MEGAN
taxonomic binning of sequence data

Overview

- Introduction
- Why use MEGAN
- using MEGAN
- Exercises

High throughput sequencing

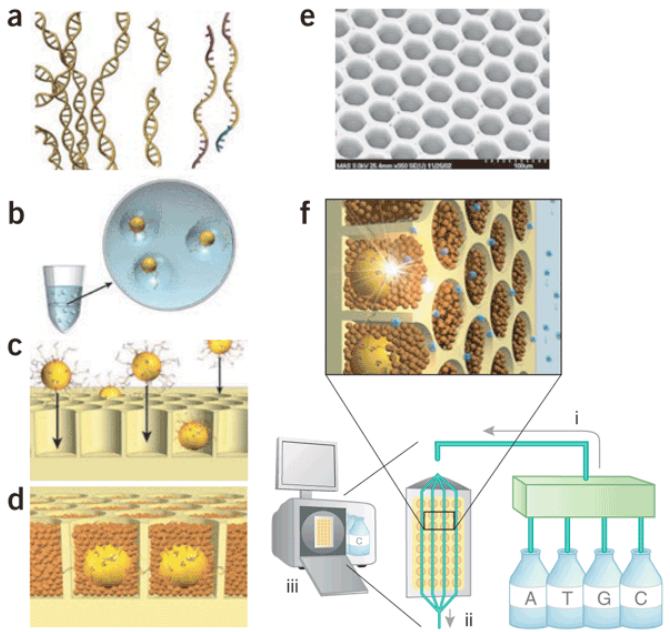
Shot-gun sequencing → huge amounts of data

454 GS-FLX generates 400-600 million bp per run

with a length of the reads between 400-500 bp

Understanding this amount of information in a quick manner?

→ classification of sequences



Sequence Classification

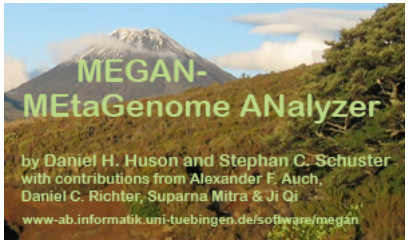
Sequence classification (binning) is the process of separating sequence data using specific information → creating bins

This information can be based on :

- Similarity e.g. MEGAN, SORT-ITEMS
- Phylogeny e.g. tools like CARMA
- Functional annotation e.g. GO classifiers

MEGAN uses similarity searches with BLAST to bin sequences into taxa.

MEGAN



Metagenome: the collective genome of all the microorganisms in am environment. (*Handelsman et al., 1998*)

Metagenomics is the study of the metagenome using high throughput sequencing.

The question:

How to determine species composition in a metagenomic dataset?

- Sequence comparison with known sequences from a database e.g. Genbank.
- Metagenomic datasets contain many sequences, so manual inspection is impossible → MEGAN is your assistant.

MEGAN

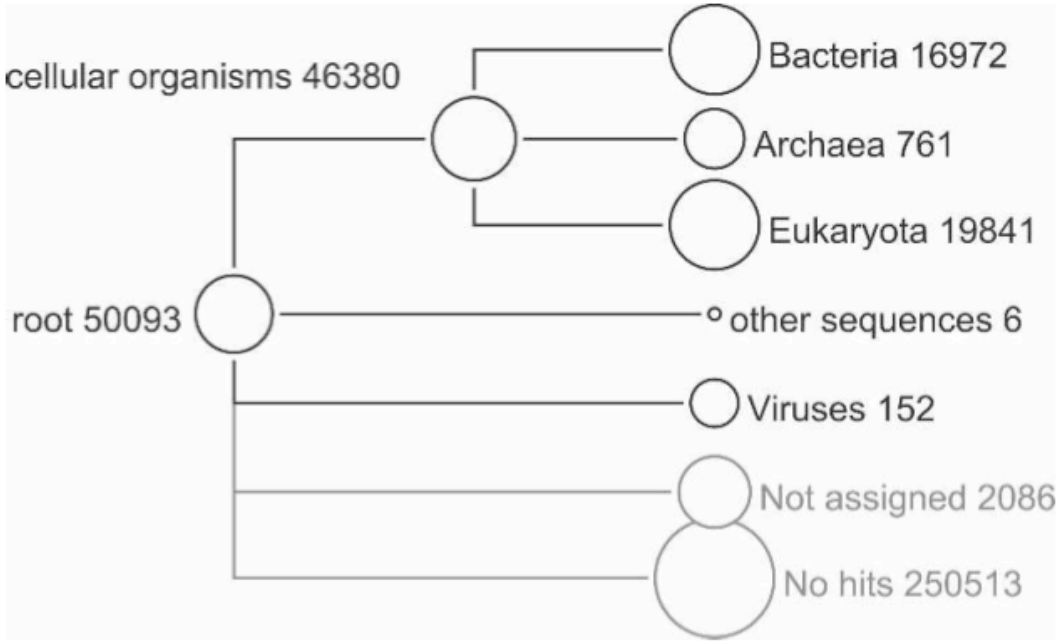


Figure 5. High-level summary of a MEGAN analysis of the mammoth data set, based on a BLASTX comparison of the 302,692 reads against the NCBI-NR database.

(Huson et al., Genome Research, 2007)

Why use Megan?

Easy to work with on a desktop / laptop computer:

Extra things needed: Java, a BLAST server (e.g. Bioportal)

MEGAN gives a visualization of BLAST results

- Study diversity
- Compare samples
- Contamination filtering
- Special gene of interest
- Extraction of sequences based on taxonomic information.

Bioportal: http://www.bioportal.uio.no//

The basics of MEGAN

MEGAN uses BLAST, a database and a taxonomy file

- BLAST N : nucleotides against a nucleotide database.
- BLAST X : Translated nucleotide
  against a protein database.
- Which Database?

  one of the many available database like the NCBI-non-redundant database, or a your own custom database.
- Taxonomy: NCBI taxonomy, or your own custom taxonomy

BLAST output file is used to bin sequences using the LCA assignment algorithm into specific taxons.

The basics of MEGAN

- The LCA algorithm = "Lowest Common Ancestor" algorithm

"In this approach, every read is assigned to some taxon. If the read aligns very specifically only to a single taxon, then it is assigned to that taxon. The less specifically a read hits taxa, the higher up in the taxonomy it is placed. Reads that hit ubiquitously may even be assigned to the root node of the NCBI taxonomy."

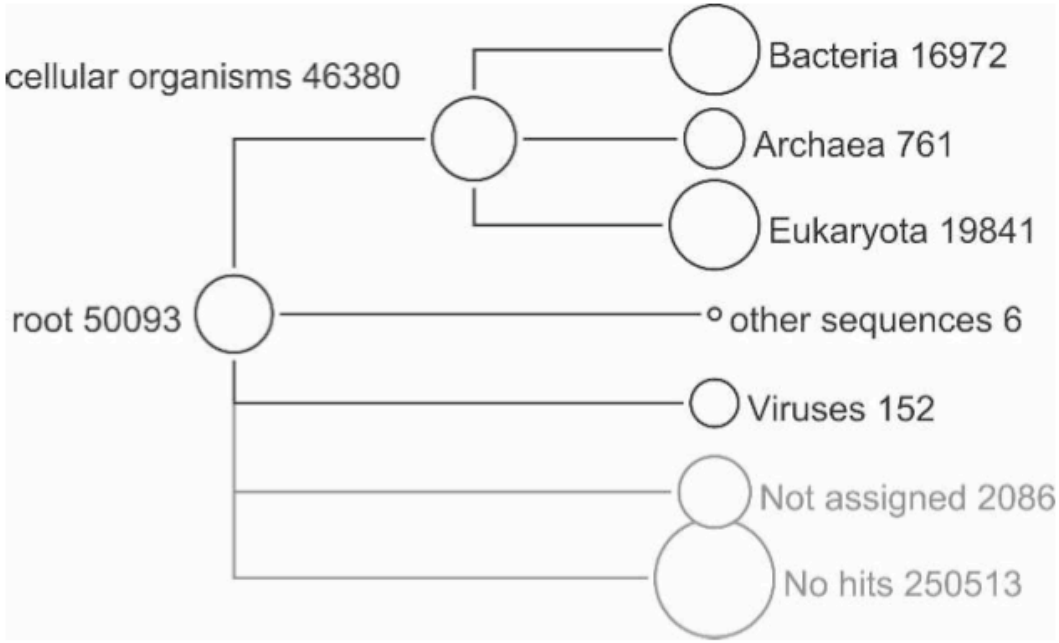
(the MEGAN manual)

The basics of MEGAN

The default LCA parameters are:

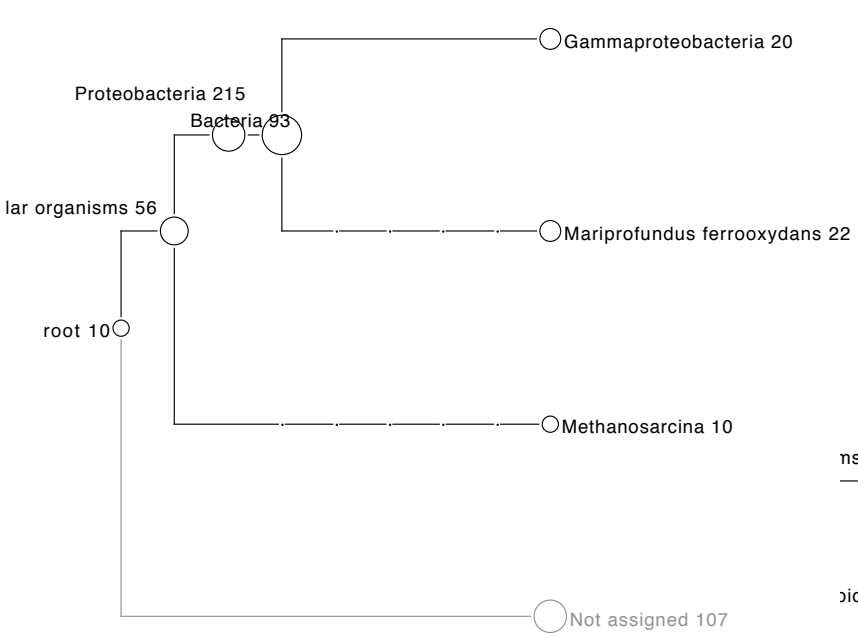
- Min Support → default = 5 reads per taxon
- Min Score → default bitscore = 35
- Min score / length → Bitscore divided by the read length d = 0
- The top percentage → The maximum percentage by which the score of a hit may fall below the best score achieved for a given read d = 10
- Win score → If a win score is set, then, for a given read, if any match exceeds the win score, only matches exceeding the win score ("winners") are used to place the given read. d = 0

The basics of MEGAN

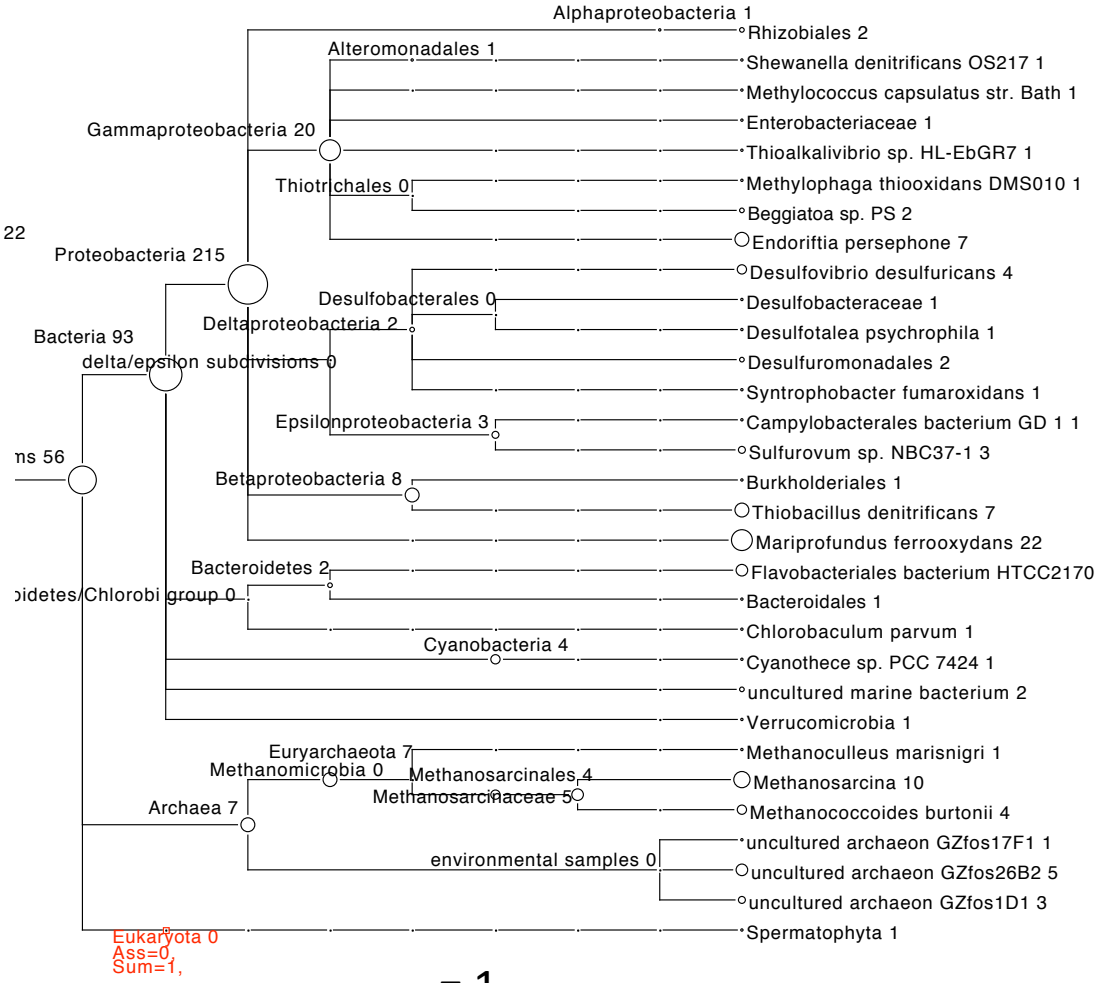


No hits: none of the blast hits reached the minimum bitscore.
Not assigned : The sequence had not enough hits to be classified to a taxon (Min support & top percentage)

