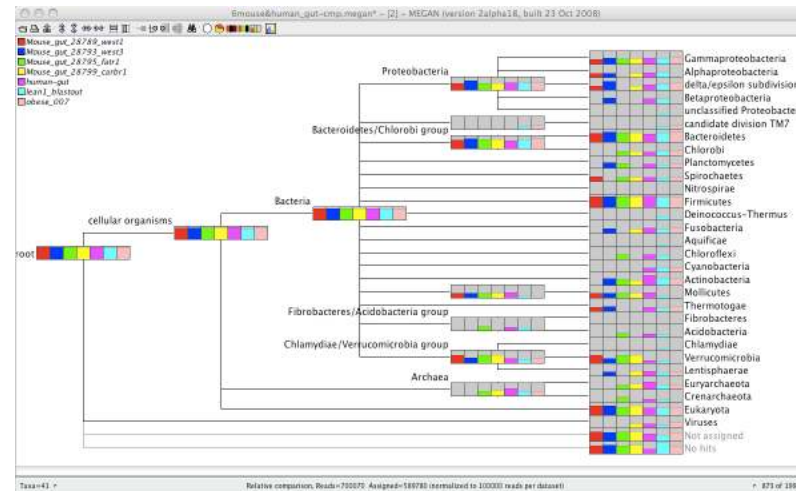


Computational Analysis of Metagenomes



Daniel H. Huson





Contents

Genomics

Sequencing

Metagenomics

Computational questions

Outlook



Genomics

Sequencing

Metagenomics

Computational questions

Outlook



Discovery of DNA

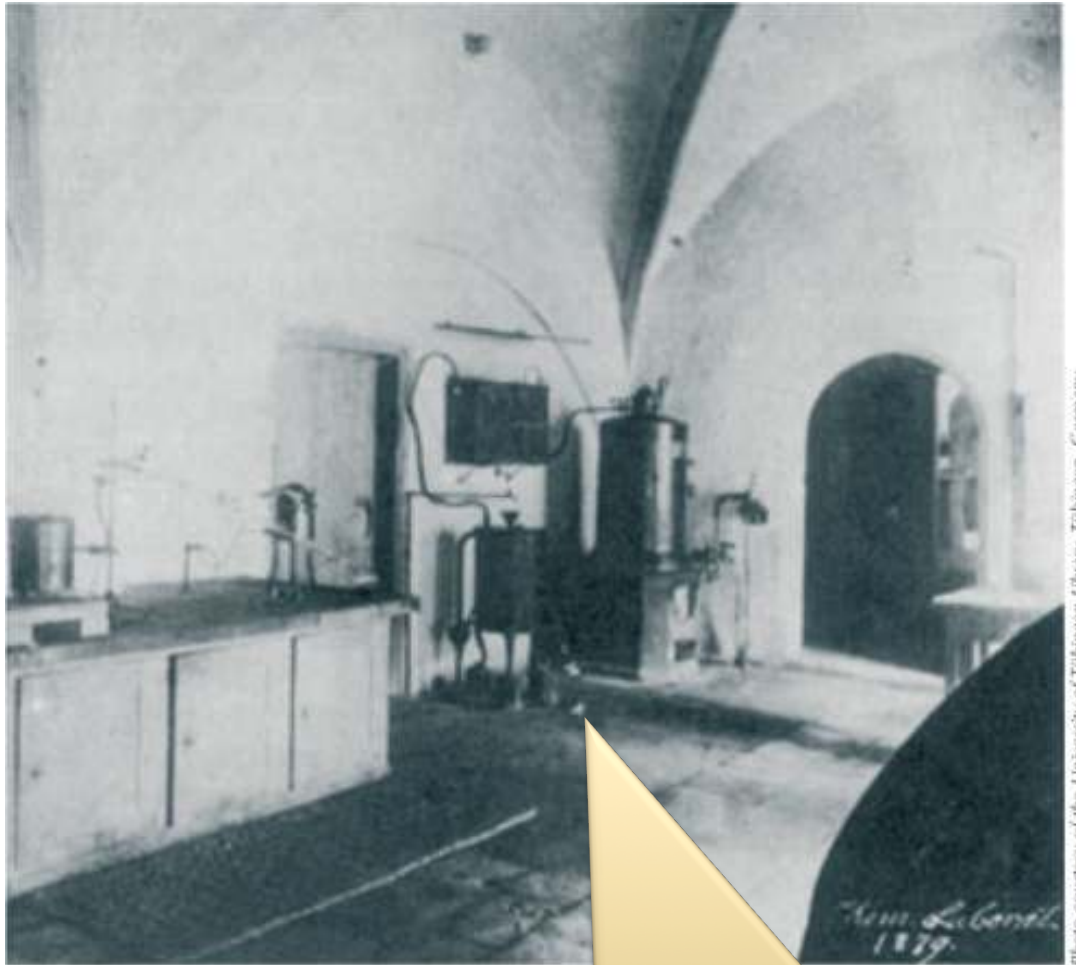
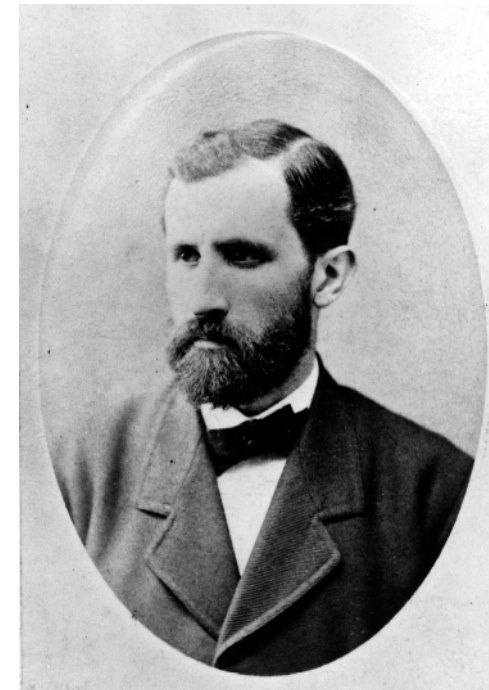


Photo courtesy of the University of Tübingen Library, Tübingen, Germany

1869: Miescher discovered DNA in the kitchen of Tübingen Castle



**Friedrich Miescher
(1844-1895)**



Genomics

Genomics is the study of the genome sequence of individual organisms



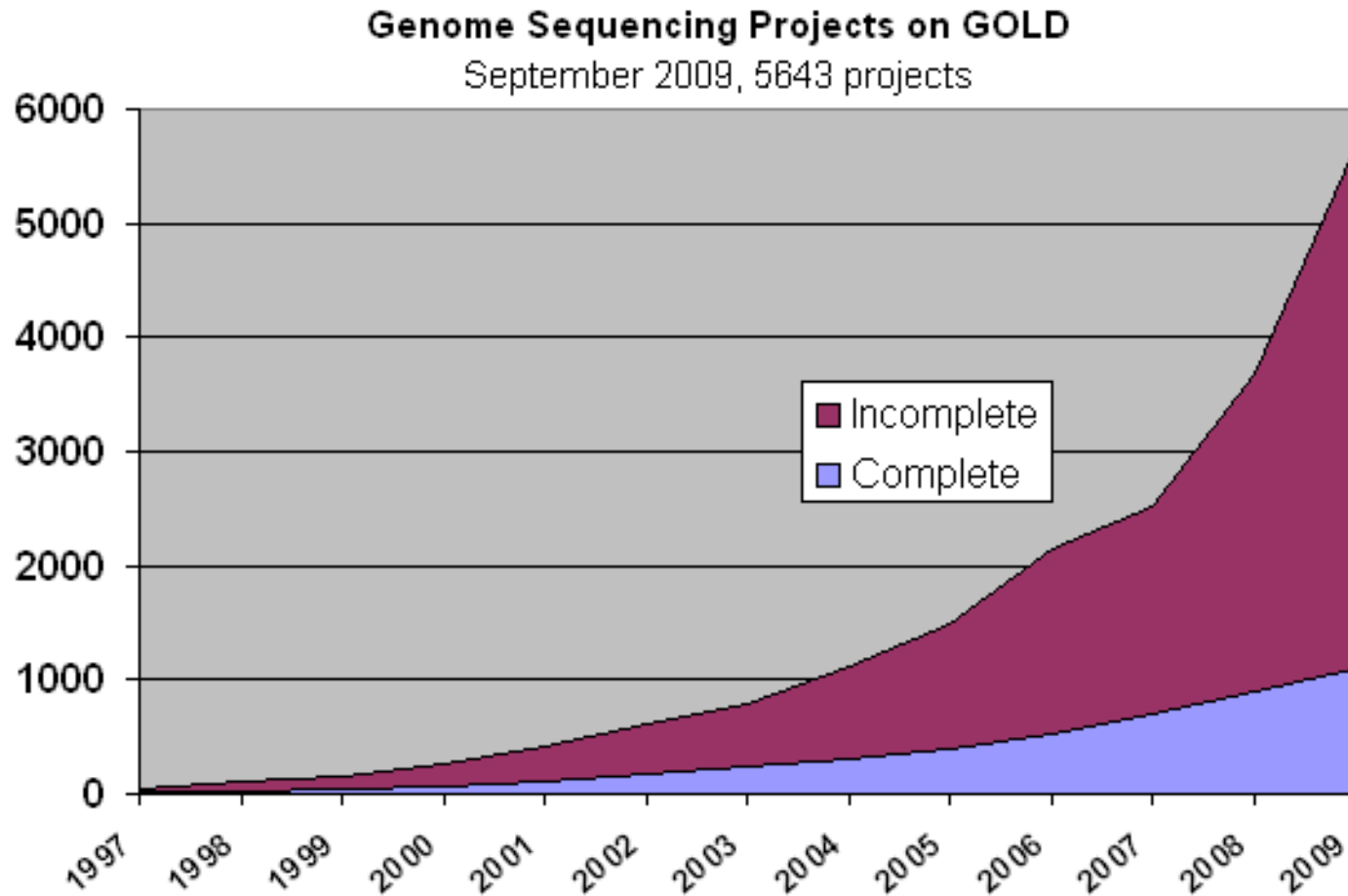
Sequencing and assembly...

Annotation and analysis...

Genome comparison...



Sequencing of Genomes

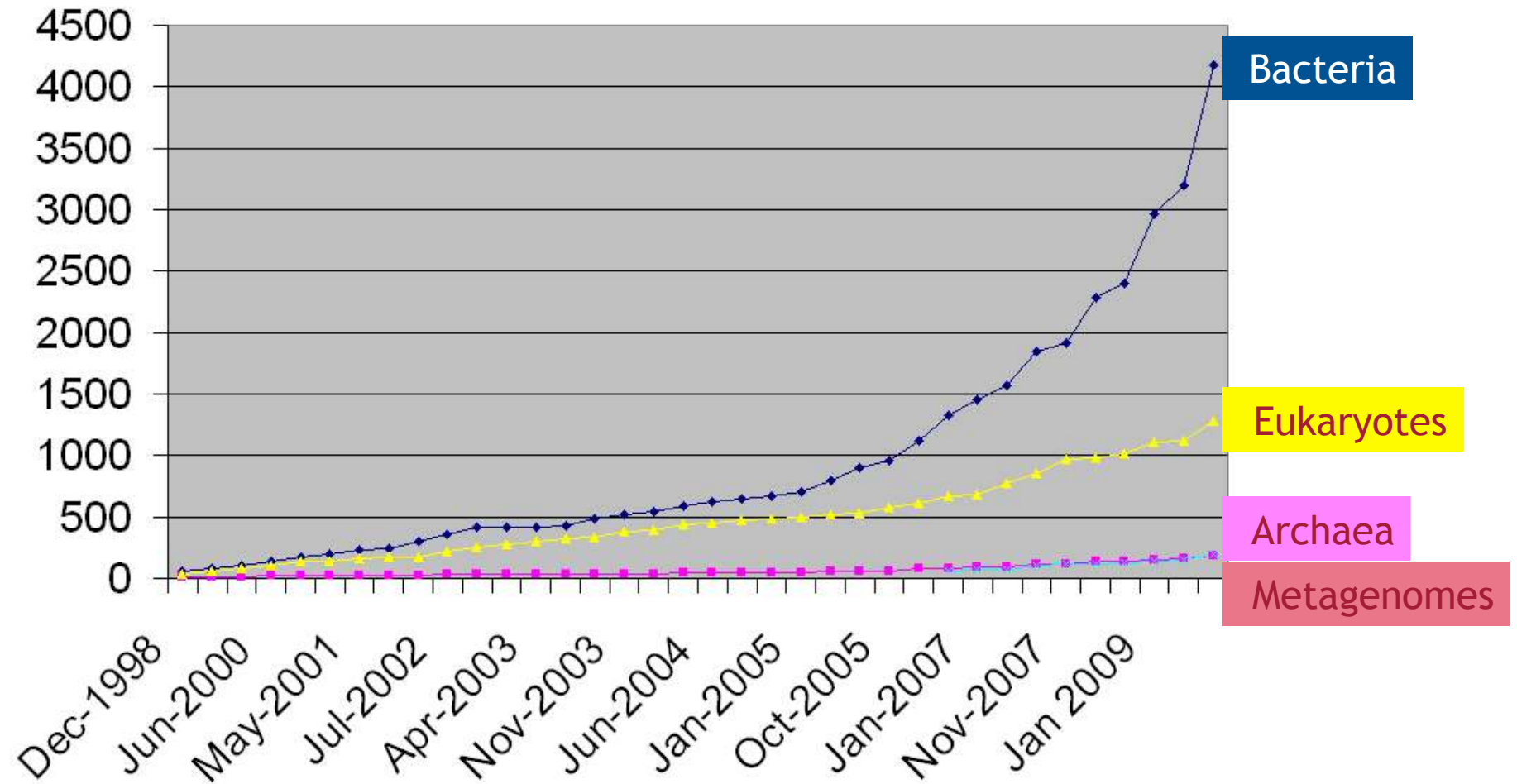


GOLD: Genomes online database
www.genomesonline.org



Genome Projects by Groups

Genome Projects on GOLD according to Phylogenetic Groups ©
September 2009 - 5831 Projects

