



genomics
aotearoa

Metagenomics
Summer School 2023

Day 3

Taxonomic classification &
phylogenetic inference

Viruses

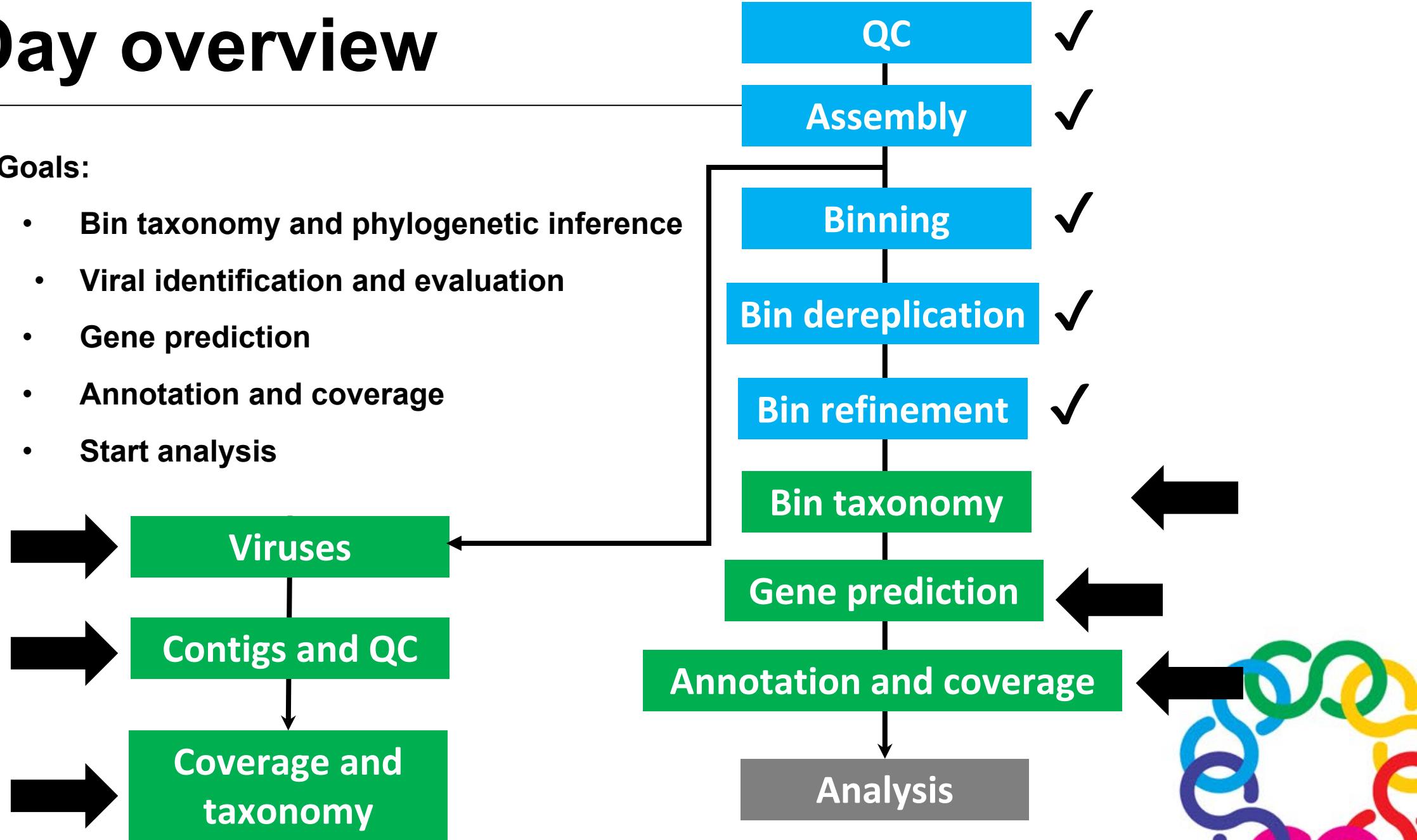
Gene prediction

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

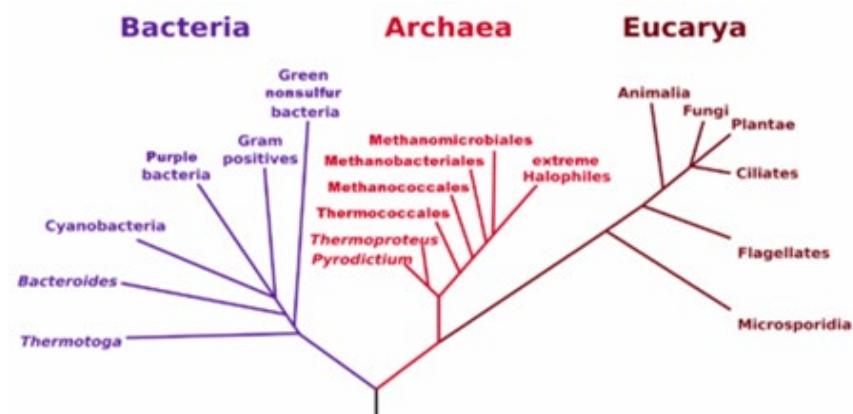
“...experience in many fields of science has demonstrated that naming is essential for accurate communication. Thus, astronomers have developed a system for naming stars, chemists a system for naming compounds, and biochemists a system for naming enzymes. All these fields share a common motivation, since 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

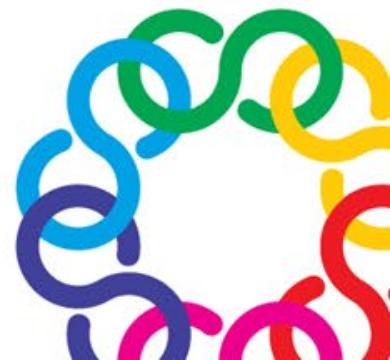


Woese et al. (1990) Proc Natl Acad Sci USA 87: 4576

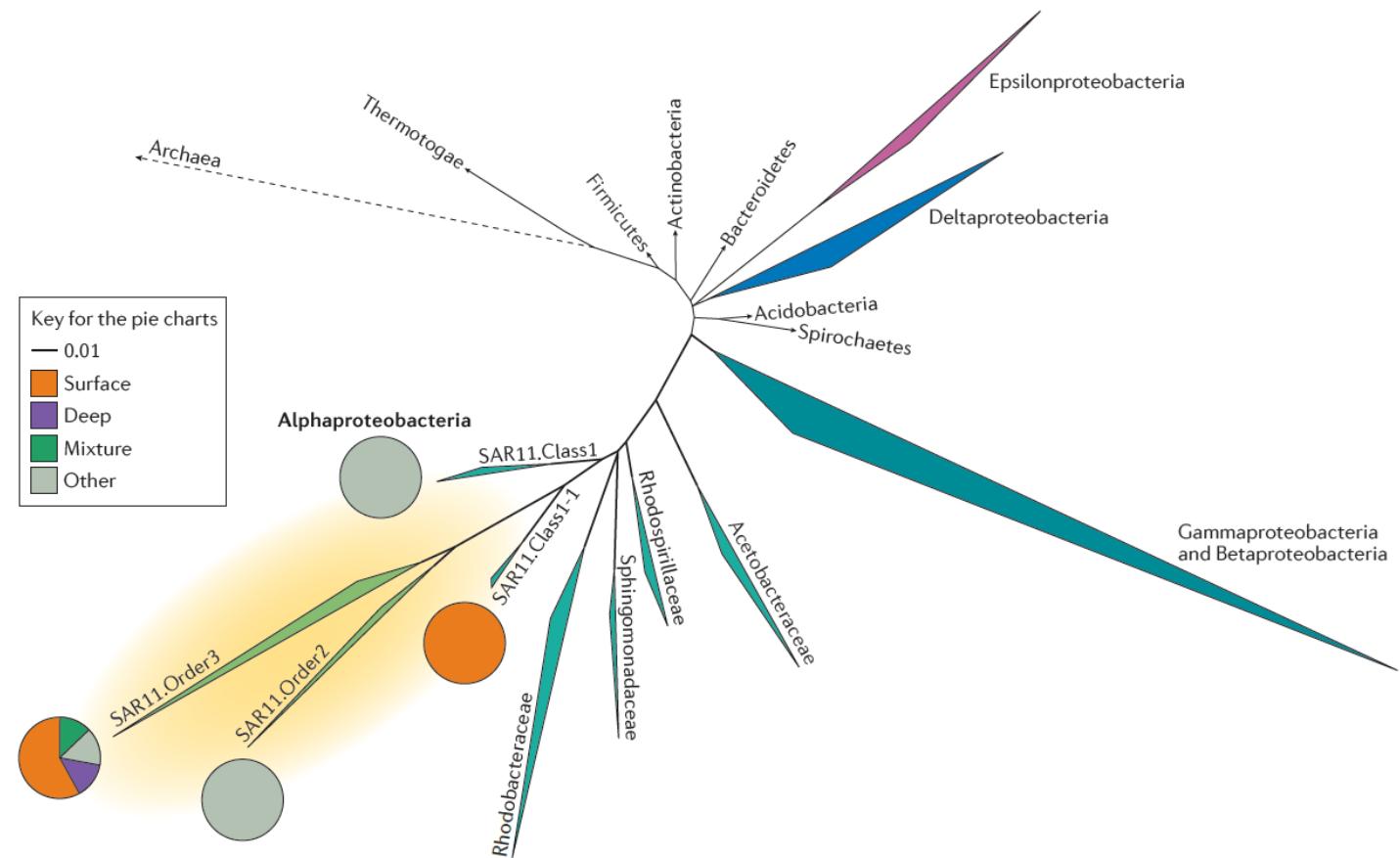


Bin taxonomic classification

- 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

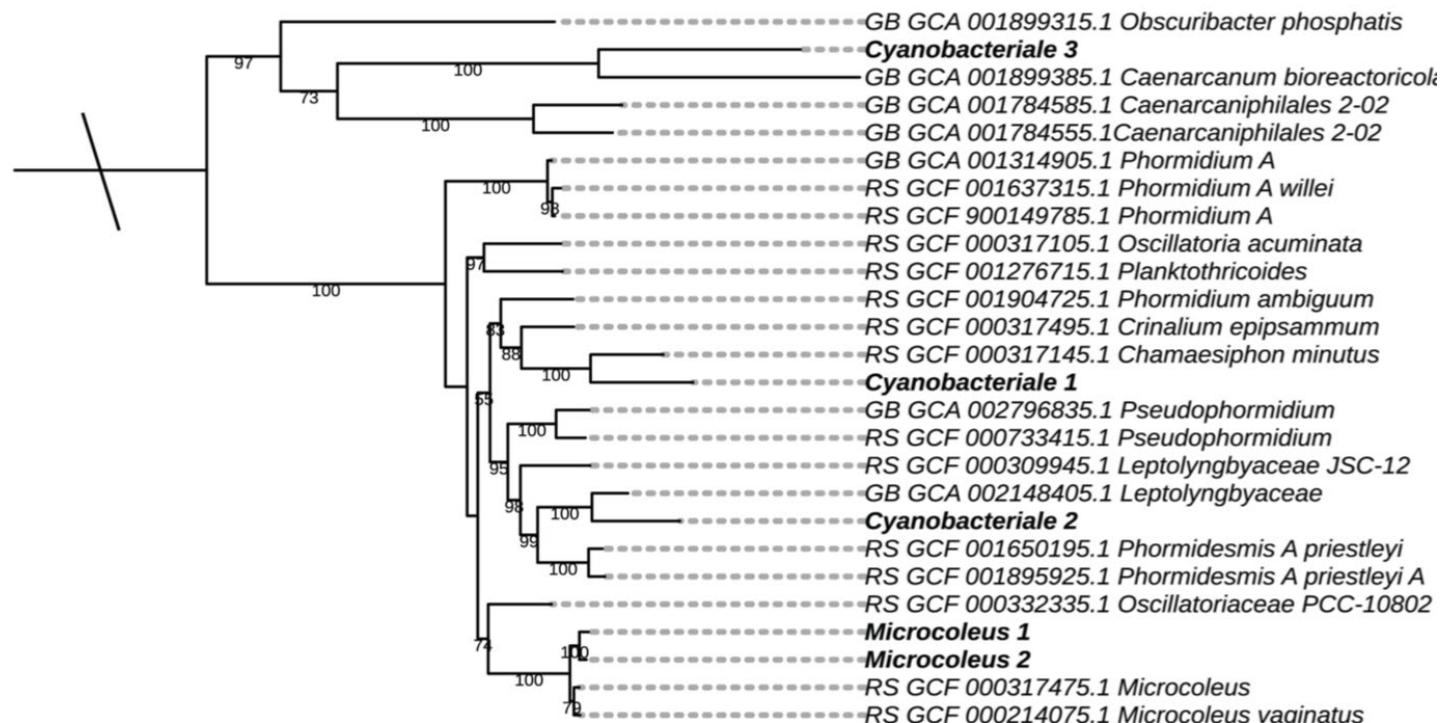


Bin taxonomic classification

Solution:

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

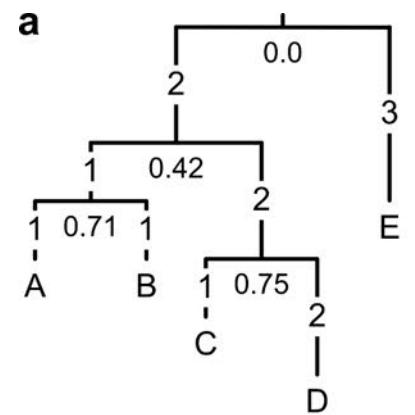
Tree scale: 0.1



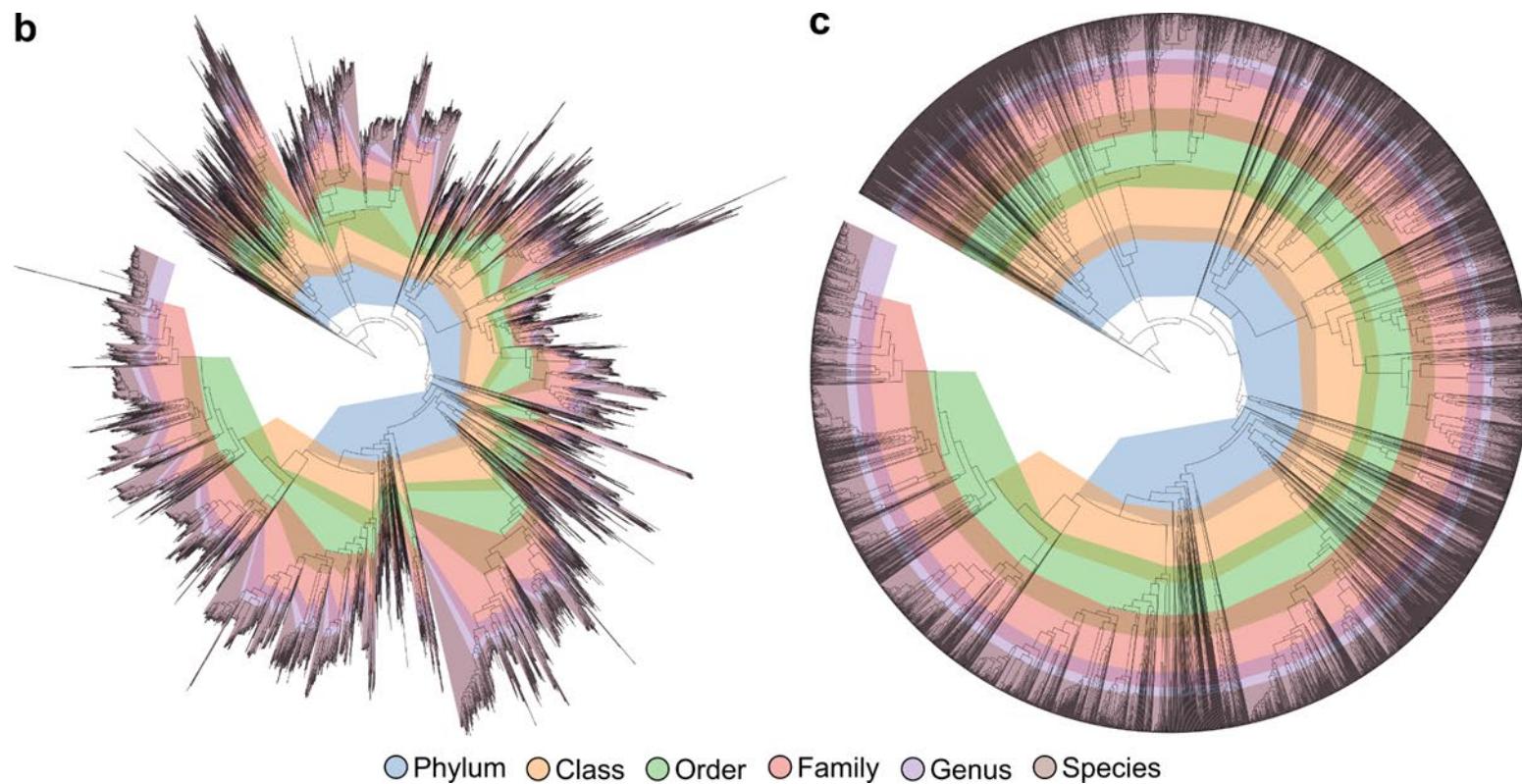
Concatenated protein sequence tree:
Phylogenetic placement of cyanobacterial genome bins
(Wai-iti River, Nelson)



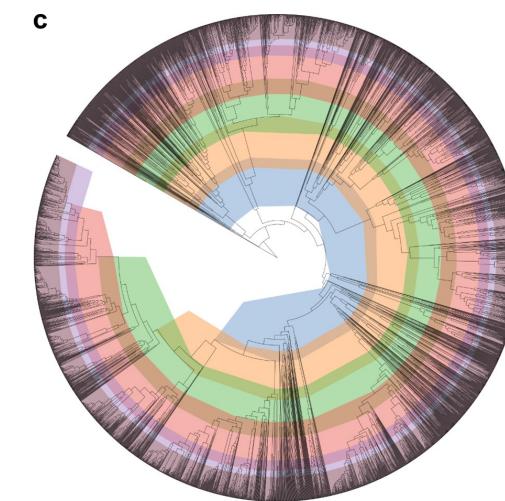
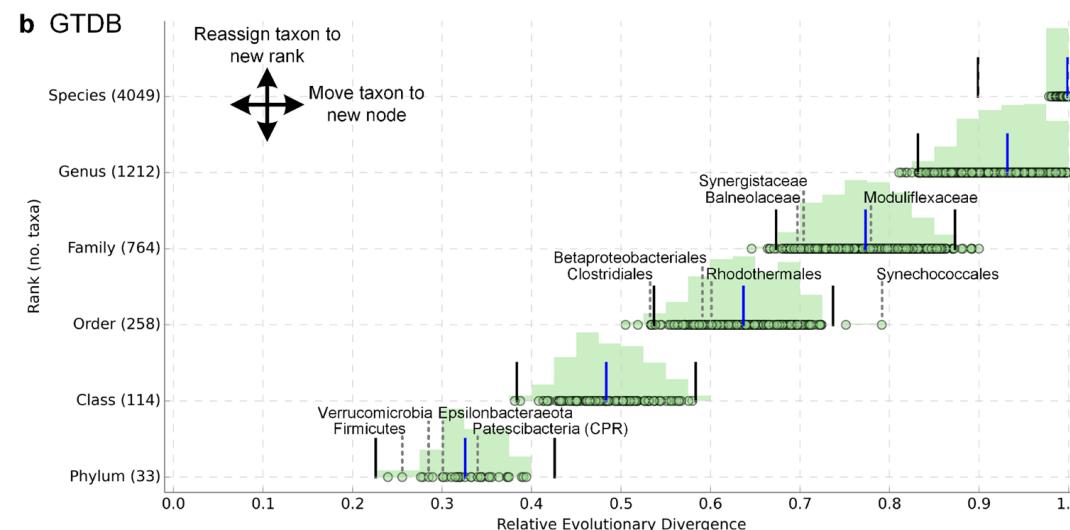
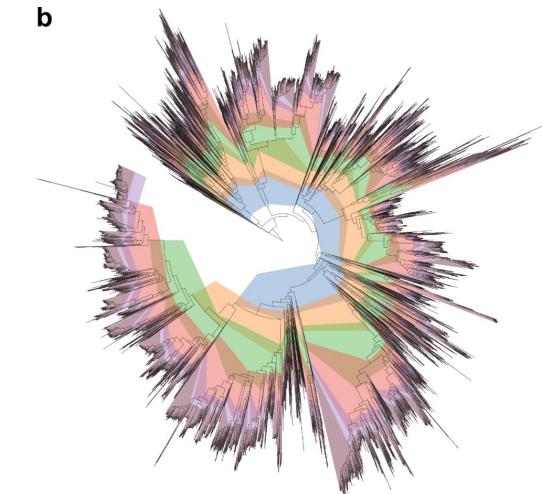
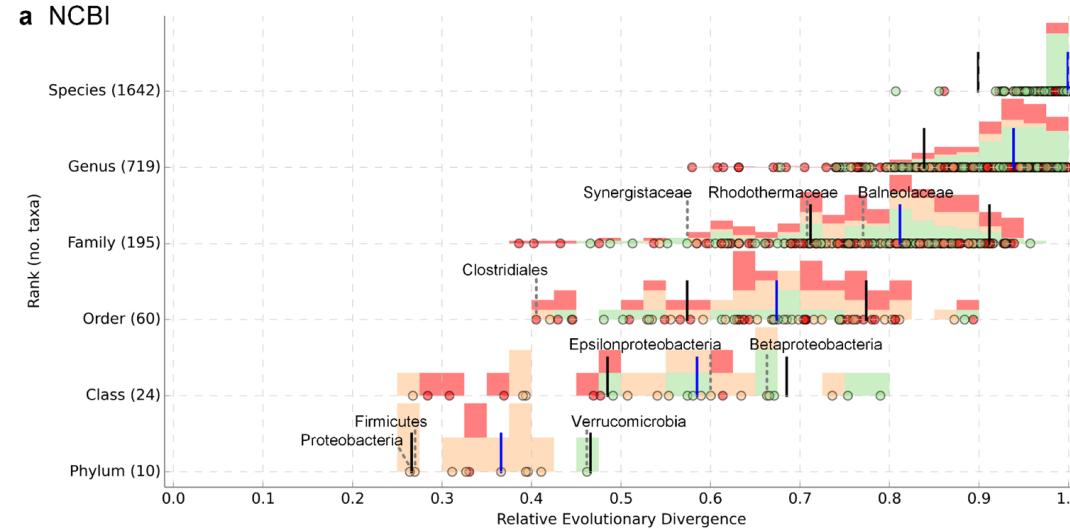
A basis for robust phylogeny



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



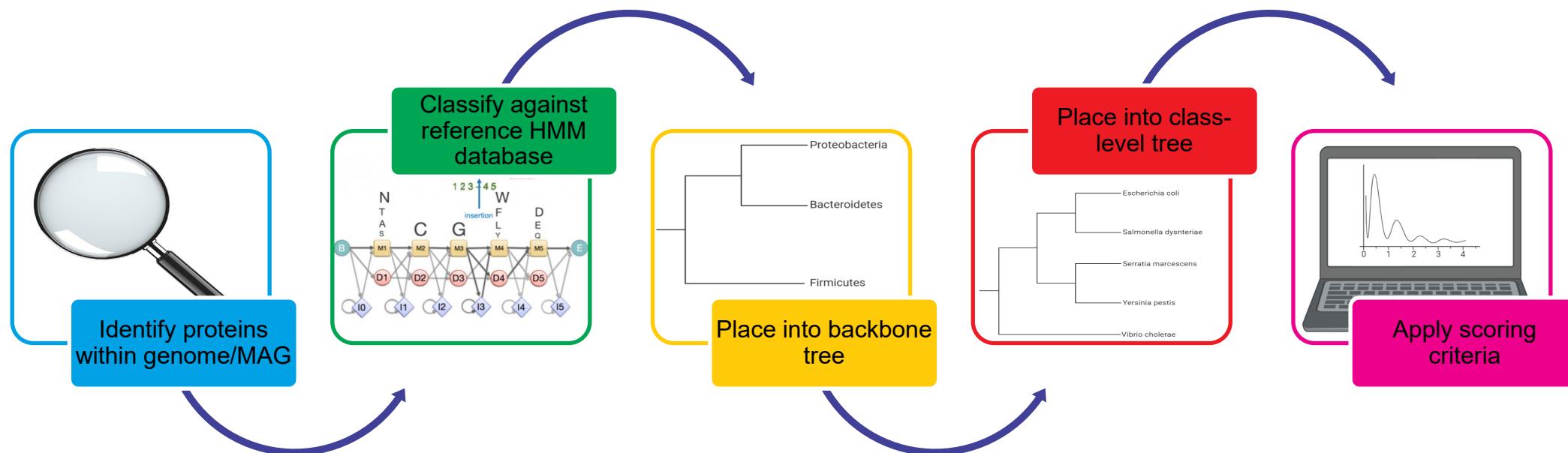
A basis for robust phylogeny



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



GTDB-TK version 2.1 – how it works



Discriminate species

Proxy for DNA-DNA hybridization

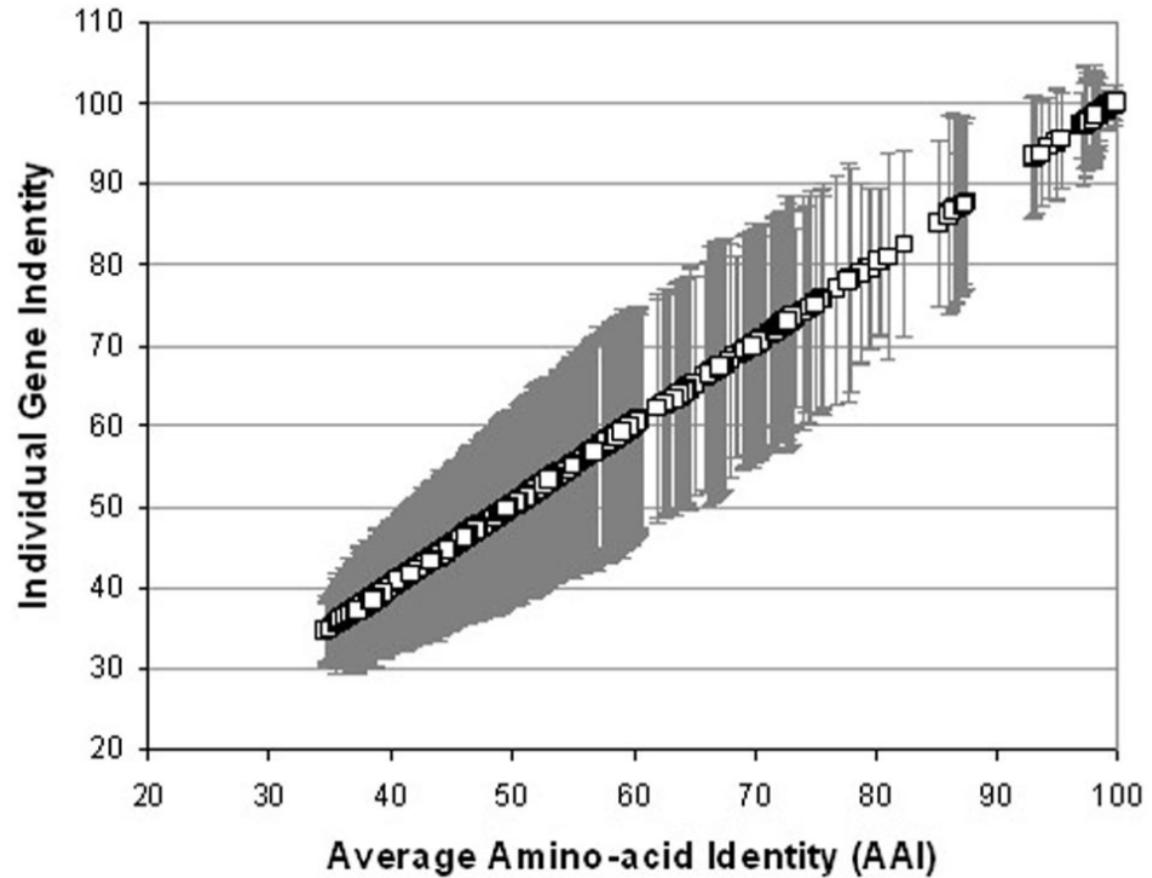
Pairwise genome comparisons:

- 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



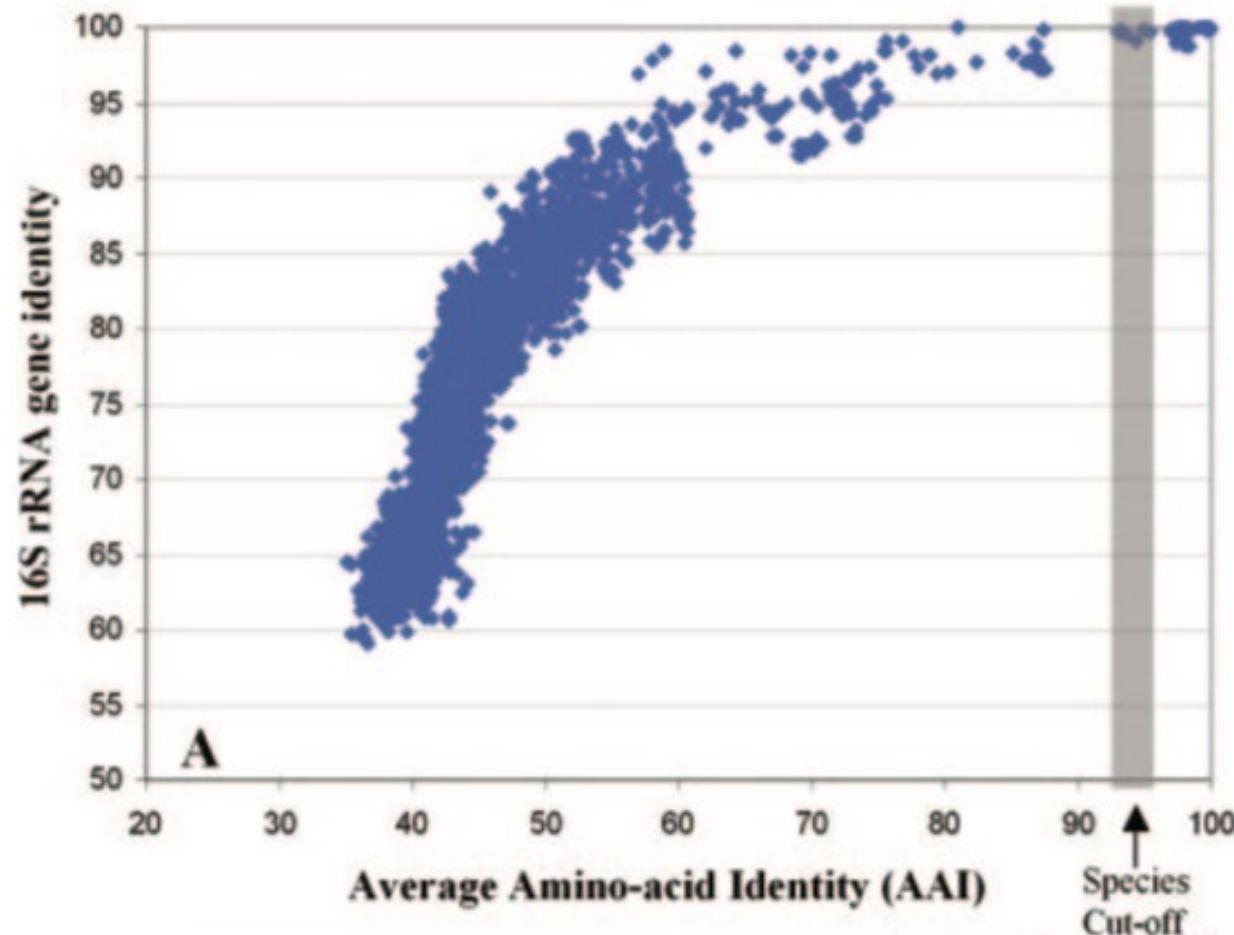
Discriminate species



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



Discriminate species



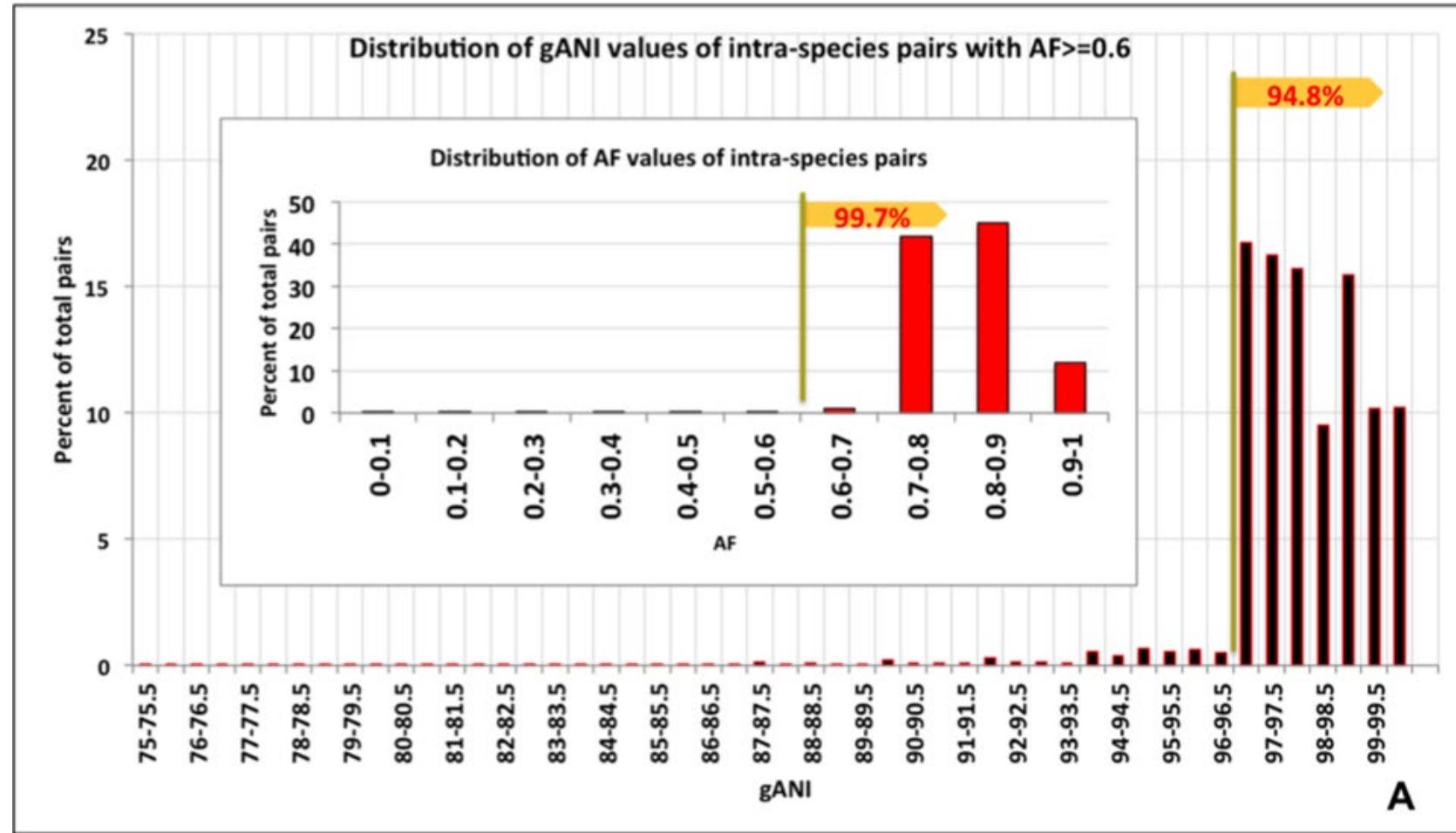
- AAI species cutoff \approx 95-96% ID
- Equivalent to 70% DNA-DNA hybridization threshold for species

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



Discriminate species

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



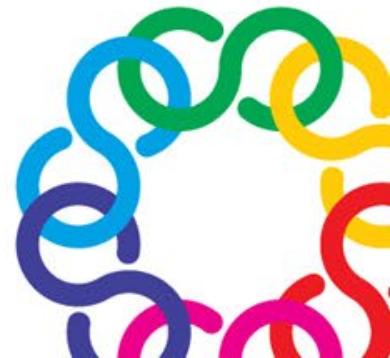
(Fig. 1A, Varghese et al., 2015, Nucleic Acids Research)



Discriminate species

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

- 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

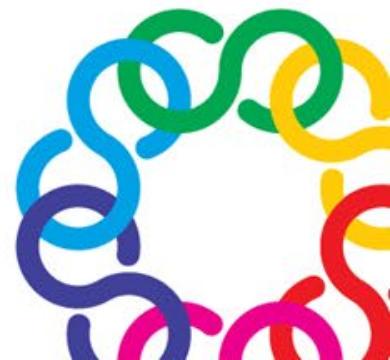
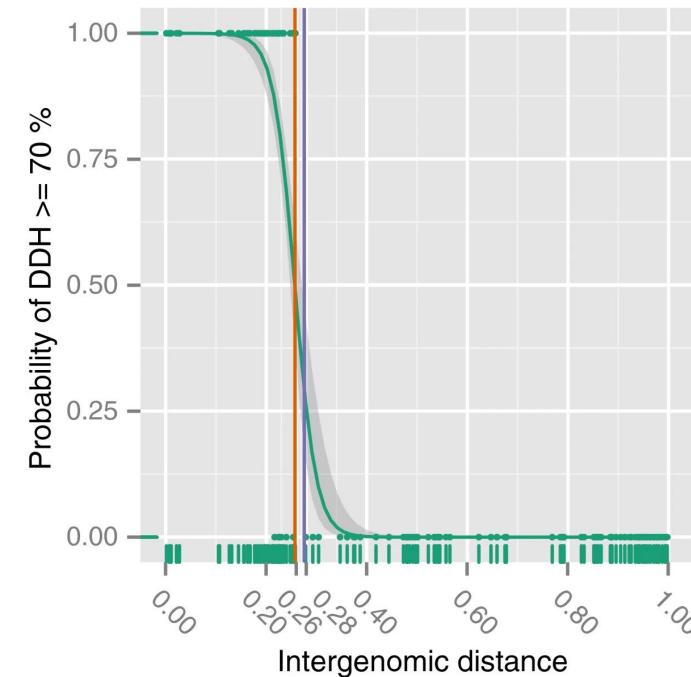


(Handley et al., 2017, ISME J)

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

K
KEGG Orthology (KO)

[Brite menu | Organism menu | Download htext | Download json]

Reference hierarchy (KO)

▼ ▼ ▼ ▼ One-click mode

▼ 09100 Metabolism

▶ 09101 Carbohydrate metabolism

▼ 09102 Energy metabolism

▶ 00190 Oxidative phosphorylation [PATH:[ko00190](#)]

▼ 00195 Photosynthesis [PATH:[ko00195](#)]

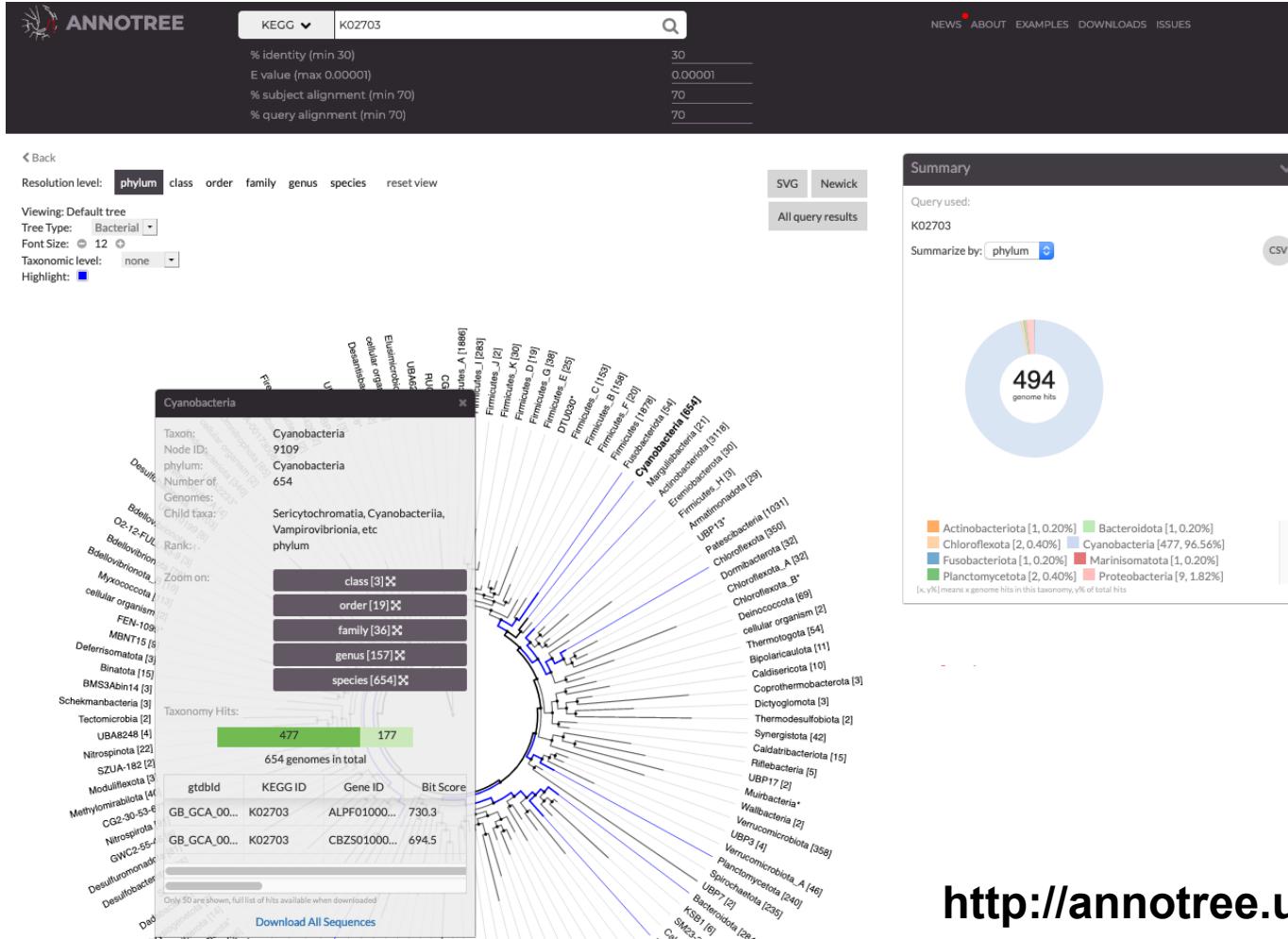
K02703 [K02703](#) psbA; photosystem II P680 reaction center D1 protein [EC:[1.10.3.9](#)]
[K02704](#) psbB; photosystem II CP47 chlorophyll apoprotein
[K02705](#) psbC; photosystem II CP43 chlorophyll apoprotein
[K02706](#) psbD; photosystem II P680 reaction center D2 protein [EC:[1.10.3.9](#)]
[K02707](#) psbE; photosystem II cytochrome b559 subunit alpha
[K02708](#) psbF; photosystem II cytochrome b559 subunit beta
[K02709](#) psbH; photosystem II PsbH protein
[K02710](#) psbI; photosystem II PsbI protein
[K02711](#) psbJ; photosystem II PsbJ protein
[K02712](#) psbK; photosystem II PsbK protein
[K02713](#) psbL; photosystem II PsbL protein
[K02714](#) psbM; photosystem II PsbM protein
[K02716](#) psbO; photosystem II oxygen-evolving enhancer protein 1
[K02717](#) psbP; photosystem II oxygen-evolving enhancer protein 2
[K08901](#) psbQ; photosystem II oxygen-evolving enhancer protein 3
[K03541](#) psbR; photosystem II 10kDa protein
[K03542](#) psbS; photosystem II 22kDa protein
[K02718](#) psbT; photosystem II PsbT protein
[K02719](#) psbU; photosystem II PsbU protein

