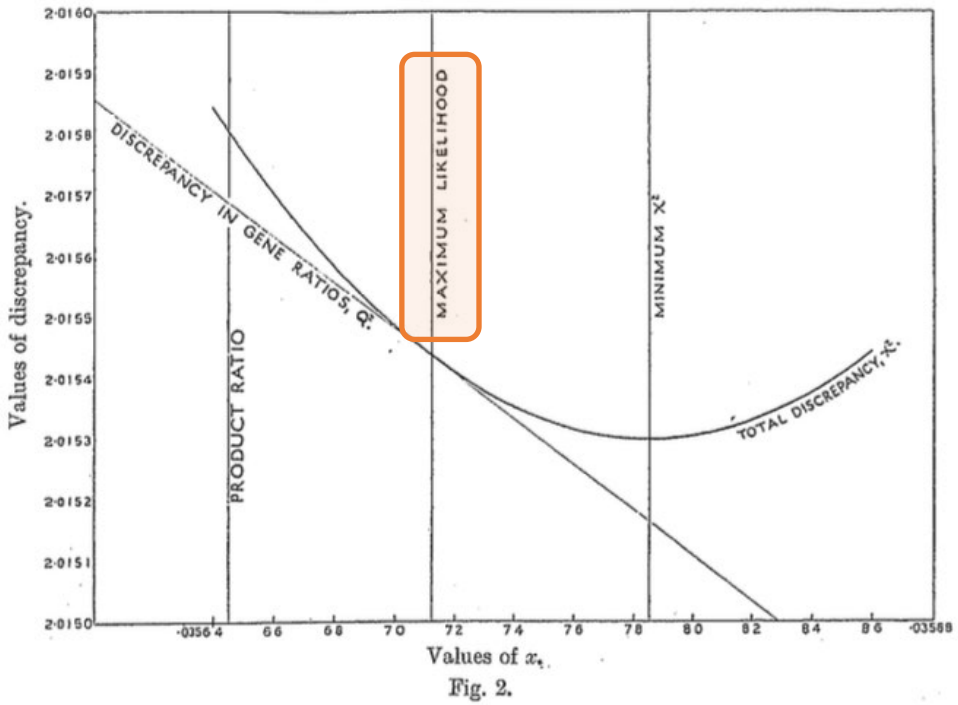


Fig. 2.

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

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

JOURNAL ARTICLE

RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies

Alexandros Stamatakis Author Notes

Bioinformatics, Volume 30, Issue 9, May 2014, Pages 1312–1313, https://doi.org/10.1093/bioinformatics/btu033
Published: 21 January 2014 **Article history ▾**

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

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

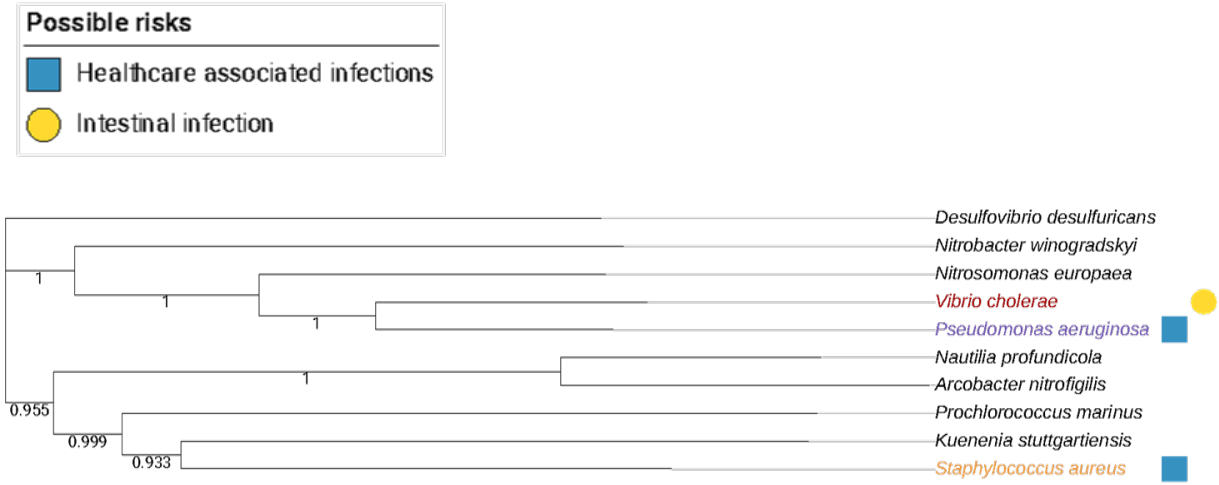
Pros: Provides better accuracy due to its advanced optimization methods (**Maximum Likelihood Optimization)**

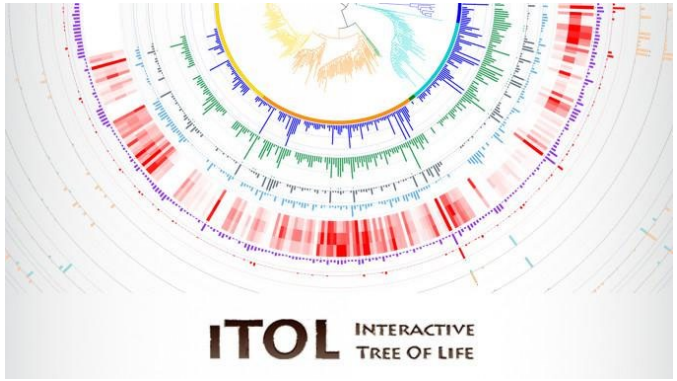
More customizations and flexibilities

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)`





Task: Build a phylogenetic tree

Go to Github MGSS webpage

Tasks:

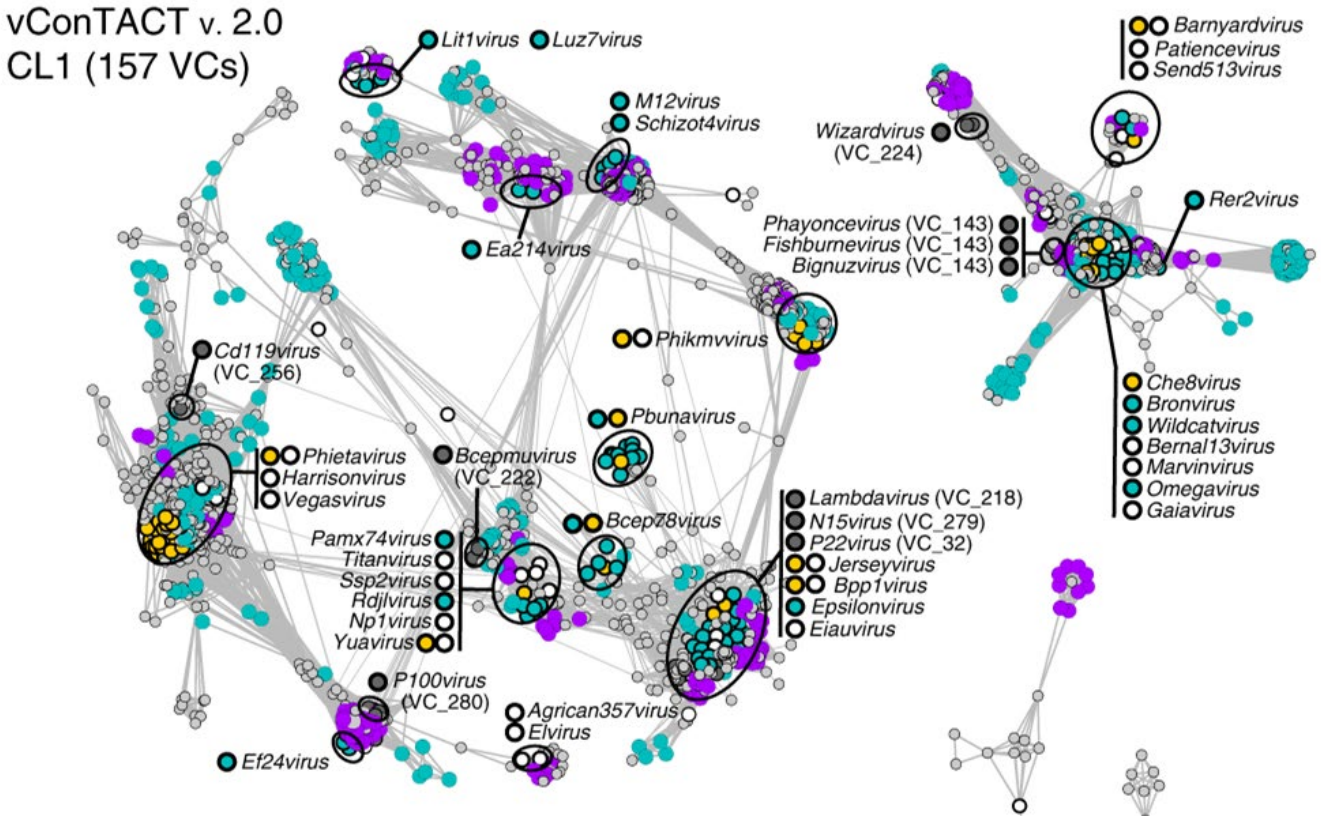
- ✓ Taxonomic classification of bins using GTDB-Tk
- Obtain MSA from GTDB-Tk output
- Display phylogenetic tree using iTOL
- Annotate your tree

Virus taxonomy

DNA virus 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

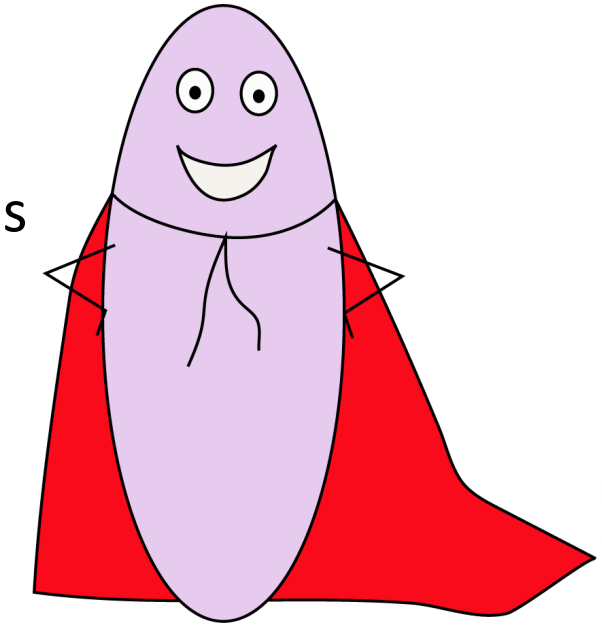
Aim:

- 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 1/10th 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/jdispatcher/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

