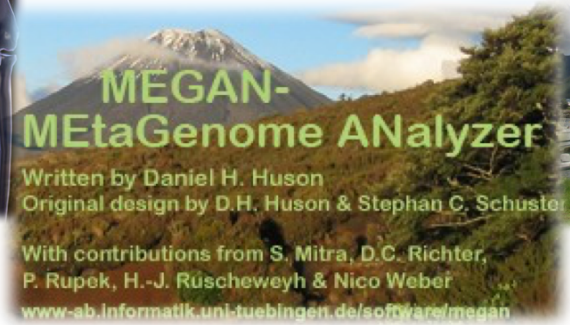
Metagenomics using MEGAN6

Who is out there?
What are they doing?
How do they compare?



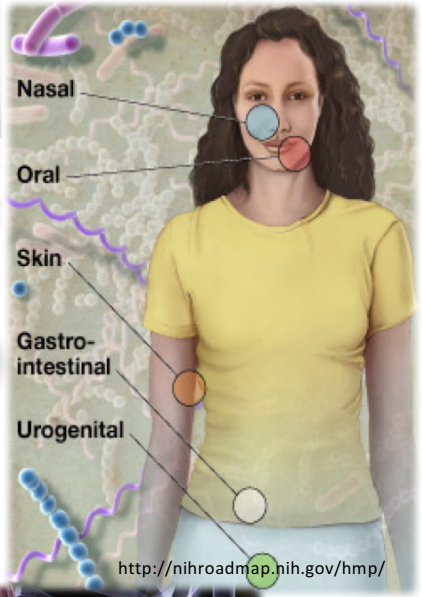
IFR Institute of Food Research

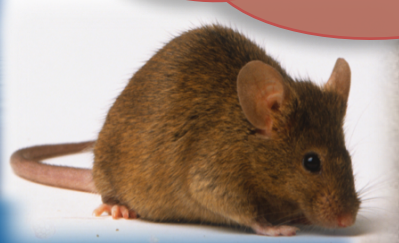
@Mitra2016

Dr. Suparna Mitra

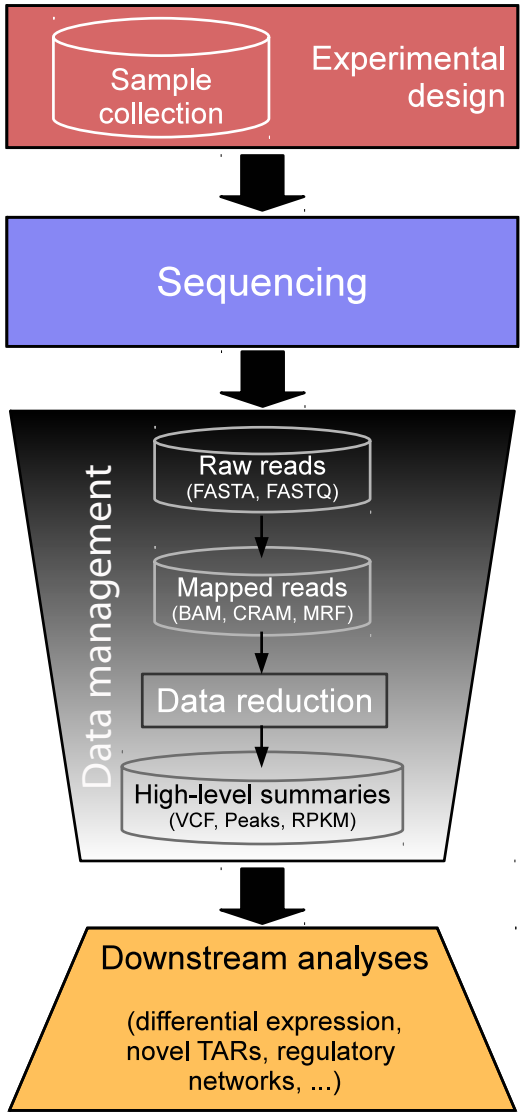
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Metagenomics

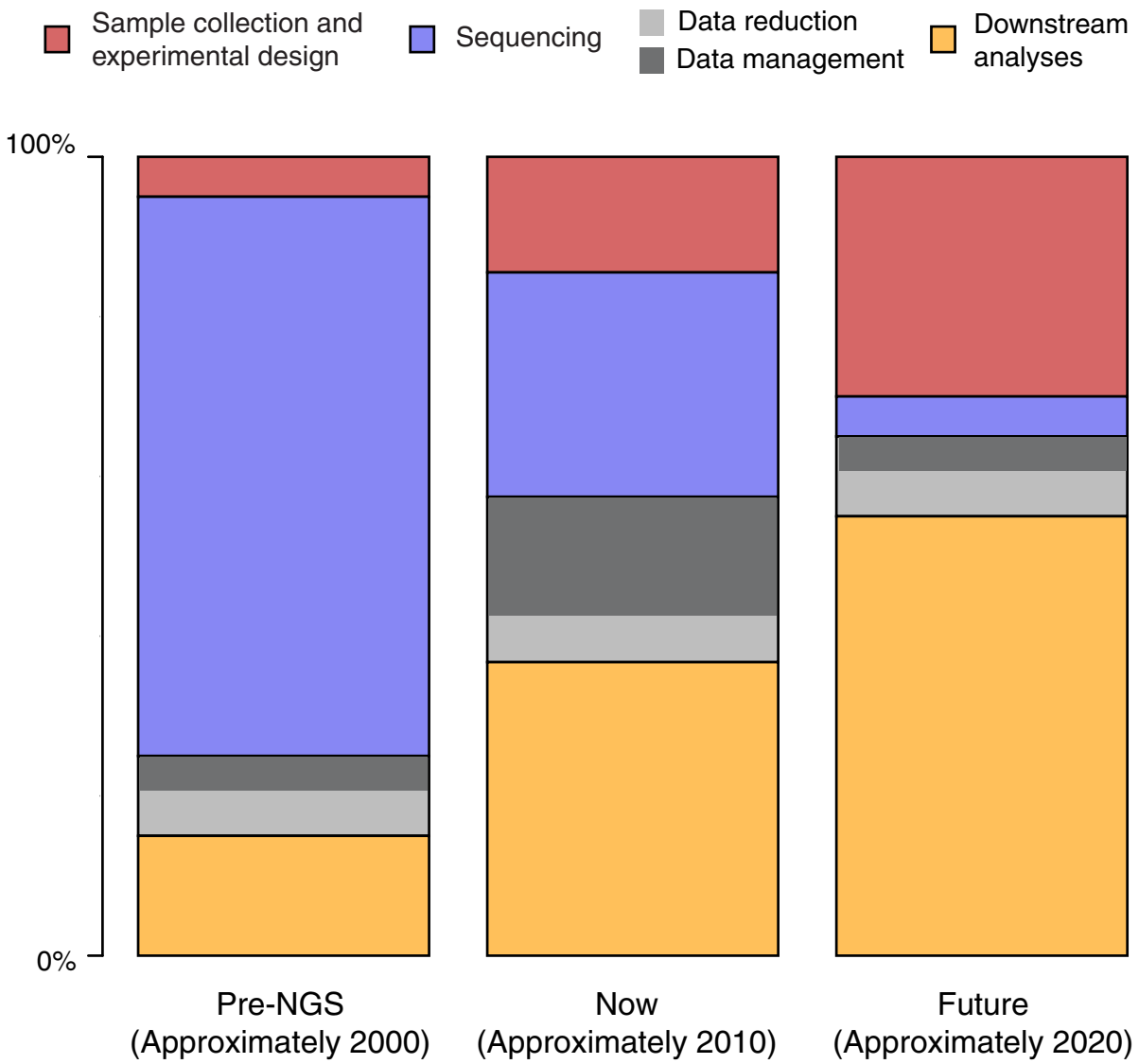




Sequencing … Bottleneck?



Sboner et al. Genome Biology 2011 12:125 doi:10.1186/gb-2011-12-8-125



Typical meta-omics study

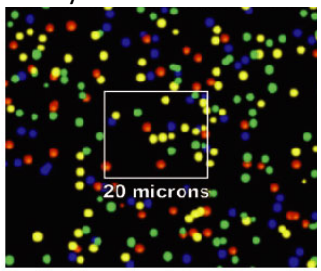
Courtesy: Dr. Daniela Drautz

Water Sample

Purify

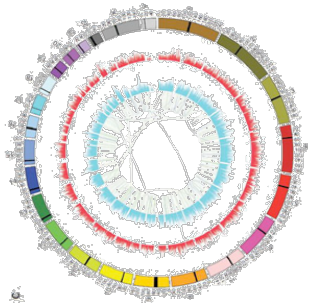
Fragment

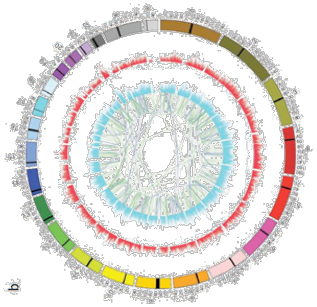
Prepare

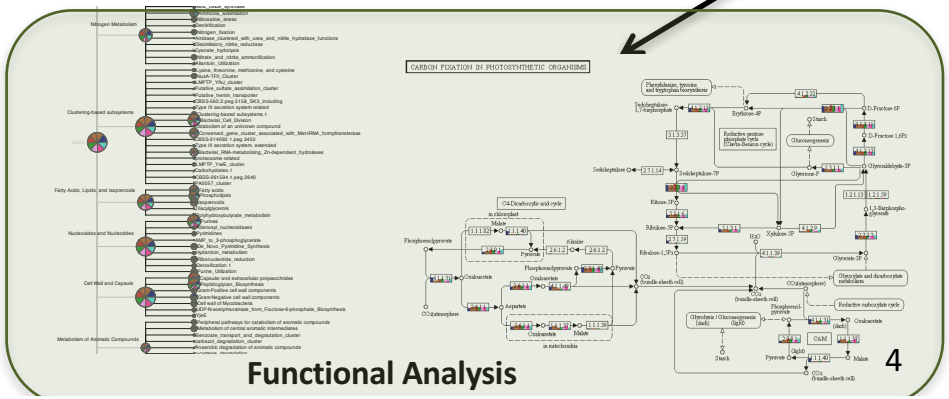


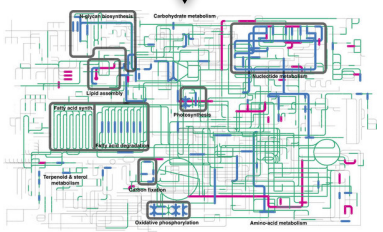
Sequencing

Genome Assembly





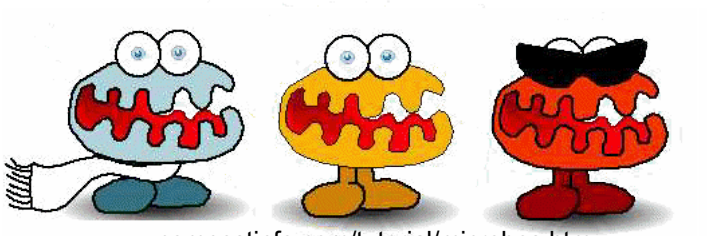




Metabolomics

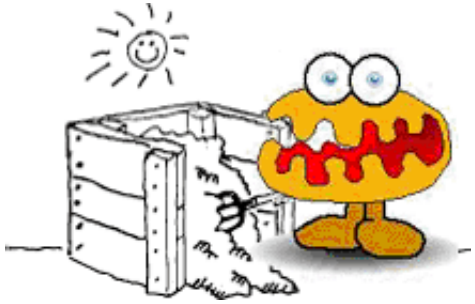
Three Basic Computational Questions

- **Who is out there?**
  - Types of organisms
  - In what proportions?

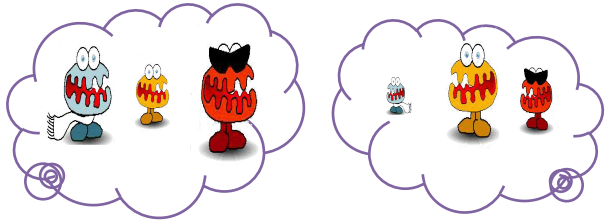


www.compostinfo.com/tutorial/microbes.htm

- **What are they doing?**
  - Types of genes
  - Which metabolic pathways?
  - In what proportions?



- **How do different samples compare?**
  - Pairwise and multiple comparisons
  - Correlations with environmental parameters?



- **Serve to answer biological or medical questions**

Who is Out There?

Taxonomic analysis

Two main approaches:

- **Targeted sequencing:**
  - **Sequence a specific gene, usually 16S rRNA, and place reads into a reference phylogeny**

- **Metagenome sequencing:**
  - **Randomly sequence DNA (or RNA) and then place reads into the NCBI taxonomy based on similarity to reference sequences**

What They are Doing?

Functional Analysis

- **A subsystem is a set of functional roles that implement a specific biological process or structural complex**
- **Genes are assigned to functional roles in subsystems using SEED**

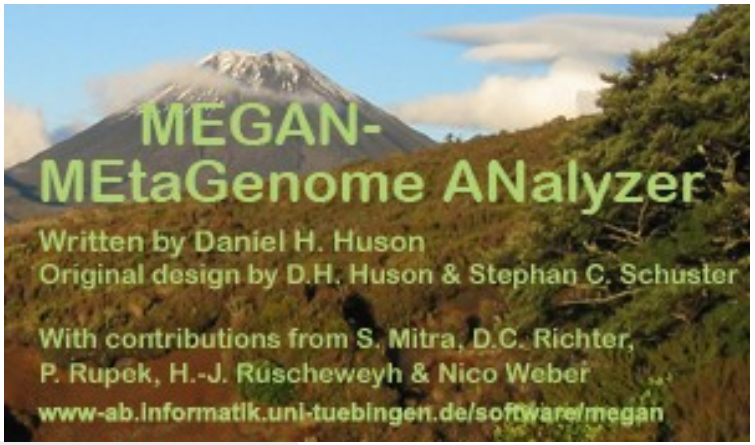
Overbeek et al., Nucleic Acids Res 33(17), 2005

- **Analysis of metabolic pathways using KEGG orthologies**

Kanehisa et al., Nucleic Acids Res 28, 2000

MEGAN – MEtaGenome ANalyzer

Huson et al, 2011



- **Interactive tool for metagenomic analysis**

MEGAN6

The most powerful interactive microbiome analysis tool

MEGAN6

The most powerful interactive microbiome analysis tool
Analyse metagenome, metatranscriptome and amplicon sequences from multiple sources

Download MEGAN6 **What's new in MEGAN6**

Microbiome analysis using a single application

MEGAN6 is a comprehensive toolbox for interactively analyzing microbiome data. All the interactive tools you need in one application.

- 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
- MEGAN parses many different types of input

Why use MEGAN6?

The software is:

1. Easy to use. MEGAN6 is a single application and all features are available through menus, toolbars and graphics. No scripting skills required.
2. Powerful. MEGAN6 allows you to work with hundreds of samples containing hundreds of millions of sequencing reads. Blast-like analysis can be performed using DIAMOND.
3. Comprehensive. MEGAN6 offers a large range of analysis tools, and is under active development.