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

KO (KEGG ORTHOLOGY) Database
Linking genomes to pathways by ortholog annotation

Menu PATHWAY BRITE MODULE KO Annotation ENZYME RModule BlastKOALA

Search KO for dsr Go

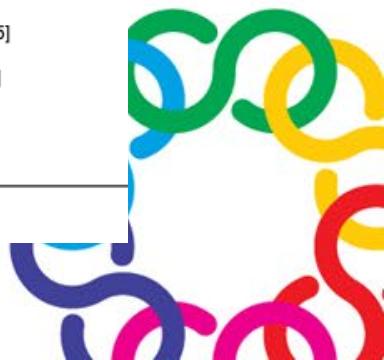
KO Database of Molecular Functions

KEGG Search ORTHOLOGY for dsr

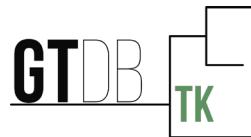
Database: ORTHOLOGY - Search term: dsr (Total 9 hits)

K04708	KDSR; 3-dehydroshinganine reductase [EC:1.1.1.102]
K07235	tusD, dsrE; tRNA 2-thiouridine synthesizing protein D [EC:2.8.1.-]
K07236	tusC, dsrF; tRNA 2-thiouridine synthesizing protein C
K07237	tusB, dsrH; tRNA 2-thiouridine synthesizing protein B
K11179	tusE, dsrC; tRNA 2-thiouridine synthesizing protein E [EC:2.8.1.-]
K11180	dsrA; dissimilatory sulfite reductase alpha subunit [EC:1.8.99.5]
K11181	dsrB; dissimilatory sulfite reductase beta subunit [EC:1.8.99.5]
K18502	dsrA; small regulatory RNA DsrA
K23077	dsrC; dissimilatory sulfite reductase related protein

DBGET integrated database retrieval system



Phylogenetics



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

bin_0_Filtered

SVLEAVSLDAWGEAIKRMRESIVKTSAGSEQFRVKTSYKKSLARLVTLNGVEVKIYDKEAIKKASLFNYTIISVNEYAVCTV
KRQTAEDIKIVKKMQUEGEGLESNVEILLEDEIENHSSWADLISDDDISNCNDSSASVSHIIIANARQTEKTINITKDVLVAD
AVTIDSLEIKEVGKEIKIVDLNIYVNRQLGIGENIMTTET----EALALDSLGTVNKAAYYKLNLHKYLATGWYNFDILFKI:
VGEIYQKEDSVAGETGKAFAKQIKLVSAFSIKSFEKVNIIDTLIELIYIRDMSYSDLKFLGYITDQKQARANKSDSKIVDG
-KQVLNHDALFLMNVLTVAAAGDIIADEQLIGGGKELGMSI-KDLNKTRGIKATAILALEIIEVKKLRAVDDEDIAKENQIEF:E
-----DSRCRIWMTRLEELWTYSDDWASWIGHTTIG-IVTYSMGNEINRIFTDEK-LLLSPWPPTDEY
VAKKRVDELMAKMFAPELLKKHLFIAPPFIFSDVKVKVKE-IAGEASDGSNRYHIVDQLFLIAGRFTAERLIKNAIKSFKA
TKVNKTI-FAHQKLETD-INASLLENSKVKGQEIEITADMTISRALKIGCIGGIIMEHSIKMNEALDTVDTHGDVIFPMVDE
LDVT-I-EVVLK-FVLAIPDLNVE--LEKVLLASDL-LFAGIDEKIISLMASAEVILQDAMLTKSEIAKALMVLIALYVA
IGKPDNAVNPAFIGNEISQERLEKEDIEADGVETVRVWMPHIALLSVRTQHNESEDRGARMVLNLTKLKAESFTTAKAEQILL
RAGQHRKFSRFNQVGRGHIDIPKEIVTTVKAGFTEVRLVIVKFVFRVWNTFADSITAIKYNAIPLGAAEIASPVA---RRAAN
IKLLWVKITVNCSSVIRKIDVIDKTEILRNKNNTIATSTHYTKMEQKAKLKLAGSP--LELRDGVVGTAKGNKGFMPCSQVY
RSMRNIGKGRSMIQIKPTKPTKWIPIFVSDEVLLVP-LKIAESNNDGTVVWKLSDMRLITMYDKKMASIAVSTTATSLVFISV
QSITLAFTELKGKFQV-IIPAANQIEAEKGVEIYTVKGDNALTKQIKLKDKEQVSRSIVETEVIGLYAFEFALNNIGIKVEIRLA
-LAVNLFPKASELEIKGYENRKIFAILEGELKLQSHKILEKAGNSNRIENIATTCKDFLIF-EIHYPDIPYEYSDIDK-NVGIT
TTSENLTDEIKYYVLQALIAAMDKEALVFKSGPENIRKETV-FAIAGKDANIAMLTH-SVPVLTKTD-FFTNAAEKTKVKNG-C

Bin_0

Bin_1

Bin_2

Bin_3

bin_0.filtered

SVLEAVSLDAWGEAIKRMRESIVKTSAGSEQFRVKTSYKKSLARLVTLNGVEVKIYDKEAIKKASLFNYTIISVNEYAVCTV
KRQTAEDIKIVKKMQUEGEGLESNVEILLEDEIENHSSWADLISDDDISNCNDSSASVSHIIIANARQTEKTINITKDVLVAD
AVTIDSLEIKEVGKEIKIVDLNIYVNRQLGIGENIMTTET----EALALDSLGTVNKAAYYKLNLHKYLATGWYNFDILFKI:
VGEIYQKEDSVAGETGKAFAKQIKLVSAFSIKSFEKVNIIDTLIELIYIRDMSYSDLKFLGYITDQKQARANKSDSKIVDG
-KQVLNHDALFLMNVLTVAAAGDIIADEQLIGGGKELGMSI-KDLNKTRGIKATAILALEIIEVKKLRAVDDEDIAKENQIEF:E
-----DSRCRIWMTRLEELWTYSDDWASWIGHTTIG-IVTYSMGNEINRIFTDEK-LLLSPWPPTDEY
VAKKRVDELMAKMFAPELLKKHLFIAPPFIFSDVKVKVKE-IAGEASDGSNRYHIVDQLFLIAGRFTAERLIKNAIKSFKA
TKVNKTI-FAHQKLETD-INASLLENSKVKGQEIEITADMTISRALKIGCIGGIIMEHSIKMNEALDTVDTHGDVIFPMVDE
LDVT-I-EVVLK-FVLAIPDLNVE--LEKVLLASDL-LFAGIDEKIISLMASAEVILQDAMLTKSEIAKALMVLIALYVA
IGKPDNAVNPAFIGNEISQERLEKEDIEADGVETVRVWMPHIALLSVRTQHNESEDRGARMVLNLTKLKAESFTTAKAEQILL
RAGQHRKFSRFNQVGRGHIDIPKEIVTTVKAGFTEVRLVIVKFVFRVWNTFADSITAIKYNAIPLGAAEIASPVA---RRAAN
IKLLWVKITVNCSSVIRKIDVIDKTEILRNKNNTIATSTHYTKMEQKAKLKLAGSP--LELRDGVVGTAKGNKGFMPCSQVY
RSMRNIGKGRSMIQIKPTKPTKWIPIFVSDEVLLVP-LKIAESNNDGTVVWKLSDMRLITMYDKKMASIAVSTTATSLVFISV
QSITLAFTELKGKFQV-IIPAANQIEAEKGVEIYTVKGDNALTKQIKLKDKEQVSRSIVETEVIGLYAFEFALNNIGIKVEIRLA
-LAVNLFPKASELEIKGYENRKIFAILEGELKLQSHKILEKAGNSNRIENIATTCKDFLIF-EIHYPDIPYEYSDIDK-NVGIT
TTSENLTDEIKYYVLQALIAAMDKEALVFKSGPENIRKETV-FAIAGKDANIAMLTH-SVPVLTKTD-FFTNAAEKTKVKNG-C

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



Phylogenetics



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.

R. A. FISHER

89

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

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

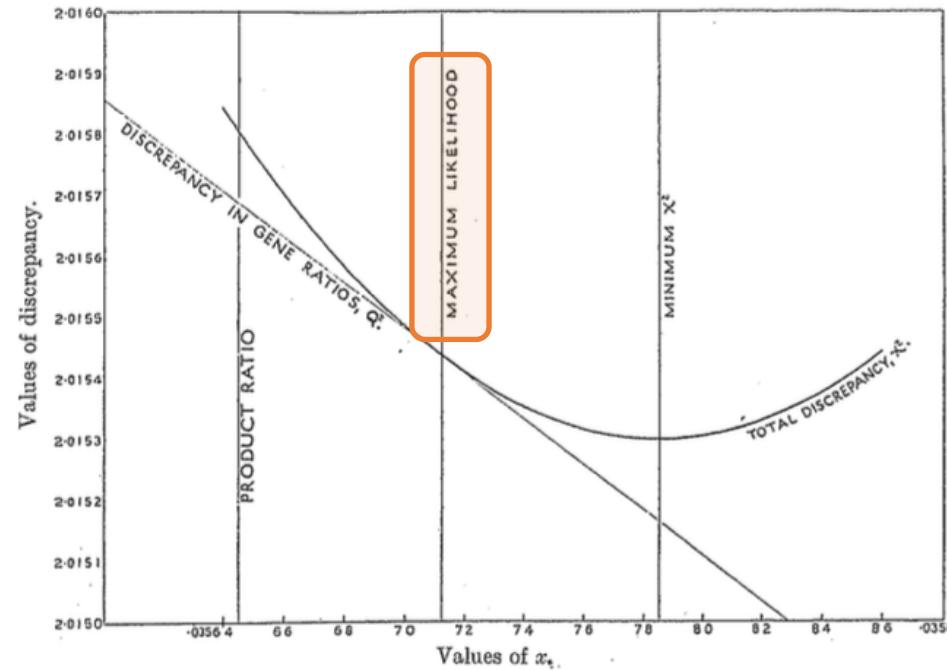


Fig. 2.

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:



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>



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



Phylogenetics



10 bins / MAGs

CPU: 02

RAM: 64MB

Time: 4 seconds



10 bins / MAGs

CPU: 22

RAM: 80MB

Time: 03:56:34

Pros: Speed, memory efficiency, easy to use

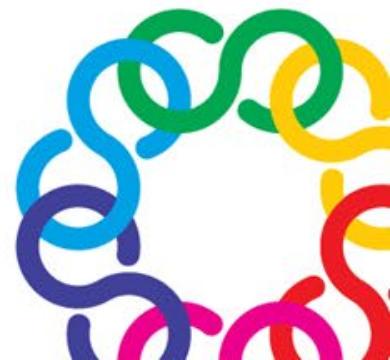
Pros: Provides better accuracy due to its advanced optimization methods (**Maximum Likelihood Optimization**)

More customizations and flexibilities

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

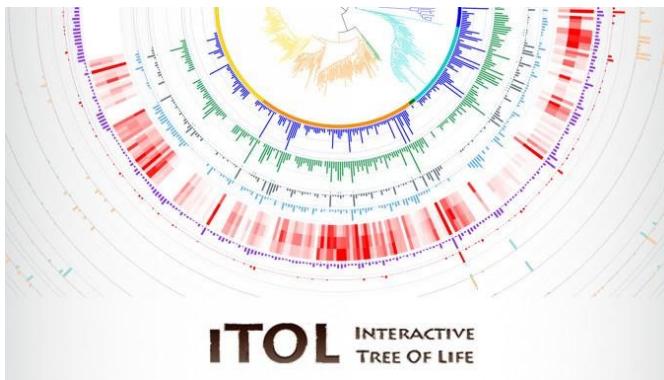
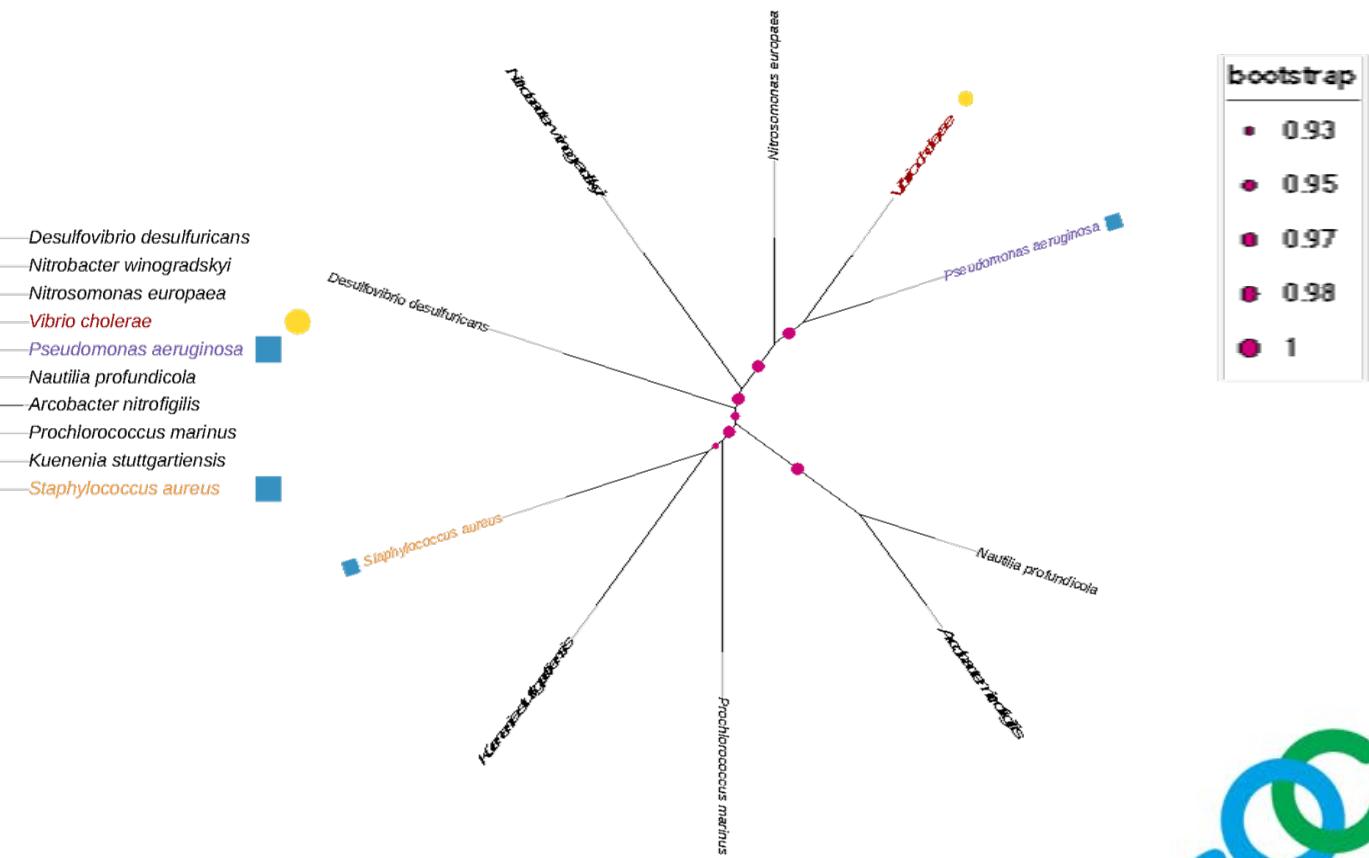
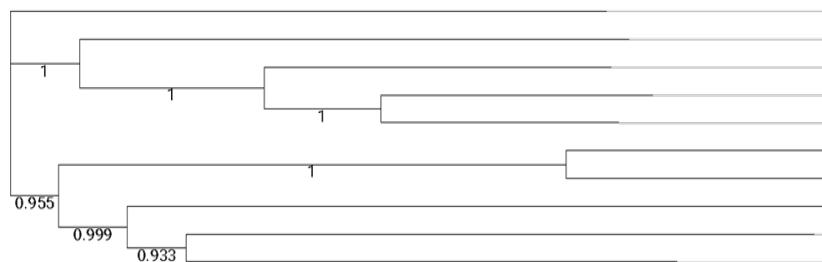
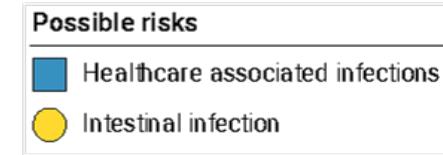
Limited customization and flexibility

