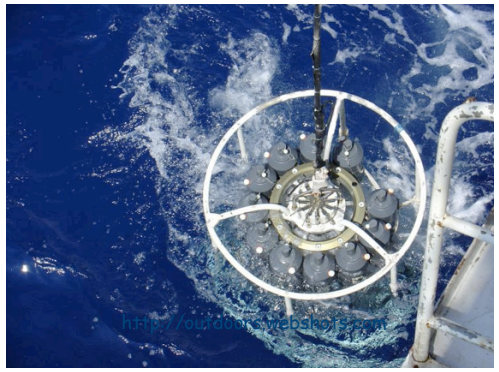
# Computational Analysis of Metagenomes

Daniel H. Huson



Eberhard Karls Universität Tübingen

Eberhard Karls Universität Tübingen

ZBIT

# Contents

- Genomics
- Sequencing
- Metagenomics
- Computational questions
- Outlook

# Contents

- Genomics
- Sequencing
- Metagenomics
- Computational questions
- Outlook

# Discovery of DNA

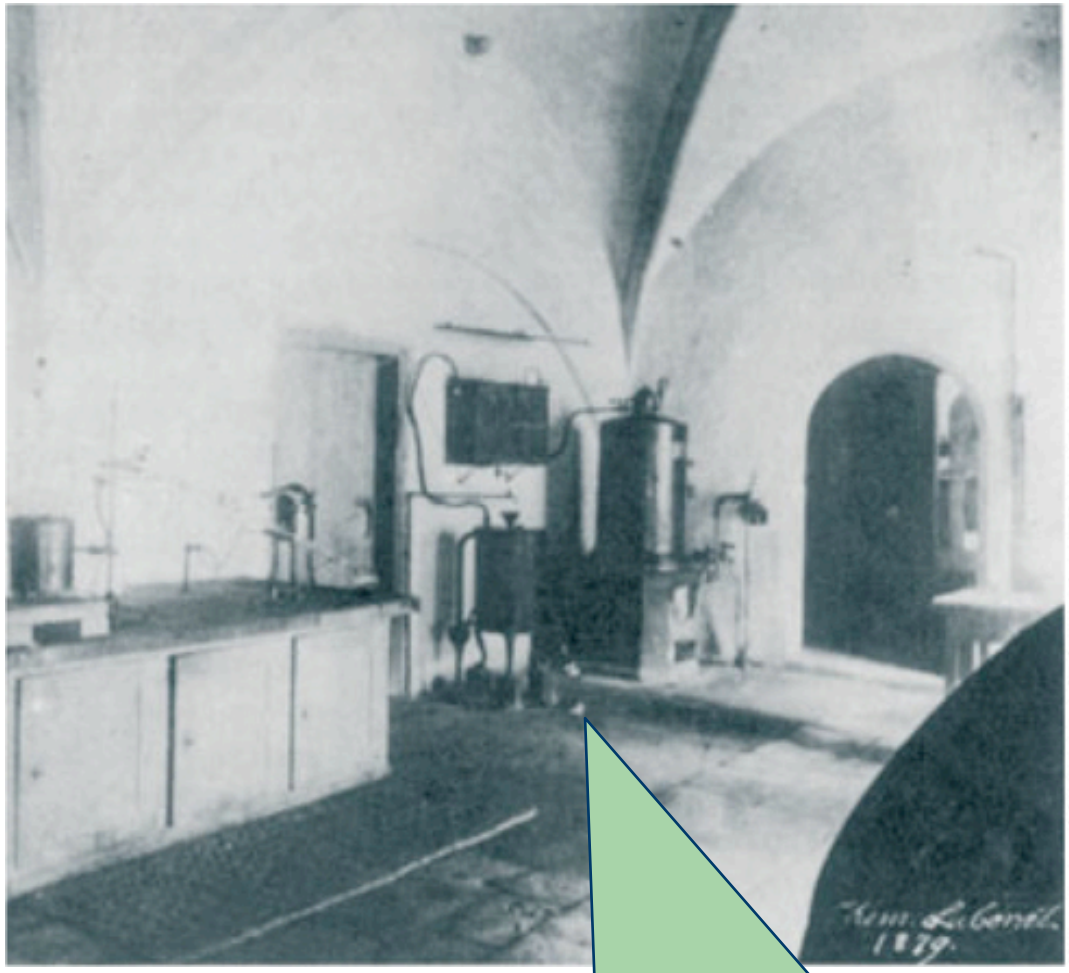
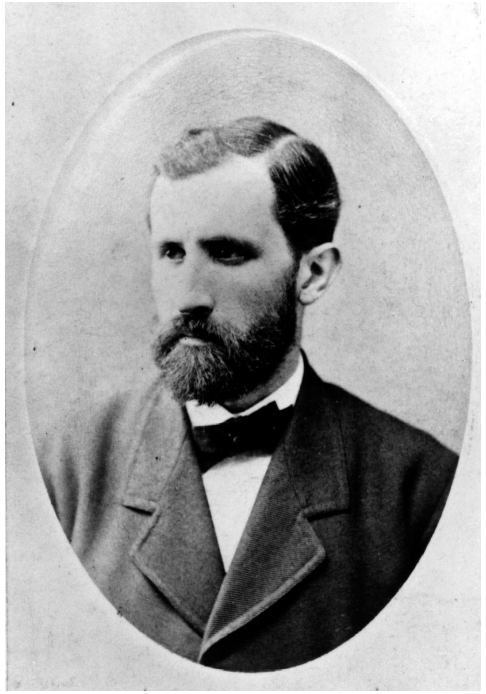


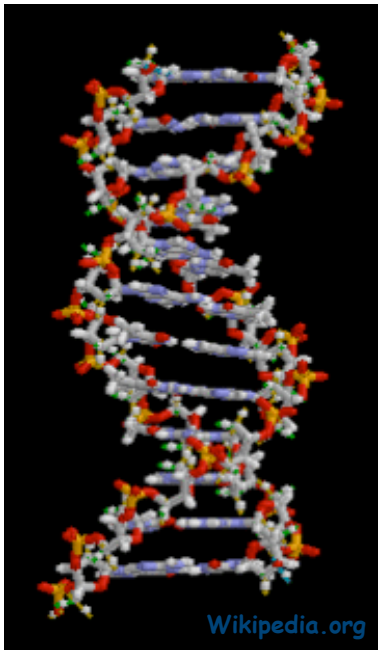
Photo courtesy of the University of Tübingen Library, Tübingen, Germany

1869: Miescher discovered DNA in the kitchen of Tübingen Castle



Friedrich Miescher
(1844-1895)

# Role of DNA

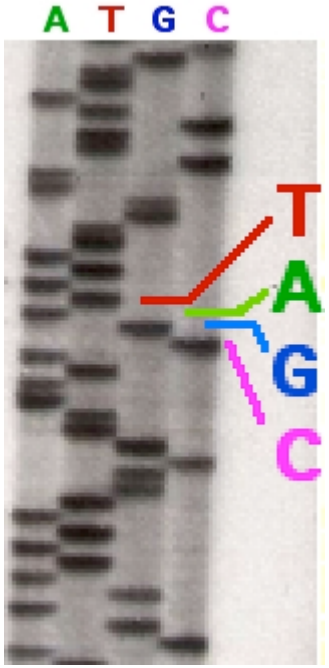



...
A - T
C - G
G - C
T - A
A - T
A - T
...

1953 Watson and Crick
- The structure of DNA is a double helix
- It is *the order of the bases* along the molecule that contains heredity information

# Sanger DNA Sequencing

1975 Frederick Sanger develops the "chain termination method" method for DNA sequencing



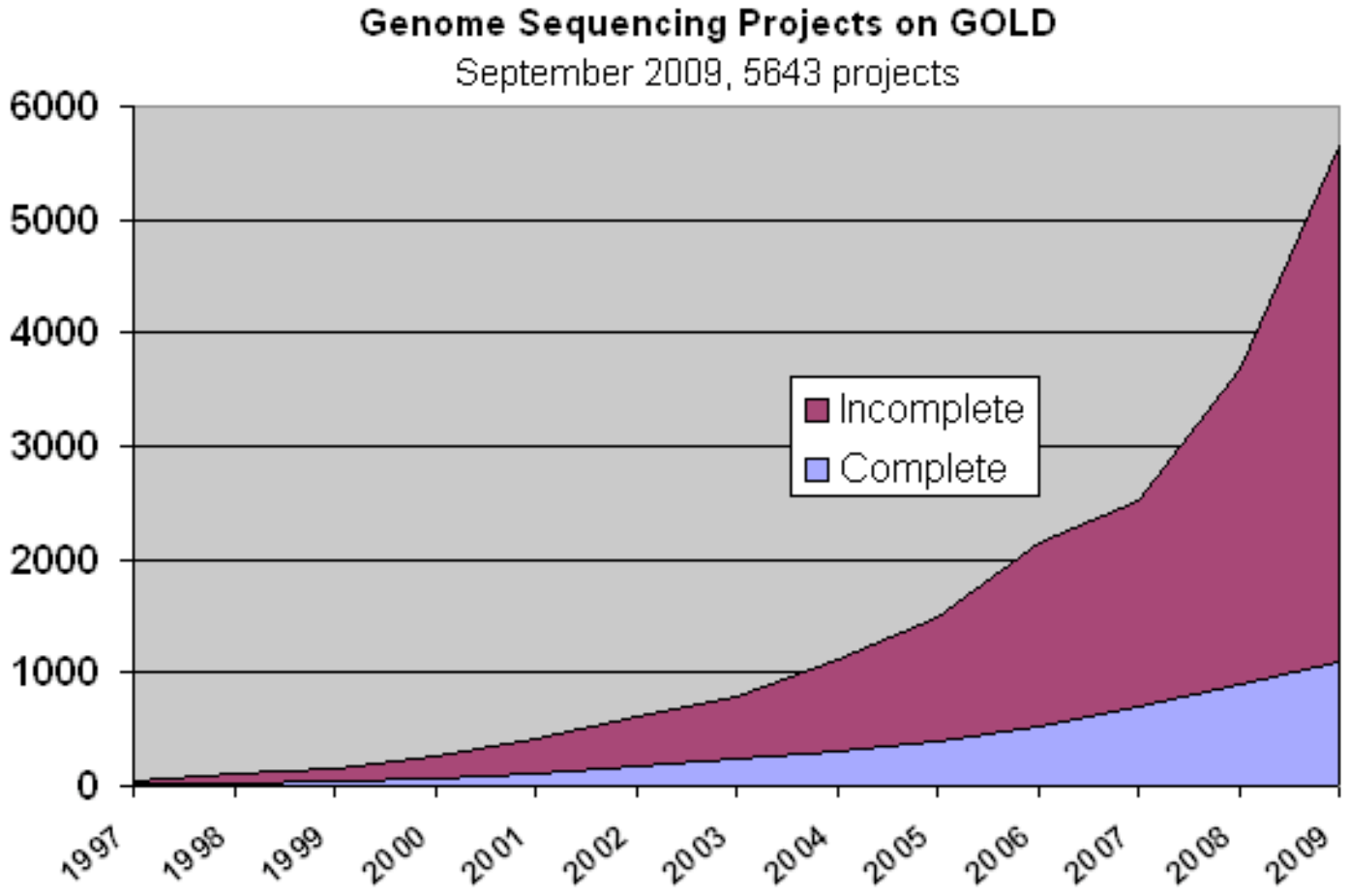

- Sanger sequencing basis of Genomics until 2005

# Genomics

- Genomics is the study of the genome sequence of individual organisms

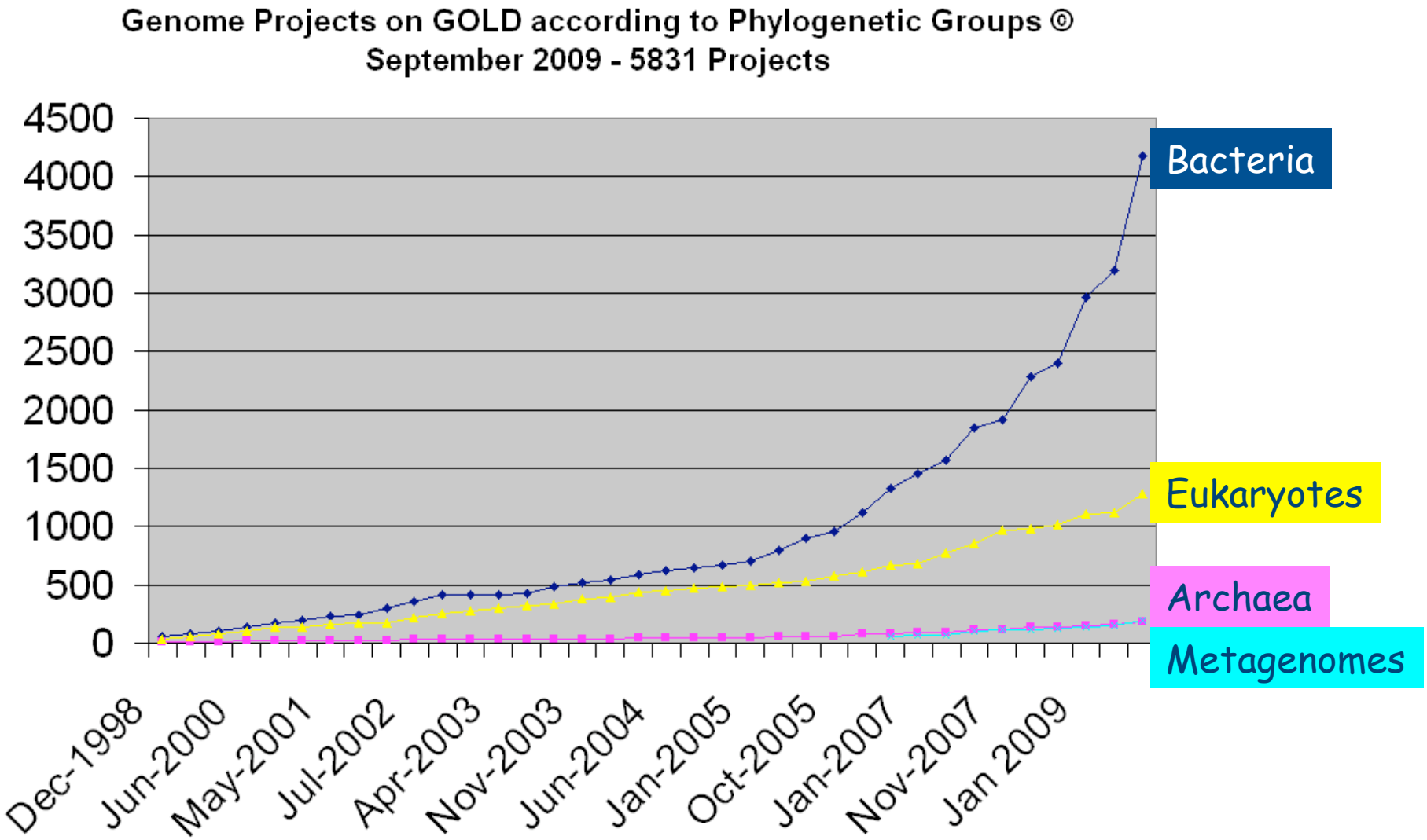
- Genome sizes:
  - Bacteria: 1-10 million bases (Mb)
  - Drosophila: 140Mb
  - Human: 3 billion bases (Gb)

# Sequencing of Genomes



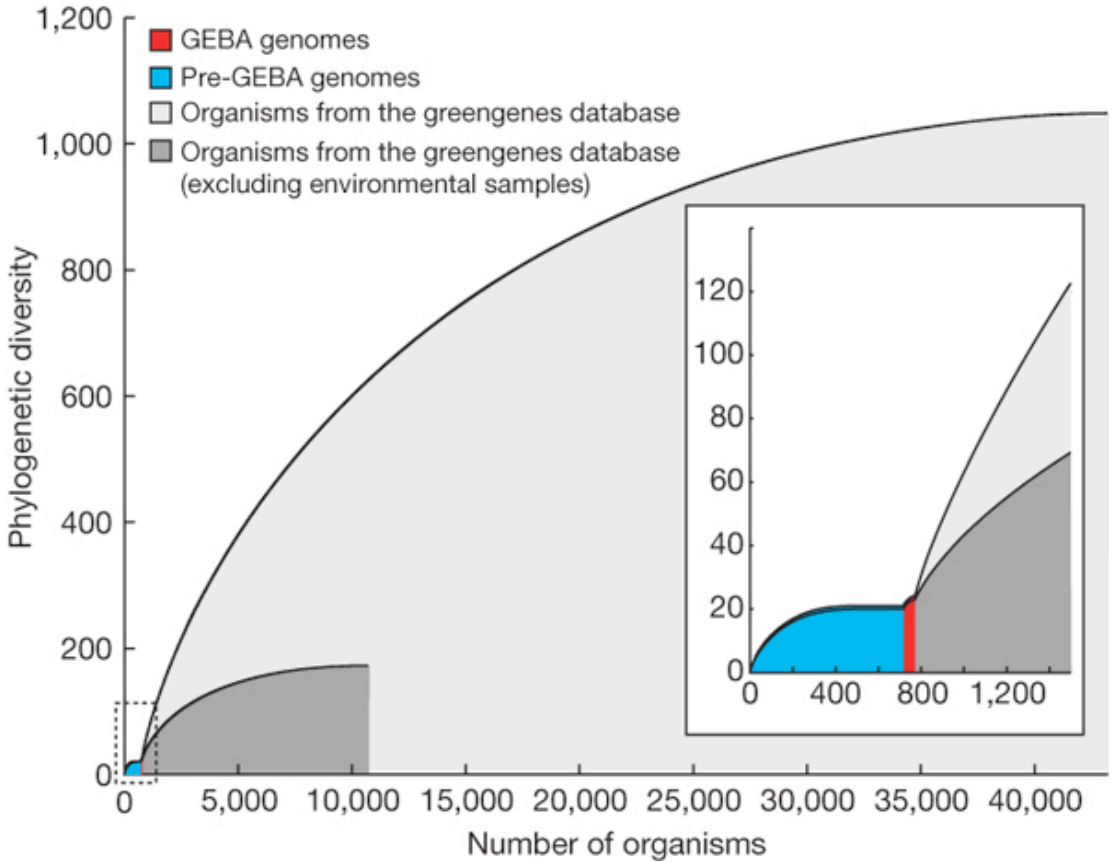

**GOLD: Genomes online database**
**www.genomesonline.org**

# Genome Projects by Groups



# The GEBA Project

- A Genomic Encyclopedia for Bacteria and Archaea
  - JGI/DSMZ project
  - Systematically sequence microbes from underrepresented clades




Dongying Wu *et al*, Nature, 2009

http://www.jgi.doe.gov/programs/GEBA/

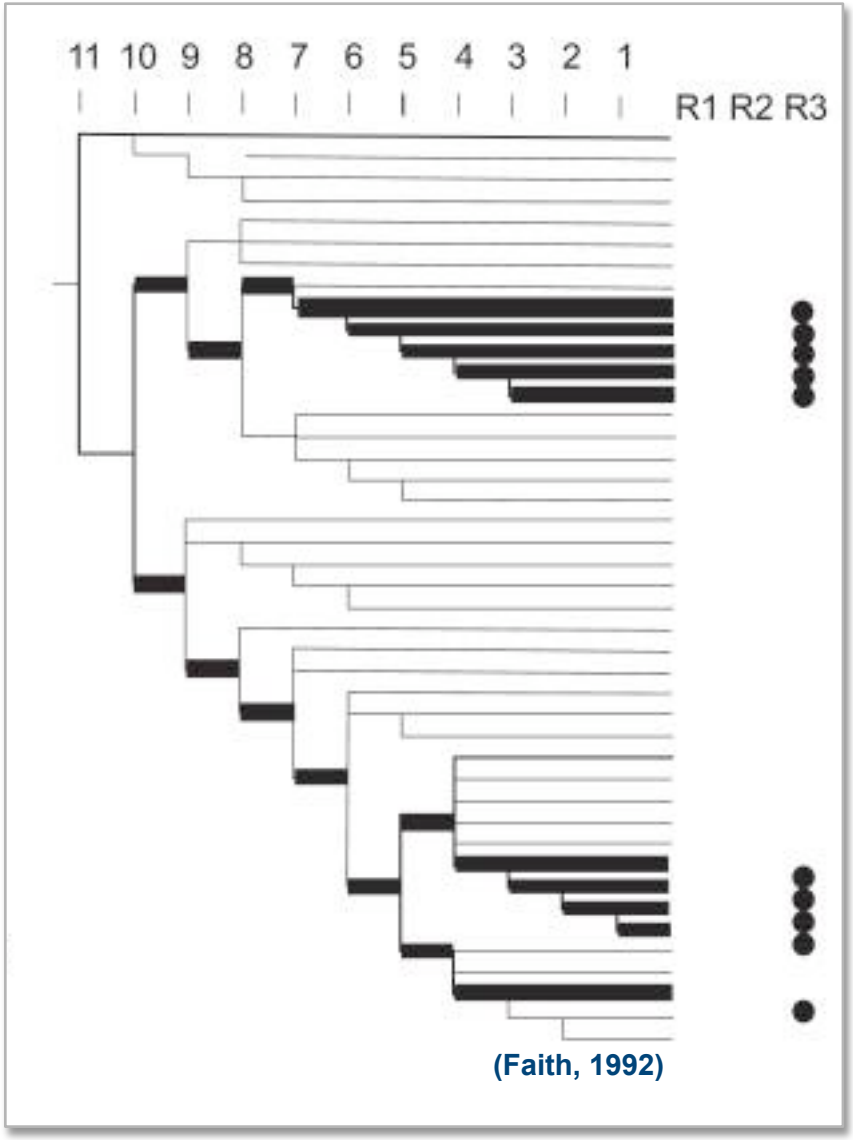
green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

# Phylogenetic Diversity (PD)

PD of a set of taxa:

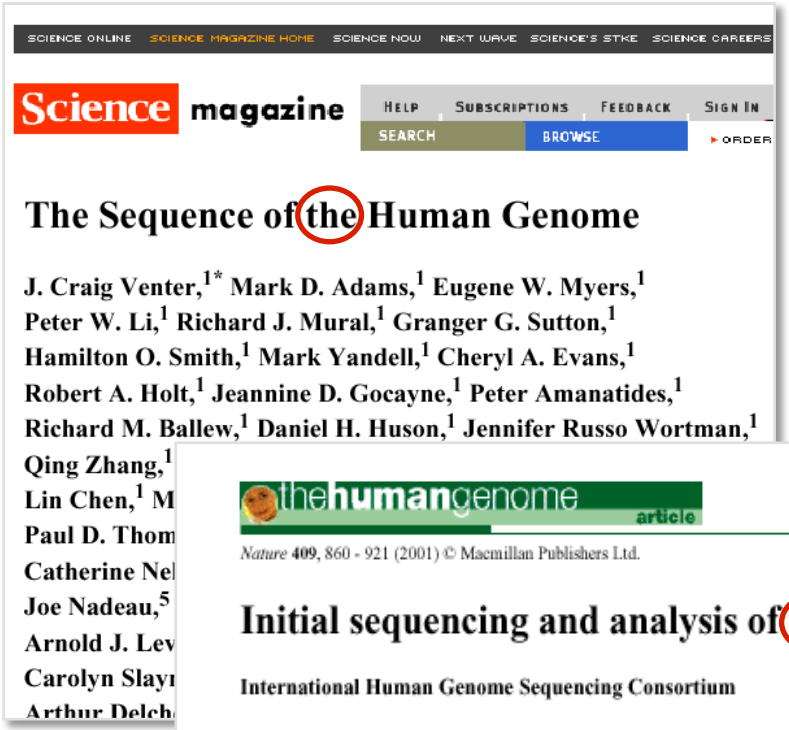
- The sum of all branches on the phylogenetic tree that spans the set

(Faith, 1992)



(Faith, 1992)

# From One to Many…


SCIENCE ONLINE SCIENCE MAGAZINE HOME SCIENCE NOW NEXT WAVE SCIENCE'S STKE SCIENCE CAREERS

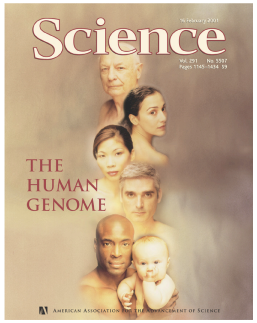
Science magazine

HELP SUBSCRIPTIONS FEEDBACK SIGN IN

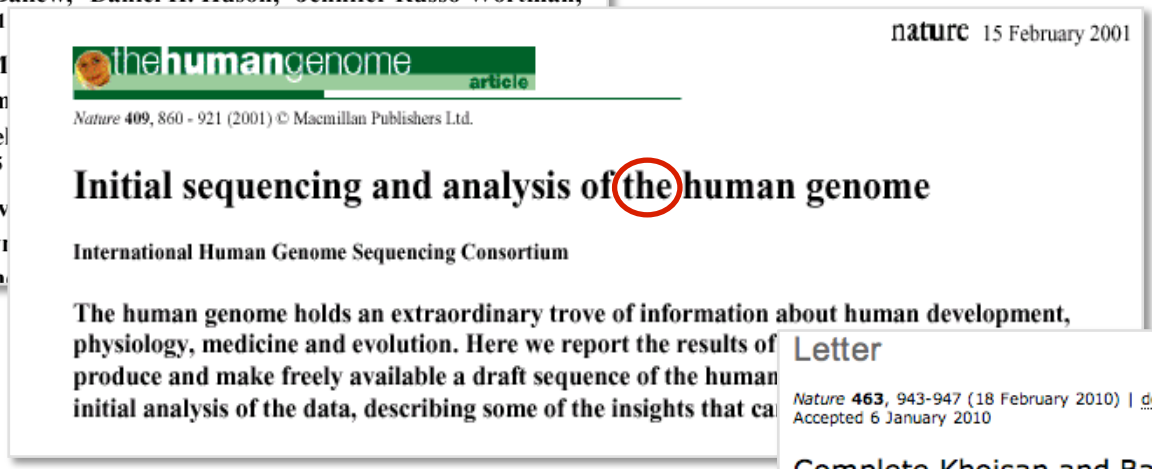
SEARCH BROWSE ORDER

The Sequence of the Human Genome

J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Lin Chen,[1] M… Paul D. Thom… Catherine Nel… Joe Nadeau,[5] Arnold J. Lev… Carolyn Slayı… Arthur Delch…




## 2001: THE Human Genome


thehumangenome article

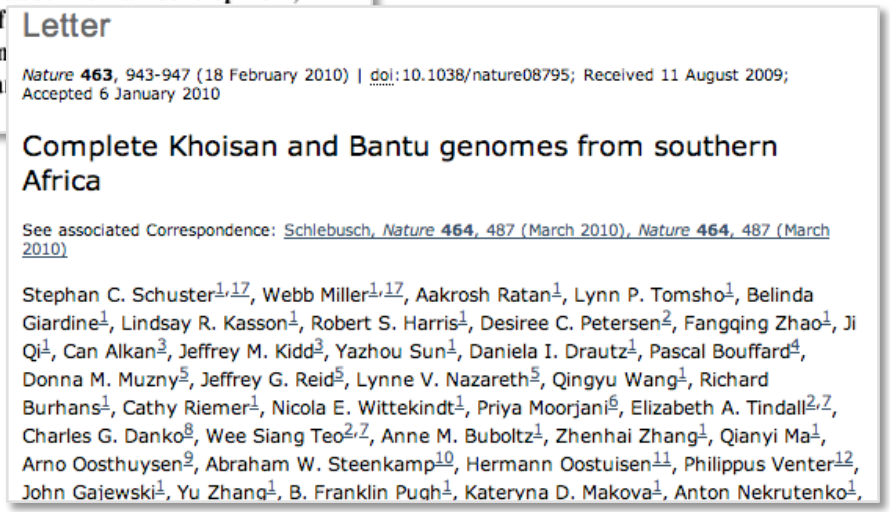
nature 15 February 2001

*Nature* **409**, 860 - 921 (2001) © Macmillan Publishers Ltd.

