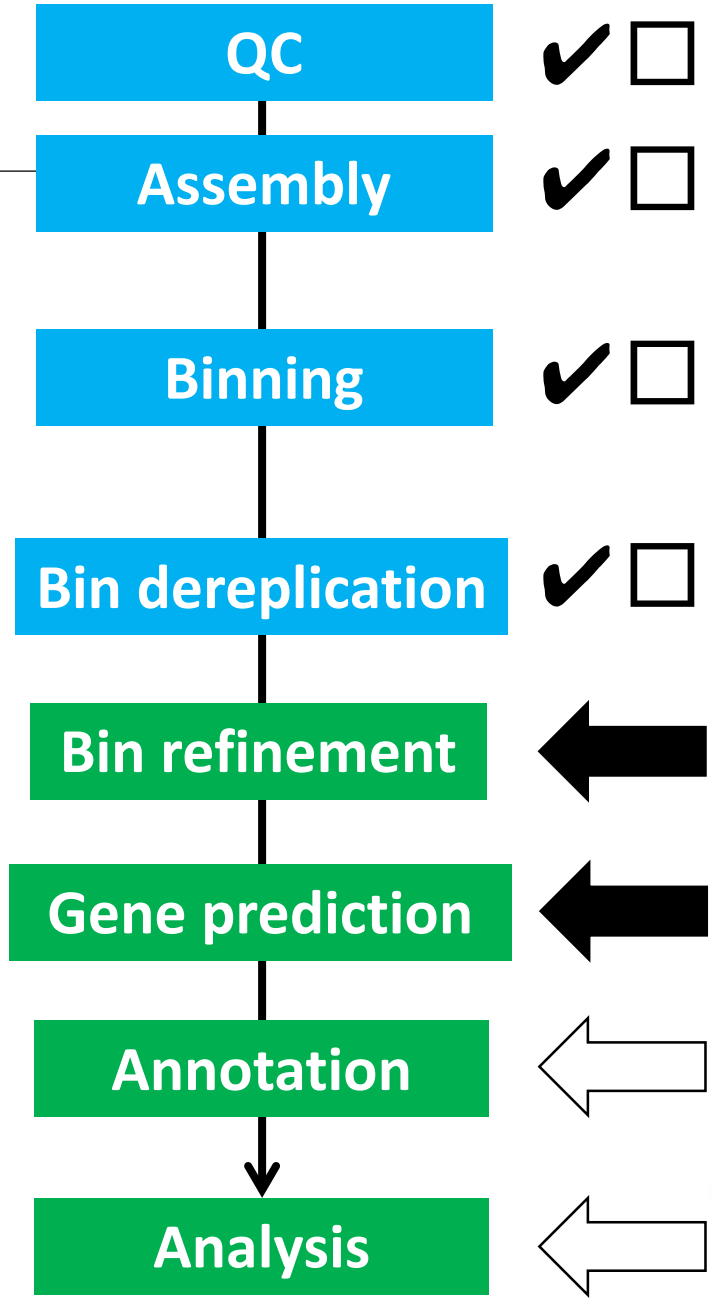
Day 3

Bin refinement
Gene prediction
Gene annotation

Day overview

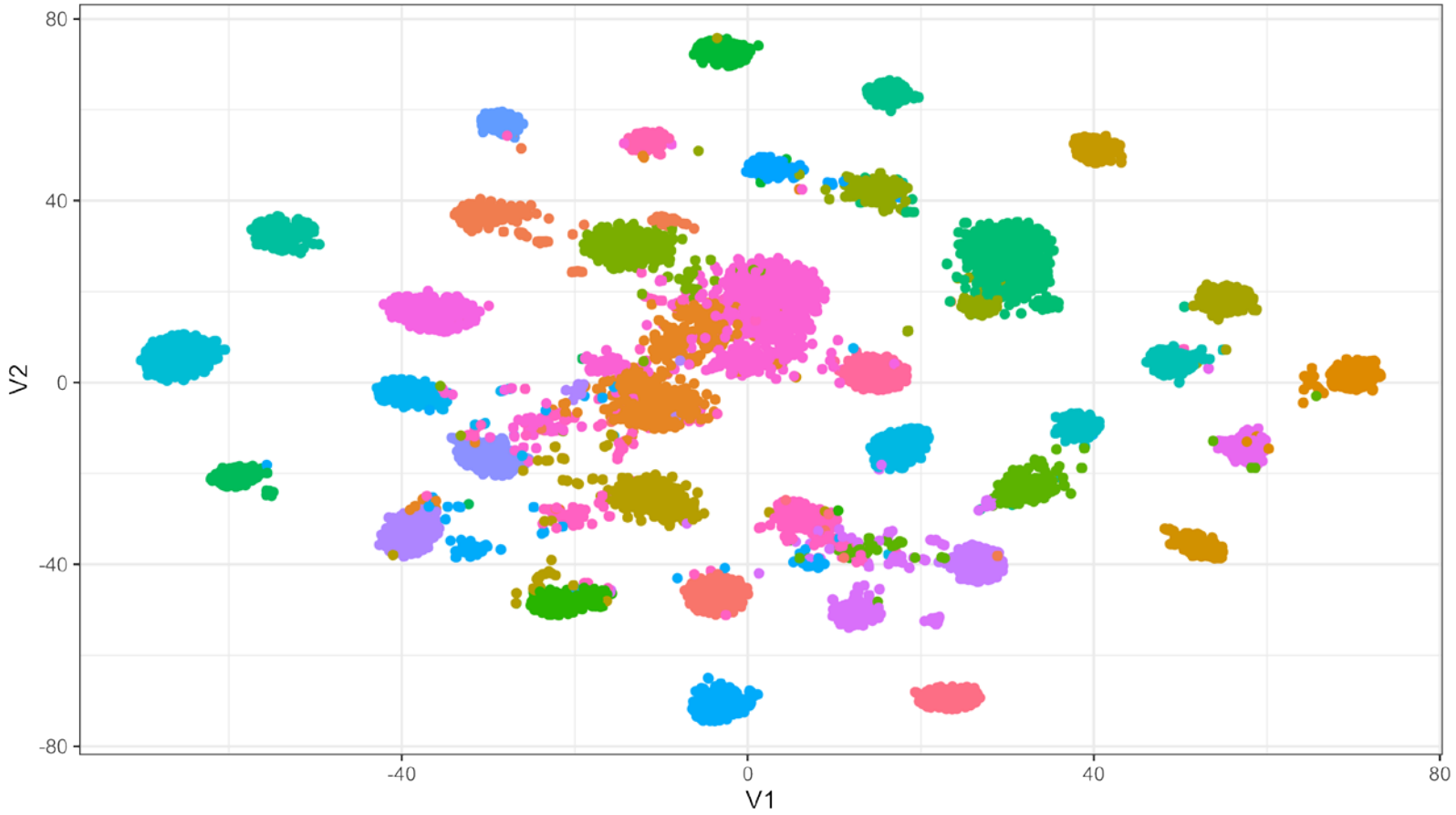
- **Goals:**
  - **Bin refinement**
  - **Gene prediction**
  - **Annotation**
  - **Start analysis**