Choose your Edition

Community Edition

Open Source version of MEGAN6

Download Community

Free download

Ultimate Edition

Premium version of MEGAN6
Additional features,
tools and support

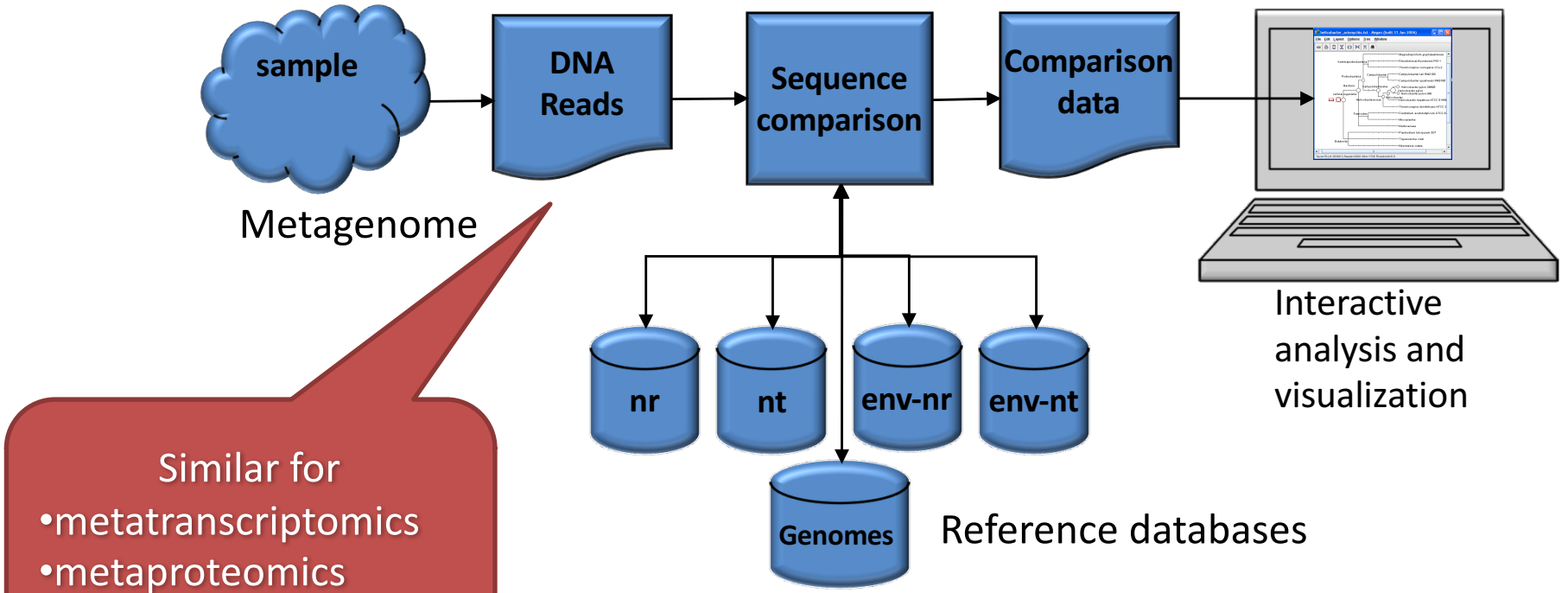
Download Ultimate

Free 30-day trial

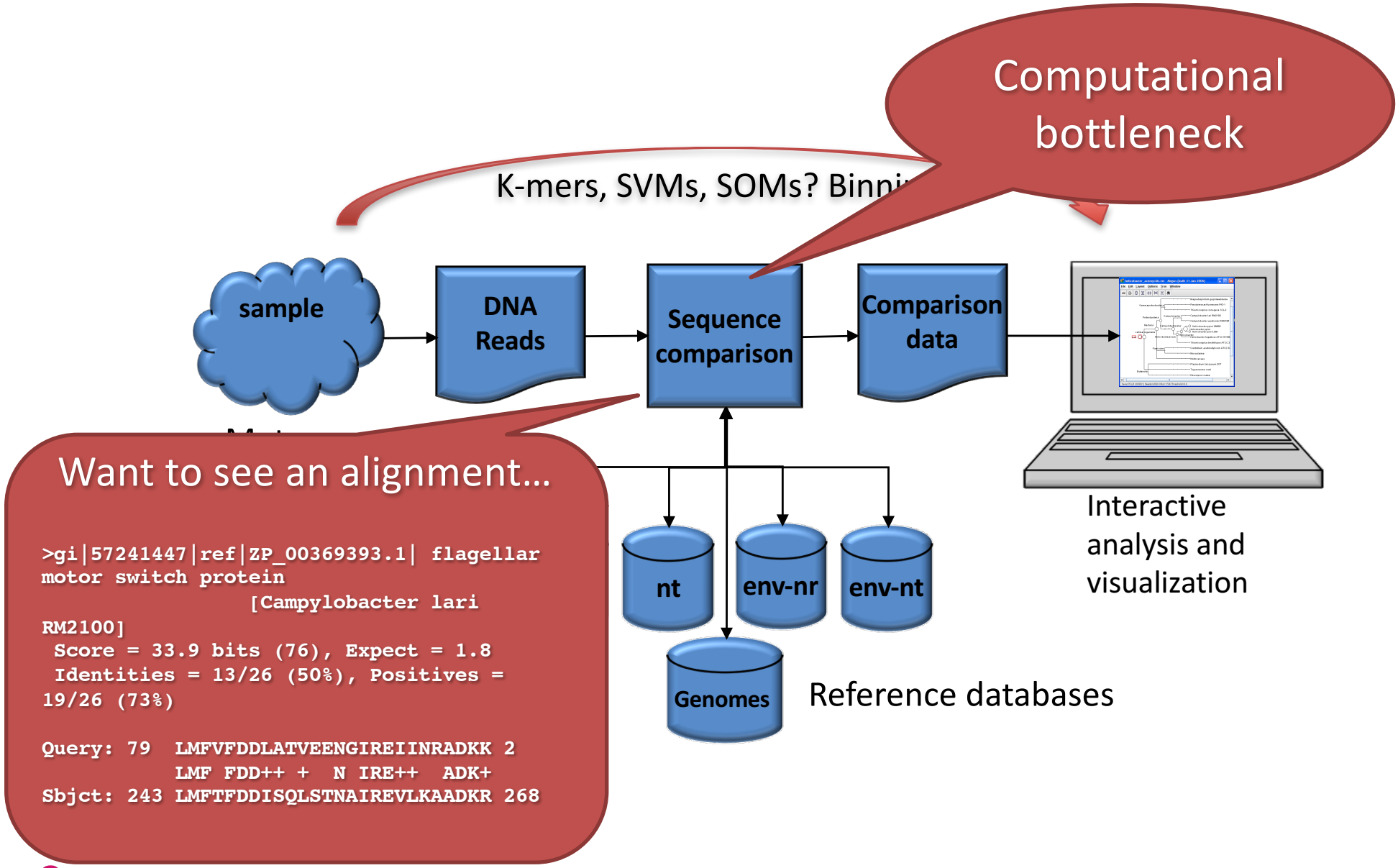
IFR Institute of Food Research

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

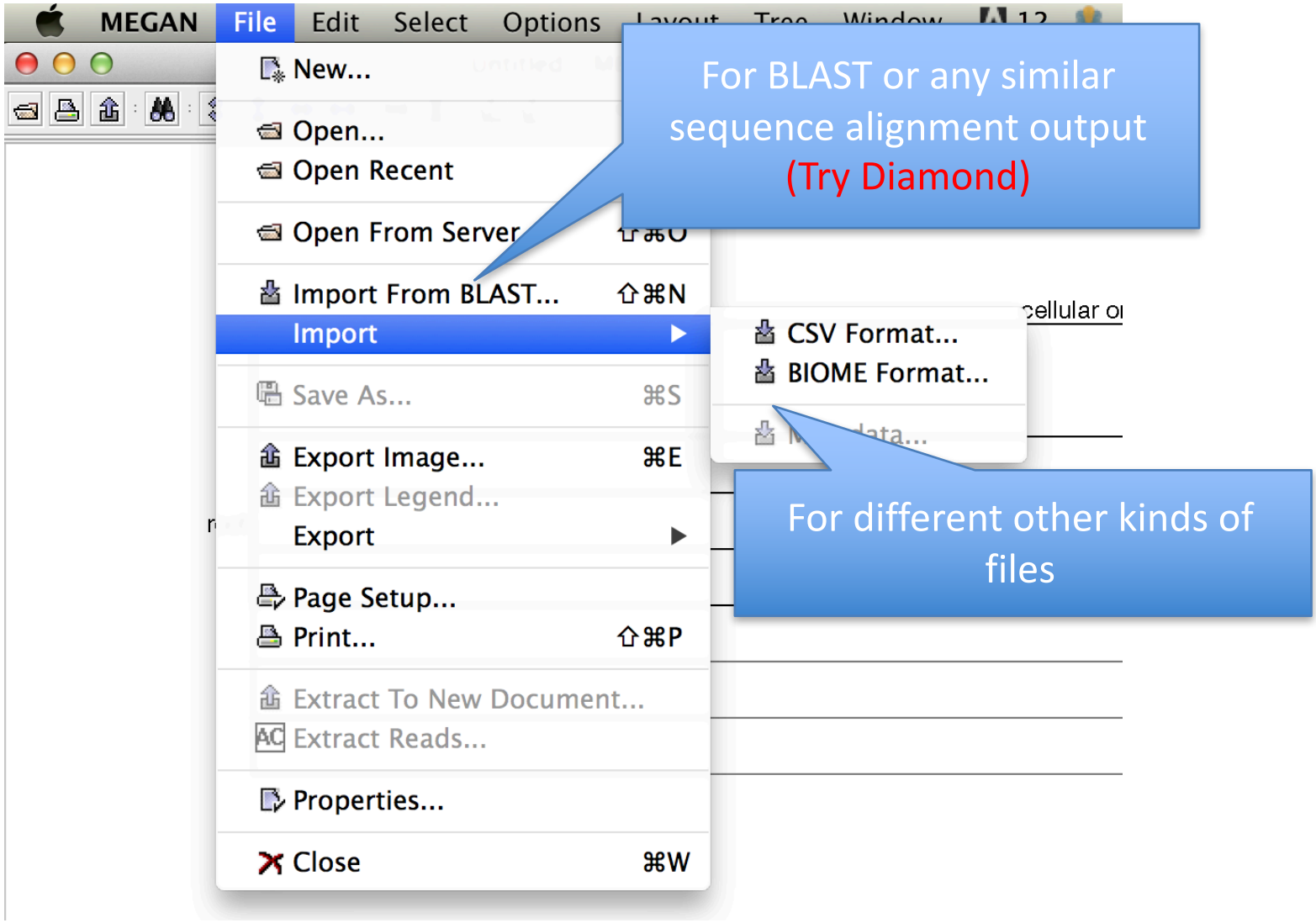
Metagenomics Pipeline



Metagenomics Pipeline

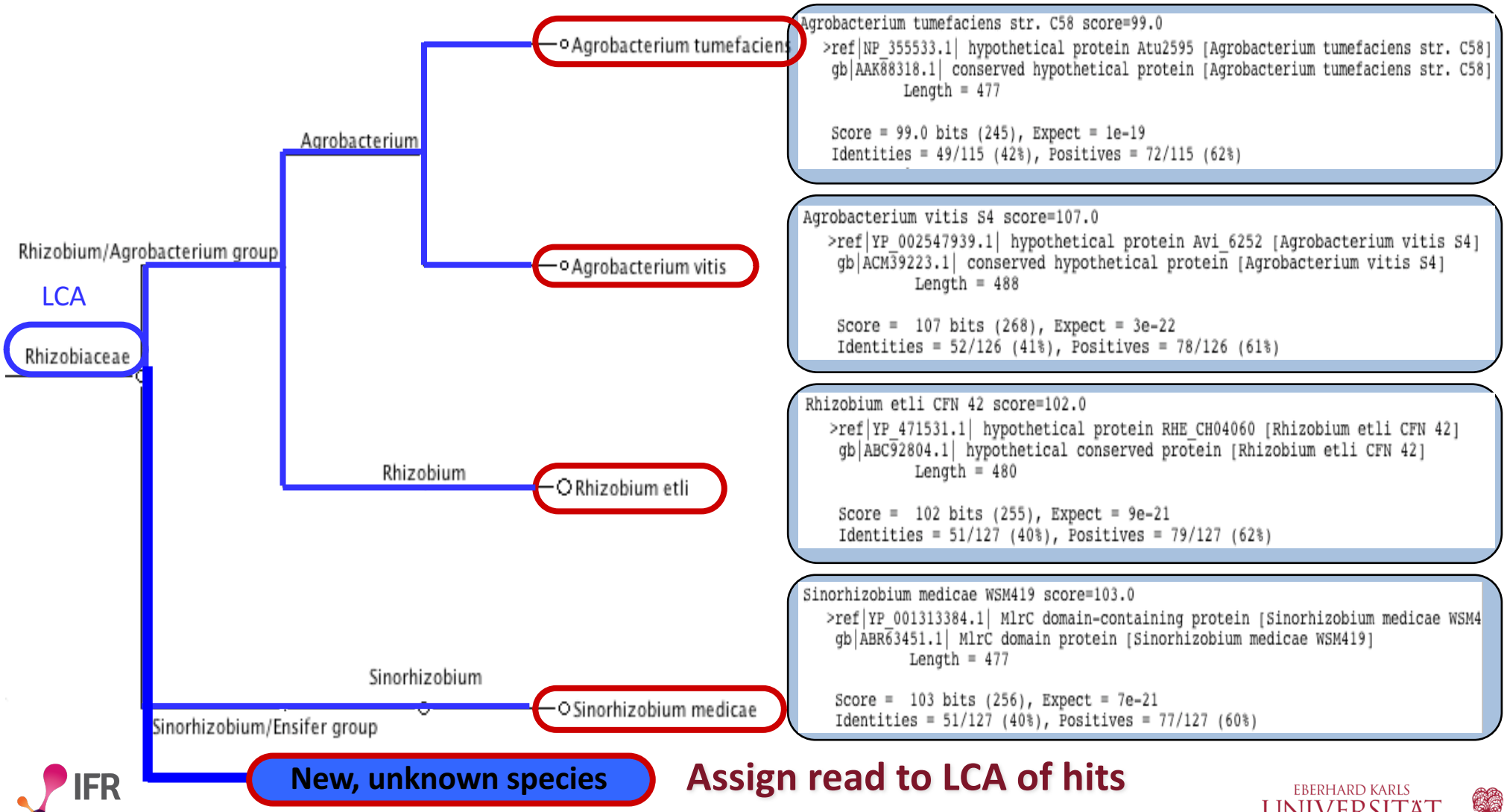


Import Options



Approach: LCA-Algorithm

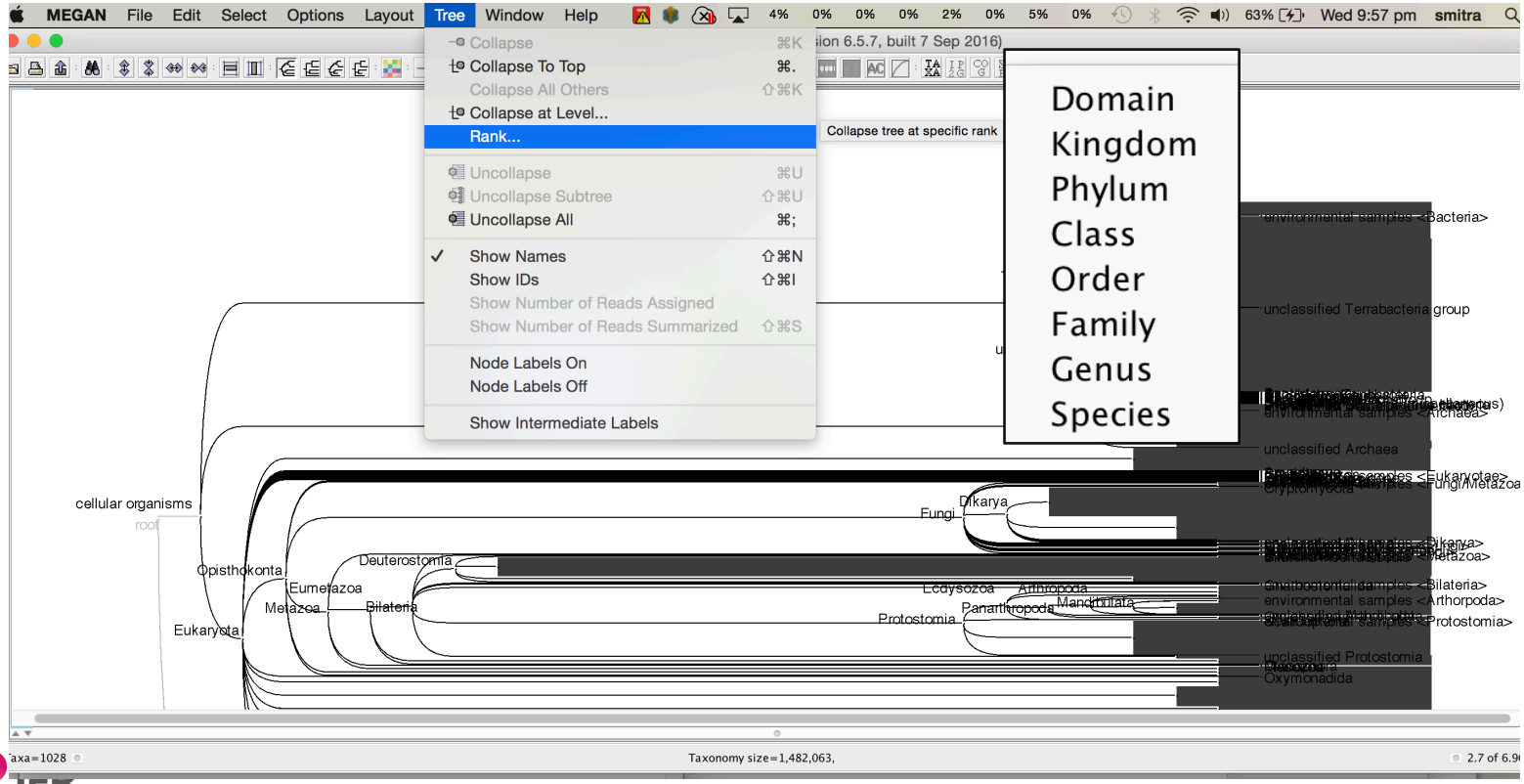
A read will often match more than one database entry

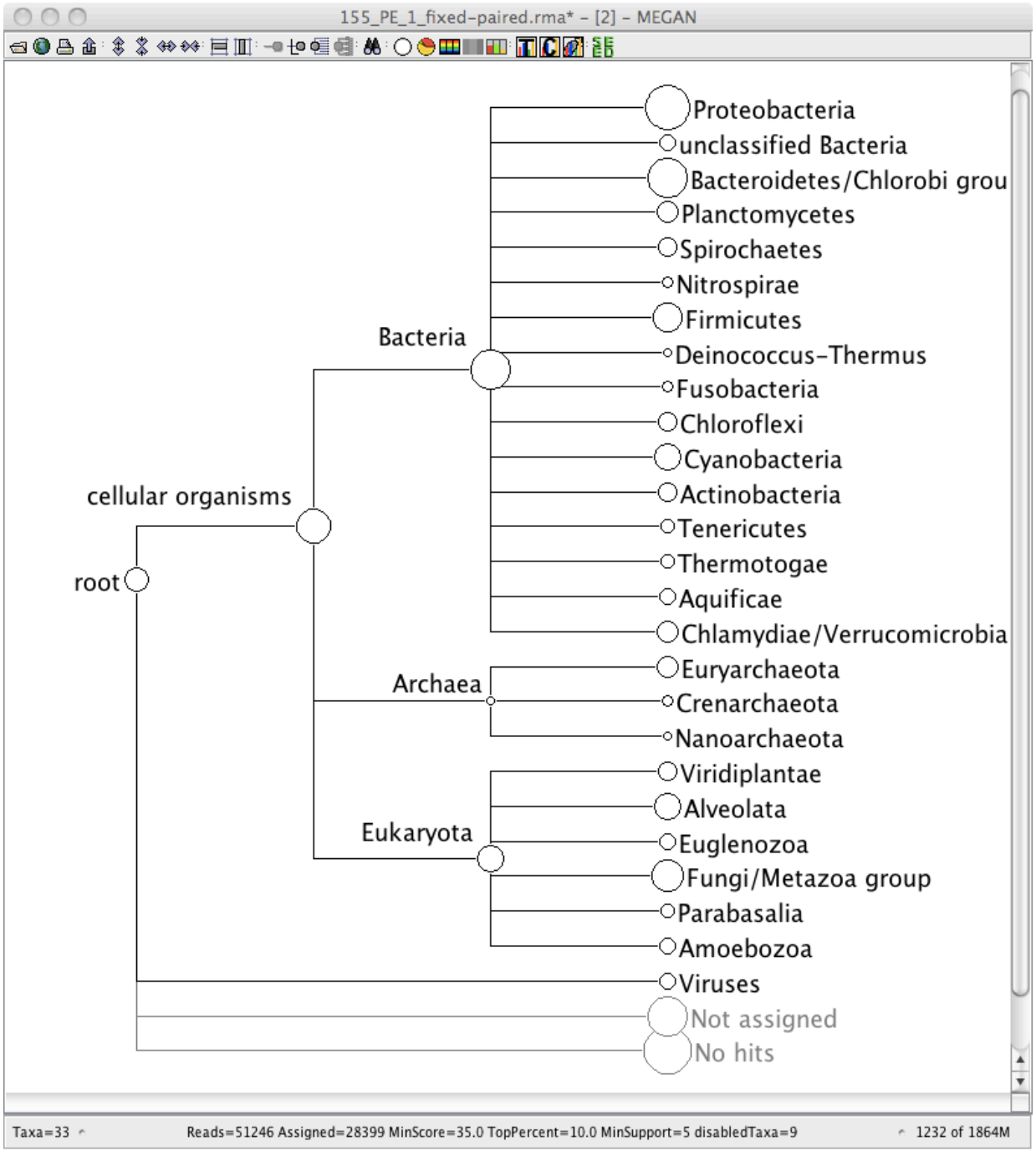


Assign read to LCA of hits

Taxonomic Binning in MEGAN

- **MEGAN uses NCBI taxonomy to organize all reads**
- **NCBI taxonomy tree is currently approx. 1,482,063 nodes and up to 25 ranks deep.**




155_PE_1_fixed-paired.rma* - [2] - MEGAN
Proteobacteria
unclassified Bacteria
Bacteroidetes/Chlorobi grou
Planctomycetes
Spirochaetes
Nitrospirae
Firmicutes
Bacteria
Deinococcus-Thermus
Fusobacteria
Chloroflexi
Cyanobacteria
cellular organisms
Actinobacteria
Tenericutes
Thermotogae
root
Aquificae
Chlamydiae/Verrucomicrobia
Euryarchaeota
Archaea
Crenarchaeota
Nanoarchaeota
Viridiplantae
Alveolata
Eukaryota
Euglenozoa
Fungi/Metazoa group
Parabasalia
Amoebozoa
Viruses
Not assigned
No hits
Taxa=33
Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9
1232 of 1864M

Data Exploration

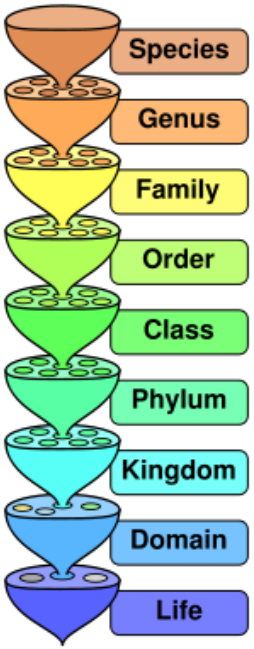
How to analyze a meta genome?