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of… produce and make freely available a draft sequence of the human… initial analysis of the data, describing some of the insights that ca…



Letter

*Nature* **463**, 943-947 (18 February 2010) | doi:10.1038/nature08795; Received 11 August 2009; Accepted 6 January 2010

Complete Khoisan and Bantu genomes from southern Africa

See associated Correspondence: Schlebusch, *Nature* **464**, 487 (March 2010), *Nature* **464**, 487 (March 2010)

Stephan C. Schuster[1,17], Webb Miller[1,17], Aakrosh Ratan[1], Lynn P. Tomsho[1], Belinda Giardine[1], Lindsay R. Kasson[1], Robert S. Harris[1], Desiree C. Petersen[2], Fangqing Zhao[1], Ji Qi[1], Can Alkan[3], Jeffrey M. Kidd[3], Yazhou Sun[1], Daniela I. Drautz[1], Pascal Bouffard[4], Donna M. Muzny[5], Jeffrey G. Reid[5], Lynne V. Nazareth[5], Qingyu Wang[1], Richard Burhans[1], Cathy Riemer[1], Nicola E. Wittekindt[1], Priya Moorjani[6], Elizabeth A. Tindall[2,7], Charles G. Danko[8], Wee Siang Teo[2,7], Anne M. Buboltz[1], Zhenhai Zhang[1], Qianyi Ma[1], Arno Oosthuysen[9], Abraham W. Steenkamp[10], Hermann Oostuisen[11], Philippus Venter[12], John Gajewski[1], Yu Zhang[1], B. Franklin Pugh[1], Kateryna D. Makova[1], Anton Nekrutenko[1],

## 2008: 1000 Genomes Project…

# Contents

- Genomics
- **Sequencing**
- Metagenomics
- Computational questions
- Outlook

# Next-Generation Sequencing Technologies


- Fuelling a rapid growth of the number and size of sequencing projects

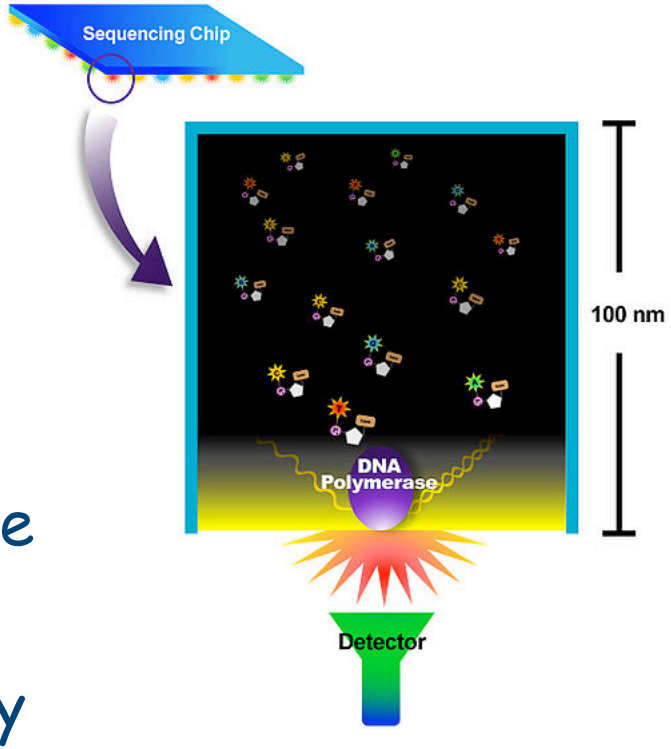
# Advances in Sequencing Technologies

- First generation (Sanger sequencing):
  - 100kb/run, read length 1000bp, 500$/Mb
- Second generation:
  - Roche/454: 450Mb/run, 400bp, 20$/Mb
  - Illumina: 35Gb/run, 100bp, 0.50$/Mb
  - SOLiD: 50Gb/run, 50bp, 0.50$/Mb
  - Heliscope: 37Gb/run, 32bp, <0.50$/Mb
- Third generation:
  - PacBio SMRT: 25Gb/run, >1000bp, ?$/Mb
- Other:
  - Ion Torrent: uses ion sensor, <100,000$

# SMRT™ Sequencing

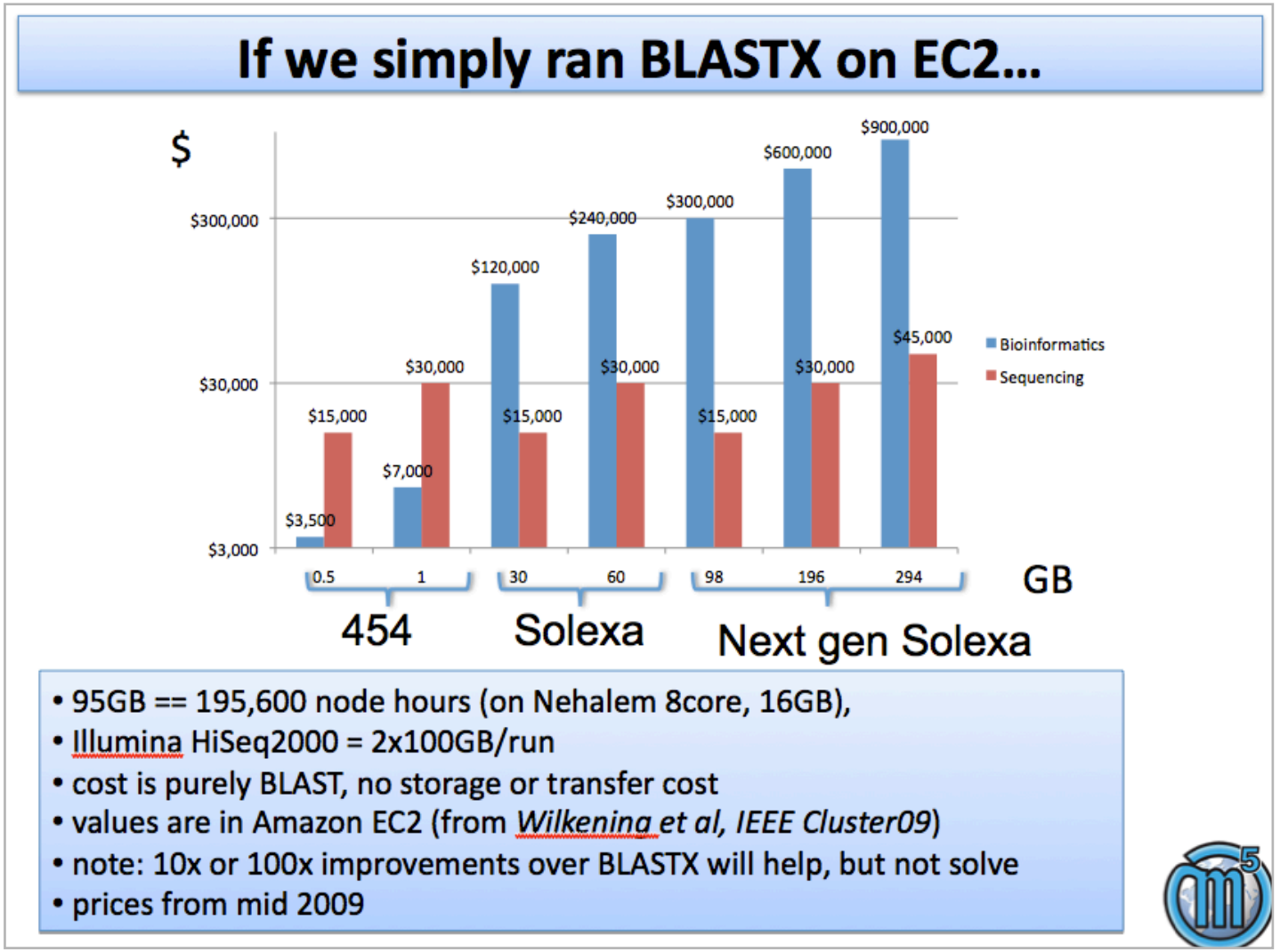
## Single Molecule Real Time Sequencing

- Observes detached fluorescent dye molecules




- Three protocols:
  - **Linear sequencing:**
    1kb reads, 10% deletion rate
  - **Circular sequencing:**
    e.g. 200bp reads, high quality
  - **Strobe sequencing:**
    e.g. 10 sections, each 100bp , each 500bp apart

# Sequencing No Longer the Bottleneck…




This slide kindly provided by Folker Meyer (Argonne National Labs)

# Contents

- Genomics
- Sequencing
- **Metagenomics**
- Computational questions
- Outlook

# How Many Species?

## Major unsolved question:

?

- Number of species on Earth?
- Cannot be answered even to within several orders of magnitude

- **Some estimations**
  - 3-50 million species of arthropods
  - 1-100 million species of nematodes

www.ucmp.berkeley.edu/arthropoda/arthropoda.html

## Once the diversity of the microbial world is catalogued, it will make astronomy look like a pitiful science

- Julian Davies, Professor Emeritus, Microbiology and Immunology, UBC

# Identified Modern Species

~1.7 million named species

- 287,655 plants, including:
  - 15,000 mosses
  - 13,025 ferns
  - 980 gymnosperms
  - 199,350 dicotyledons
  - 59,300 monocotyledons
- 74,000-120,000 fungi
- 10,000 lichens

- 5,700 prokaryotes

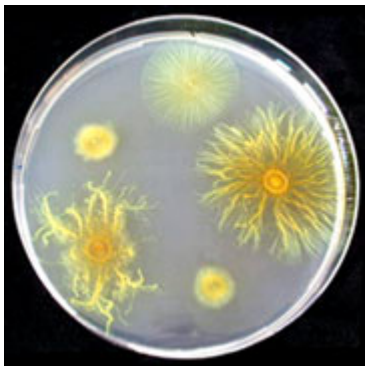
- ~1,250,000 animals, including:
  - 1,190,200 invertebrates:
    - 950,000 insects
    - 70,000 mollusks
    - 40,000 crustaceans
    - 130,200 others
  - 58,808 vertebrates:
    - 29,300 fish
    - 5,743 amphibians
    - 8,240 reptiles
    - 10,234 birds
    - 5,416 mammals

Source: http://en.wikipedia.org/wiki/Biodiversity

Sequences for ~200,000

# Metagenomics

- "The study of the DNA of uncultured organisms"
- > 99% of all microbes cannot be cultured
- *A genome:*
  - Entire genetic information of a single organism
- *A metagenome:*
  - Entire genetic information of a community of organisms



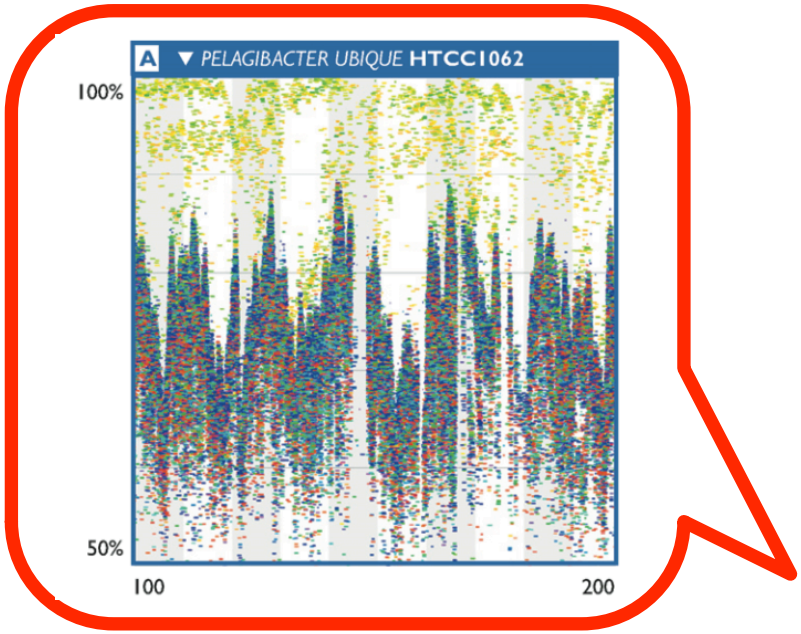
www.innovations-report.de

# Typical Sources of Metagenomes

- Soil samples
- Sea water samples
- Seabed samples
- Air samples
- Medical samples
- Ancient bones
- Human microbiome

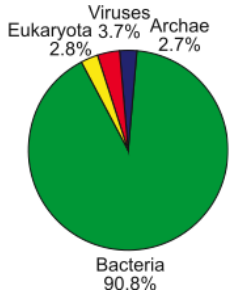



# Global Ocean Sampling Expedition




## Rusch *et al.* (2007):

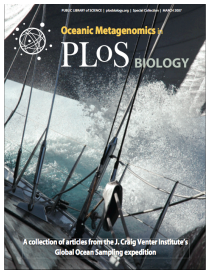
- 41 samples
- Size filtered 0.1-0.8μm
- Sanger sequencing
  - 7.7 million reads
  - length ~822bp
  - ~ 5.9Gb sequence
- Low abundance of *clonal* organisms




| | |
|---|---|
| Alpha *Proteobacteria* | 0.32 |
| Unclassified *Proteobacteria* | 0.155 |
| Gamma *Proteobacteria* | 0.132 |
| *Bacteroidetes* | 0.13 |
| *Cyanobacteria* | 0.079 |
| *Firmicutes* | 0.075 |
| *Actinobacteria* | 0.046 |
| Marine Group A | 0.022 |
| Beta *Proteobacteria* | 0.017 |
| OP11 | 0.008 |
| Unclassified *Bacteria* | 0.008 |
| Delta *Proteobacteria* | 0.005 |
| *Planctomycetes* | 0.002 |
| Epsilon *Proteobacteria* | 0.001 |

## Yooseph *et al.* (2007):

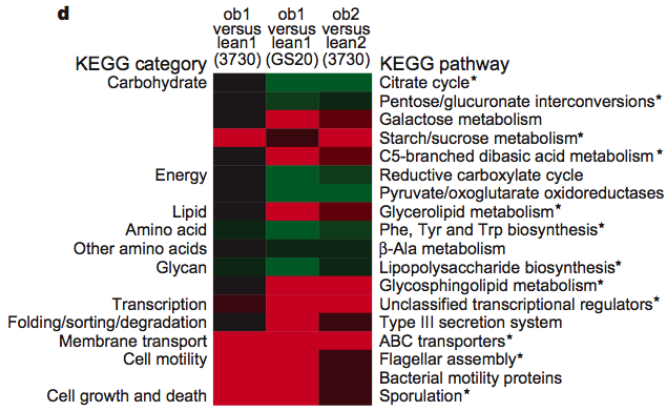
- 6 million proteins
  - linear rate of discovery



# Gut Microbiota

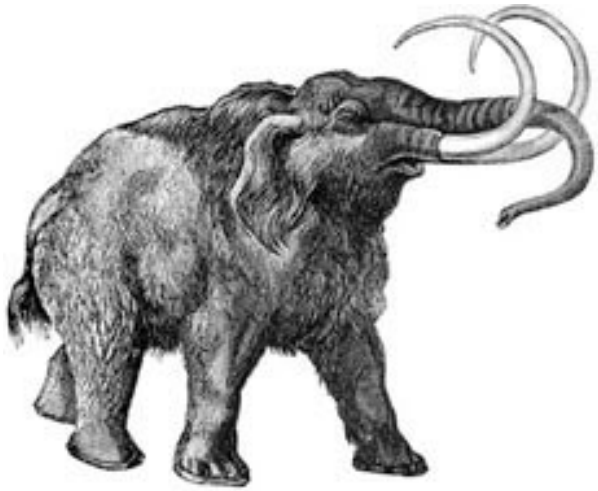

- Turnbaugh *et al* (2006)
- Caecal microbial DNA of *ob/ob*, *ob/+*, *+/+* mice
- Sanger sequencing:
  - 39.5 Mb
  - read length 750 bp
- 454 sequencing:
  - 160 Mb
  - read length 93 bp

- Obesity-associated gut microbiome
  - Change in relative abundance of Bacteroidetes and Firmicutes
  - Change in functional capacity (toward energy harvesting)



# Mammoth Project

- DNA collected from permafrost mammoth (28,000 years old)
- DNA extracted from 1 gram of bone
- 454 sequencing:
  - ~302,000 reads
  - ~95 bp length
- > 50% mammoth

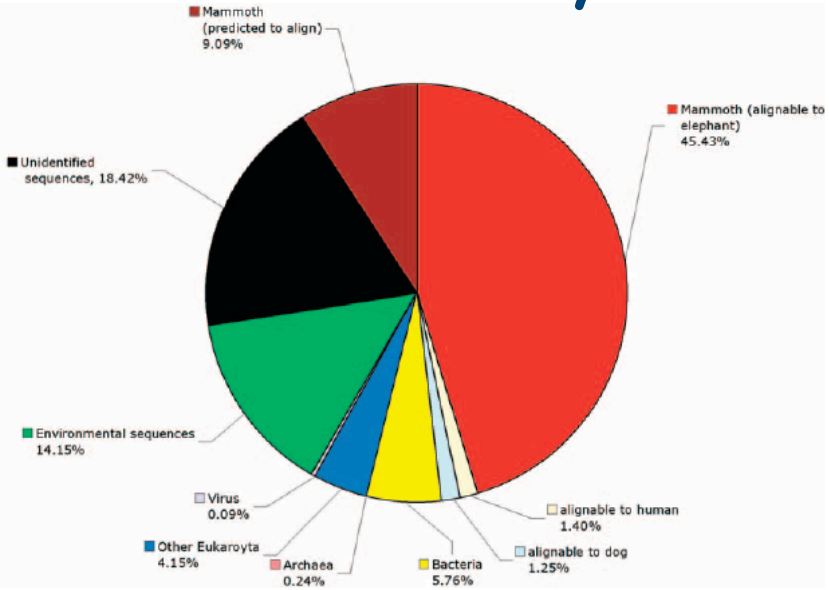


REPORTS

## Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA

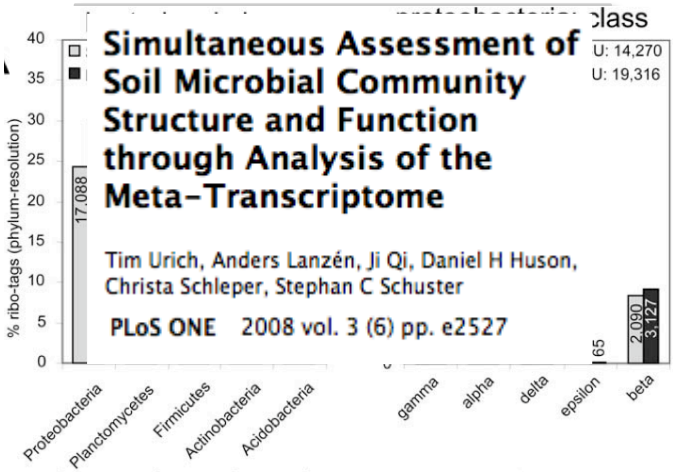
Hendrik N. Poinar,[1,2,3*] Carsten Schwarz,[1,2] Ji Qi,[4] Beth Shapiro,[5] Ross D. E. MacPhee,[6] Bernard Buigues,[7] Alexei Tikhonov,[8] Daniel H. Huson,[9] Lynn P. Tomsho,[4] Alexander Auch,[9] Markus Rampp,[10] Webb Miller,[4] Stephan C. Schuster[4*]

## Taxonomic analysis

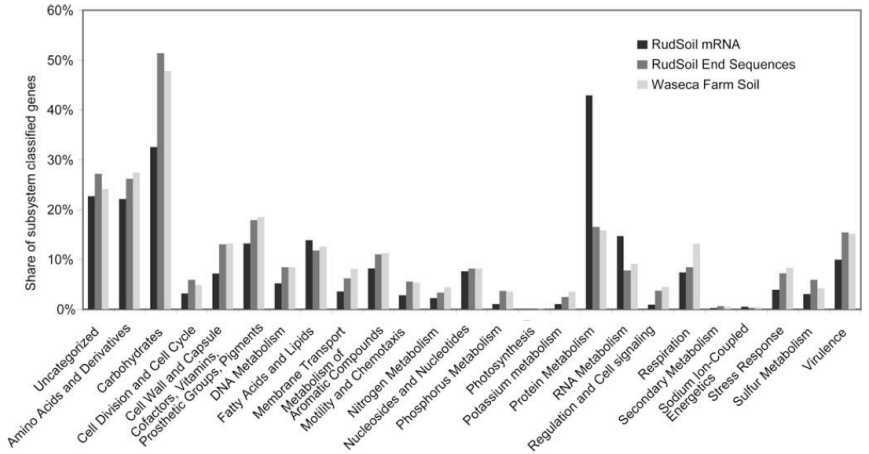


# "Meta Transcriptomics" of Soil

- Urich *et al* (2008):
  - RNA randomly reverse transcribed into cDNA
  - No PCR or cloning
  - 454 sequencing:
    - ~ 250,000 sequences
    - ~ 98 bp length
  - RNA types:
    - ~ 75% rRNA *tags*
    - ~ 8% mRNA *tags*
    - ~ 17% unassigned