Bin refinement

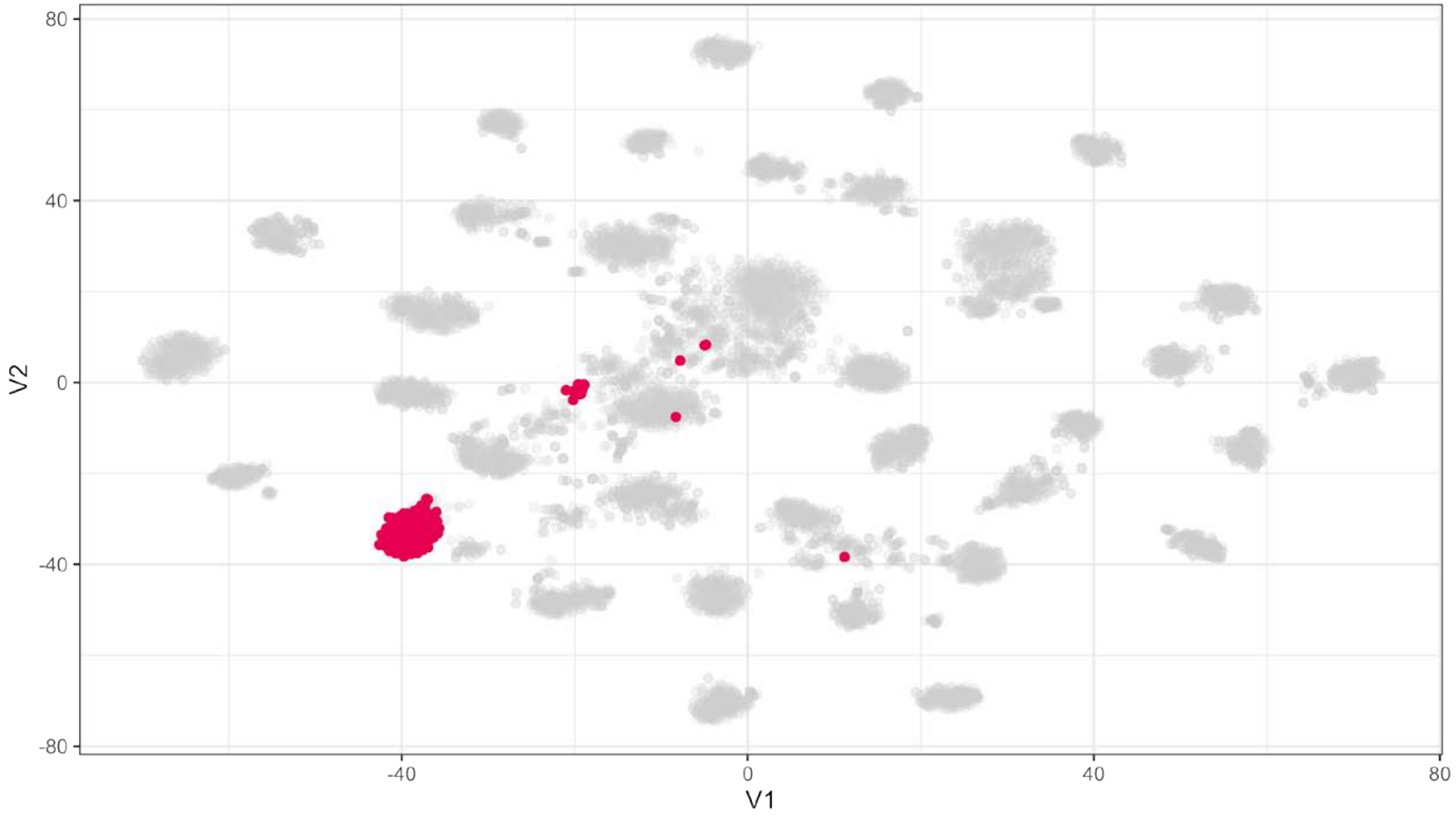
VizBin

- **Inspect bin assignments**



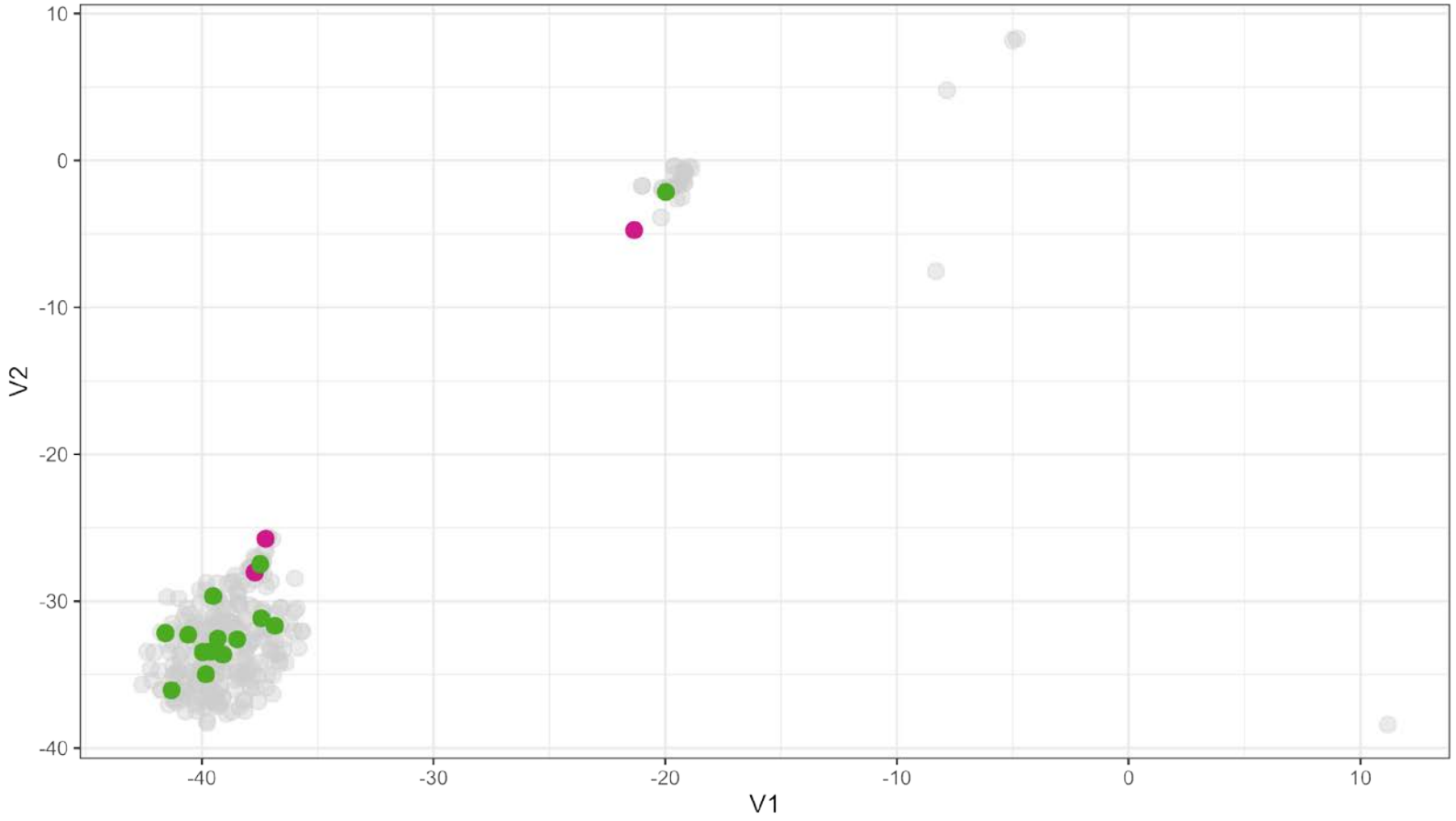
VizBin

- **Use graphical user interface to assign bins or reassign contigs**



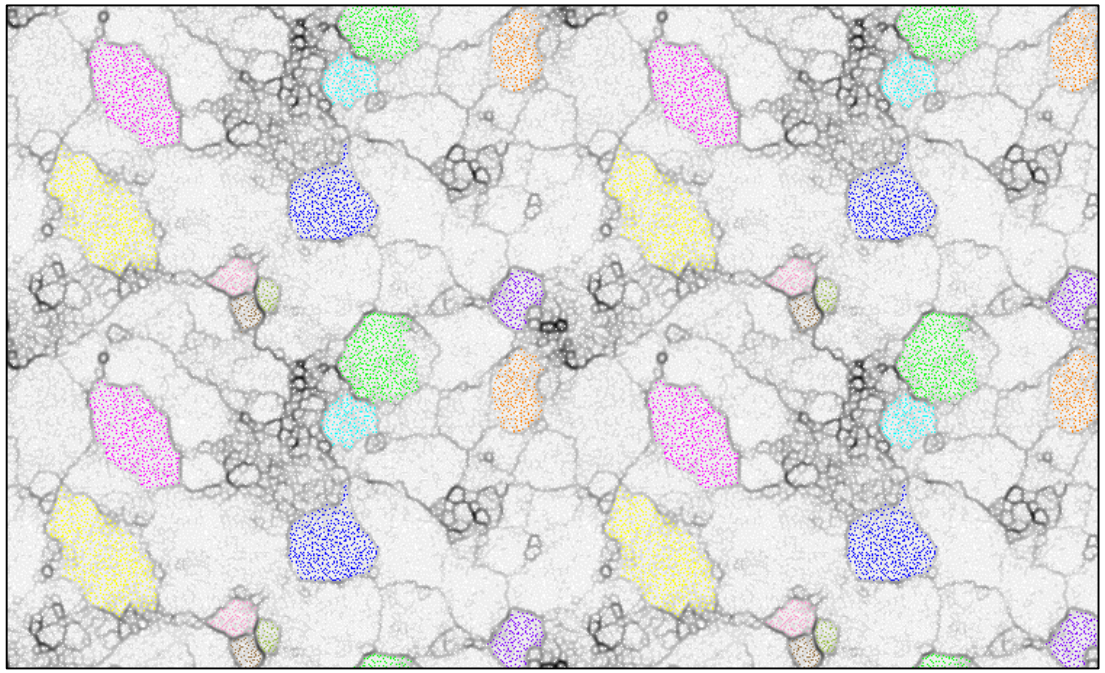
VizBin

- **Identify contigs with unstable placement**



Alternative: ESOM

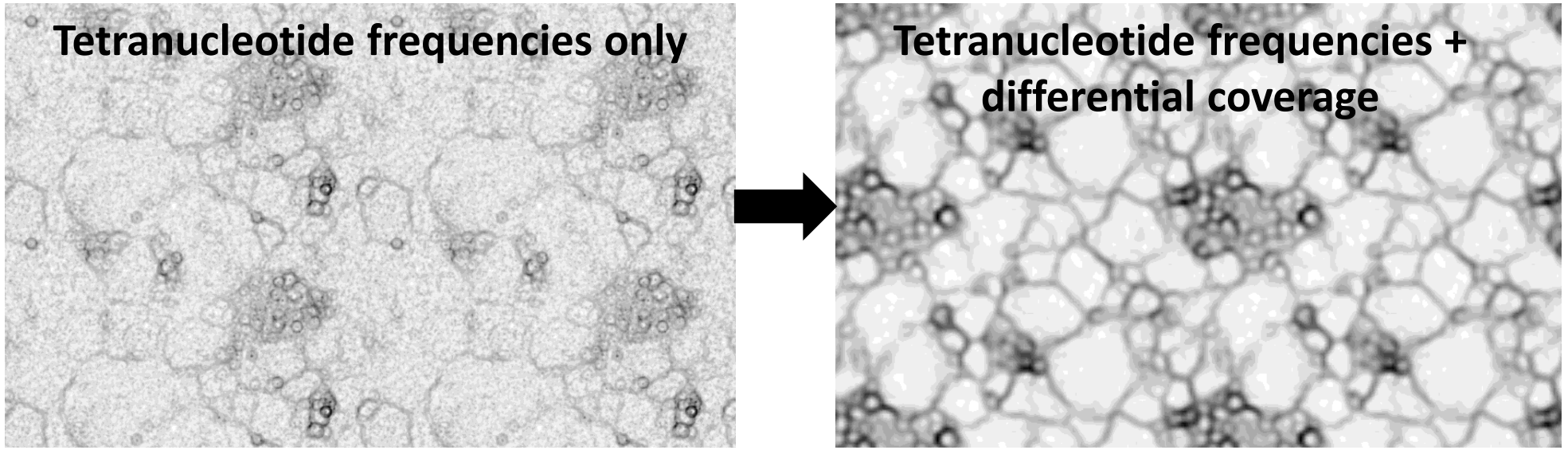
- **Emergent Self Organizing Map (ESOM)**
- **esomana: http://databionic-esom.sourceforge.net/user.html**



Gulf of Mexico seafloor communities (Handley et al., 2017, ISME J)

Alternative: ESOM

- **Flexible: use whatever data you want, e.g.: tetranucleotide frequencies, coverage, both**



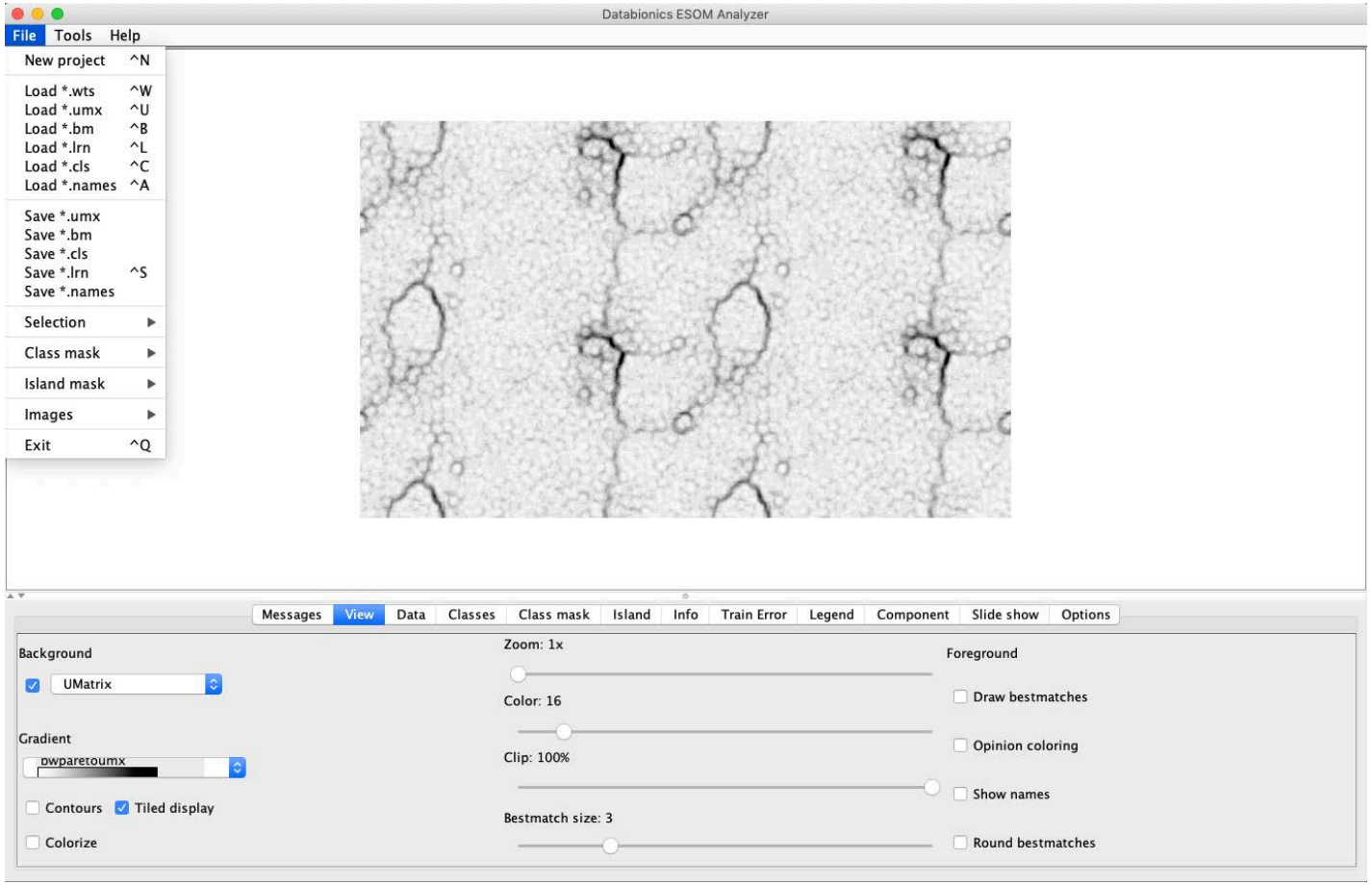
Indistinct bin boundaries between highly similar genomes

Clear bin boundaries between highly similar genomes using spatial gradient data

Gulf of Mexico seafloor communities (Handley et al., 2017, ISME J)