- https://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

- https://genemark.bme.gatech.edu/
- GeneMark is a family of gene prediction tools
- Genomic sequences can be analysed either by the self-training program GeneMarkS-2 (sequences >50 kb) or using Heuristic Models by GeneMark.hmm
- 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)
  - Standard for prokaryotes and plant plastid
- If genes are too short, uses alternative code 4 (TGA not a stop codon)
  - Mycoplasma/Spiroplasma (also mold and protozoa)
- Will not try code 25 (must manually select), but will issue warning if genes are short
  - Candidate Phyla Radiation/Patescibacteria and Candidate Division SR1 (uncultured aquatic/host-associated microbiome)

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

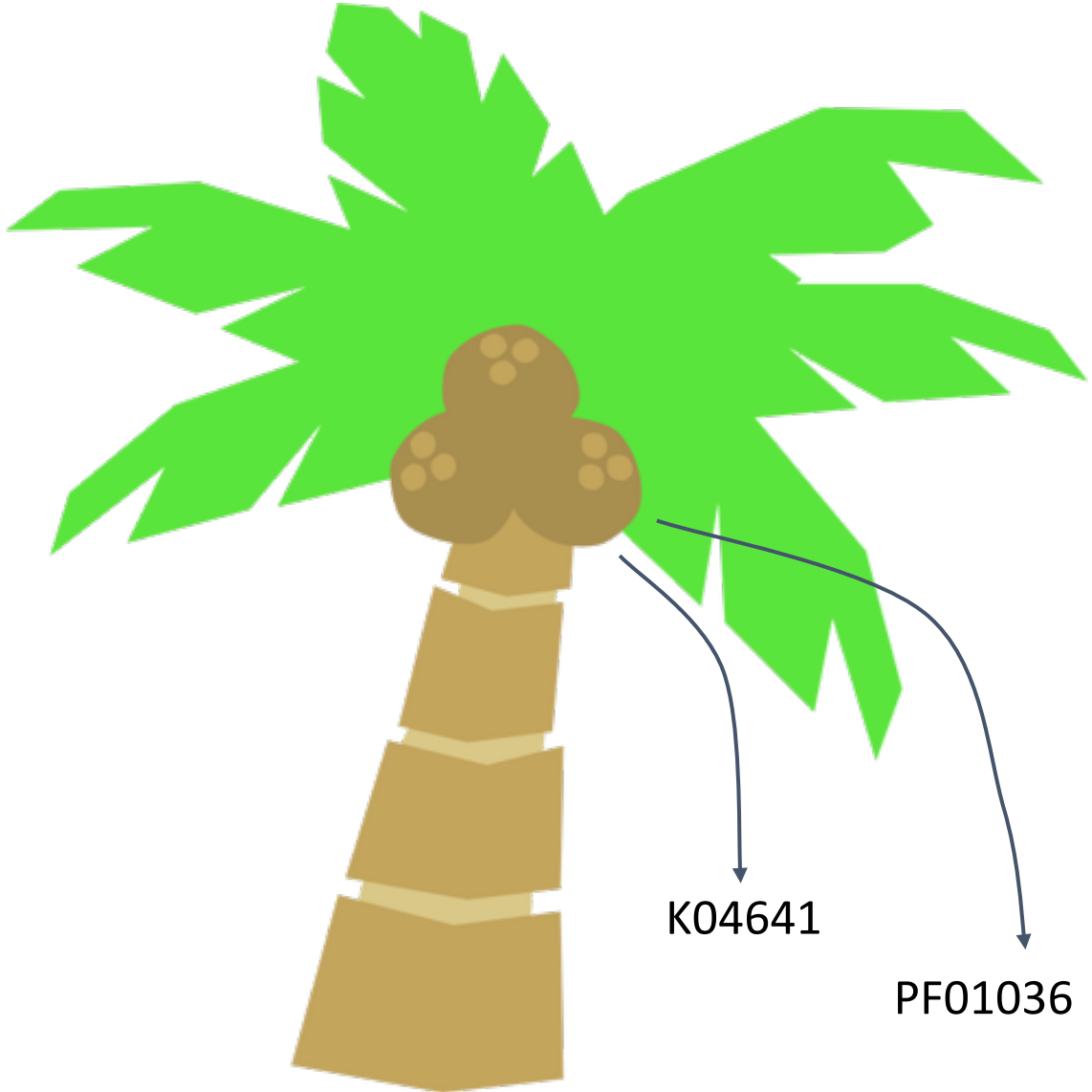
Task: Gene prediction

Go to Github MGSS webpage

Tasks:

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

Gene annotation (part 1)

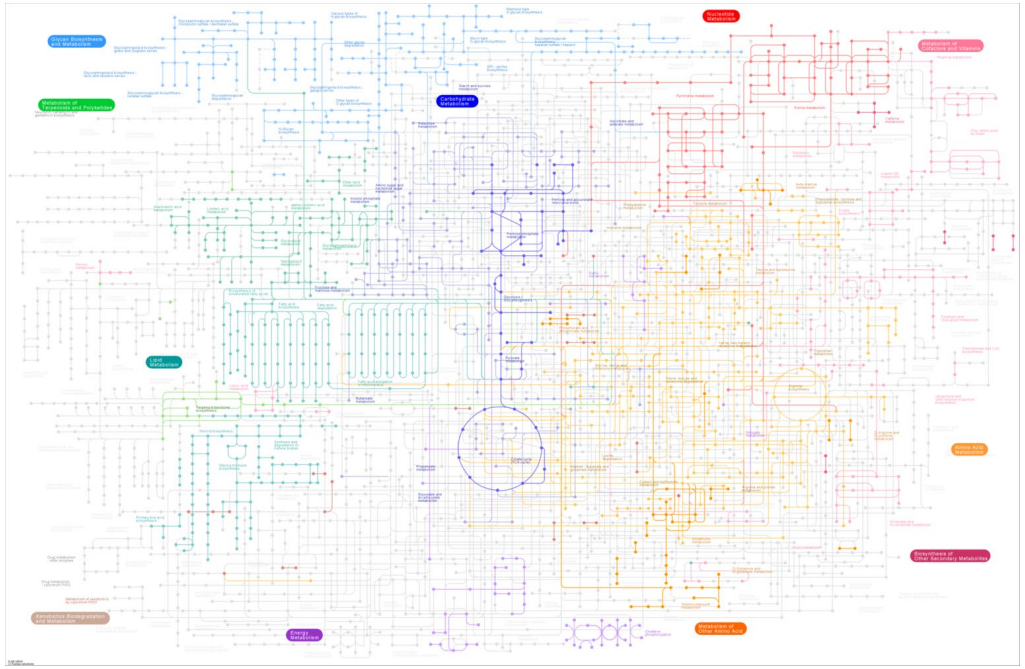


"You exist in the context of all in which you live and what came before you"

KH

Gene annotation

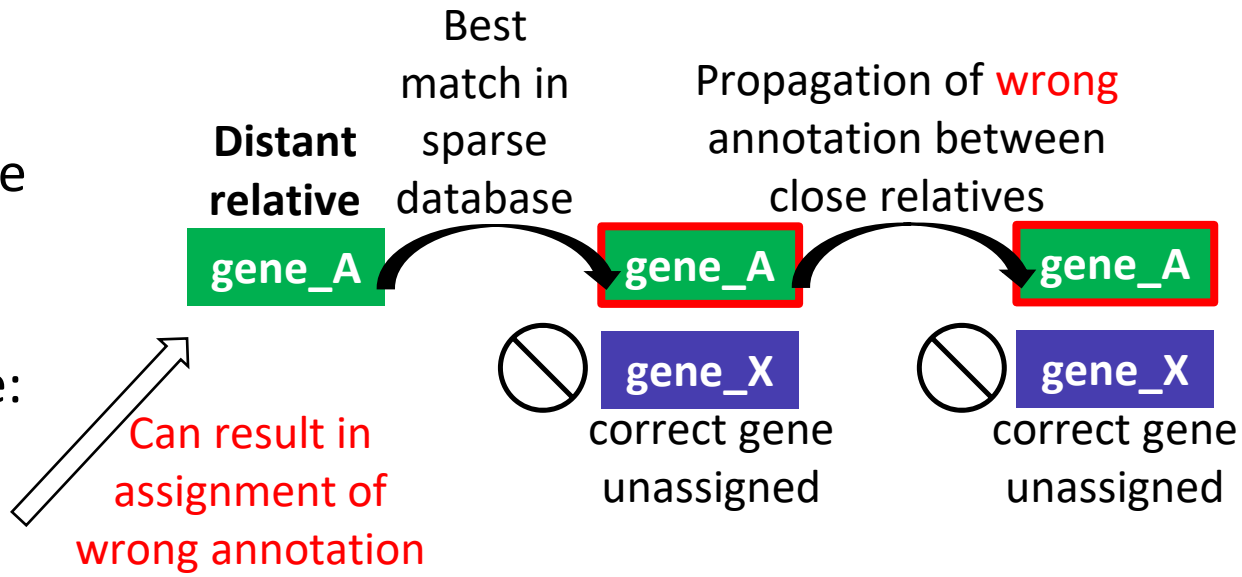
- Genome annotation attempts to predict gene function
- Predicted genes or protein sequences are compared against a curated set of reference sequences for which function is known, or is strongly suspected



Gene annotation

Caveat:

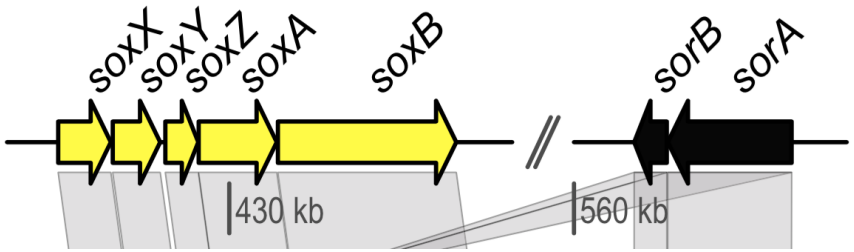
- Annotations are dependent on the reference database
- Environmental genomes can have:
  - Genes with distant homology matches to unrelated taxa
  - Large numbers of "hypothetical" gene annotations (= genes of unknown function)