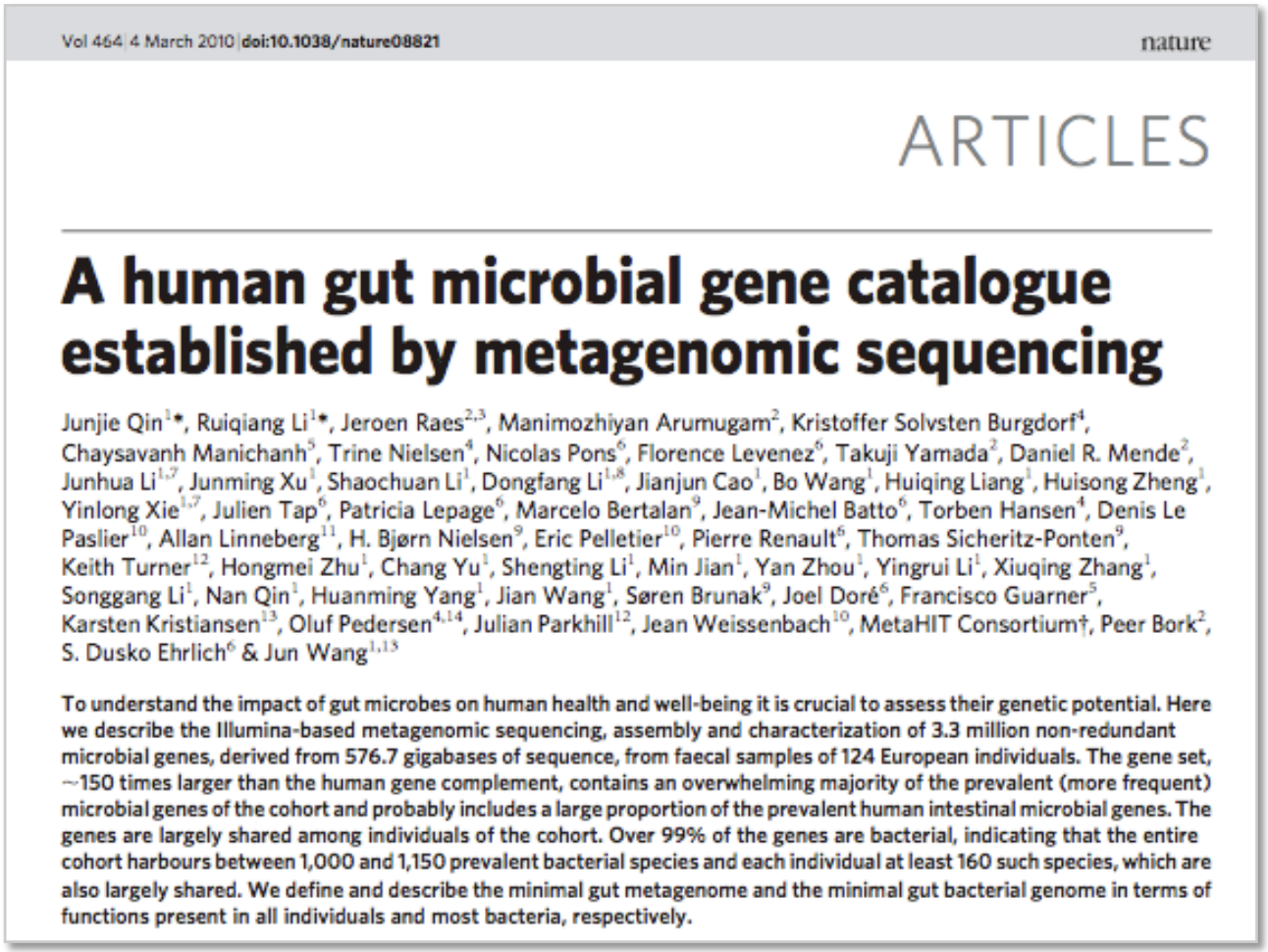



rRNA analysis




mRNA analysis

# Large-Scale Human Gut Analysis



Vol 464 | 4 March 2010 | doi:10.1038/nature08821 nature

ARTICLES

## A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

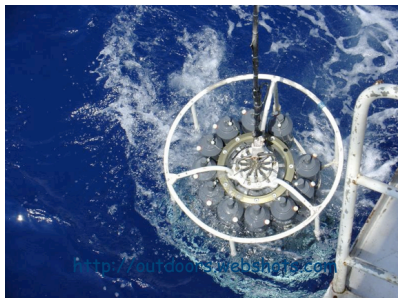
To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

- 576Gb of sequence from 124 individuals

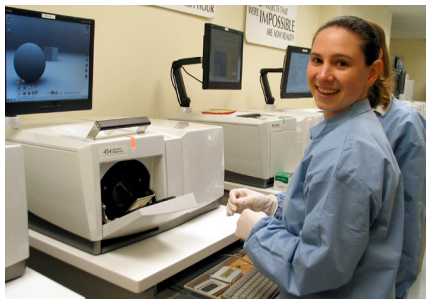
# Contents

- Genomics
- Sequencing
- Metagenomics
- **Computational questions**
- Outlook

# Metagenome Analysis



Environmental sample

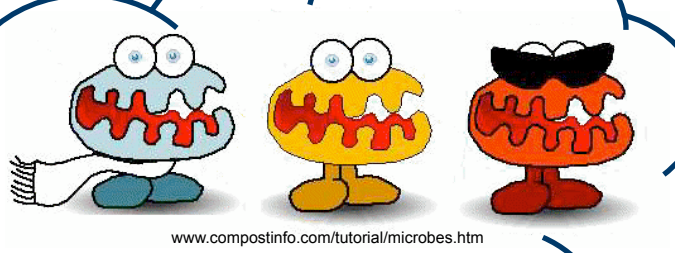


High-throughput DNA sequencing

10 million sequences

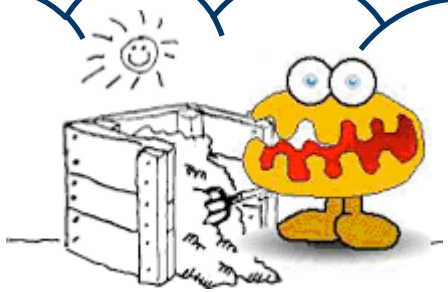
Basic computational analysis

10 000 hours



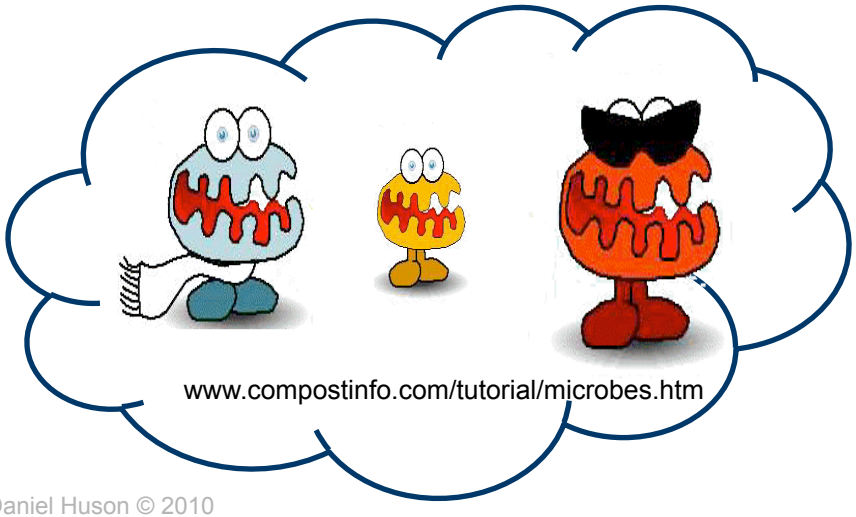
www.compostinfo.com/tutorial/microbes.htm

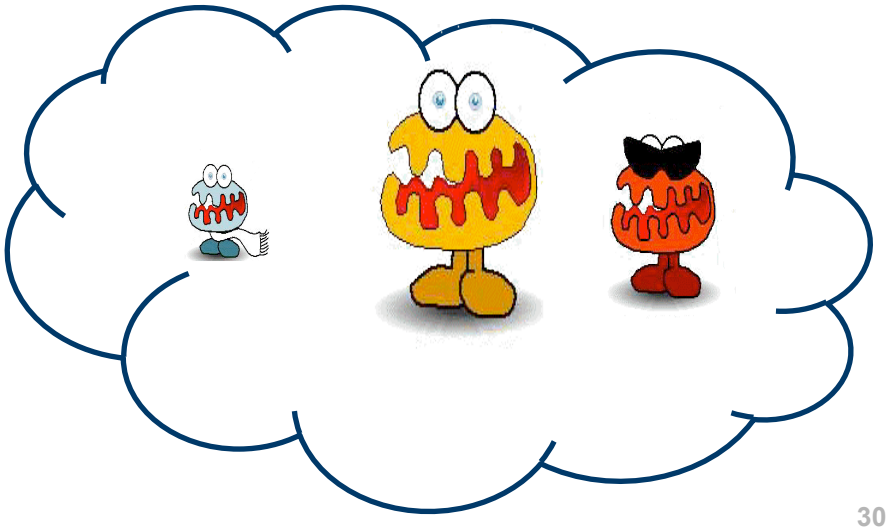
Q1: Who is out there?



Q2: What are they doing?

# Q3: How Do They Compare?

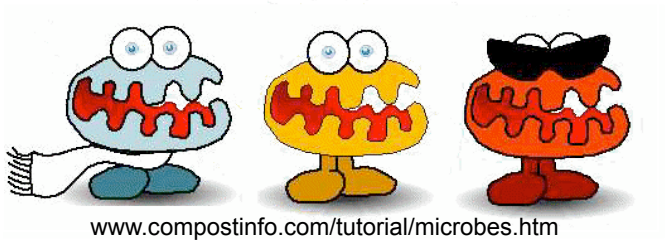



# Additional Questions

- How to cluster reads by relatedness (using machine learning techniques)?
- How to assemble metagenome data?
- Gene prediction?
- Faster sequence comparison
- etc

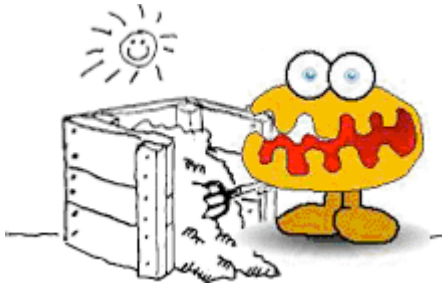
# Three Basic Computational Questions

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

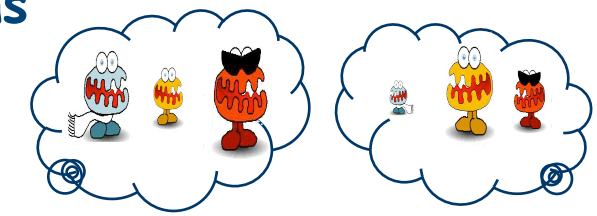


www.compostinfo.com/tutorial/microbes.htm

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



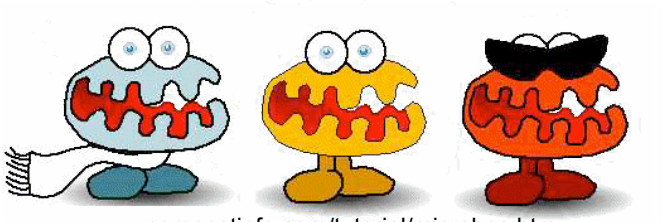
- How do different samples compare?
  - Pairwise and multiple comparisons
  - Correlations with environmental parameters?



- Serve to answer biological or medical questions

# Three Basic Computational Questions

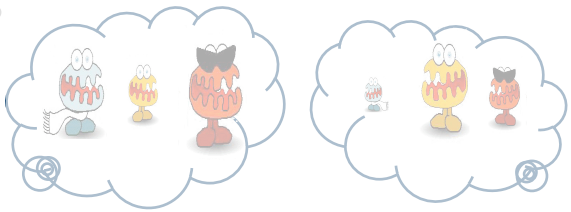
- **Who is out there?**
  - Types of organisms
  - In what proportions?



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?

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

# Who is Out There?

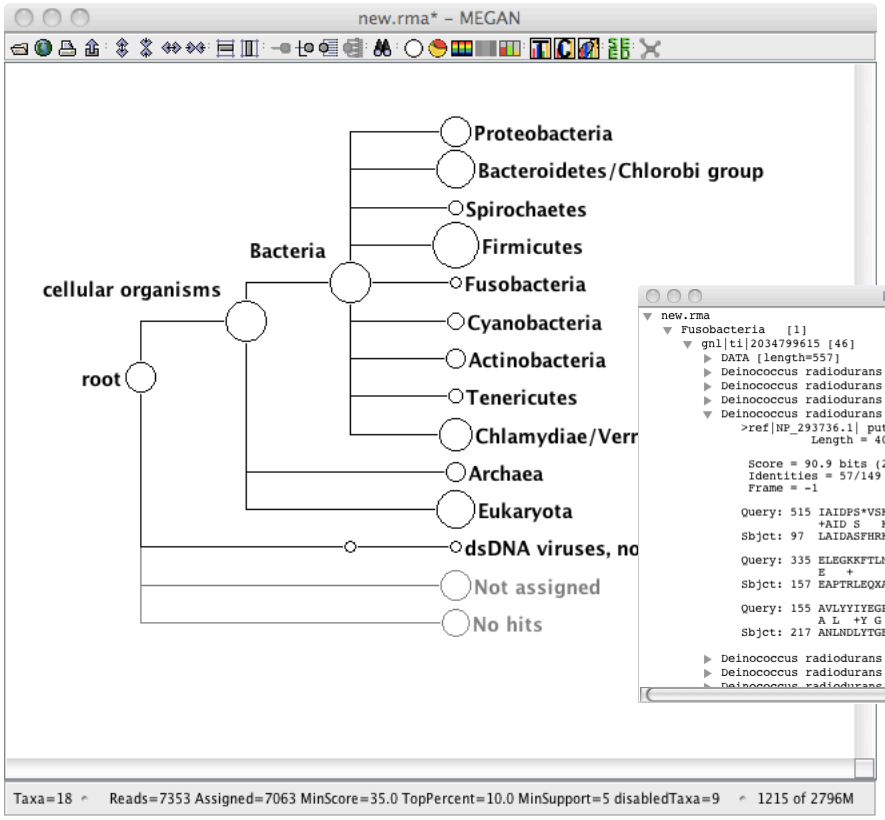
Main tool: Similarity search

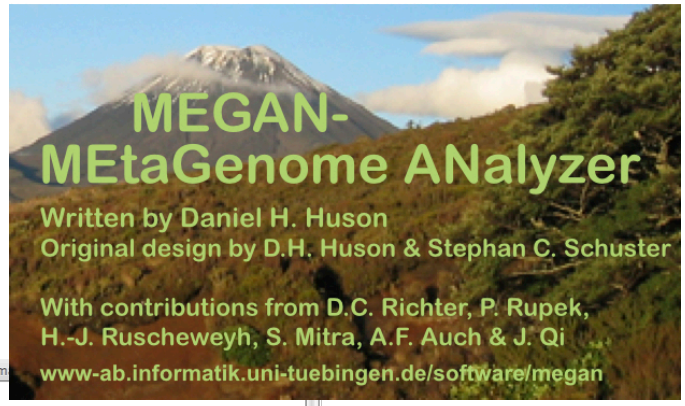
For every DNA (or cDNA) read:

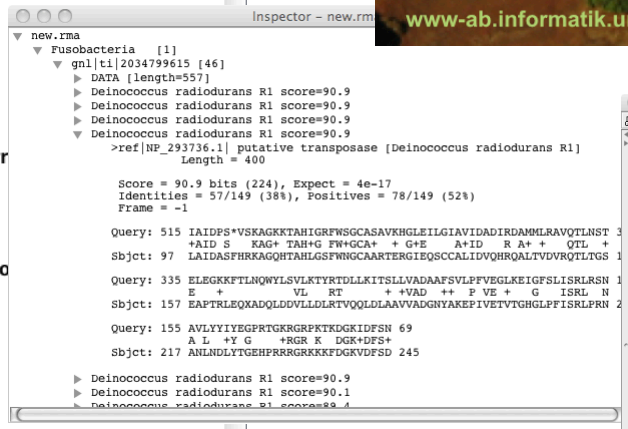
- Find significant matches to sequences in a reference database
- Use matches to place read in NCBI taxonomy

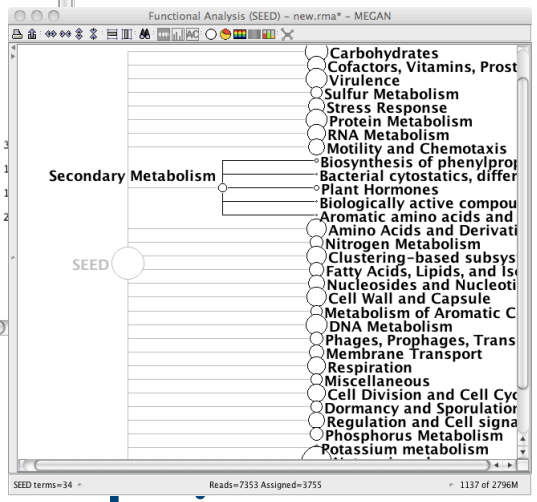
# MEGAN – MEtaGenome ANalyzer

Huson et al, 2007



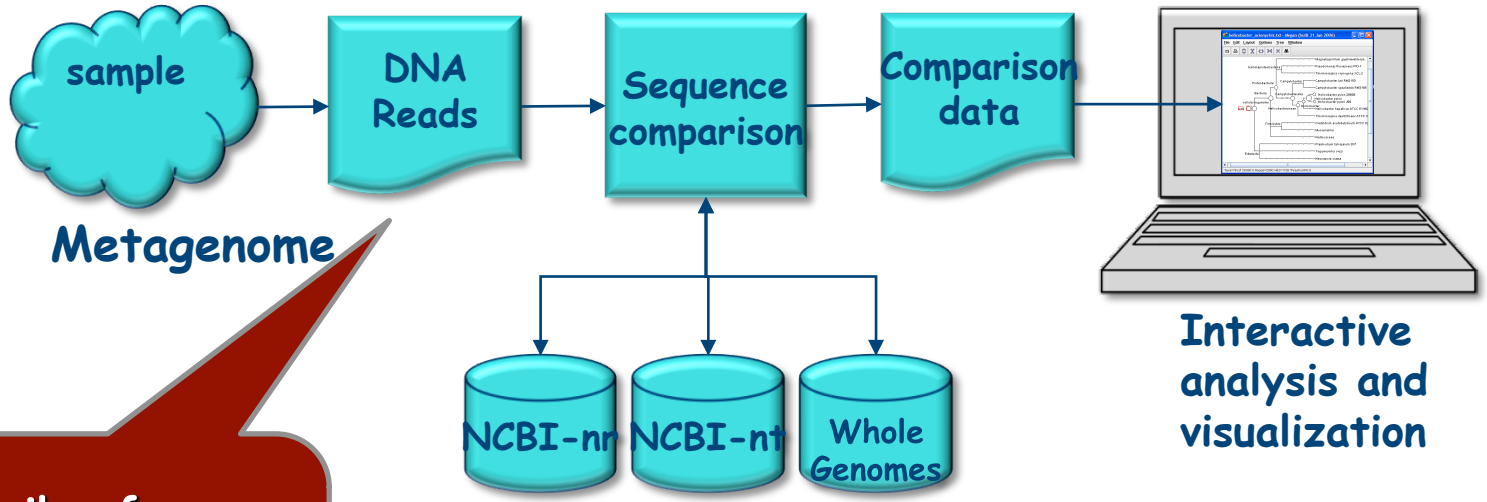




- **Interactive tool for metagenomic analysis**
  **(Version 4, to be released Nov 2010)**

# Metagenomics Pipeline

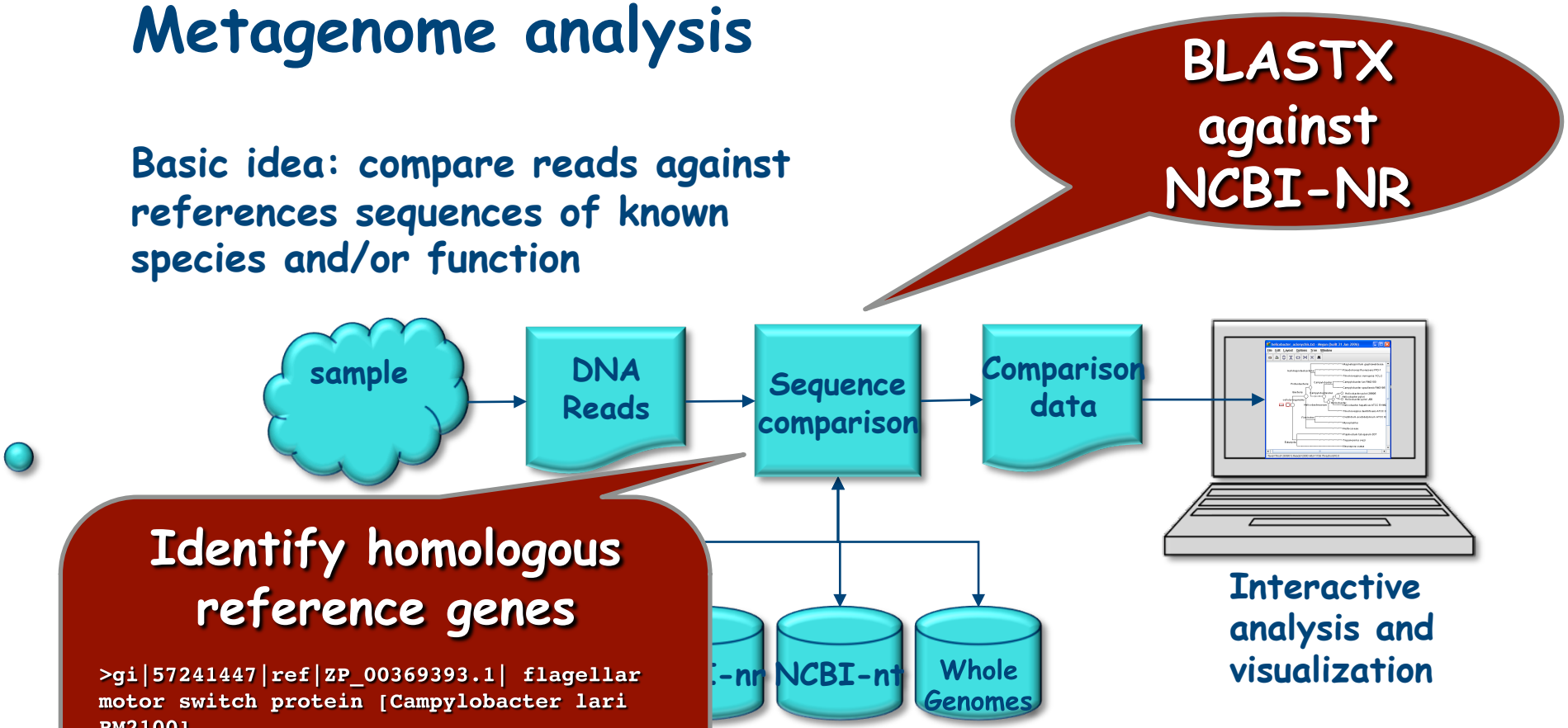



Similar for
- metatranscriptomics
- metaproteomics
- amplicon sequencing

# Identifying Taxa and Genes

## Metagenome analysis

Basic idea: compare reads against references sequences of known species and/or function

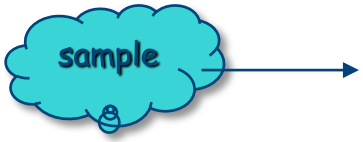



**Identify homologous reference genes**

```
>gi|57241447|ref|ZP_00369393.1| flagellar
motor switch protein [Campylobacter lari
RM2100]
 Score = 33.9 bits (76), Expect = 1.8
 Identities = 13/26 (50%), Positives =
19/26 (73%)

Query: 79  LMFVFDDLATVEENGIREIINRADKK 2
           LMF FDD++ +  N IRE++  ADK+
Sbjct: 243 LMFTFDDISQLSTNAIREVLKAADKR 268
```

# Sequence Comparison

- DNA Read




ACTGTGCACGTTGACGTAAGTTT...CGTGT

- Align to reference sequences, e.g. BLASTX against NR database:

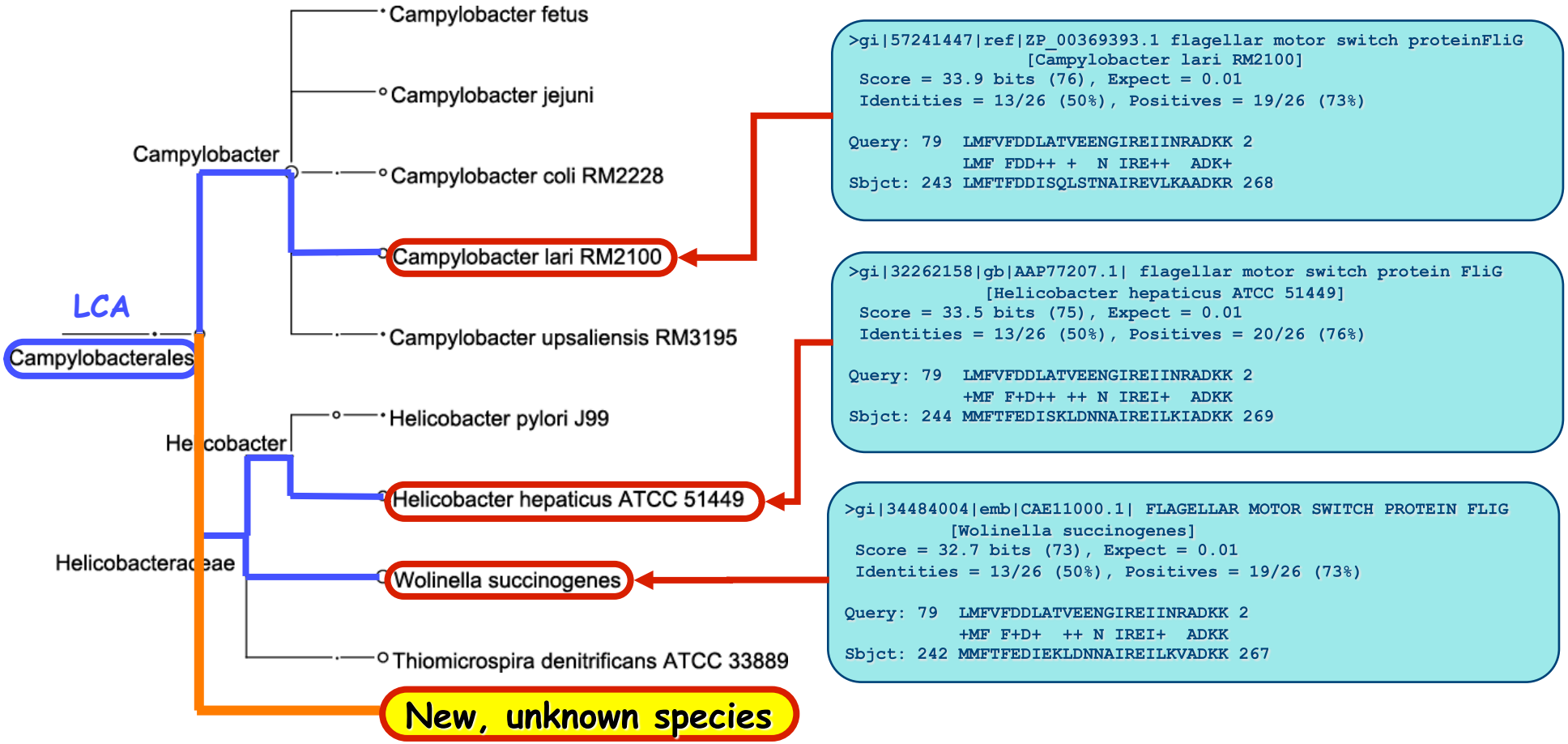
```
>gi|57241447|ref|ZP_00369393.1| flagellar motor switch protein
                    Campylobacter lari RM2100|
 Score = 33.9 bits (76), Expect =0.01
 Identities = 13/26 (50%), Positives = 19/26 (73%)

Query: 79   LMFVFDDLATVEENGIREIINRADKK 2
            LMF FDD++ +  N IRE++  ADK+
Sbjct: 243 LMFTFDDISQLSTNAIREVLKAADKR 268
```

- Indicates gene content:

**Campylobacter lari RM2100**

# Taxonomic Placement Using LCA

## A read will often match more than one database entry:




**LCA approach: Assign read to LCA of hits in taxonomy**

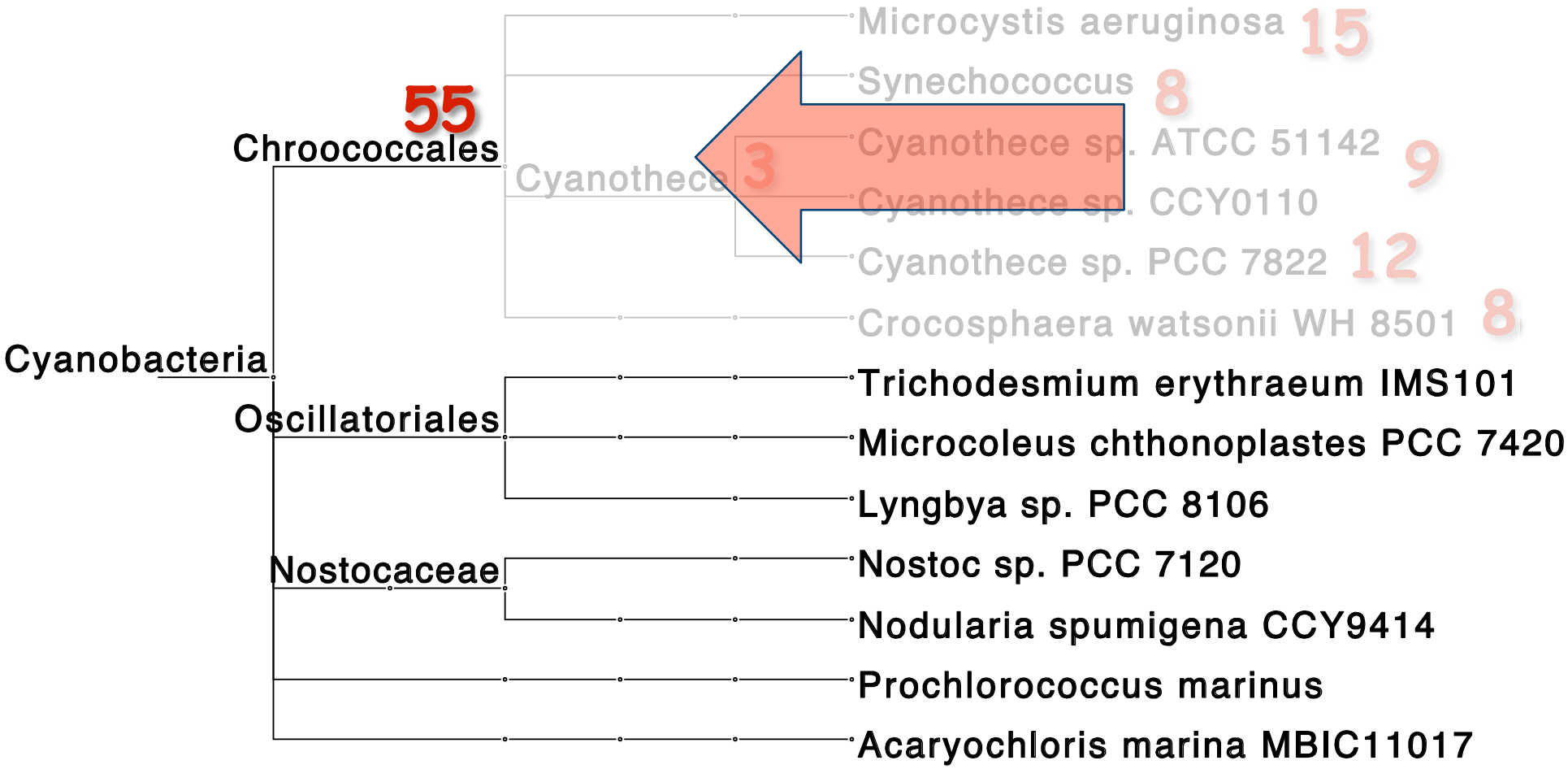
# Taxonomic Placement Using LCA

- For each DNA read:
  - Determine which gene sequences it matches
  - Corresponding species are assumed to contain the gene
  - Place read on the LCA of species

- Is placement by gene content or phylogenetic footprint

- Robust against false positive placements
- Robust against (known cases) of horizontal gene transfer

# Minimum Support Filter

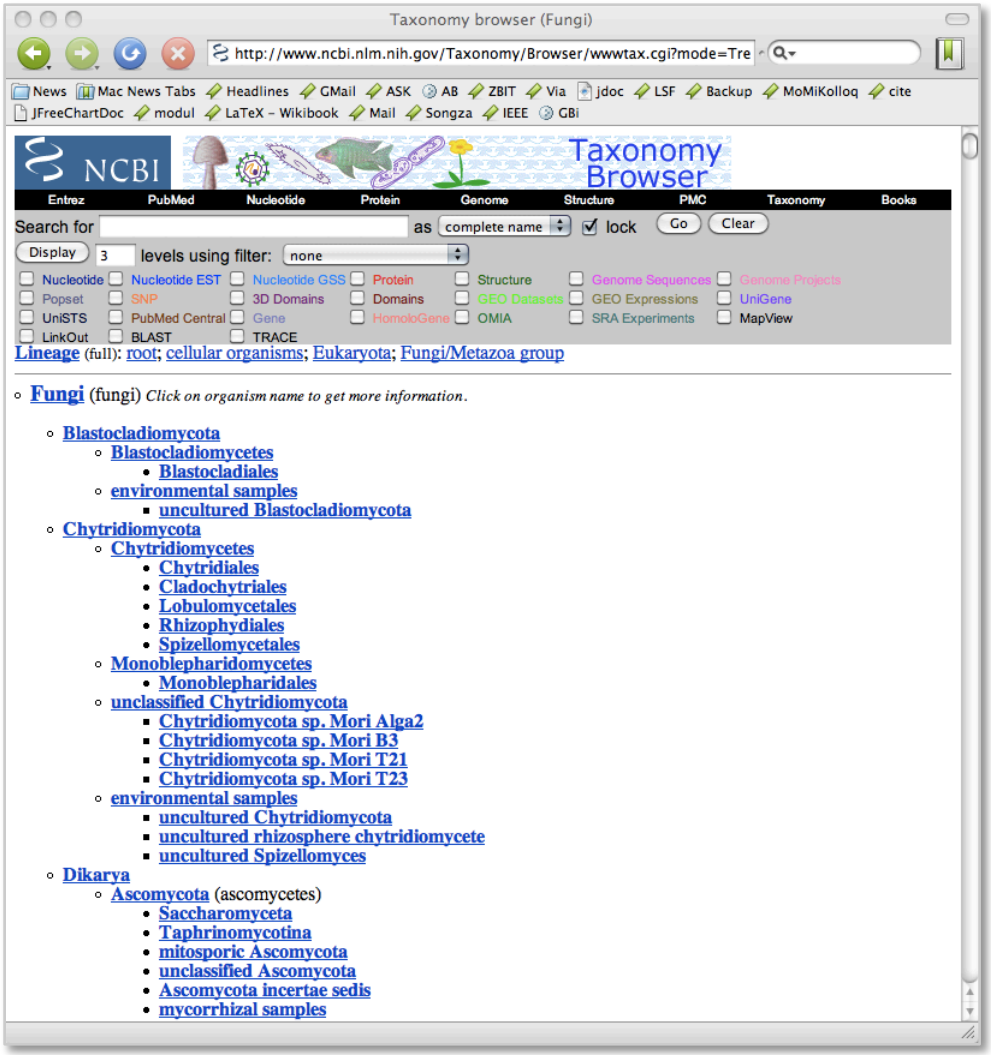



- Require at least e.g. 50 reads on a node

# Taxonomic Analysis using NCBI Taxonomy

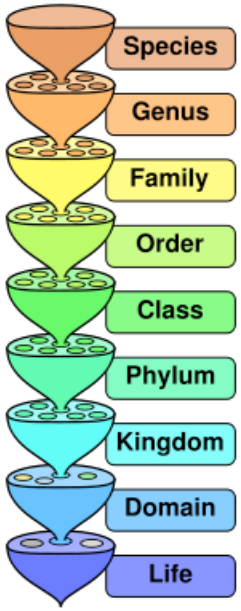
## NCBI taxonomy:

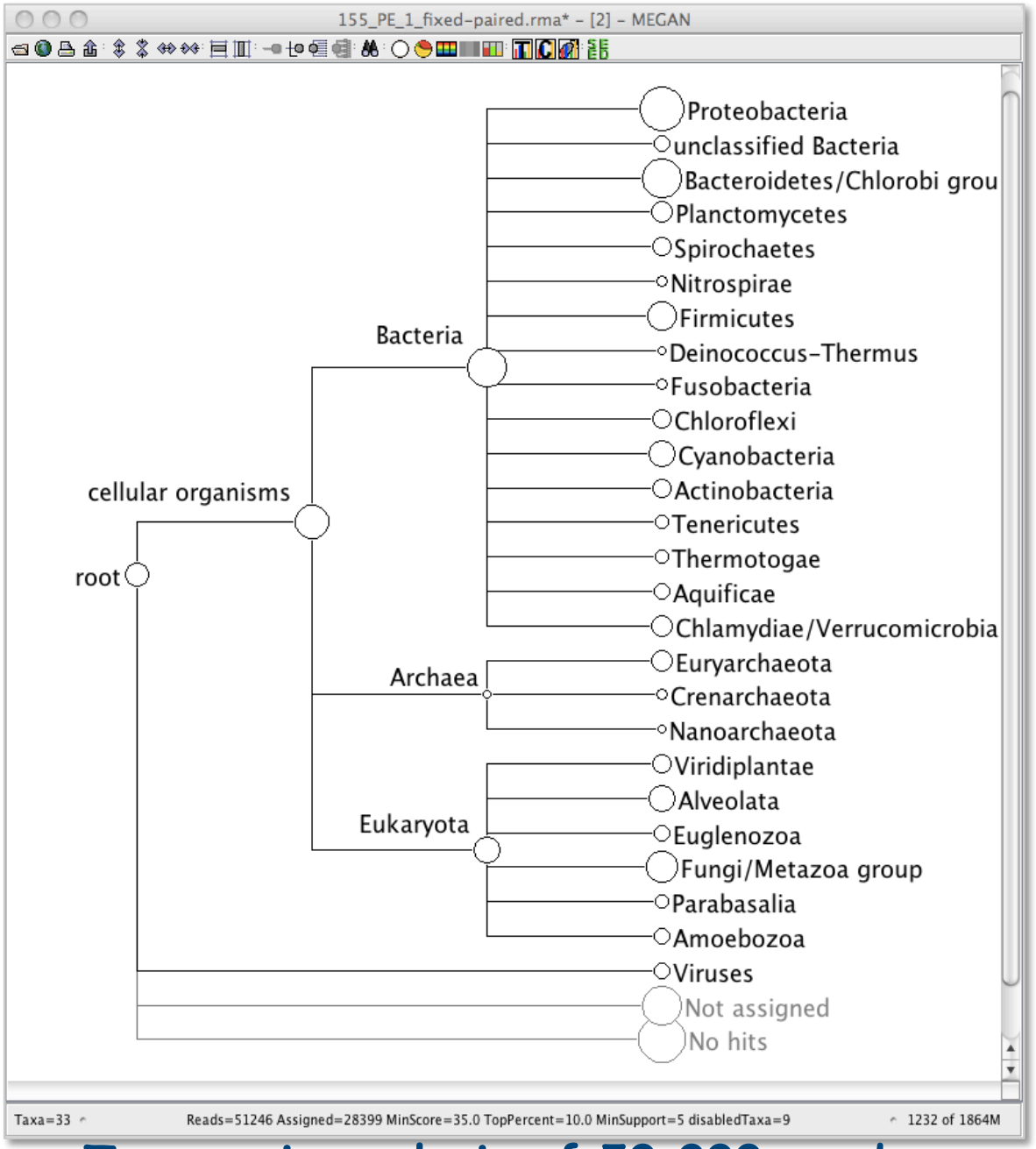
- Contains all species represented by some sequence
- >560,000 nodes
- (2007: 280,000 nodes)
- King Phillip Came Over For Green Soup… (and more)



# Organize and Visualize

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

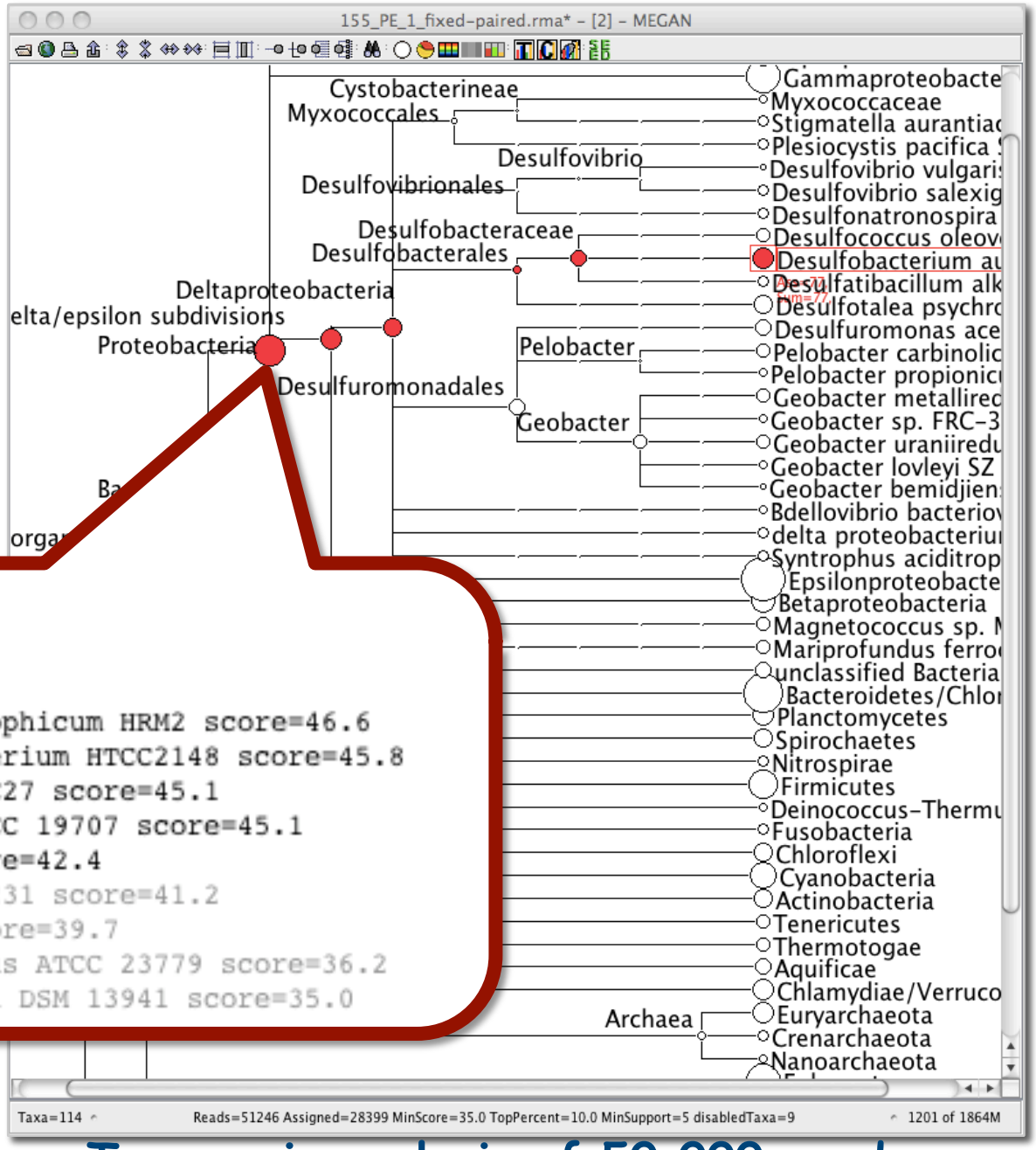



Taxonomic analysis of 50,000 reads

# Organize and Visualize

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

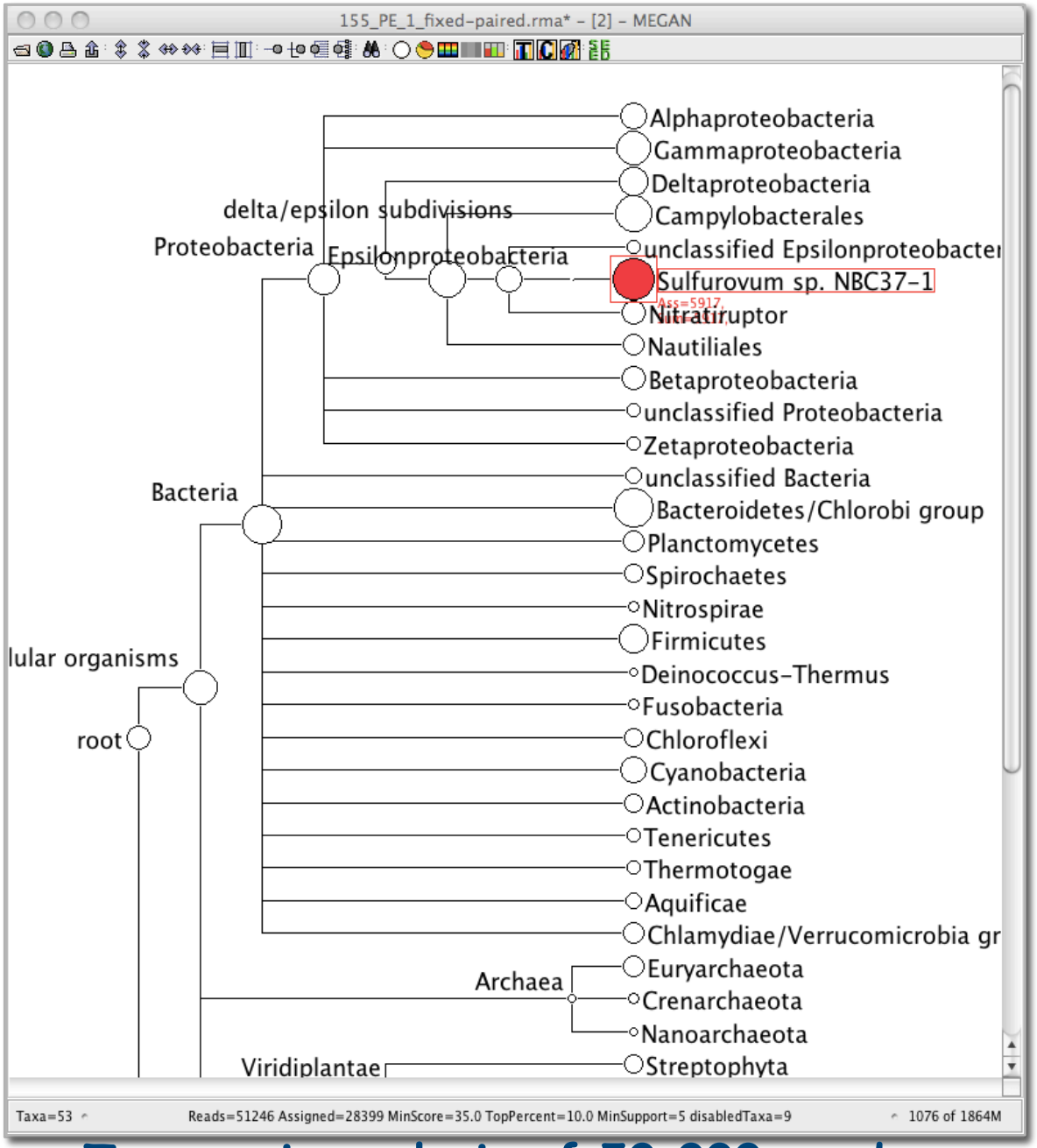



GDEG1CX11G0QZ8.1 [9]
- DATA [length=233]
- Desulfobacterium autotrophicum HRM2 score=46.6
- marine gamma proteobacterium HTCC2148 score=45.8
- Nitrosococcus oceani AFC27 score=45.1
- Nitrosococcus oceani ATCC 19707 score=45.1
- Roseovarius sp. 217 score=42.4
- Nitrococcus mobilis Nb-231 score=41.2
- Geobacter lovleyi SZ score=39.7
- Herpetosiphon aurantiacus ATCC 23779 score=36.2
- Roseiflexus castenholzii DSM 13941 score=35.0