Alternative: ESOM

Topographic map of clusters



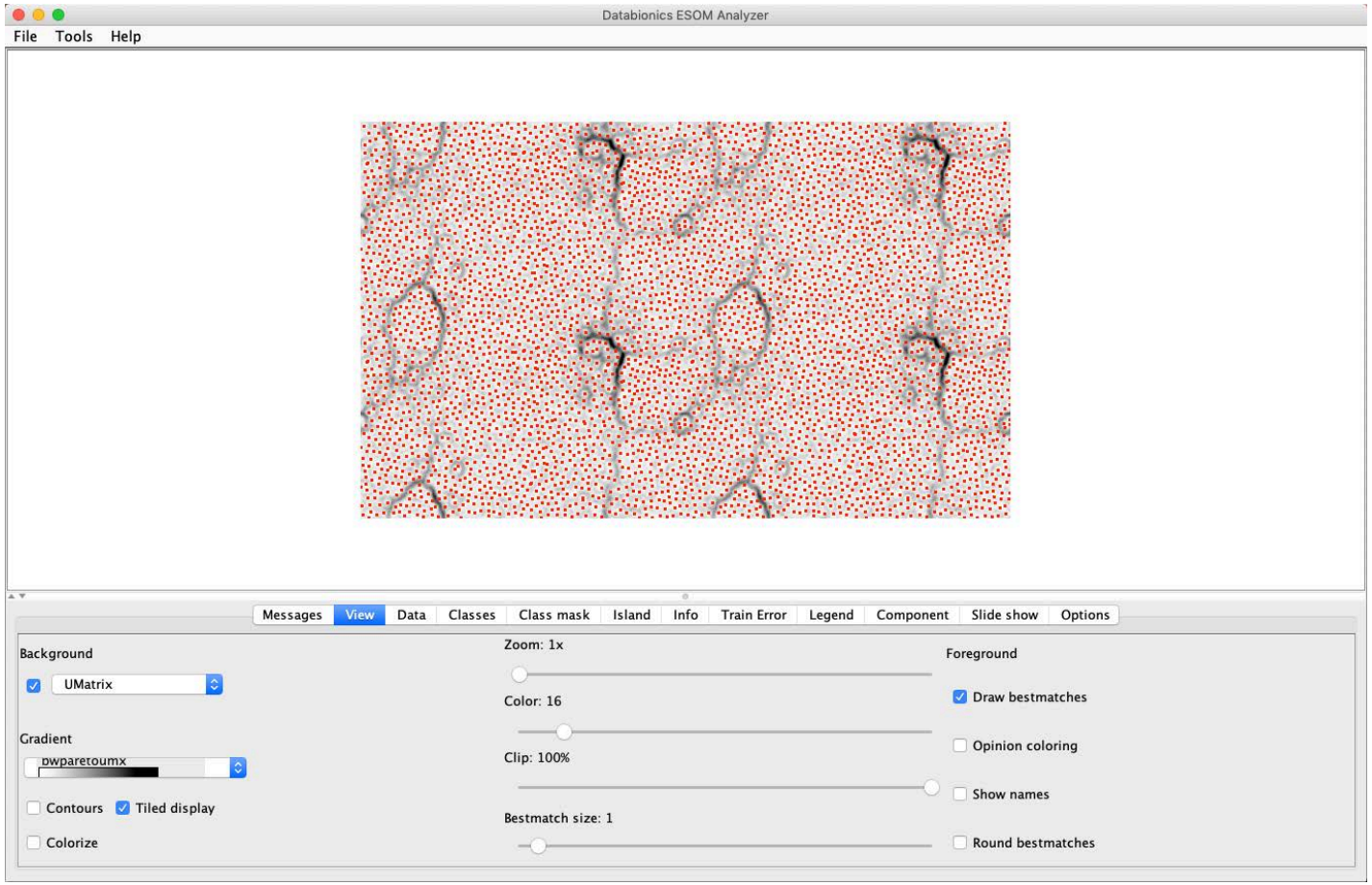
- Uses graphical user interface
- Must supply own prepared data (e.g. pre-calculate tetranucleotide frequencies)

- Dark lines = bin boundaries
- Strong lines = strong bin divisions

Colorado groundwater communities (Handley et al., 2013, ISME J)

Alternative: ESOM

Map with contig fragments shown in red



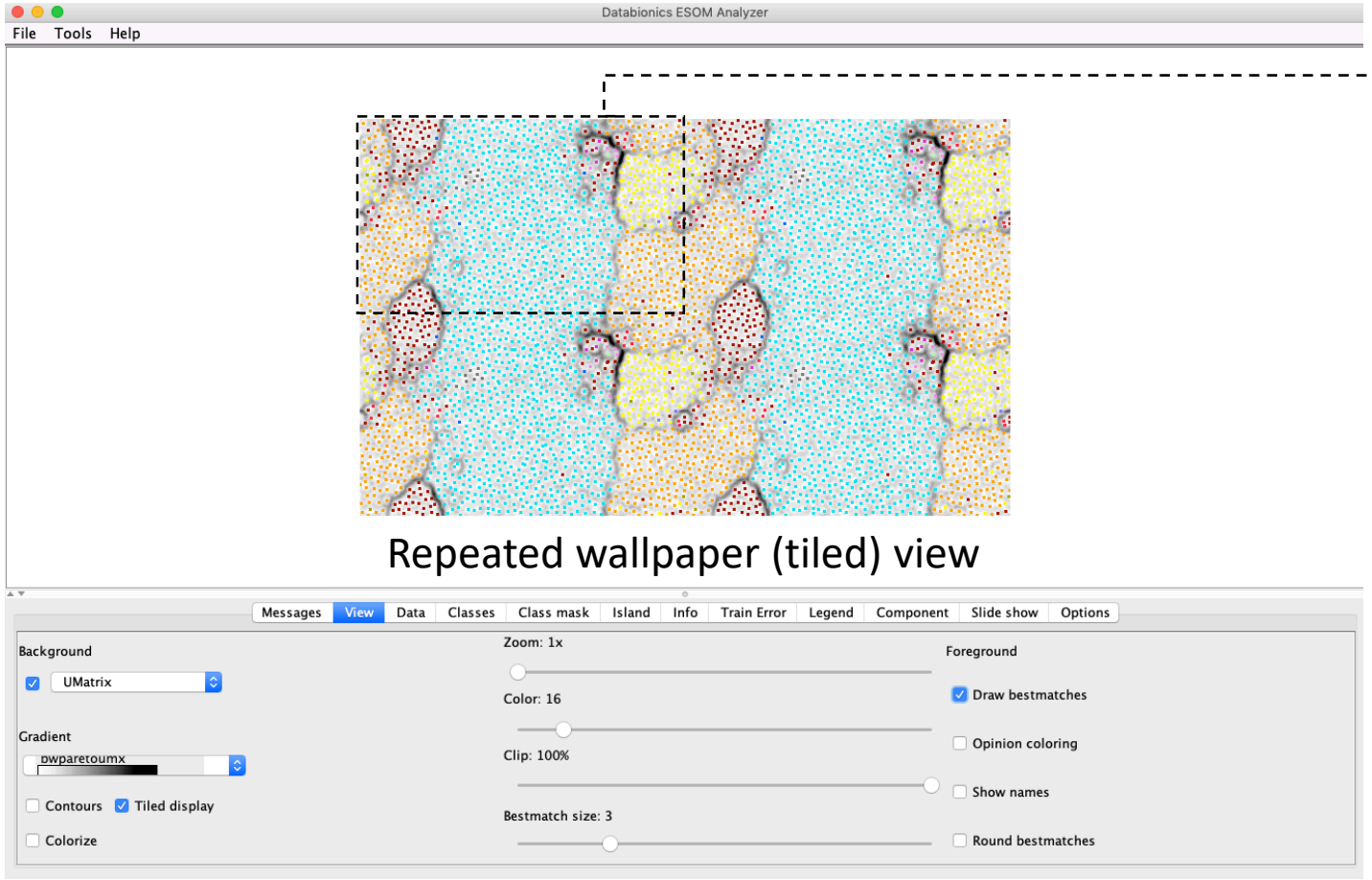
- Uses graphical user interface
- Must supply own prepared data (e.g. pre-calculate tetranucleotide frequencies)

- Dark lines = bin boundaries
- Strong lines = strong bin divisions

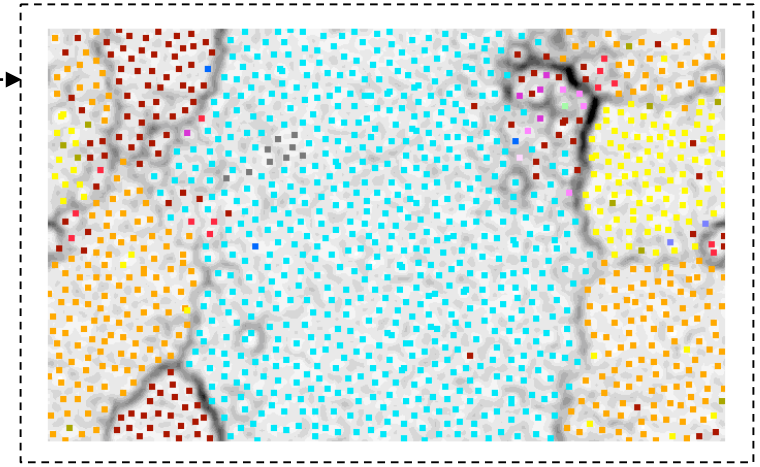
Colorado groundwater communities (Handley et al., 2013, ISME J)

Alternative: ESOM

Map with contig fragments shown and coloured by bin assignment



Repeated wallpaper (tiled) view

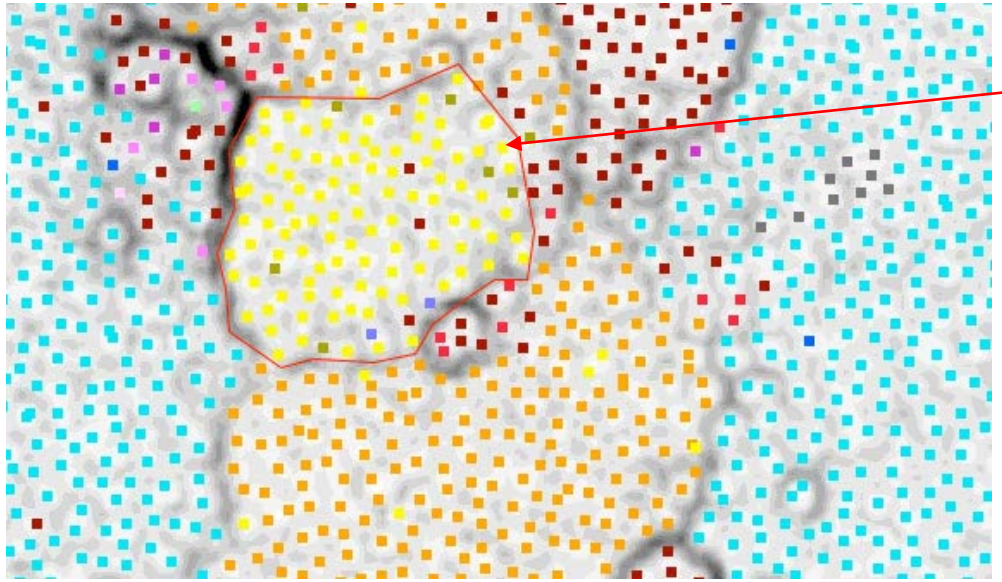


Non-repeating (un-tiled) view

Colorado groundwater communities (Handley et al., 2013, ISME J)

Alternative: ESOM

Map with contig fragments shown and coloured by bin assignment

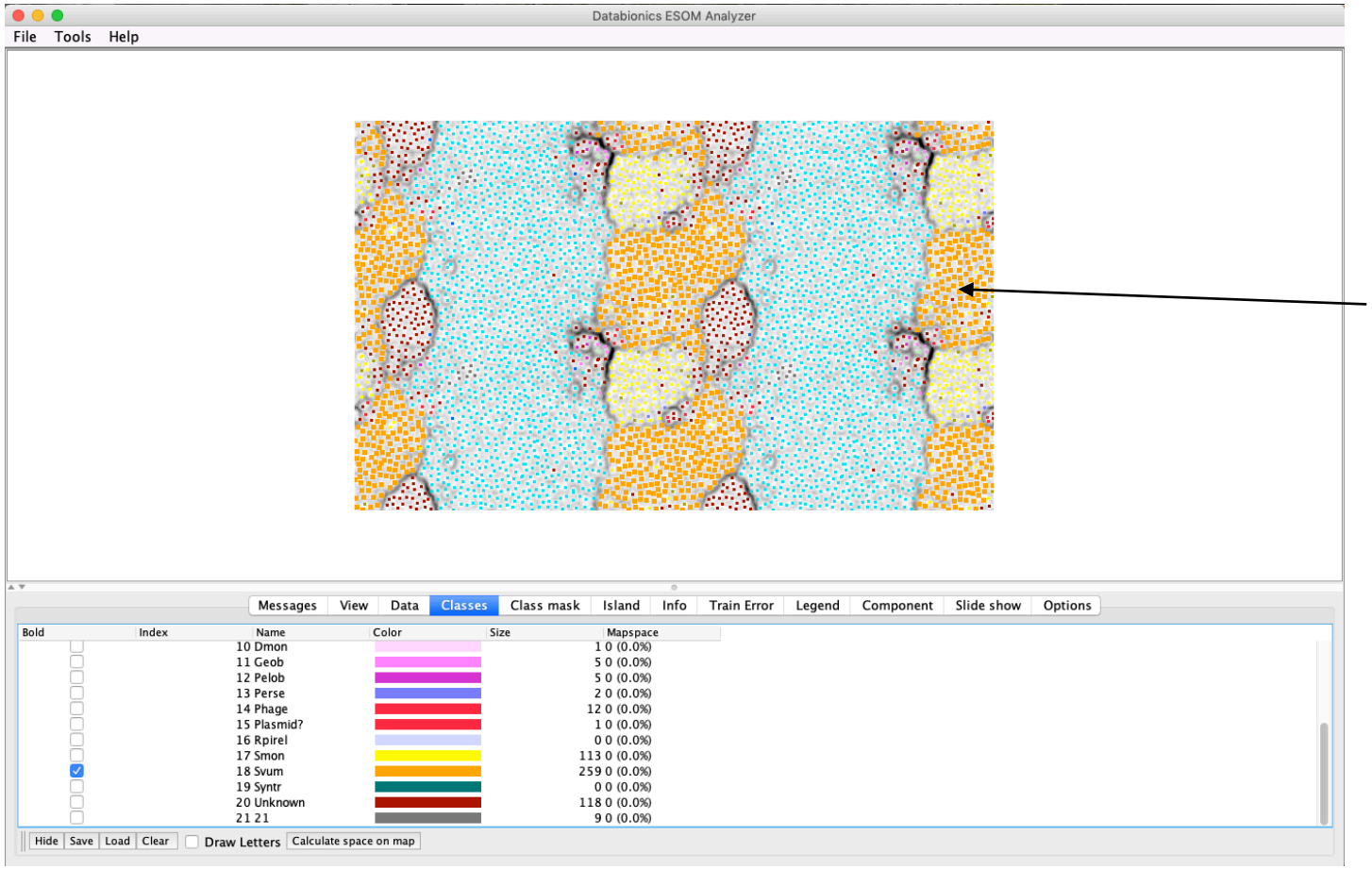


Like VizBin, bins are selected by manually drawing around boundary

Colorado groundwater communities (Handley et al., 2013, ISME J)