Playing with min support



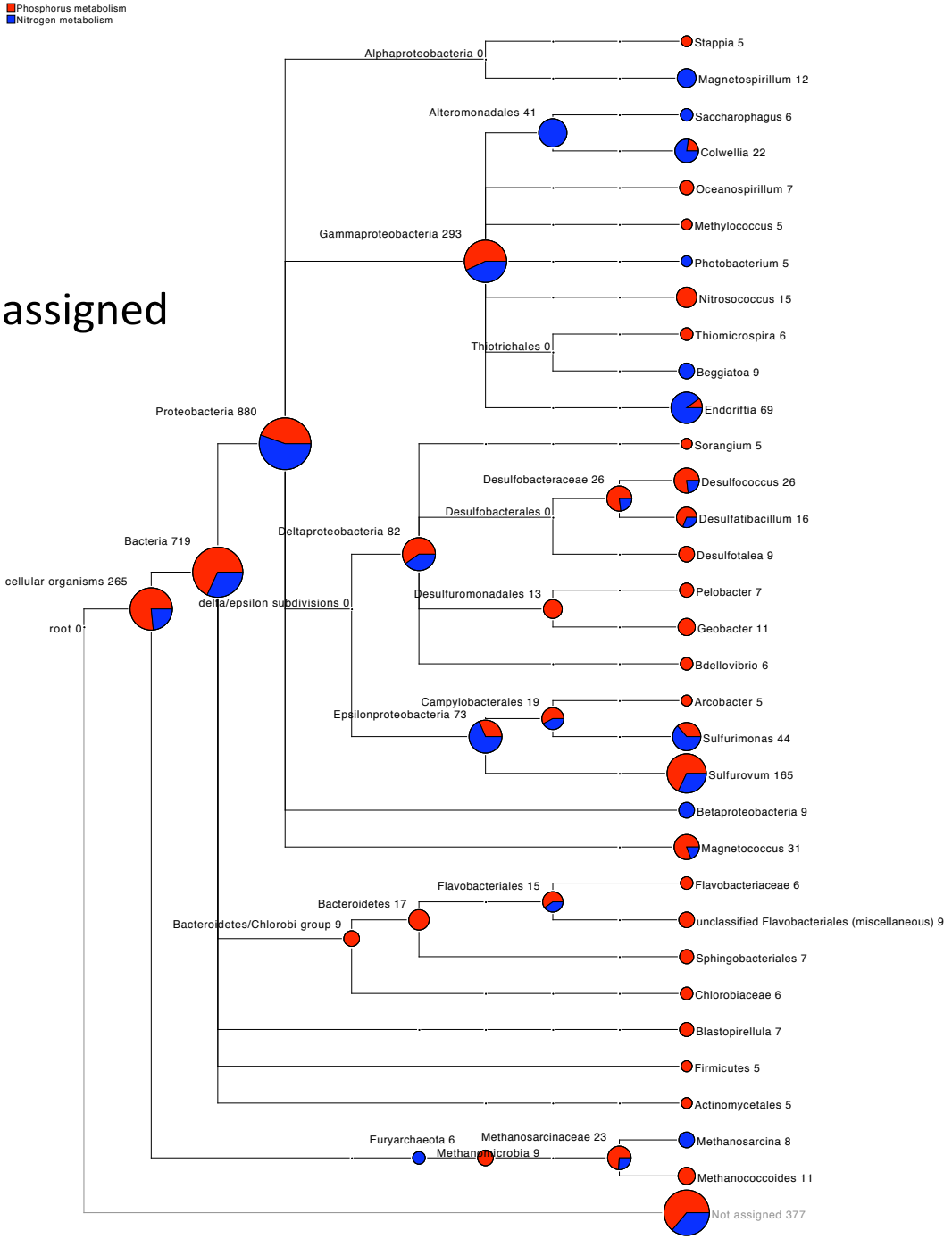
= 10



= 1

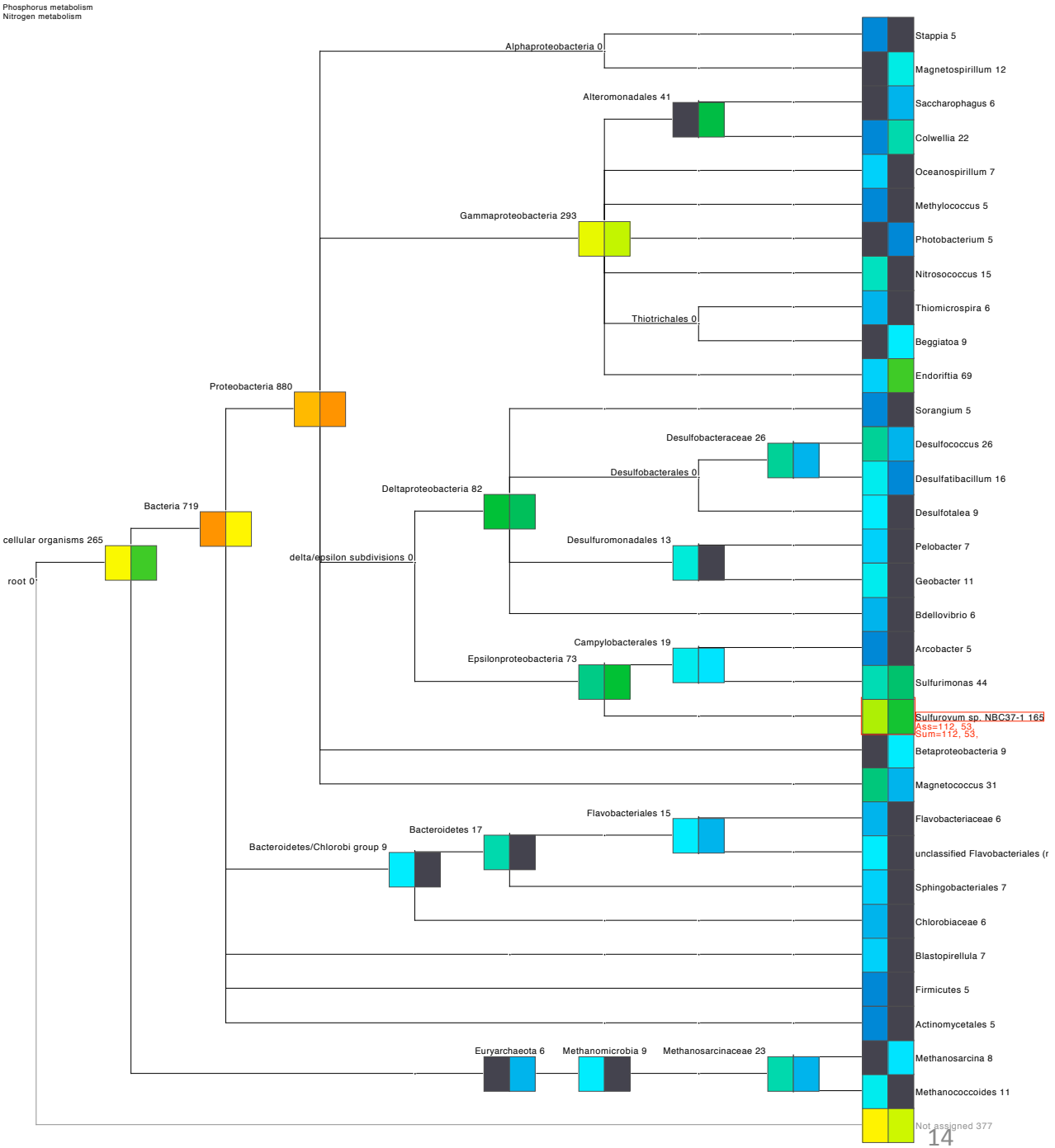
An example

Comparison between reads assigned to Phosphorus metabolism and Nitrogen metabolism

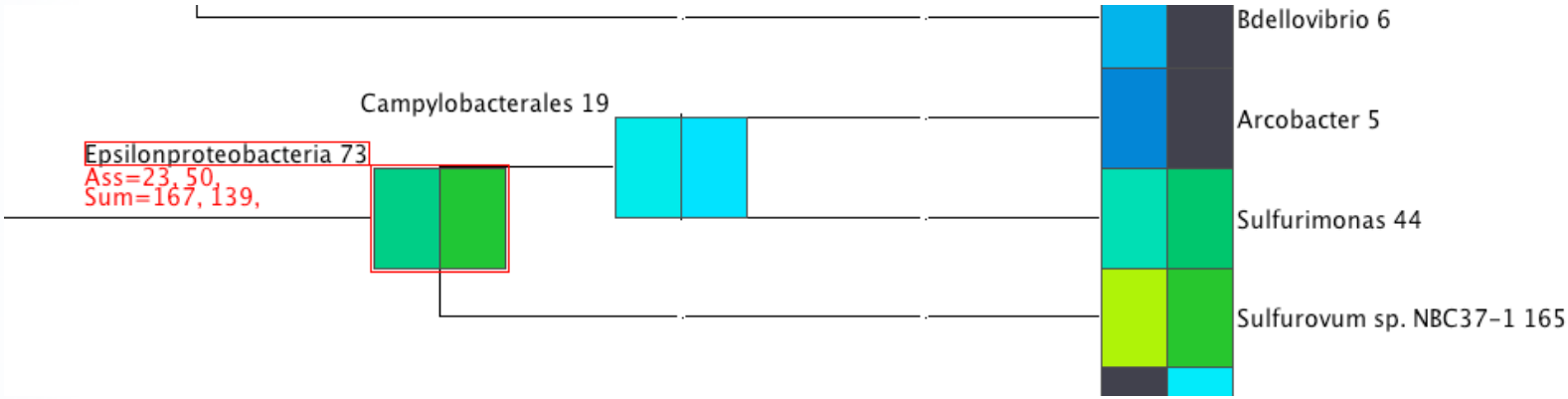


Reads were annotated using MG-RAST

An example


1: Phosphorus metabolism
2: Nitrogen metabolism
Alphaproteobacteria 0
Alteromonadales 41
Gammaproteobacteria 293
Thiotrichales 0
Proteobacteria 880
Bacteria 719
cellular organisms 265
root 0
Deltaproteobacteria 82
delta/epsilon subdivisions 0
Desulfobacteraceae 26
Desulfobacterales 0
Desulfuromonadales 13
Campylobacterales 19
Epsilonproteobacteria 73
Bacteroidetes 17
Bacteroidetes/Chlorobi group 9
Flavobacteriales 15
Euryarchaeota 6
Methanomicrobia 9
Methanosarcinaceae 23
Stappia 5
Magnetospirillum 12
Saccharophagus 6
Colwellia 22
Oceanospirillum 7
Methylococcus 5
Photobacterium 5
Nitrosococcus 15
Thiomicrospira 6
Beggiatoa 9
Endoriftia 69
Sorangium 5
Desulfococcus 26
Desulfatibacillum 16
Desulfotalea 9
Pelobacter 7
Geobacter 11
Bdellovibrio 6
Arcobacter 5
Sulfurimonas 44
Sulfurovum sp. NBC37-1 165
Ass=112, 53,
Sum=112, 53,
Betaproteobacteria 9
Magnetococcus 31
Flavobacteriaceae 6
unclassified Flavobacteriales (r
Sphingobacteriales 7
Chlorobiaceae 6
Blastopirellula 7
Firmicutes 5
Actinomycetales 5
Methanosarcina 8
Methanococcoides 11
Not assigned 377

An example



Reads at node Epsilonbacteria: total in comparison 73
Assigned = 23, 50
Summarized = 167, 139

What do these numbers mean?

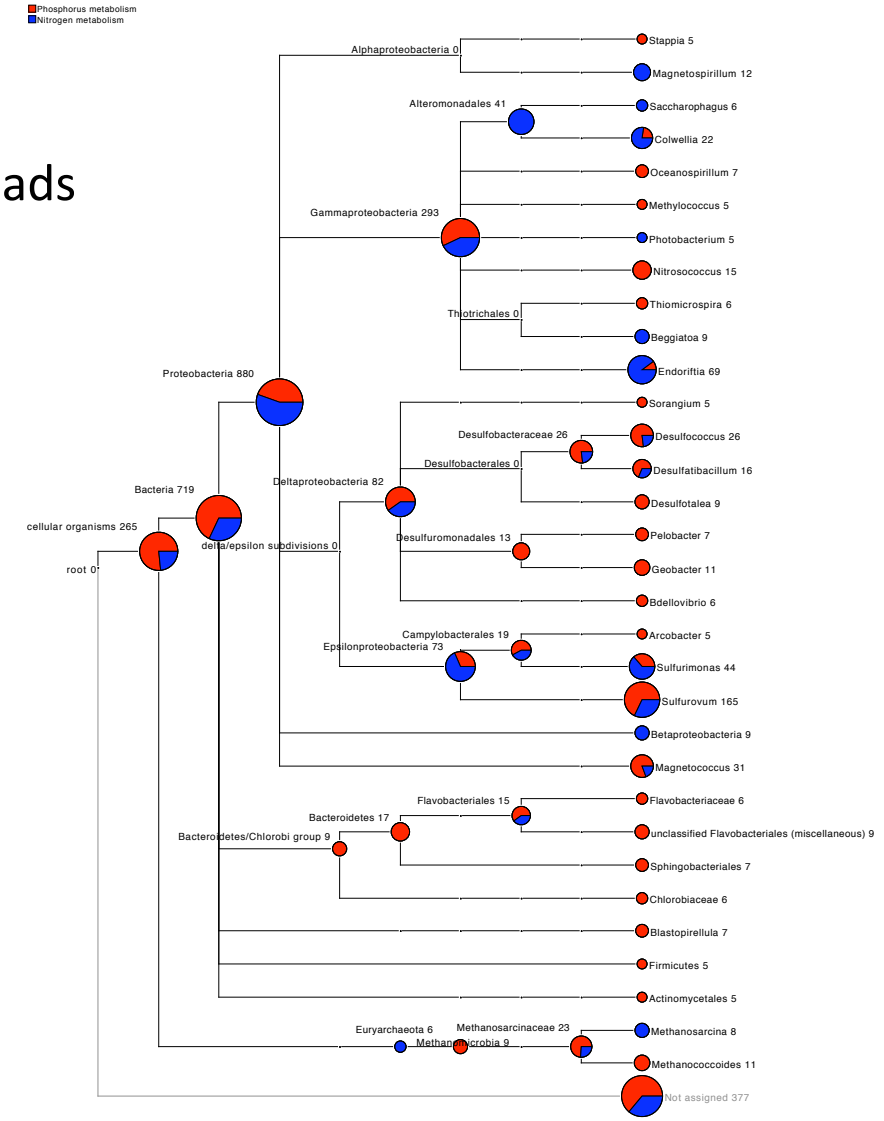
Other tools in MEGAN

- Comparing datasets
- GO annotation
- Rarefaction analysis
- Extracting sequences
- Identification of Clusters of Orthologous Groups of proteins (COGs).
- Microbial attributes

Comparing datasets

- Absolute numbers
- Normalized → 100.000 reads

More than 2 sets can be Compared.