Taxonomic analysis of 50,000 reads

# Interact and Summarize

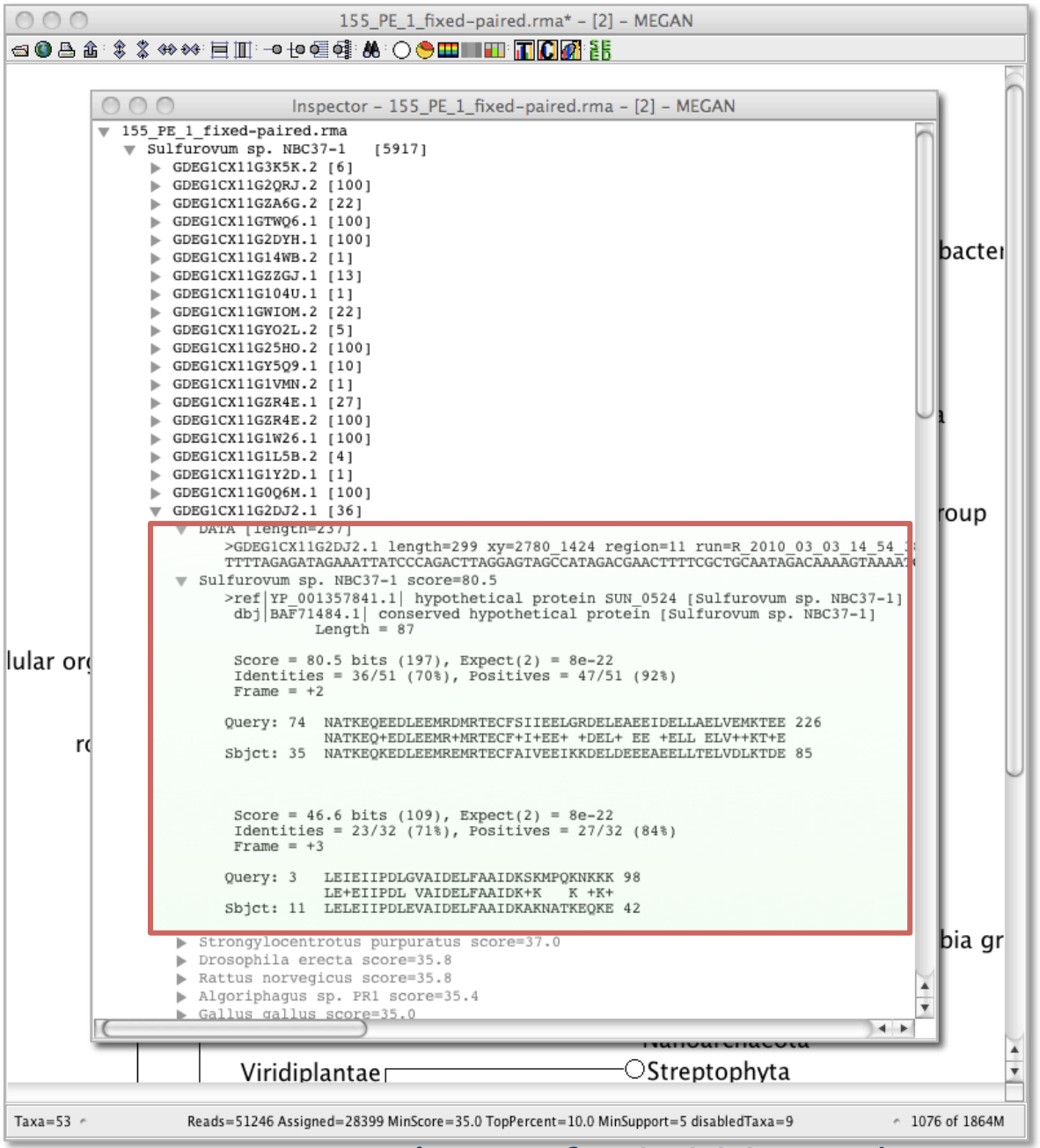
- ✓ Search for nodes of interest




Taxonomic analysis of 50,000 reads

# Interact and Summarize

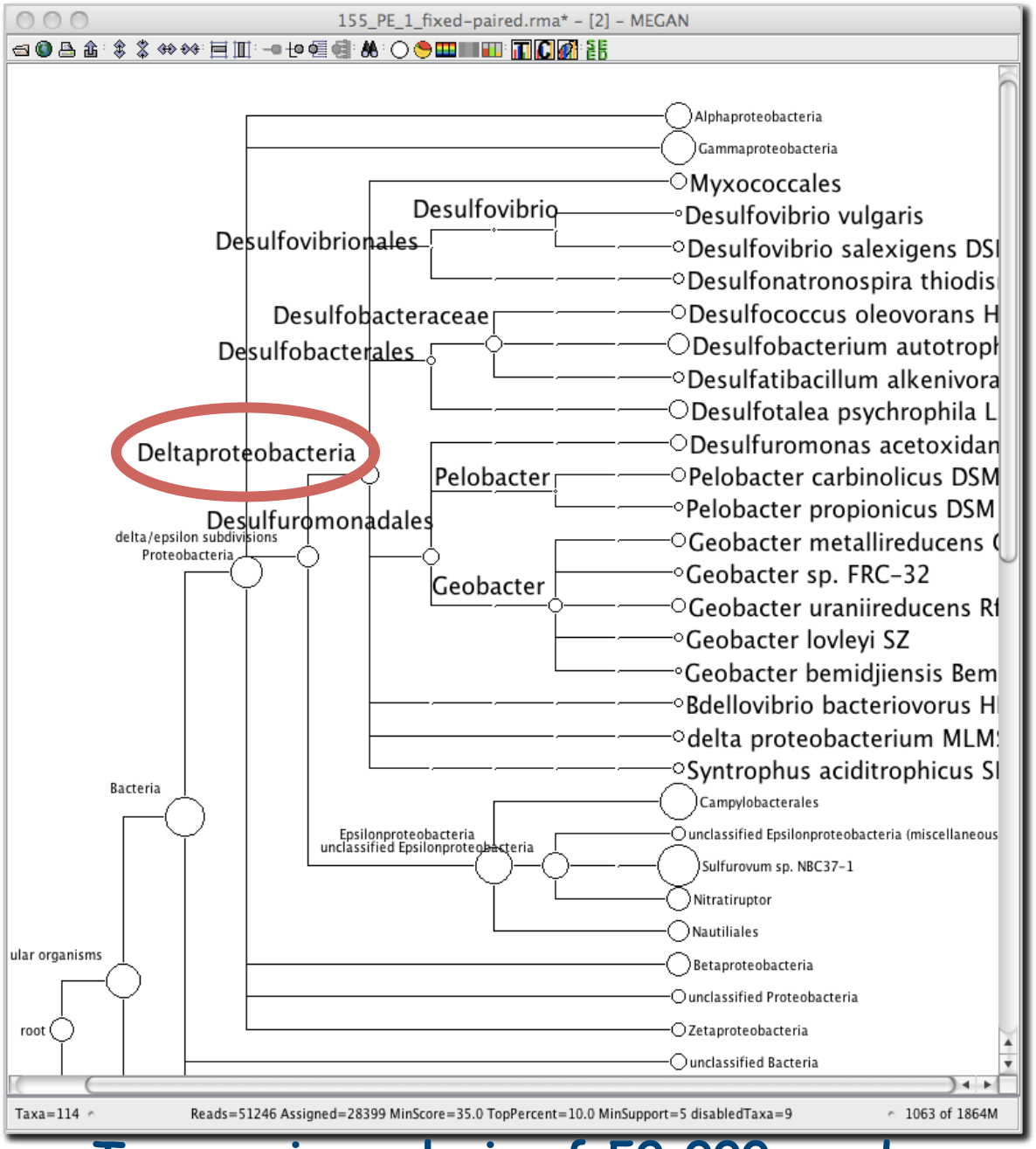
- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node




Taxonomic analysis of 50,000 reads

# Interact and Summarize

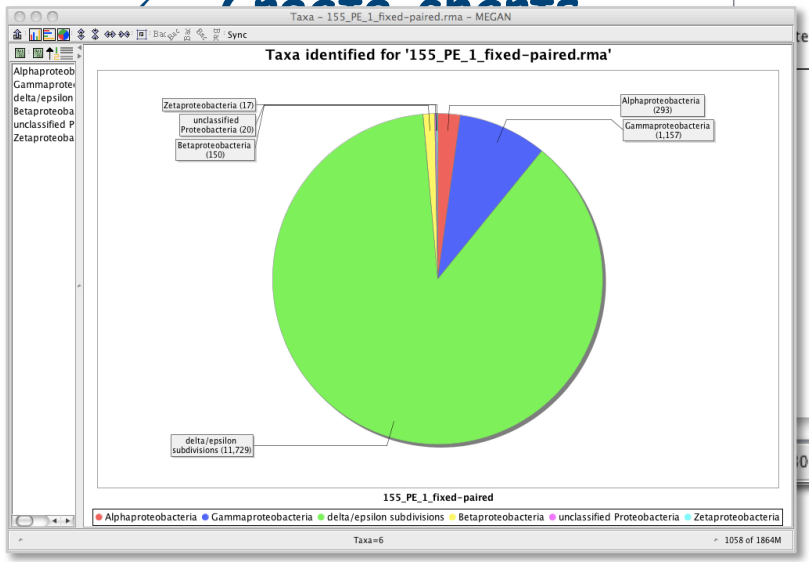
- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node
- ✓ Collapse and un-collapse parts of the tree

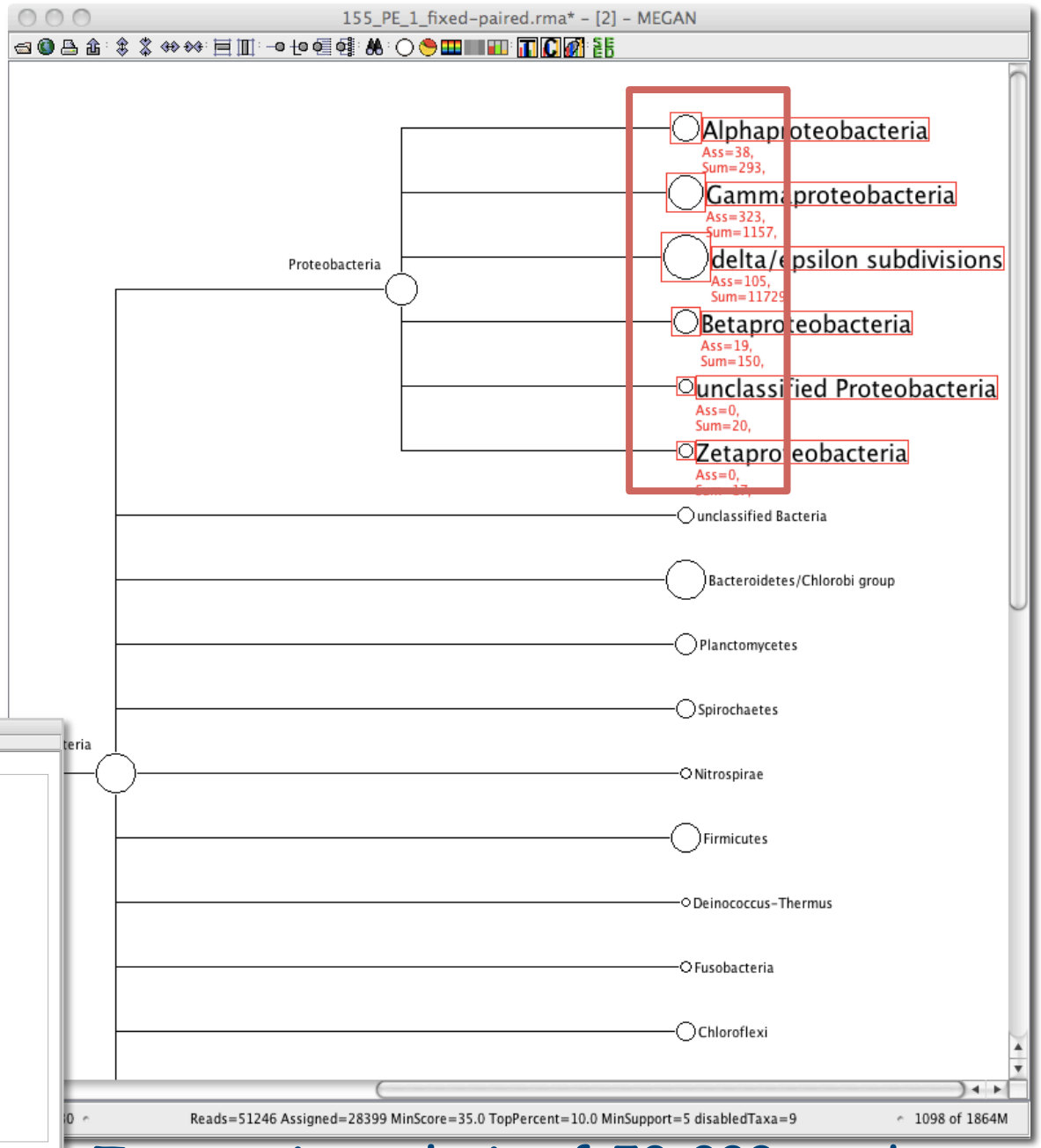



Taxonomic analysis of 50,000 reads

# Interact and Summarize

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

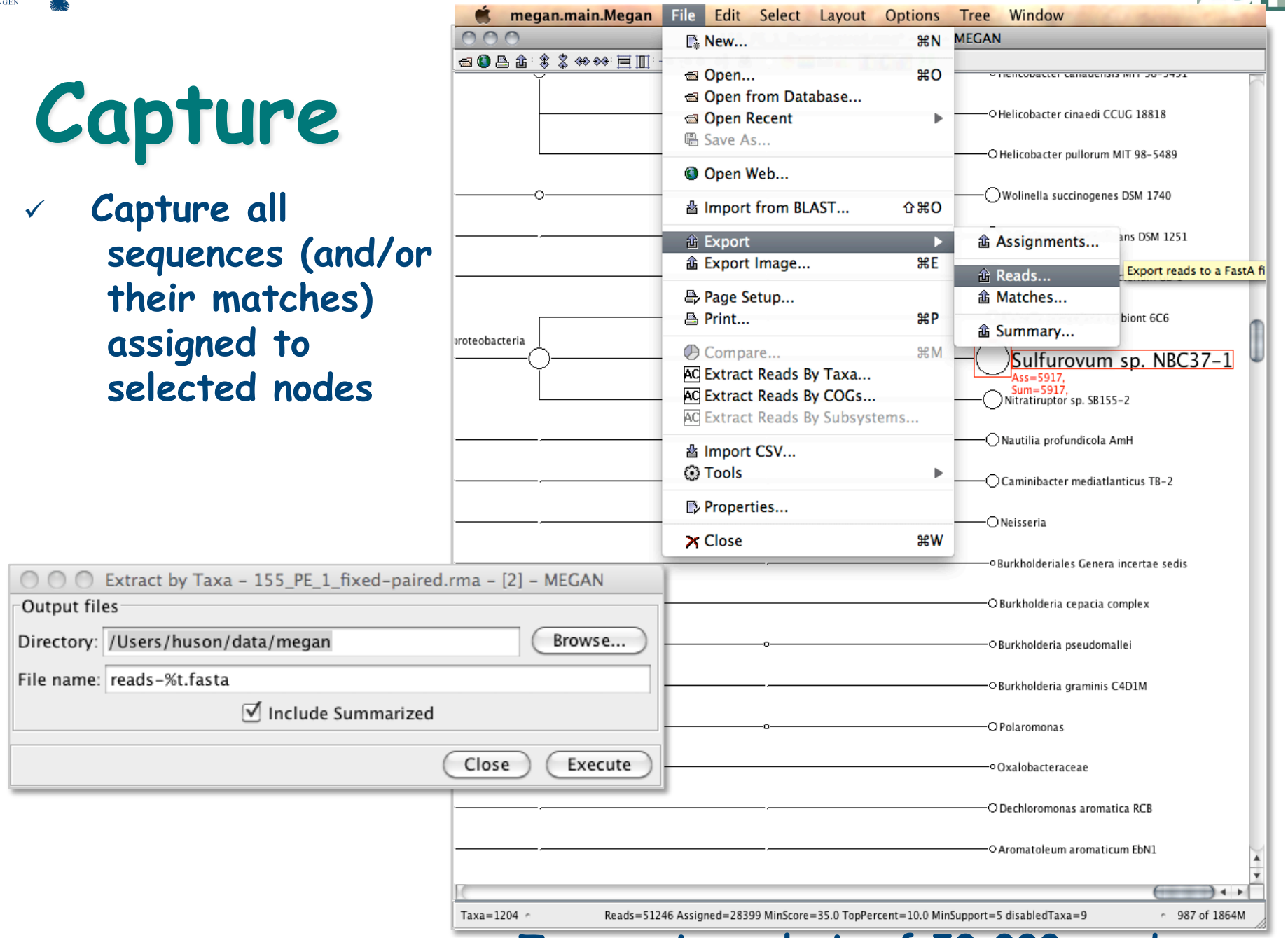



Taxonomic analysis of 50,000 reads

# Capture

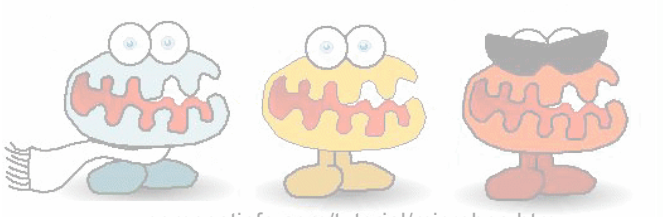
- ✓ Capture all sequences (and/or their matches) assigned to selected nodes




Taxonomic analysis of 50,000 reads

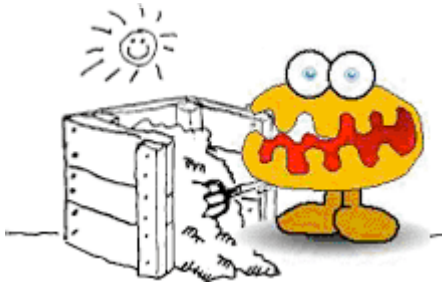
# Three Basic Computational Questions

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

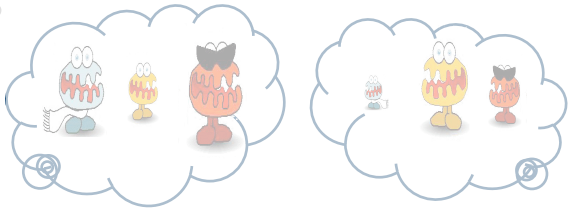


www.compostinfo.com/tutorial/microbes.htm

- **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?

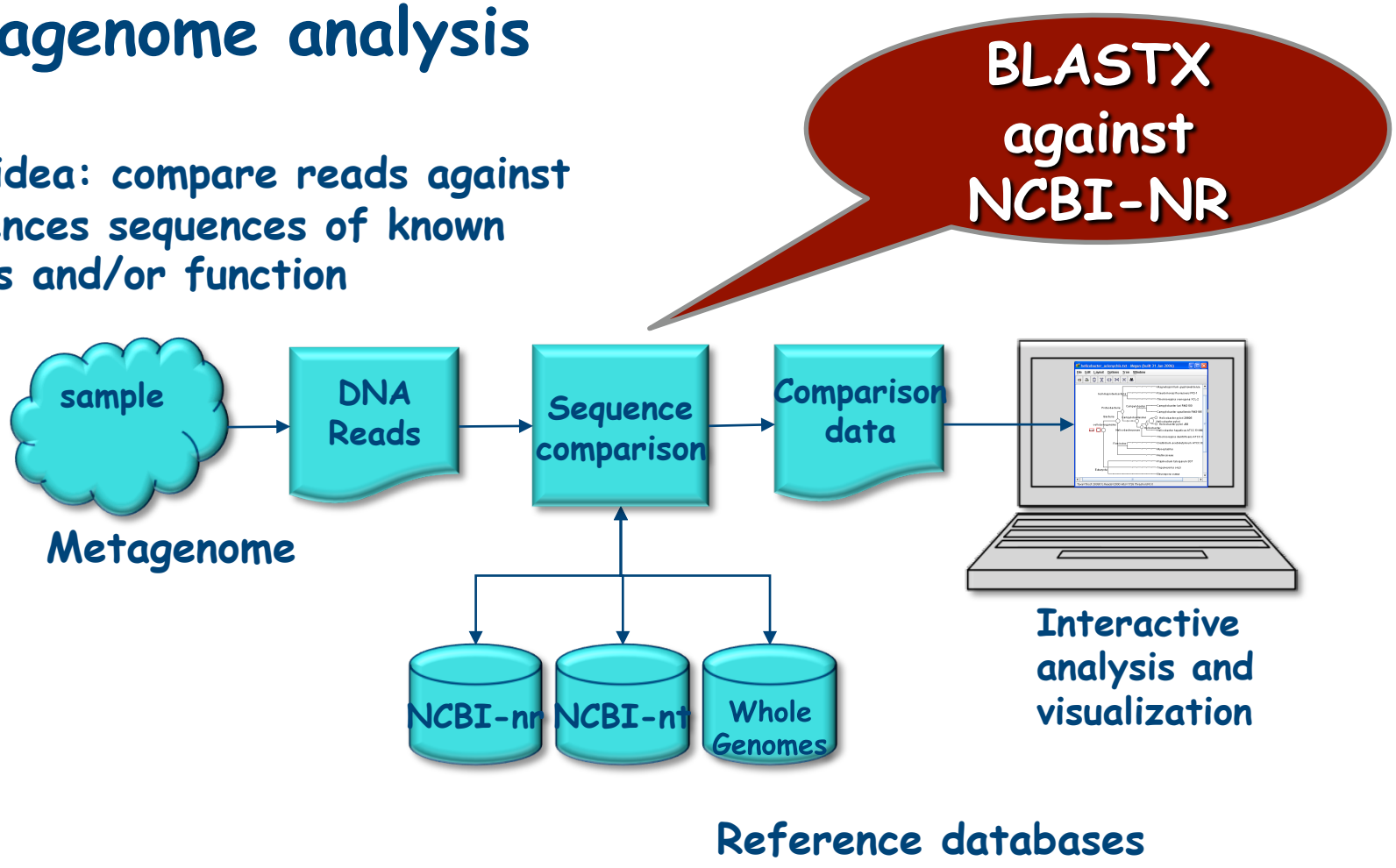


- Serve to answer biological or medical questions

# Identifying Taxa and Genes

## Metagenome analysis

Basic idea: compare reads against references sequences of known species and/or function