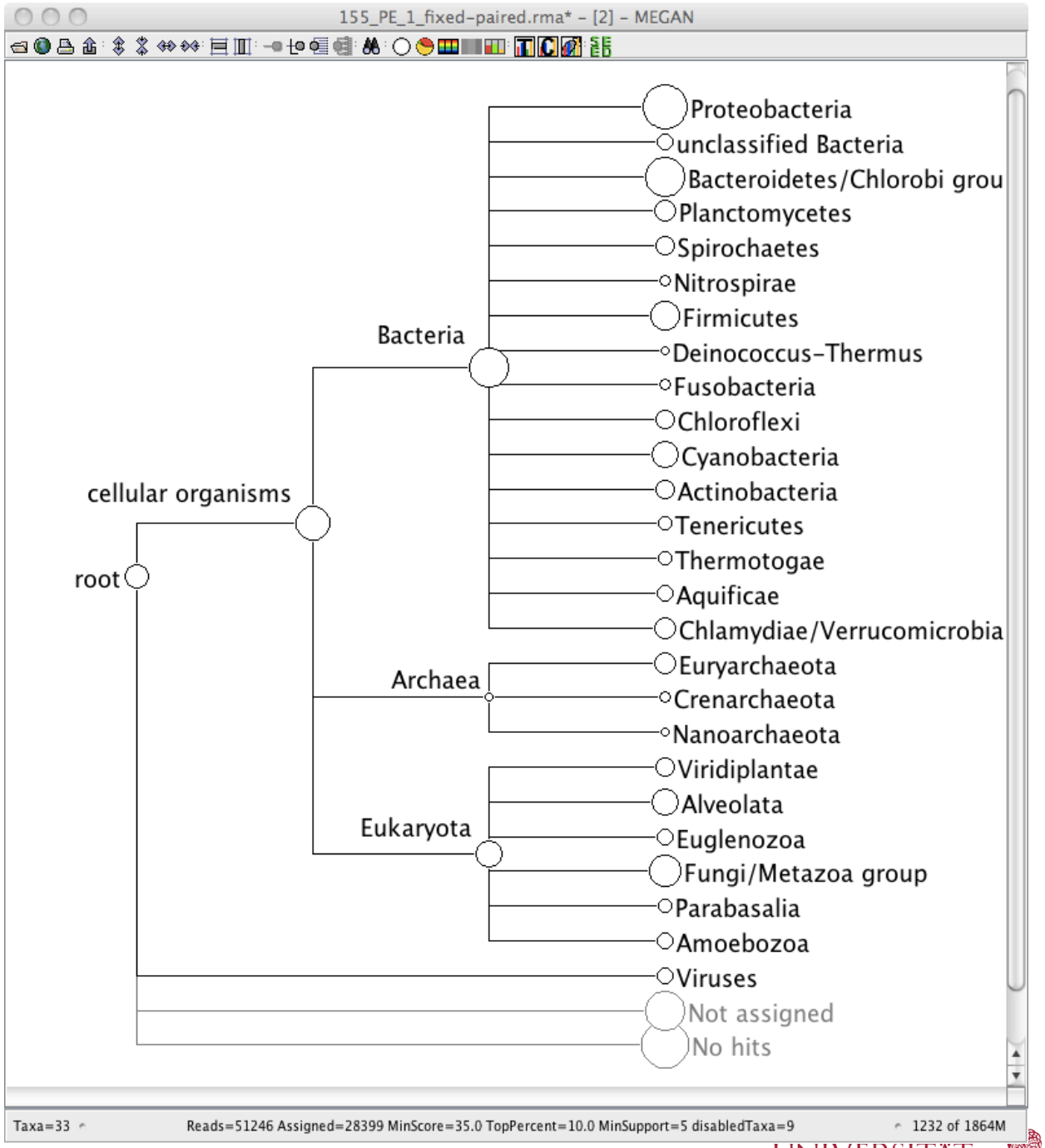
- Organize
- Visualize
- Interact
- Summarize
- Capture
- Compare

Organize and visualize

Taxonomical analysis

- ✓ **Use NCBI taxonomy to structure sequences by evolutionary relatedness of organisms**





IFR Institute of Food Research

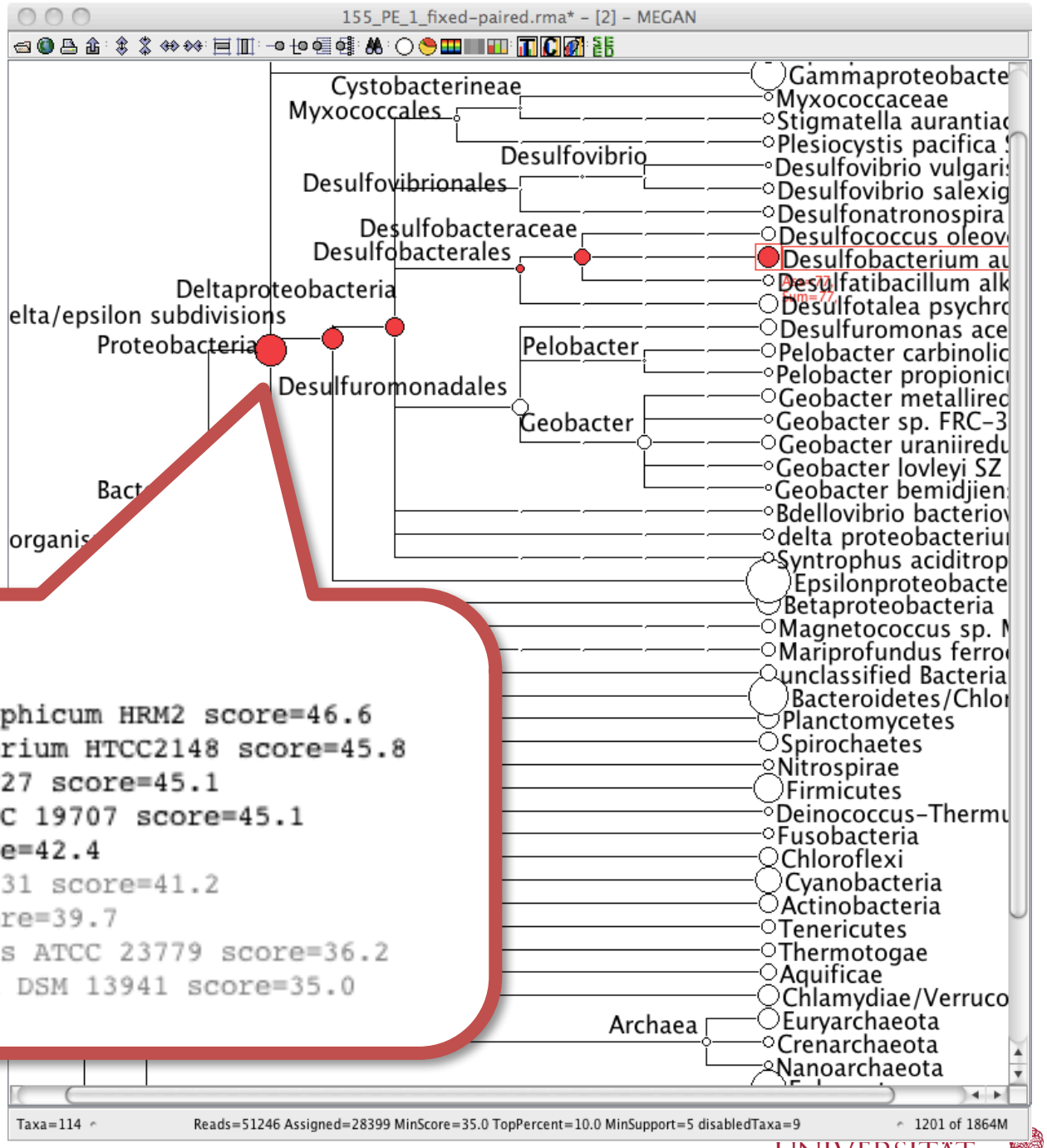
@Mitra2016

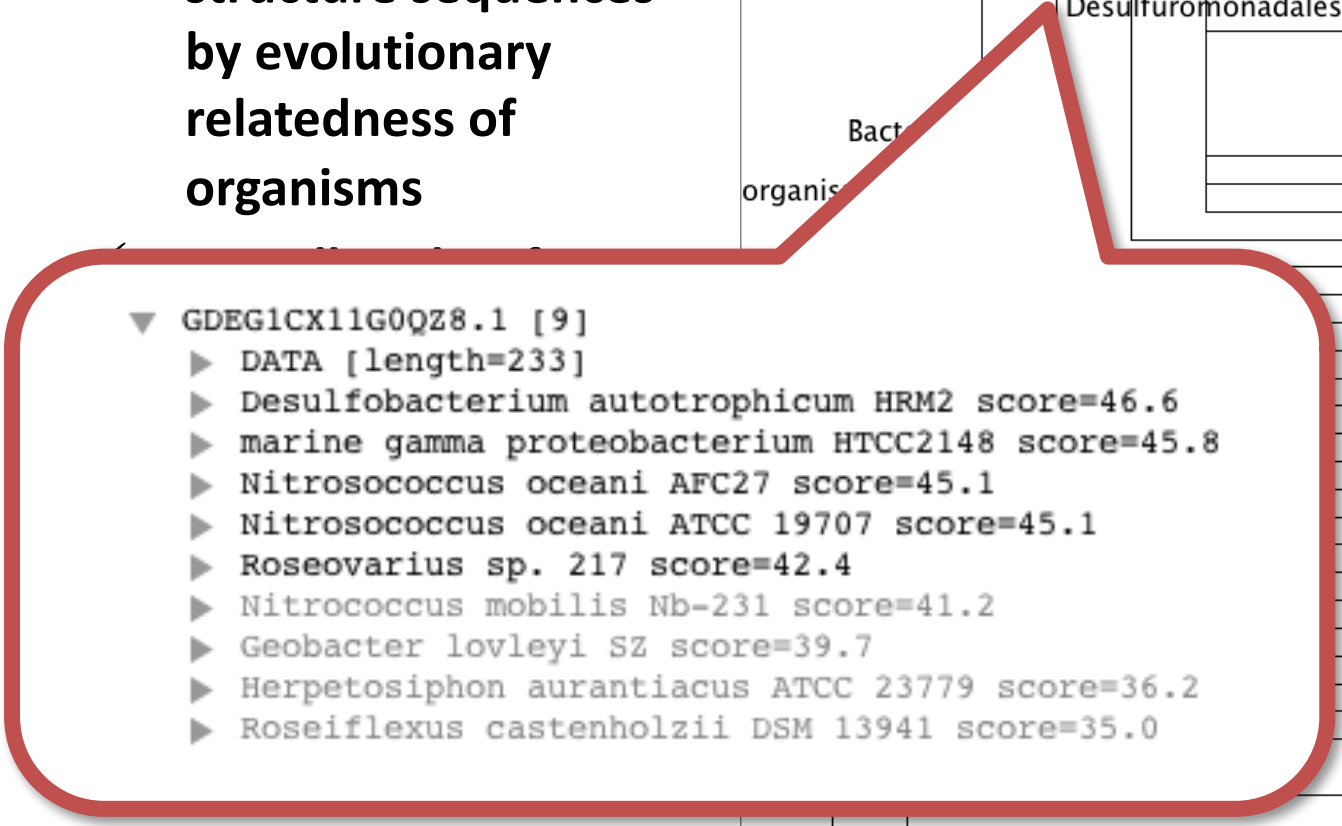
UNIVERSITÄT TÜBINGEN

Organize and visualize

Taxonomical analysis

- ✓ **Use NCBI taxonomy to structure sequences by evolutionary relatedness of organisms**

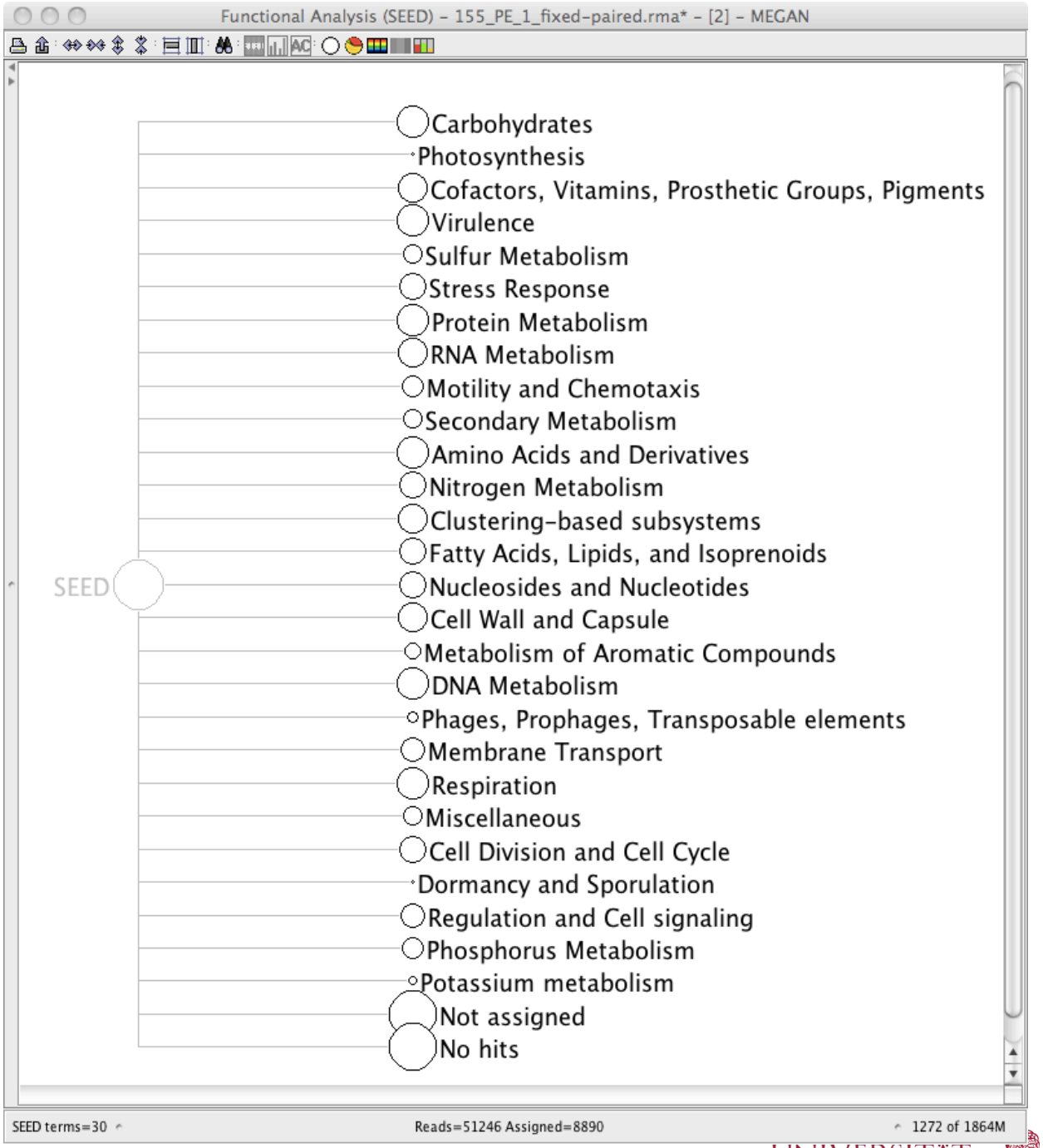




Organize and visualize

Functional analysis

- ✓ **Use SEED classification (like MG-RAST) to structure sequences by subsystems**



The SEED
Home of the SEED.

www.theseed.org

SEED: Overbeek et al.,
Nucleic Acids Res 33 (17), 2005

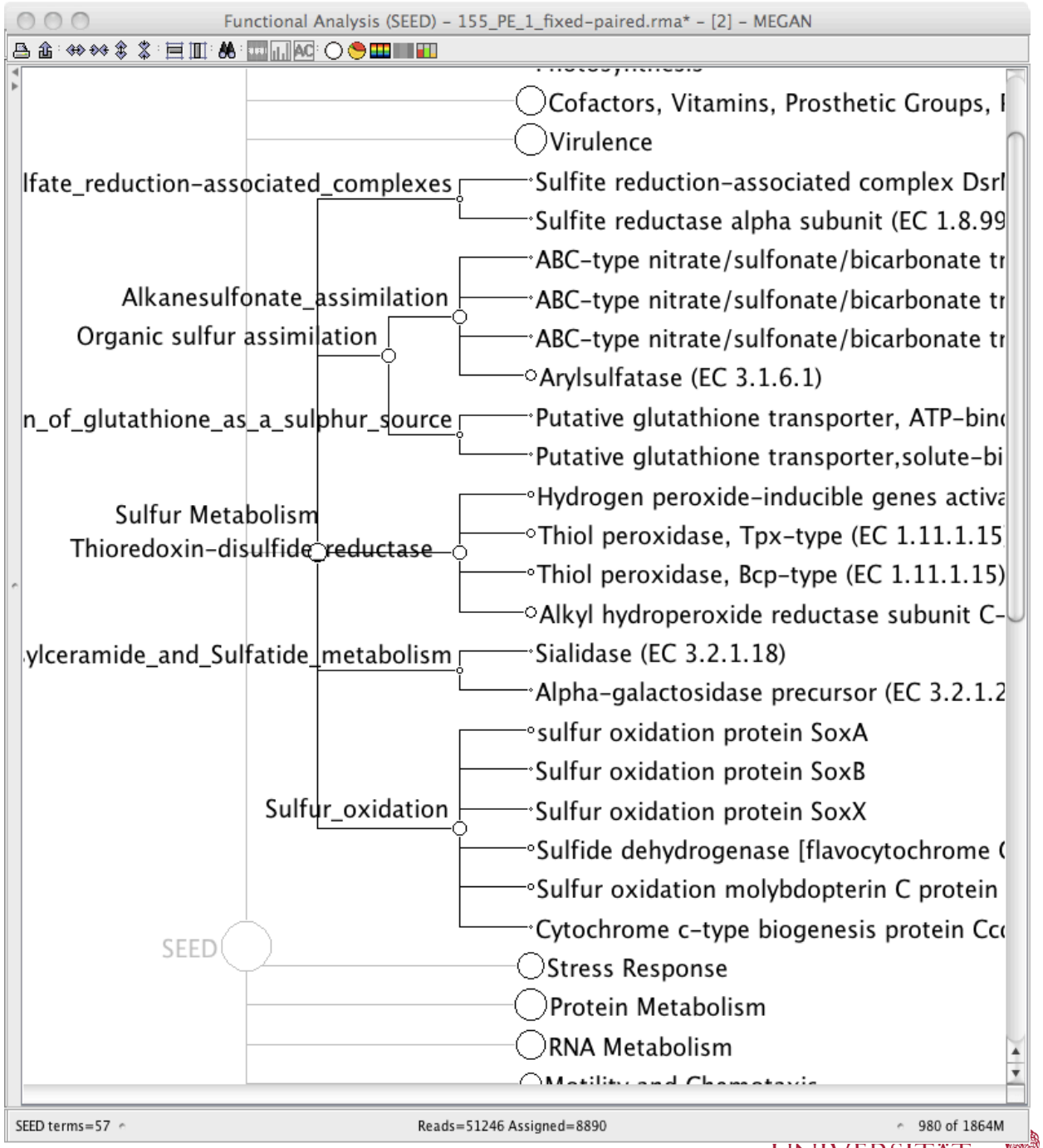
Organize and visualize

Functional analysis

- ✓ **Use SEED classification (like MG-RAST) to structure sequences by subsystems**
- ✓ **… and by functional roles**

The SEED
Home of the SEED.