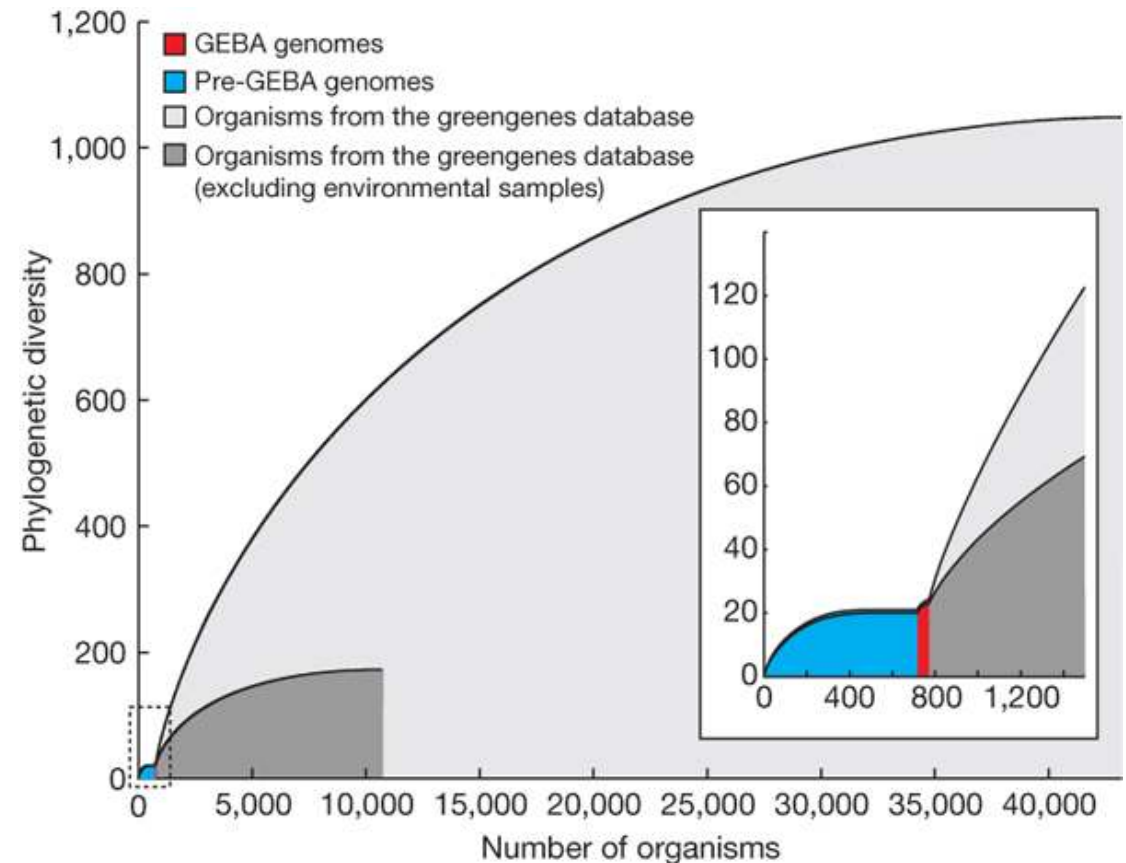




The GEBA Project

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/>

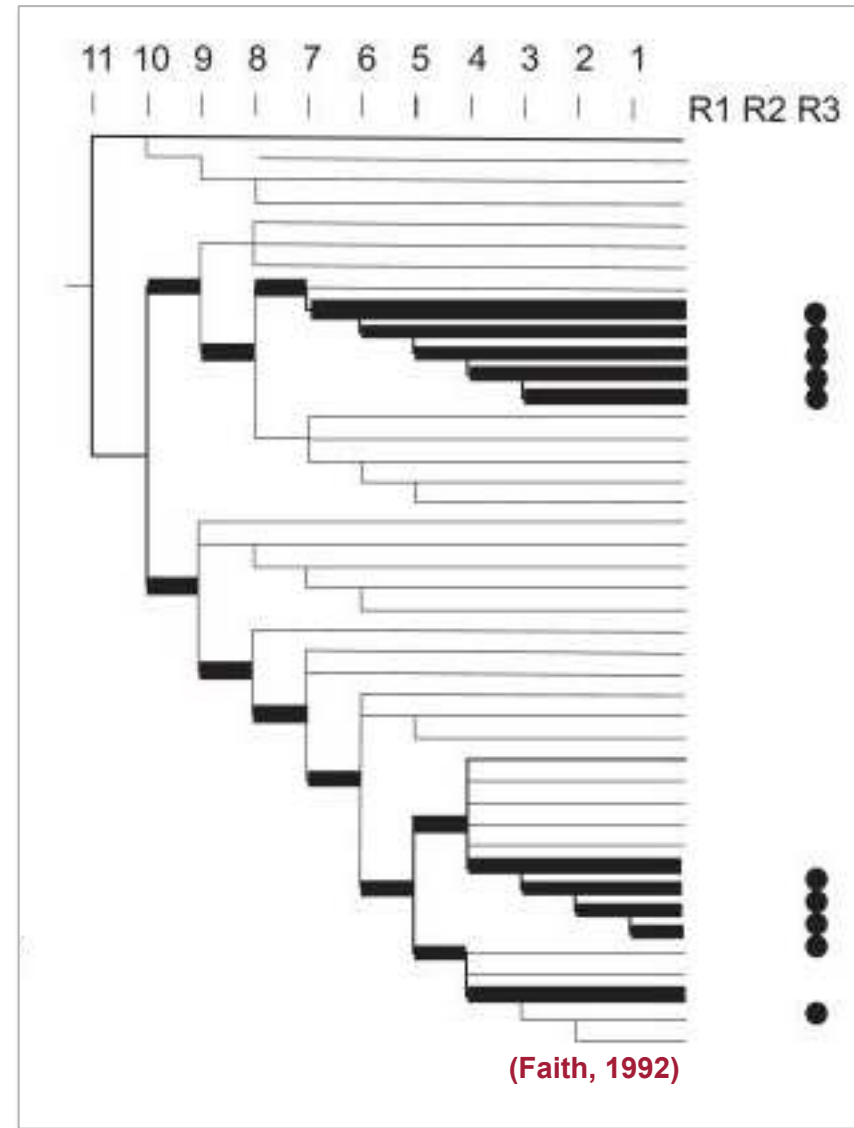


Phylogenetic Diversity (PD)

PD of a set of taxa:

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

(Faith, 1992)





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:

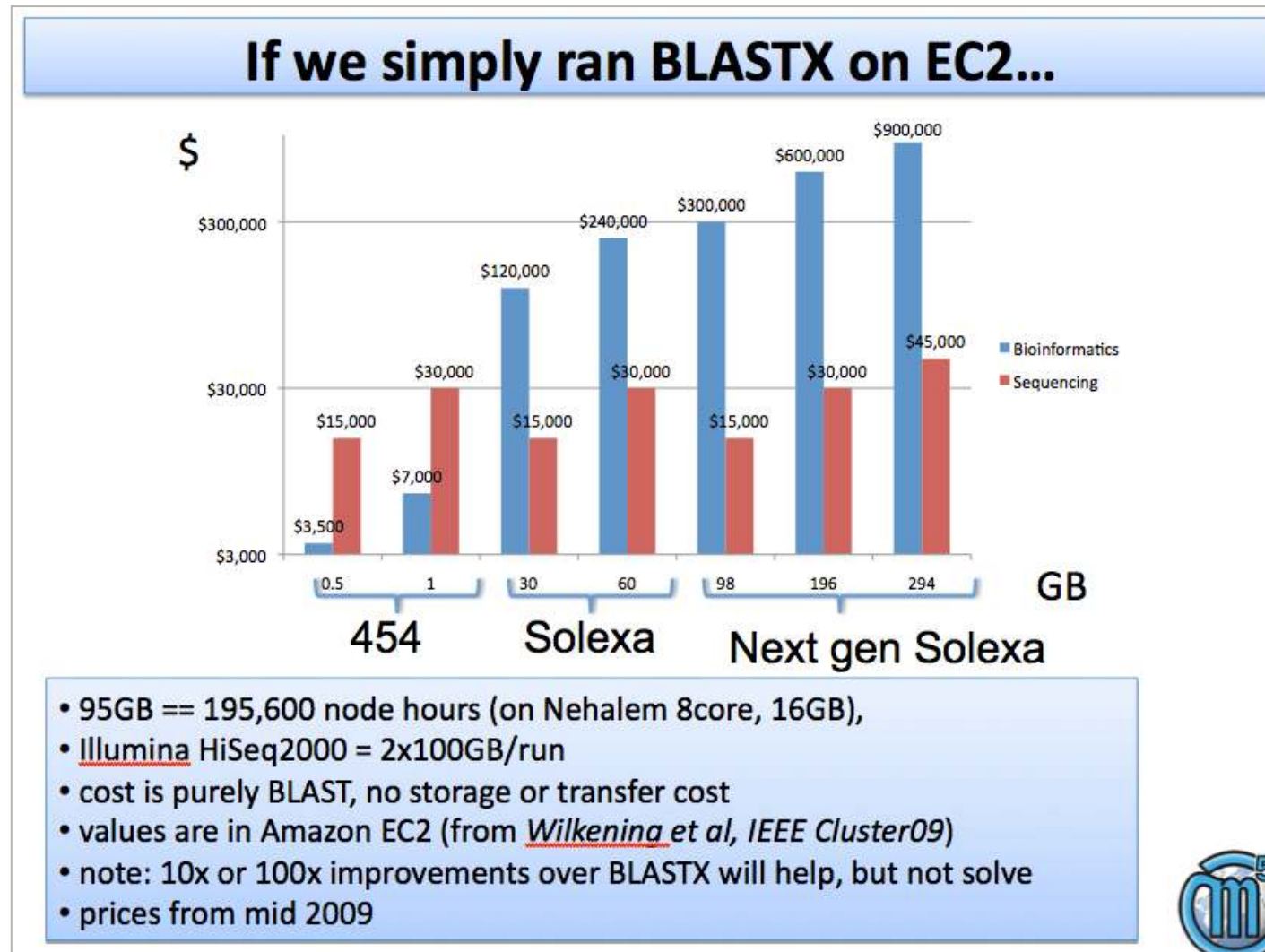
PacBio **SMRT**: 25Gb/run, >1000bp, ?\$/Mb

Other:

Ion Torrent: uses ion sensor, 100bp per run, 10x cheaper than 454



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>

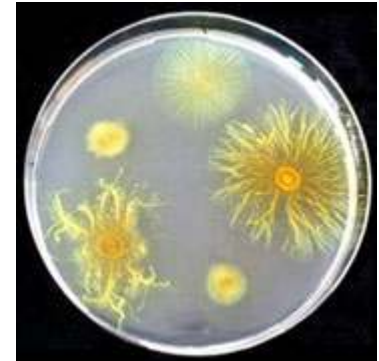
Sequences for ~200,000



Metagenomics

“The study of the DNA of uncultured organisms”

> 99% of all microbes cannot be cultured



www.innovations-report.de

A **genome**:

Entire genetic information of a single organism

A **metagenome**:

Entire genetic information of a community of organisms

A **metatranscriptome**:

Entire collection of transcripts of a community at a given time



Typical Sources of Metagenomes

Soil samples



Sea water samples



Seabed samples

Air samples

Medical samples

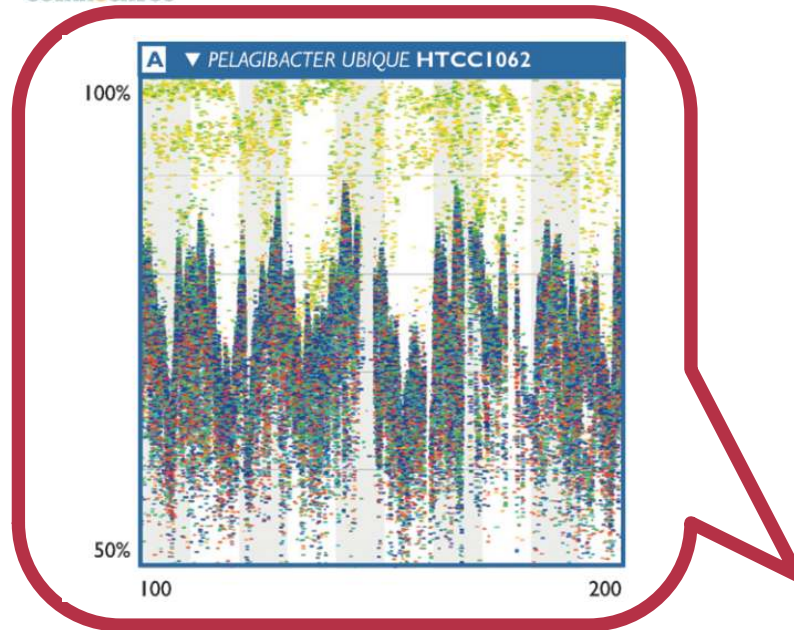


Ancient bones

Human microbiome



Global Ocean Sampling Expedition



Rusch *et al.* (2007):