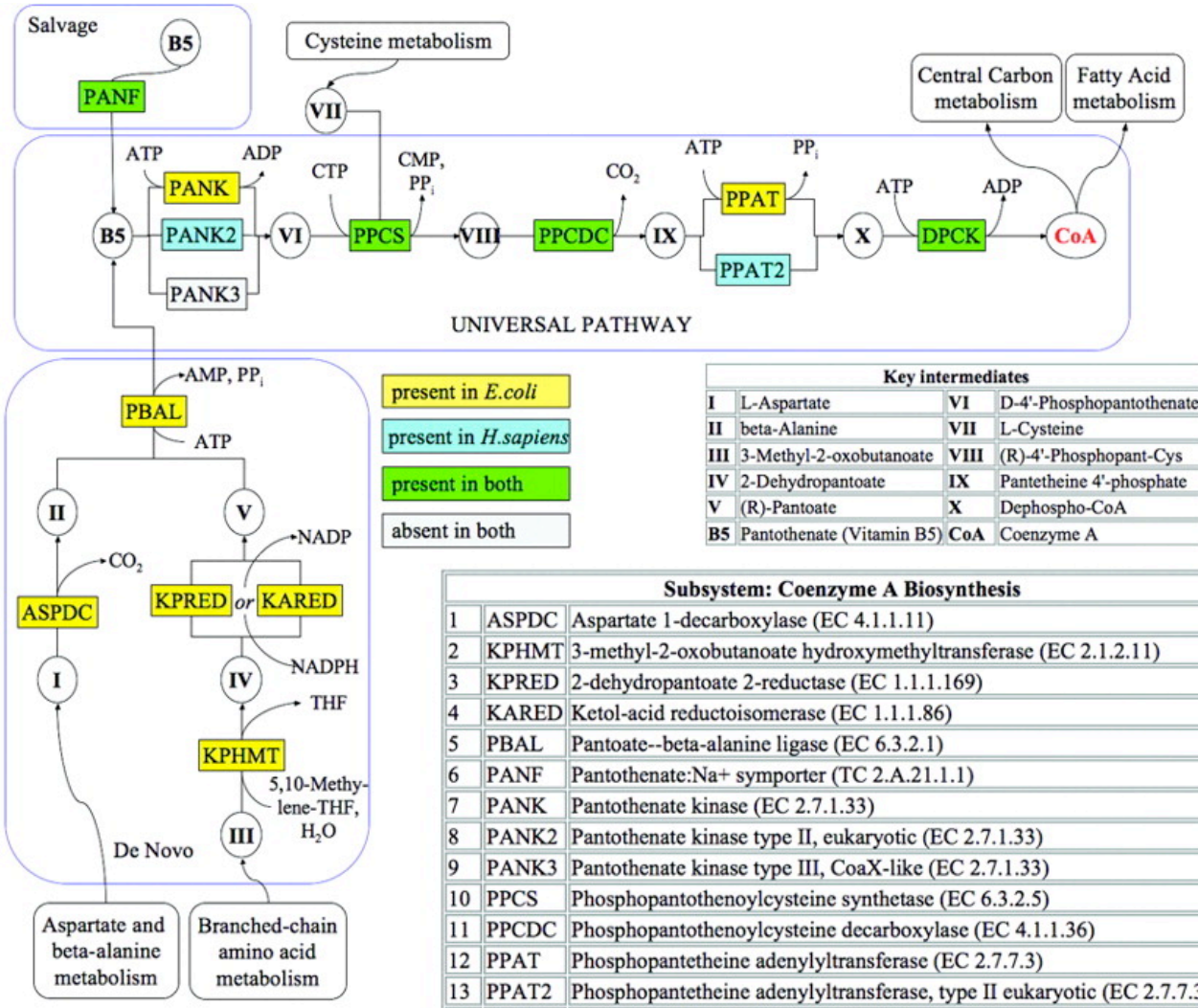


# Functional Analysis using SEED

- The SEED classification assigns genes to functional roles in subsystems
- A subsystem is a set of functional roles that implement a specific biological process or structural complex
- RAST and MG-RAST: Rapid annotation using subsystem technology
- Graph has ~10,000 nodes and edges
- www.theSEED.org

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

# Example of a Subsystem




| Key intermediates | | | |
|---|---|---|---|
| I | L-Aspartate | VI | D-4'-Phosphopantothenate |
| II | beta-Alanine | VII | L-Cysteine |
| III | 3-Methyl-2-oxobutanoate | VIII | (R)-4'-Phosphopant-Cys |
| IV | 2-Dehydropantoate | IX | Pantetheine 4'-phosphate |
| V | (R)-Pantoate | X | Dephospho-CoA |
| B5 | Pantothenate (Vitamin B5) | CoA | Coenzyme A |

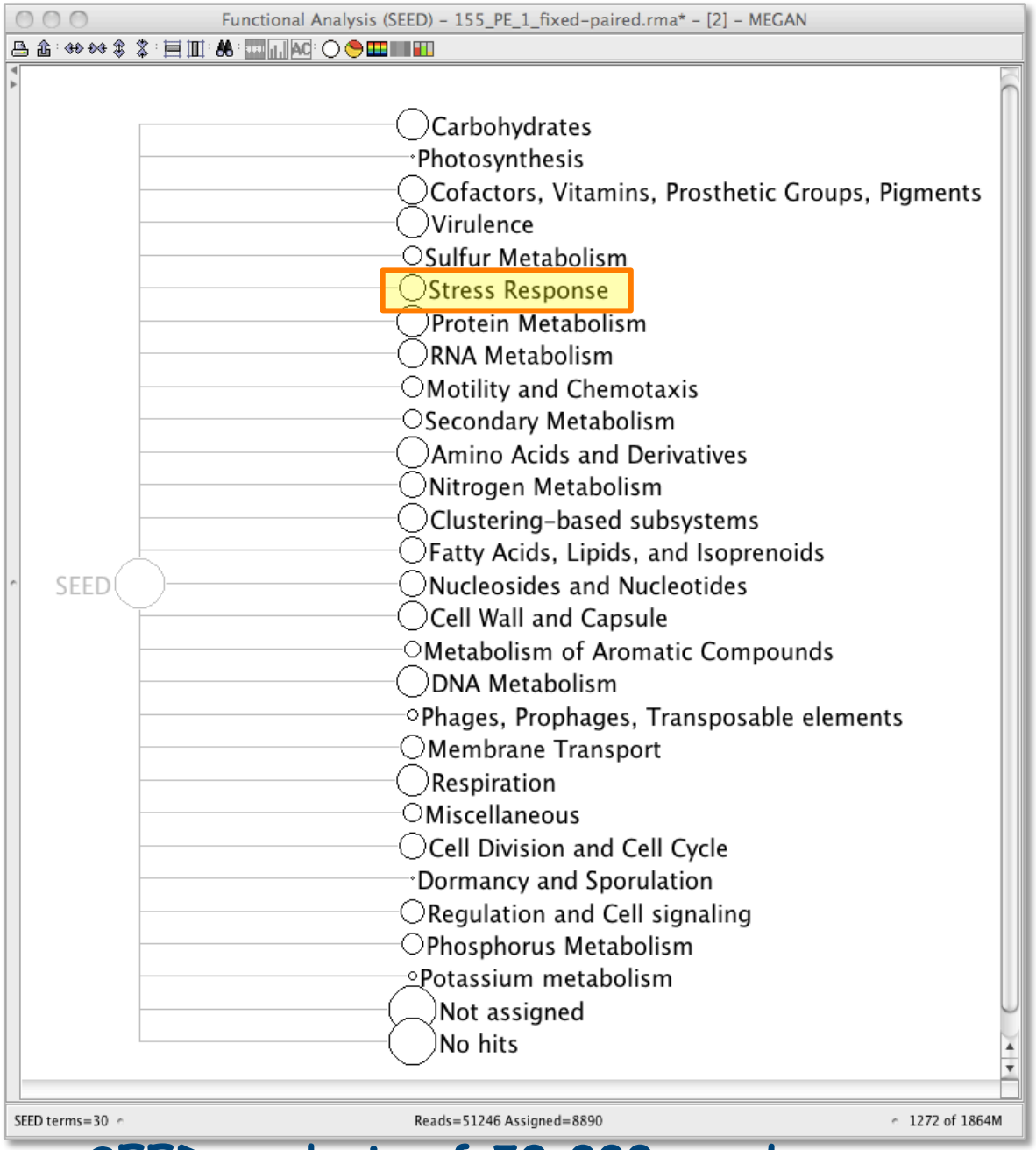
| Subsystem: Coenzyme A Biosynthesis | | |
|---|---|---|
| 1 | ASPDC | Aspartate 1-decarboxylase (EC 4.1.1.11) |
| 2 | KPHMT | 3-methyl-2-oxobutanoate hydroxymethyltransferase (EC 2.1.2.11) |
| 3 | KPRED | 2-dehydropantoate 2-reductase (EC 1.1.1.169) |
| 4 | KARED | Ketol-acid reductoisomerase (EC 1.1.1.86) |
| 5 | PBAL | Pantoate--beta-alanine ligase (EC 6.3.2.1) |
| 6 | PANF | Pantothenate:Na+ symporter (TC 2.A.21.1.1) |
| 7 | PANK | Pantothenate kinase (EC 2.7.1.33) |
| 8 | PANK2 | Pantothenate kinase type II, eukaryotic (EC 2.7.1.33) |
| 9 | PANK3 | Pantothenate kinase type III, CoaX-like (EC 2.7.1.33) |
| 10 | PPCS | Phosphopantothenoylcysteine synthetase (EC 6.3.2.5) |
| 11 | PPCDC | Phosphopantothenoylcysteine decarboxylase (EC 4.1.1.36) |
| 12 | PPAT | Phosphopantetheine adenylyltransferase (EC 2.7.7.3) |
| 13 | PPAT2 | Phosphopantetheine adenylyltransferase, type II eukaryotic (EC 2.7.7.3) |
| 14 | DPCK | Dephospho-CoA kinase (EC 2.7.1.24) |

## Coenzyme A Biosynthesis Subsystem

# Organize and Visualize

## Functional analysis

- ✓Use SEED classification to bin sequences by subsystems




SEED analysis of 50,000 reads

The SEED
Home of the SEED.

www.theseed.org

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

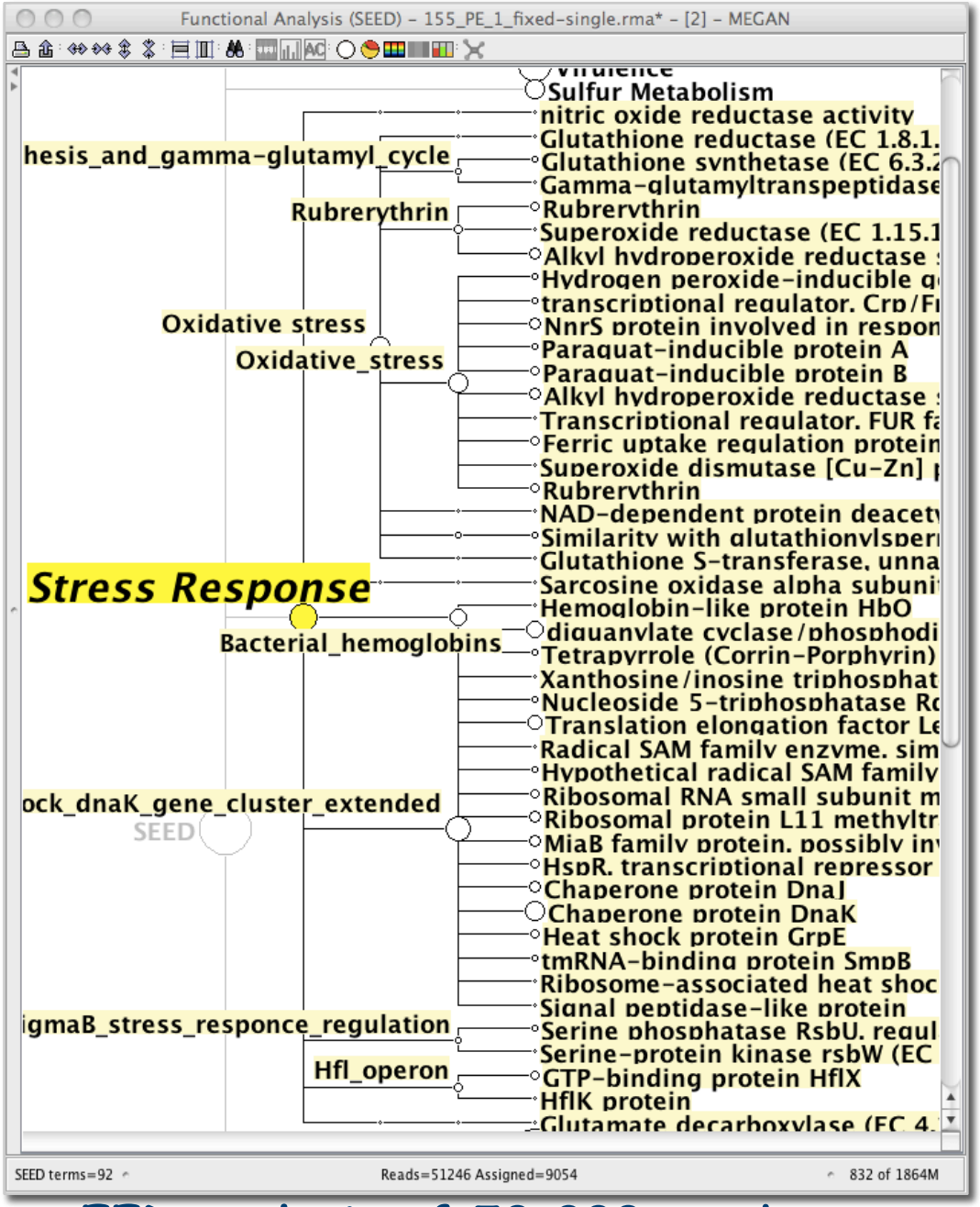
# Organize and Visualize

## Functional analysis

- ✓Use SEED classification to bin sequences by subsystems
- ✓… and by functional roles


www.theseed.org

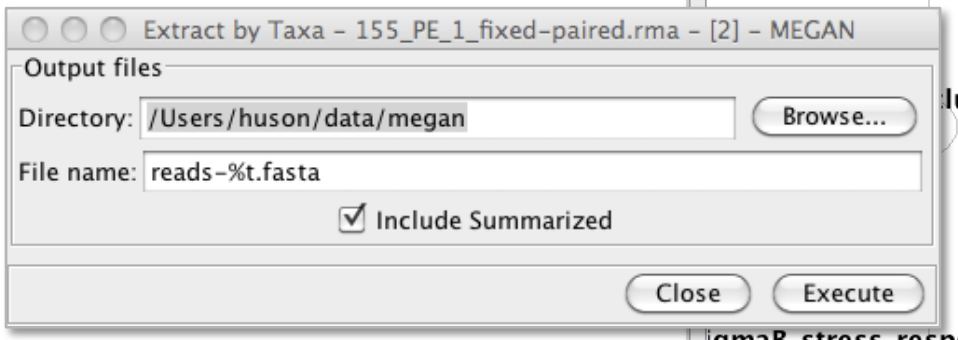
SEED: Overbeek et al., Nucleic Acids Res 33 (17), 2005

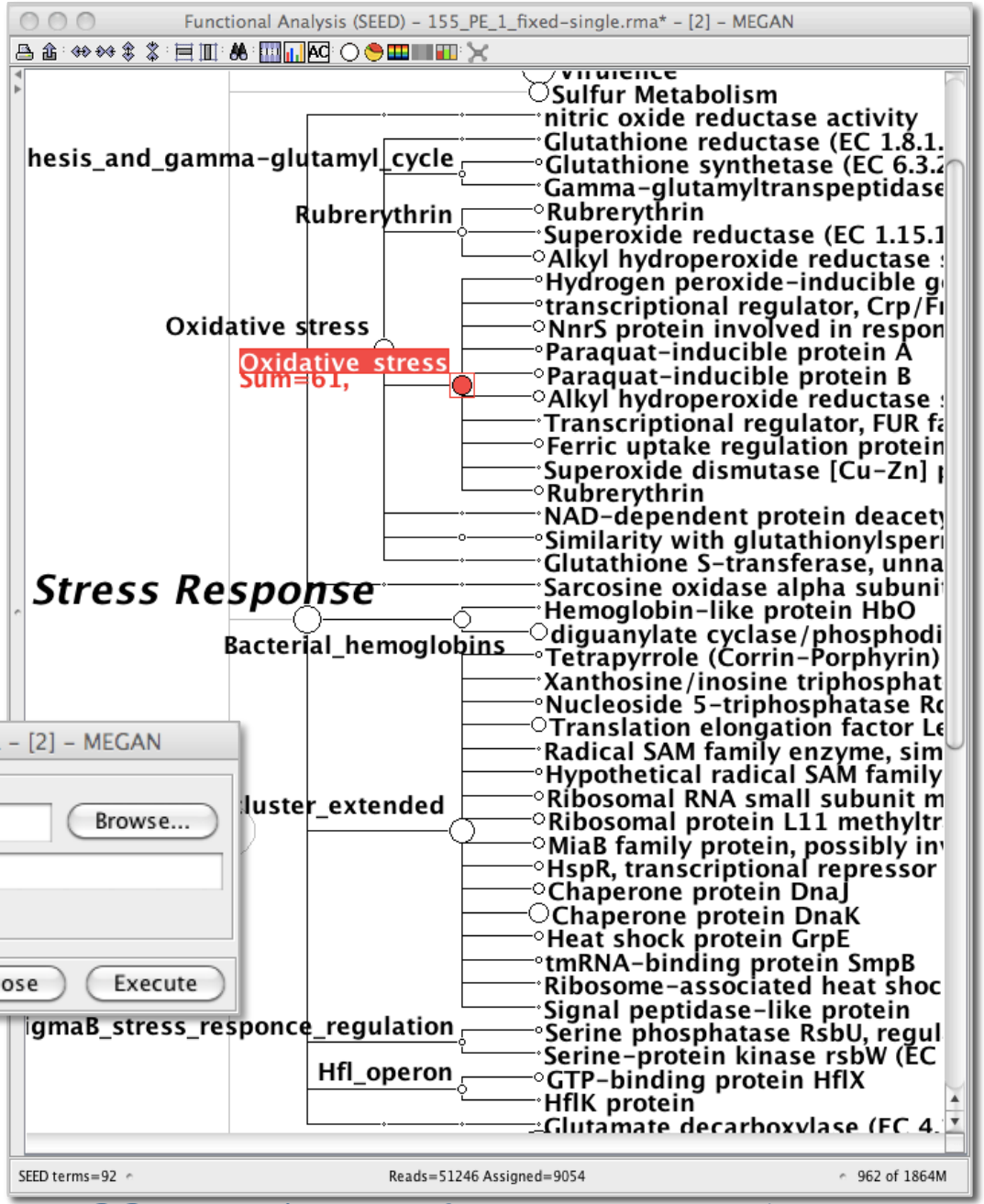



SEED analysis of 50,000 reads

# Capture

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



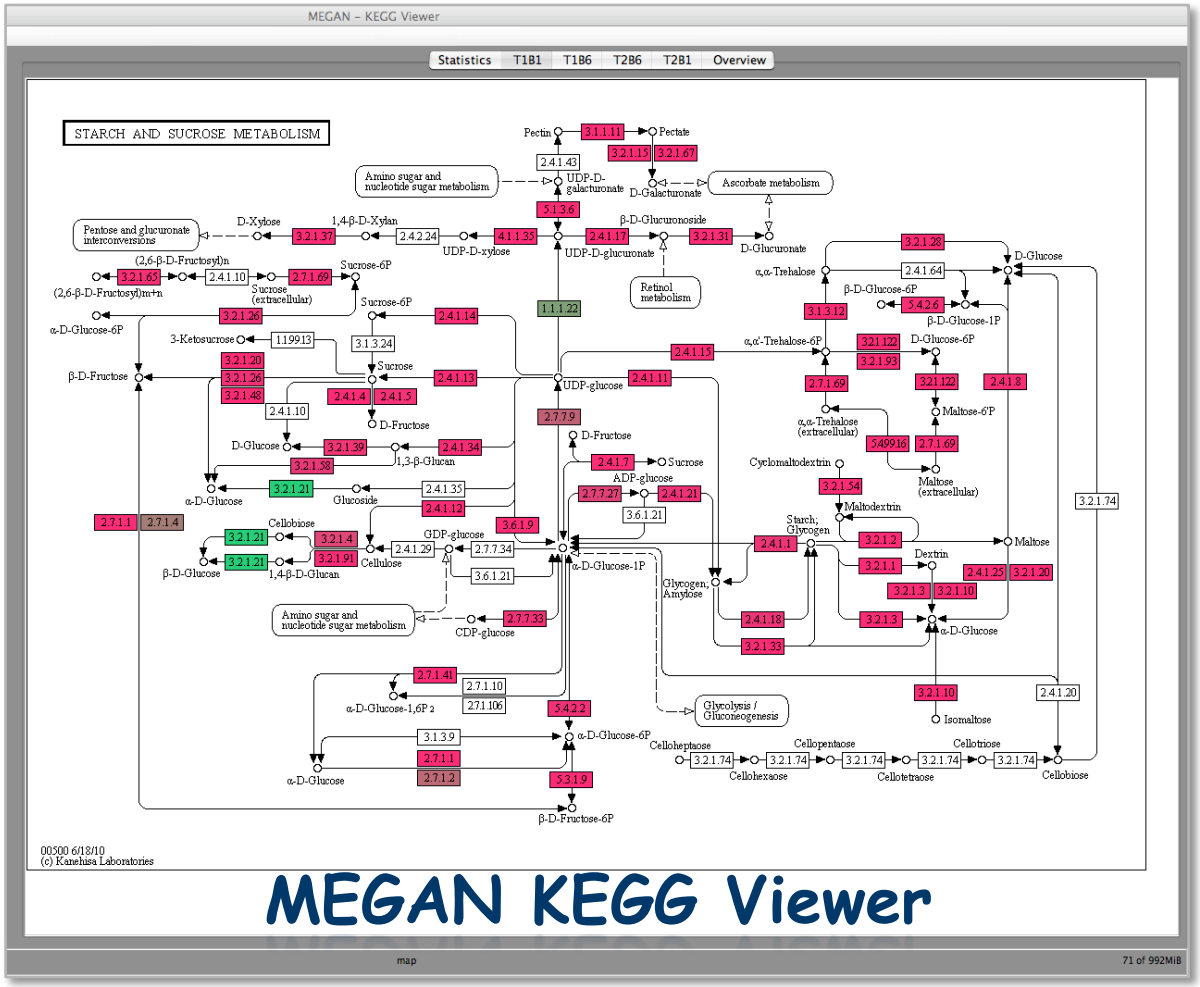

SEED analysis of 50,000 reads

# Pathway Analysis

## Functional analysis

- ✓Use KEGG pathways to bin sequences by their presence in pathways


http://www.genome.jp/kegg/

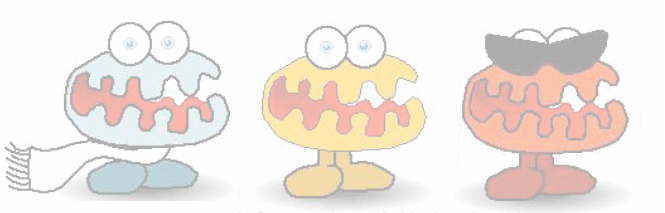



MEGAN KEGG Viewer

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

# Three Basic Computational Questions

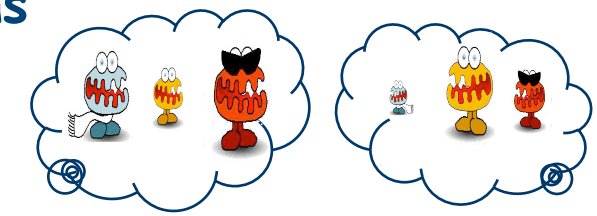
- Who is out there?
  - Types of organisms
  - In what proportions?