www.theseed.org

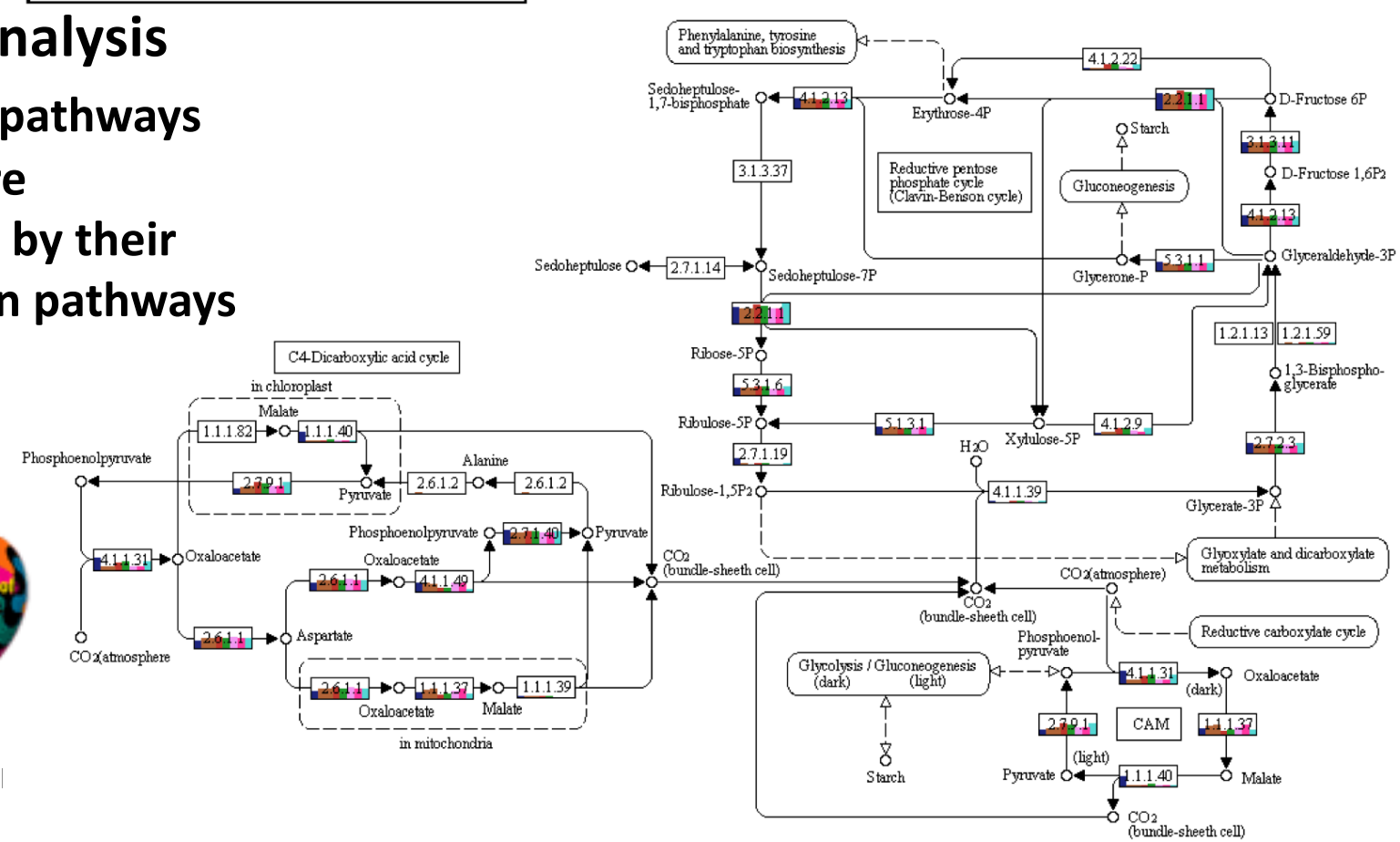


SEED: Overbeek et al.,
Nucleic Acids Res 33 (17), 2005

Organize and visualize

Functional analysis

- ✓ **Use KEGG pathways to structure sequences by their presence in pathways**

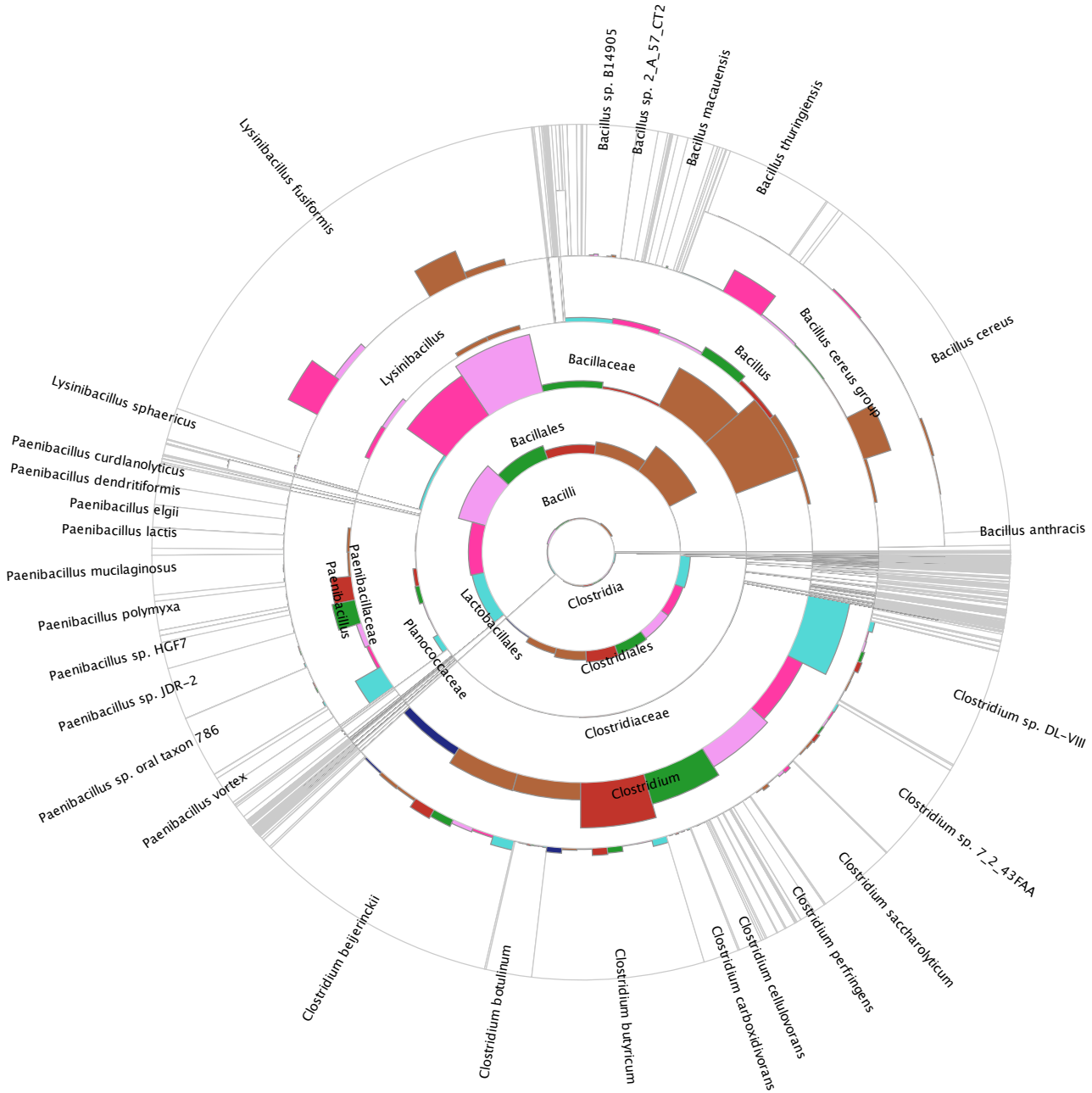


http://www.genome.j

KEGG: Kanehisa et al,
Nucleic Acids Res. 38, D355-D360 (2010)