41 samples

Size filtered 0.1-0.8 μ m

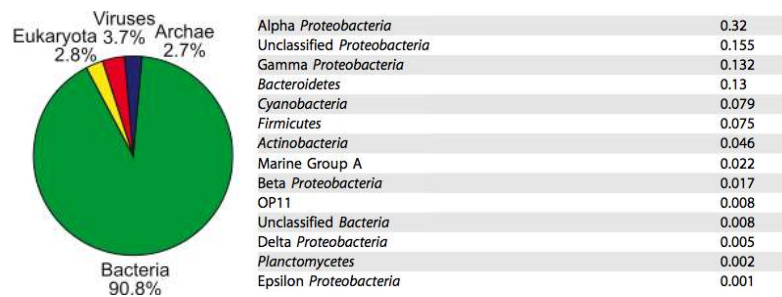
Sanger sequencing

7.7 million reads

length ~822bp

~ 5.9Gb sequence

Low abundance of *clonal* organisms



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

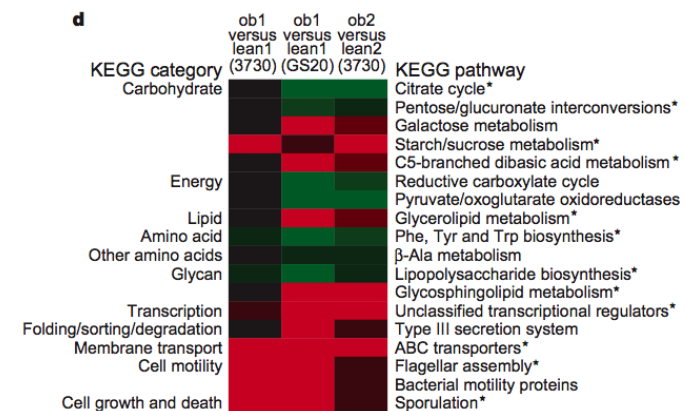
read length 93 bp



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)

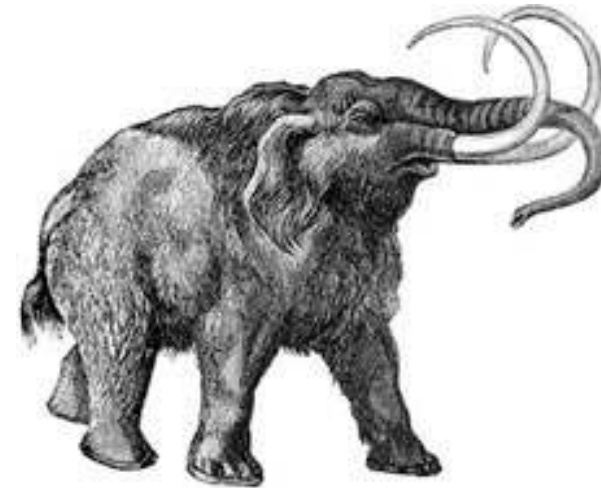
DNA extracted from 1 gram of
bone

454 sequencing:

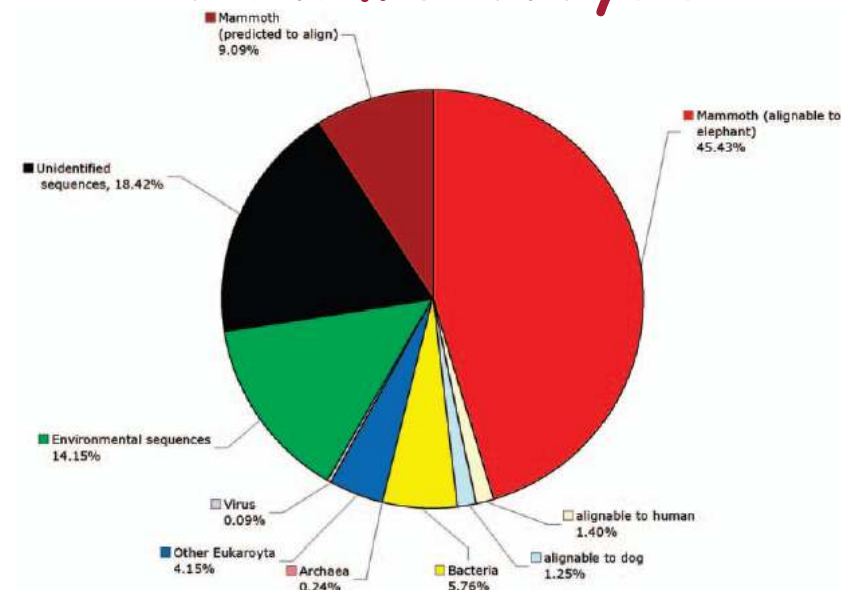
~302,000 reads

~95 bp length

> 50% mammoth



Taxonomic analysis



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*}



“Meta Transcriptomics” of Soil

Urich *et al* (2008):

RNA randomly reverse
transcribed into cDNA

No PCR or cloning

454 sequencing:

~ 250,000 sequences

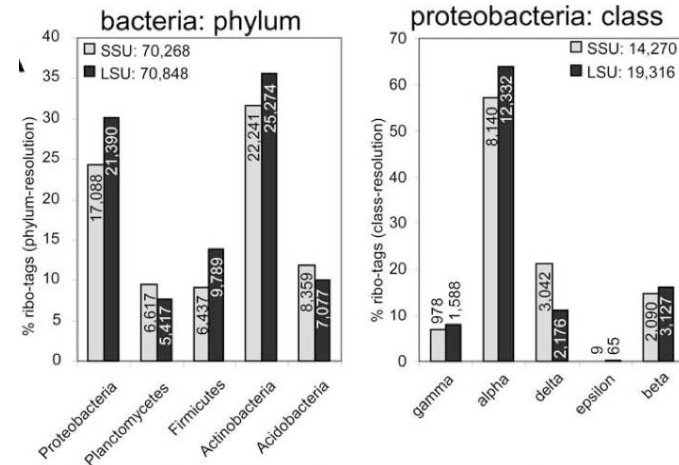
~ 98 bp length

RNA types:

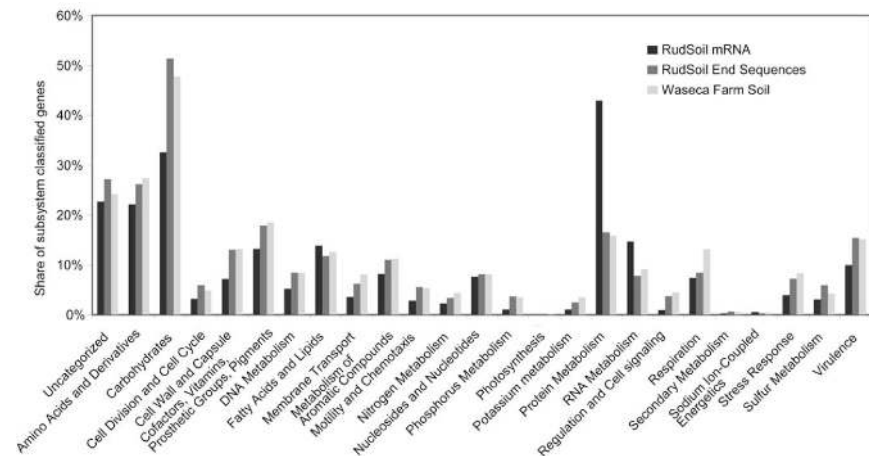
~ 75% rRNA *tags*

~ 8% mRNA *tags*

~ 17% unassigned



rRNA analysis



mRNA analysis



Large-Scale Human Gut Analysis

Taxonomic analysis using MEGAN

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin^{1*}, Ruiqiang Li^{1*}, Jeroen Raes^{2,3}, Manimozhiyan Arumugam², Kristoffer Solvsten Burgdorf⁴, Chaysavanh Manichanh⁵, Trine Nielsen⁴, Nicolas Pons⁶, Florence Levenez⁹, Takuji Yamada², Daniel R. Mende², Junhua Li^{1,7}, Junming Xu¹, Shaochuan Li¹, Dongfang Li^{1,8}, Jianjun Cao¹, Bo Wang¹, Huiqing Liang¹, Huisong Zheng¹, Yinlong Xie^{1,7}, Julien Tap⁶, Patricia Lepage⁶, Marcelo Bertalan⁹, Jean-Michel Batto⁶, Torben Hansen⁴, Denis Le Paslier¹⁰, Allan Linneberg¹¹, H. Bjørn Nielsen⁹, Eric Pelletier¹⁰, Pierre Renault⁶, Thomas Sicheritz-Ponten⁹, Keith Turner¹², Hongmei Zhu¹, Chang Yu¹, Shengting Li¹, Min Jian¹, Yan Zhou¹, Yingrui Li¹, Xiuqing Zhang¹, Songgang Li¹, Nan Qin¹, Huanming Yang¹, Jian Wang¹, Søren Brunak⁹, Joel Doré⁶, Francisco Guarner⁵, Karsten Kristiansen¹³, Oluf Pedersen^{4,14}, Julian Parkhill¹², Jean Weissenbach¹⁰, MetaHIT Consortium†, Peer Bork², S. Dusko Ehrlich⁶ & Jun Wang^{1,15}

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

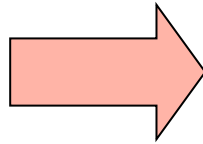
Outlook



Metagenome Analysis

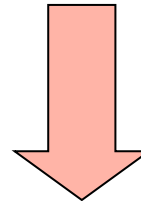


Environmental
sample



High-throughput
DNA sequencing

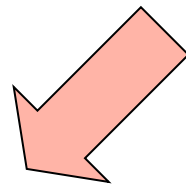
10 million sequences



Basic computational
analysis



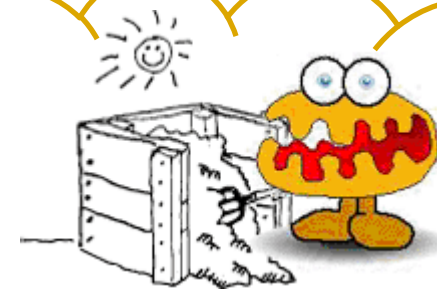
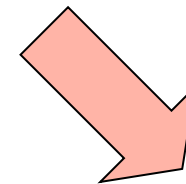
10 000
hours



www.compostinfo.com/tutorial/microbes.htm

Q1: Who is out there?

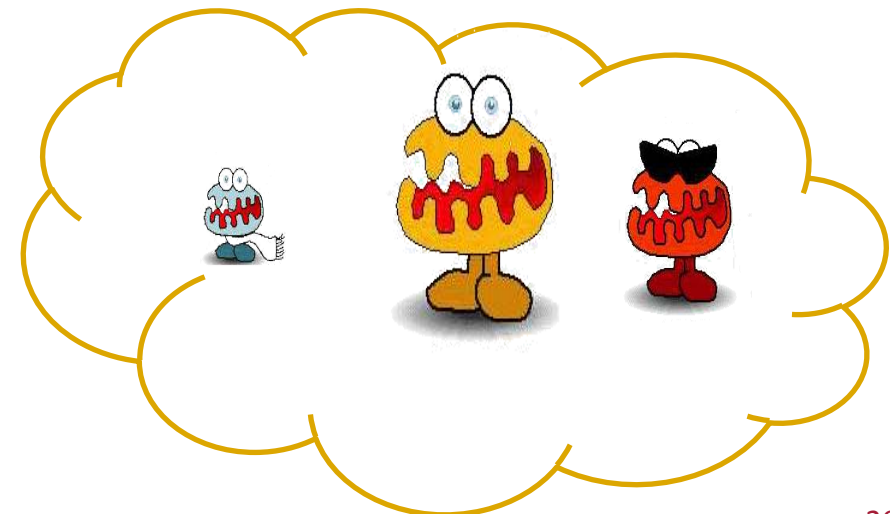
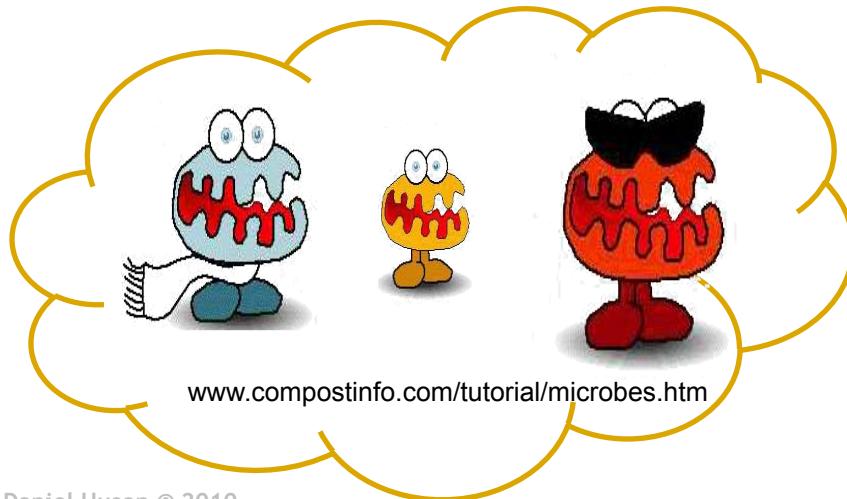
Daniel Huson © 2010



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?



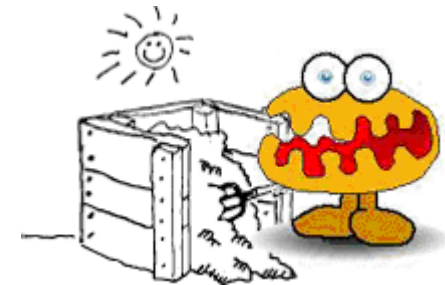
www.compostinfo.com/tutorial/microbes.htm

What are they doing?

Types of genes

Which metabolic pathways?

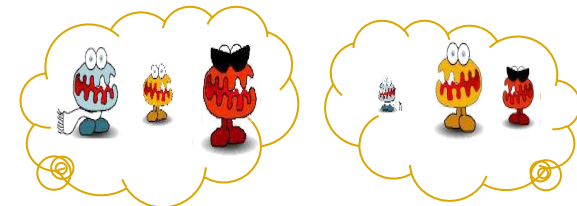
In what proportions?