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

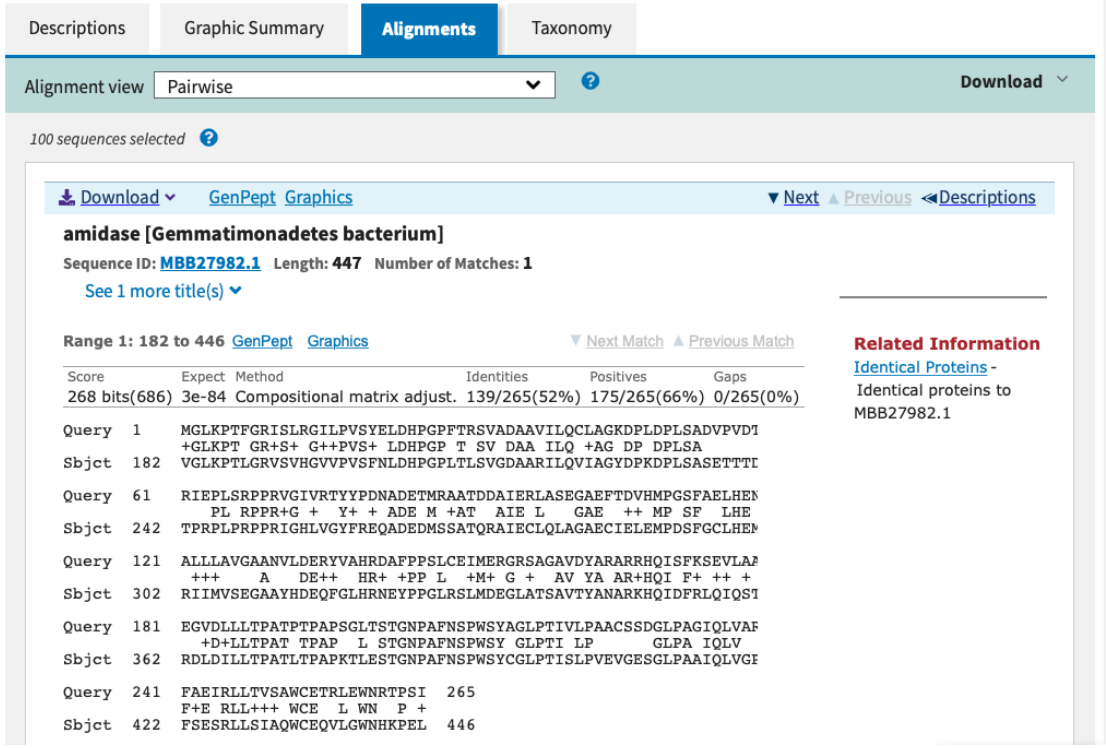
There are two main ways to perform gene annotation with protein sequences:

- 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 (https://www.ebi.ac.uk/interpro/)
    - Focused more on protein domains based on hidden Markov models
- TIGRfam (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

- 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 (**Uni**Prot **Ref**erence Clusters) and SwissProt (https://www.uniprot.org/)
    - Protein clustering at different levels (e.g. UniRef100, UniRef90, UniRef50)
    - SwissProt is manually curated and experimentally confirmed
- BioCyc (https://biocyc.org) and MetaCyc (https://metacyc.org) (Subscription required)
    - BioCyc contains organism specific pathway/genome Databases + software tools
    - MetaCyc contains pathways shared between multiple organisms

Gene annotation

Databases for specific functions

- dbCAN3 (https://bcb.unl.edu/dbCAN2/) [Carbohydrate-active enzymes, FASTA + HMM]
  - Web server for automated annotation with downloadable databases
- TCDB (https://tcdb.org/) [Transporters, FASTA + HMM]
  - Classification system with extensive literature summaries and specialist prediction tools
- MEROPS (https://www.ebi.ac.uk/merops/) [Peptidases & inhibitors, FASTA]
  - Also includes peptidase inhibitors
- CARD (https://card.mcmaster.ca/home) [Antimicrobial resistance, FASTA]
  - Extensive, curated, and updated AMR gene collection
    (Free for academic use, license required for commercial use)

Gene annotation

Interproscan (https://interproscan-docs.readthedocs.io/en/latest/index.html)
(Jones et al. 2014, Bioinformatics, 30(9))

- Protein family annotation tool mainly using HMMs
- Annotates across 14 HMM databases
- Uses curated thresholds of member databases for functional assignments
  - E-values reported but often redundant
- Unifies functions and signatures across multiple databases
- Also has additional analyses for protein localisation
  (user must possess license and database files)

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**
- **Signal peptide prediction using SignalP6**

Online resources and data analysis

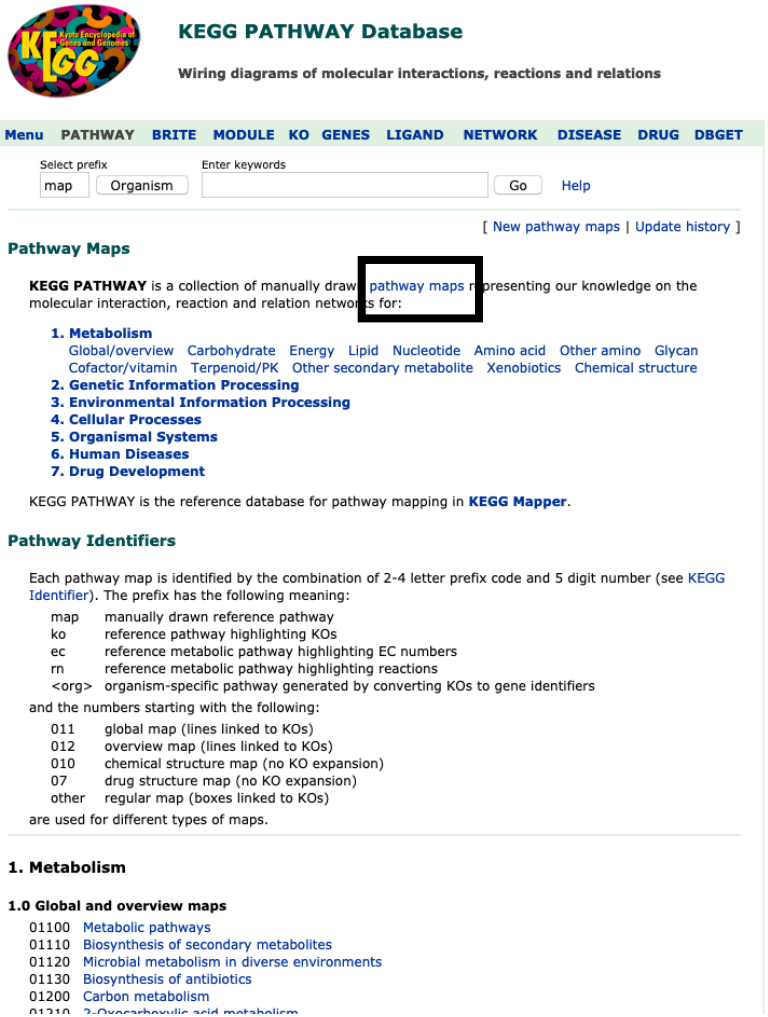
Gene annotation

Web-based annotation tools:

- Web BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)
- 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 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/kegg2.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:

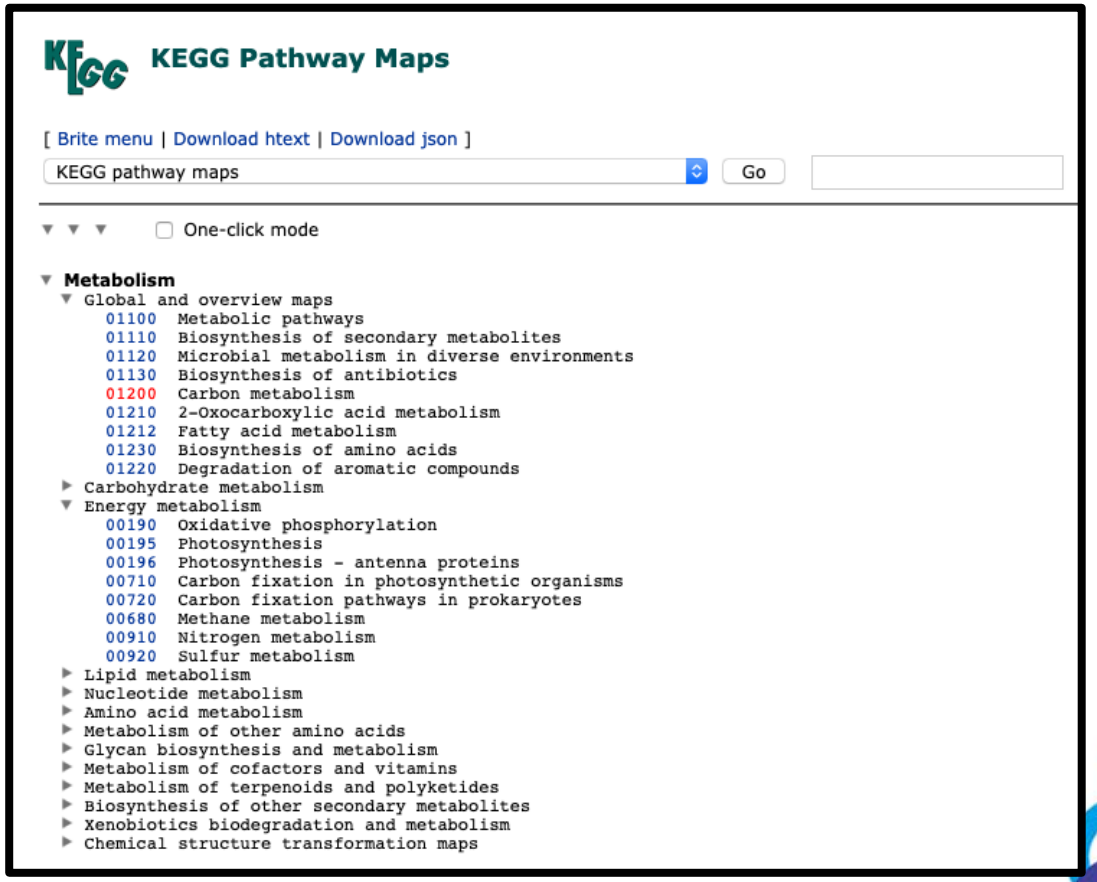
011	global map (lines linked to KOs)
012	overview map (lines linked to KOs)
010	chemical structure map (no KO expansion)
07	drug structure map (no KO expansion)
other	regular map (boxes linked to KOs)

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

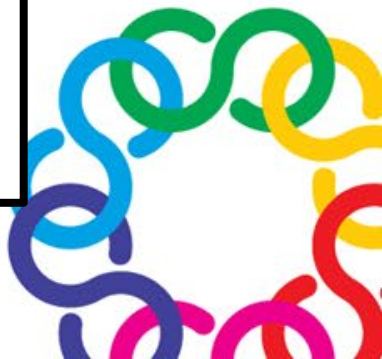
[Brite menu | Download htext | Download json]