Alternative: ESOM

Map with contig fragments shown and coloured by bin assignment



- Select pre-assigned bin to highlight contig fragments
- Choose/change bin colours
- Example: *Sulfurovum* bin highlighted

Colorado groundwater communities (Handley et al., 2013, ISME J)

Task: Work with VizBin

Use VizBin to:

- Prepare input files for VizBin
- Project high-dimensional data down into a 2D plot
- Pick refined bins

Gene prediction

Gene prediction and annotation

- Genome prediction 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

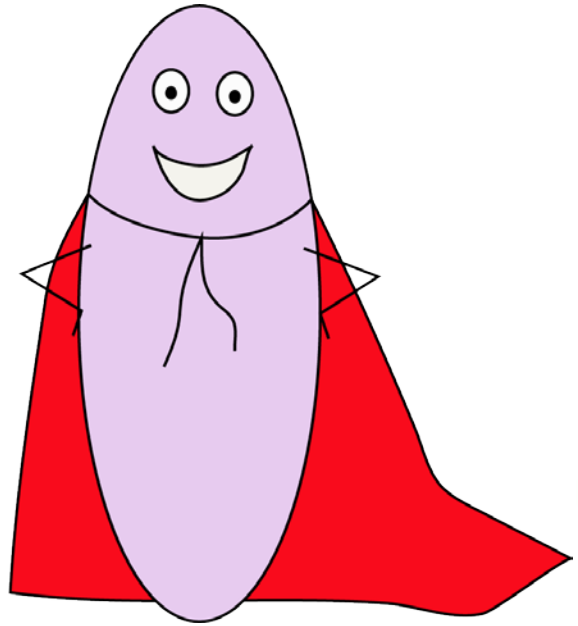
Aim:

- 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 algorithms:**

- Markov Models
- Dynamic Programming
- Linear discriminant analysis
- Linguist methods
- Neural Network

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/
- 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)

- http://compbio.ornl.gov/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

Preparing data for gene prediction

1. Identify and prepare input files for each gene prediction tool (Prodigal, FragGeneScan, MeTaxa2 and Aragorn)
2. Configure parameters for gene prediction

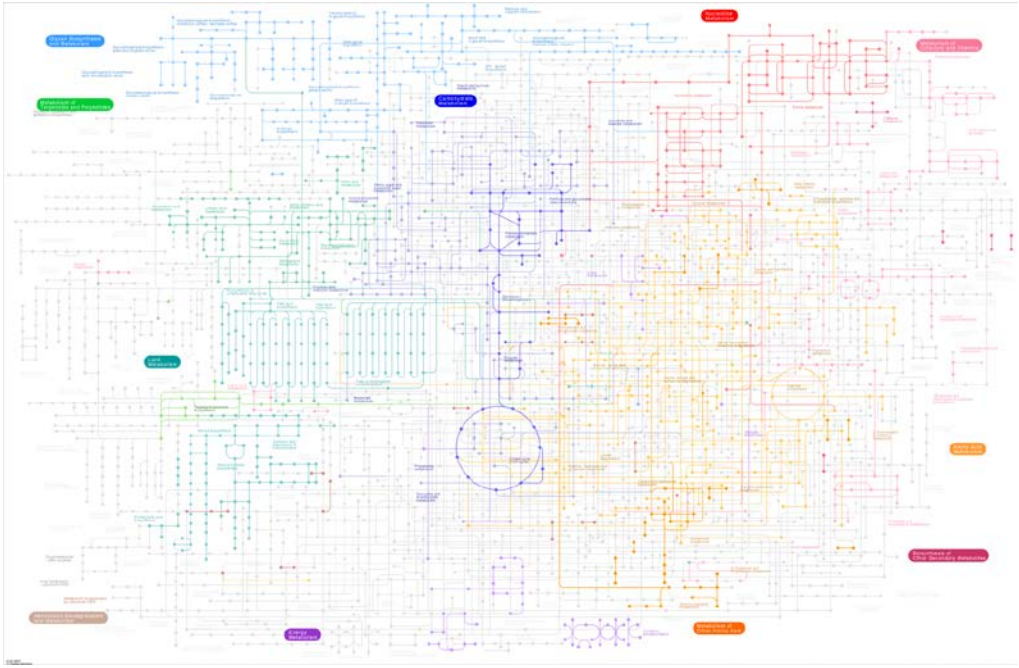
Perform gene prediction

1. Run each job directly from the node (no slurm script required)

Gene annotation

Gene annotation

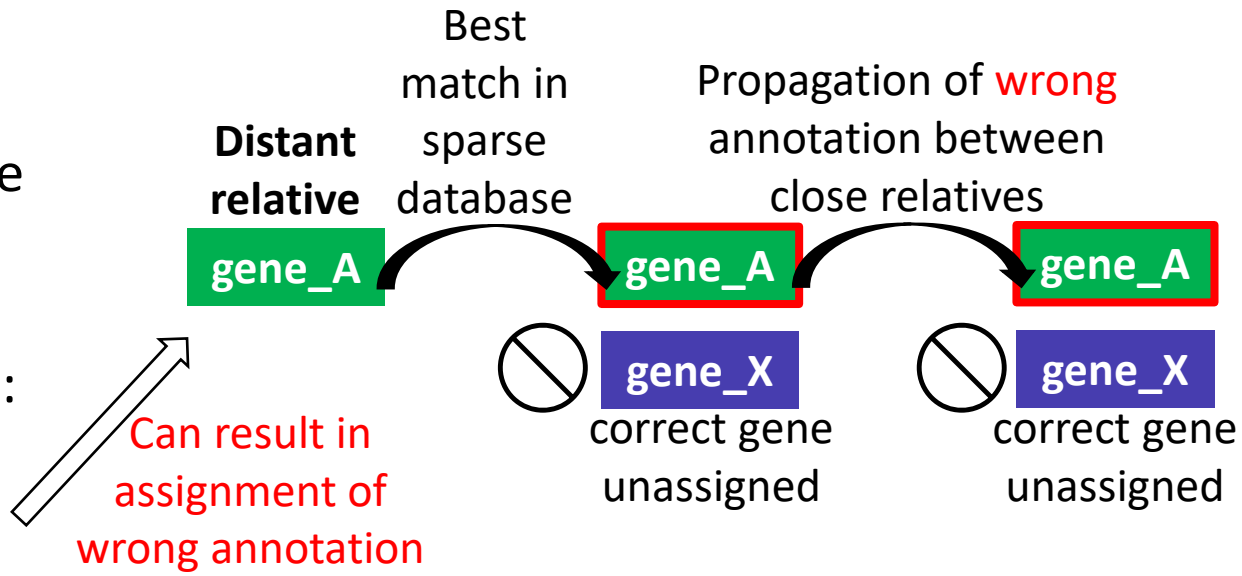
- Genome annotation attempts to predict gene function
- Predicted genes or protein sequences are compared against a curated set of reference sequences for which function is known, or is strongly suspected



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

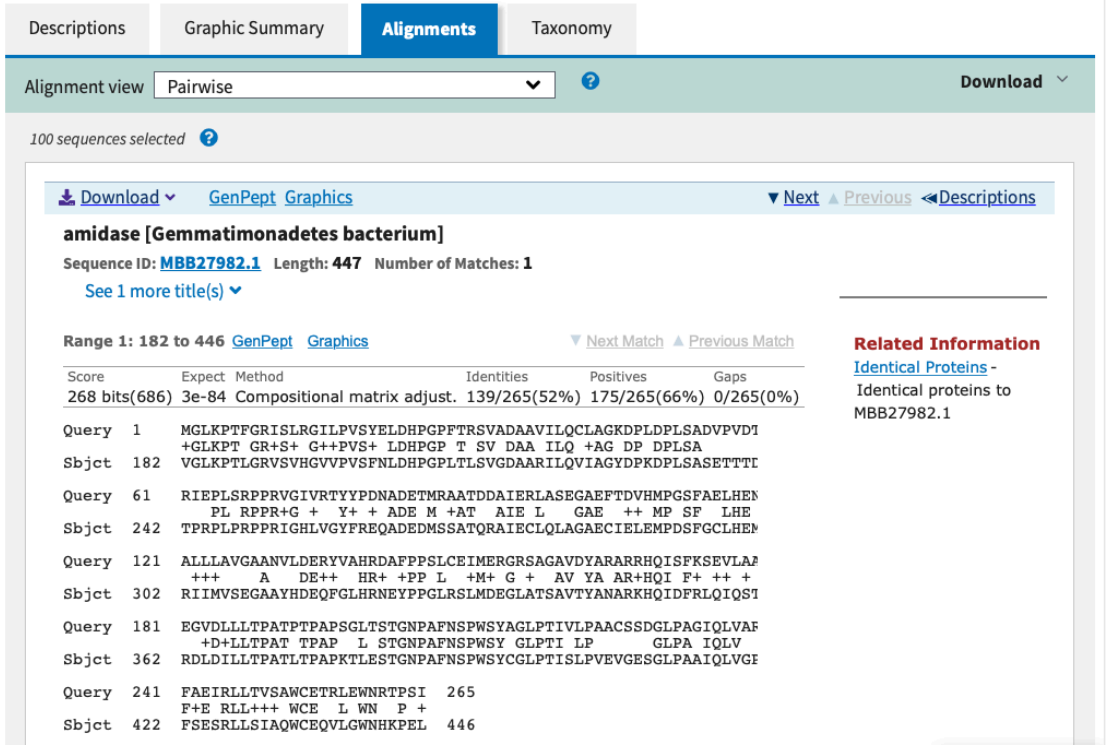
There are two main ways to perform gene annotation with protein sequences:

- 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://pfam.xfam.org)

  Focused more on protein domains based on hidden Markov models

- TIGRfam (https://www.jcvi.org/tigrfams)

  Database of protein family definitions based on hidden Markov models

Gene annotation

Common functional databases (continued)

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

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

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

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

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

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

  Subscriptions are required to access most of BioCyc

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

  2722 pathways from 3009 different organisms

Gene annotation

Graphical User Interface – MEGAN

- Toolbox for interactively analyzing microbiome data.
  - Taxonomic analysis using the NCBI taxonomy or a customized taxonomy such as SILVA
  - Functional analysis using InterPro2GO, SEED, eggNOG or KEGG
  - Bar charts, word clouds, Voronoi tree maps and many other charts
  - PCoA, clustering and networks
  - Supports metadata

https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/megan6/

Gene annotation

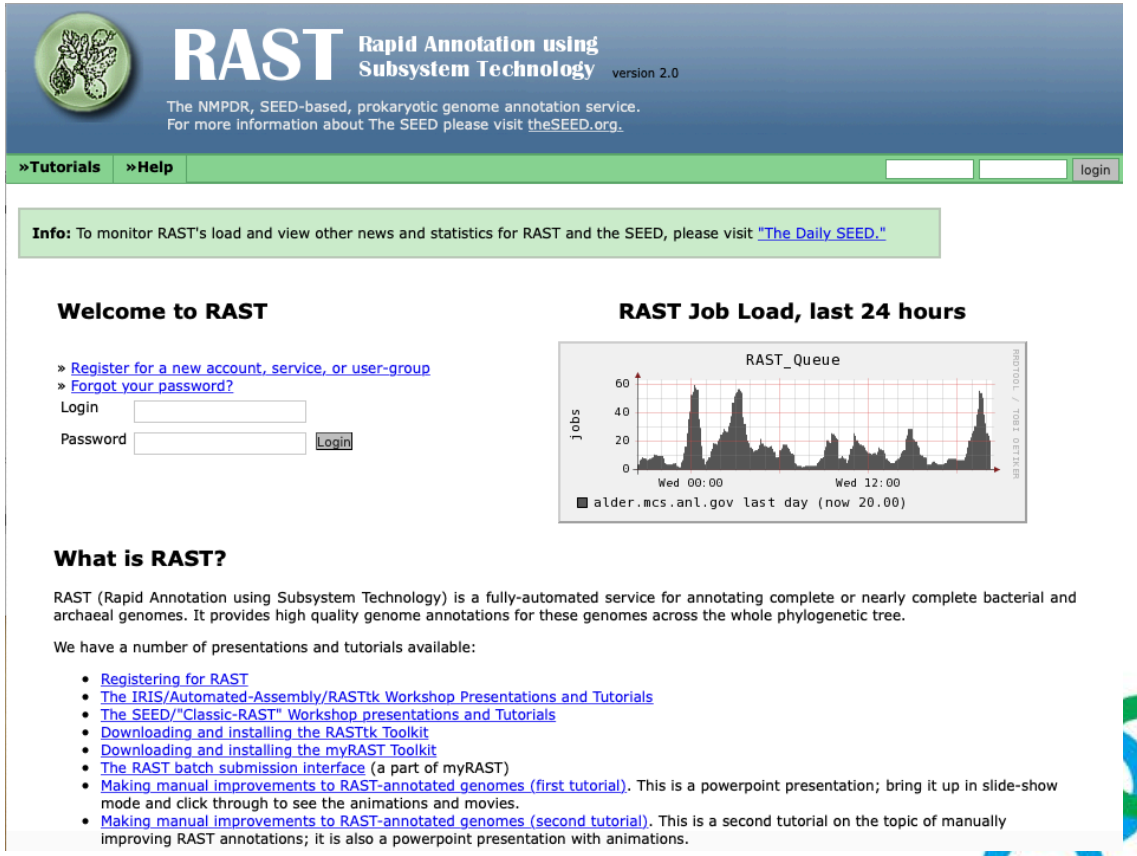
Some web-based annotation tools:

- 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

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

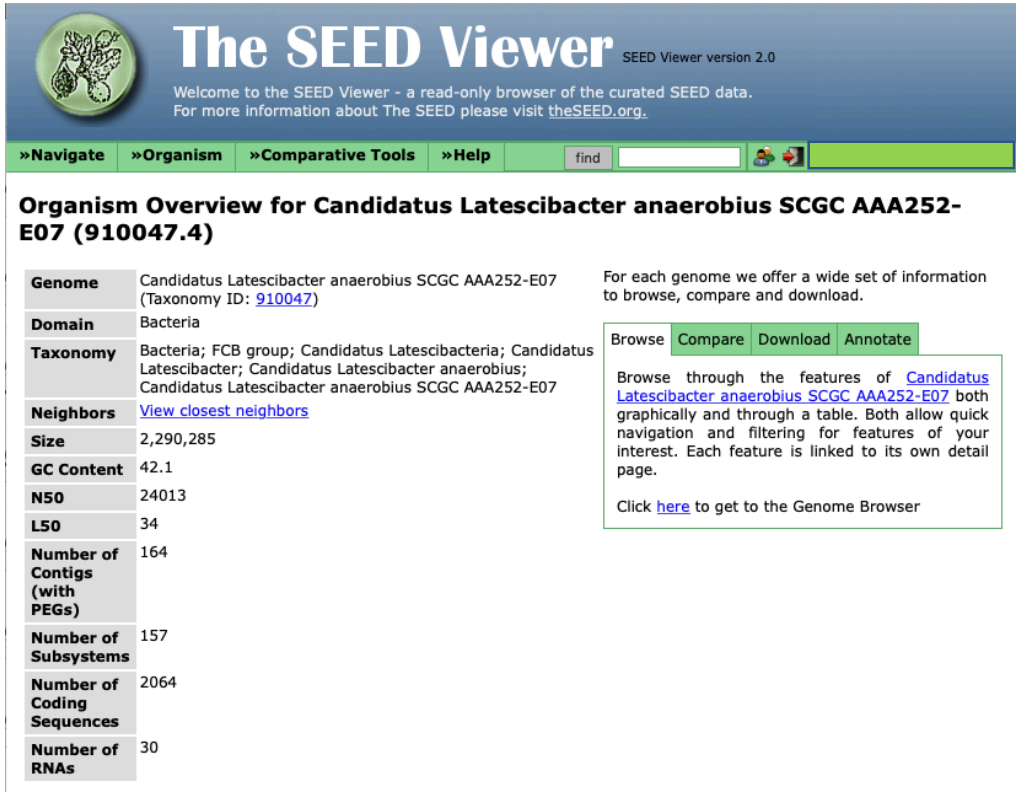
- Fast annotation (~1 genome/day)
- Can use for individual genome bins
- It works well for genomes similar to large groups of reference genomes
- As usual: requires manual curation after initial annotation

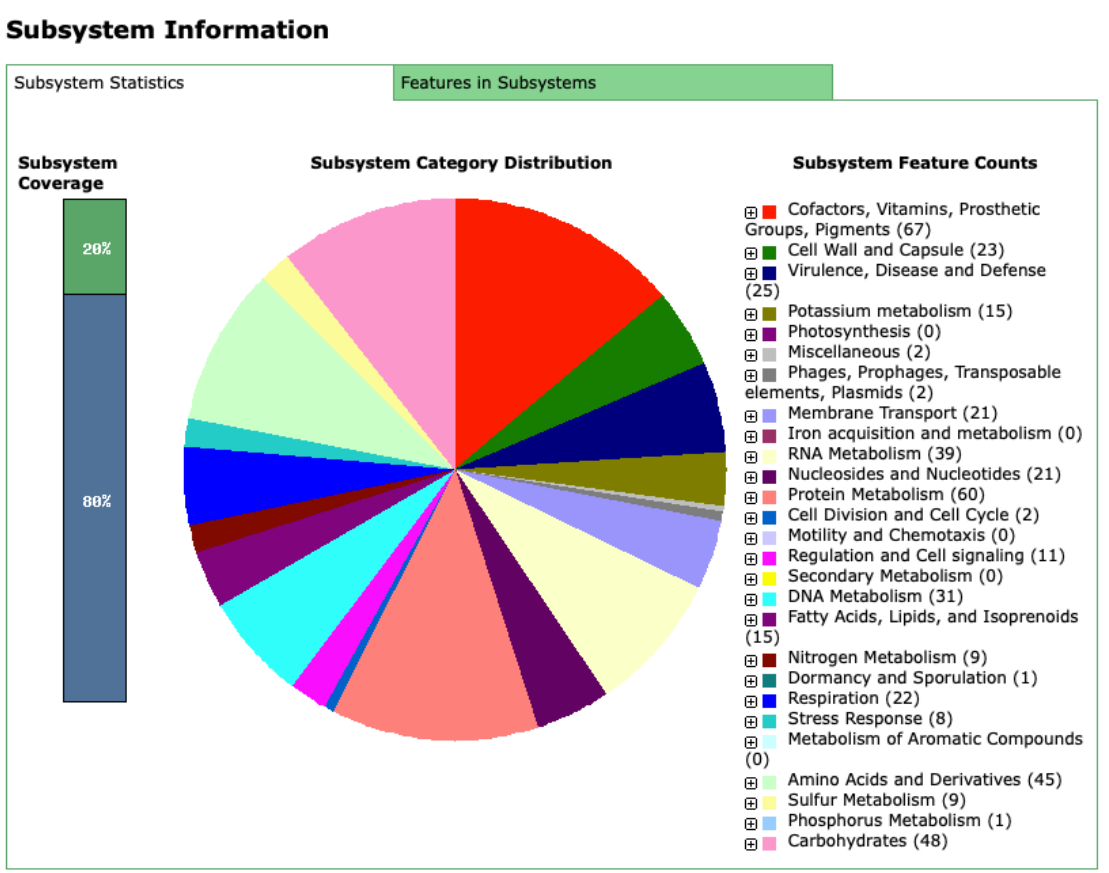


http://rast.theseed.org/FIG/rast.cgi

Gene annotation

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





Task: Gene annotation

Preparing data for gene annotation

1. Identify and prepare input files for gene annotation with Diamond
2. Configure parameters for gene annotation

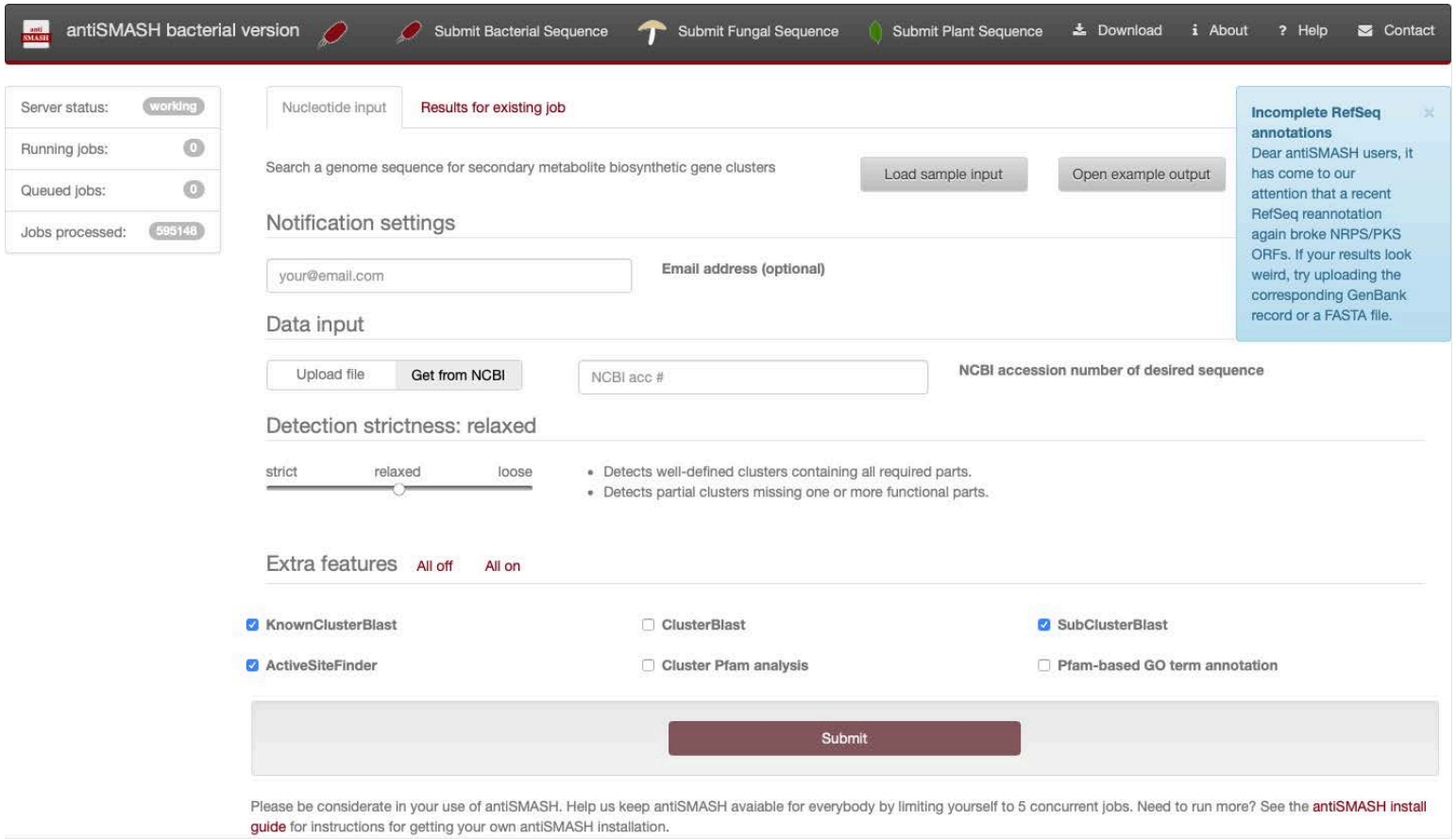
Perform gene annotation

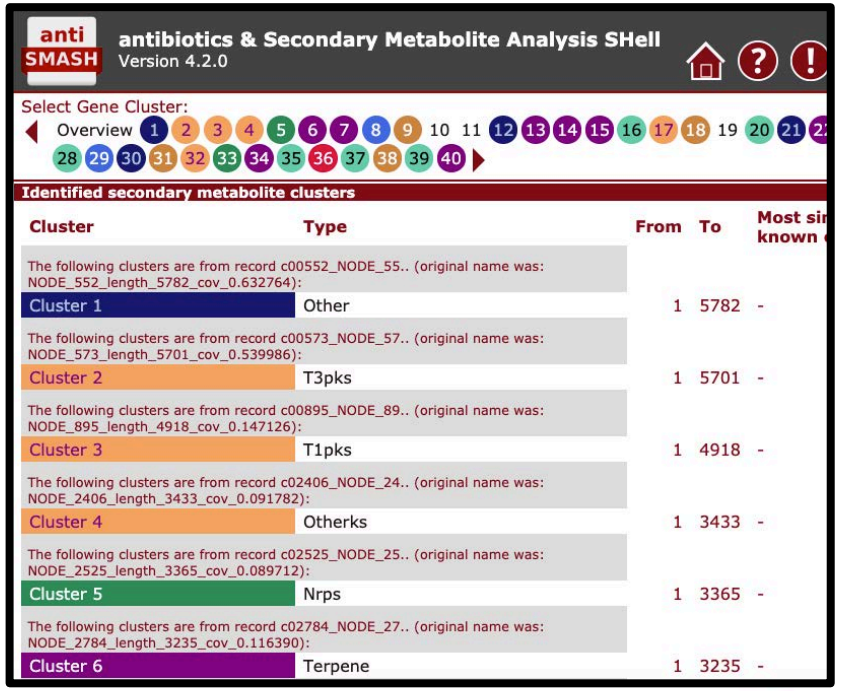
1. Prepare an annotation job to run under slurm
2. Use MEGAN to explore gene networks