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 as basis 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



Alternative Approaches

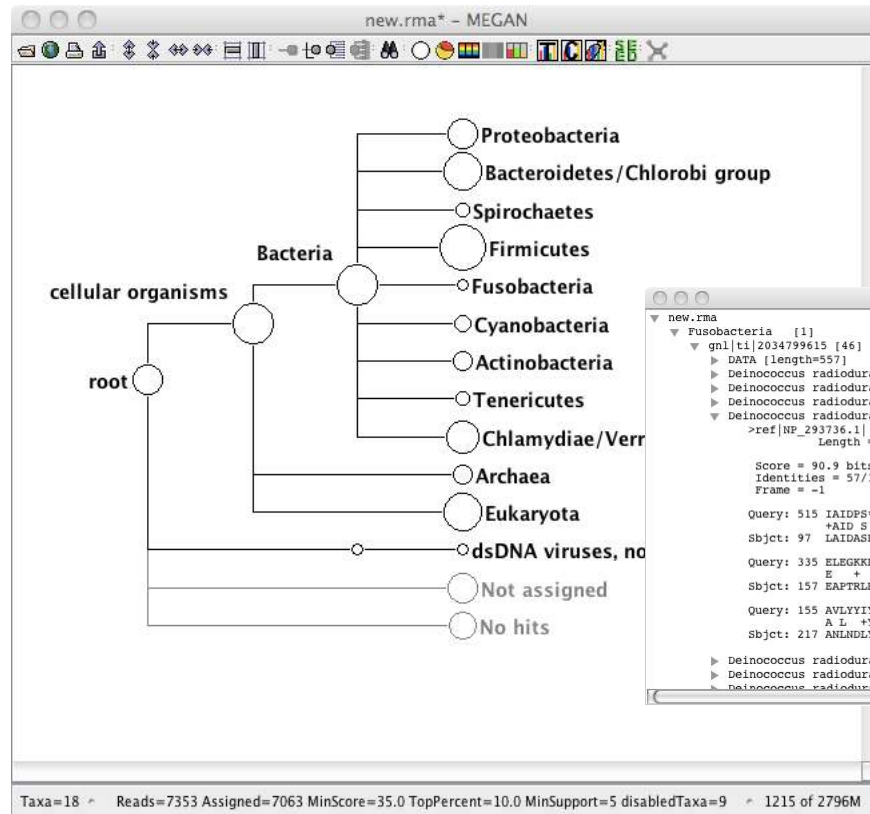
Based on signatures of data such as k-mer counts, use machine-learning method such as HMMs, SVMs or neural networks to classify reads

Hybrid approach: use fast machine-learning approach to place reads on high-level taxa, then use comparative approach restricted to those taxa



MEGAN – MEtaGenome ANalyzer

Huson et al, 2007

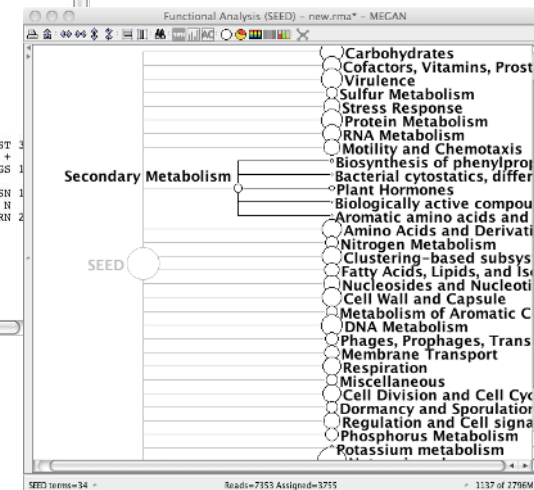
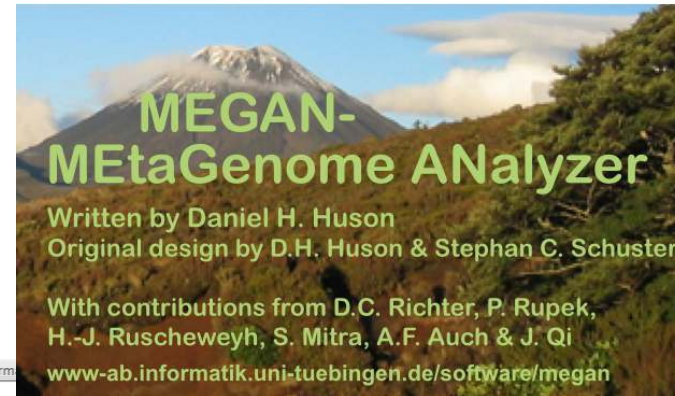


Inspector - new.rma

```

new.rma
└─ Fusobacteria [1]
   └─ gnl|ti|2034799615 [46]
      └─ DATA [length=557]
         └─ Deinococcus radiodurans R1 score=90.9
            └─ Deinococcus radiodurans R1 score=90.9
               └─ Deinococcus radiodurans R1 score=90.9
                  └─ Deinococcus radiodurans R1 score=90.9
                     └─ >ref|NP_293736.1| putative transposase [Deinococcus radiodurans R1]
                        └─ Length = 400
                           └─ Score = 90.9 bits (224), Expect = 4e-17
                              └─ Identities = 57/149 (38%), Positives = 78/149 (52%)
                                 └─ Frame = -1
                                    └─ Query: 515 IAIDPS*VSKAGKTAHIGRPNWGCASAVKHGLEILGIAVIDADIRDAHMLRAVOTLNST
                                       └─ +AID S KAG+ TAH+G FW+GCA+ + G+E A+ID R A+ + QTL +
                                          └─ Sb|ct: 97 LAIDASPHRAGGHTAHLSFWNGCAARTERGIEQSCCALIDVQHRQALTVDVQRTLTGS
                                             └─ Query: 335 ELEGKFTLNQWLSVLKTYRTDOLLKITSLLVADAAFSVLPFVEGLKEIGFSLISRLRN
                                                └─ E + + VL RT + +VAD ++ F VE + G ISRL N
                                                   └─ Sb|ct: 157 EAPTRLEQXADQLDDVLLDLRTVQQLDLAAVADGNAYAKEFIVETVTGHLPLISRLPN
                                                      └─ Query: 155 AVLYIYIEGPRTRGRGRPKTKDGKIDFSN 69
                                                         └─ A L +Y G +RGR K DGR+DFS+
                                                            └─ Sb|ct: 217 ANHLDLYTGEHPRRGRKKKFDGKVDPSD 245
                                                               └─ Deinococcus radiodurans R1 score=90.9
                                                                  └─ Deinococcus radiodurans R1 score=90.1
                                                                     └─ Deinococcus radiodurans R1 score=89.4

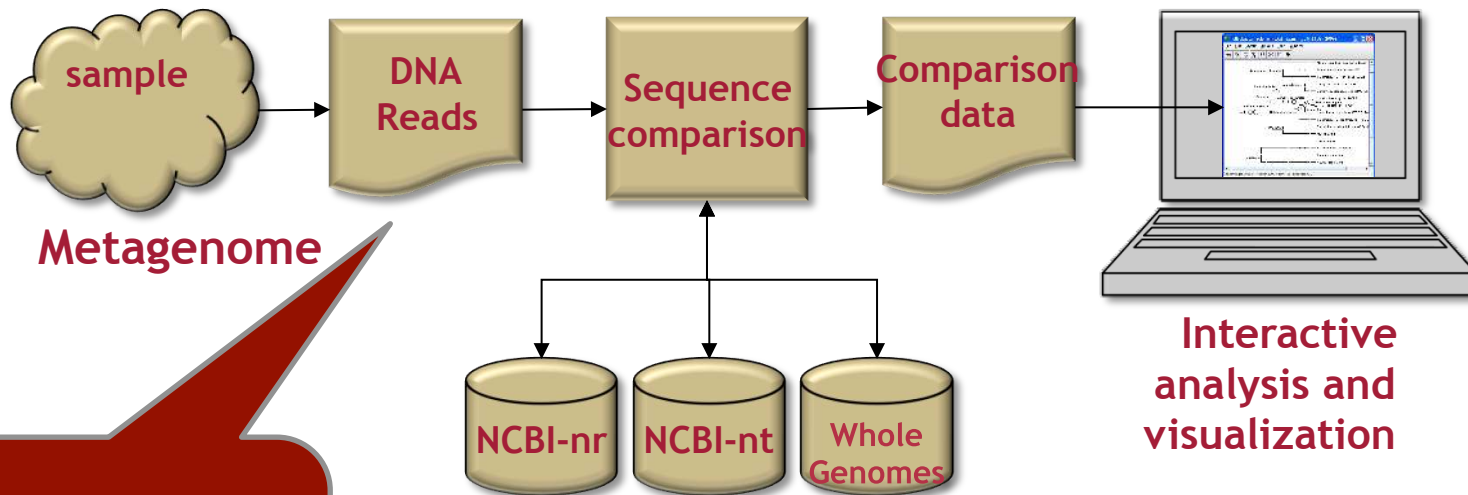
```



- Interactive tool for metagenomic analysis
(Version 4beta, released Dec 2010)



Metagenomics Pipeline



Similar for

- metatranscriptomics
- metaproteomics
- amplicon sequencing

Reference databases

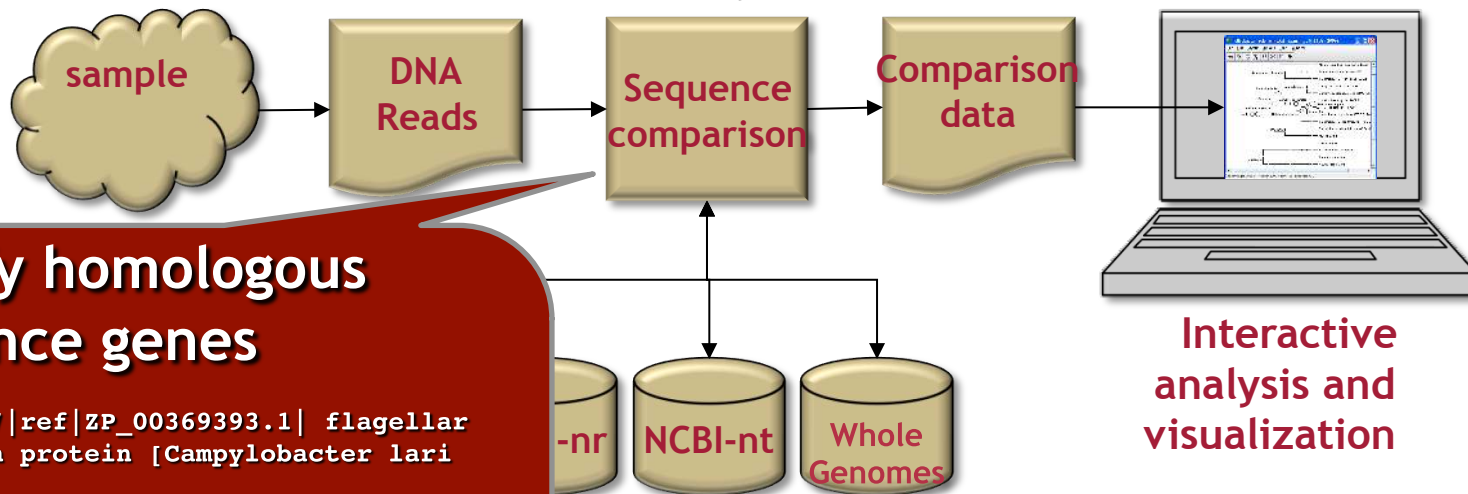


Identifying Taxa and Genes

Metagenome analysis

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

BLASTX
against NCBI-
NR



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
```

Reference databases



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
```

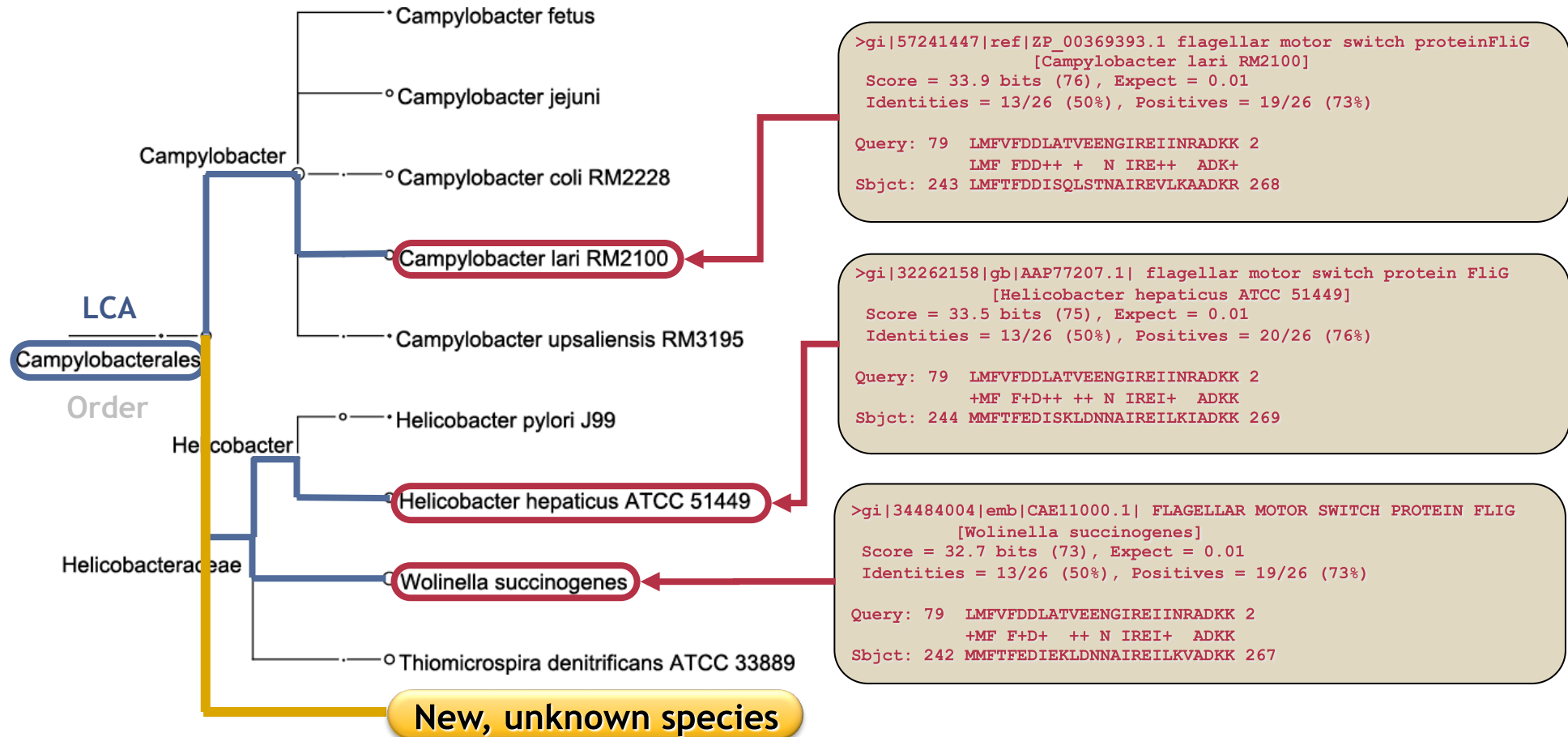
Indicates gene content:

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

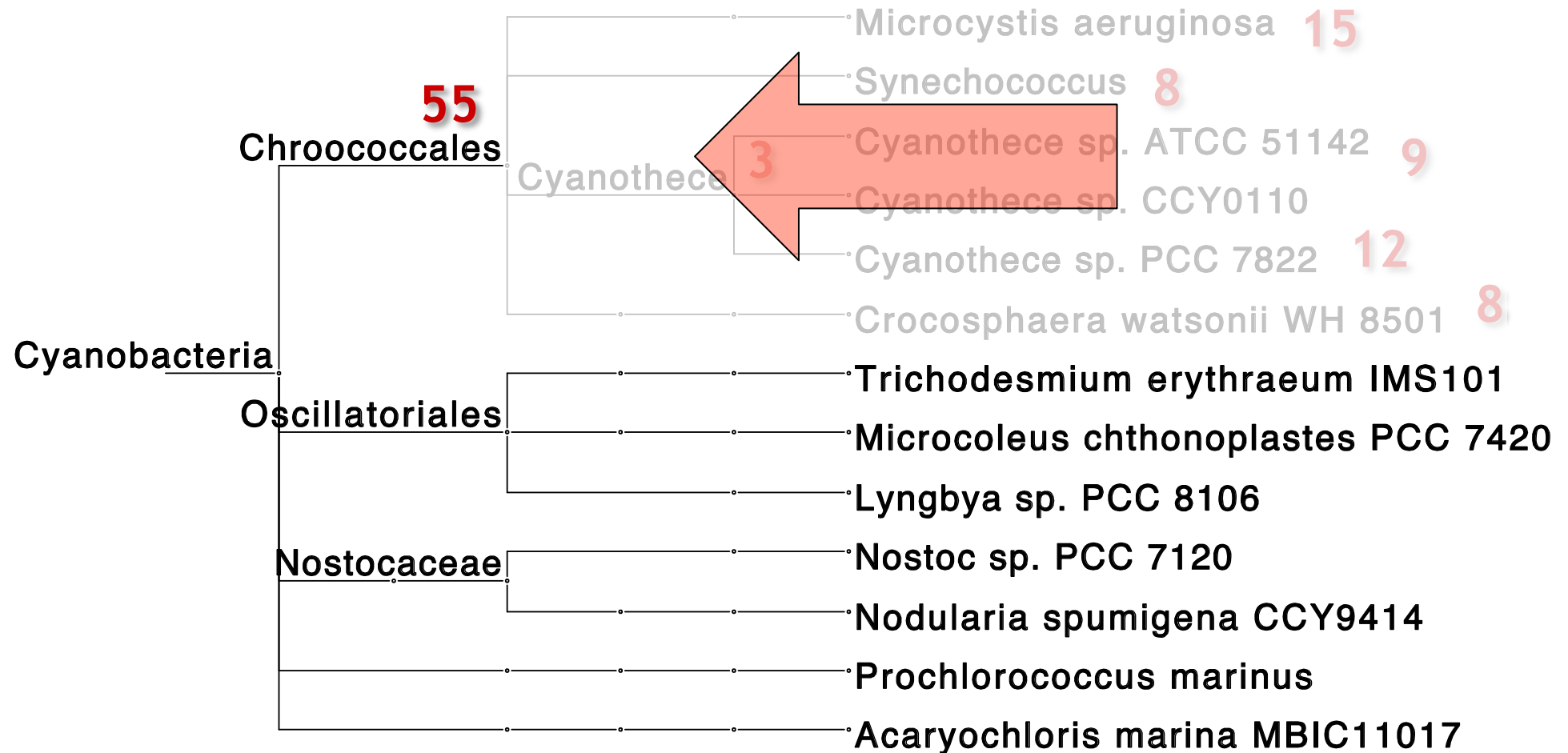
Is placement by **gene content** or “phylogenetic footprint”

Robust against false positive placements

Robust against (known cases) of horizontal gene transfer



Minimum Support Filter



Require at least 50 reads on a node, say



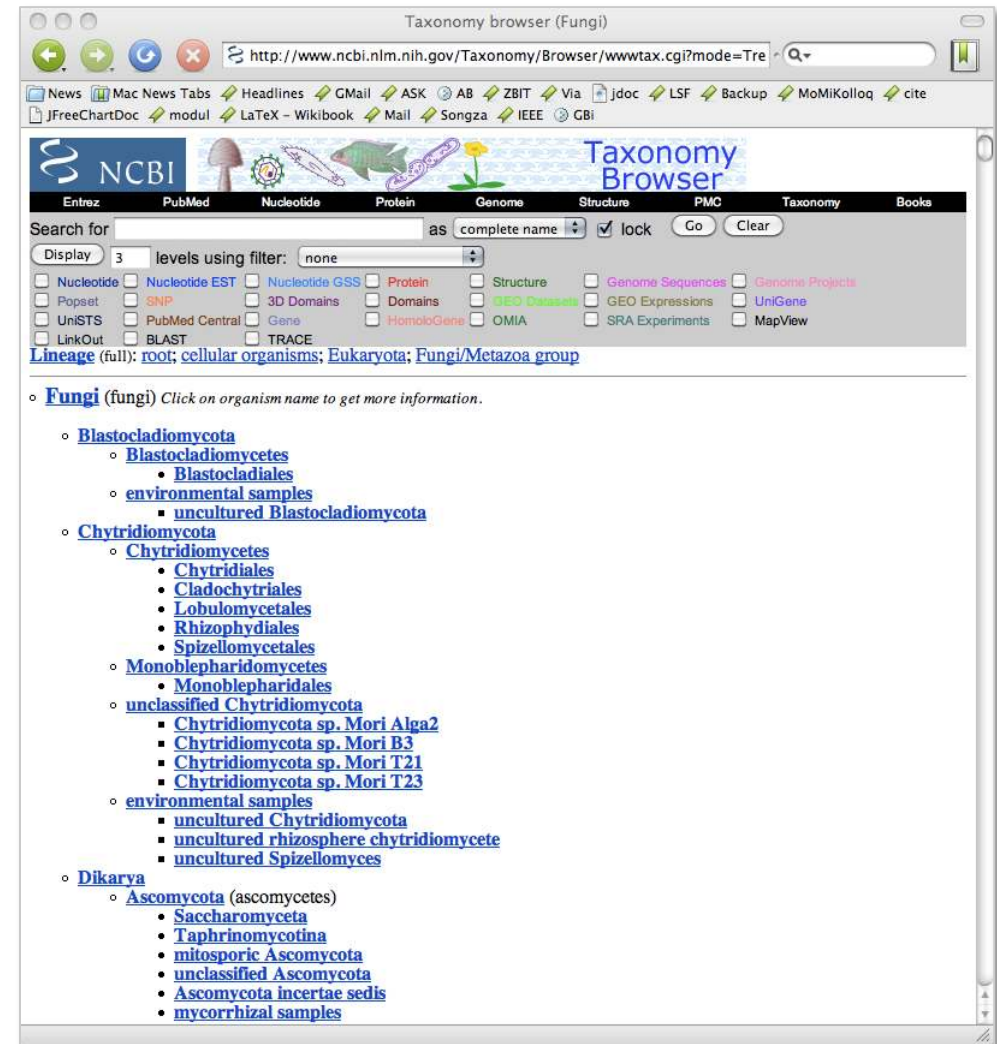
Taxonomic Analysis using NCBI Taxonomy

NCBI taxonomy:

Contains all species
represented by some
sequence

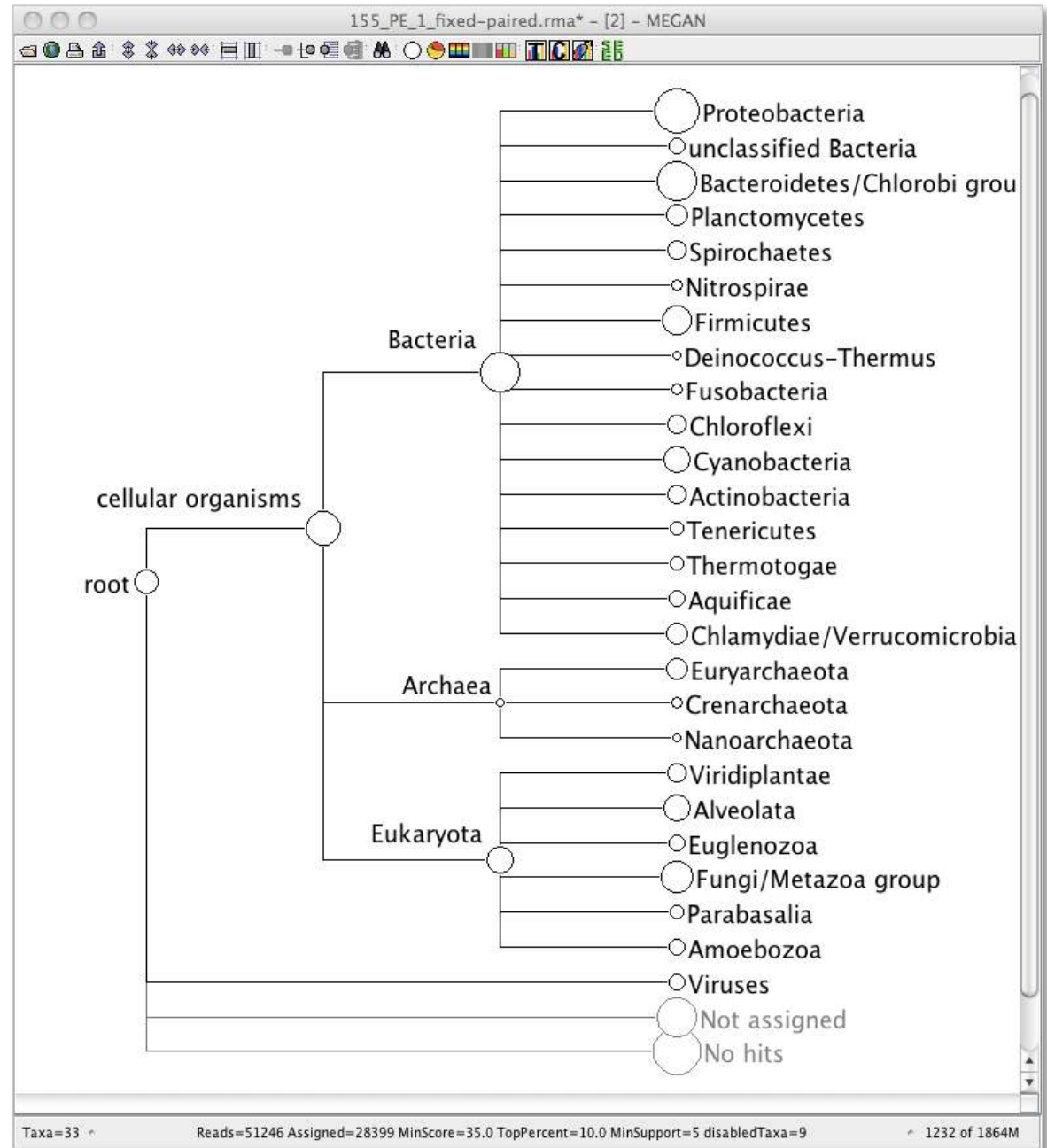
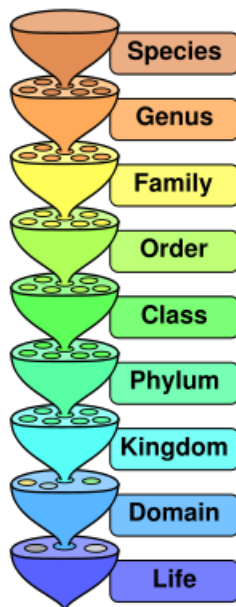
>560,000 nodes
(2007: 280,000 nodes)

King Phillip Came
Over For Green
Soup... (and more)