KEGG pathway maps | Go

One-click mode

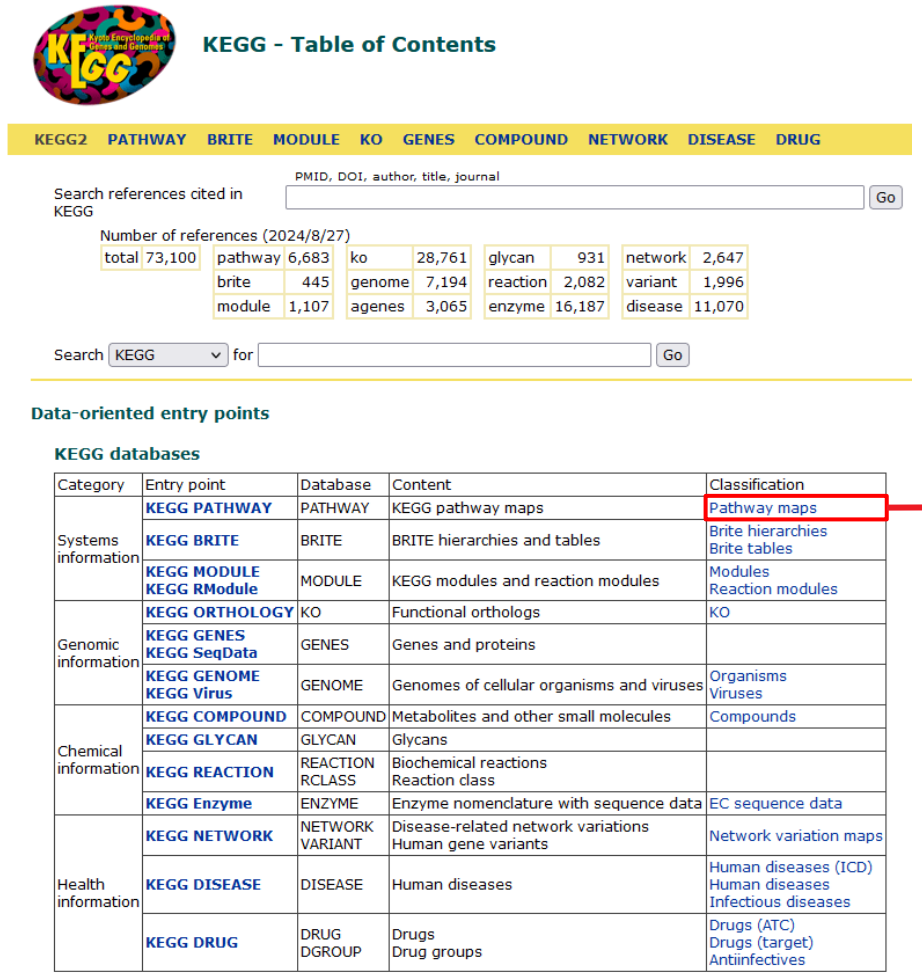
Metabolism

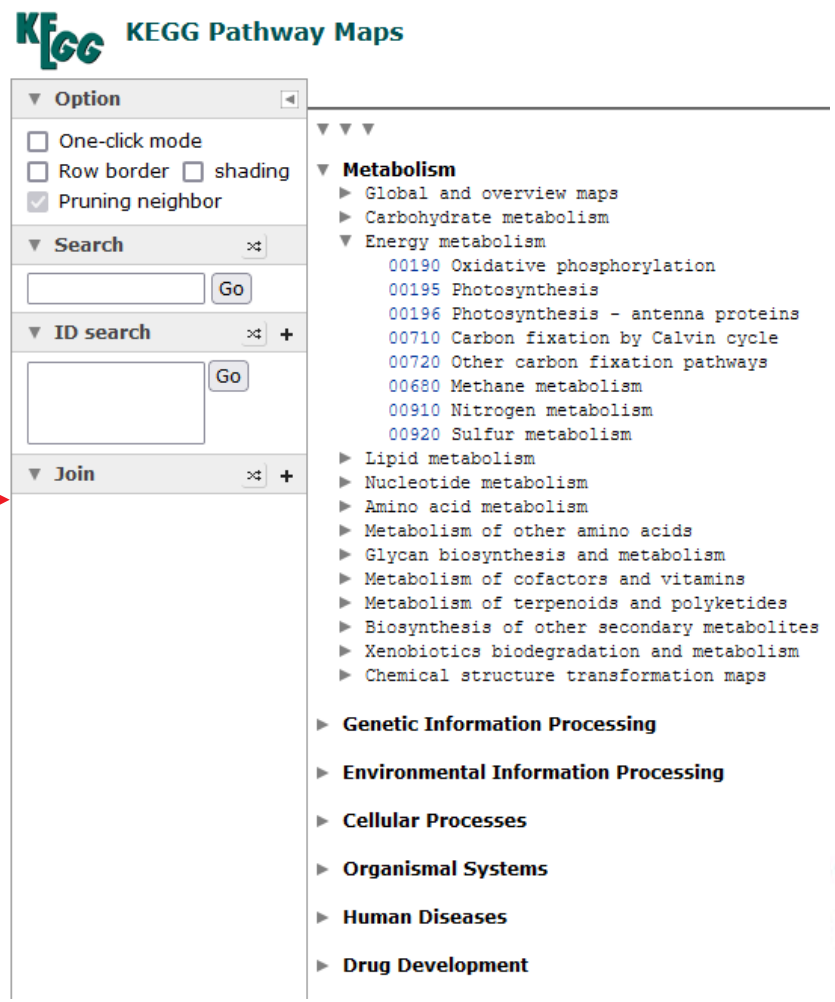
- Global and overview maps
  - 01100 Metabolic pathways
  - 01110 Biosynthesis of secondary metabolites
  - 01120 Microbial metabolism in diverse environments
  - 01130 Biosynthesis of antibiotics
  - 01200 Carbon metabolism
  - 01210 2-Oxocarboxylic acid metabolism
  - 01212 Fatty acid metabolism
  - 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

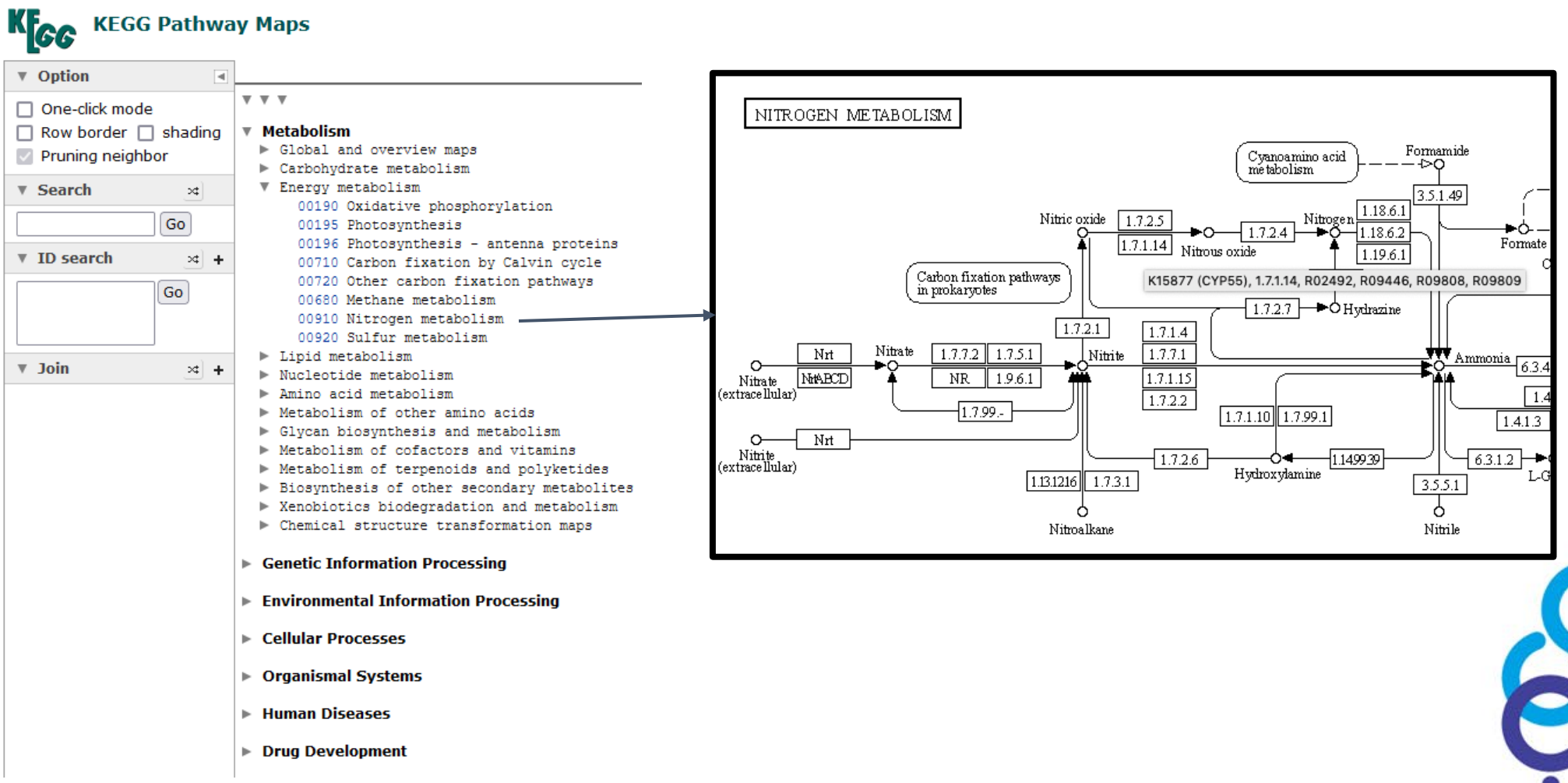
KEGG: https://www.genome.jp/kegg/kegg2.html