Online resources and data analysis

Gene annotation

Identification of Biosynthetic Gene Clusters with **antiSMASH**

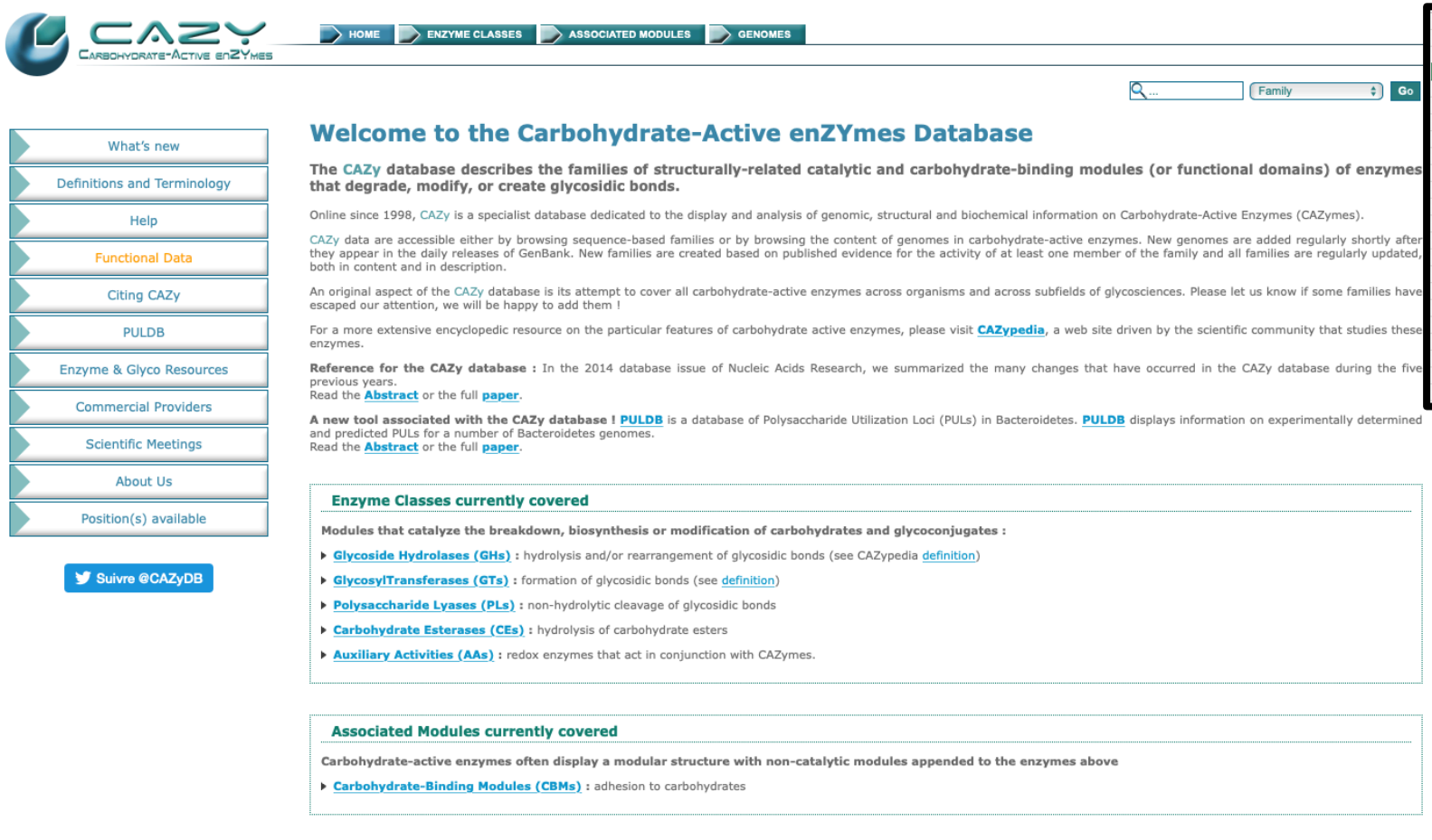




https://antismash.secondarymetabolites.org/

Gene annotation

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

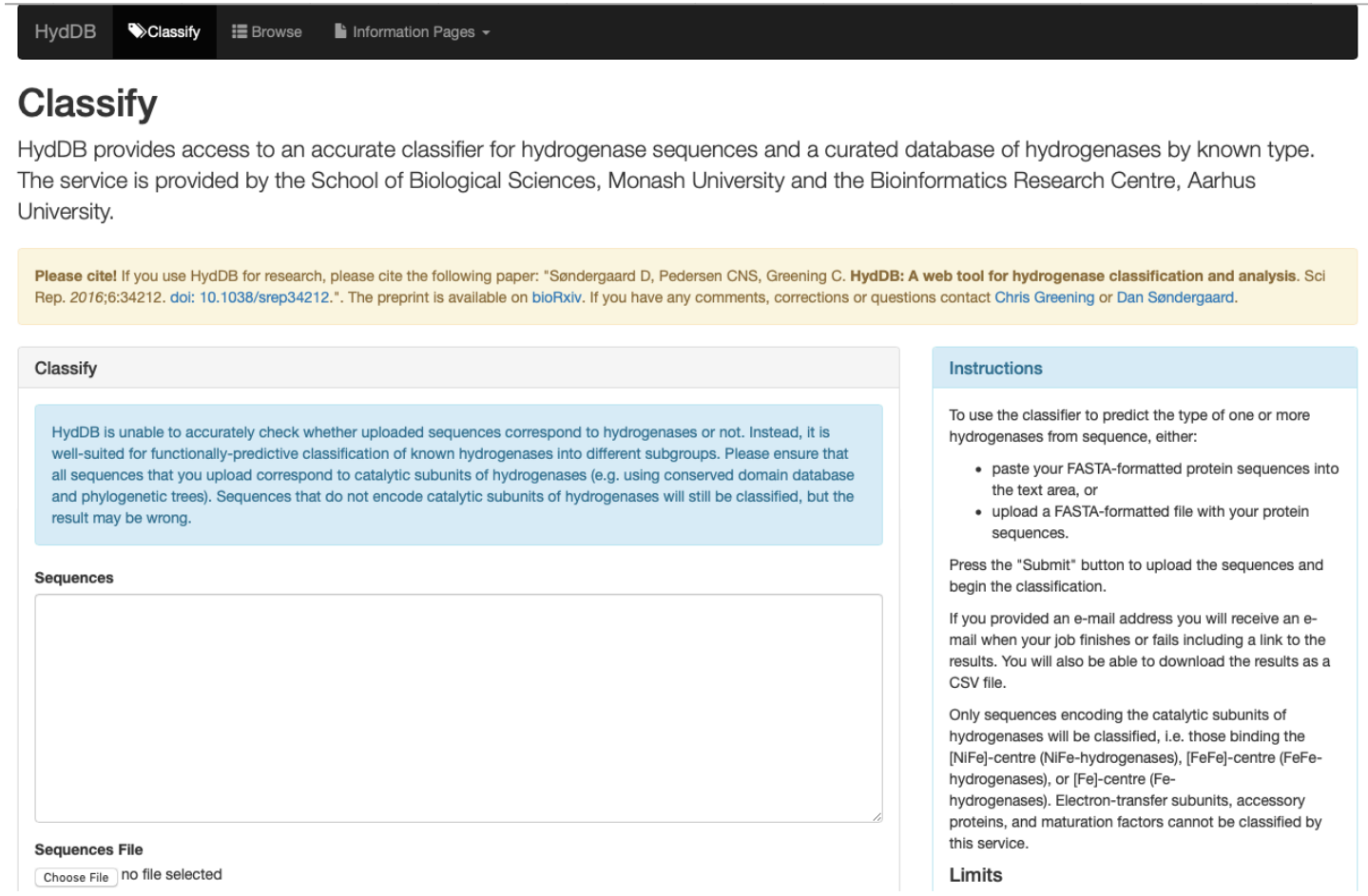


Query	Subject	E-value	Subject start	Subject end	Query start	Query end	Covered fraction
fig|6666666.197029.peg.1003	GH109.hmm	6.90E-12	1	117	2	112	0.920634921
fig|6666666.197029.peg.1015	GT4.hmm	3.00E-35	8	157	193	340	0.93125
fig|6666666.197029.peg.1034	PL17.hmm	0.00017	2	84	577	648	0.589928058
fig|6666666.197029.peg.1083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

http://www.cazy.org

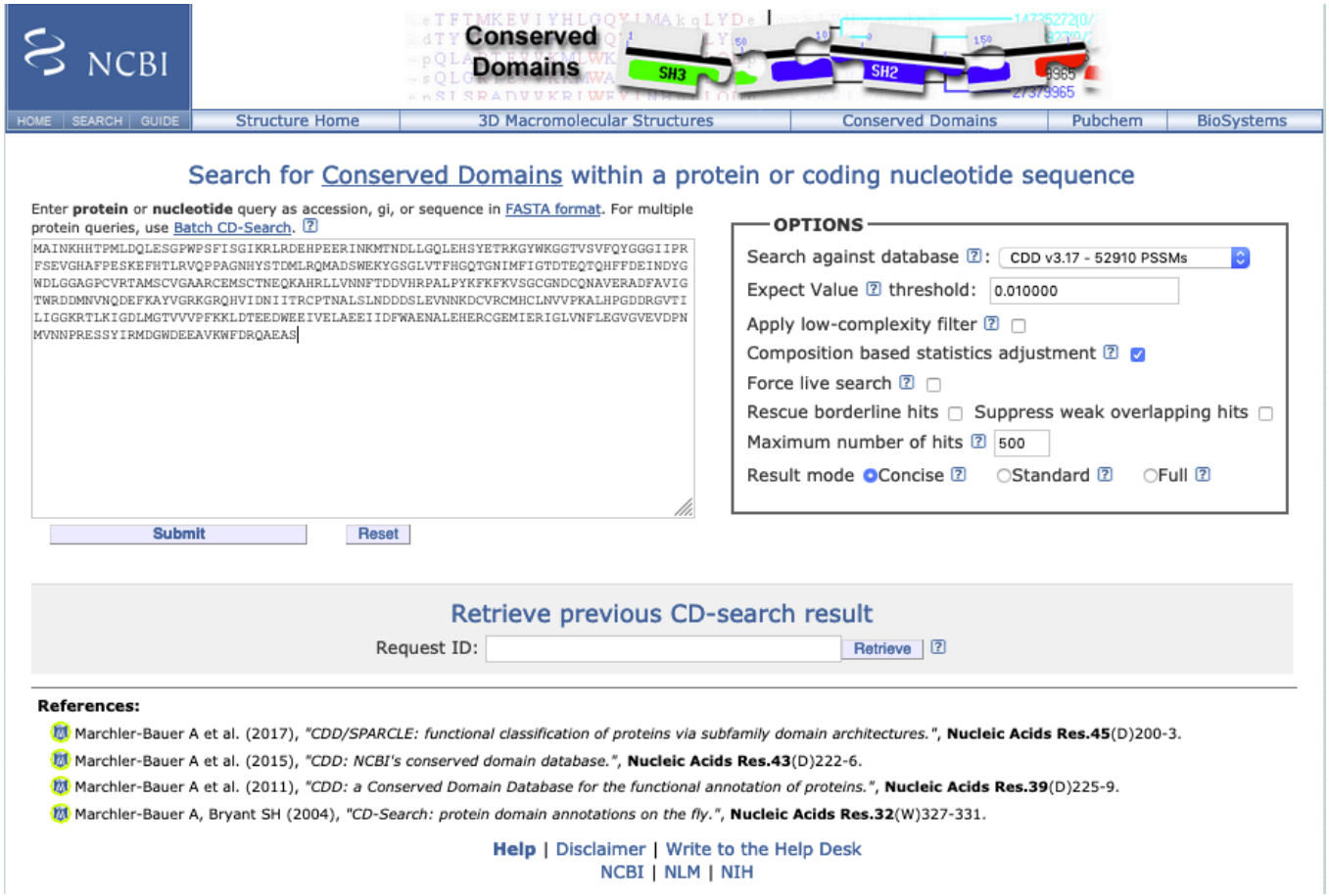
Gene annotation

Accurate classifier for hydrogenase sequences - **HydDB**



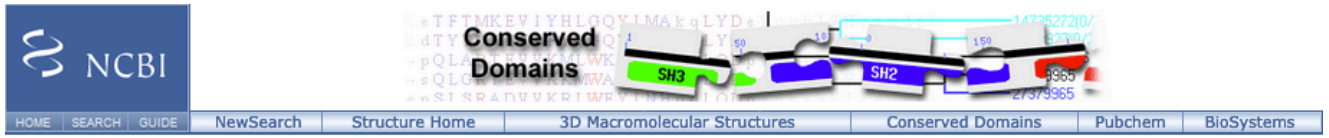
https://services.birc.au.dk/hyddb/

NCBI Conserved Domain Search



Search nucleotide/protein sequence(s) for conserved domains

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

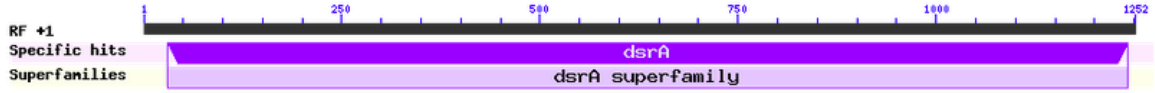


Conserved domains on [lcl|1]