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

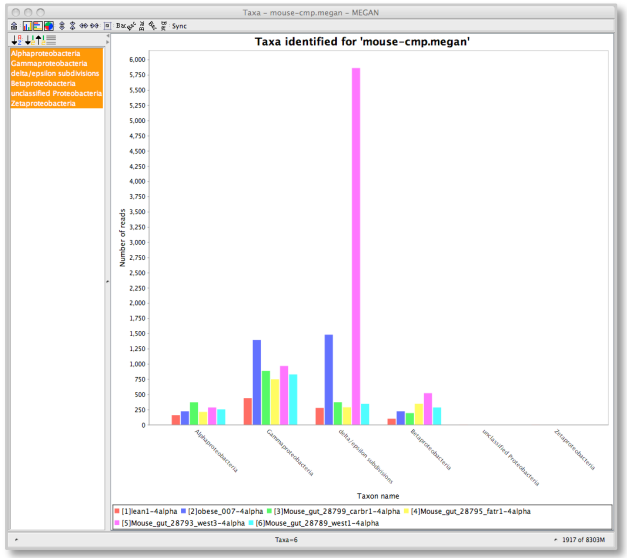
# Compare

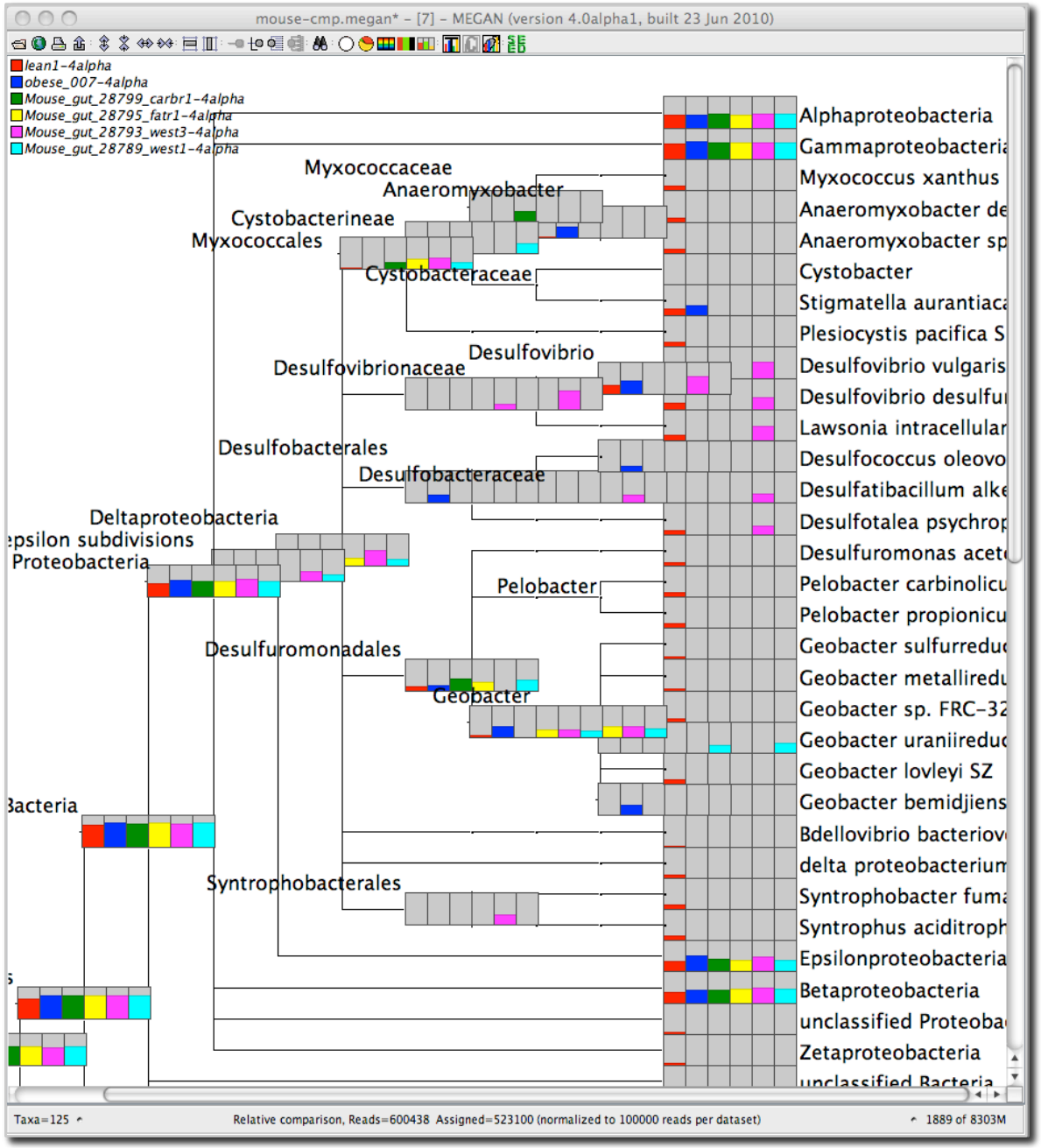
Display multiple datasets simultaneously

✓Taxonomical comparison

✓Interact

✓... and summarize

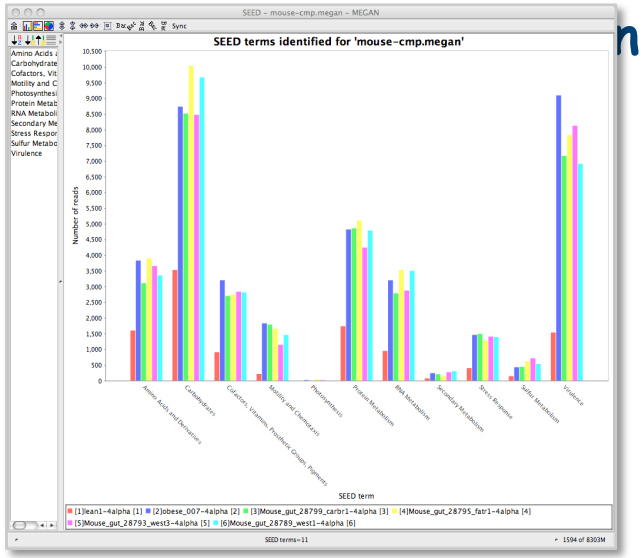


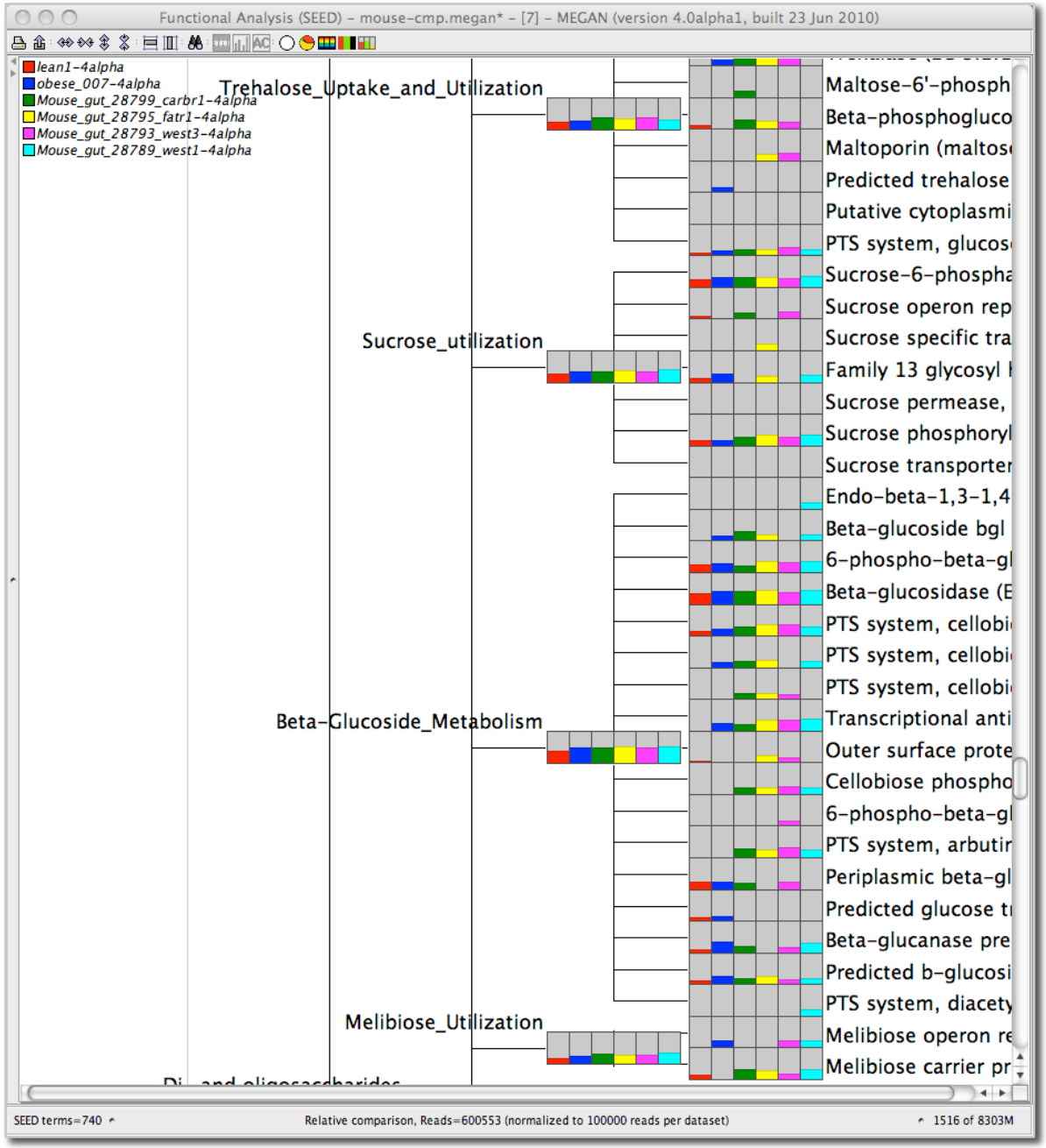



# Compare

Display multiple datasets simultaneously

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

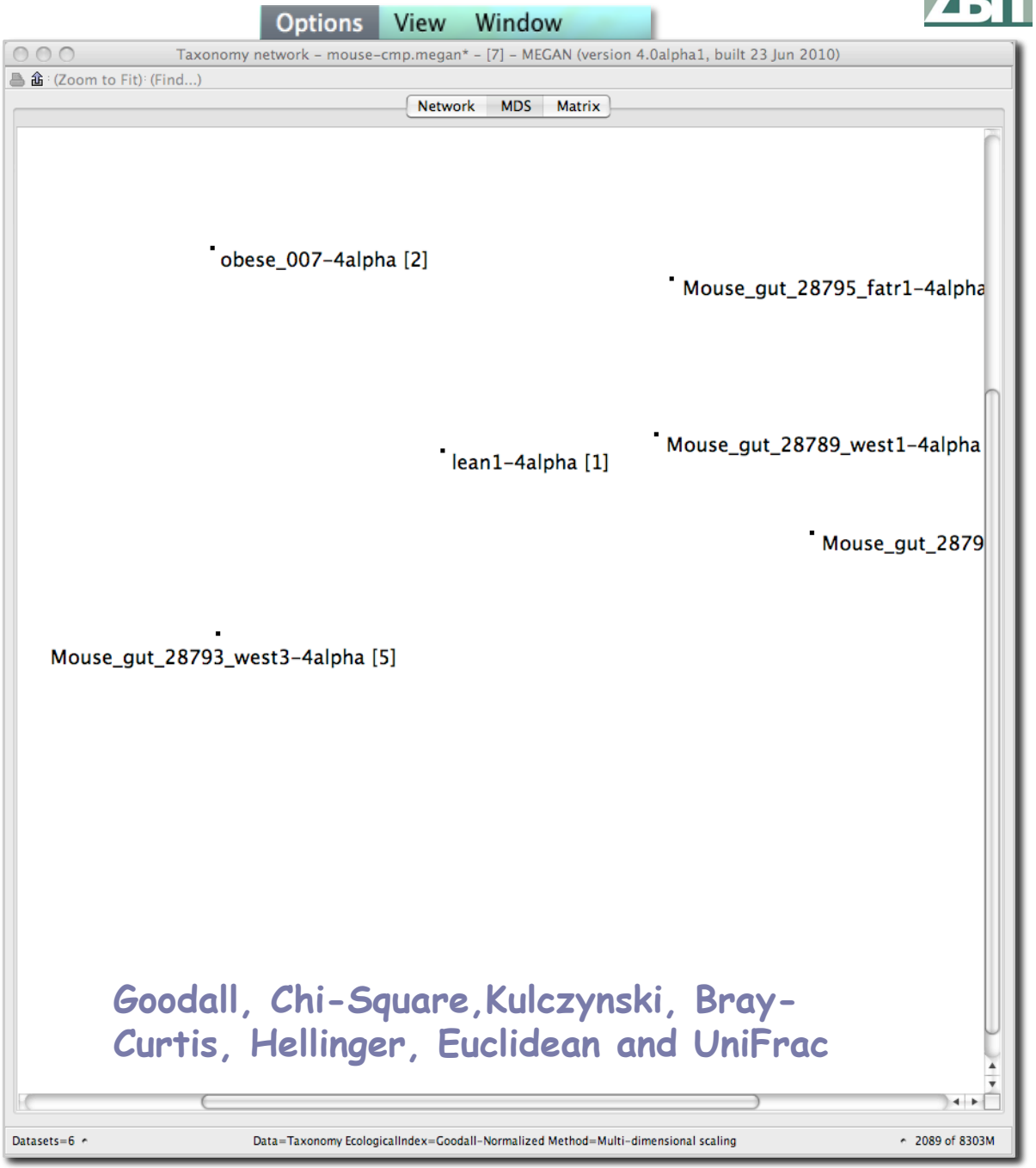


# Compare

High-level comparison:

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



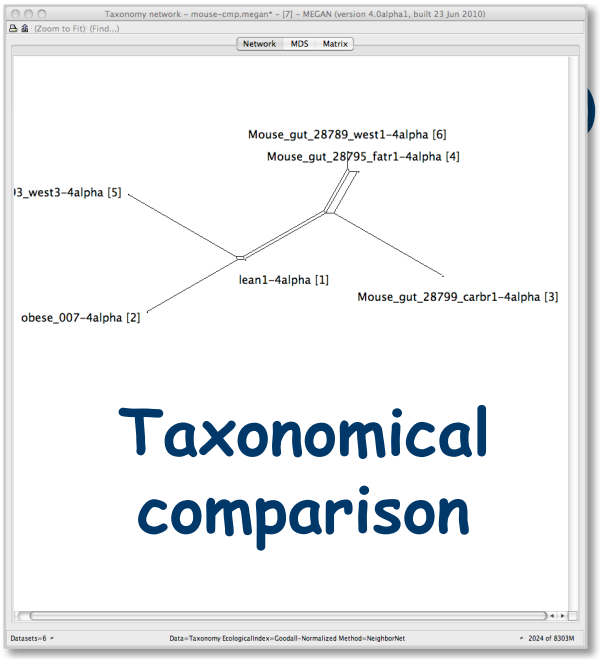

Mitra, Gilbert, Field and Huson, ISME J, 2010

Neighbor-net:
Bryant and Moulton, 2003

# Compare

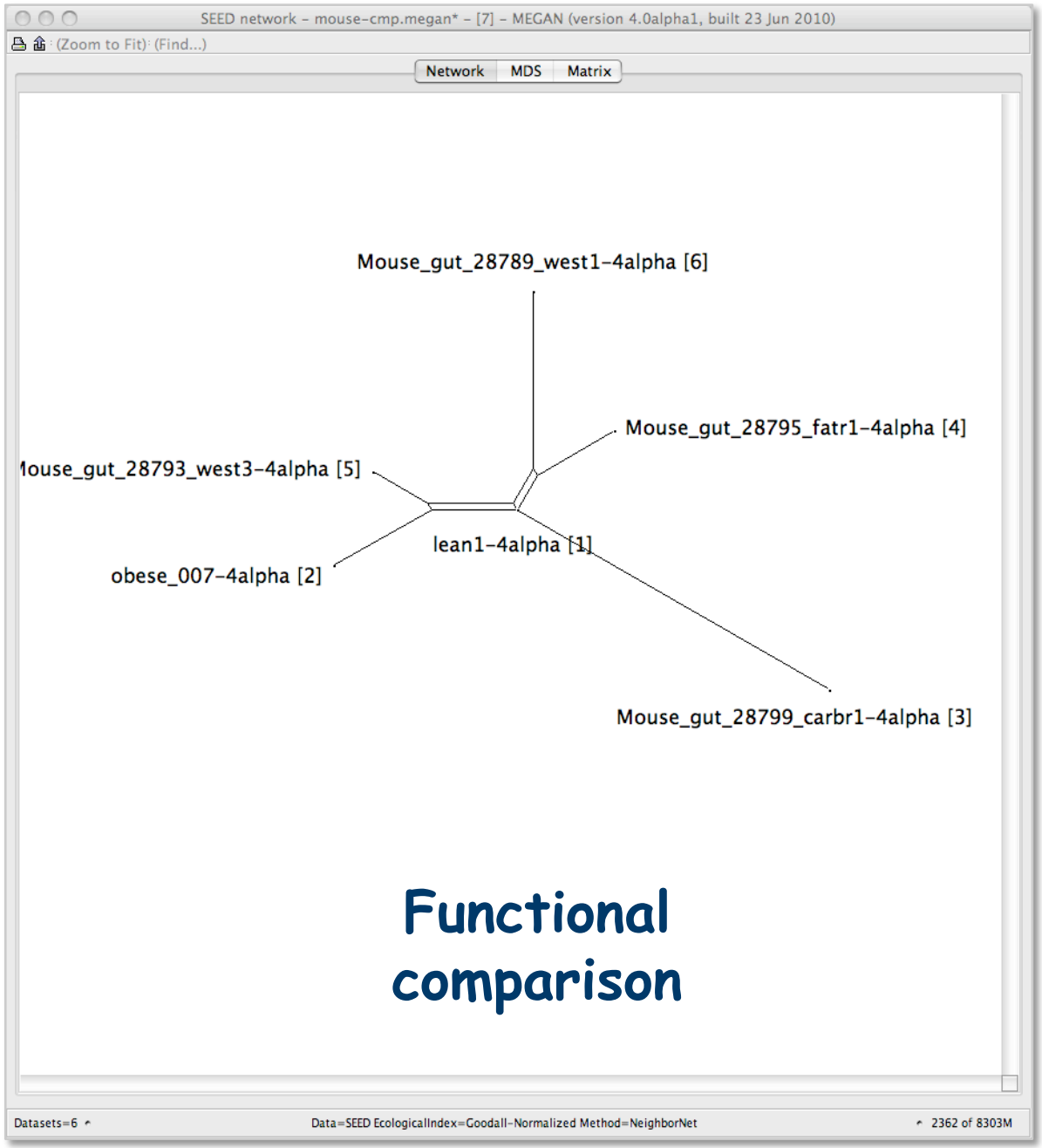
## High-level comparison:

- ✓ Select taxa




**Taxonomical comparison**

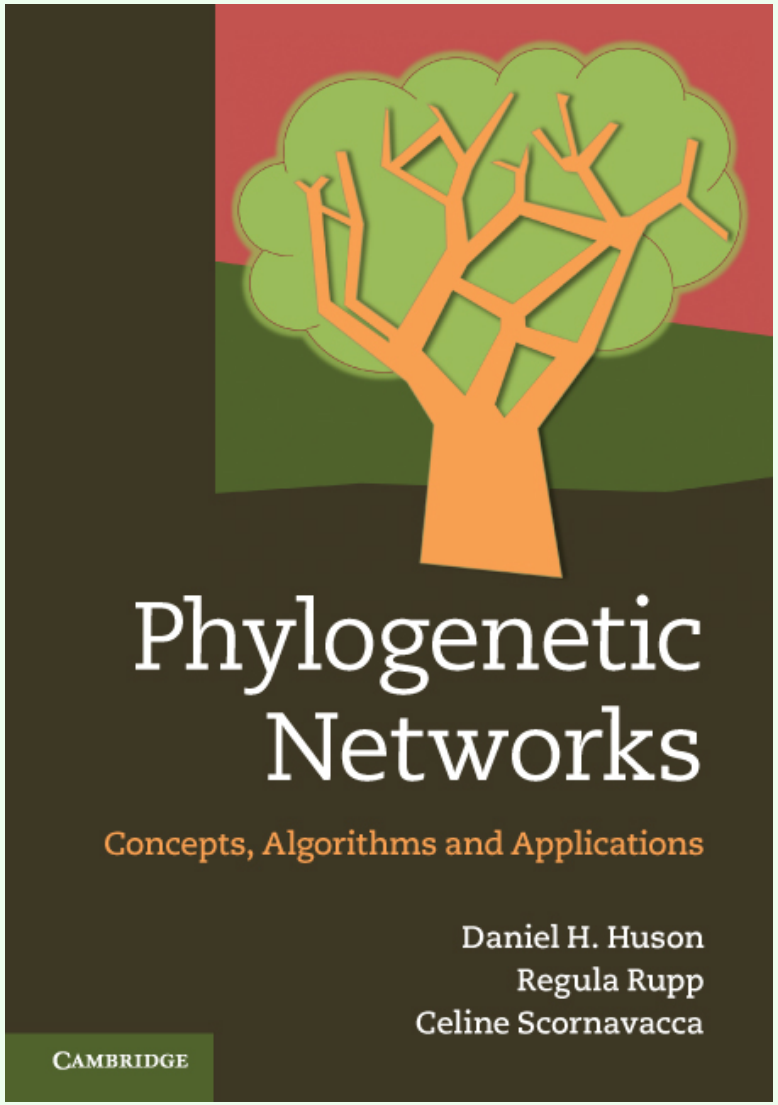
- ✓ Select functions…




**Functional comparison**

# Phylogenetic Networks

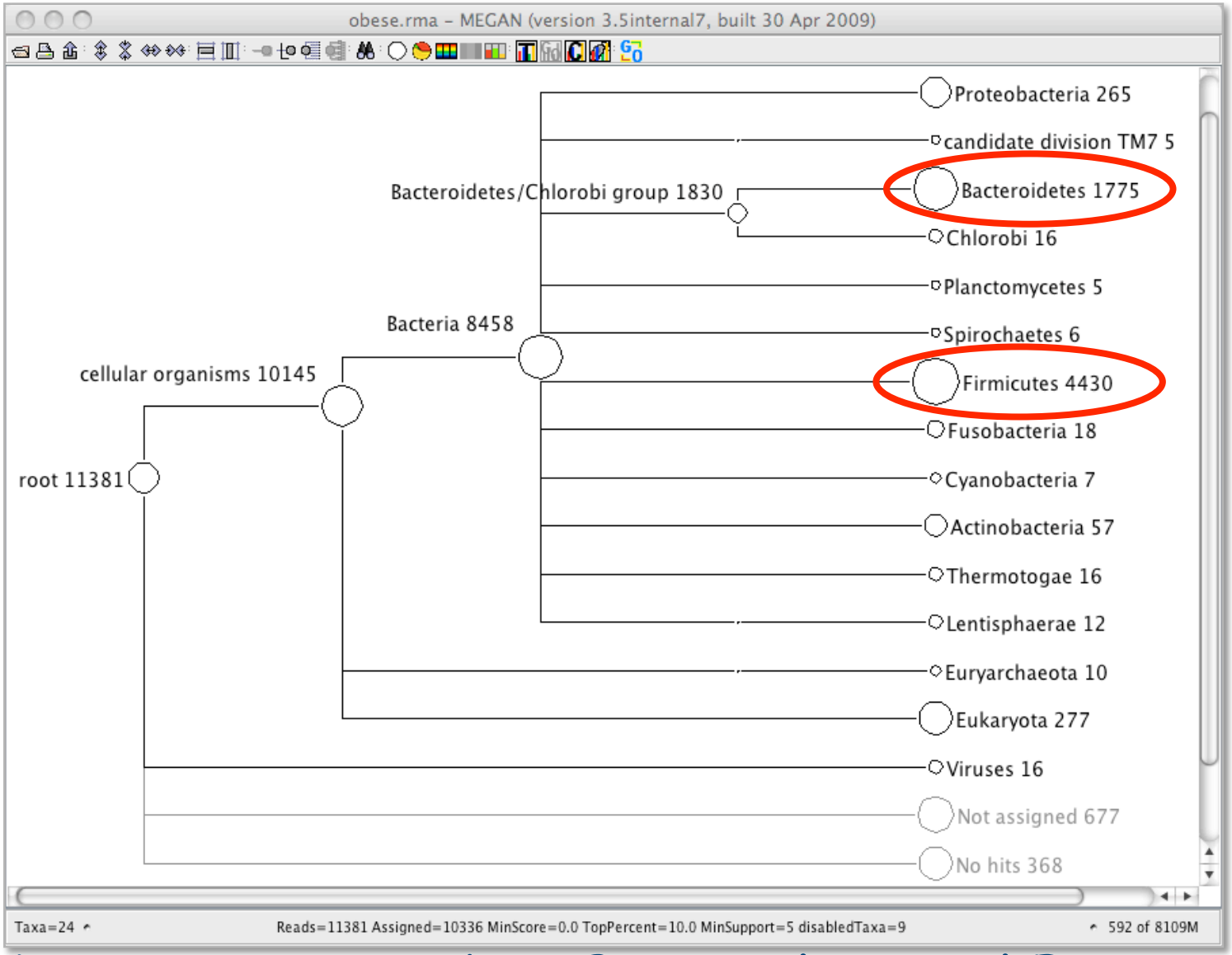
## Concepts, Algorithms and Applications




- ~ 360 pages
  - 55 lemmas
  - 20 theorems
  - 50 algorithms
  - 85 exercises
  - 15 applications
  - 190 figures
- ~ 40 EUR