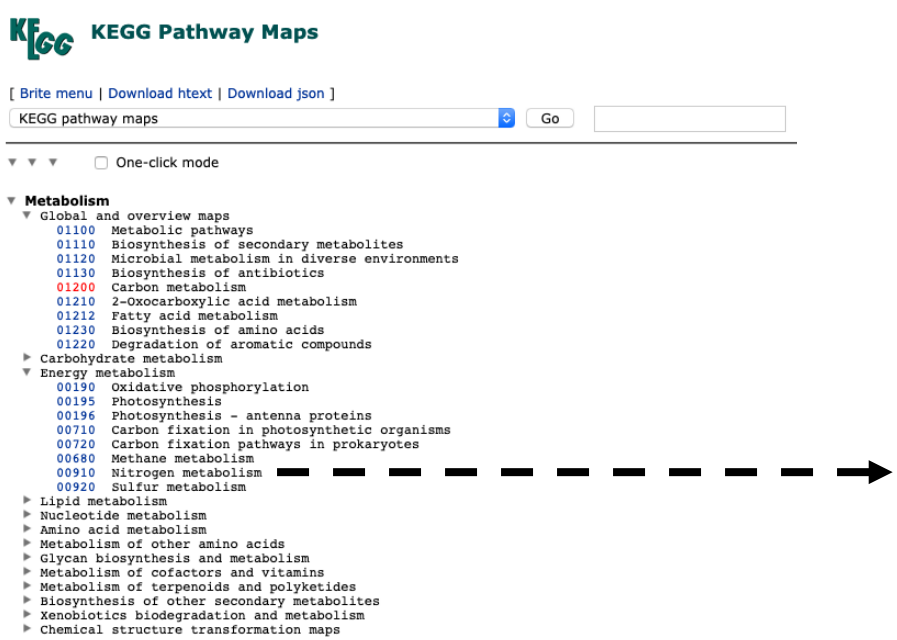
Gene annotation

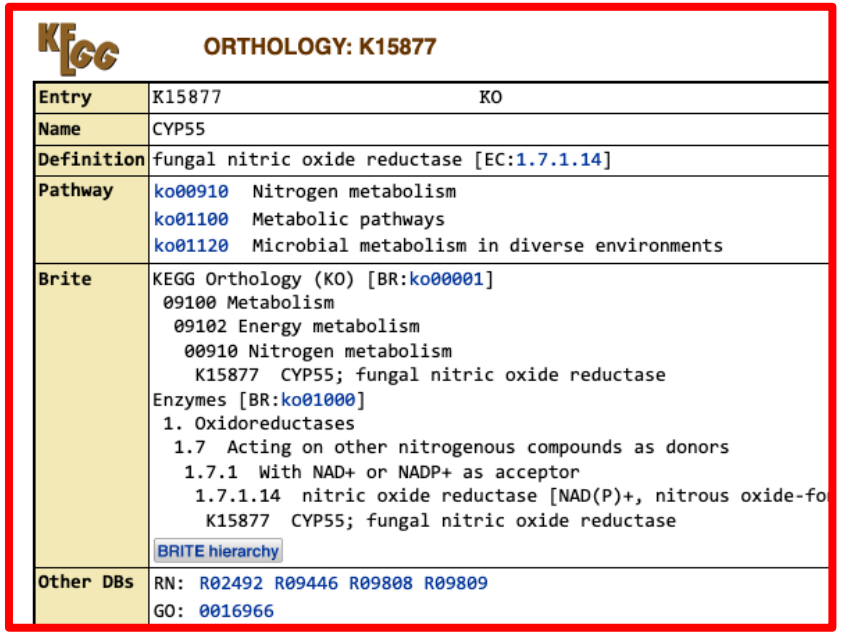
KEGG: https://www.genome.jp/kegg/pathway.html



Gene annotation

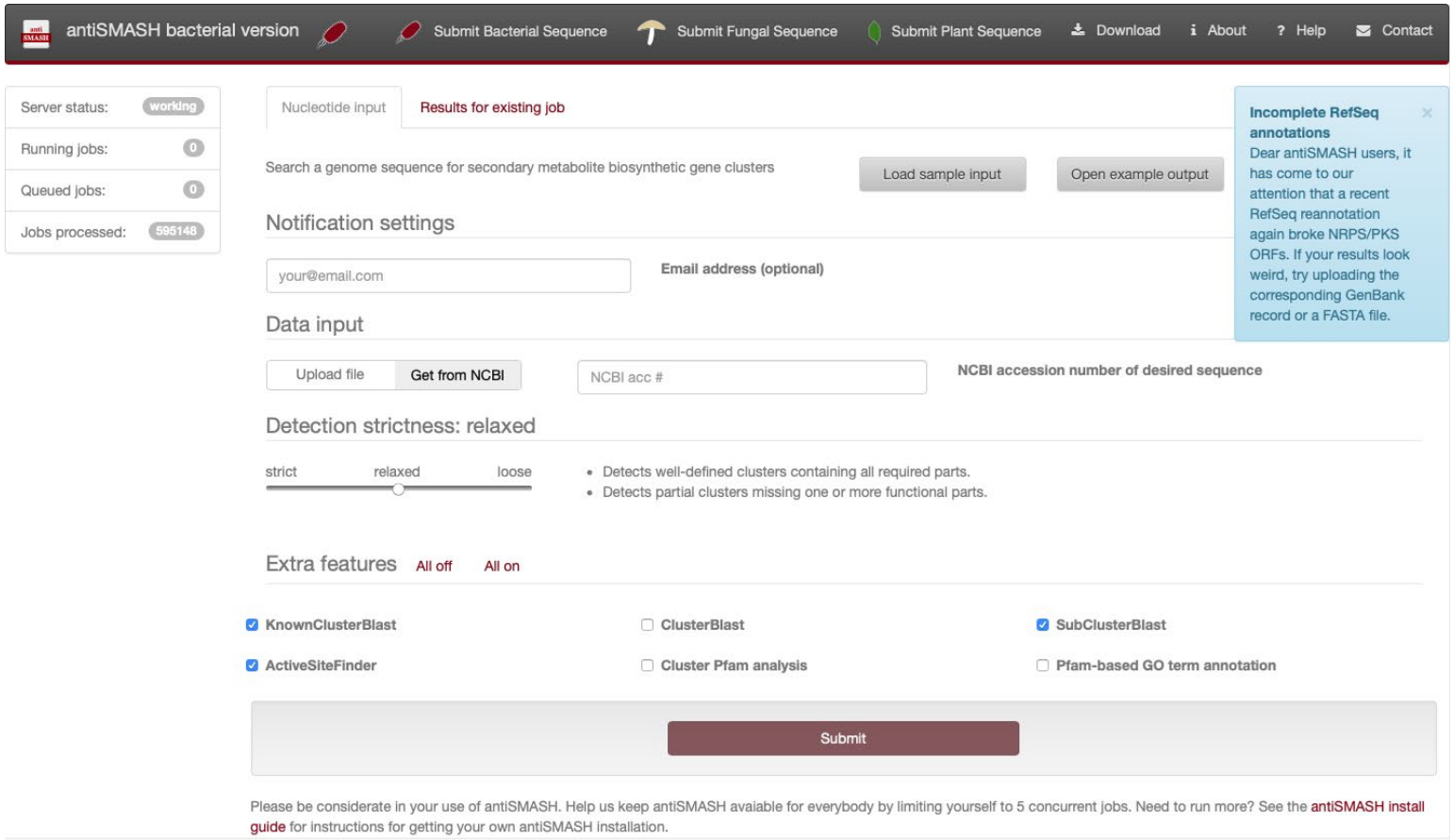
KEGG: https://www.genome.jp/kegg/pathway.html

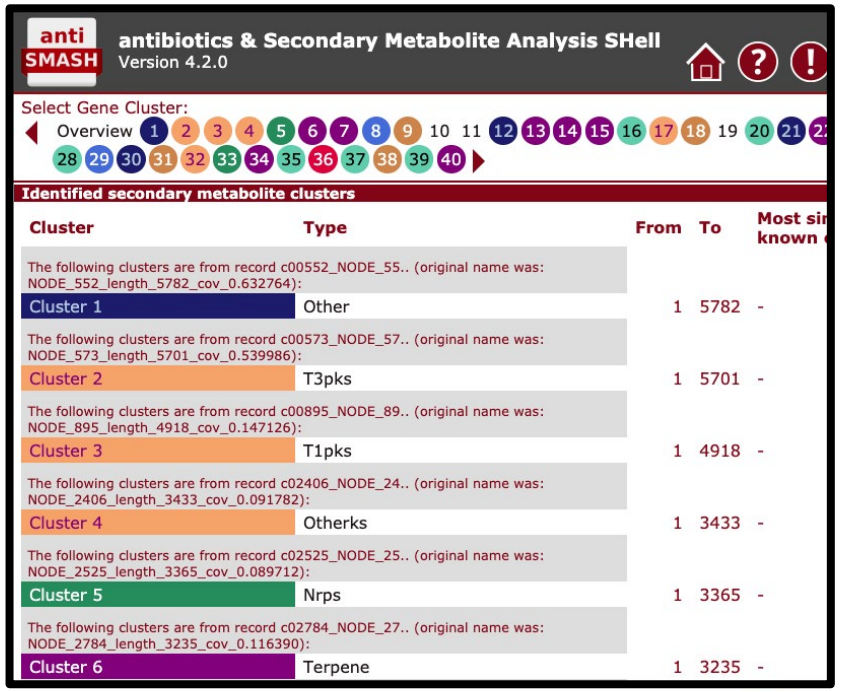




Gene annotation

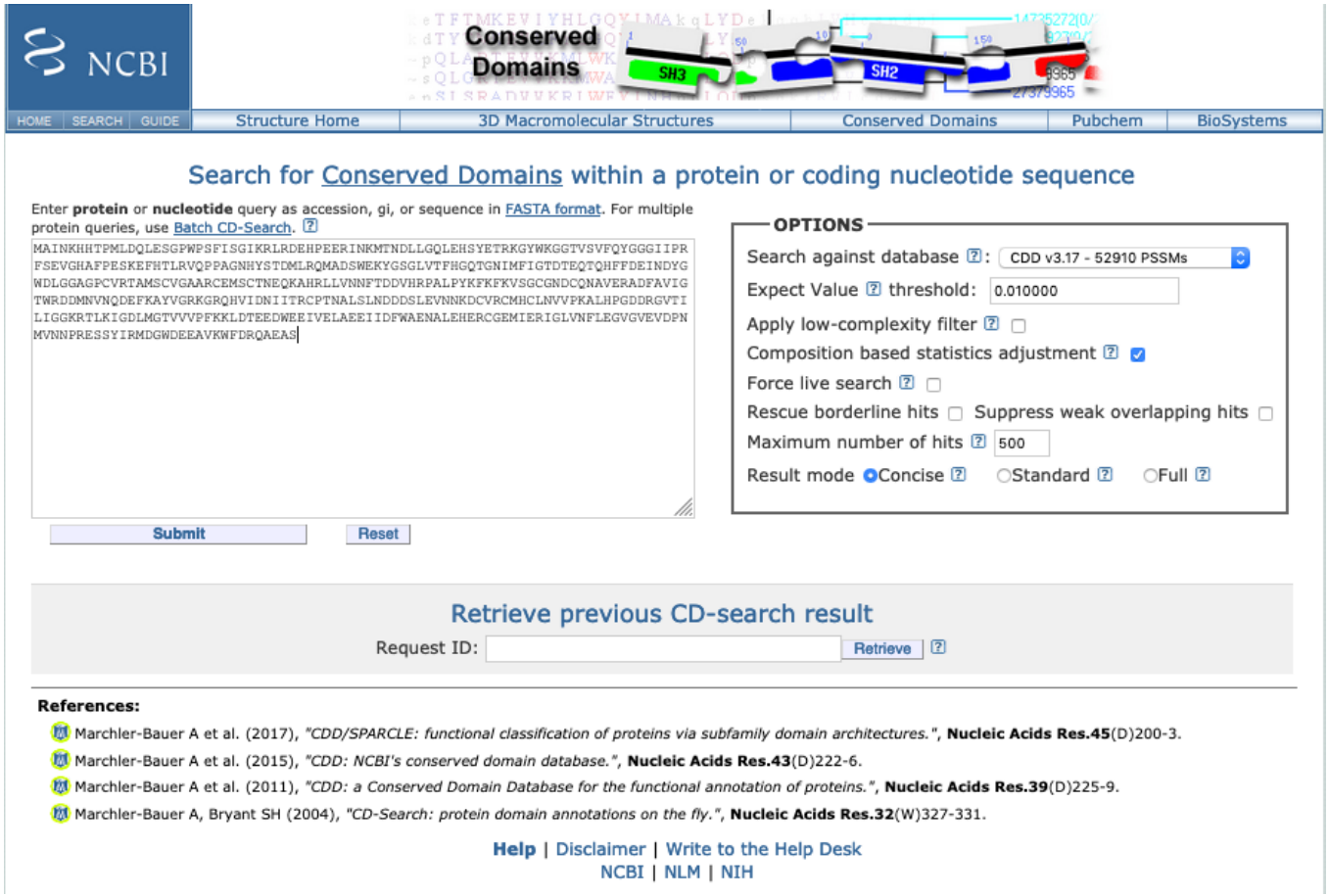
Identification of Biosynthetic Gene Clusters with **antiSMASH**