Organize and Visualize

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



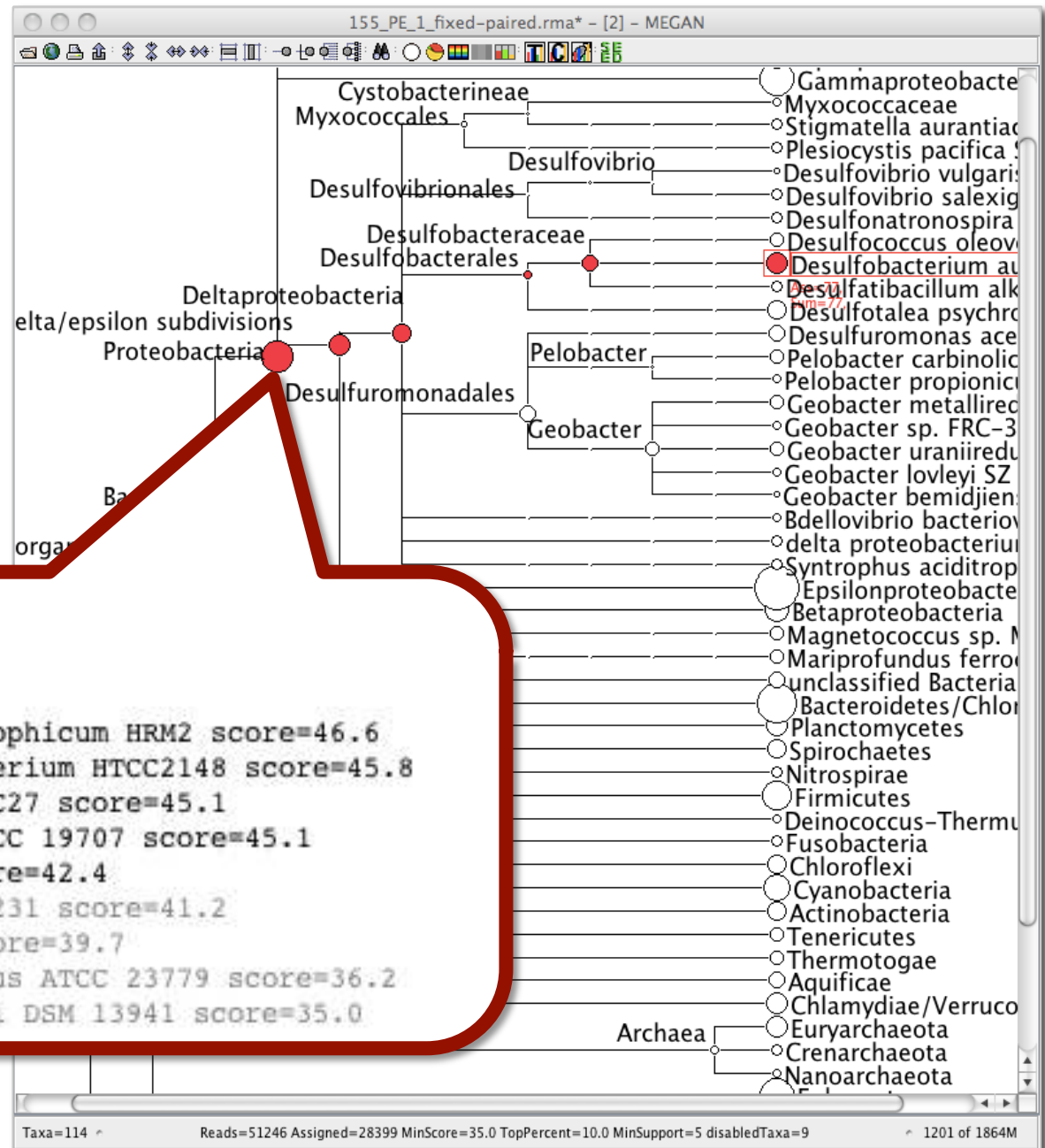
Taxonomic analysis of 50,000 reads

Organize and Visualize

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

```

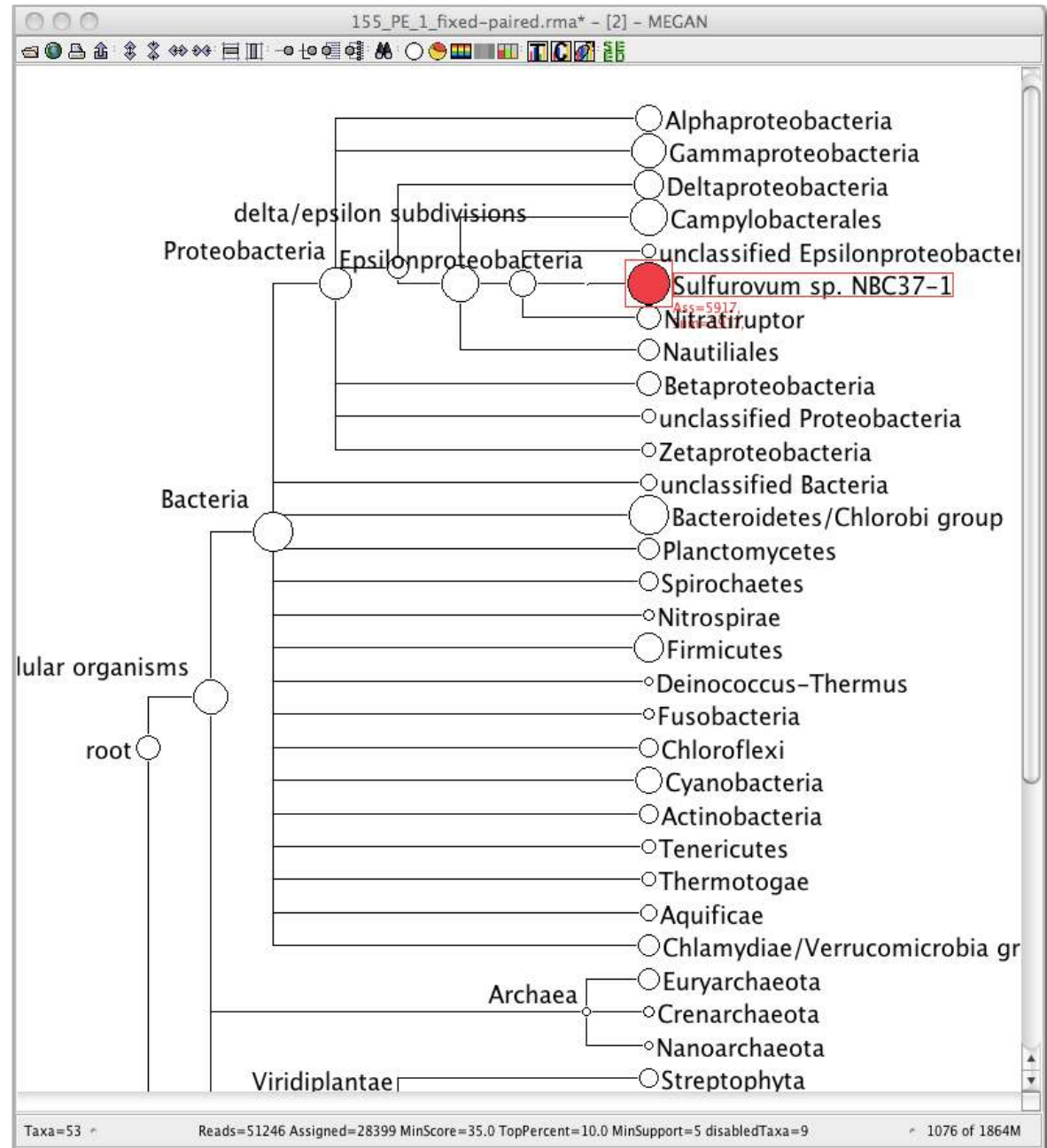
▼ GDEG1CX11G0QZ8.1 [9]
  ► DATA [length=233]
  ► Desulfobacterium autotrophicum HRM2 score=46.6
  ► marine gamma proteobacterium HTCC2148 score=45.8
  ► Nitrosococcus oceani AFC27 score=45.1
  ► Nitrosococcus oceani ATCC 19707 score=45.1
  ► Roseovarius sp. 217 score=42.4
  ► Nitrococcus mobilis Nb-231 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

155_PE_1_fixed-paired.rma* - [2] - MEGAN

Inspector - 155_PE_1_fixed-paired.rma - [2] - MEGAN

```

155_PE_1_fixed-paired.rma
└─ Sulfurovum sp. NBC37-1 [5917]
   └─ GDEG1CX11G3K5K.2 [6]
      └─ GDEG1CX11G2QRJ.2 [100]
         └─ GDEG1CX11GZA6G.2 [22]
            └─ GDEG1CX11GTWQ6.1 [100]
               └─ GDEG1CX11G2DYH.1 [100]
                  └─ GDEG1CX11G14WB.2 [1]
                     └─ GDEG1CX11GZZGJ.1 [13]
                        └─ GDEG1CX11G104U.1 [1]
                           └─ GDEG1CX11GWIOM.2 [22]
                              └─ GDEG1CX11GYO2L.2 [5]
                                 └─ GDEG1CX11G25HO.2 [100]
                                    └─ GDEG1CX11GY5Q9.1 [10]
                                       └─ GDEG1CX11G1VMN.2 [1]
                                          └─ GDEG1CX11GZR4E.1 [27]
                                             └─ GDEG1CX11GZR4E.2 [100]
                                                └─ GDEG1CX11G1W26.1 [100]
                                                   └─ GDEG1CX11G1L5B.2 [4]
                                                      └─ GDEG1CX11G1Y2D.1 [1]
                                                         └─ GDEG1CX11G0Q6M.1 [100]
                                                            └─ GDEG1CX11G2DJ2.1 [36]
                                                               └─ DATA [length=237]
                                                                  >GDEG1CX11G2DJ2.1 length=299 xy=2780 1424 region=11 run=R 2010_03_03_14_54_16
                                                                  TTTTAGAGATAGAAATTATCCAGACTTAGGAGTAGCCATAGACGAACCTTTTCGCTGCAATAGACAAAAGTAAAT
                                                                  └─ Sulfurovum sp. NBC37-1 score=80.5
                                                                     >ref|YP_001357841.1| hypothetical protein SUN_0524 [Sulfurovum sp. NBC37-1]
                                                                     dbj|BAF71484.1| conserved hypothetical protein [Sulfurovum sp. NBC37-1]
                                                                     Length = 87

                                                                     Score = 80.5 bits (197), Expect(2) = 8e-22
                                                                     Identities = 36/51 (70%), Positives = 47/51 (92%)
                                                                     Frame = +2

                                                                     Query: 74 NATKEQEEDLEEMRDMRTECFSIIEELGRDELEAEEIDELLAELVEMKTEE 226
                                                                     NATKEQ+EDLEEMR+MRTECF+I+EE+ +DEL+ EE +ELL ELV++KT+E
                                                                     Sbjct: 35 NATKEQKEDLEEMRDMRTECFATVEEIKKDELDEEEAEELLTELVDLKTDE 85

                                                                     Score = 46.6 bits (109), Expect(2) = 8e-22
                                                                     Identities = 23/32 (71%), Positives = 27/32 (84%)
                                                                     Frame = +3

                                                                     Query: 3 LEIEIIPDLGVAIDELFAAIDKSKMPQKNKKK 98
                                                                     LE+EIIPDL VAIDELFAAIDK+K K +K+
                                                                     Sbjct: 11 LELEIIPDLVAIDELFAAIDKAKNATKEQKE 42