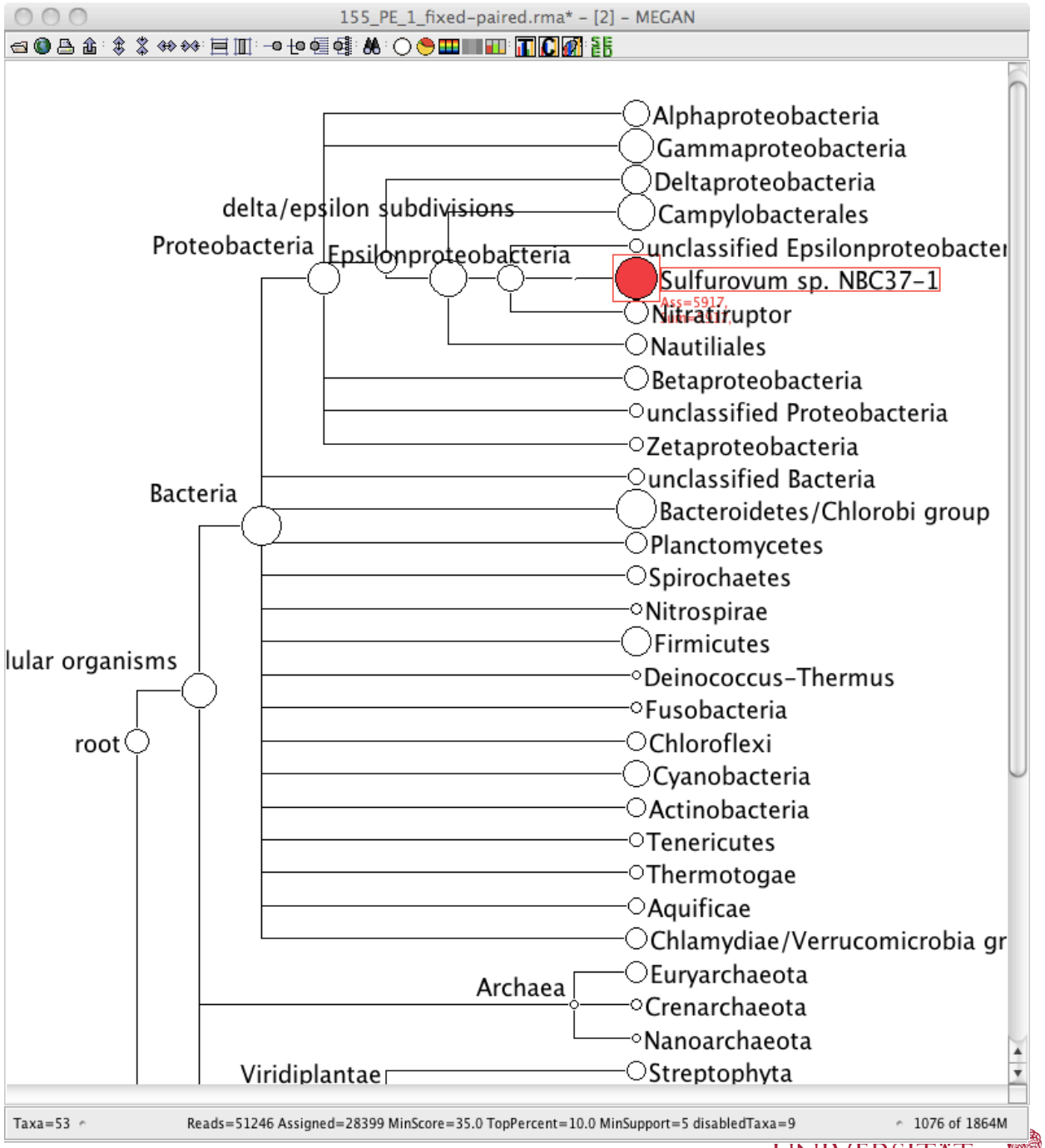
MEGAN KEGG VIEWER

Various Chart Views



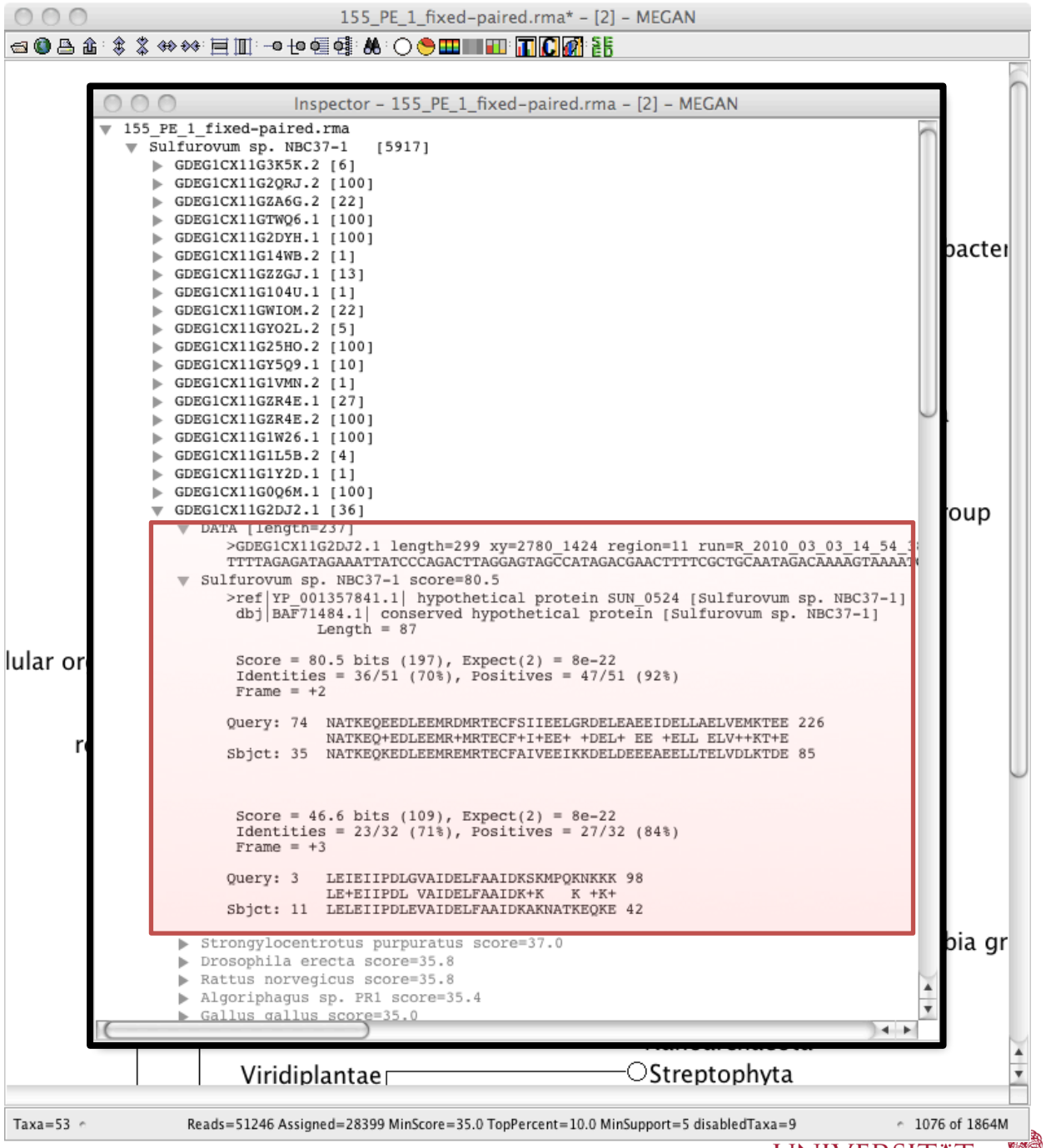
Interact and summarize

- ✓ **Search for nodes of interest**



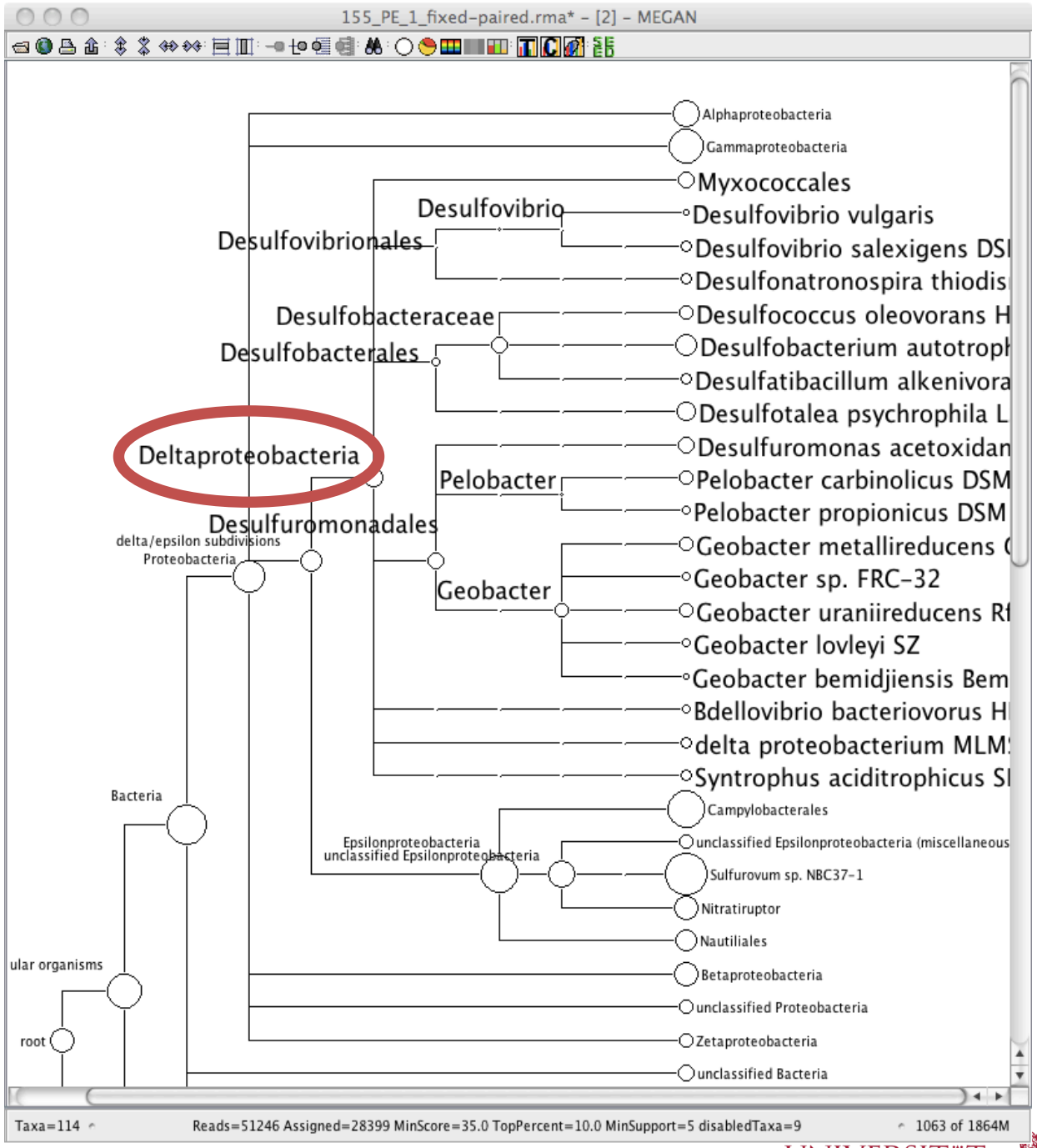
Interact and summarize

- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node



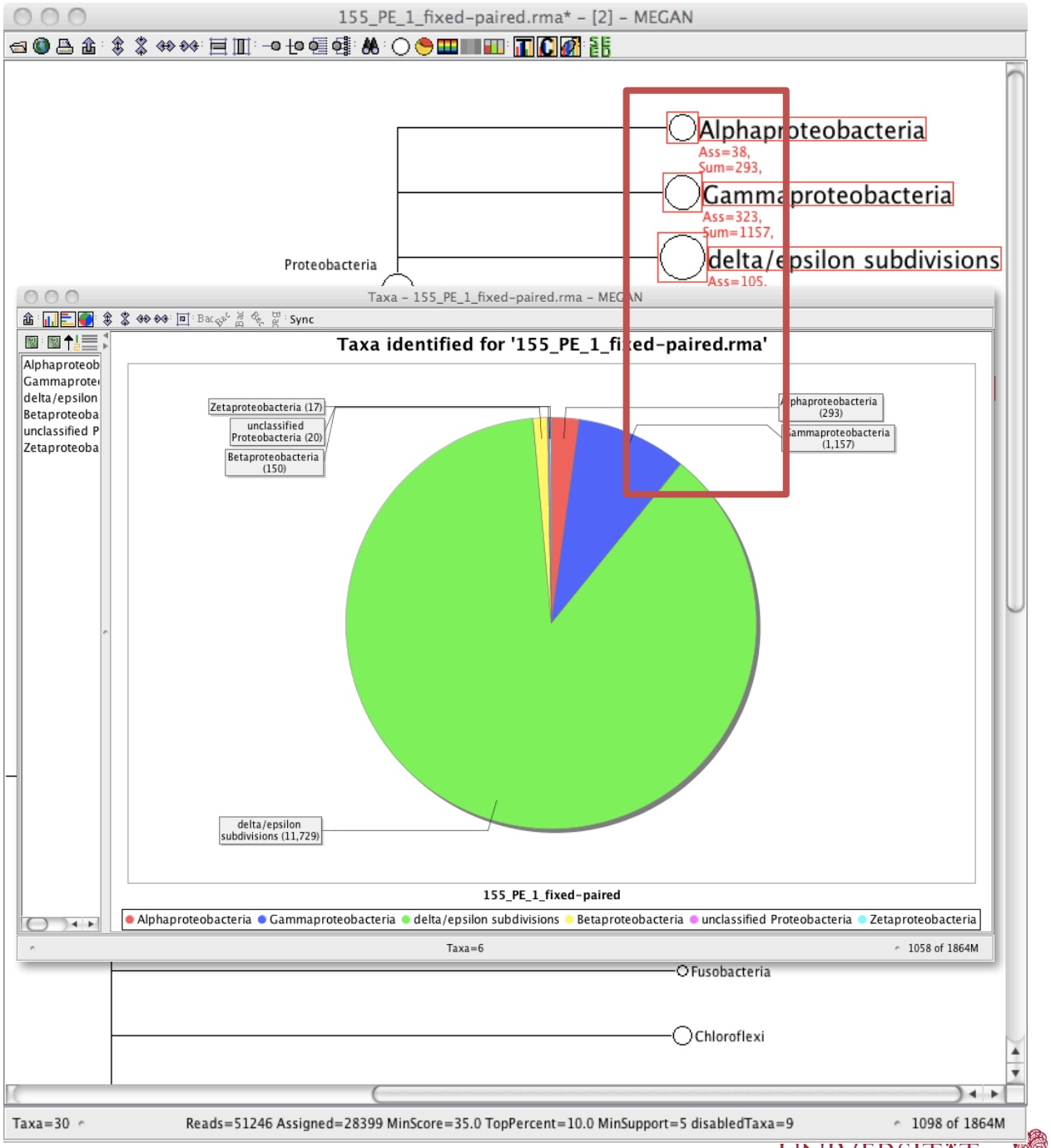
Interact and summarize

- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node
- ✓ Collapse and un-collapse parts of the tree



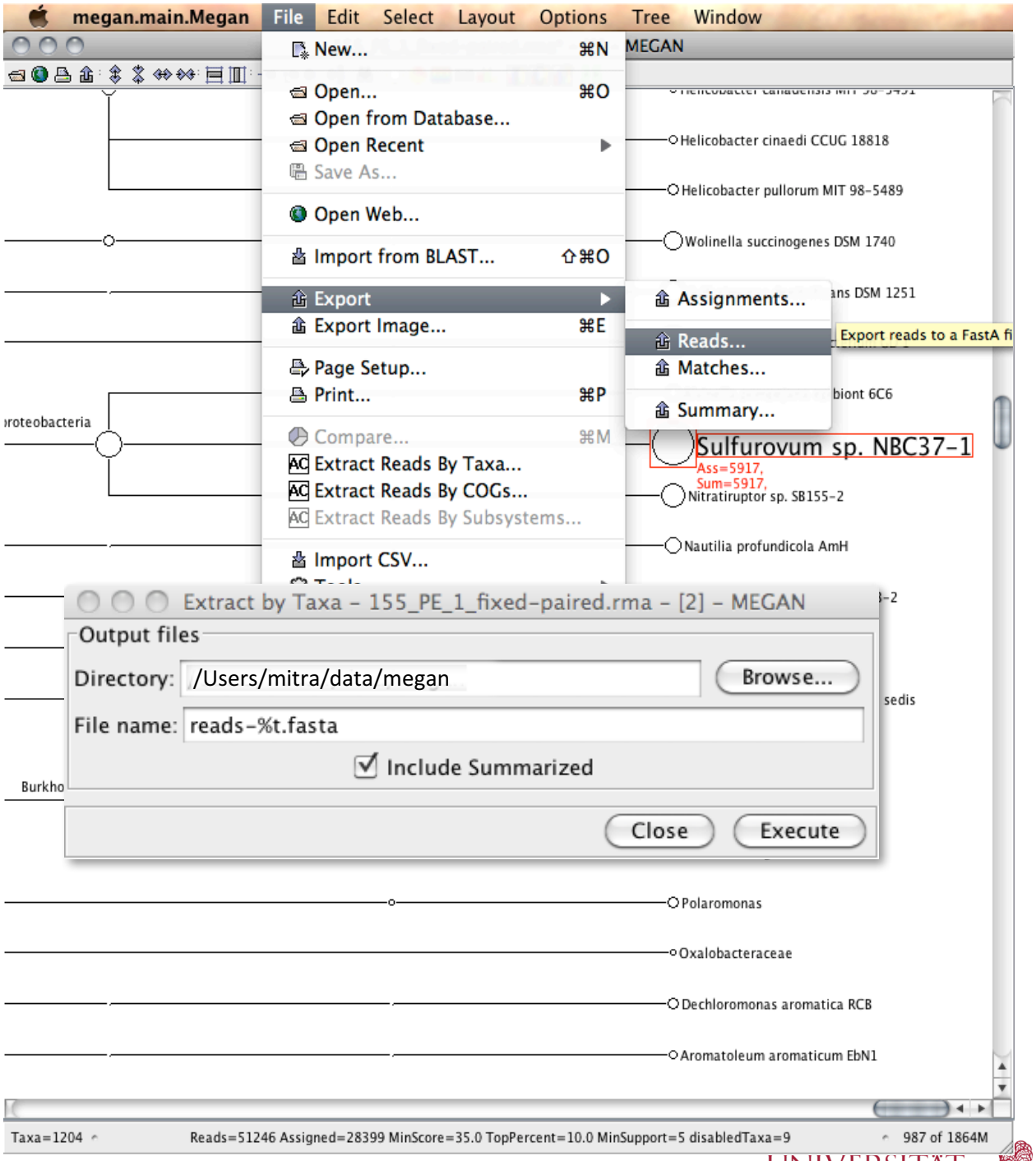
Interact and summarize

- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node
- ✓ Collapse and un-collapse parts of the tree
- ✓ Create charts



Capture

- ✓ Capture all sequences (and/or their matches) assigned to selected nodes
- ✓ Taxonomy
- ✓ SEED
- ✓ KEGG



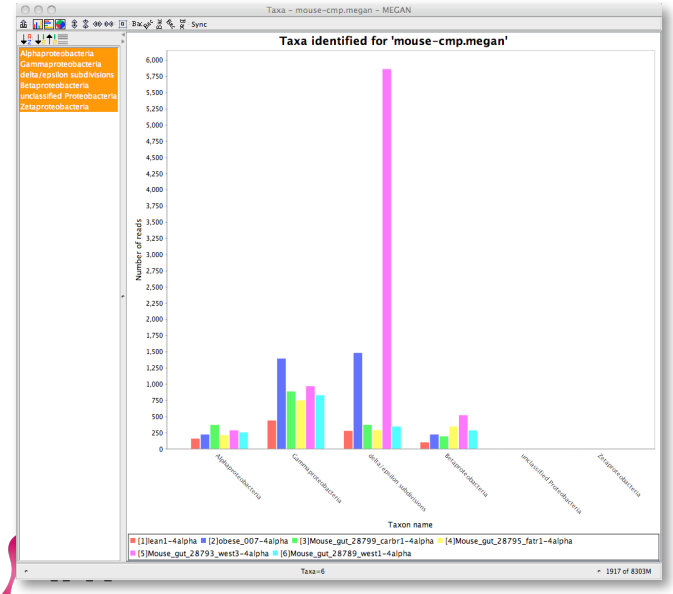
Comparative Metagenomics

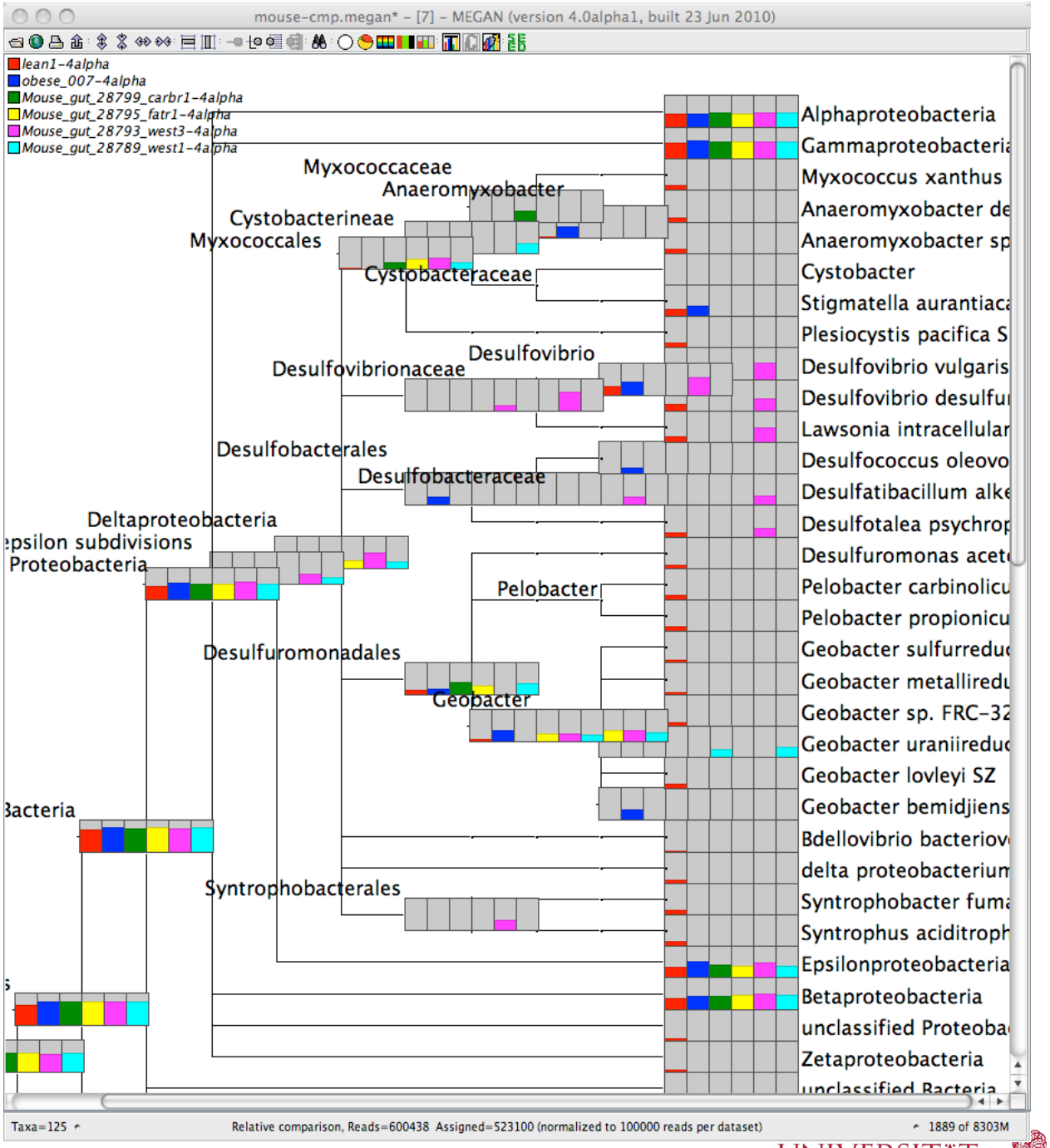
- **Why would one want to?**
  - **Structure of microbial communities**
    - **Time series studies**
    - **Based on geography**
  - **Clinical research to correlate microbes with disease**
    - **Comparing individuals**
    - **Different time point**
    - **Different drug influence**
    - **…..**
- **Methods for the comparison is needed**

Visual Comparison

Display multiple datasets simultaneously

- ✓ **Taxonomical comparison**
- ✓ **Interact**
- ✓ **... and summarize**

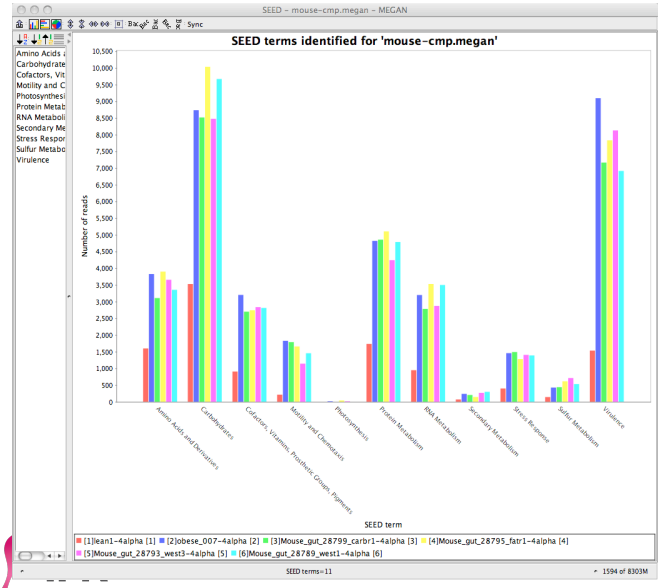




Visual Comparison

Display multiple datasets simultaneously

- ✓ **Taxonomical comparison**
- ✓ **Interact**
- ✓ **... and summarize**
- ✓ **Functional comparison**



Pair-wise Comparison

obese1
lean1

Bacteria
Ass=8614, 7864,
Sum=31566, 34220,
UPv=0.04479146
DPv=0.0

Bacteroidetes/Chlorobi group
Ass=76, 265,
Sum=3377, 9979,
UPv=0.0
DPv=0.030413214

Bacteroidetes
Ass=439, 1594,
Sum=3269, 9657,
UPv=0.97989726
DPv=7.451327E-4

Flavobacteriales

Flavobacteria

Sphingobacteriales

Bacteroidales
Ass=654, 2552,
Sum=2714, 7771,
UPv=1.0
DPv=8.859298E-23

Planctomycetales

Spirochaetales

Bacillales
Ass=14, 17,
Sum=272, 208,
UPv=9.810656E-14
DPv=0.872849

Bacilli
Ass=70, 69,
Sum=1247, 1778,
UPv=0.0
DPv=7.798125E-15

Lactobacillales
Ass=91, 92,
Sum=905, 1501,
UPv=2.6991946E-15
DPv=1.3411514E-24

Firmicutes
Ass=2029, 1284,
Sum=18111, 13478,
UPv=0.0
DPv=0.0

Clostridia
Ass=444, 270,
Sum=14742, 10308,
UPv=1.6338725E-26
DPv=0.31346714

Clostridiales
Ass=6506, 4500,
Sum=14209, 9974,
UPv=0.11779715
DPv=0.0

Thermoanaerobacterales

Deinococci

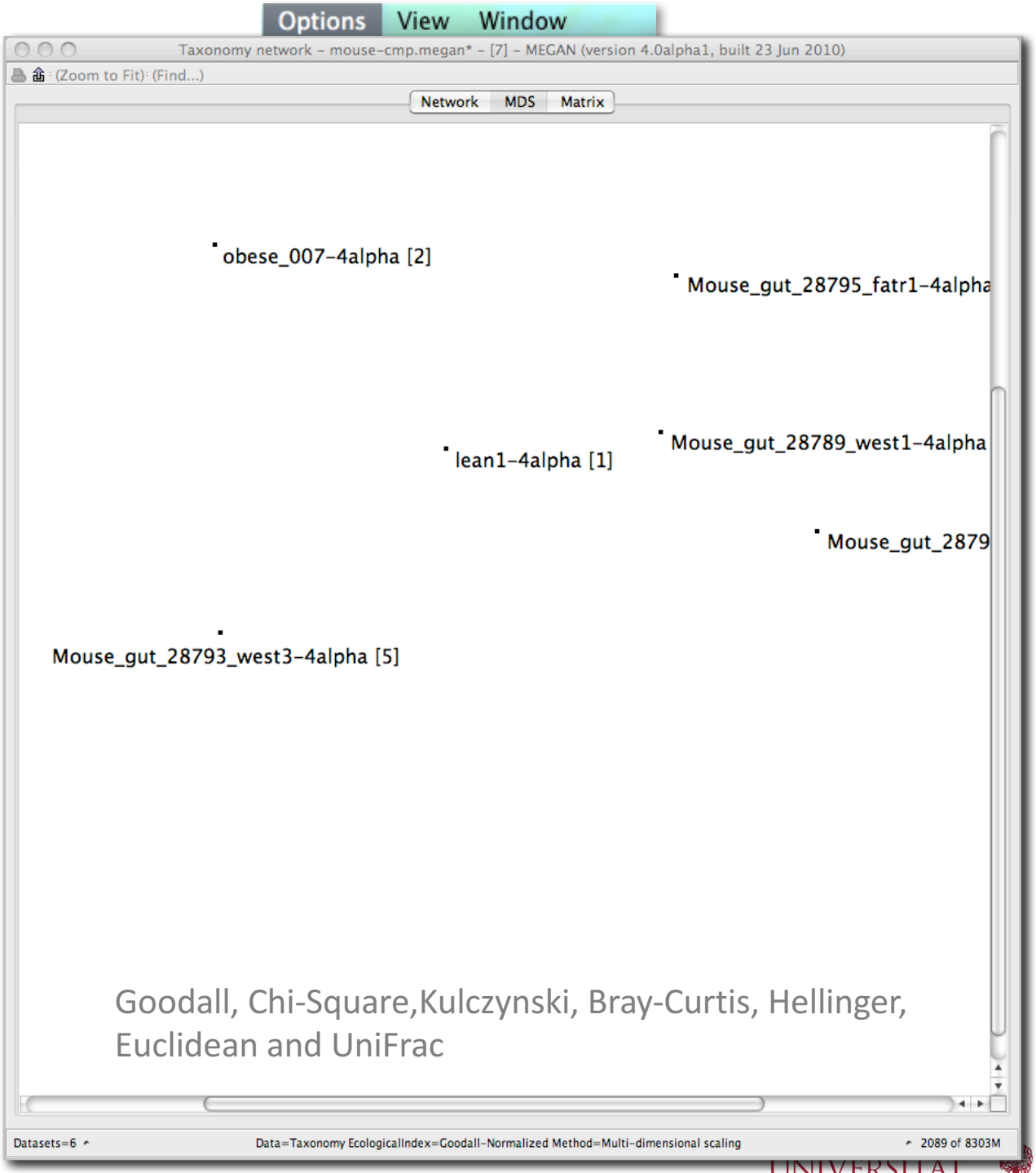
Methylophilaceae
Nitrosomonadaceae
unclassified Proteobacteria
Mariprofundaceae
candidate division TM7
Flavobacteriaceae
unclassified Flavobacteriales
unclassified Flavobacteria
Flexibacteraceae
Cyclobacteriaceae
Sphingobacteriaceae
Rhodothermaceae
Prevotellaceae
Bacteroidaceae
Porphyromonadaceae
Rikenellaceae
environmental samples
Chlorobiaceae
Planctomycetaceae
unclassified Planctomycetales
Spirochaetaceae
Leptospiraceae
Brachyspiraceae
Nitrospiraceae
Staphylococcaceae
Bacillaceae
Paenibacillaceae
Listeriaceae
Bacillales Family XII. Incertae Sedis
Leuconostocaceae
Streptococcaceae
Enterococcaceae
Lactobacillaceae
Lachnospiraceae
Veillonellaceae
Peptostreptococcaceae
Clostridiaceae
Eubacteriaceae
Peptococcaceae
Syntrophomonadaceae
Heliobacteriaceae
Clostridiales Family XVIII. Incertae Se
Ruminococcaceae
Halanaerobiaceae
Thermoanaerobacteraceae
Thermoanaerobacterales Family III. Ir
Erysipelotrichaceae
Thermaceae
Deinococcaceae

n
l., 2006

Multiple Comparison

High-level comparison:

- ✓ Select taxa
- ✓ Compute ecological indices (distances)
- ✓ Represent distances using neighbor-net
- ✓ … or MDS



Mitra, Gilbert, Field and Huson, ISME J, 2010

Neighbor-net: Bryant and Moulton, 2003

IFR Institute of Food Research

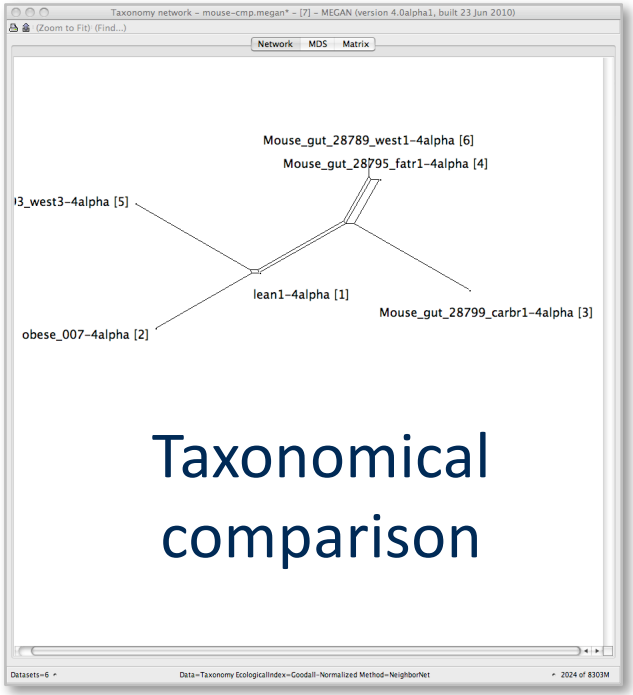
@Mitra2016

UNIVERSITÄT TÜBINGEN

Multiple Comparison

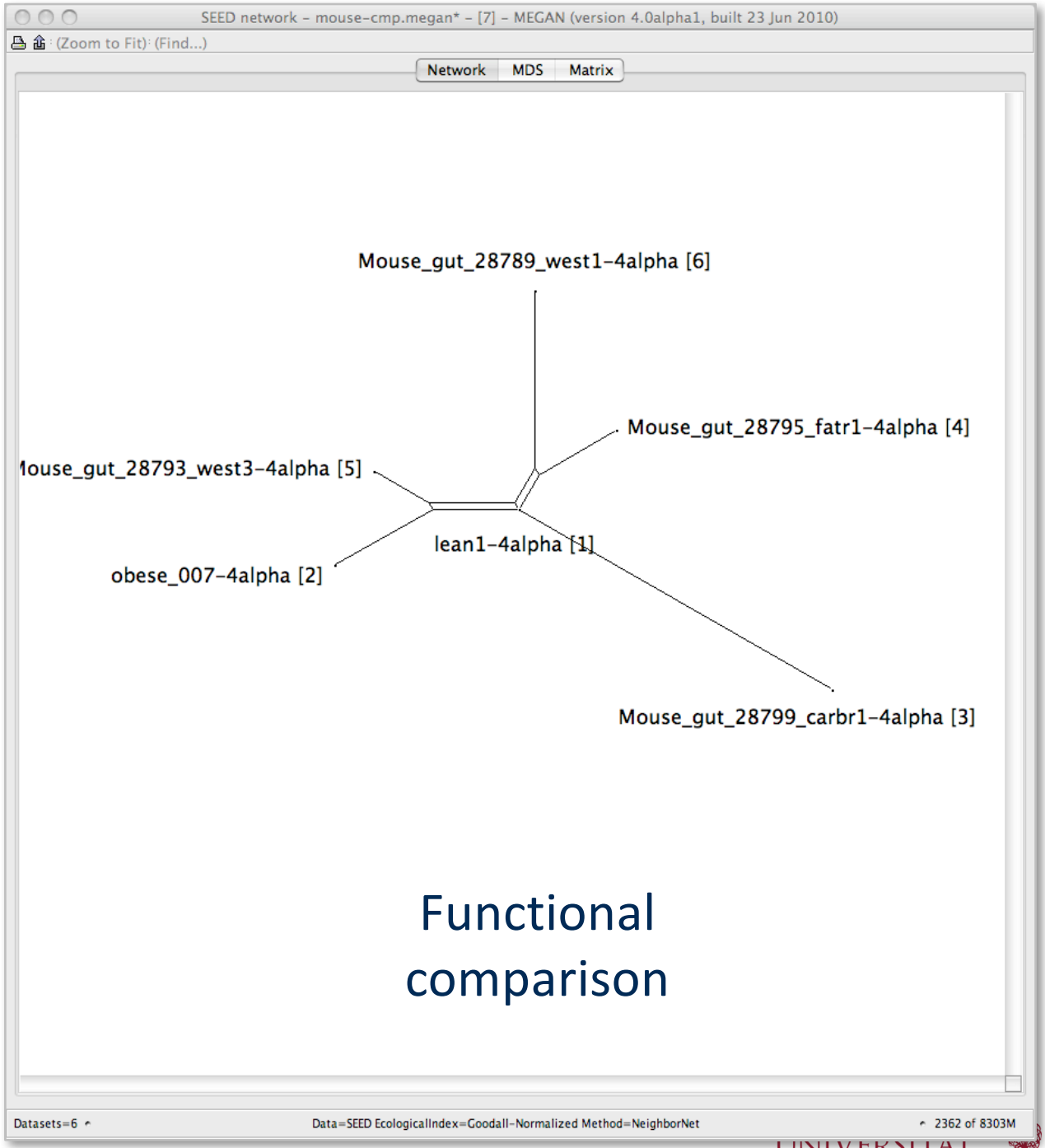
High-level comparison:

✓ **Select taxa**



Taxonomical comparison

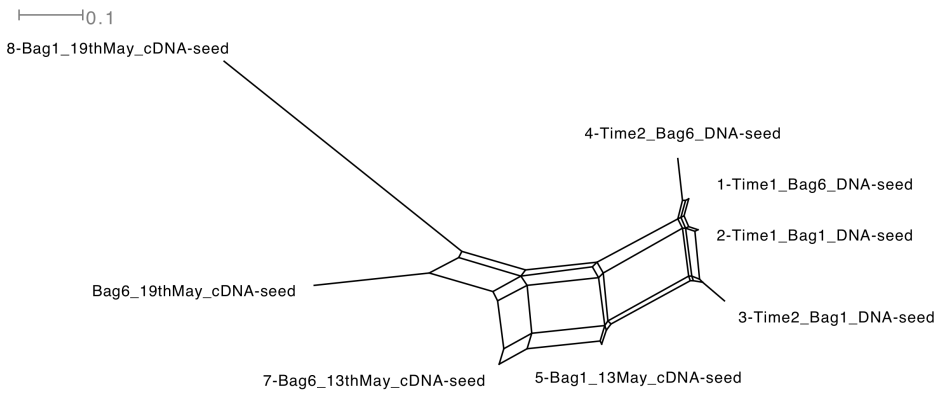
✓ **Select functions…**



Functional comparison

Network Visualization

- Metagenomes do not evolve along a tree
- Numerous environmental factors may result in incompatible signals
- Neighbor-net method is used to compute an unrooted network between multiple datasets

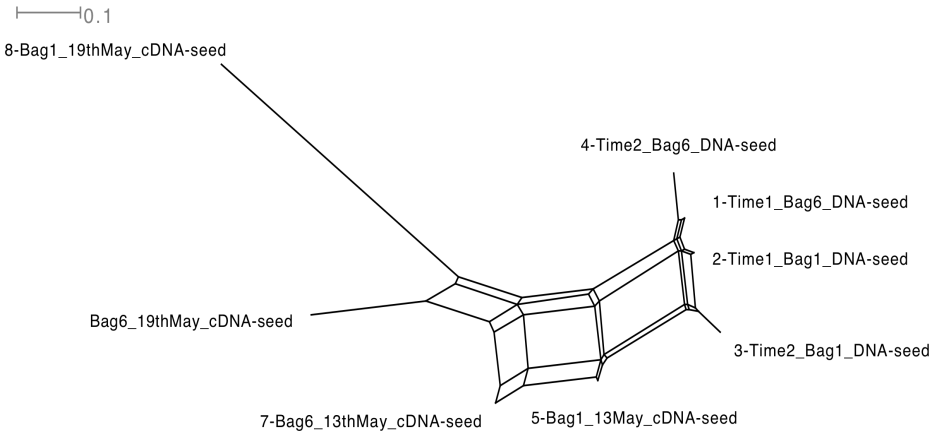


Mitra et al., 2010

Computational Approach

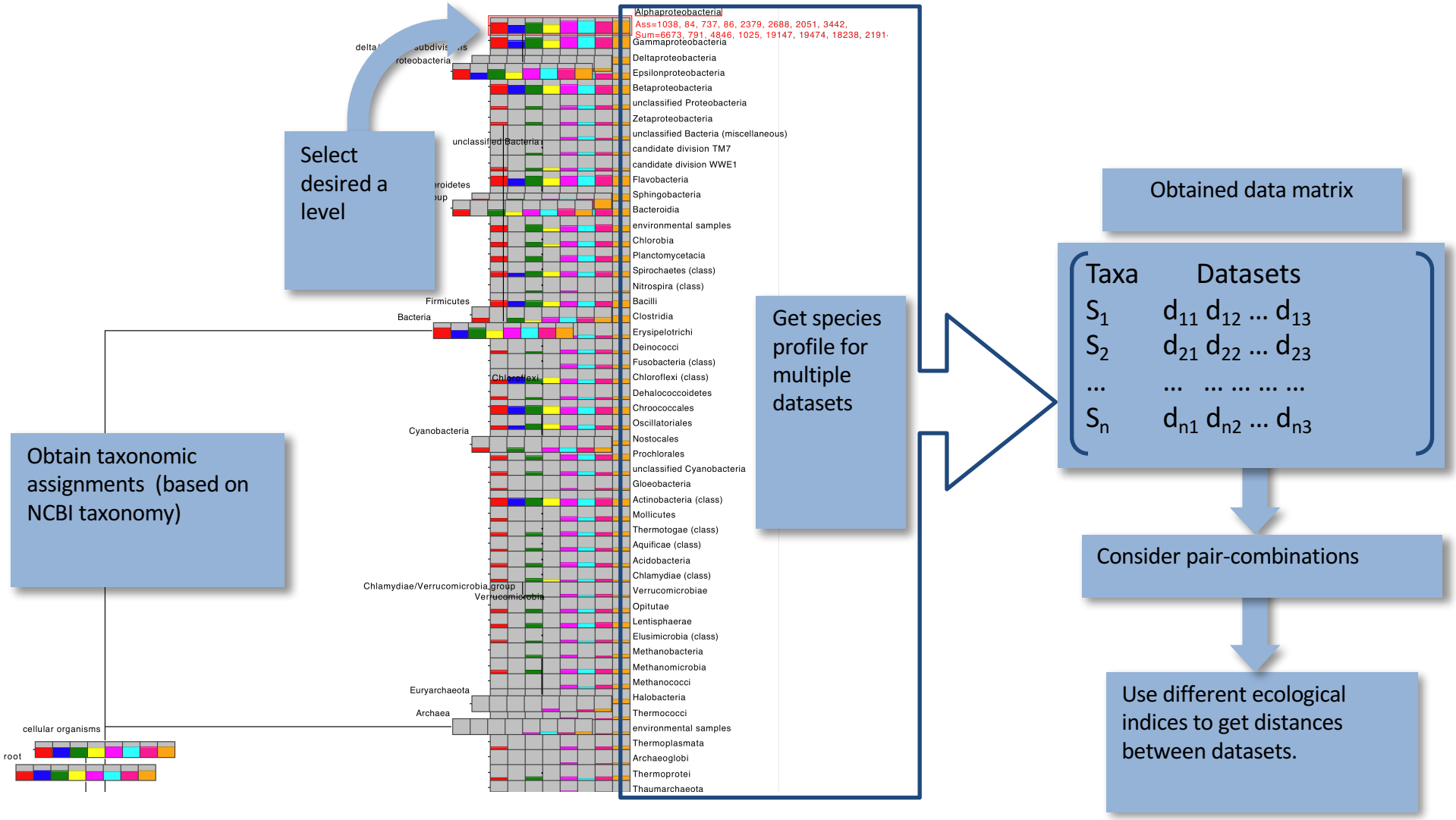
- Data matrix obtained from taxonomic or functional profile
- Distances obtained from pair combinations using different ecological indices

$$\mathcal{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & \dots & \dots & d_{1n} \\ \vdots & d_{22} & & & & \vdots \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & & & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & \dots & \dots & d_{nn} \end{pmatrix}.$$



Mitra et al., 2010

Computational Approach



Based on NCBI taxonomy

Reads are assigned to gene content

Performance

- **Euclidean distance**
- **Kulczynski distance**
- **Bray-Curtis distance**
- **Chi-squared distance**
- **Hellinger distance**
- **Goodall's index**

Dominated by the large values

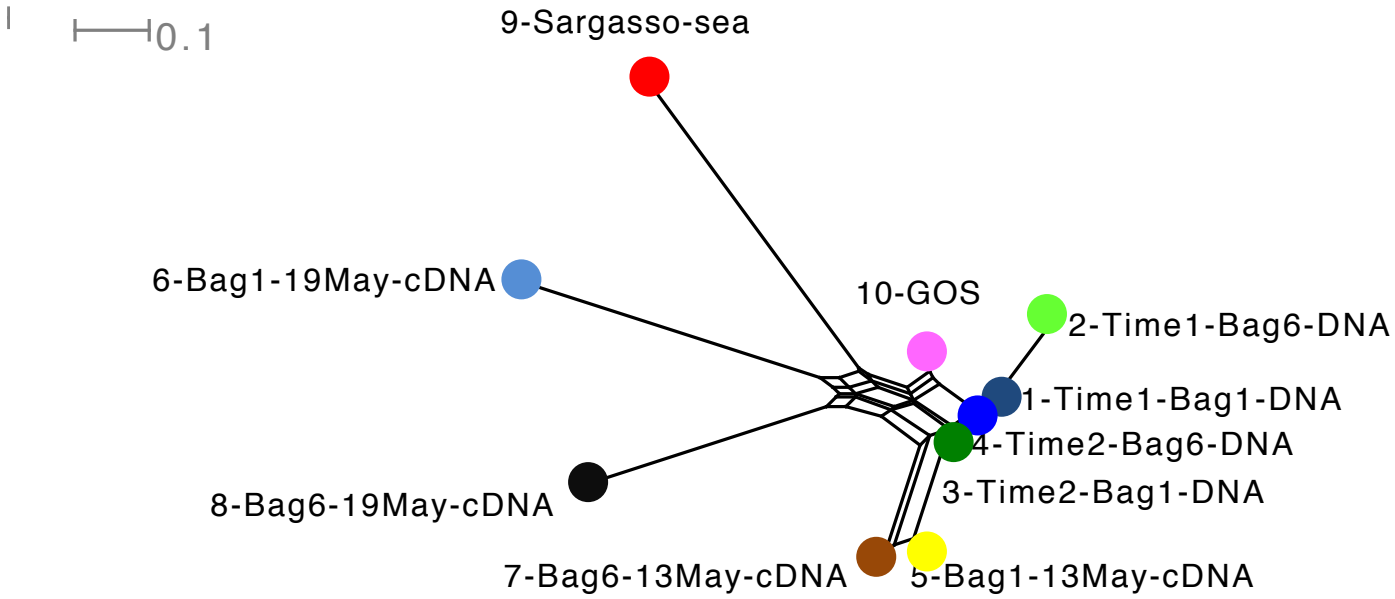
Based on differences in the proportions of values

Non-parametric measure for multivariate datasets
(effective for the study of rare-taxa)

Robustness

Comparison of 10 marine datasets

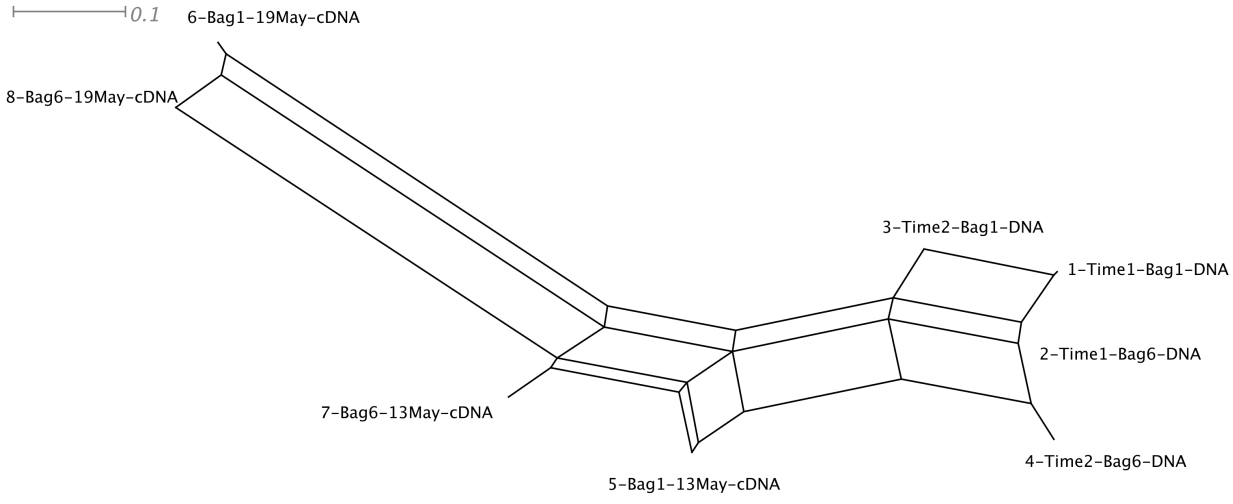
Similarity analysis at different levels of taxonomy using ecological indices