https://antismash.secondarymetabolites.org/

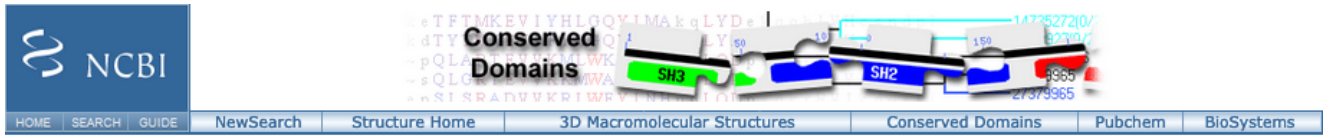
NCBI Conserved Domain Search



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

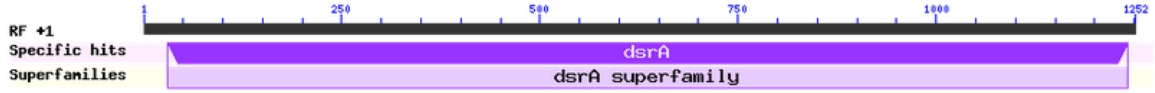


Conserved domains on [lcl|1]

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

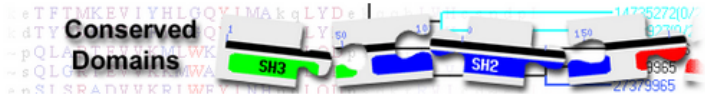
```
                    10        20        30        40        50        60        70        80
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1              11 LDQLESGPWPSFISGIKRLRDEHPEERINKMTNDLLGQLEHSYETRKGYWKGGTVSVFQYGGGIIPRFSEVGHAFPESKE 90
Cdd:TIGR02064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

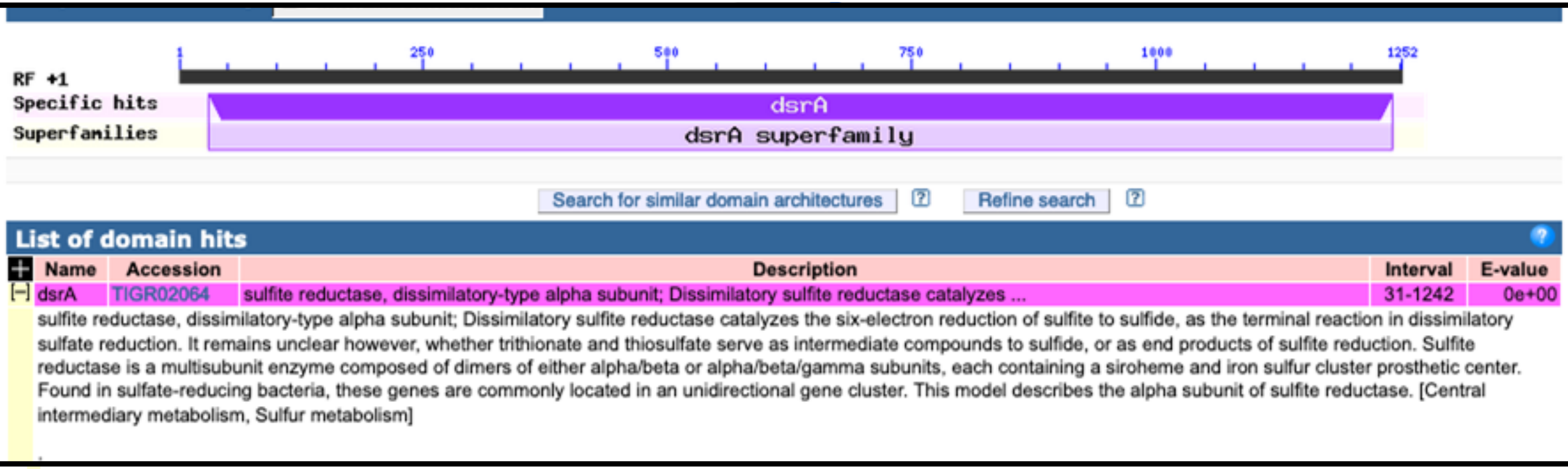
References:

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





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]

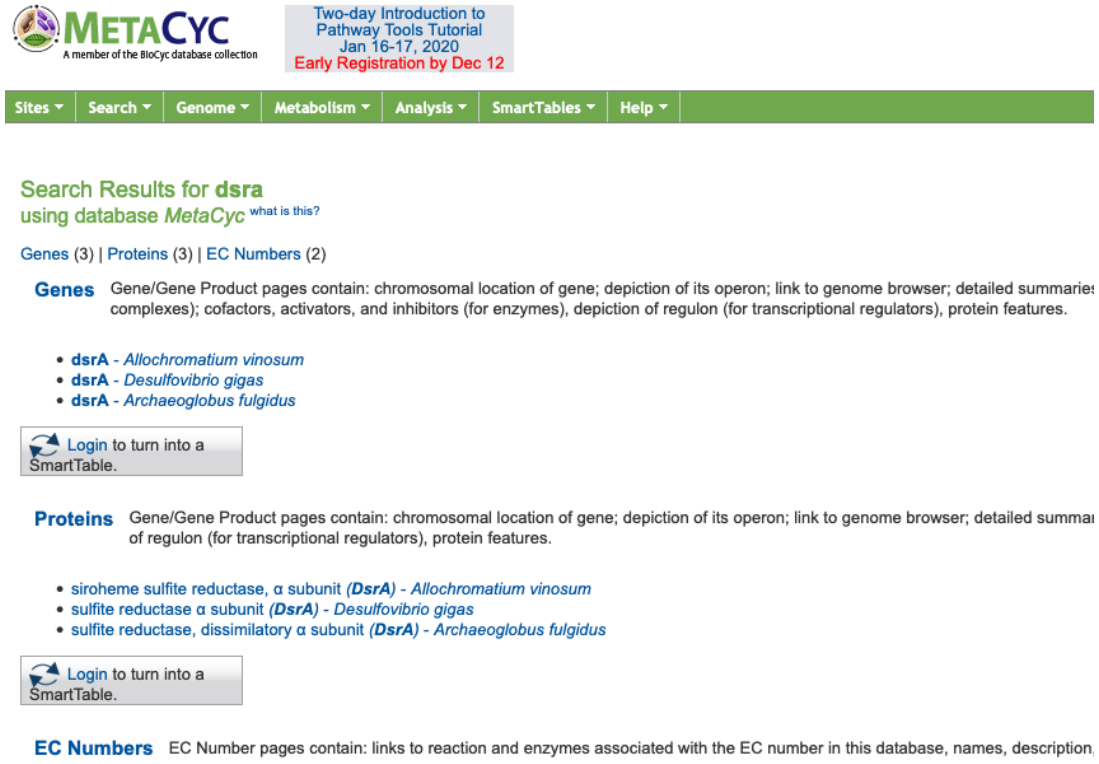
```
                    170       180       190       200       210       220       230       240
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1           171 VGAARCEMSCTNEQKAHRLLVNNFTDDVHRPALPYKFKFKVSGCGNDCQNAVERADFAVIGTWRDDMNVNQDEFKAYVGR 250
Cdd:TIGR02064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

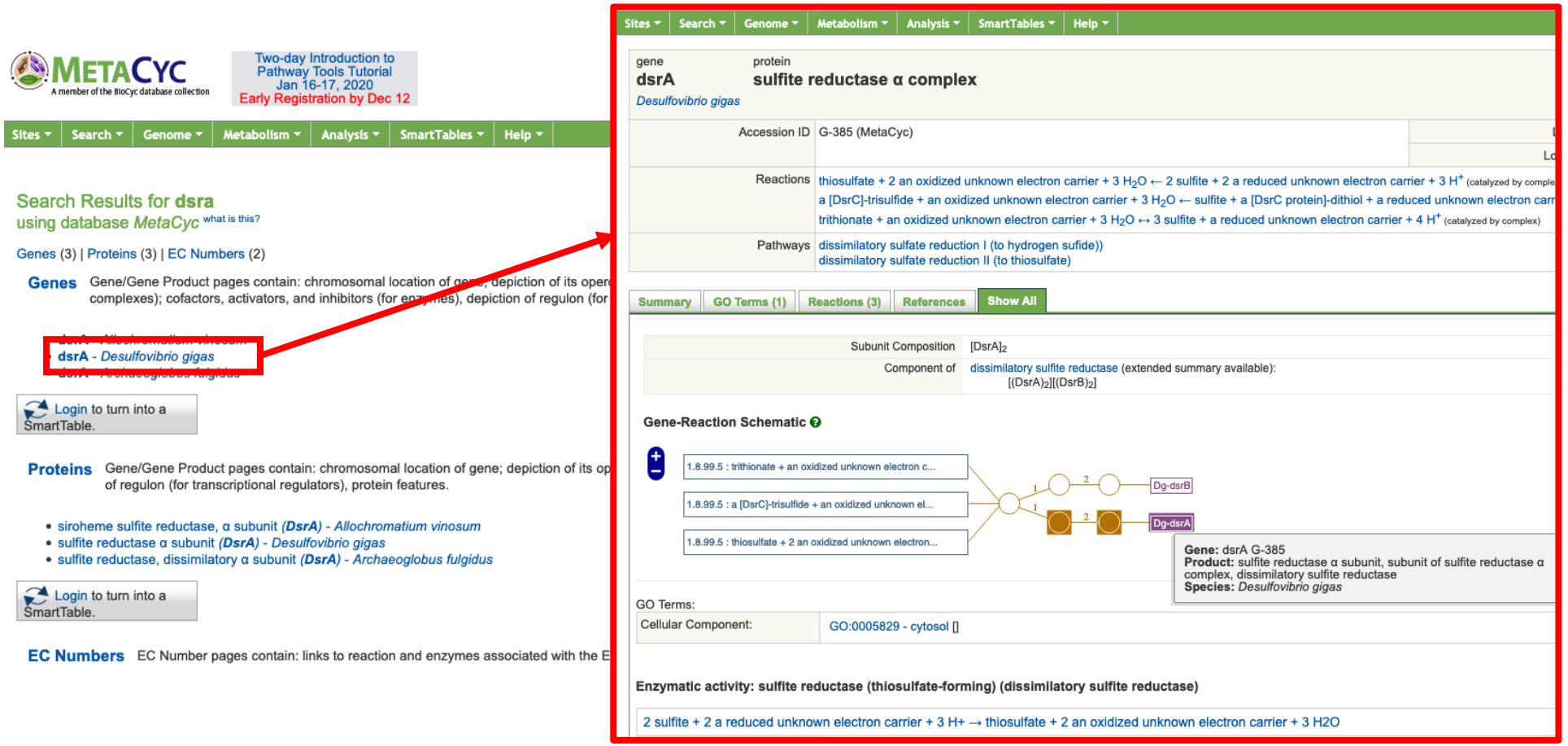
Metacyc: experimentally curated metabolic pathways