View: Concise Results

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

Graphical summary Zoom to residue level show extra options »



Search for similar domain architectures | Refine search

List of domain hits

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

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

:

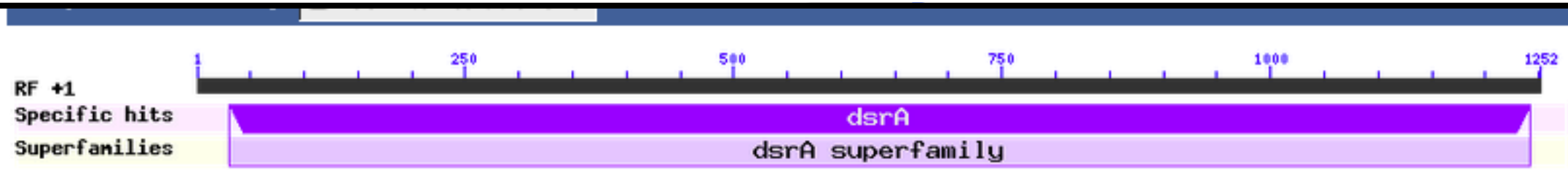
Pssm-ID: 273948 [Multi-domain] **Cd Length:** 402 **Bit Score:** 667.31 **E-value:** 0e+00

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

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

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

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

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

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

References:

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

Conserved Domain Search results:
dsrA gene of *Solemya velum* gill symbiont strain WH