Comparing datasets

- MEGAN offers a statistical test to compare if the number of reads found in a taxon between two datasets is different.

  → but it is only visual.

- If you want to test it yourself

  → extract assignments after comparison

  → table with taxa and the number of reads per taxon for each dataset.

GO annotation

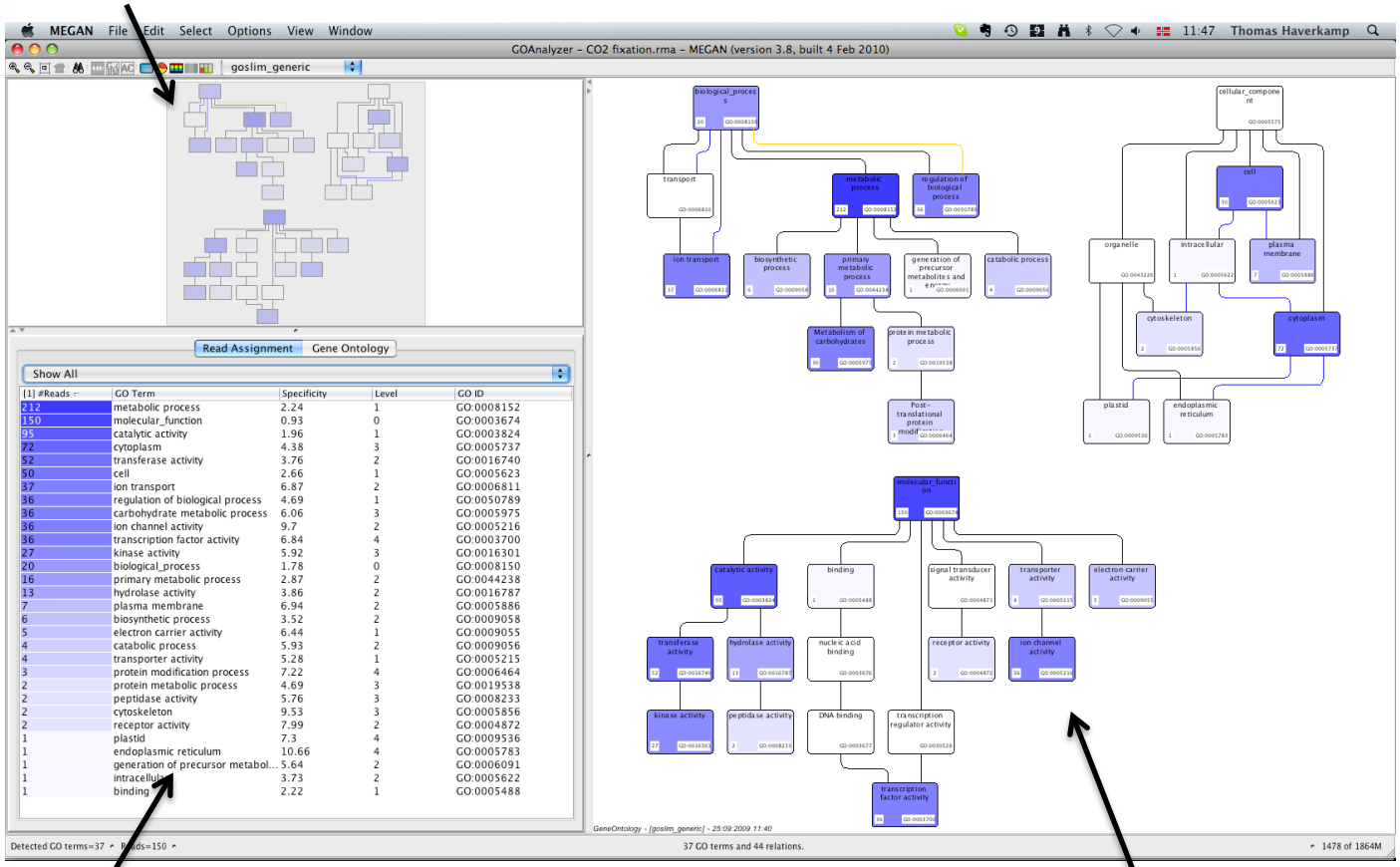
The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO collaborators are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

http://www.geneontology.org/

How is this done in MEGAN?