Cons: Require more computational resources and time

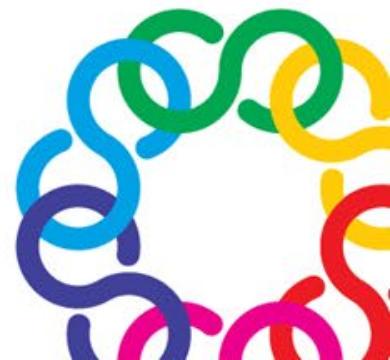


Phylogenetics

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](#)

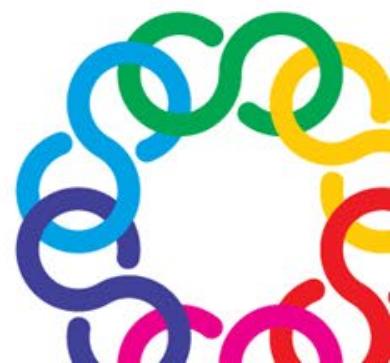
16k Accesses | 112 Citations | 73 Altmetric | [Metrics](#)

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

Article | [Open Access](#) | Published: 21 December 2020

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

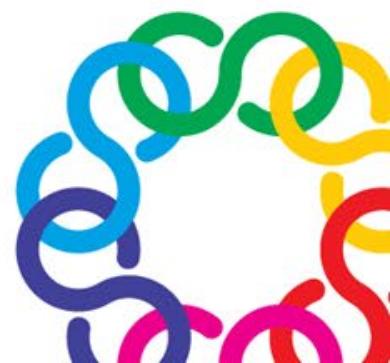
[Stephen Nayfach](#) , [Antonio Pedro Camargo](#), [Frederik Schulz](#), [Emiley Eloe-Fadrosh](#), [Simon Roux](#) & [Nikos C. Kyriides](#) 

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

Inspect viral sequence quality

CheckV

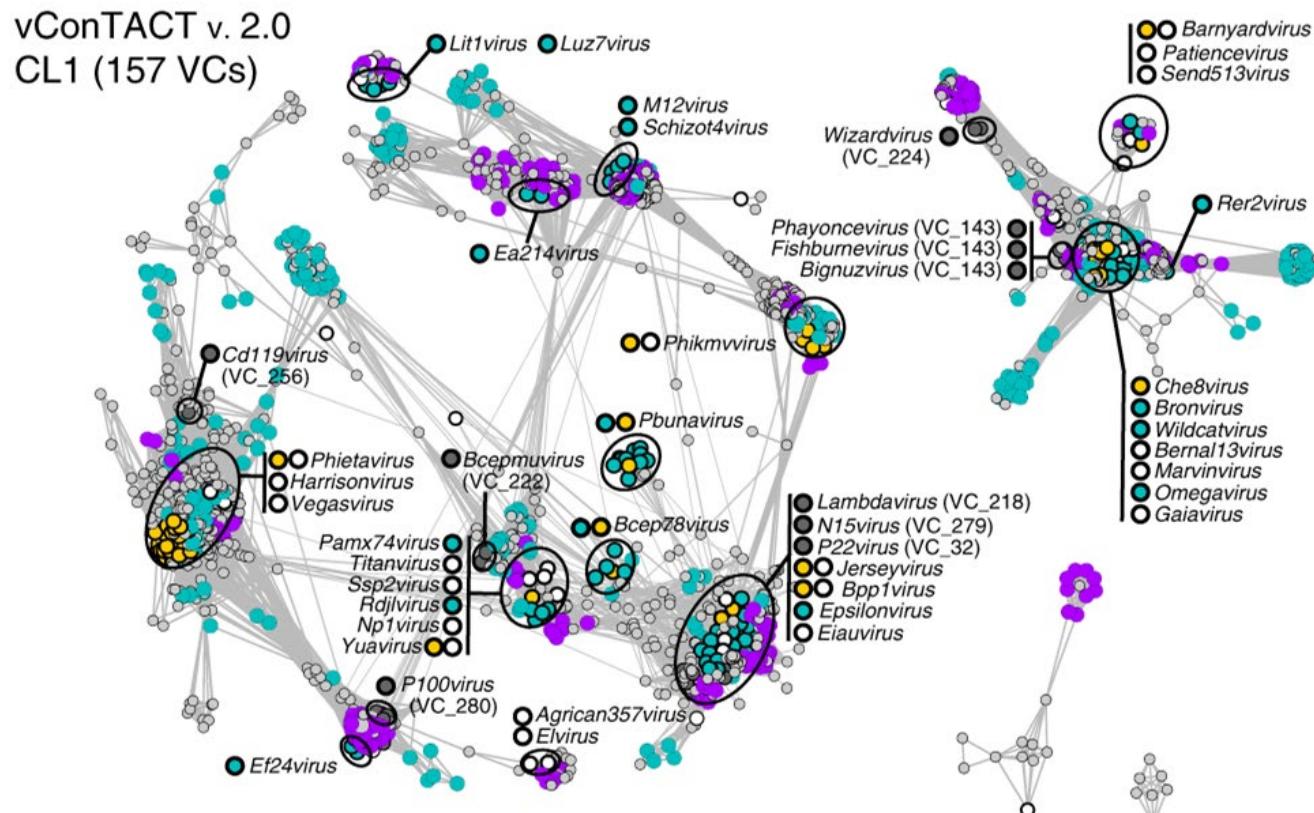
- Removes host regions based on hits to multiple databases and GC content
- Estimates completeness and quality by comparing against known and complete virus sequences
- Identifies closed genomes (direct terminal repeats)



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

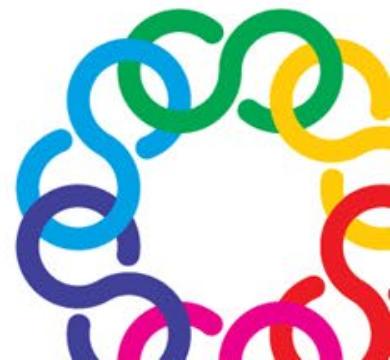
[Go to Github MGSS webpage](#)

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

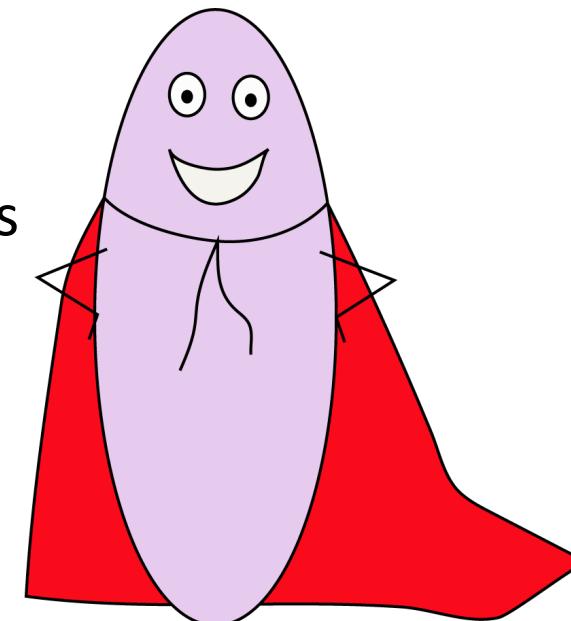
Aim:

- 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 1/10th 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

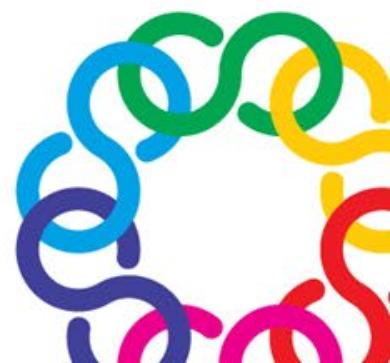
- Homology:
 - Search by sequence similarity to homologous sequences
 - Based on the assumption that functional regions are more conserved evolutionarily than non-functional regions
- *Ab initio*:
 - Search by content: find genes by statistical properties that distinguish protein-coding DNA from non-coding DNA
 - Search by signals/sites, e.g. promoters, start and stop codons



Gene prediction

Homology: Sequence similarity searches

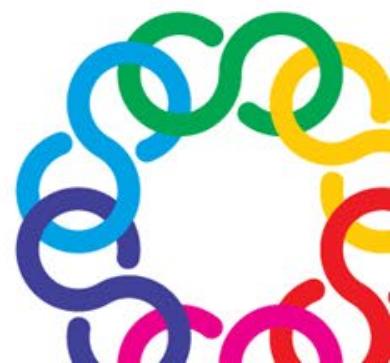
- Finding similarity in gene sequences between expressed sequence tags (ESTs), proteins, or other genomes to the input genome
- Local alignment:
 - BLAST family tools: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
 - Global alignment
 - GeneWise: <https://www.ebi.ac.uk/Tools/psa/genewise/>



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

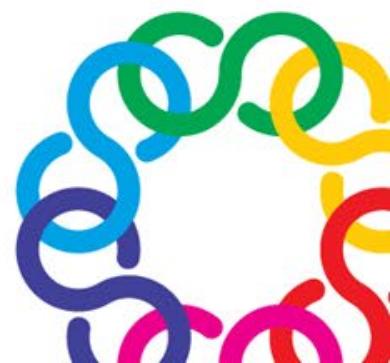
- <http://ccb.jhu.edu/software/glimmer/index.shtml>
- Interpolated Markov model (IMM) DNA discriminator
- Log-likelihood that a given interval on a DNA sequence was generated by a model of coding versus non-coding DNA



Gene prediction

Markov Model Based Algorithms: GeneMark/GeneMarkHMM/MetaGeneMark

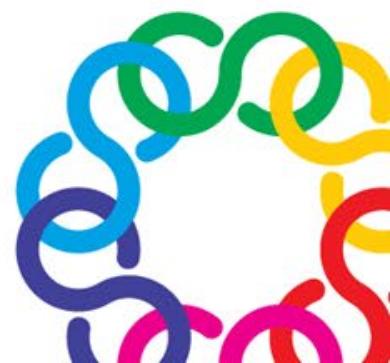
- <http://exon.gatech.edu/GeneMark/>
- GeneMark is a family of gene prediction tools
- Genomic sequences can be analysed either by the self-training program GeneMarkS (sequences >50 kb) or using Heuristic Models by GeneMarkHMM
- Pre-trained model parameters are available for many species
- Metagenomics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

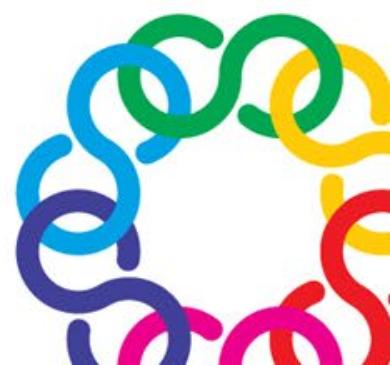
- **Caveat:** unusual genetic codes
 - First uses genetic code 11 (stop codons TAA, TGA, TAG)
 - If genes are too short, uses alternative code 4 (TGA not a stop codon)
 - Will not try code 25, but will issue warning if genes are short
 - Must manually select code 25



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

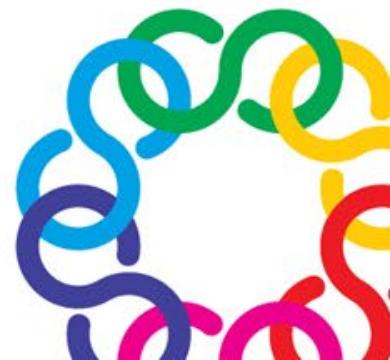
[Go to Github MGSS webpage](#)

Tasks:

- Predict open reading frames and protein sequences

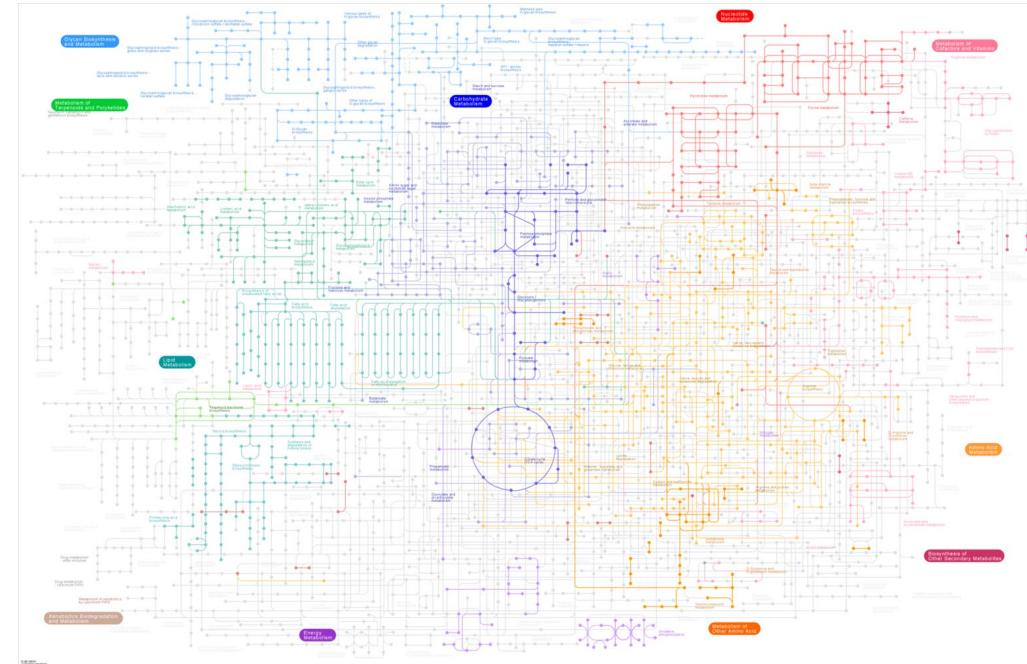


Gene annotation (part 1)