Top level | Phylum | Class | Order | Family | Genus | Species

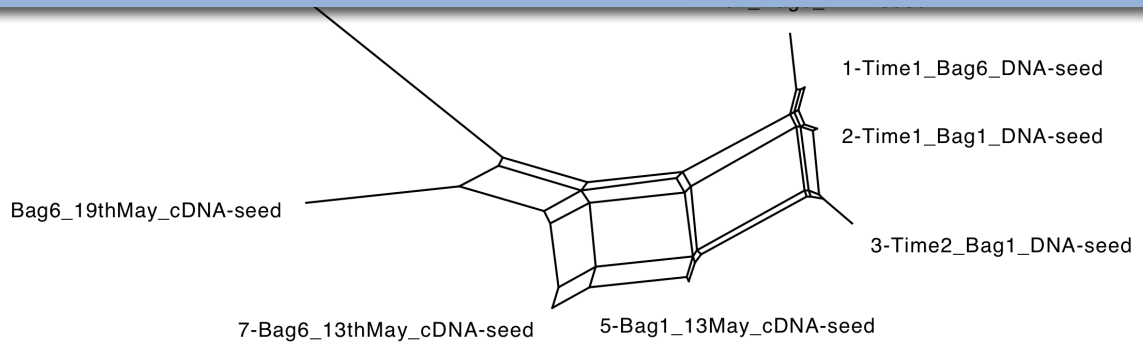
Robustness

Comparison of eight marine metagenome



Network obtained from taxonomic profile

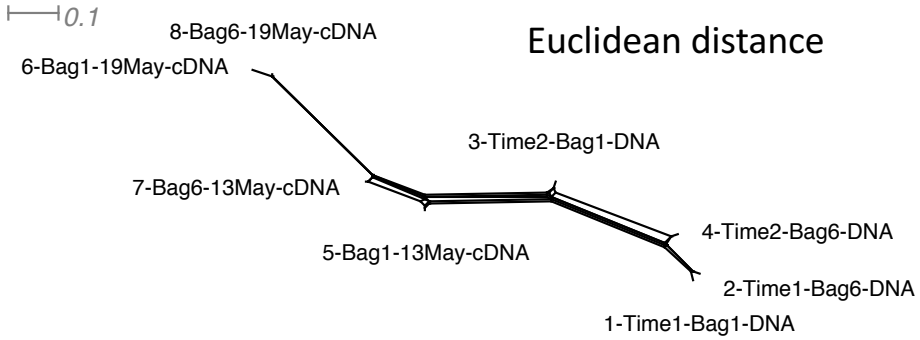
Clusters are quite similar for both taxonomic and functional content…

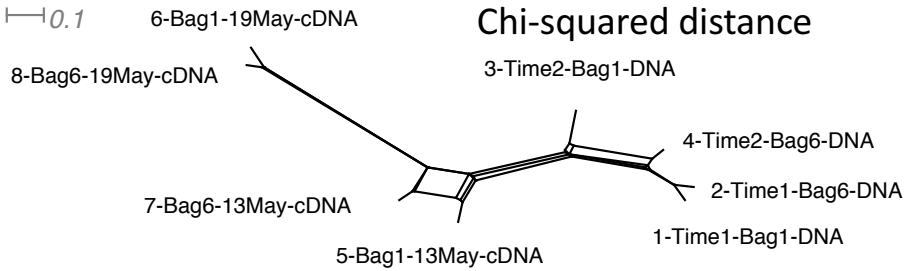


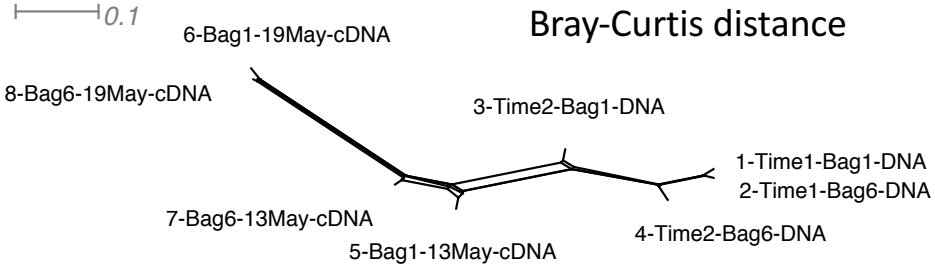
Network obtained from functional profile

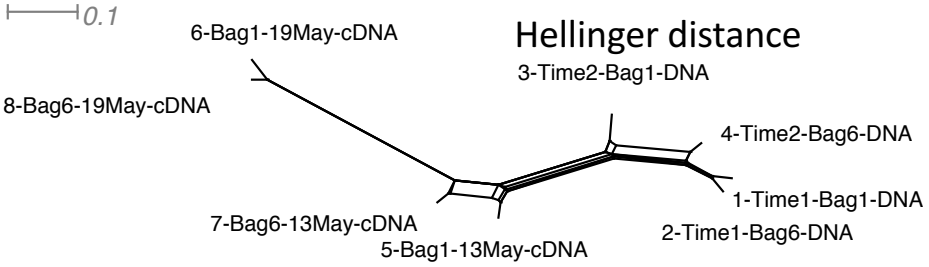
Performance

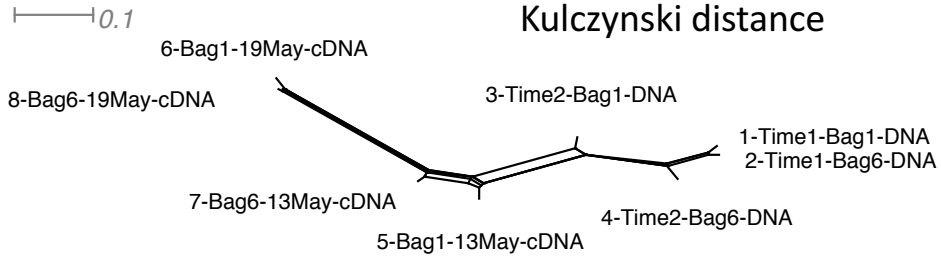
Comparison of eight marine metagenomes using six ecological measures

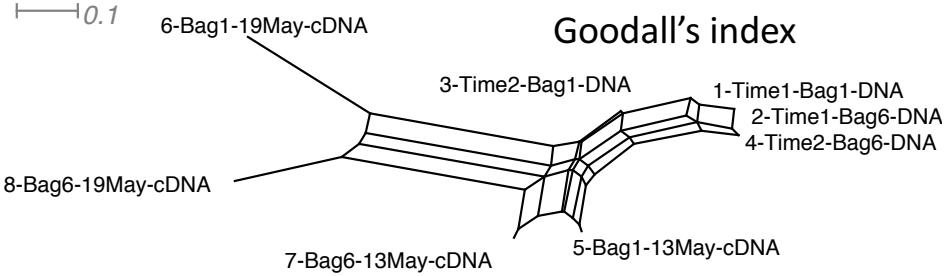










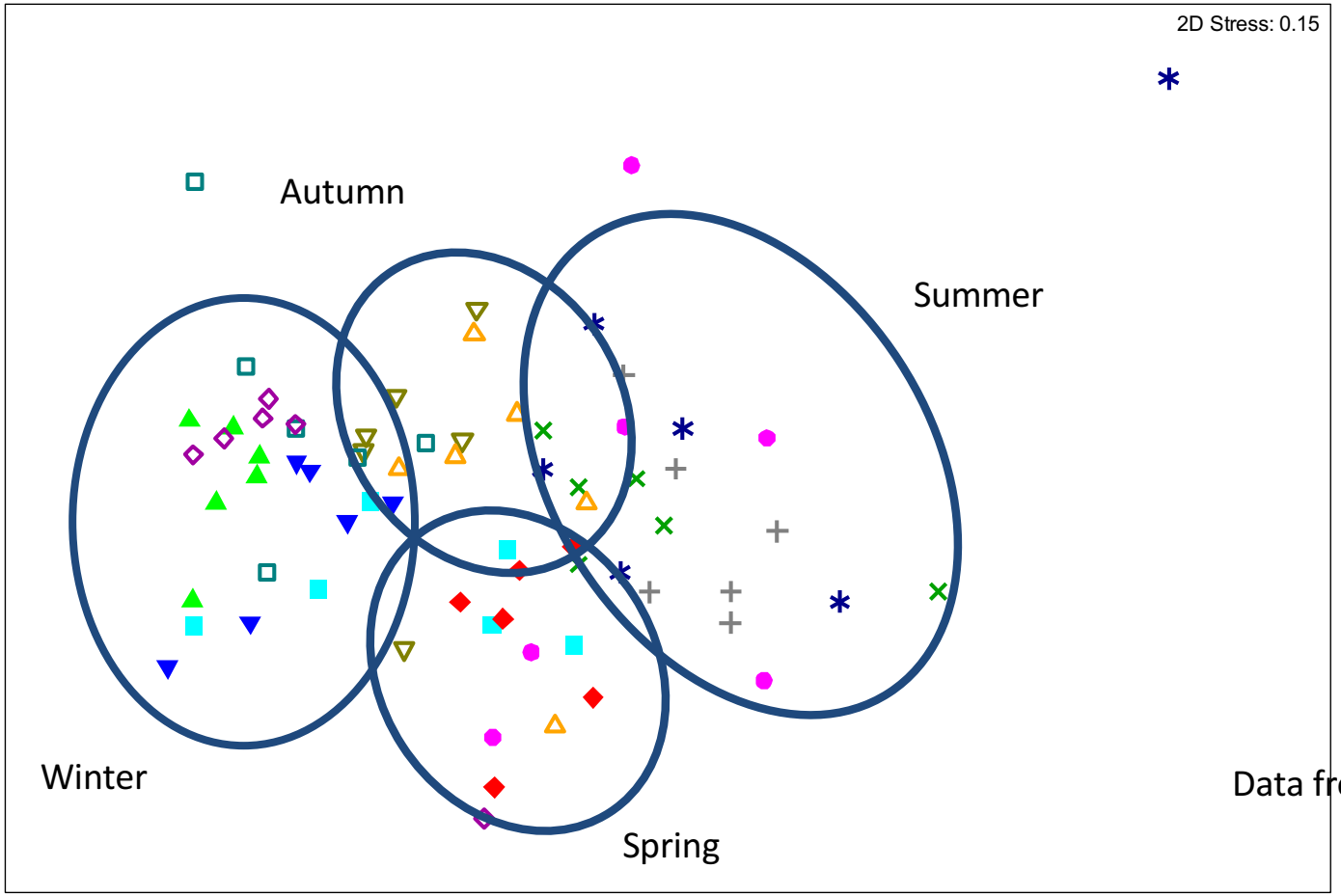


Correlation Study

Correlation between samples and different seasons

6 years (for 12 month) time series data from Western English Channel

NMDS plot

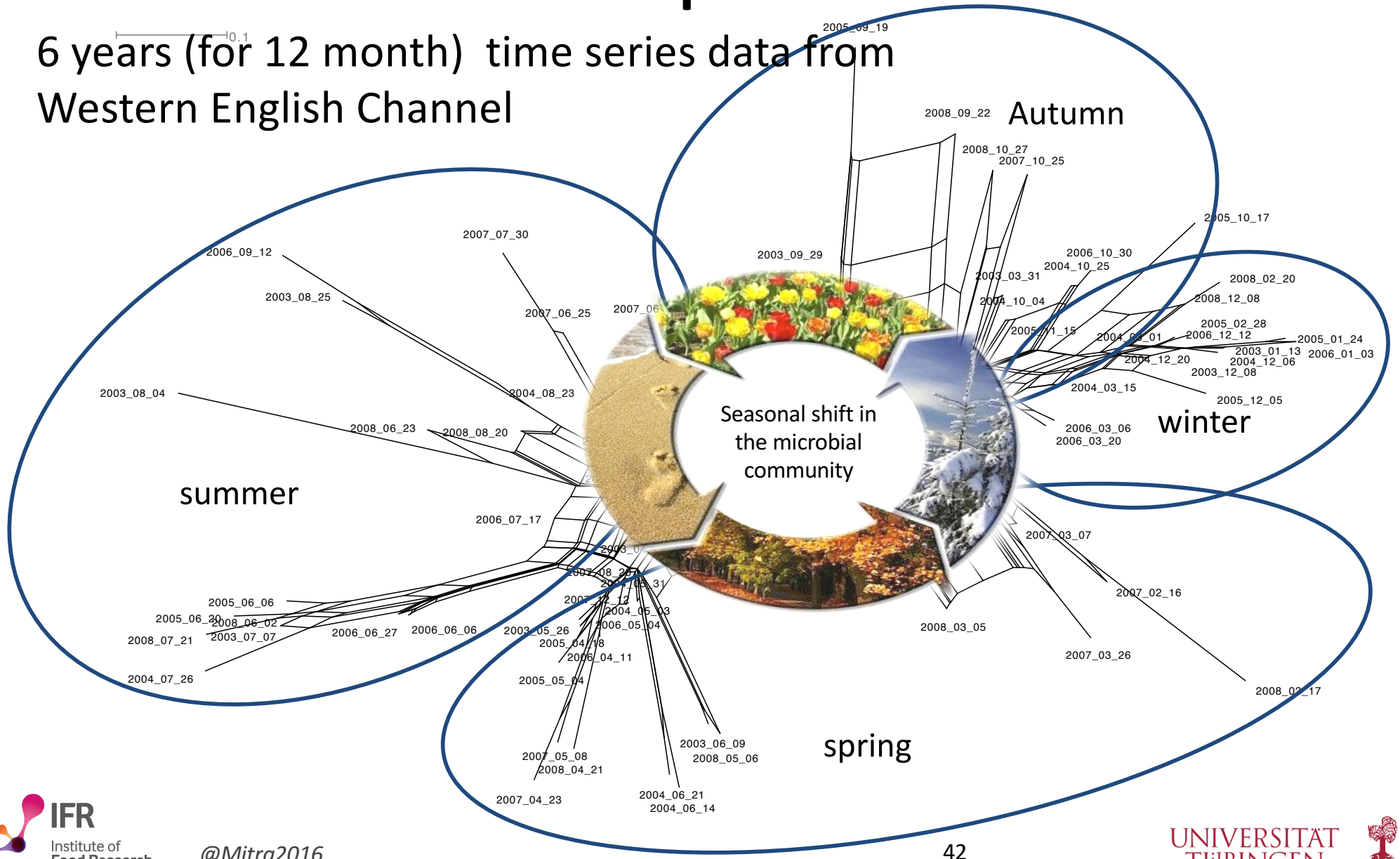


Data from Gilbert et al., 2009

Correlation Study

Correlation between samples and different seasons

6 years (for 12 month) time series data from Western English Channel



Bioinformatics and complexity

Bioinformatics

- Metagenomic shotgun data
  - One billion reads
  - – 15 samples

- Alignment against NCBI-NR protein reference database
  - BLASTX on single server: ~15 years (Altschul *et al*, 1990)
  - RAPSearch2 on single server: ~3.5 months (Zhao *et al*, 2012)
- This takes too long…

DIAMOND

- ✓ DIAMOND is a “drop in” replacement for BLASTX
- ✓ Much, much faster…
- ✓ Very similar sensitivity to BLASTX
- ✓ Software available for download here:

  http://ab.inf.uni-tuebingen.de/software/diamond

- ✓

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink[1], Chao Xie[2,3] &
Daniel H Huson[1,2]