                                                                     Strongylocentrotus purpuratus score=37.0
                                                                     Drosophila erecta score=35.8
                                                                     Rattus norvegicus score=35.8
                                                                     Algoriphagus sp. PR1 score=35.4
                                                                     Gallus gallus score=35.0

```

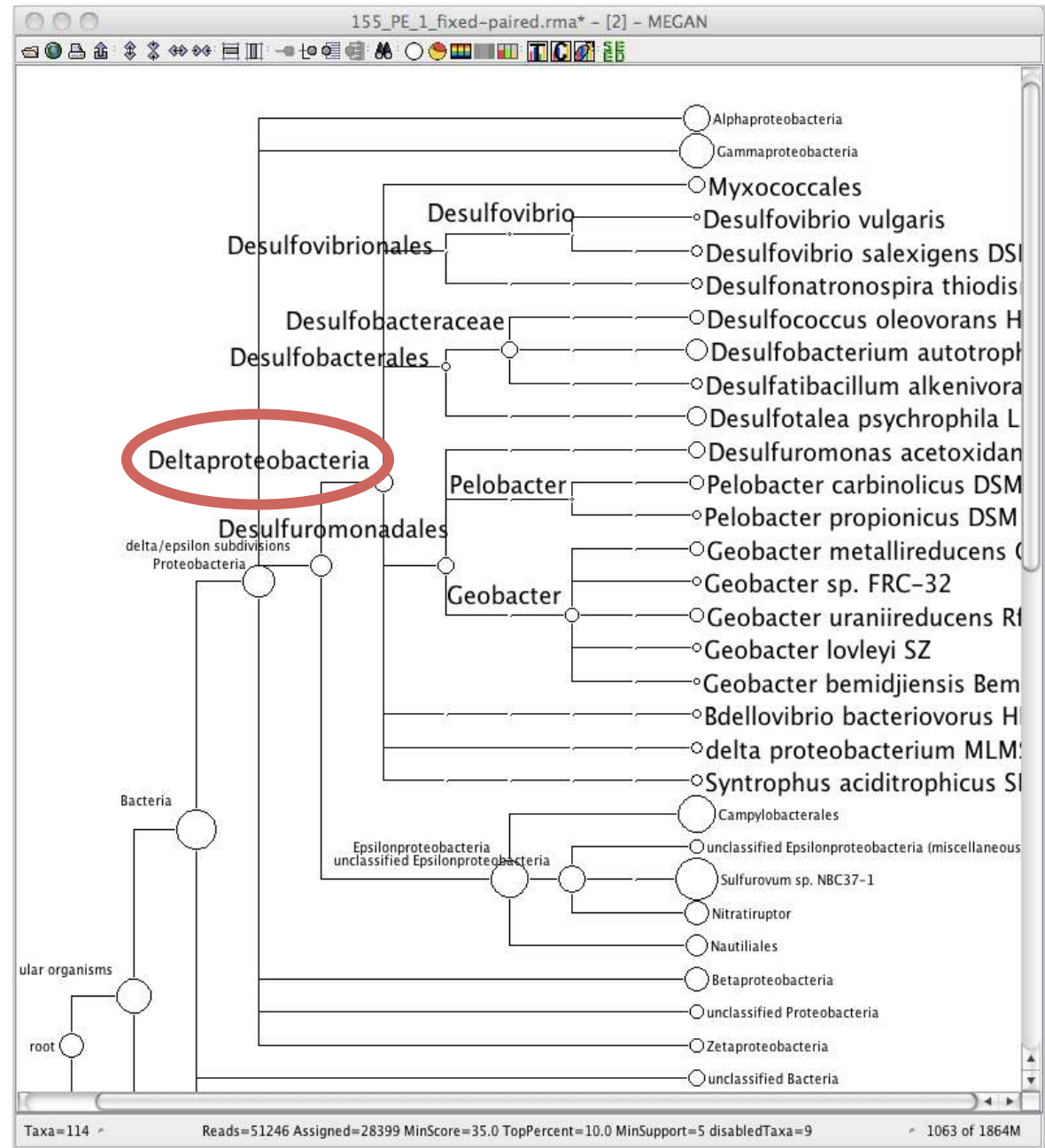
Viridiplantae Streptophyta

Taxa=53 Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9 1076 of 1864M

Taxonomic analysis of 50,000 reads

Interact and Summarize

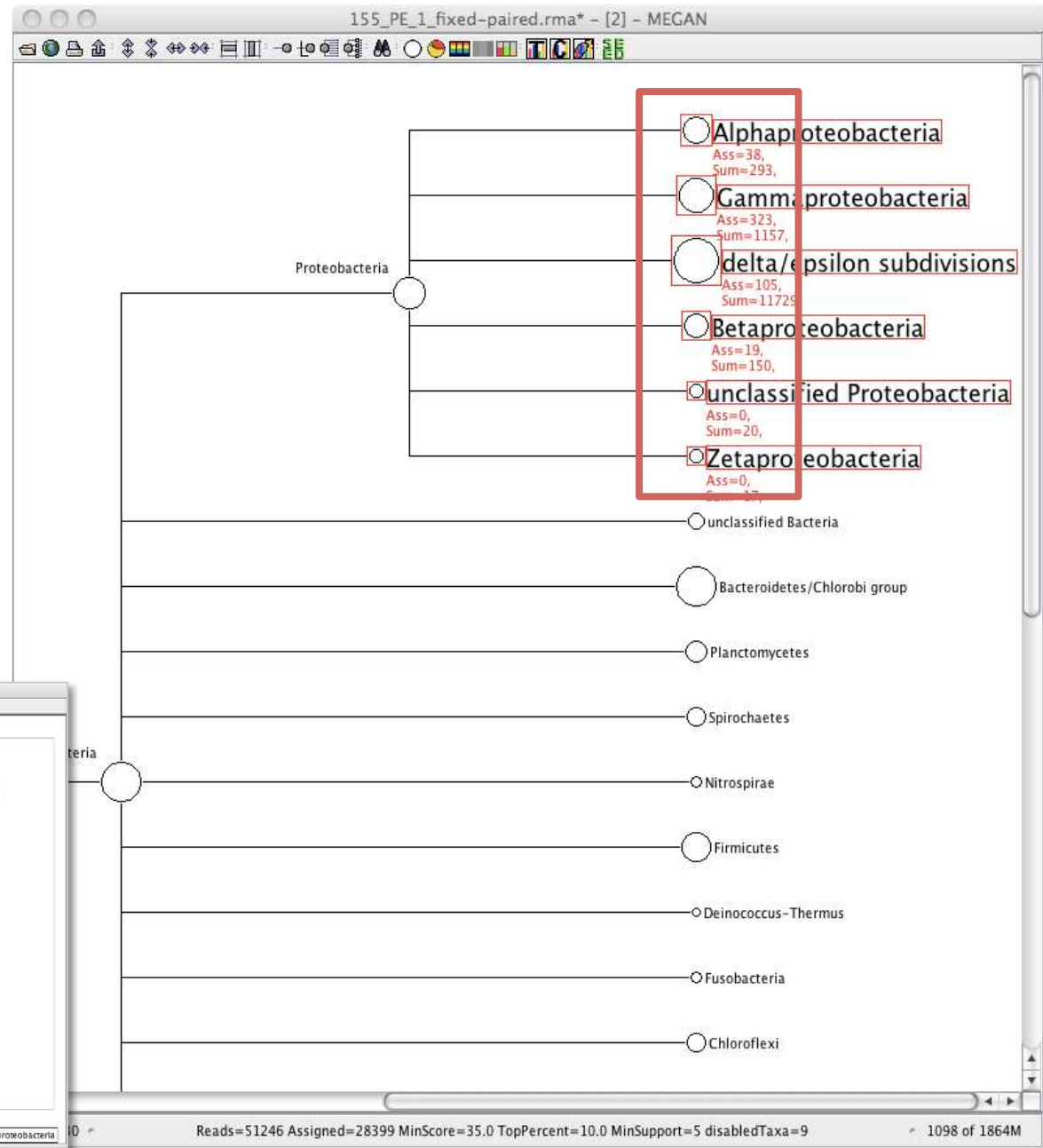
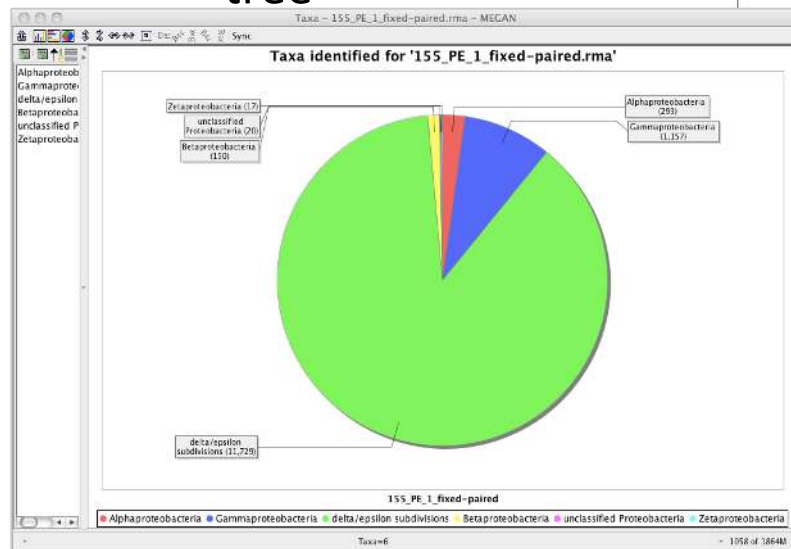
- ✓ Search for nodes of interest
- ✓ Inspect sequences assigned to a node
- ✓ Collapse and un-collapse parts of the tree



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



Capture

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

The screenshot displays the MEGAN software interface. The main window shows a taxonomic tree on the left and a list of sequences on the right. The 'Export' menu is open, showing options like 'Export', 'Export Image...', 'Page Setup...', 'Print...', 'Compare...', 'Extract Reads By Taxa...', 'Extract Reads By COGs...', 'Extract Reads By Subsystems...', 'Import CSV...', 'Tools', 'Properties...', and 'Close'. The 'Extract Reads By