GO annotation

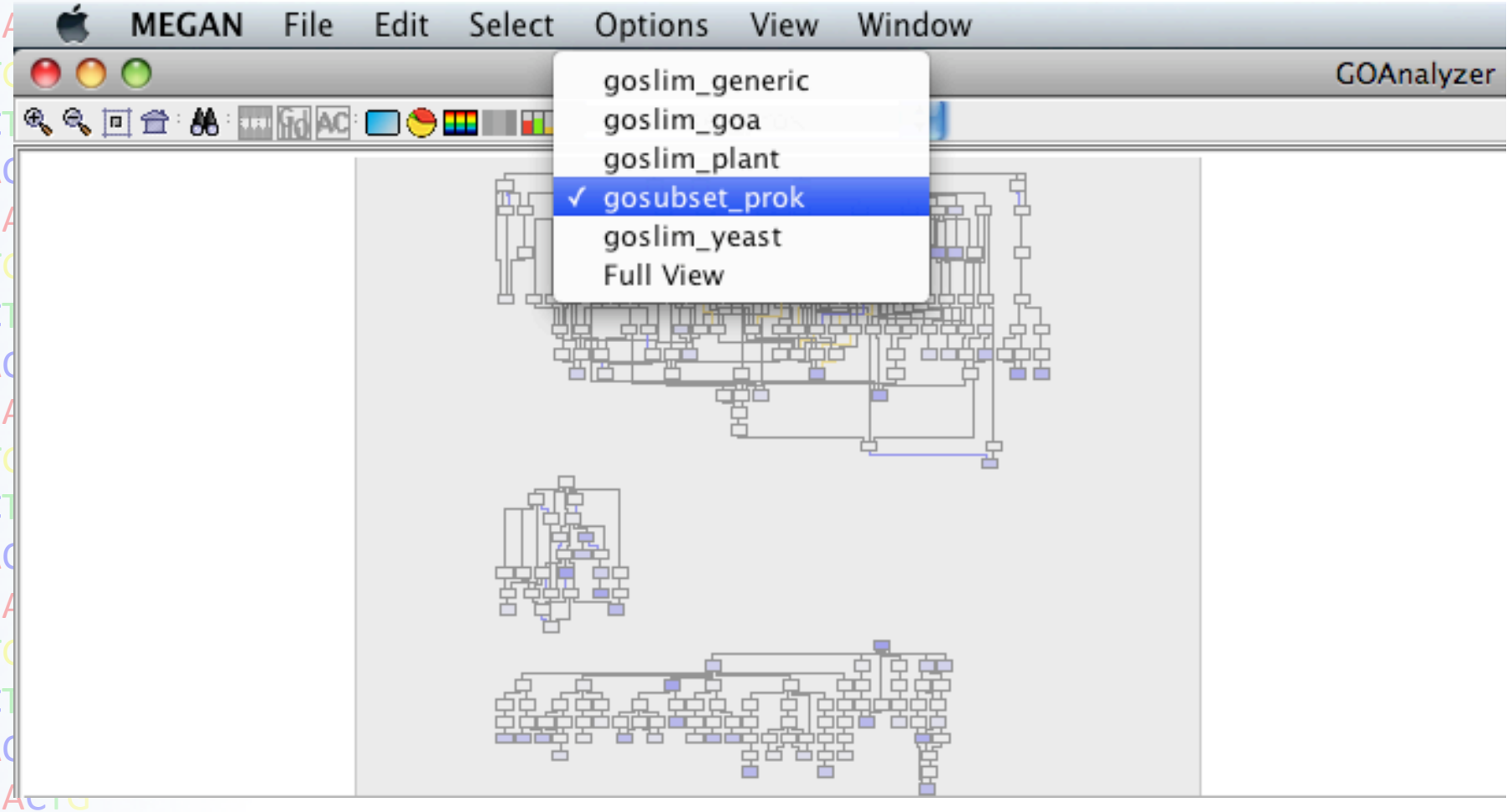
Overview



GO terms

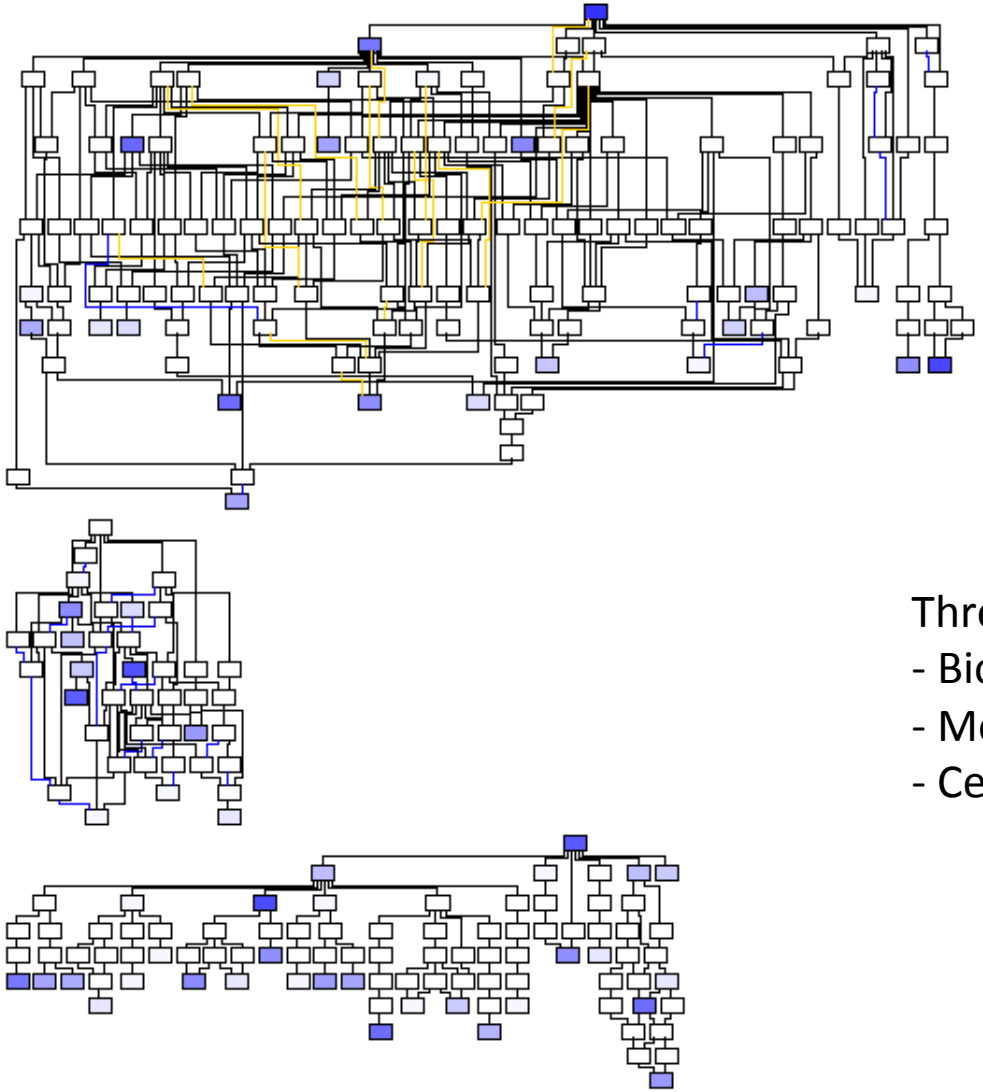
zoomable
GO map

GO annotation



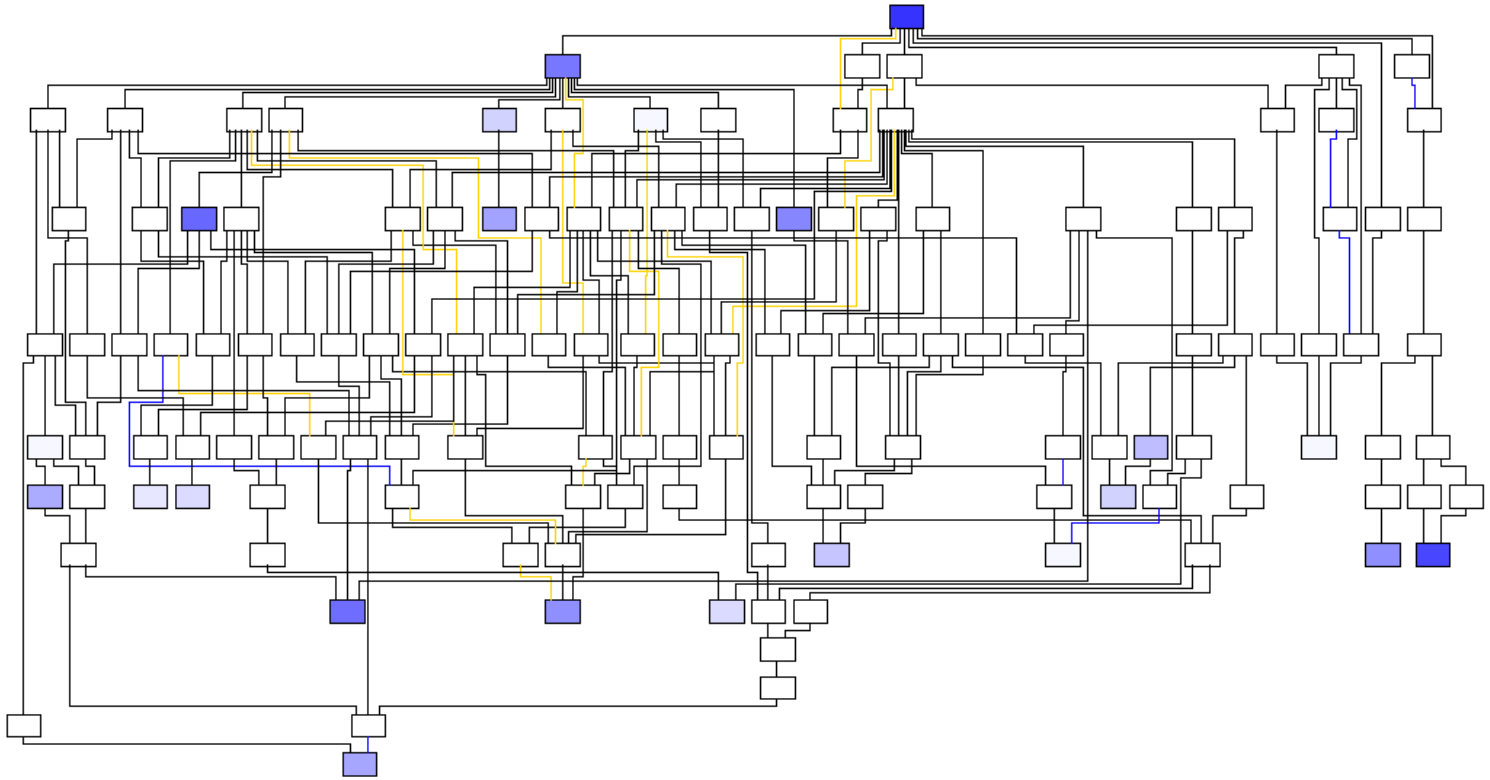
Choose the right GO map for analyzing your data