Gene annotation

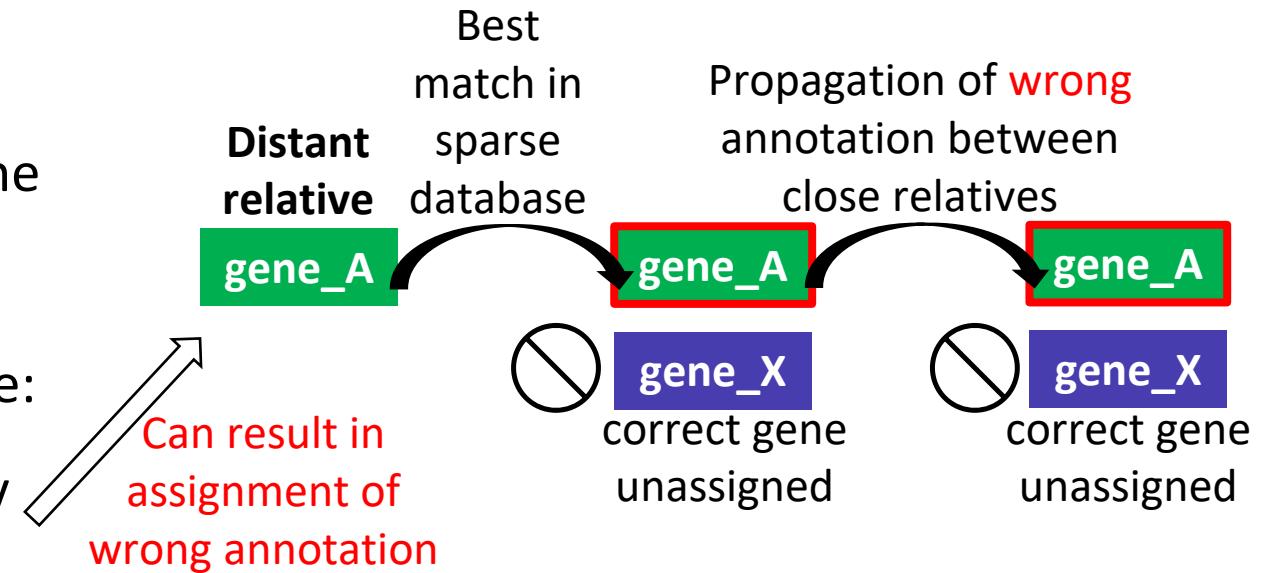
- Genome annotation attempts to predict gene function
- Predicted genes or protein sequences are compared against a curated set of reference sequences for which function is known, or is strongly suspected



Gene annotation

Caveat:

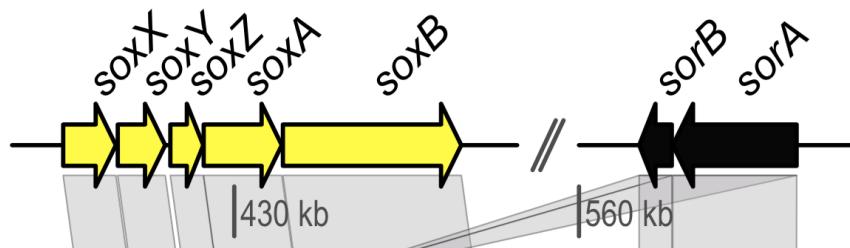
- Annotations are dependent on the reference database
- Environmental genomes can have:
 - Genes with distant homology matches to unrelated taxa
 - Large numbers of “hypothetical” gene annotations (= genes of unknown function)



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

Screenshot of a BLAST search results page. The top navigation bar includes 'Descriptions', 'Graphic Summary', **Alignments**, and 'Taxonomy'. The 'Alignments' tab is selected. Below it, the 'Alignment view' dropdown is set to 'Pairwise'. A green banner at the top indicates '100 sequences selected'. The main content area shows a search result for 'amidase [Gemmatimonadetes bacterium]'. The sequence ID is MBB27982.1, with a length of 447 and one match found. A link to 'See 1 more title(s)' is present. The alignment details for the first hit are shown, including the range (182 to 446), download links for 'GenPept' and 'Graphics', and 'Next Match' and 'Previous Match' buttons. The 'Related Information' section lists 'Identical Proteins - Identical proteins to MBB27982.1'. The alignment table shows the query sequence (MBB27982.1) and subject sequence (amidase) with their respective amino acid sequences and matching positions.

Score	Expect	Method	Identities	Positives	Gaps
268 bits(686)	3e-84	Compositional matrix adjust.	139/265(52%)	175/265(66%)	0/265(0%)
Query 1	MGLKPTFGRISLRGILPVSYELDHPGPFTRSVADAALQCLAGKDPPLDPLSADVPVDI +GLKP <small>T</small> GR+S+ G++PVS+ LDHPGP T SV DAA ILQ +AG DP DPLSA				
Sbjct 182	VGLKP <small>T</small> LGRVSVHGVVPSFNLDHPGPLTLSVGDARILQVIAGYDPKDPPLSASETTTI				
Query 61	RIEPLSRPPRVRGVIVRTYYPPDNNADETMRAATDDAIERLASEGAEFTDVHMPGSFAELHEN PL RPPR+G + Y+ +ADE M +AT AIE L GAE ++ MP SF LHE				
Sbjct 242	TPRPLPRPPRIGHLVGYFREQADEMSSATQRAIECLOLAGAECLEMPDSFGCLHEN#				
Query 121	ALLLAVGAANVLDERGYVAHRDAFPPLCEIMERGRSAGAVDYARARRHQISFKSEVLA# +++ A DE++ HR+ +PP L +M+ G + AV YA AR+HQI F+ ++ +				
Sbjct 302	RIIMVSEGAAYHDEQFGLHRNEYPPGLRSLMDEGLATSAVTYANARKHQIDFRLQIQSI#				
Query 181	EGVDLILLTPATPTPAPSGLTSTGNPAFNPSWSYAGLPTIVLPAACSSDGLPAGIQQLVAF +D+LLTPAT TPAP L STGNPAFNPSWSY GLPTI LP GLPA IQLV				
Sbjct 362	RDLIDLITTPATLTPAKTLESTGNPAFNPSWSYCGLPTISLPVEVGESGLPAAIQLVGE#				
Query 241	FAEIRLLTVSAWCETRLEWNRTSSI 265 F+E RLL+++ WCE L WN P +				
Sbjct 422	FSESRLLSIAQWCEQVLGWNHKPEL 446				



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 (Protein ANalysis THrough Evolutionary Relationships) 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

 [HOME](#) [ENZYME CLASSES](#) [ASSOCIATED MODULES](#) [GENOMES](#)

Welcome to the Carbohydrate-Active enZYmes Database

The CAZY database describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds.

Online since 1998, CAZY is a specialist database dedicated to the display and analysis of genomic, structural and biochemical information on Carbohydrate-Active Enzymes (CAZymes). CAZY data are accessible either by browsing sequence-based families or by browsing the content of genomes in carbohydrate-active enzymes. New genomes are added regularly shortly after they appear in the daily releases of GenBank. New families are created based on published evidence for the activity of at least one member of the family and all families are regularly updated, both in content and in description.

An original aspect of the CAZY database is its attempt to cover all carbohydrate-active enzymes across organisms and across subfields of glycosciences. Please let us know if some families have escaped our attention, we will be happy to add them !

For a more extensive encyclopedic resource on the particular features of carbohydrate active enzymes, please visit [CAZypedia](#), a web site driven by the scientific community that studies these enzymes.

Reference for the CAZY database : In the 2014 database issue of Nucleic Acids Research, we summarized the many changes that have occurred in the CAZY database during the five previous years.
Read the [Abstract](#) or the full [paper](#).

A new tool associated with the CAZY database ! PULDB is a database of Polysaccharide Utilization Loci (PULs) in Bacteroidetes. PULDB displays information on experimentally determined and predicted PULs for a number of Bacteroidetes genomes.
Read the [Abstract](#) or the full [paper](#).

Enzyme Classes currently covered

Modules that catalyze the breakdown, biosynthesis or modification of carbohydrates and glycoconjugates :

- [Glycoside Hydrolases \(GHS\)](#) : hydrolysis and/or rearrangement of glycosidic bonds (see CAZypedia [definition](#))
- [GlycosylTransferases \(GTS\)](#) : formation of glycosidic bonds (see [definition](#))
- [Polysaccharide Lyases \(PLs\)](#) : non-hydrolytic cleavage of glycosidic bonds
- [Carbohydrate Esterases \(CEs\)](#) : hydrolysis of carbohydrate esters
- [Auxiliary Activities \(AAs\)](#) : redox enzymes that act in conjunction with CAZymes.

Associated Modules currently covered

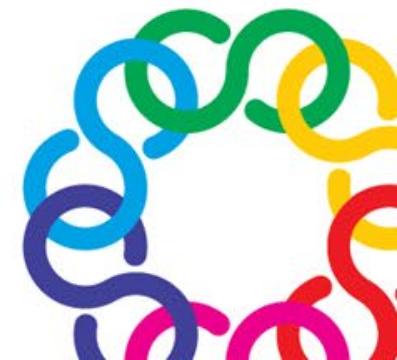
Carbohydrate-active enzymes often display a modular structure with non-catalytic modules appended to the enzymes above

- [Carbohydrate-Binding Modules \(CBMs\)](#) : adhesion to carbohydrates

[Suivre @CAZYDB](#)

<http://www.cazy.org>

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	0.94444444
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



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

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

[» Tutorials](#) [» Help](#) [login](#)

Info: To monitor RAST's load and view other news and statistics for RAST and the SEED, please visit ["The Daily SEED."](#)

Welcome to RAST

[» Register for a new account, service, or user-group](#)
[» Forgot your password?](#)

Login
Password Login

RAST Job Load, last 24 hours

RAST Queue

Jobs

Wed 00:00 Wed 12:00

■ alder.mcs.anl.gov last day (now 20.00)

What is RAST?

RAST (Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating complete or nearly complete bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylogenetic tree.

We have a number of presentations and tutorials available:

- [Registering for RAST](#)
- [The IRIS/Automated-Assembly/RASTtk Workshop Presentations and Tutorials](#)
- [The SEED/"Classic-RAST" Workshop presentations and Tutorials](#)
- [Downloading and installing the RASTtk Toolkit](#)
- [Downloading and installing the myRAST Toolkit](#)
- [The RAST batch submission interface \(a part of myRAST\)](#)
- [Making manual improvements to RAST-annotated genomes \(first tutorial\)](#). This is a powerpoint presentation; bring it up in slide-show mode and click through to see the animations and movies.
- [Making manual improvements to RAST-annotated genomes \(second tutorial\)](#). This is a second tutorial on the topic of manually improving RAST annotations; it is also a powerpoint presentation with animations.

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

Gene annotation

RAST Annotation Server (Rapid Annotation using Subsystem Technology)

The SEED Viewer SEED Viewer version 2.0

Welcome to the SEED Viewer - a read-only browser of the curated SEED data. For more information about The SEED please visit theSEED.org.

>Navigate >Organism >Comparative Tools >Help find

Organism Overview for *Candidatus Latescibacter anaerobius* SCGC AAA252-E07 (910047.4)

Genome *Candidatus Latescibacter anaerobius* SCGC AAA252-E07 (Taxonomy ID: [910047](#))

Domain Bacteria

Taxonomy Bacteria; FCB group; *Candidatus Latescibacteria*; *Candidatus Latescibacter*; *Candidatus Latescibacter anaerobius*; *Candidatus Latescibacter anaerobius* SCGC AAA252-E07

Neighbors [View closest neighbors](#)

Size 2,290,285

GC Content 42.1

N50 24013

L50 34

Number of Contigs (with PEGs) 164

Number of Subsystems 157

Number of Coding Sequences 2064

Number of RNAs 30

For each genome we offer a wide set of information to browse, compare and download.

Browse Compare Download Annotate

Browse through the features of [Candidatus Latescibacter anaerobius SCGC AAA252-E07](#) both graphically and through a table. Both allow quick navigation and filtering for features of your interest. Each feature is linked to its own detail page.