https://metacyc.org/

Gene annotation

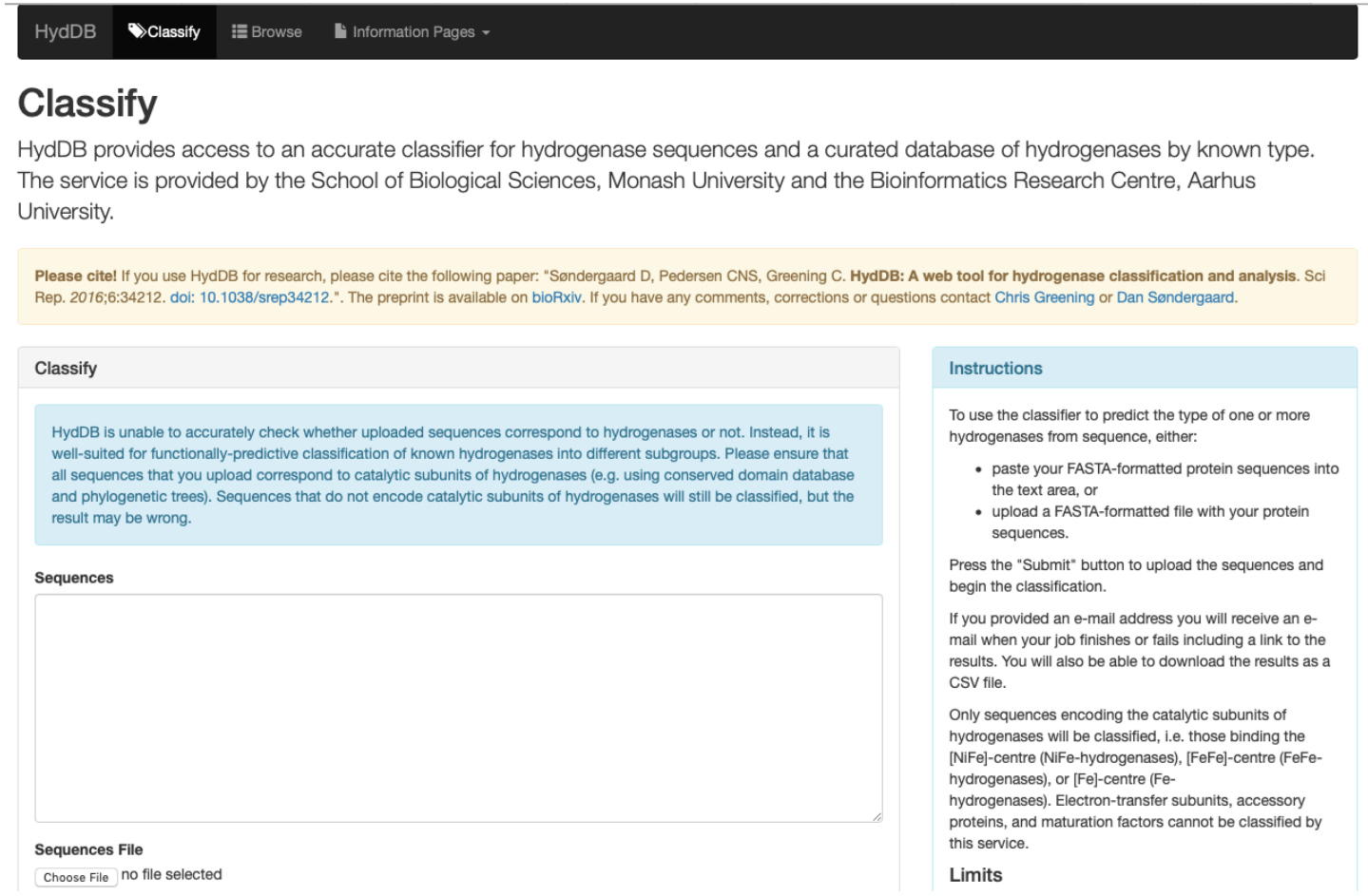
Metacyc: experimentally curated metabolic pathways



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

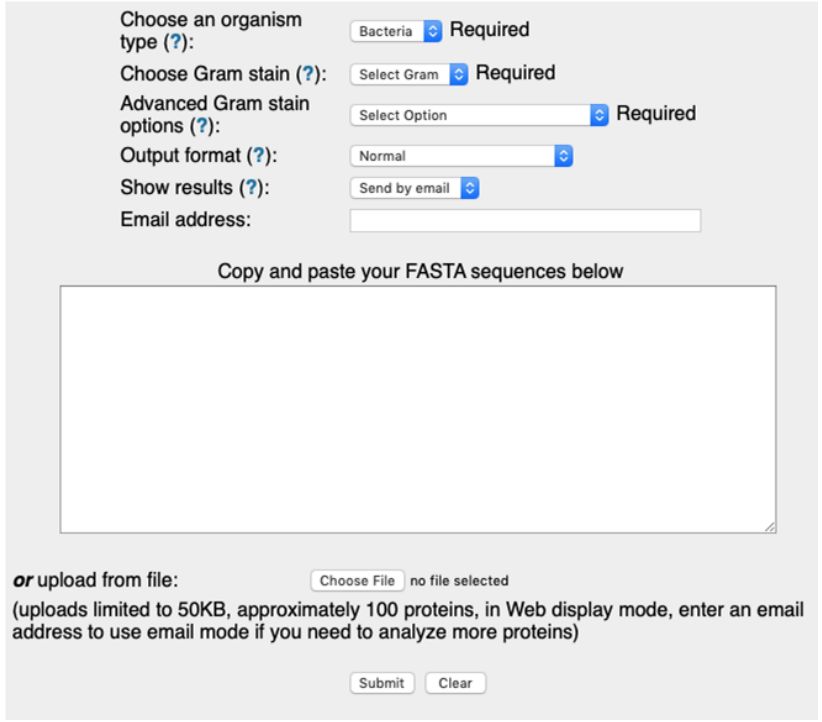
Submit a Sequence to PSORTb version 3.0.2

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.

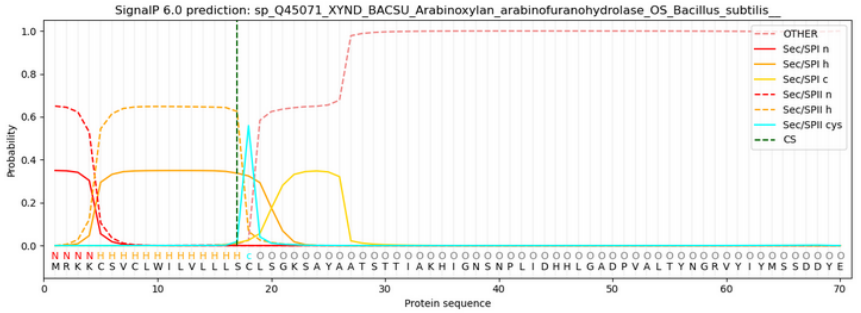
Submit Clear fields

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



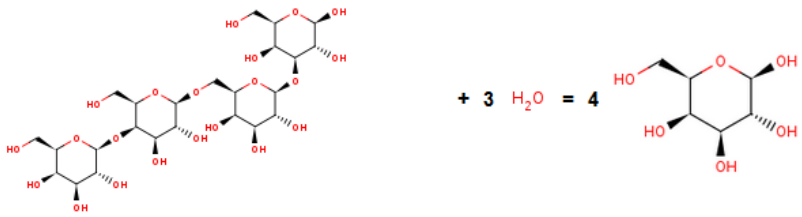
Gene annotation

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/
- **ANNOTREE:** http://annotree.uwaterloo.ca/annotree/

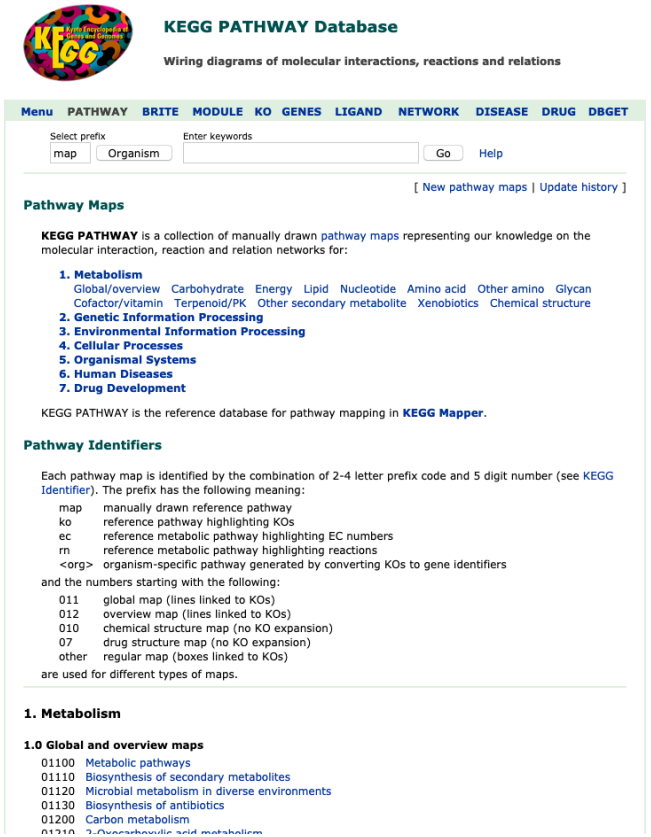
Task: Gene annotation

Tasks:

- **View KEGG annotation in KEGG website**

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

- View functional distribution in ANNOTREE



Science for Communities

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HIKINA WHAKATUTUKI

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF BIOLOGICAL SCIENCES

Delving into plastisphere metagenomes: The joys of the unassembled metagenomic analyses

Jessica Wallbank

NeSI
New Zealand eScience Infrastructure

AIM²
Aotearoa Impacts & Mitigation of Microplastics

Outline of my project

- Multi-omics approach
  - Amplicon data
  - Metagenomic data
- 7 substrate types

(5 plastics, glass and seawater)

- 3 timepoints

(3, 9 and 12 months)

- 106 samples in total

Amplicon data revealed potential ‘plastic degrading microbes’ were present in very low abundances.