GO annotation

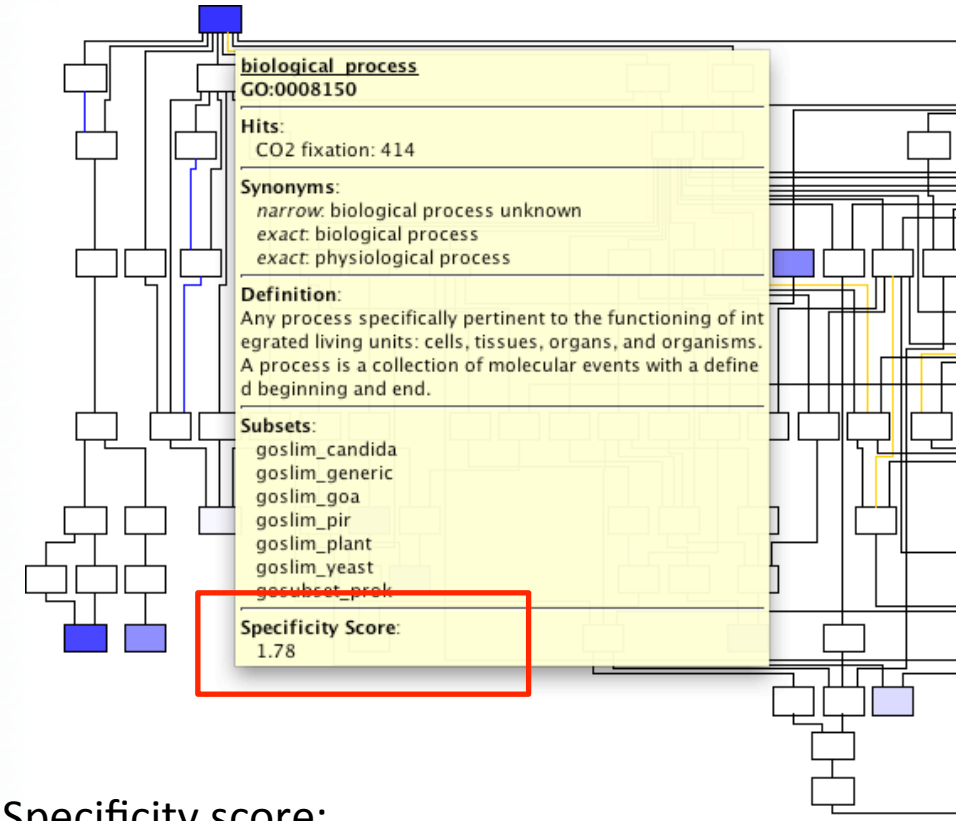


Three maps:
- Biological processes
- Molecular functions
- Cellular components

Biological processes

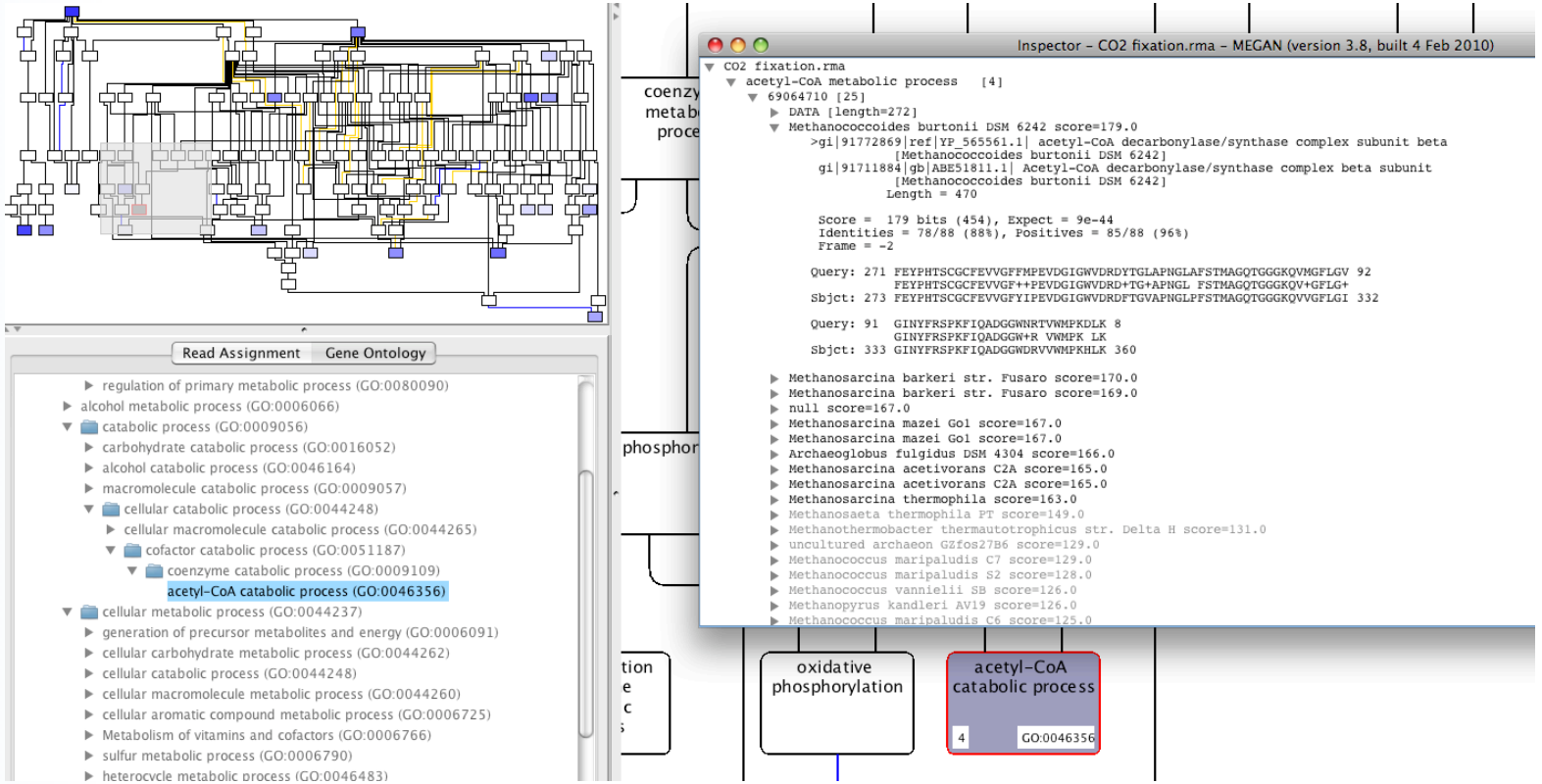


GO annotation



Specificity score:
The score of a term reflects the frequency of gene annotations to that term (or to descendants in the sub graph of that term).
Terms often used for annotated gene products are assigned with a lower specificity than infrequently used terms.

Acetyl-CoA



The reads belonging to specific node can be extracted.

GO annotation

Read Assignment | Gene Ontology

Show All

[1] #Reads	GO Term	Specificity	Level	GO ID
36	regulation of transcription, DNA-dependent	5.81	8	GO:0006355
36	transcription factor activity	6.84	4	GO:0003700
35	transketolase activity	13.68	4	GO:0004802
35	chloride transport	13.72	7	GO:0006821
27	voltage-gated chloride channel activity	13.94	9	GO:0005247
24	glycolate oxidase complex	17.87	7	GO:0009339
21	fructose-bisphosphate aldolase activity	12.67	5	GO:0004332
19	carbon utilization by fixation of carbon dioxide	8.34	3	GO:0015977
16	ribulose-bisphosphate carboxylase activity	8.62	5	GO:0016984
16	pentose-phosphate shunt, non-oxidative branch	14.68	12	GO:0009052
16	ribose-5-phosphate isomerase activity	13.53	5	GO:0004751
14	triose-phosphate isomerase activity	12.86	5	GO:0004807
14	Glucose metabolism	8.42	6	GO:0006006
11	ATPase activity	6.71	7	GO:0016887
8	transporter activity	5.28	1	GO:0005215
8	acetyl-CoA metabolic process	10.55	5	GO:0006084
8	catalytic activity	1.96	1	GO:0003824
7	plasma membrane	6.94	4	GO:0005886
6	thiamin biosynthetic process	11.29	7	GO:0009228
5	serine-type endopeptidase activity	9.14	6	GO:0004252
5	electron carrier activity	6.44	1	GO:0009055
5	intrinsic to membrane	5.22	5	GO:0031224
4	carbon utilization	8.29	2	GO:0015976
4	acetyl-CoA catabolic process	10.64	6	GO:0046356
3	protein amino acid phosphorylation	7.68	8	GO:0006468
3	glycerol metabolic process	11.39	6	GO:0006071
3	intracellular	3.73	3	GO:0005622
2	phosphoribulokinase activity	15.38	5	GO:0008974
2	Removal of aminoterminal propeptides from gamma-...	6.53	6	GO:0006508
2	ion transmembrane transporter activity	6.22	4	GO:0015075
2	kinesin complex	16.23	10	GO:0005871
2	transmembrane receptor activity	9.43	4	GO:0004888

Rarefaction analysis

Rarefaction analysis:
Sequence similarity is used to generate rarefaction curves.

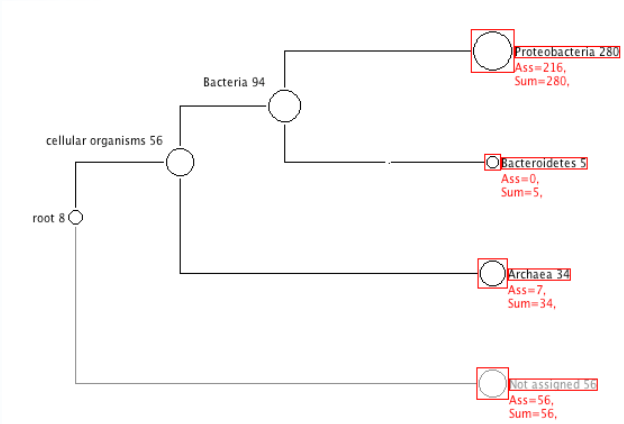
With shotgun data this is not possible, sequences are not similar.

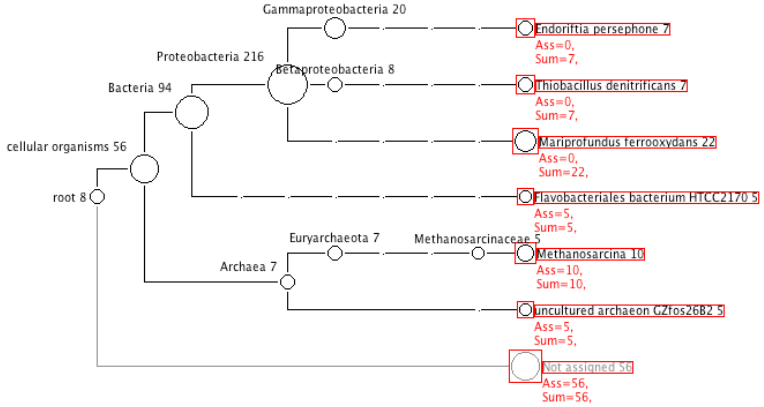
MEGAN samples sequences and counts how many “leaves” are found after binning.

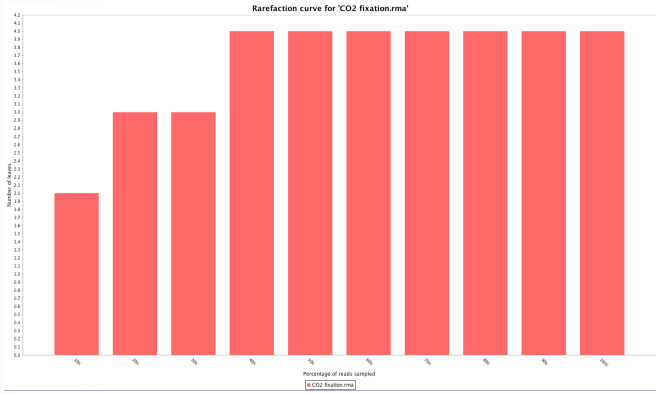
Calculation in MEGAN depends on how the taxonomy is represented!!!