Click [here](#) to get to the Genome Browser

Subsystem Information

Subsystem Statistics Features in Subsystems

Subsystem Coverage

Subsystem Category Distribution

Subsystem Feature Counts

- Cofactors, Vitamins, Prosthetic Groups, Pigments (67)
- Cell Wall and Capsule (23)
- Virulence, Disease and Defense (25)
- Potassium metabolism (15)
- Photosynthesis (0)
- Miscellaneous (2)
- Phages, Prophages, Transposable elements, Plasmids (2)
- Membrane Transport (21)
- Iron acquisition and metabolism (0)
- RNA Metabolism (39)
- Nucleosides and Nucleotides (21)
- Protein Metabolism (60)
- Cell Division and Cell Cycle (2)
- Motility and Chemotaxis (0)
- Regulation and Cell signaling (11)
- Secondary Metabolism (0)
- DNA Metabolism (31)
- Fatty Acids, Lipids, and Isoprenoids (15)
- Nitrogen Metabolism (9)
- Dormancy and Sporulation (1)
- Respiration (22)
- Stress Response (8)
- Metabolism of Aromatic Compounds (0)
- Amino Acids and Derivatives (45)
- Sulfur Metabolism (9)
- Phosphorus Metabolism (1)
- Carbohydrates (48)



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

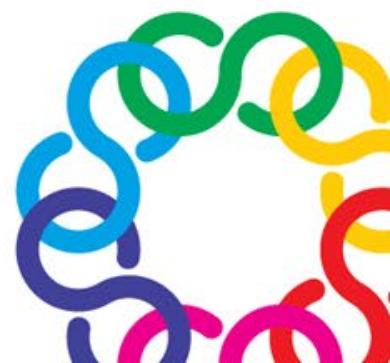


Task: Gene annotation

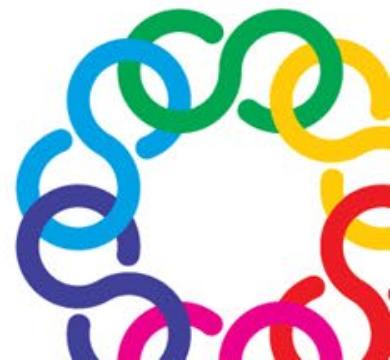
[Go to Github MGSS webpage](#)

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 Enter keywords
map Organism 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
01210 2-Oxocarboxylic acid metabolism
01212 Fatty acid metabolism
01230 Biosynthesis of amino acids
01220 Degradation of aromatic compounds

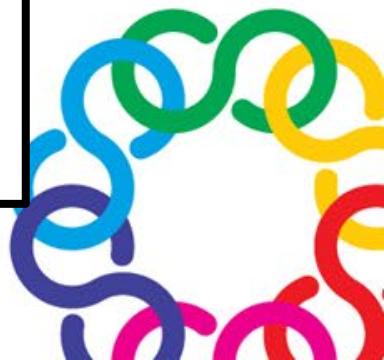
KEGG Pathway Maps

[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>

KEGG Pathway Maps

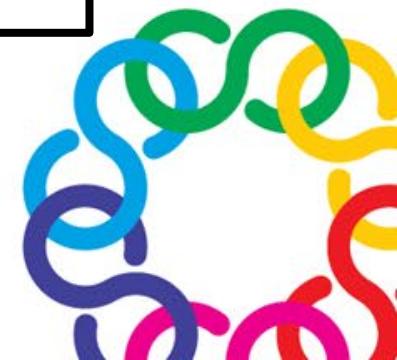
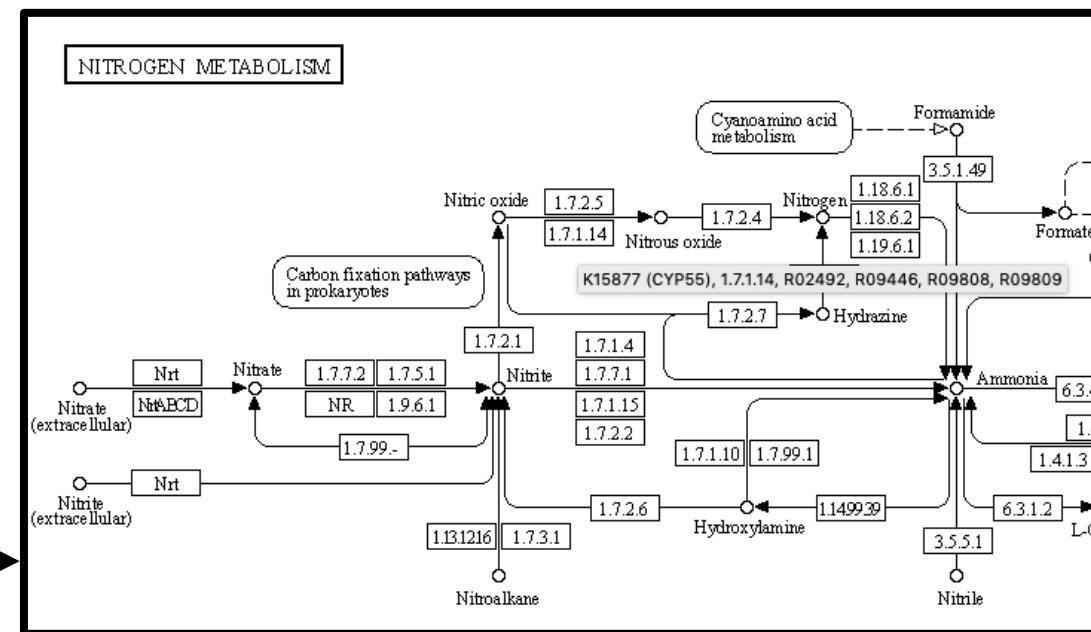
[Brite menu | Download htext | Download json]

KEGG pathway maps

One-click mode

▼ Metabolism

- ▼ Global and overview maps
- 00100 Metabolic pathways
- 00110 Biosynthesis of secondary metabolites
- 00120 Microbial metabolism in diverse environments
- 00130 Biosynthesis of antibiotics
- 00140 Carbon metabolism
- 00150 2-Oxocarboxylic acid metabolism
- 00160 Fatty acid metabolism
- 00170 Biosynthesis of amino acids
- 00180 Degradation of aromatic compounds
- Carbohydrate metabolism
- Energy 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>

KEGG Pathway Maps

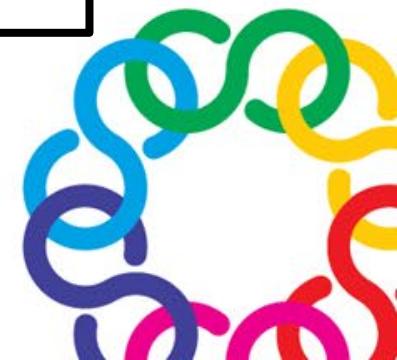
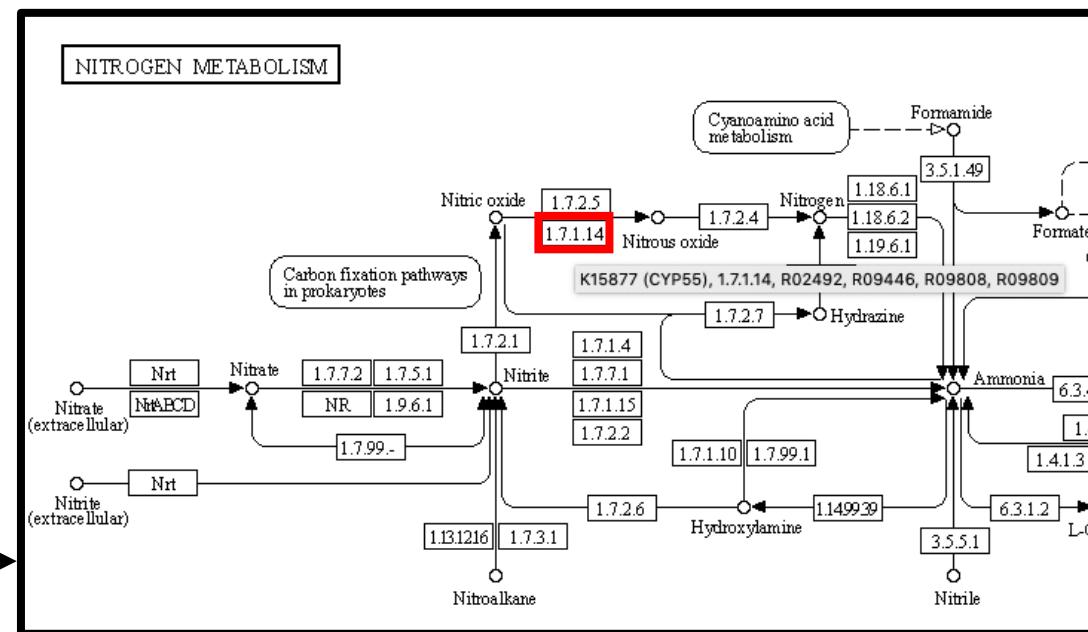
[Brite menu | Download htext | Download json]

KEGG pathway maps

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
- Carbon fixation pathways in prokaryotes
- 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>

KEGG Pathway Maps

[Brite menu | Download htext | Download json]

KEGG pathway maps

▼ ▼ ▼ 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

→

ORTHOLOGY: K15877	
Entry	K15877 KO
Name	CYP55
Definition	fungal nitric oxide reductase [EC:1.7.1.14]
Pathway	ko00910 Nitrogen metabolism ko01100 Metabolic pathways ko01120 Microbial metabolism in diverse environments
Brite	KEGG Orthology (KO) [BR:ko00001] 09100 Metabolism 09102 Energy metabolism 00910 Nitrogen metabolism K15877 CYP55; fungal nitric oxide reductase Enzymes [BR:ko01000] 1. Oxidoreductases 1.7 Acting on other nitrogenous compounds as donors 1.7.1 With NAD+ or NADP+ as acceptor 1.7.1.14 nitric oxide reductase [NAD(P)+, nitrous oxide-fo K15877 CYP55; fungal nitric oxide reductase BRITE hierarchy
Other DBs	RN: R02492 R09446 R09808 R09809 GO: 0016966



Gene annotation

Identification of Biosynthetic Gene Clusters with antiSMASH

antiSMASH bacterial version Submit Bacterial Sequence Submit Fungal Sequence Submit Plant Sequence Download About Help Contact

Server status: working
Running jobs: 0
Queued jobs: 0
Jobs processed: 585148

Nucleotide input Results for existing job

Search a genome sequence for secondary metabolite biosynthetic gene clusters Load sample input Open example output

Incomplete RefSeq annotations
Dear antiSMASH users, it has come to our attention that a recent RefSeq reannotation again broke NRPS/PKS ORFs. If your results look weird, try uploading the corresponding GenBank record or a FASTA file.

Notification settings
Email address (optional): your@email.com

Data input
Upload file Get from NCBI NCBI acc # NCBI accession number of desired sequence

Detection strictness: relaxed
strict relaxed loose
• Detects well-defined clusters containing all required parts.
• Detects partial clusters missing one or more functional parts.

Extra features All off All on
 KnownClusterBlast ClusterBlast SubClusterBlast
 ActiveSiteFinder Cluster Pfam analysis Pfam-based GO term annotation

Submit

Please be considerate in your use of antiSMASH. Help us keep antiSMASH available for everybody by limiting yourself to 5 concurrent jobs. Need to run more? See the [antiSMASH install guide](#) for instructions for getting your own antiSMASH installation.

antiSMASH antibiotics & Secondary Metabolite Analysis SHell Version 4.2.0

Select Gene Cluster: Overview 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

Identified secondary metabolite clusters

Cluster	Type	From	To	Most similar known
Cluster 1	Other	1	5782	-
Cluster 2	T3pks	1	5701	-
Cluster 3	T1pks	1	4918	-
Cluster 4	Otherks	1	3433	-
Cluster 5	Nrps	1	3365	-
Cluster 6	Terpene	1	3235	-

<https://antismash.secondarymetabolites.org/>



NCBI Conserved Domain Search

Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in [FASTA format](#). For multiple protein queries, use [Batch CD-Search](#).

MAINKHHITPMLDQLESGFWPSFISGIKRLRDEHPEERINKMTNDLLGQLEHSYETRKGYWKGGTVSVFQYGGGIIPRFSEVGHAFPEKFHTLRLVQPAGNHYSTMLRQMADEWERYGSGLVTFHQQTGNIMF1GTDTEQTHFFDEINDYGWLGGAGPCVRTAMSVCVGAARCEMSCTNEQKAIRLLVNNFTDDVHRLPALPYKFKFKVSGCGNDQCNAVERADFAVIGTWRDDMNVNQDEFKAYGRKGRQHVIDNIITRCPTNALSINDDDSLEVNNKDCVRCMHCCLNVPKALHPGDRRGVTILIGGKRTLKGDLMGTVVVVFKKLDTEEDWEELAEEIIDFWAENALEHERCGERMIERIGLNFLEGVGVEVDPMVNPNPRESSYIRMDGWDEEAVKWFDRQAEAS|

OPTIONS

Search against database: CDD v3.17 - 52910 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter:

Composition based statistics adjustment:

Force live search:

Rescue borderline hits: Suppress weak overlapping hits:

Maximum number of hits: 500

Result mode: Concise Standard Full

Submit **Reset**

Retrieve previous CD-search result

Request ID: Retrieve

References:

- [1] Marchler-Bauer A et al. (2017), "CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.", *Nucleic Acids Res.* **45**(D)200-3.
- [2] Marchler-Bauer A et al. (2015), "CDD: NCBI's conserved domain database.", *Nucleic Acids Res.* **43**(D)222-6.
- [3] Marchler-Bauer A et al. (2011), "CDD: a Conserved Domain Database for the functional annotation of proteins.", *Nucleic Acids Res.* **39**(D)225-9.
- [4] Marchler-Bauer A, Bryant SH (2004), "CD-Search: protein domain annotations on the fly.", *Nucleic Acids Res.* **32**(W)327-331.

[Help](#) | [Disclaimer](#) | [Write to the Help Desk](#)
[NCBI](#) | [NLM](#) | [NIH](#)

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

Search nucleotide/protein sequence(s) for conserved domains



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

Category	dsrA (%)
RF +1	100
Specific hits	~80
Superfamilies	~70

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 sulfite 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

1 Cdd:TIGR02064	<p>10 20 30 40 50 60 70 80</p> <p>.....*....*....*....*....*....*....*.... </p> <p>11 LDQLESGPWPSFISGIXKRLRDEHPEERINXKMTNDLIGQLEHSEYETRKGYWKGGTVSVFQYGGCIIPRFSEVGHAFFESKE 90</p> <p>1 LDQLEKGPWPSPVSIEKKTAAYRADYQVVPDPELLGVLLELSYDEERKTHWKGGLIVSVPFGYGGIVIGRYSDQGEKPFPGVAE 80</p>
1 Cdd:TIGR02064	<p>90 100 110 120 130 140 150 160</p> <p>.....*....*....*....*....*....*....*.... </p> <p>91 FHTLRVQPPAGNHYSTDMLRQMADSWEKYGGSLVTFHGQTGNIMFIGIDTEOHDFFEINDYWGDLGGAGPCVRATMSC 170</p> <p>81 FHTVRAQPSGFYSTDYLRLQLCDWWEKYGGSLTNFHGQTGDIVFLGTQTPQLQEIFEELTNLGTDLGGGSNRLTPESC 160</p>
1 Cdd:TIGR02064	<p>170 180 190 200 210 220 230 240</p> <p>.....*....*....*....*....*....*....*.... </p> <p>171 VGAARCEMSCTNEQKAHRLVNNFTDDVHRPALPYKFKFVSGCGNDQNCAVERADFAVIGTWRDDMMNVNQDEFKAYVGR 250</p> <p>161 VGPARCEFACYDTLKACYELTMYEQDELHRPAFPYKFKFSGCPNDCVAAIARSDFAVIGTWKDDIKVQEAVKYIAQ 240</p>
1 Cdd:TIGR02064	<p>250 260 270 280 290 300 310 320</p> <p>.....*....*....*....*....*....*....*.... </p> <p>251 KGKRQHVIDNIITRCPTNALSLNDDDSLEVNNKDCVRCMHCLNVPKALHPGDDRGVTILIGGRKTRLKIIGDLMGTVVVPPF 330</p> <p>241 WGKFDFIENEVNVNRCPTKAISWDGSKELSIDNRECVRCMHCKNMKPALKHPGDERGVVTILIGGKAPILDGAQMGWVVVPPF 320</p>
1 Cdd:TIGR02064	<p>330 340 350 360 370 380 390 400</p> <p>.....*....*....*....*....*....*....*.... </p> <p>331 kldTEEDWEEIVELAEEIIDFWAENALEHERCGEMIERIGLVNPLEGVGVVEVDPMVNPNPRESSYIRMDGWEDEAVKWD 410</p> <p>321 --EAEPPYDEIKELEVKEIIDWWDEEGKNRERIGETIKRGLQKPLEVIGIEPDPQMVKEPRTPNPYIFFKVEDEVPGGWDA 398</p>
1	<p>411 RQAE 414</p> <p>GJL-TIGR02064 399 RQAE 402</p>

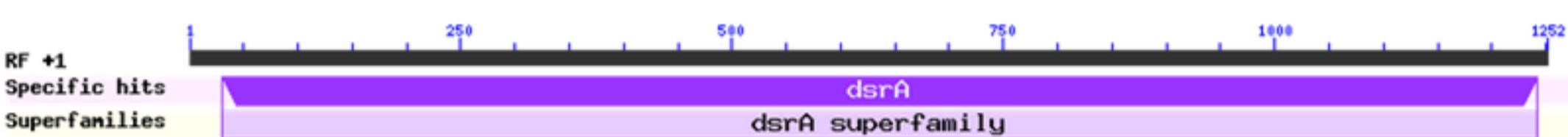
Conserved Domain Search results:

dsrA gene of *Solemya velum* symbiont strain WH



References:

Marchler-Bauer A et al. (2012). "CDD/SPARCLE: functional classification of proteins via subfamily-domain architectures." *Nucleic Acids Res* 40(D):D200-3.