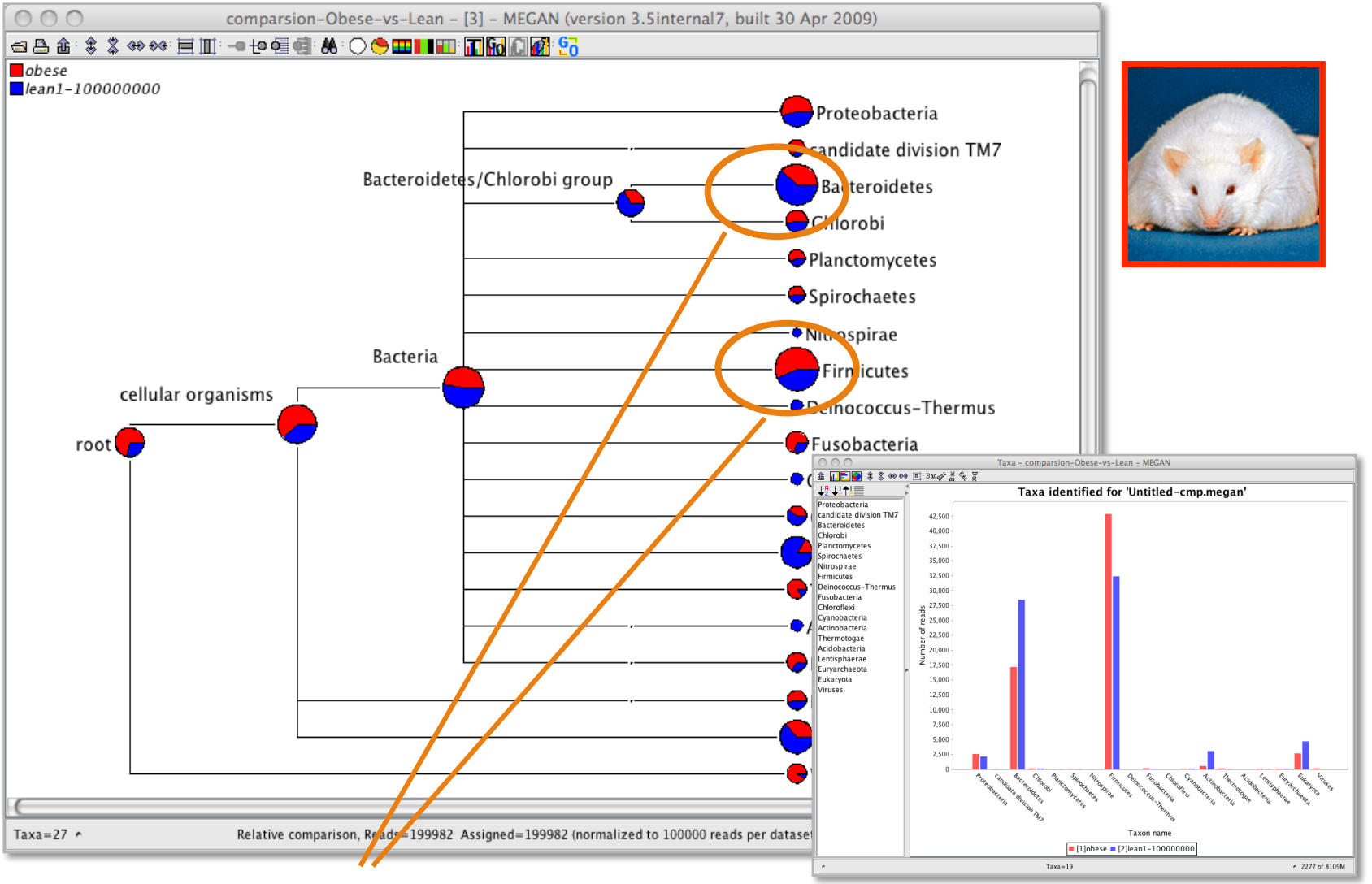
Dec 2010

# Example: Mouse Gut Microbiome




Dominant gut microbes: Bacteroidetes and Firmicutes

# Comparative Analysis






Change in proportions of these two phyla

Eberhard Karls Universität Tübingen

ZBIT

## Basic Principles

# How to analyze a meta genome?

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

# Contents

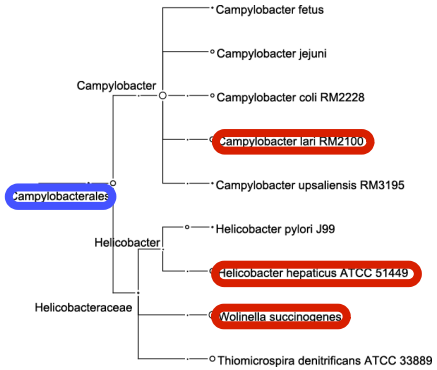
- Genomics
- Sequencing
- Metagenomics
- **More computational questions**
- Outlook

# Single Reads vs Paired Reads

In metagenomics:

- Use single reads or paired reads?
- In the latter case, short clones or long clones?

# Taxonomical Analysis Based on Gene Content

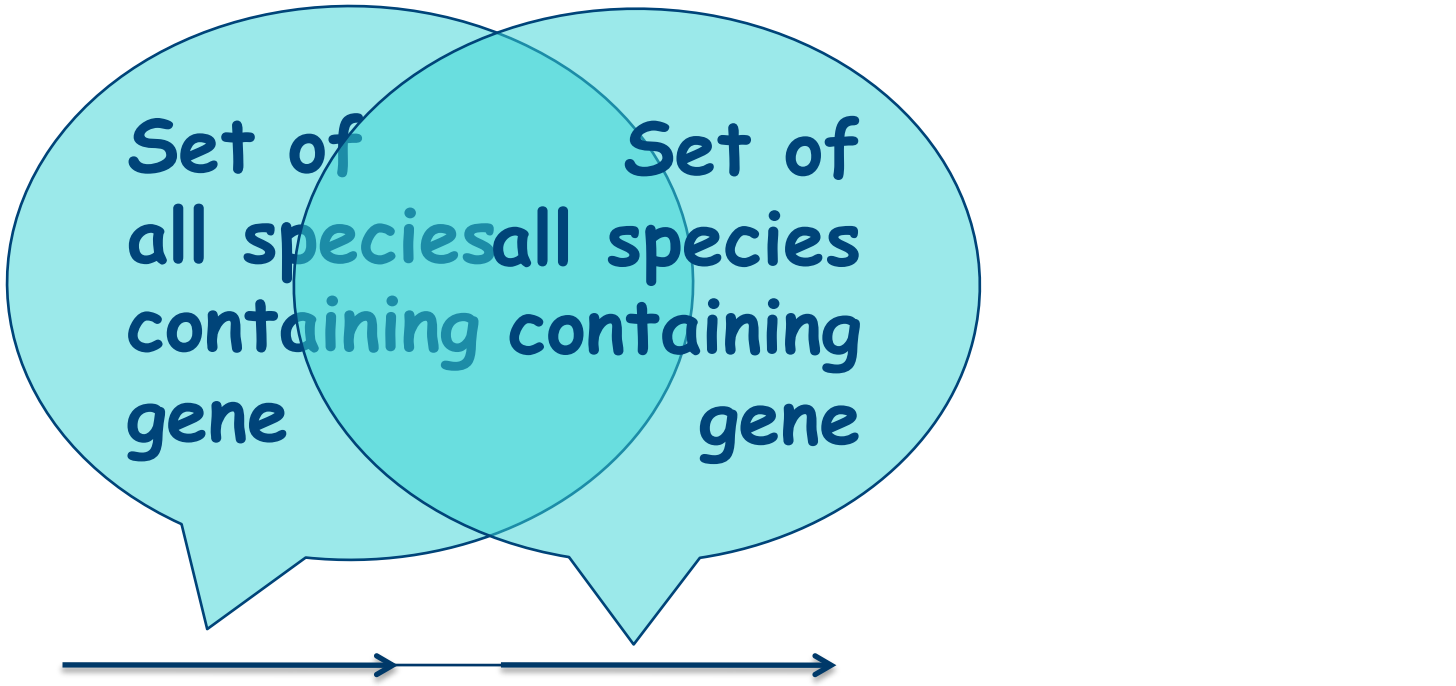




Unknown source genome

## Use LCA to assign to (higher-rank) taxon

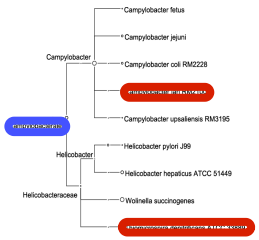
# Taxonomical Analysis Based on Gene Content

## Short clones or long clones?



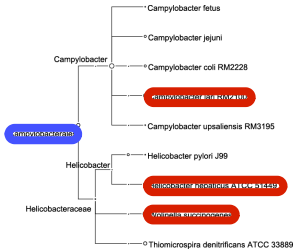
# Taxonomical Analysis Based on Gene Content

## Claim: long clones are more specific

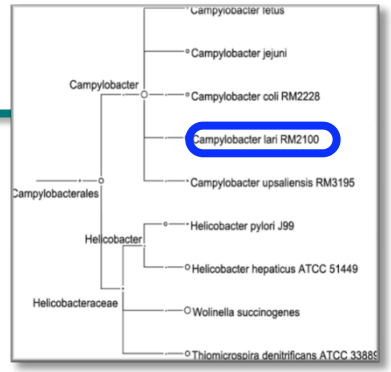




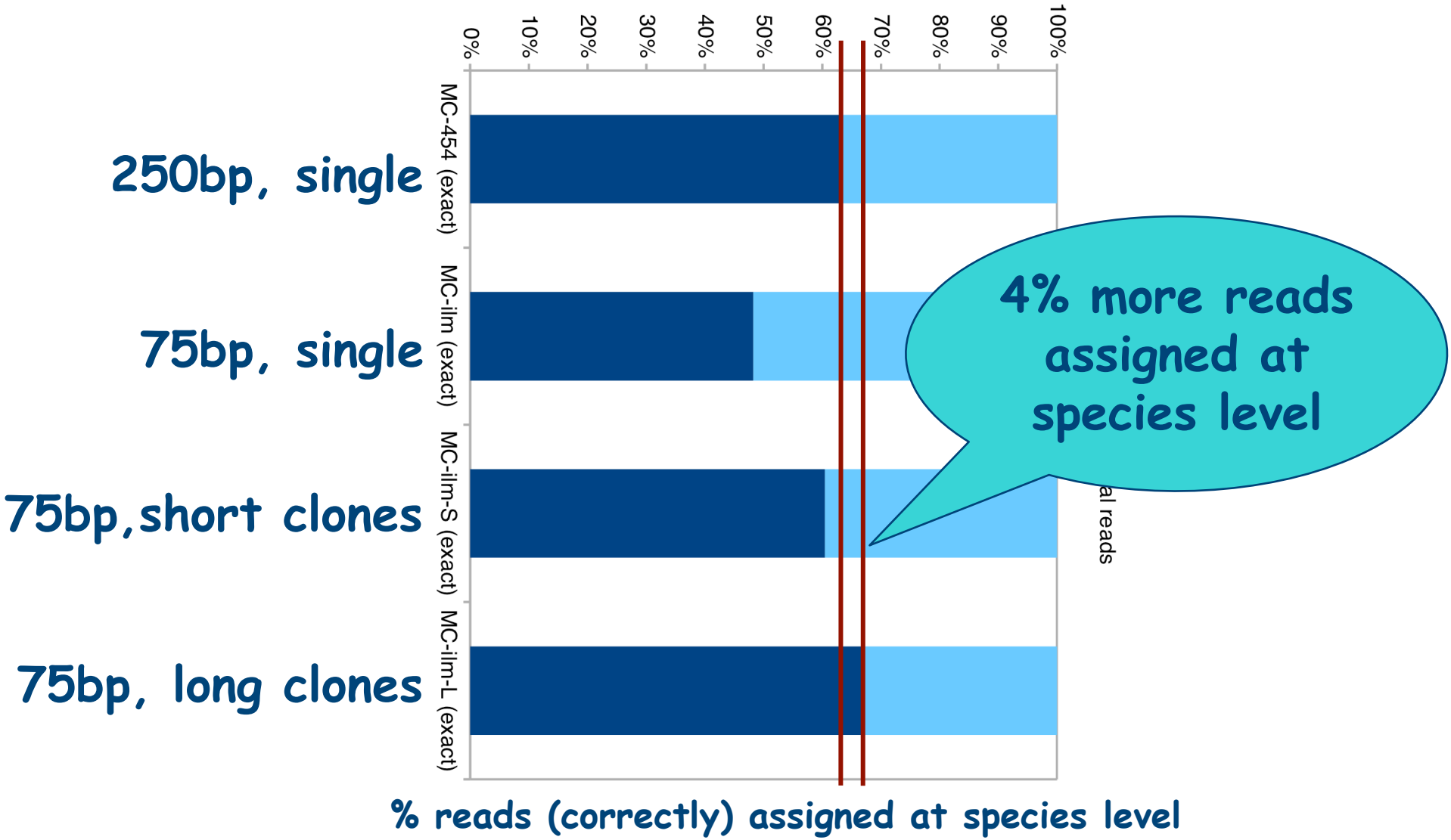





## LCA of intersection of species:



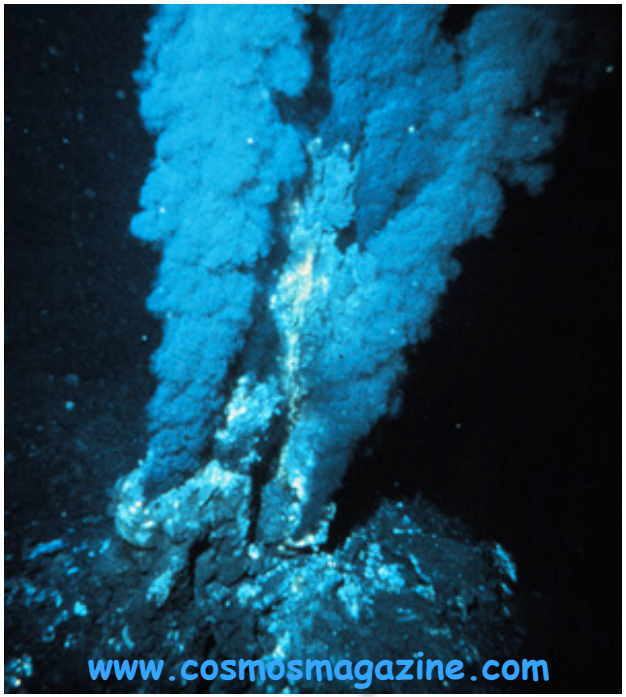
# Simulated Performance 454 *vs* Illumina




MetaSim – simulator (Richter et al, 2008)

# Ocean Seabed Sample

- Joint work with Ida Steen (Bergen)



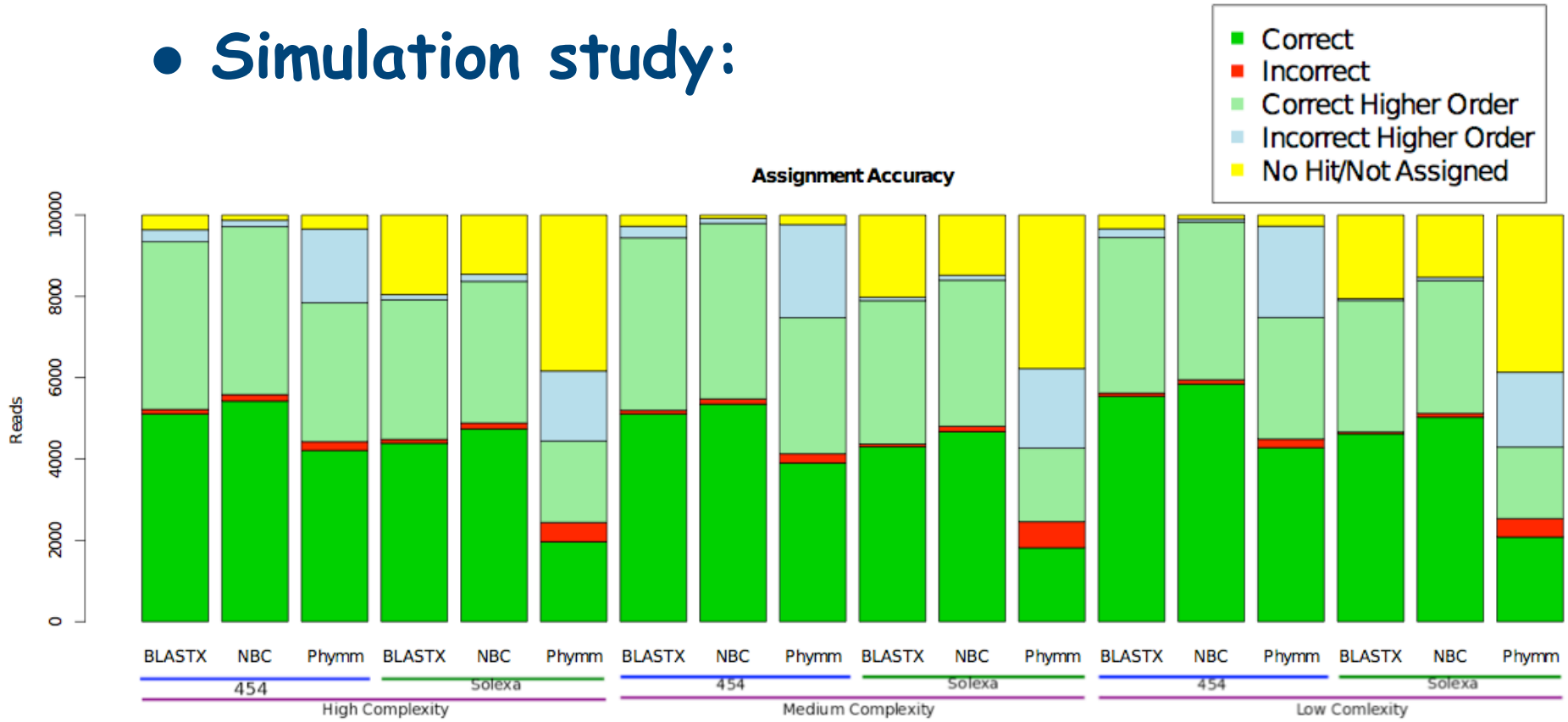

- Currently analyzing large set of 454 paired reads (7kb clones) collected from an Arctic hot vent at 3km depth

# Hybrid Approach

- Machine-learning based classification approaches are much faster than BLASTX
- Biologists want to see alignments and they are needed for functional analysis
- Hybrid approach:
  - Use taxonomic classifier to perform taxonomic binning
  - BLASTX reads only against assigned taxa
- Study of NBC (Rosen *et al* 2008) and Phymm (Brady & Salzberg, 2009)

# Hybrid Approach

- Simulation study:

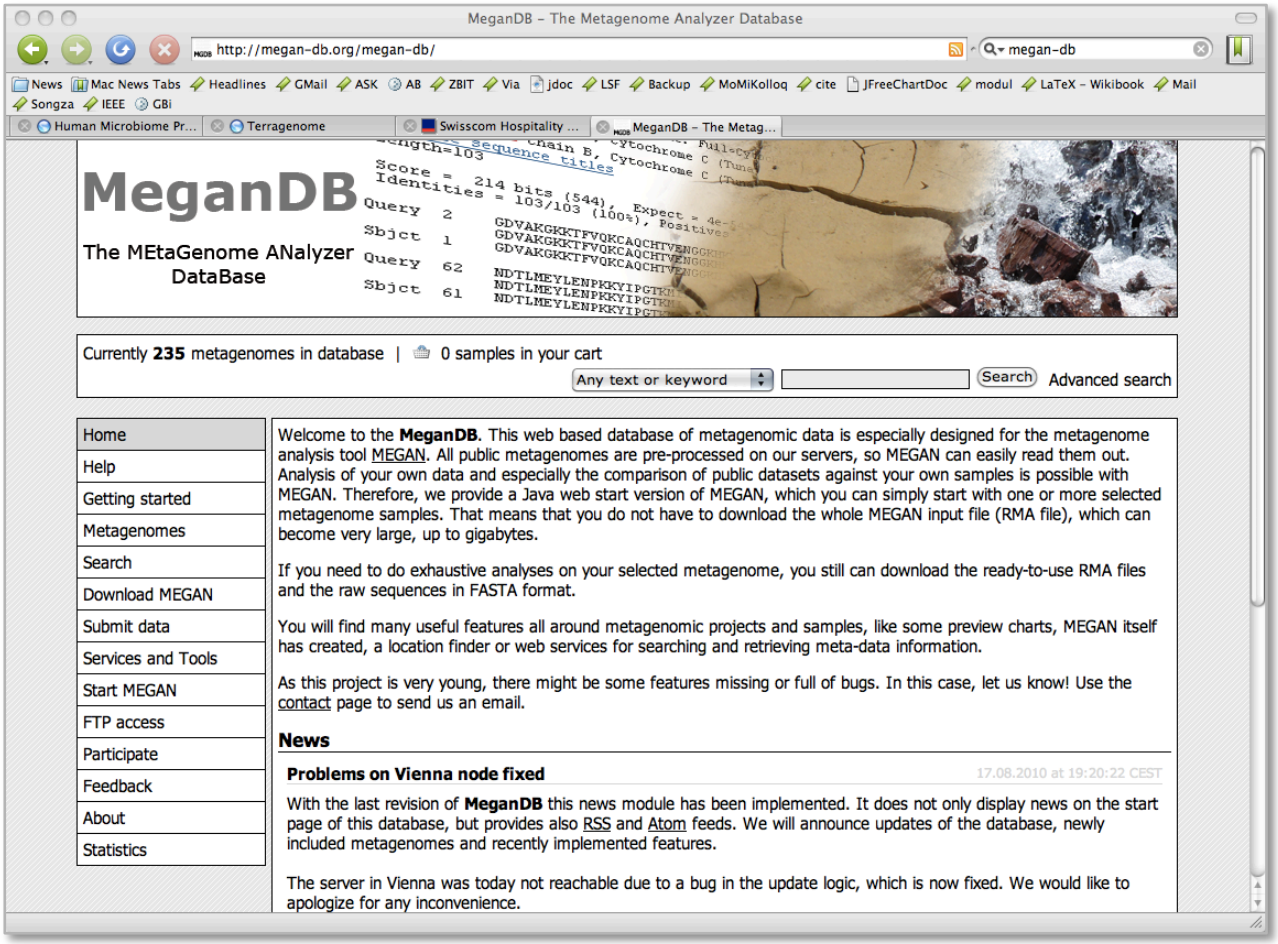



- ~10x speed-up over full BLASTX
- Increased accuracy using NCB
- Decreased accuracy using Phymm

Weber et al, submitted

# MEGAN-DB

- Database of precomputed MEGAN files




- www.megan-db.org

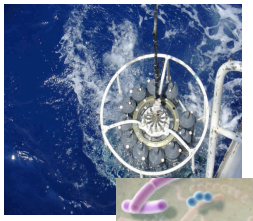
Joint work with Thomas Rattei and Simon Domke

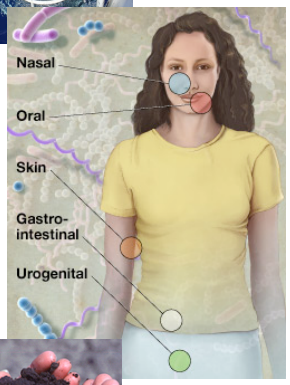
# Contents

- Genomics
- Sequencing
- Metagenomics
- Computational questions
- **Outlook**

# Computational Challenges…

- Global Ocean Sampling
  - www.jcvi.org/cms/research/projects/gos
- Human Microbiome Project
  - nihroadmap.nih.gov/hmp/
- Terragenome Consortium






Terabases of sequences

- Survey of the Earth microbiome

- Petabases of sequences
- Processing and storage of exabytes of data

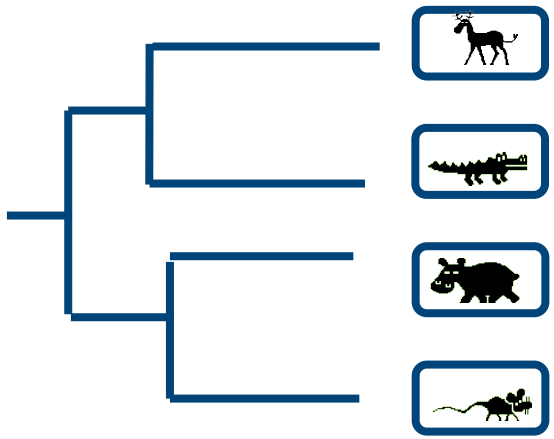
(mega, giga, tera, peta, exa…)

# Computational Challenges…

- Data storage and access
- Tools for navigating metagenome data and metadata
- Ever faster analysis methods
- How to learn across multiple datasets?
- How to build a model of the Earth microbiome?

# Joint Work With:

- Tübingen: Suparna Mitra, Daniel Richter, Nico Weber & Max Schubach
- Penn State: Stephan Schuster and Qi Ji
- Vienna: Tim Urich, Christa Schleper, Thomas Rattei, Simon Domke
- Bergen: Ida Steen and Anders Lanzen



www-ab.informatik.uni-tuebingen.de