If you collapse one node your rarefaction results may change

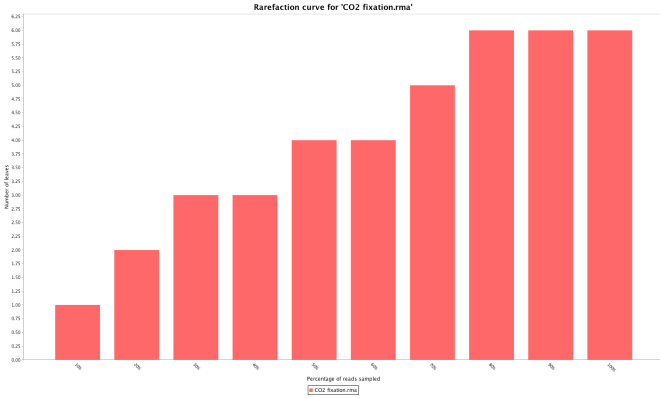
Rarefaction analysis







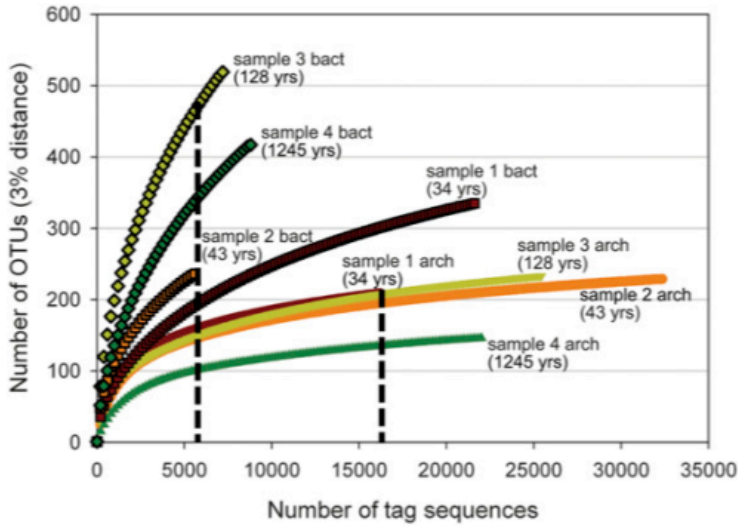
phylum



species

Rarefaction analysis

- Make sure you know what level of taxonomy you use.
- Uncollapse your taxonomy to species level.
- Comparison of rarefaction curves of different datasets is possible.
- For amplicon sequences you can better use: Unifrac, Mothur, or EstimateS.

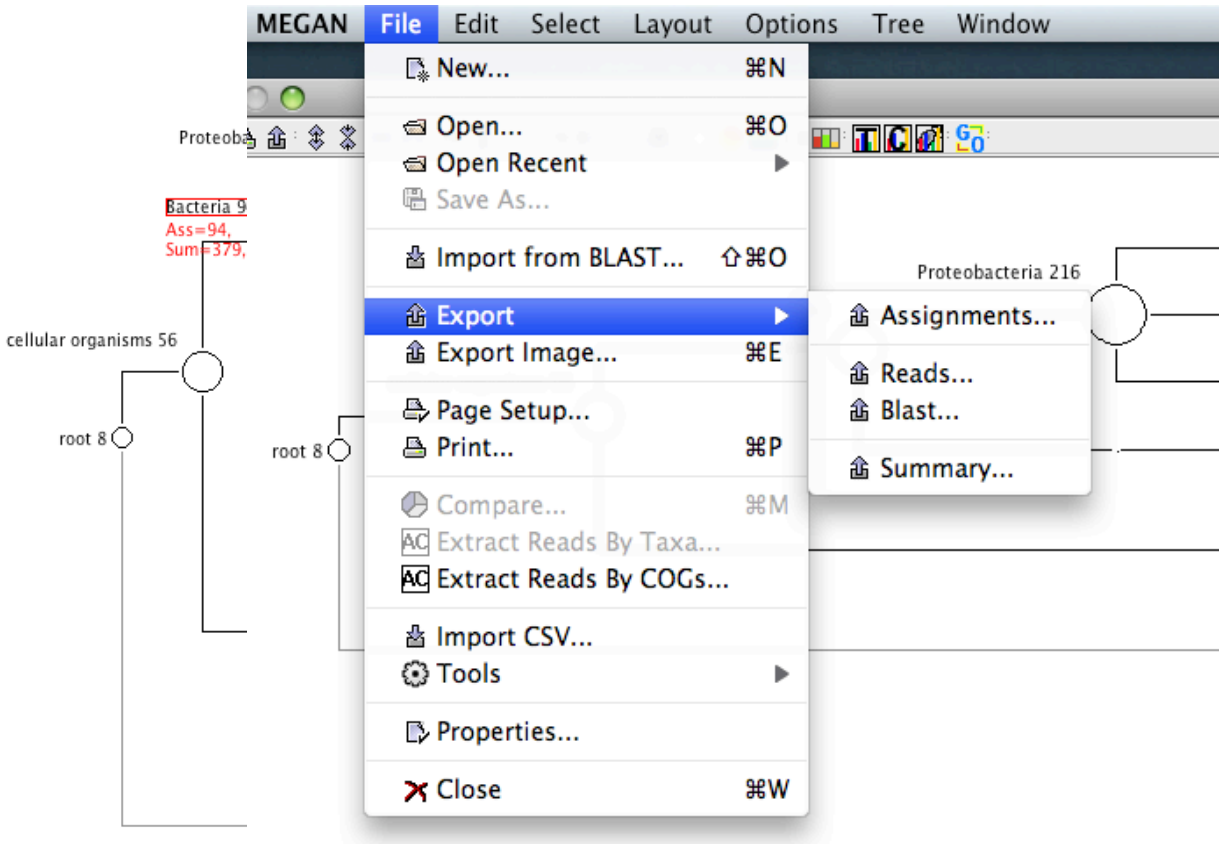


Brazelton et al., 2009

Extracting sequences

MEGAN offers several ways of extracting sequences

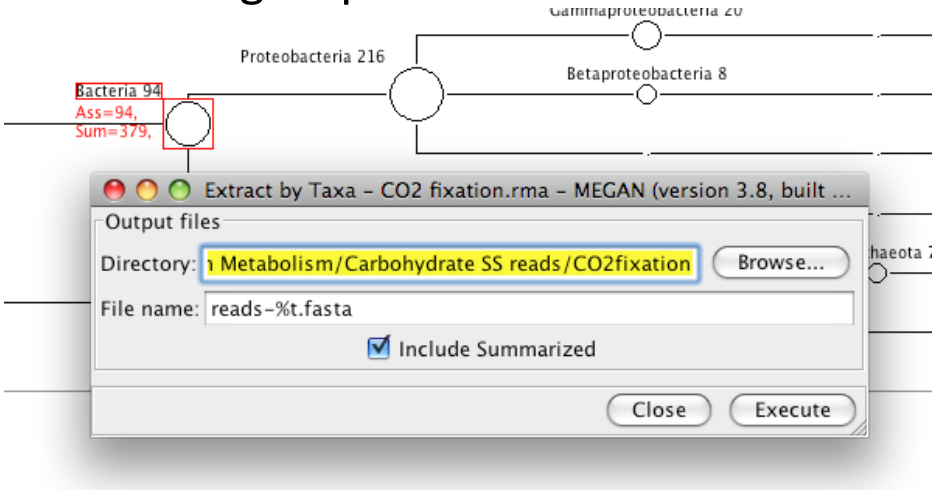
- Export reads



Extracting sequences

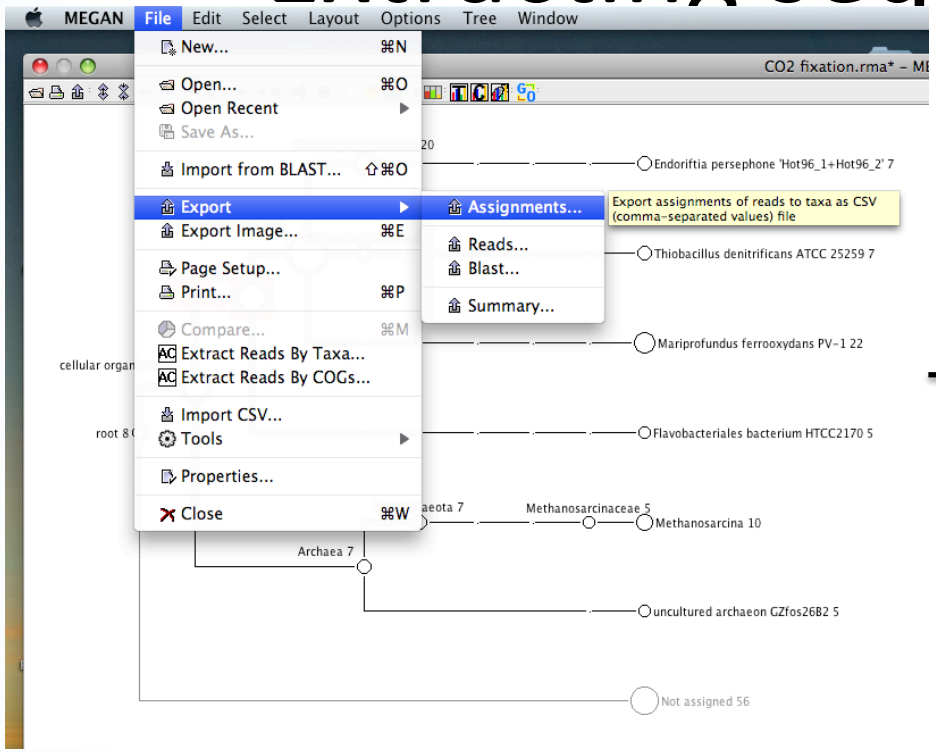
MEGAN offers several ways of extracting sequences

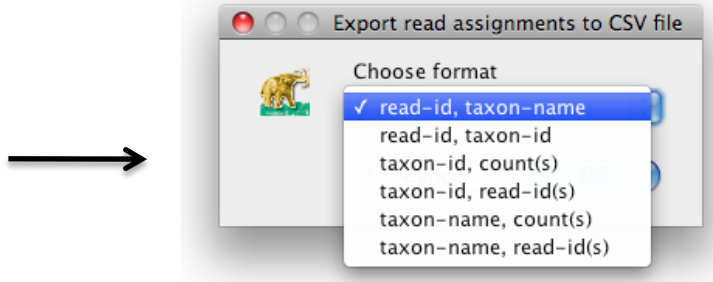
- Export reads
- Extract reads per Taxa.