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 ... 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]	31-1242	0e+00

170 180 190 200 210 220 230 240
*....|.....*....|.....*....|.....*....|.....*....|.....*....|.....*....|.....*....|
 1 171 VGAACMCSCTNEQKAHRLVNNFTDDHRPALPKFKFVSCGGCNDQNAVERPAFAGIVTWRDMVNQDEFKAYVGR 250
 Odd: MTCP02056 161 VGPBAPRBCXQYTFKACSYTPEPDRYKVKFVSCGGCNDQNAVERPAFAGIVTWRDMVNQDEFKAYVGR 240

250 260 270 280 290 300 310 320
*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|.....*.....|
 1 251 KGRHVIDN1ITRCPTNLNSLNDDSLLEVNNKDCVRCMHCLNVNPVKHLPGDDRGVITLILIGKKRTLK1GDLMTGVVPFKF 330
 252 MGRHVIDN1ITRCPTNLNSLNDDSLLEVNNKDCVRCMHCLNVNPVKHLPGDDRGVITLILIGKKRTLK1GDLMTGVVPFKF 330

330 340 350 360 370 380 390 400
*....|.....*....|.....*....|.....*....|.....*....|.....*....|.....*....|
 1 331 kLDTEDEWDRIVELAEEIIDFWAENALEHERCGEMIERIGLIVNFILEGVGVVEVDPPNNVNNPRESSTYIRMDGWDDEAVKWFED 410
 332 DLEDDWDRIVELAEEIIDFWAENALEHERCGEMIERIGLIVNFILEGVGVVEVDPPNNVNNPRESSTYIRMDGWDDEAVKWFED 300

1 411 RQAE 414
Cdd:TIGR02064 289 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

The screenshot shows the Metacyc website interface. At the top left is the Metacyc logo with the text "A member of the BioCyc database collection". To its right is a blue banner for a "Two-day Introduction to Pathway Tools Tutorial" on "Jan 16-17, 2020" with "Early Registration by Dec 12". Below the banner is a green navigation bar with links: "Sites", "Search", "Genome", "Metabolism", "Analysis", "SmartTables", and "Help".

Search Results for **dsra**

using database **MetaCyc** what is this?

[Genes \(3\)](#) | [Proteins \(3\)](#) | [EC Numbers \(2\)](#)

Genes Gene/Gene Product pages contain: chromosomal location of gene; depiction of its operon; link to genome browser; detailed summaries; complexes); cofactors, activators, and inhibitors (for enzymes), depiction of regulon (for transcriptional regulators), protein features.

- [dsrA - *Allochromatium vinosum*](#)
- [dsrA - *Desulfovibrio gigas*](#)
- [dsrA - *Archaeoglobus fulgidus*](#)

[Login](#) to turn into a SmartTable.

Proteins Gene/Gene Product pages contain: chromosomal location of gene; depiction of its operon; link to genome browser; detailed summaries; regulon (for transcriptional regulators), protein features.

- siroheme sulfite reductase, α subunit (*DsrA*) - *Allochromatium vinosum*
- sulfite reductase α subunit (*DsrA*) - *Desulfovibrio gigas*
- sulfite reductase, dissimilatory α subunit (*DsrA*) - *Archaeoglobus fulgidus*

[Login](#) to turn into a SmartTable.

EC Numbers EC Number pages contain: links to reaction and enzymes associated with the EC number in this database, names, description.

<https://metacyc.org/>



Gene annotation

Metacyc: experimentally curated metabolic pathways

Two-day Introduction to Pathway Tools Tutorial Jan 16-17, 2020 Early Registration by Dec 12

Sites | Search | Genome | Metabolism | Analysis | SmartTables | Help

Search Results for **dsra** using database **MetaCyc** what is this?

Genes (3) | Proteins (3) | EC Numbers (2)

Genes Gene/Gene Product pages contain: chromosomal location of gene; depiction of its operon; complexes; cofactors, activators, and inhibitors (for enzymes); depiction of regulon (for transcriptional regulators).

dsrA - Allochromatium vinosum
dsrA - Desulfovibrio gigas
dsrA - Archaeoglobus fulgidus

Proteins Gene/Gene Product pages contain: chromosomal location of gene; depiction of its operon; complexes; cofactors, activators, and inhibitors (for enzymes); depiction of regulon (for transcriptional regulators), protein features.

- siroheme sulfite reductase, α subunit (**DsrA**) - *Allochromatium vinosum*
- sulfite reductase α subunit (**DsrA**) - *Desulfovibrio gigas*
- sulfite reductase, dissimilatory α subunit (**DsrA**) - *Archaeoglobus fulgidus*

EC Numbers EC Number pages contain: links to reaction and enzymes associated with the EC number.

Sites | Search | Genome | Metabolism | Analysis | SmartTables | Help

gene **dsrA** protein **sulfite reductase α complex**
Desulfovibrio gigas

Accession ID G-385 (MetaCyc)

Reactions thiosulfate + 2 an oxidized unknown electron carrier + 3 H₂O → 2 sulfite + 2 a reduced unknown electron carrier + 3 H⁺ (catalyzed by complex a [DsrC]-trisulfide + an oxidized unknown electron carrier + 3 H₂O → sulfite + a [DsrC protein]-dithiol + a reduced unknown electron carrier + 3 H⁺)
trithionate + an oxidized unknown electron carrier + 3 H₂O → 3 sulfite + a reduced unknown electron carrier + 4 H⁺ (catalyzed by complex)

Pathways dissimilatory sulfate reduction I (to hydrogen sulfide)
dissimilatory sulfate reduction II (to thiosulfate)

Summary | GO Terms (1) | Reactions (3) | References | Show All

Subunit Composition [DsrA]₂
Component of dissimilatory sulfite reductase (extended summary available): [(DsrA)₂][(DsrB)₂]

Gene-Reaction Schematic

+ 1.8.99.5 : trithionate + an oxidized unknown electron c...
- 1.8.99.5 : a [DsrC]-trisulfide + an oxidized unknown el...
1.8.99.5 : thiosulfate + 2 an oxidized unknown electron...
1 2 Dg-dsrB
1 2 Dg-dsrA

Gene: dsrA G-385
Product: sulfite reductase α subunit, subunit of sulfite reductase α complex, dissimilatory sulfite reductase
Species: *Desulfovibrio gigas*

GO Terms:
Cellular Component: GO:0005829 - cytosol

Enzymatic activity: sulfite reductase (thiosulfate-forming) (dissimilatory sulfite reductase)

2 sulfite + 2 a reduced unknown electron carrier + 3 H⁺ → thiosulfate + 2 an oxidized unknown electron carrier + 3 H₂O

<https://metacyc.org/>

Gene annotation

Accurate classifier for hydrogenase sequences - HydDB

HydDB  Classify  Browse  Information Pages ▾

Classify

HydDB provides access to an accurate classifier for hydrogenase sequences and a curated database of hydrogenases by known type. The service is provided by the School of Biological Sciences, Monash University and the Bioinformatics Research Centre, Aarhus University.

Please cite! If you use HydDB for research, please cite the following paper: "Søndergaard D, Pedersen CNS, Greening C. **HydDB: A web tool for hydrogenase classification and analysis**. Sci Rep. 2016;6:34212. doi: 10.1038/srep34212.". The preprint is available on [bioRxiv](#). If you have any comments, corrections or questions contact [Chris Greening](#) or [Dan Søndergaard](#).

Classify

HydDB is unable to accurately check whether uploaded sequences correspond to hydrogenases or not. Instead, it is well-suited for functionally-predictive classification of known hydrogenases into different subgroups. Please ensure that all sequences that you upload correspond to catalytic subunits of hydrogenases (e.g. using conserved domain database and phylogenetic trees). Sequences that do not encode catalytic subunits of hydrogenases will still be classified, but the result may be wrong.

Sequences

Sequences File

Choose File no file selected

Instructions

To use the classifier to predict the type of one or more hydrogenases from sequence, either:

- paste your FASTA-formatted protein sequences into the text area, or
- upload a FASTA-formatted file with your protein sequences.

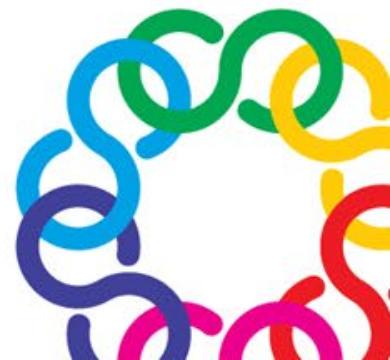
Press the "Submit" button to upload the sequences and begin the classification.

If you provided an e-mail address you will receive an e-mail when your job finishes or fails including a link to the results. You will also be able to download the results as a CSV file.

Only sequences encoding the catalytic subunits of hydrogenases will be classified, i.e. those binding the [NiFe]-centre (NiFe-hydrogenases), [FeFe]-centre (FeFe-hydrogenases), or [Fe]-centre (Fe-hydrogenases). Electron-transfer subunits, accessory proteins, and maturation factors cannot be classified by this service.

Limits

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



Gene annotation

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



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/>

Choose an organism type (?): Required

Choose Gram stain (?): Required

Advanced Gram stain options (?): Required

Output format (?):

Show results (?):

Email address:

Copy and paste your FASTA sequences below

or upload from file: no file selected
(uploads limited to 50KB, approximately 100 proteins, in Web display mode, enter an email address to use email mode if you need to analyze more proteins)



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	
------------	--------------	------	------------------	-----	-----------------	----------	-----------	--

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

Submit data

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

i Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 5000.

i 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:

No file selected.

Organism

Eukarya
 Other
i "Eukarya" only predicts Sec/SPI SPs.

Output format:

Long output
 Short output (no figures)

Model mode:

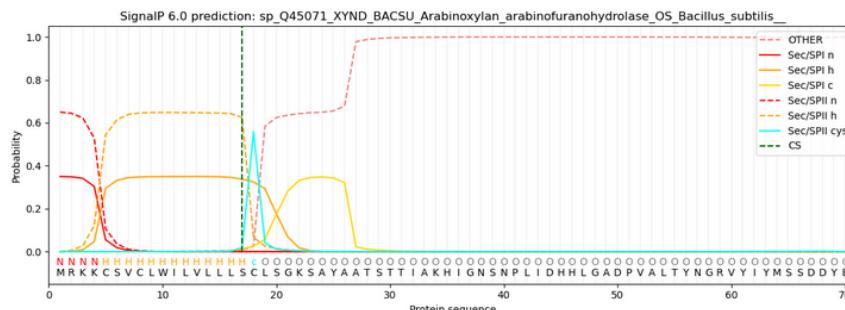
Fast

Slow

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

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

Information on EC 3.2.1.23 - beta-galactosidase

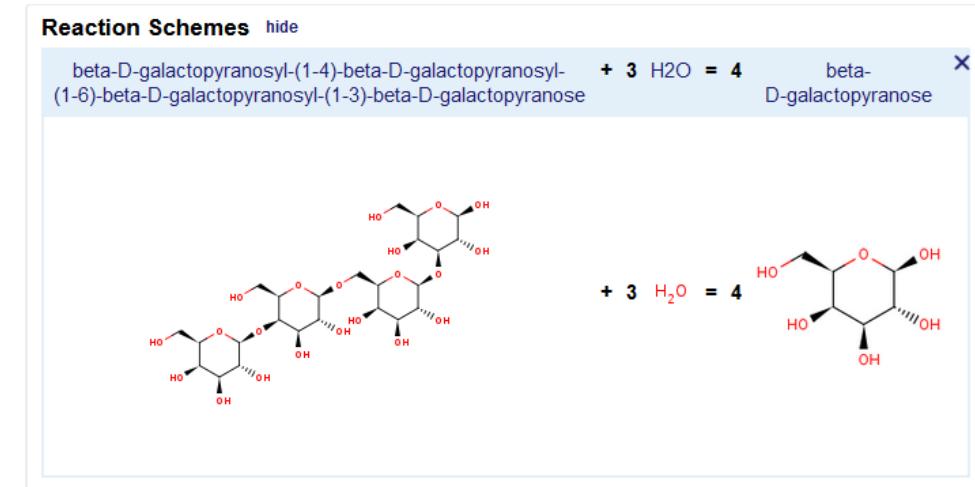
for references in articles please use BRENDA:EC3.2.1.23

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](#)

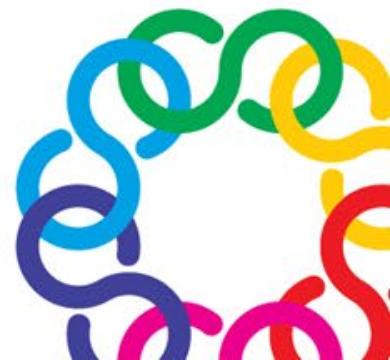
PATHWAY SOURCE ▲▼	PATHWAYS ▲▼
BRENDA	metabolism of disaccharids
KEGG	Galactose metabolism , Glycosaminoglycan degradation , Glycosphingolipid biosynthesis - ganglio series , Other glycan degradation , Sphingolipid metabolism
-	Sphingolipid metabolism



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 Enter keywords

[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
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
01210	2-Oxocarboxylic acid 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