Aims of my project

- Compare taxonomic assignments from metagenomic data to amplicon data
- Investigate the broad-scale functional potential within microbial communities
- Determine whether core functions were conserved regardless of substrate/timepoint
- Identify potential genes conferring plastic degradation

Methods

Sample collection and DNA extractions

Shotgun sequencing

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Taxonomic assignment

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Metaxa2

- Unassembled reads

- Identifies and extracts partial rRNA sequences

- small subunit (SSU; 16S/18S) rRNA gene
- large subunit (LSU; 23S/28S) rRNA gene

- Assigns taxonomy to archaeal, bacterial, eukaryotic, mitochondrial or chloroplast origin

Assembled but not binned

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Assembly

metaSPAdes

Assembled but not binned

Sample collection and DNA extractions

Shotgun sequencing

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Taxonomic assignment using Metaxa2

Assembly

metaSPAdes

hyattpd/**Prodigal**

Prediction of open reading frames (ORFs)

diamond

.daa

Annotating sequences by aligning contigs with reference databases

Assembled but not binned

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Assembly

metaSPAdes

hyattpd/**Prodigal**

Prediction of open reading frames (ORFs)

diamond

.daa

Annotating sequences by aligning contigs with reference databases

Read mapping

BBMap / Bowtie

Count data

featureCounts

Unassembled

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Prediction of open reading frames (ORFs)

gatech-genemark/ **MetaGeneMark-2**

hyattpd/**Prodigal**

gaberoo/ **FragGeneScan**

Unassembled

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Annotating sequences by aligning them with reference databases

diamond

.daa

Prediction of open reading frames (ORFs)

gatech-genemark/ **MetaGeneMark-2**

hyattpd/**Prodigal**

gaberoo/ **FragGeneScan**

Unassembled

Taxonomic assignment using Metaxa2

Sample collection and DNA extractions

Quality checking, read trimming and adapter removal

FastQC

.fastq .gz

usadellab/ **Trimmomatic**

Shotgun sequencing

Annotating sequences by aligning them with reference databases

diamond

.daa

MEGANIZER

MEGAN6 community edition

Prediction of open reading frames (ORFs)

gatech-genemark/ **MetaGeneMark-2**

hyattpd/**Prodigal**

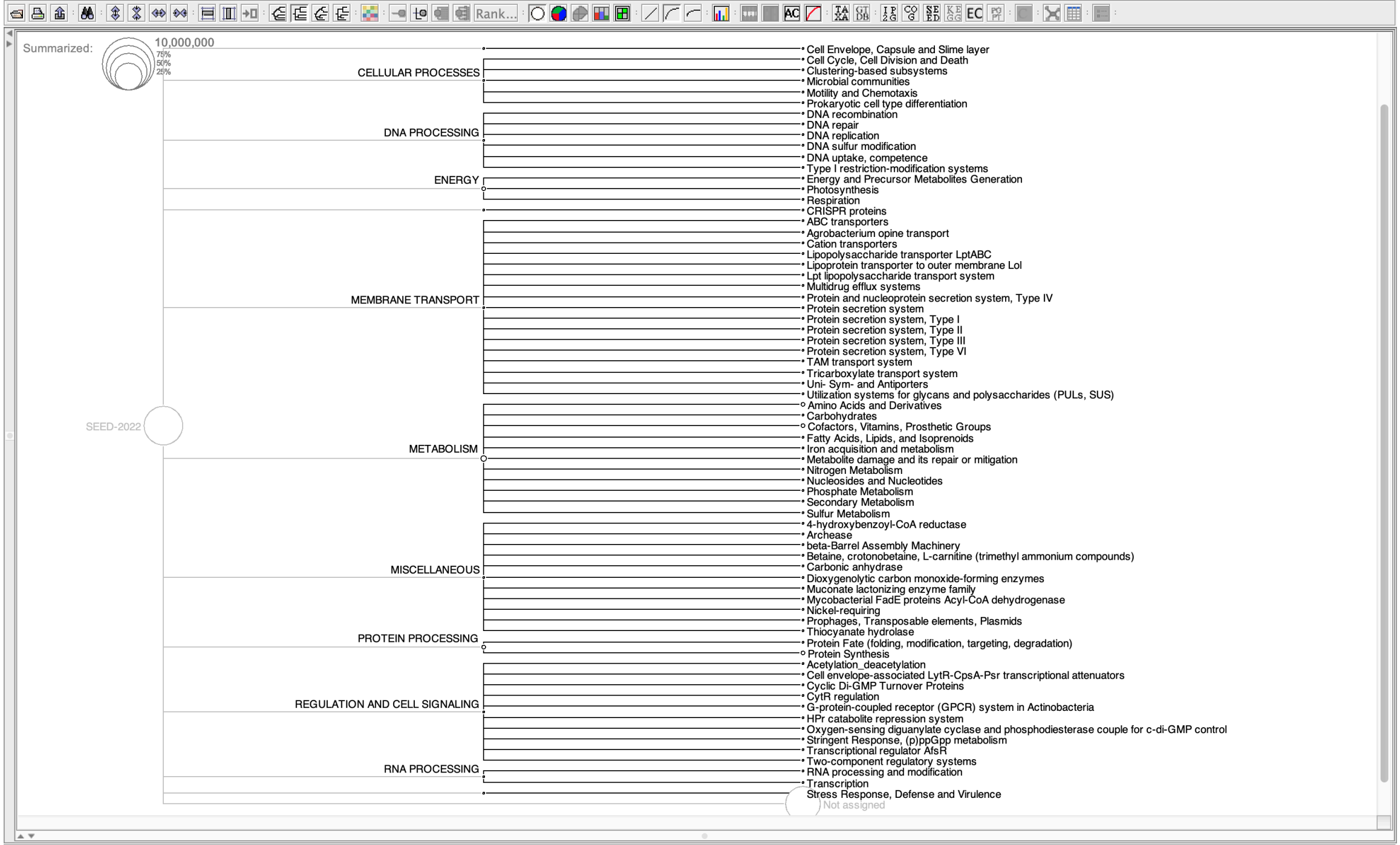
gaberoo/ **FragGeneScan**

MEGAN6

- Software for interactive metagenomics analyses
- Specialised for unassembled reads
- Command-line tools
- Taxonomic analyses – NCBI
- Functional analyses - SEED, InterPro2GO, eggNOG

Community edition is free

Ultimate edition requires a paid subscription but includes up-to-date KEGG pathways


Summarized:
10,000,000
75%
50%
25%
SEED-2022
CELLULAR PROCESSES
Cell Envelope, Capsule and Slime layer
Cell Cycle, Cell Division and Death
Clustering-based subsystems
Microbial communities
Motility and Chemotaxis
Prokaryotic cell type differentiation
DNA PROCESSING
DNA recombination
DNA repair
DNA replication
DNA sulfur modification
DNA uptake, competence
Type I restriction-modification systems
ENERGY
Energy and Precursor Metabolites Generation
Photosynthesis
Respiration
CRISPR proteins
MEMBRANE TRANSPORT
ABC transporters
Agrobacterium opine transport
Cation transporters
Lipopolysaccharide transporter LptABC
Lipoprotein transporter to outer membrane Lol
Lpt lipopolysaccharide transport system
Multidrug efflux systems
Protein and nucleoprotein secretion system, Type IV
Protein secretion system
Protein secretion system, Type I
Protein secretion system, Type II
Protein secretion system, Type III
Protein secretion system, Type VI
TAM transport system
Tricarboxylate transport system
Uni- Sym- and Antiporters
Utilization systems for glycans and polysaccharides (PULs, SUS)
METABOLISM
Amino Acids and Derivatives
Carbohydrates
Cofactors, Vitamins, Prosthetic Groups
Fatty Acids, Lipids, and Isoprenoids
Iron acquisition and metabolism
Metabolite damage and its repair or mitigation
Nitrogen Metabolism
Nucleosides and Nucleotides
Phosphate Metabolism
Secondary Metabolism
Sulfur Metabolism
MISCELLANEOUS
4-hydroxybenzoyl-CoA reductase
Archease
beta-Barrel Assembly Machinery
Betaine, crotonobetaine, L-carnitine (trimethyl ammonium compounds)
Carbonic anhydrase
Dioxygenolytic carbon monoxide-forming enzymes
Muconate lactonizing enzyme family
Mycobacterial FadE proteins Acyl-CoA dehydrogenase
Nickel-requiring
Prophages, Transposable elements, Plasmids
Thiocyanate hydrolase
PROTEIN PROCESSING
Protein Fate (folding, modification, targeting, degradation)
Protein Synthesis
REGULATION AND CELL SIGNALING
Acetylation_deacetylation
Cell envelope-associated LytR-CpsA-Psr transcriptional attenuators
Cyclic Di-GMP Turnover Proteins
CytR regulation
G-protein-coupled receptor (GPCR) system in Actinobacteria
HPr catabolite repression system
Oxygen-sensing diguanylate cyclase and phosphodiesterase couple for c-di-GMP control
Stringent Response, (p)ppGpp metabolism
Transcriptional regulator AfsR
Two-component regulatory systems
RNA PROCESSING
RNA processing and modification
Transcription
Stress Response, Defense and Virulence
Not assigned

Visualisation of broad-scale functioning (SEED)

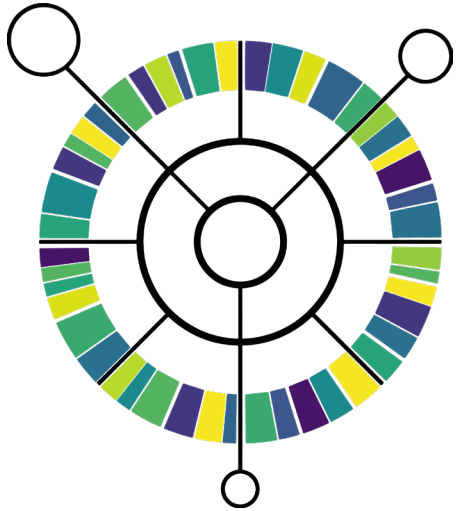
Count tables were normalised before Bray-Curtis dissimilarity matrices were calculated.

Specialised gene annotations

Plastic DB

Item	Update
Microorganisms	753
Proteins	219
Last update	13/05/2024
Last addition	Protein 00219

CARD
(Comprehensive
Antibiotic Resistance
Database)



Custom diamond databases can be made and used with diamond.

Identification of genes associated with plastic degradation

Item	Update
Microorganisms	753
Proteins	219
Last update	13/05/2024
Last addition	Protein 00219

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**