This <u>only</u> works when you created your MEGAN taxonomy using both The Blastout file and your original fasta sequences.

Extracting sequences

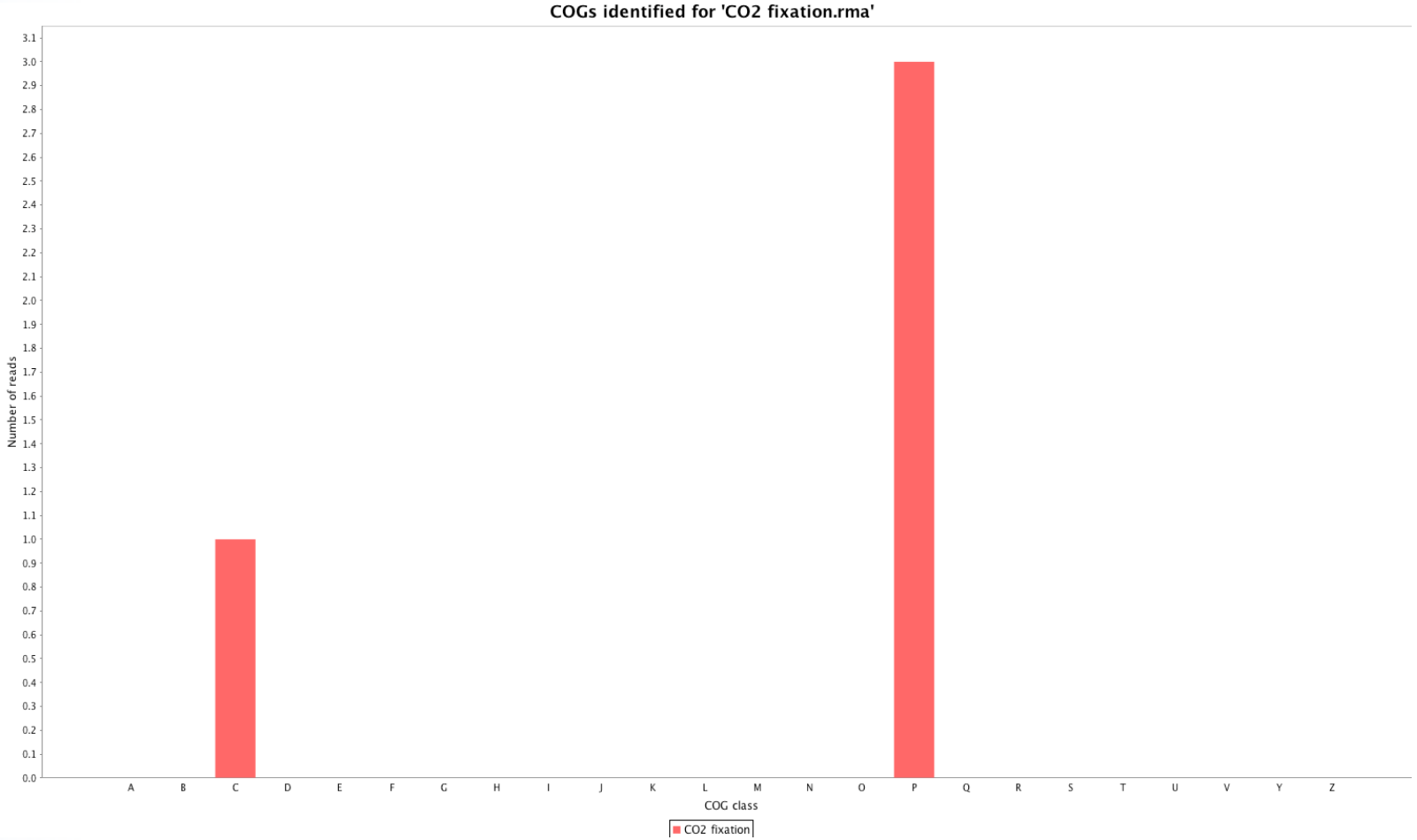




Extracting assignments can be interesting for analysis later on.

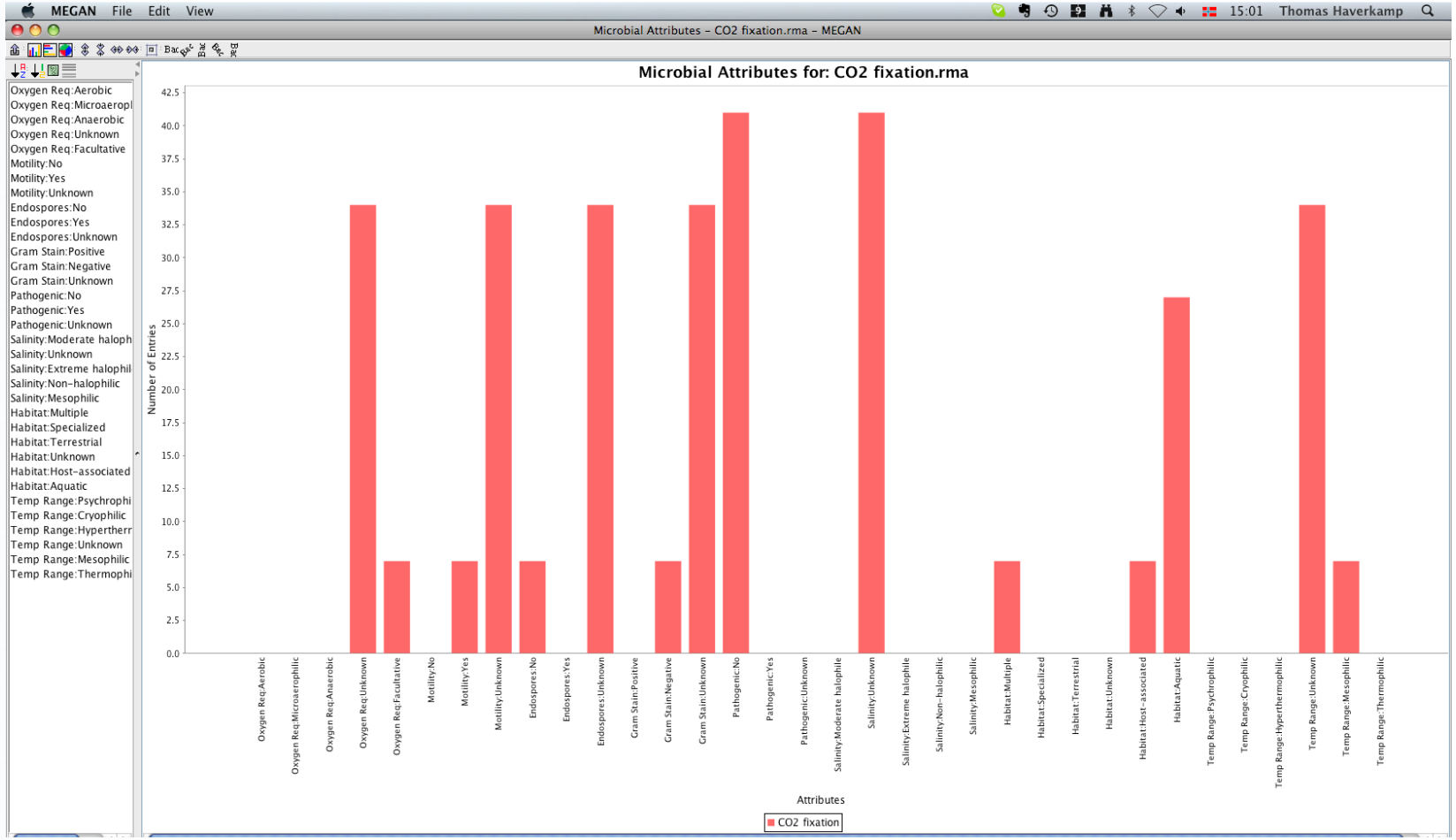
Taxon-ID: numeric ID for taxon, not the name !!!

Identification of Clusters of Orthologous Groups of proteins



Not very easy to use, since it needs you to search what COGs are actually found, GO analyzer is easier.

Microbial attributes



Is dependent of the information in the NCBI database on the species found in your dataset. Usefulness is limited.

MEGAN exercise

Two datasets:

- Carbon metabolism – 3795 reads
- Nitrogen metabolism – 999 reads

Both datasets contain annotated reads extracted from a deep sea sediment metagenome.

Annotation was done using the MetaGenomic RAST Server.

(MG-RAST: Meyer et al., 2008)

Any questions?

Adopted from: Thomas Haverkamp