NATURE METHODS 2015

IFR Institute of Food Research

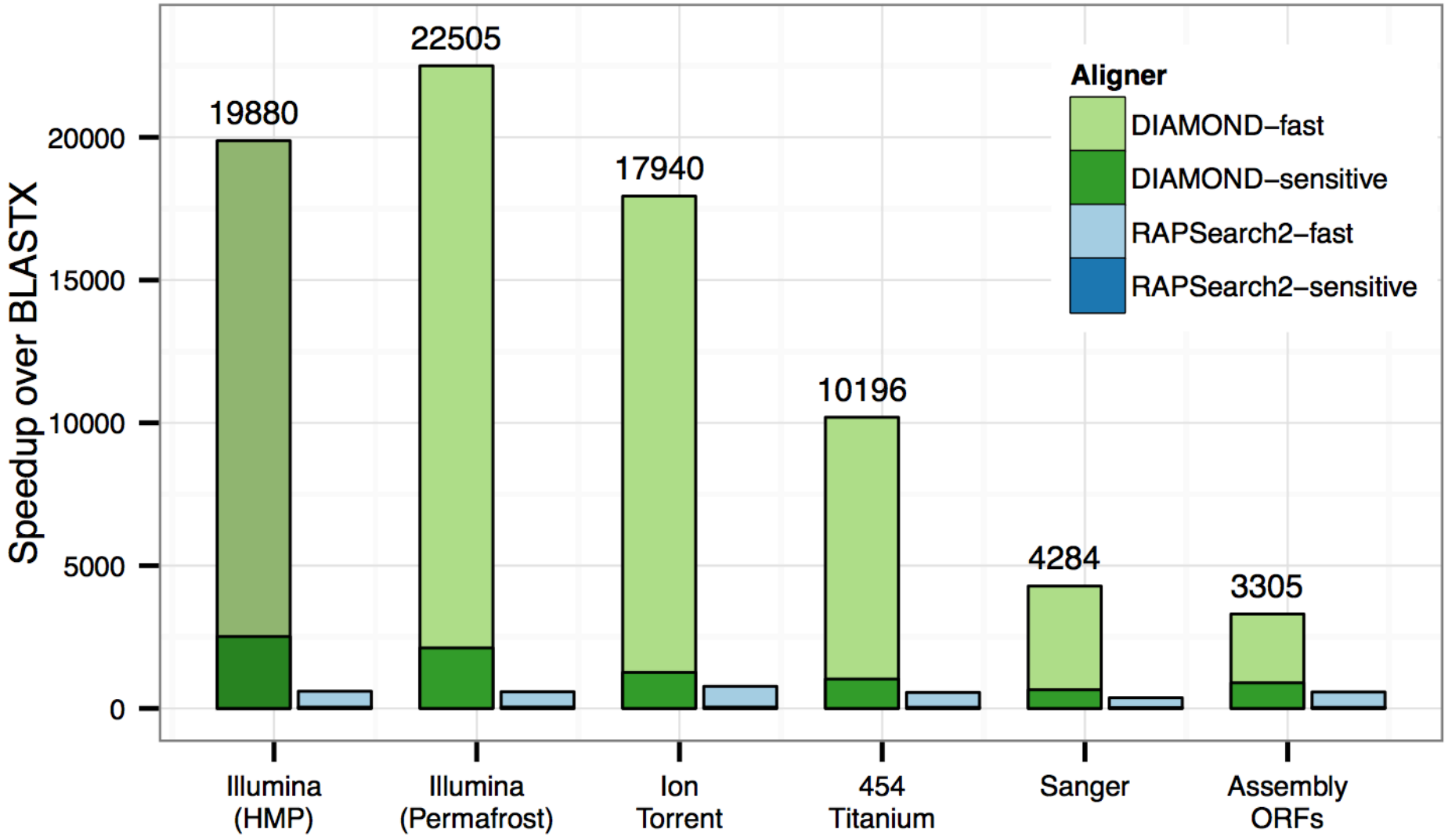
@Mitra2016

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Performance

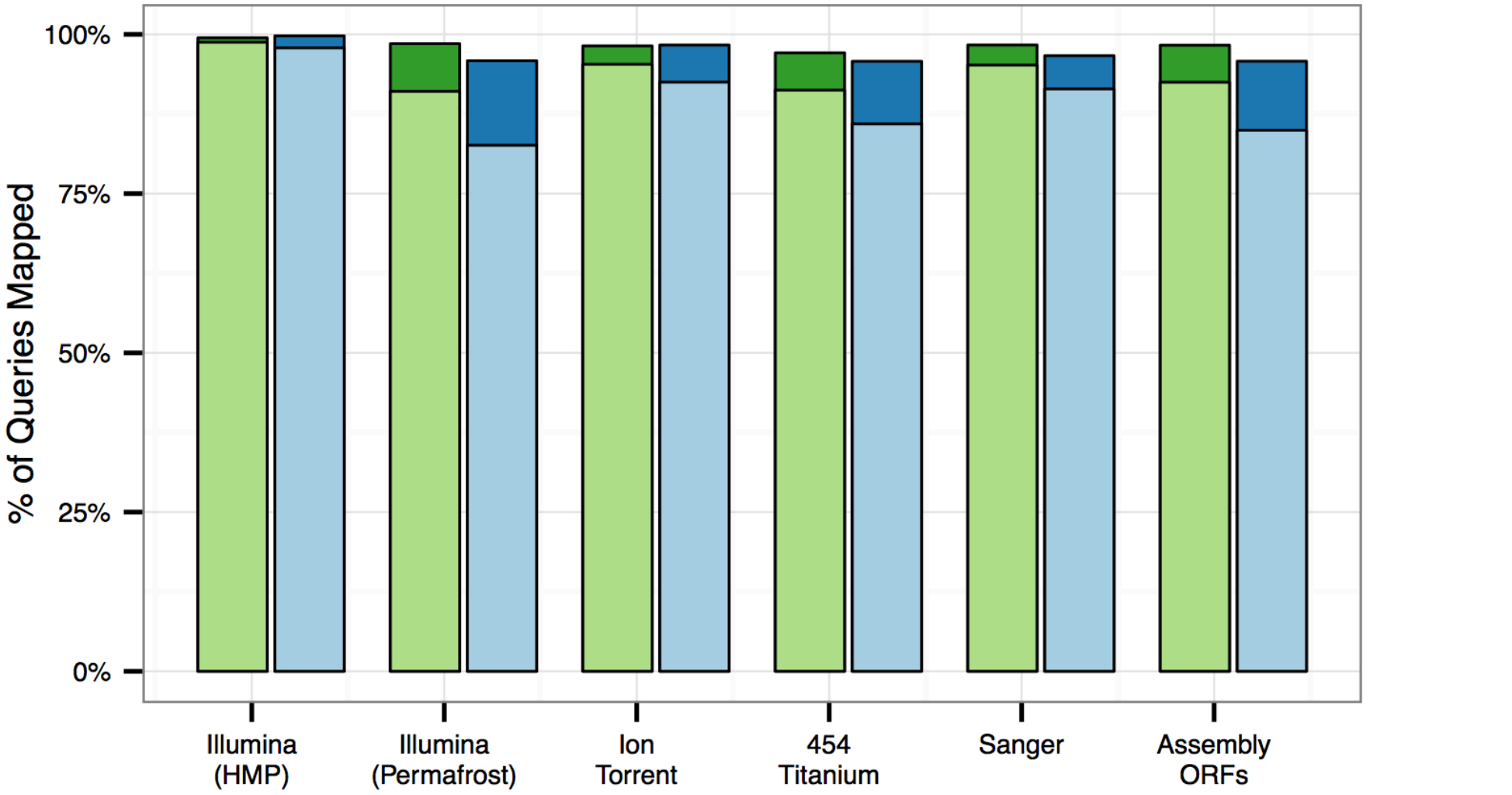
- DIAMOND is **20 000x** faster than BLASTX on Illumina HiSeq reads

DIAMOND - Speedup



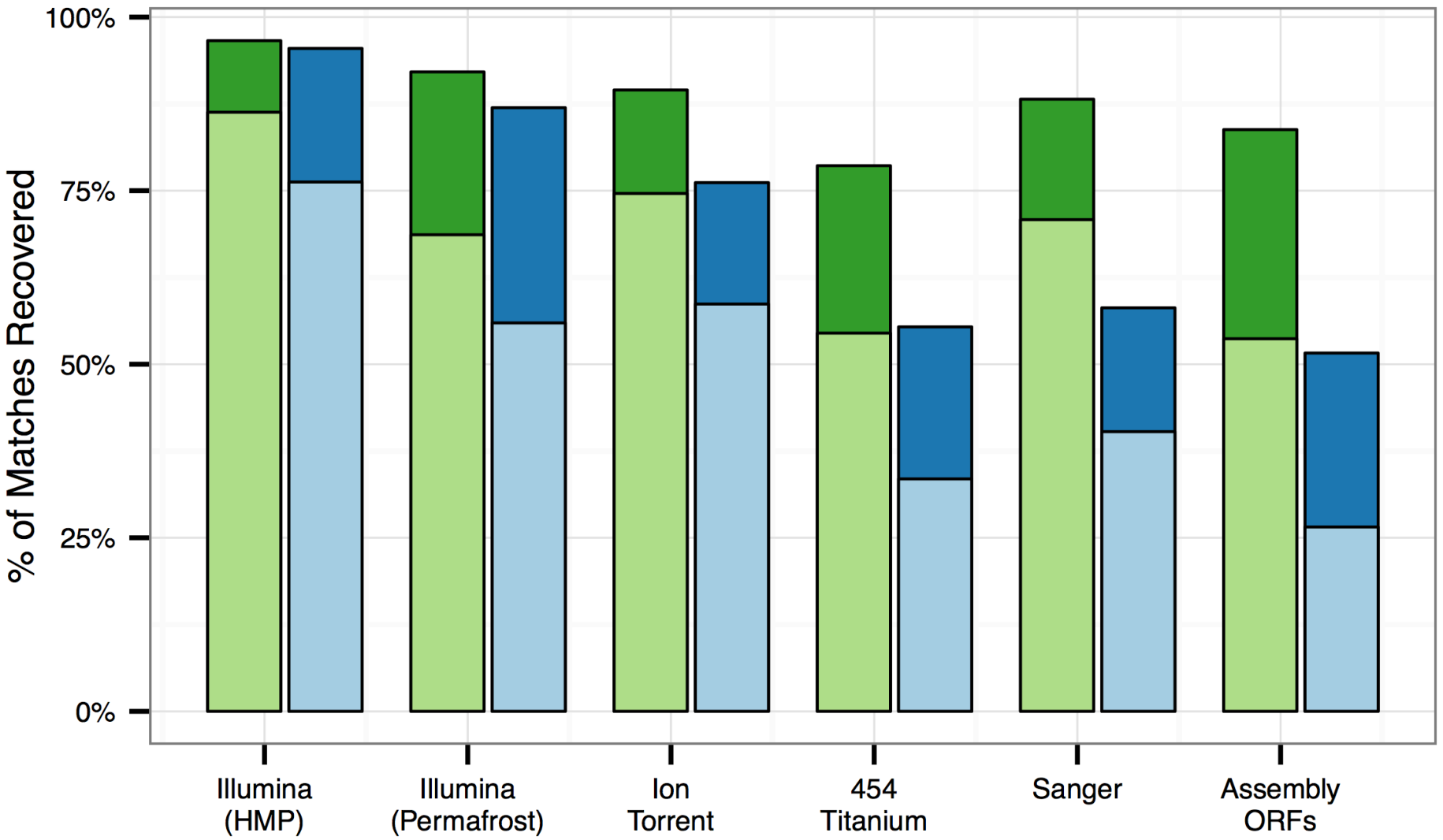
DIAMOND – Sensitivity

Percent of reads with a BLASTX match that also have a DIAMOND or RAPSearch2 match (e-value≤10^{-3})



DIAMOND – Sensitivity 2

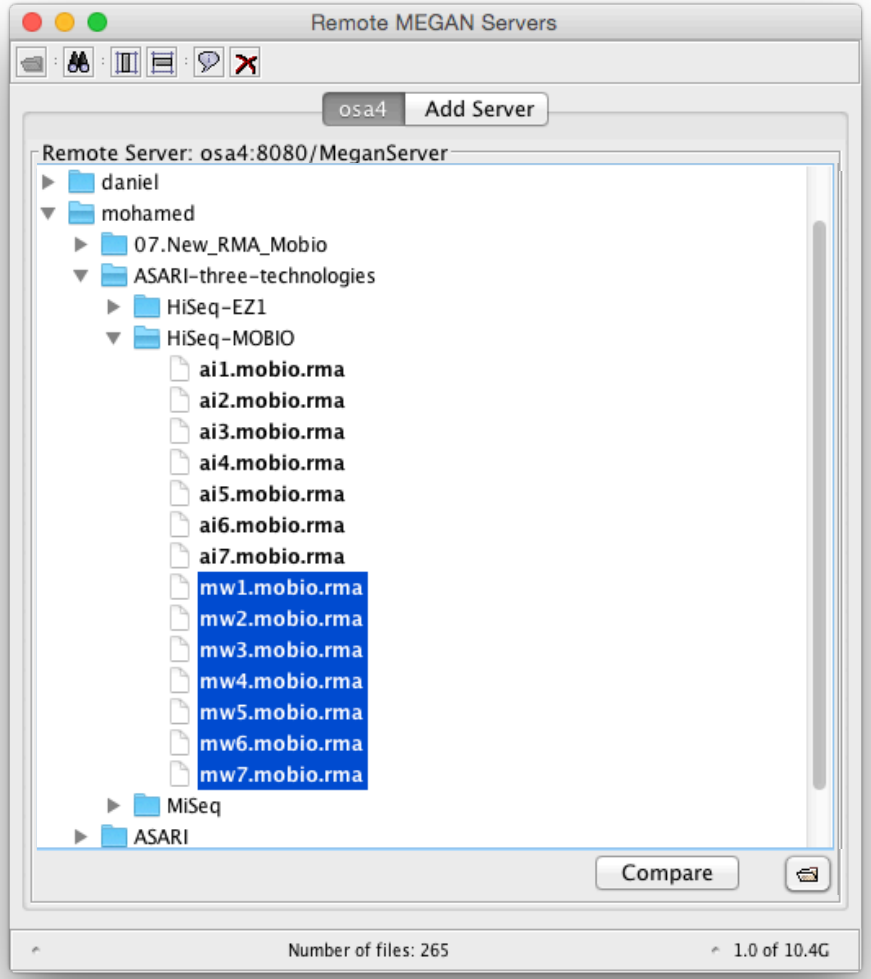
Percentage of BLASTX matches recovered
(≤250 matches per query, e-value ≤10^{-3})



Large Numbers of Large Samples

- Typical project has 100s of samples, each 10s of GBs
- How to work with such data?
- MeganServer software allows remote access from within MEGAN6

(Hans Ruscheweyh)



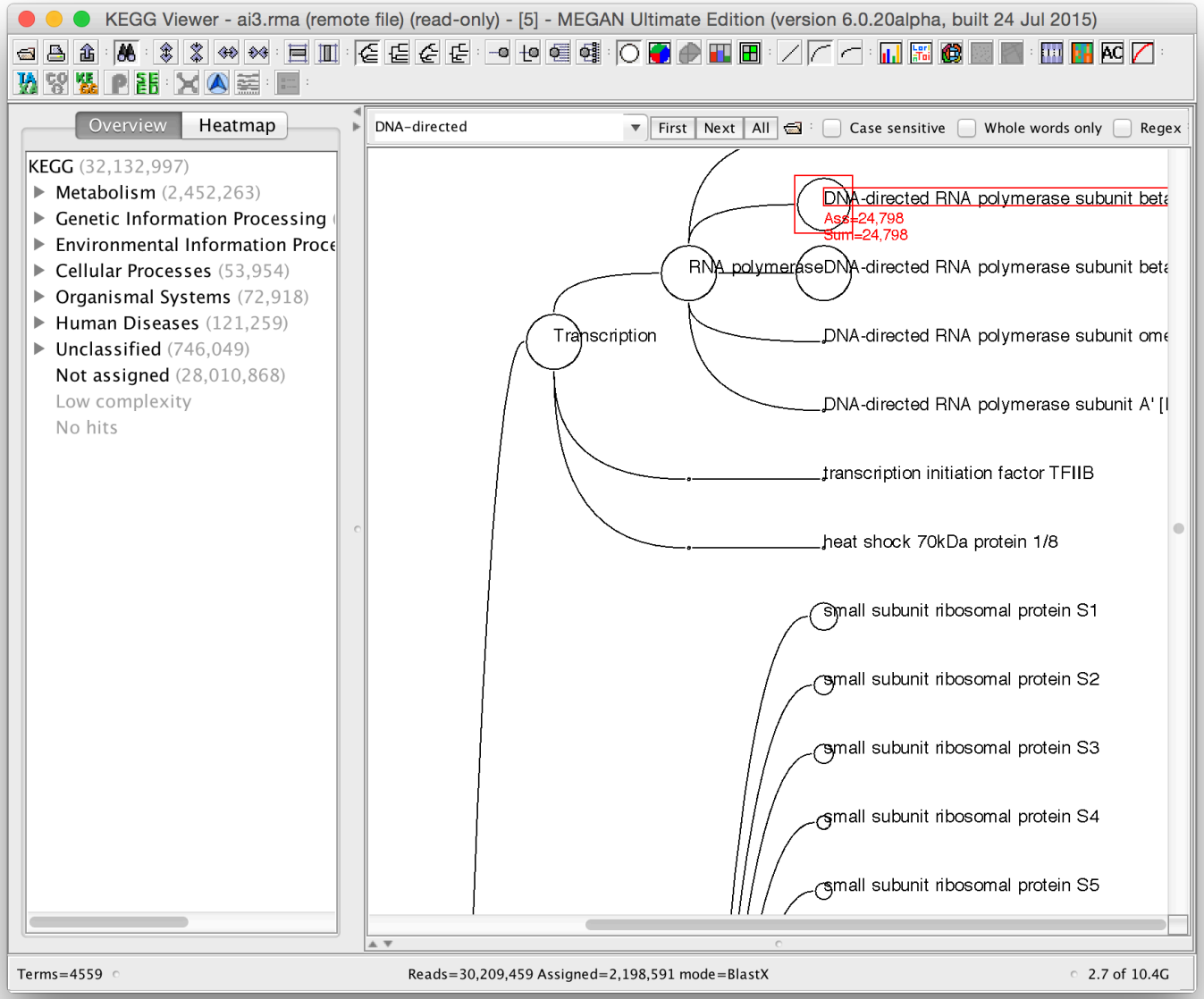
Metagenome Assembly

- Assembly of multiple genomes from metagenome data is very difficult
- Data reduction: first assemble, then BLAST
- **Now:** first (DIAMOND-)BLAST, then assemble

Gene-Centric Assembly

- Assemble gene-by-gene…
- Use DIAMOND to align reads to protein sequences
- Assemble all reads that align to a given gene
- Do this on-the-fly in MEGAN

Gene-Centric Assembly, rpoB



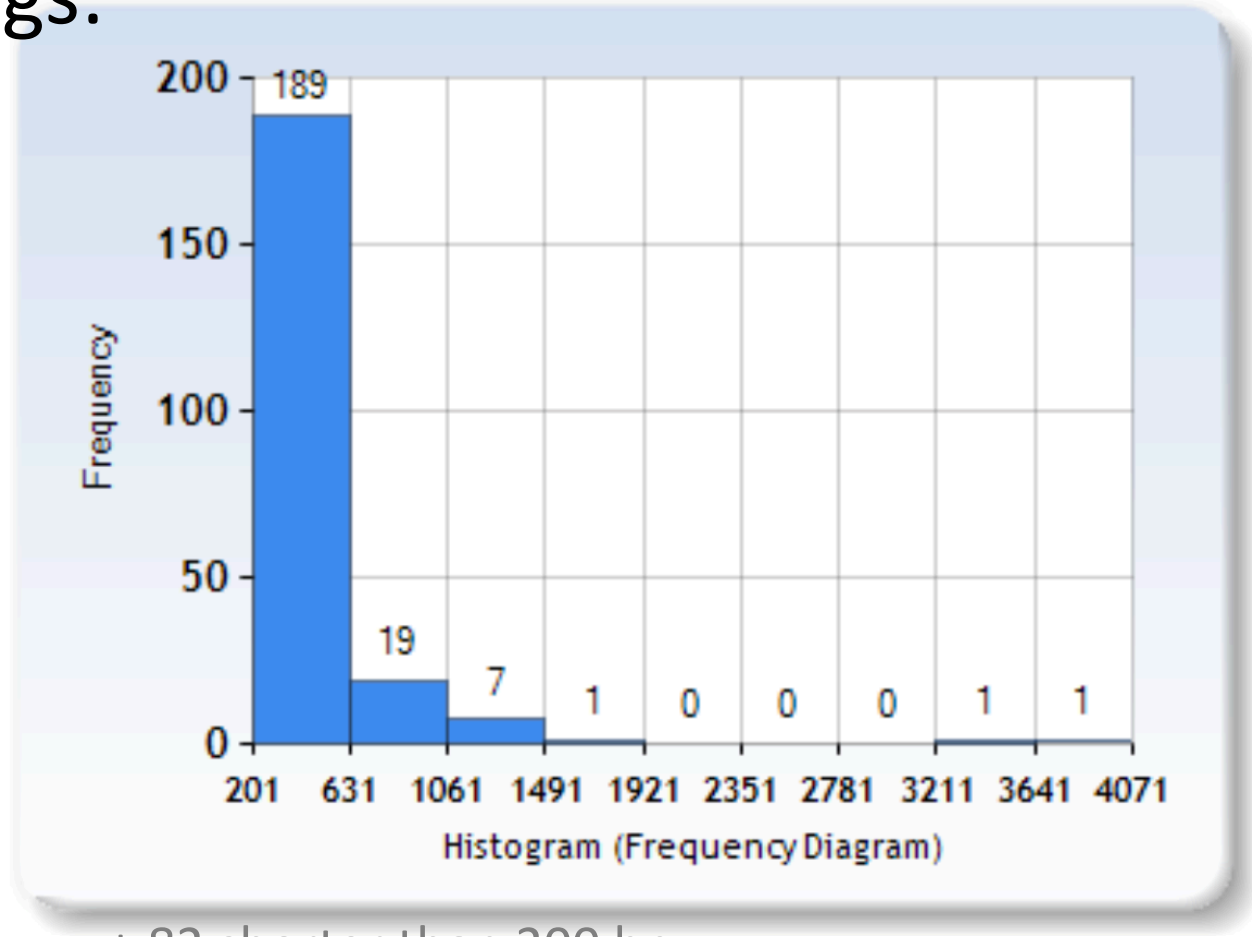
Human gut sample, 30 mio reads, 24,798 assigned to rpoB

Gene-Centric Assembly, rpoB

Human gut sample, 30 mio reads, 24,798 assigned to rpoB

Gene-Centric Assembly, rpoB

- Gene-centric assembly of 24,798 reads, contigs:



+ 82 shorter than 200 bp

Human gut sample, 30 mio reads, 24,798 assigned to rpoB

Will allow you to:

- **Organize**
- **Visualize**
- **Interact with**
- **Summarize**
- **Capture**
- **Compare**

your metagenome data

http://ab.inf.uni-tuebingen.de/software/megan6/

Thank You ☺

Acknowledgements

- **Tübingen, Germany**
  - **Prof. Daniel Huson**
  - **Hans-Joachim Ruscheweyh**
  - **Max Schubach**
  - **Rewati Tapra**
  - **Lindsay Hall**
  - **Cristina Alcon**
- **SCELSE, Singapore**
  - **Prof. Stephan Schuster**
  - **Dr. Rohan Williams**
- **Karlsruhe, Germany**
  - **Dr. Bernhard Klar**
- **Chicago, USA**
  - **Dr. Jack Gilbert**