Search for similar domain architectures | Refine search

List of domain hits

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

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

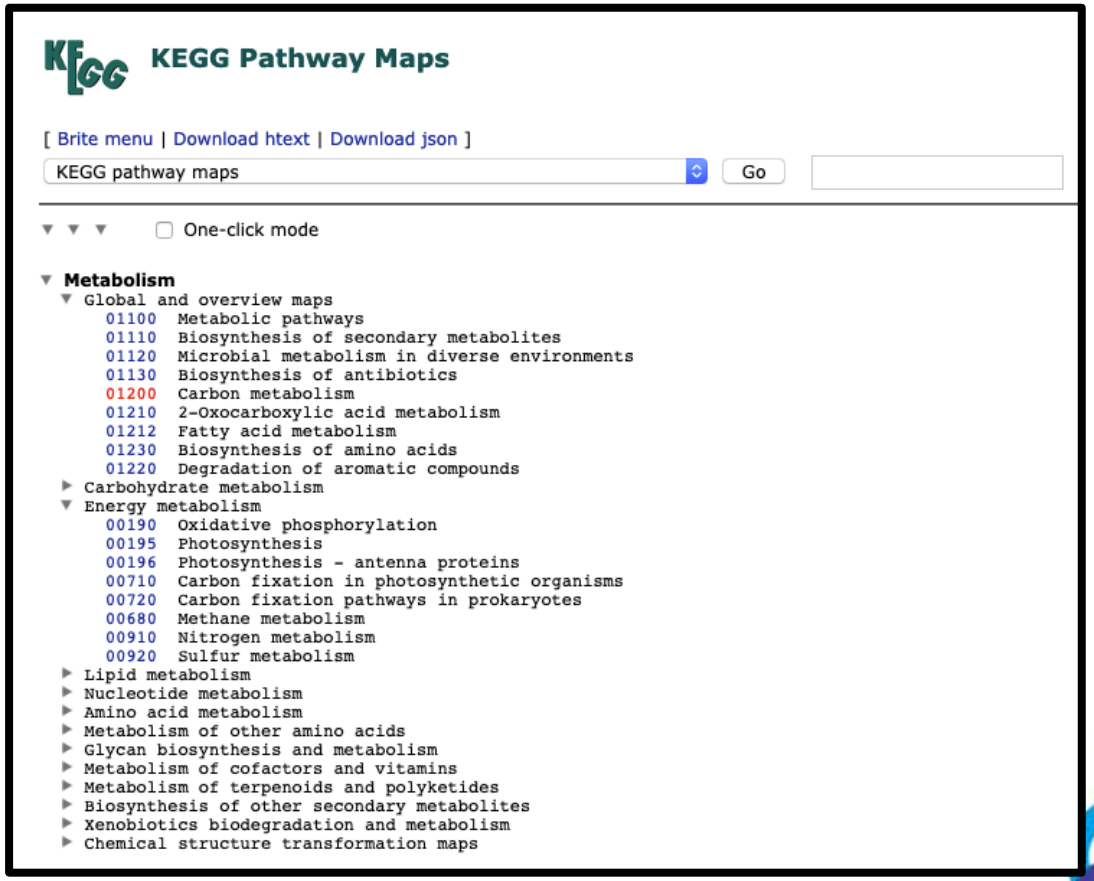
```
                    170       180       190       200       210       220       230       240
                    ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
1           171 VGAARCEMSCTNEQKAHRLLVNNFTDDVHRPALPYKFKFKVSGCGNDCQNAVERADFAVIGTWRDDMNVNQDEFKAYVGR 250
Cdd:TIGR02064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

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

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

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

References:

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

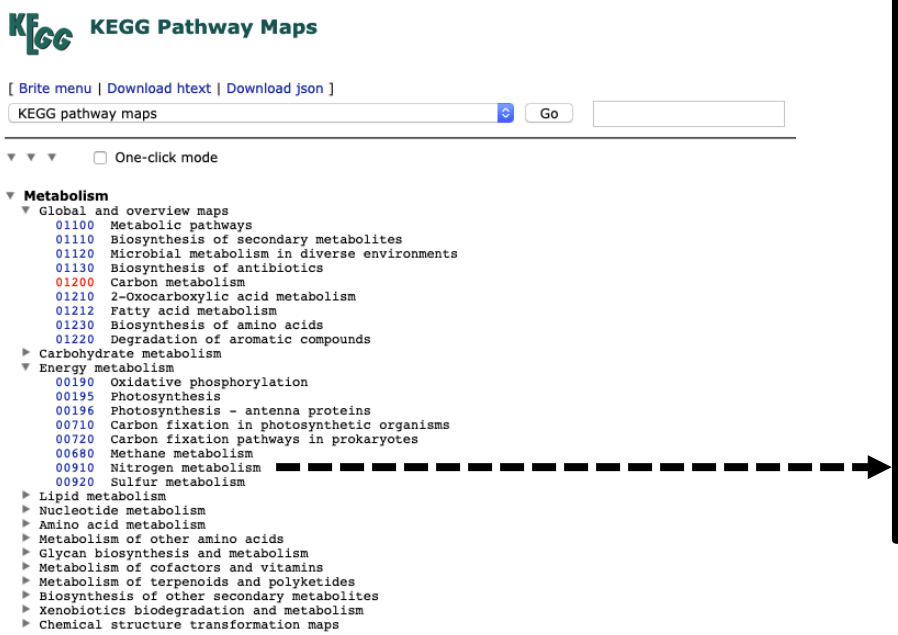
Gene annotation

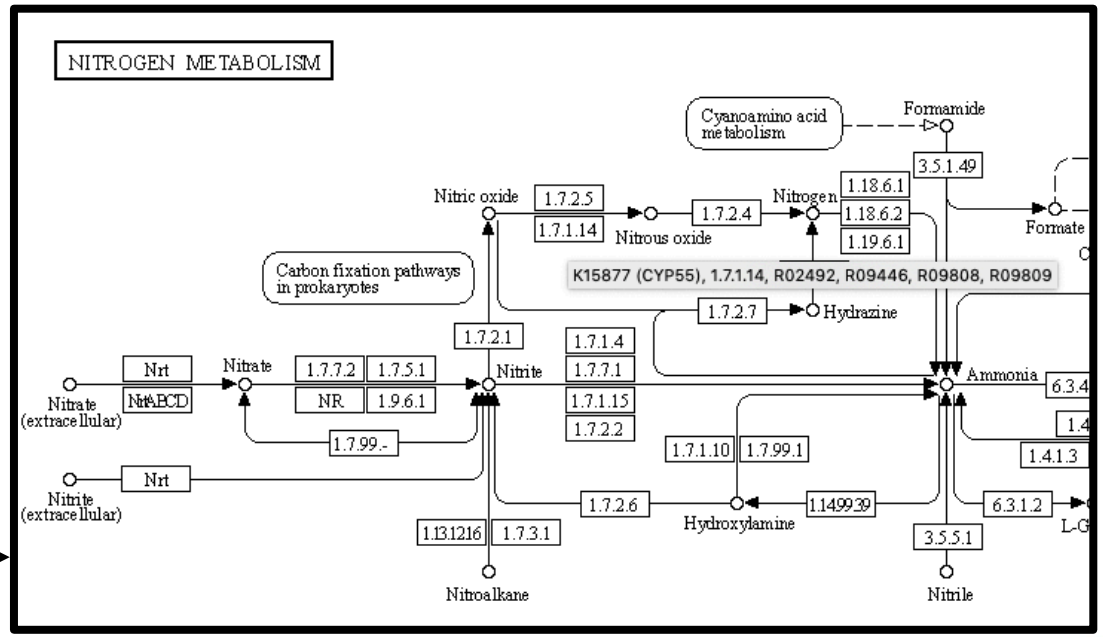
KEGG: https://www.genome.jp/kegg/pathway.html



Gene annotation

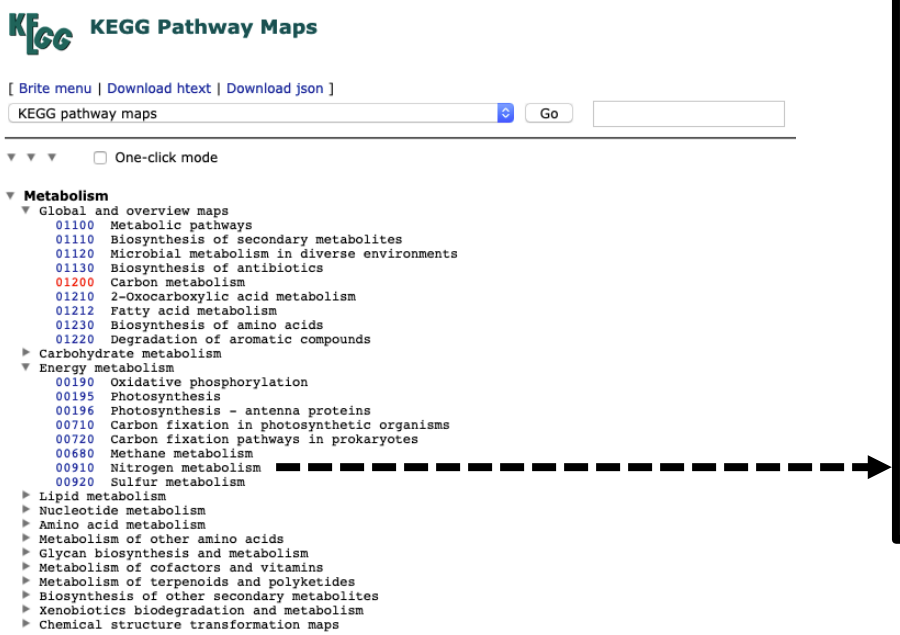
KEGG: https://www.genome.jp/kegg/pathway.html

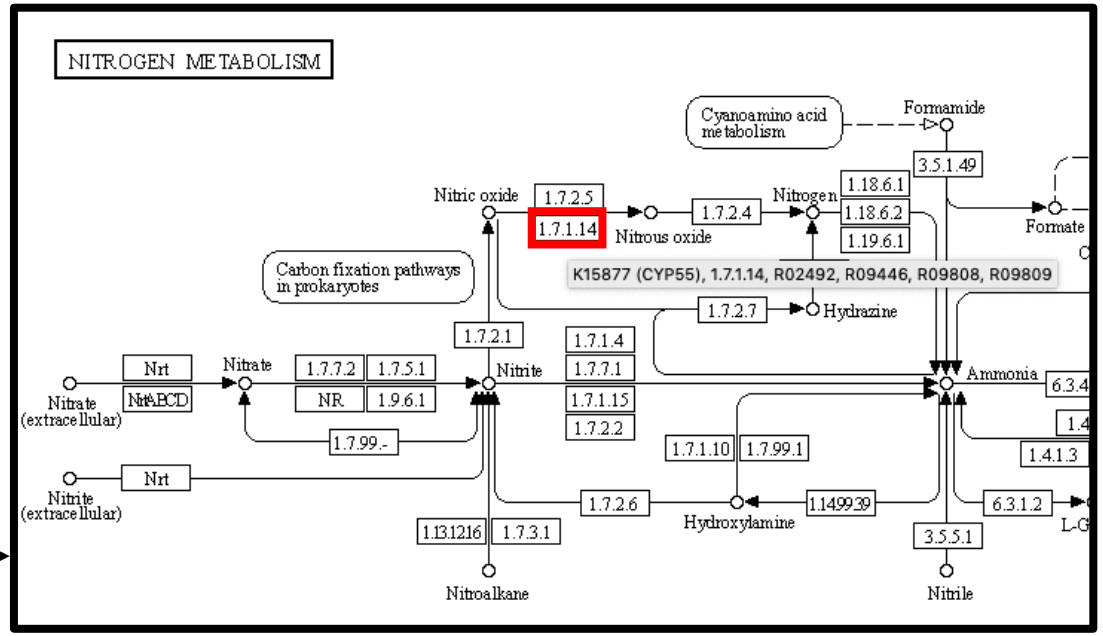




Gene annotation

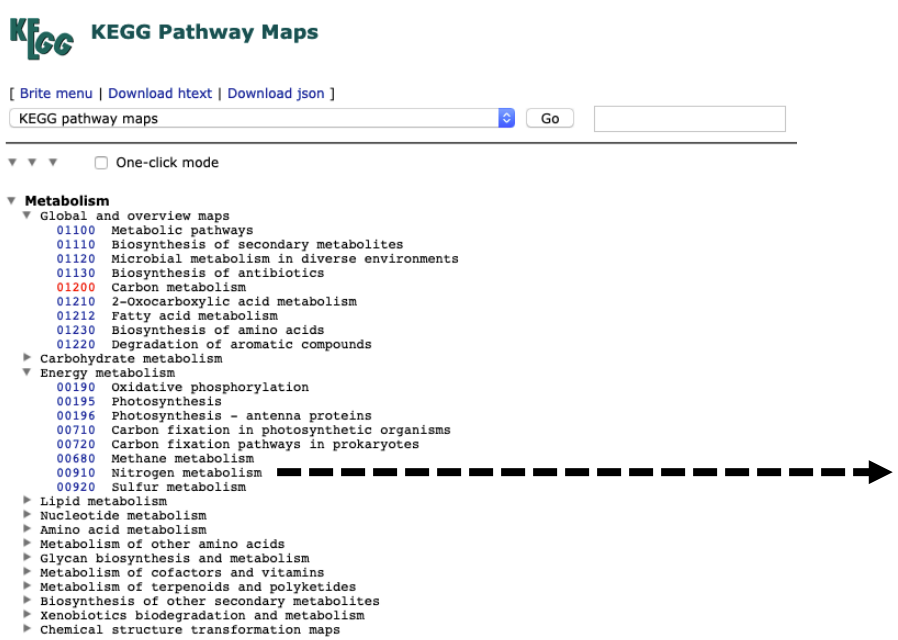
KEGG: https://www.genome.jp/kegg/pathway.html

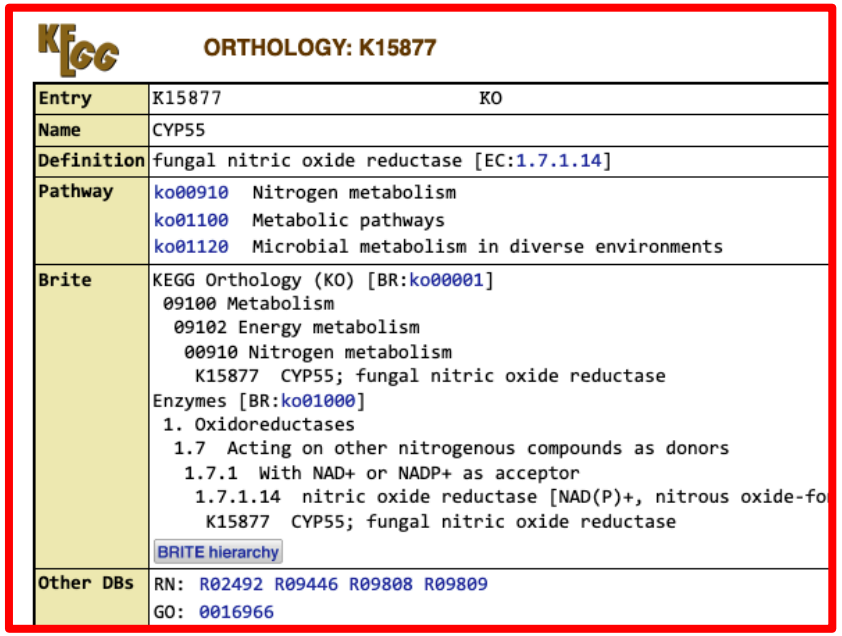




Gene annotation

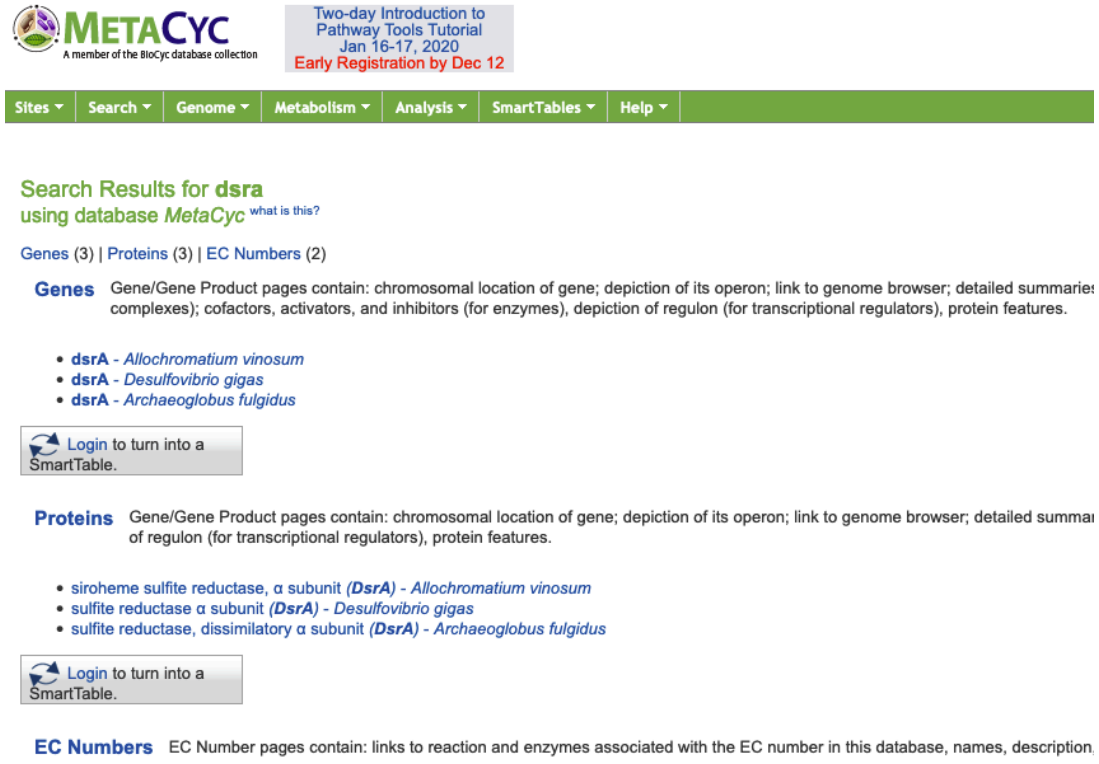
KEGG: https://www.genome.jp/kegg/pathway.html





Gene annotation

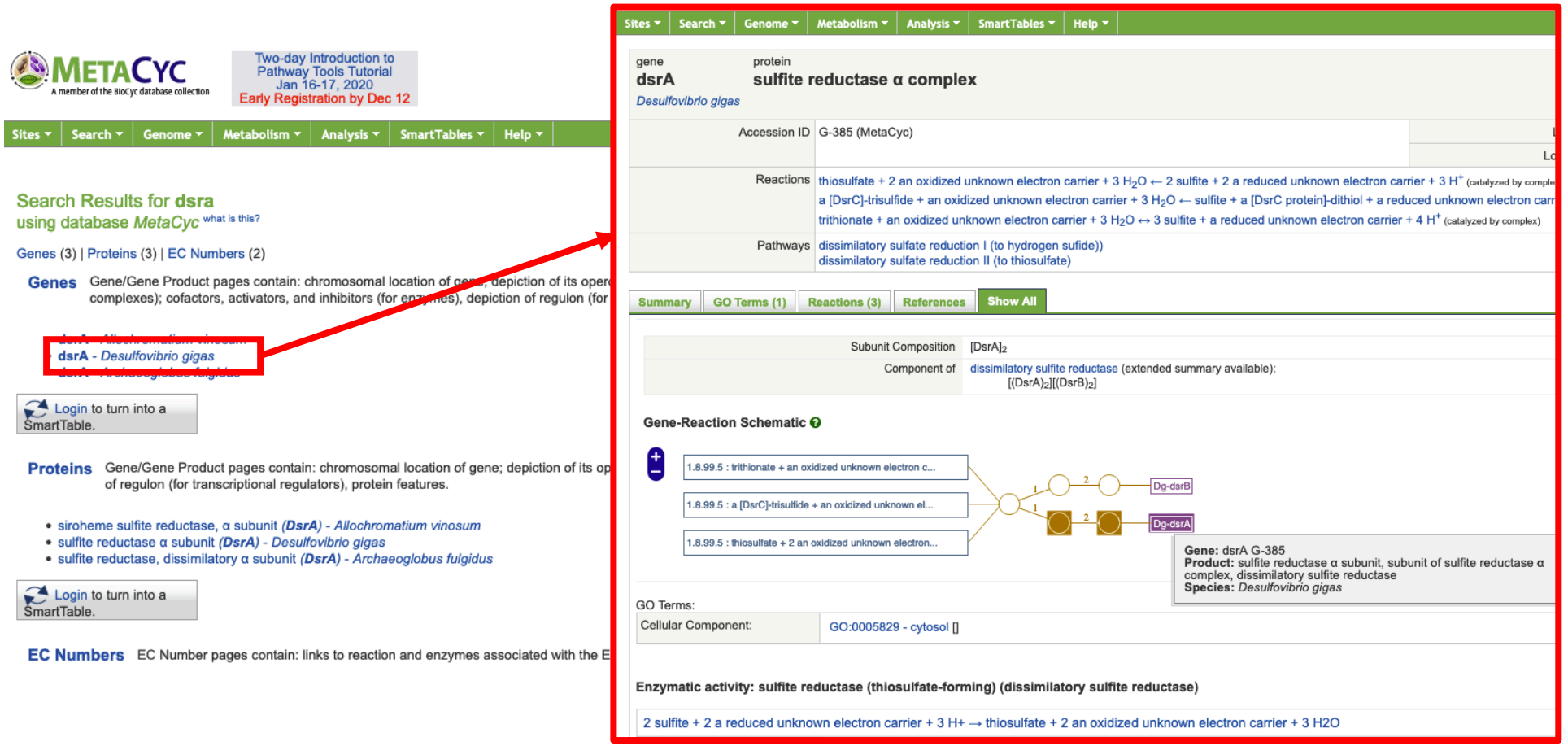
Metacyc: experimentally curated metabolic pathways



https://metacyc.org/

Gene annotation

Metacyc: experimentally curated metabolic pathways



https://metacyc.org/

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**?)

Choose an organism type (?): Bacteria Required
Choose Gram stain (?): Select Gram Required
Advanced Gram stain options (?): Select Option Required
Output format (?): Normal
Show results (?): Send by email
Email address:

Copy and paste your FASTA sequences below

or upload from file: Choose 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)

Submit Clear

http://psort.org/

Task: Analyze data for group work

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-Ljundahl)
8. **Non-polar flagella expression due to a chromosomal deletion**
9. **Plasmid-encoded antibiotic resistance**
10. **Aerobic (versus anaerobic) 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/

Optional: work with own data, or continue group task