Taxa...' option is highlighted. Below the main window, the 'Extract by Taxa - 155_PE_1_fixed-paired.rma - [2] - MEGAN' dialog box is open. It shows the output files directory as '/Users/huson/data/megan' and the file name as 'reads-%t.fasta'. The 'Include Summarized' checkbox is checked. The 'Execute' button is highlighted.

Taxonomic analysis of 50,000 reads

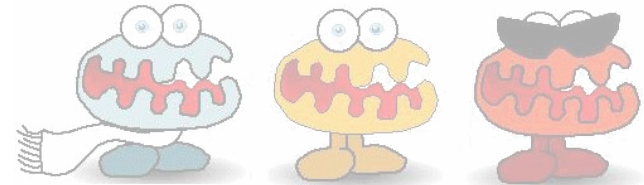


Three Basic Computational Questions

Who is out there?

Types of organisms

In what proportions?



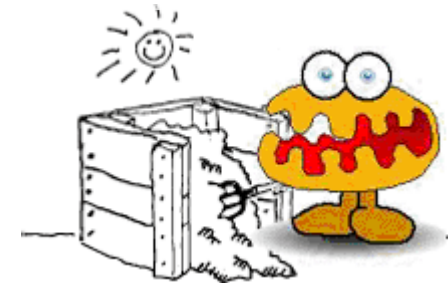
www.compostinfo.com/tutorial/microbes.htm

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

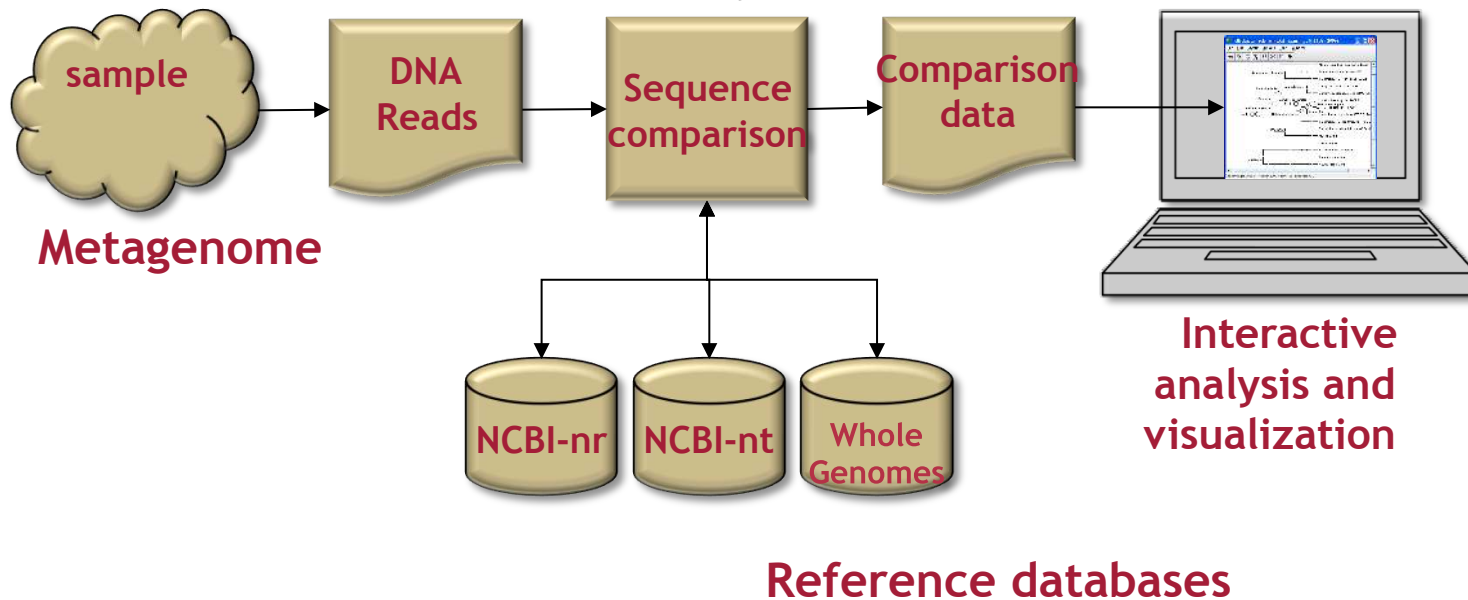


Identifying Taxa and Genes

Metagenome analysis

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

BLASTX
against NCBI-
NR





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 $\approx 13,000$ nodes

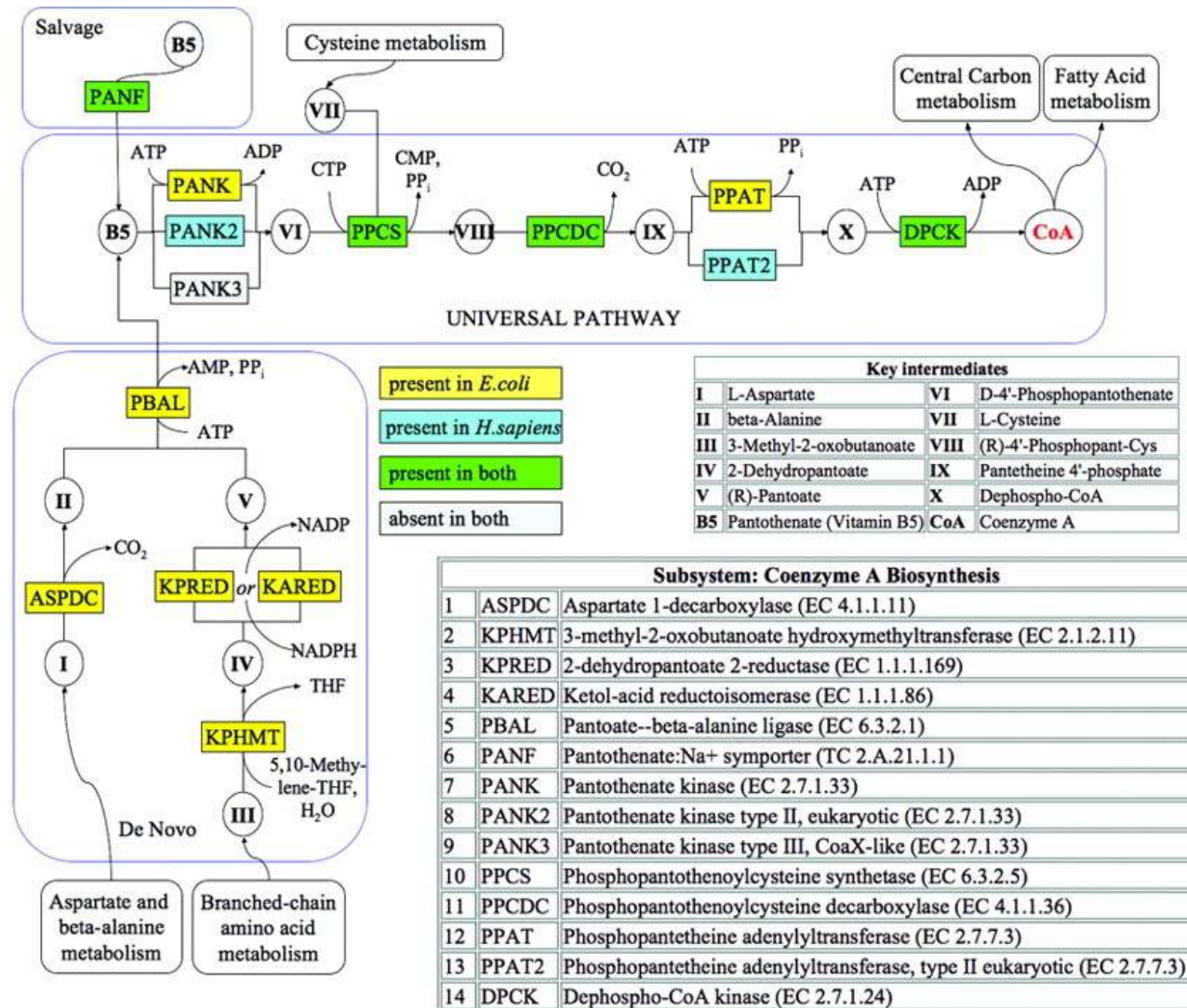
www.theSEED.org



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



Example of a Subsystem



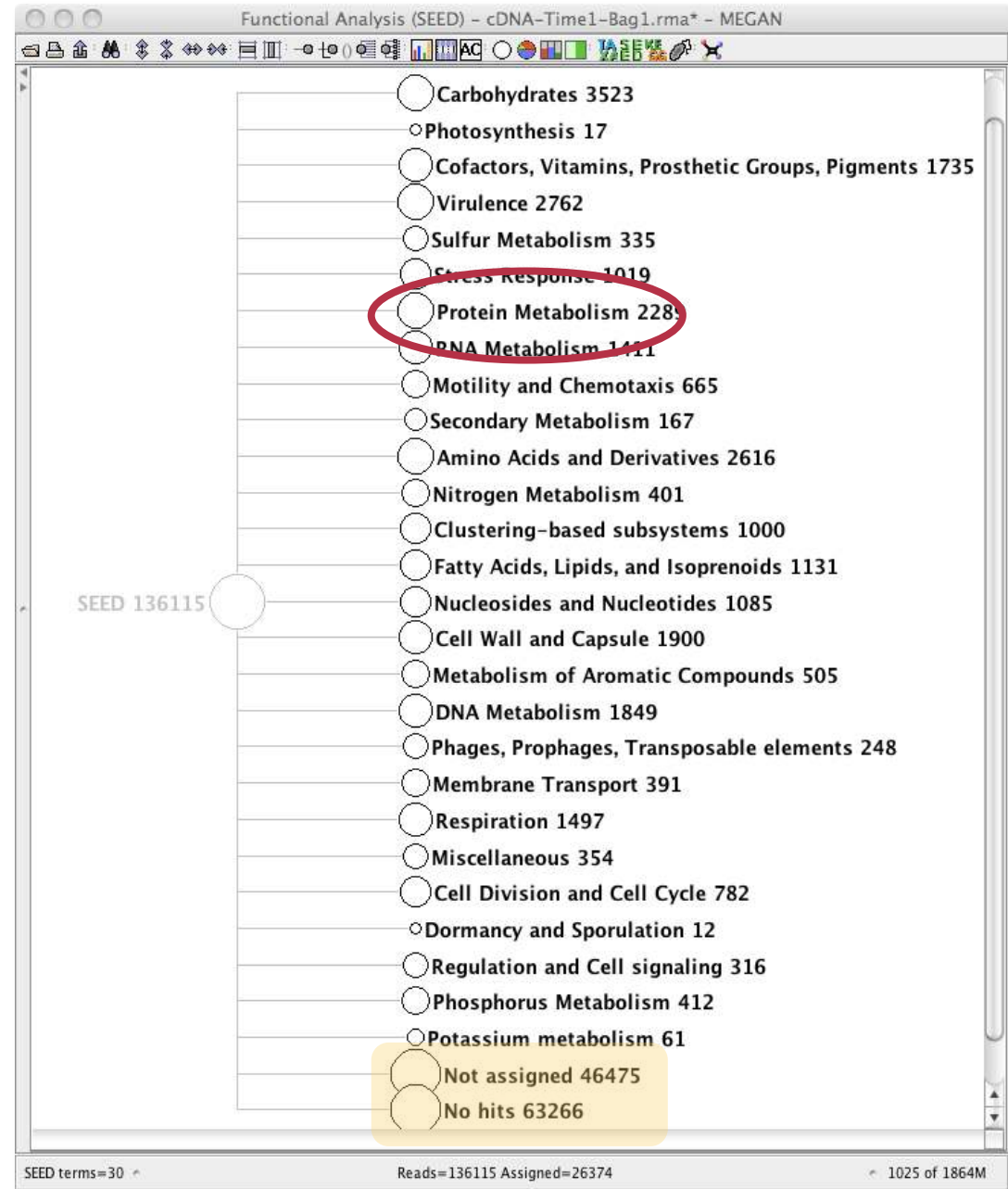
Subsystem: Coenzyme A Biosynthesis

SEED Functional Analysis

MEGAN's SEED-based analysis

Overview

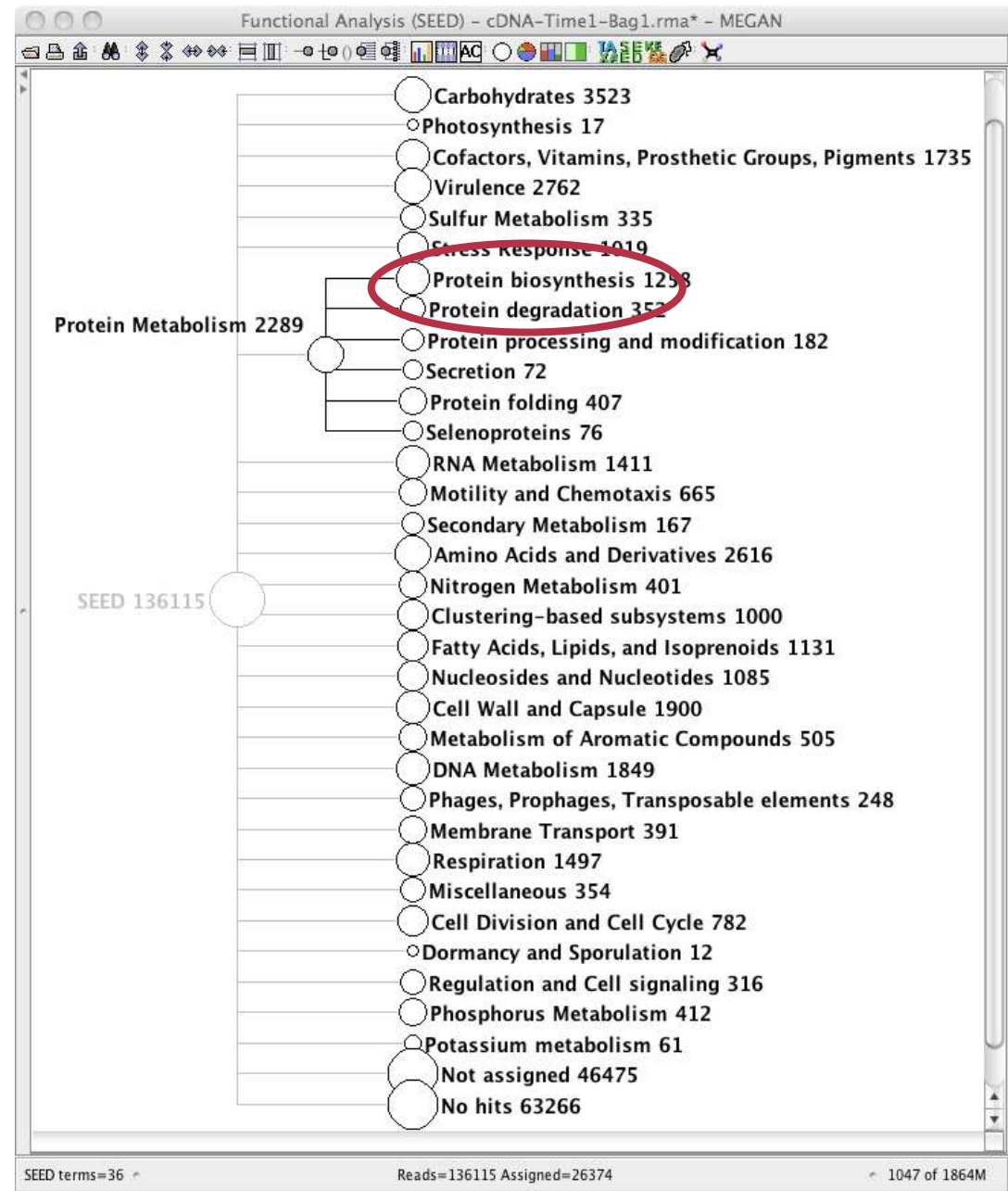
136 000 Marine
cDNAs (Gilbert 2008)
26 000 assigned



SEED analysis of 136 000 reads

SEED Functional Analysis

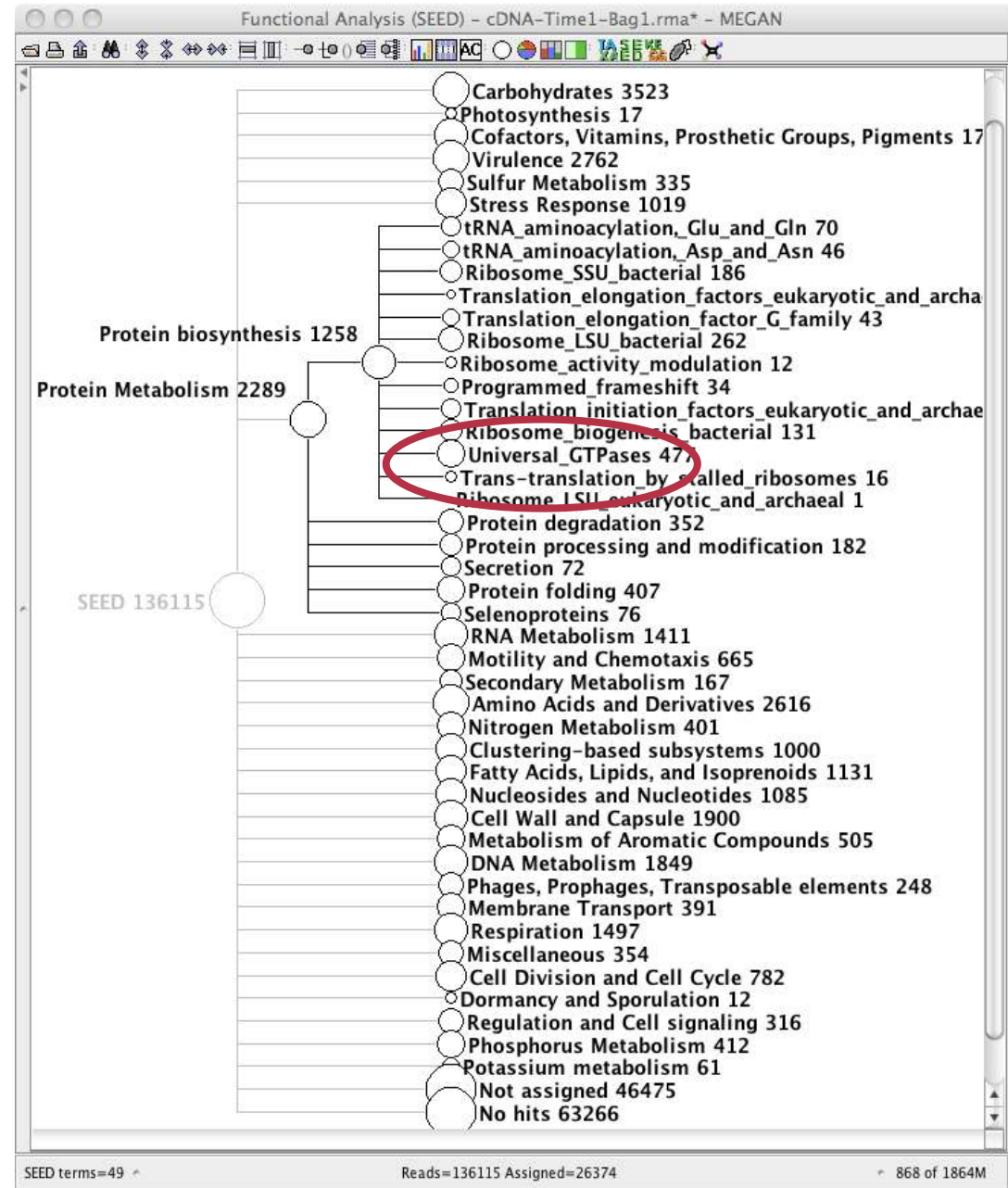
Details of Protein Metabolism



SEED analysis of 136 000 reads

SEED Functional Analysis

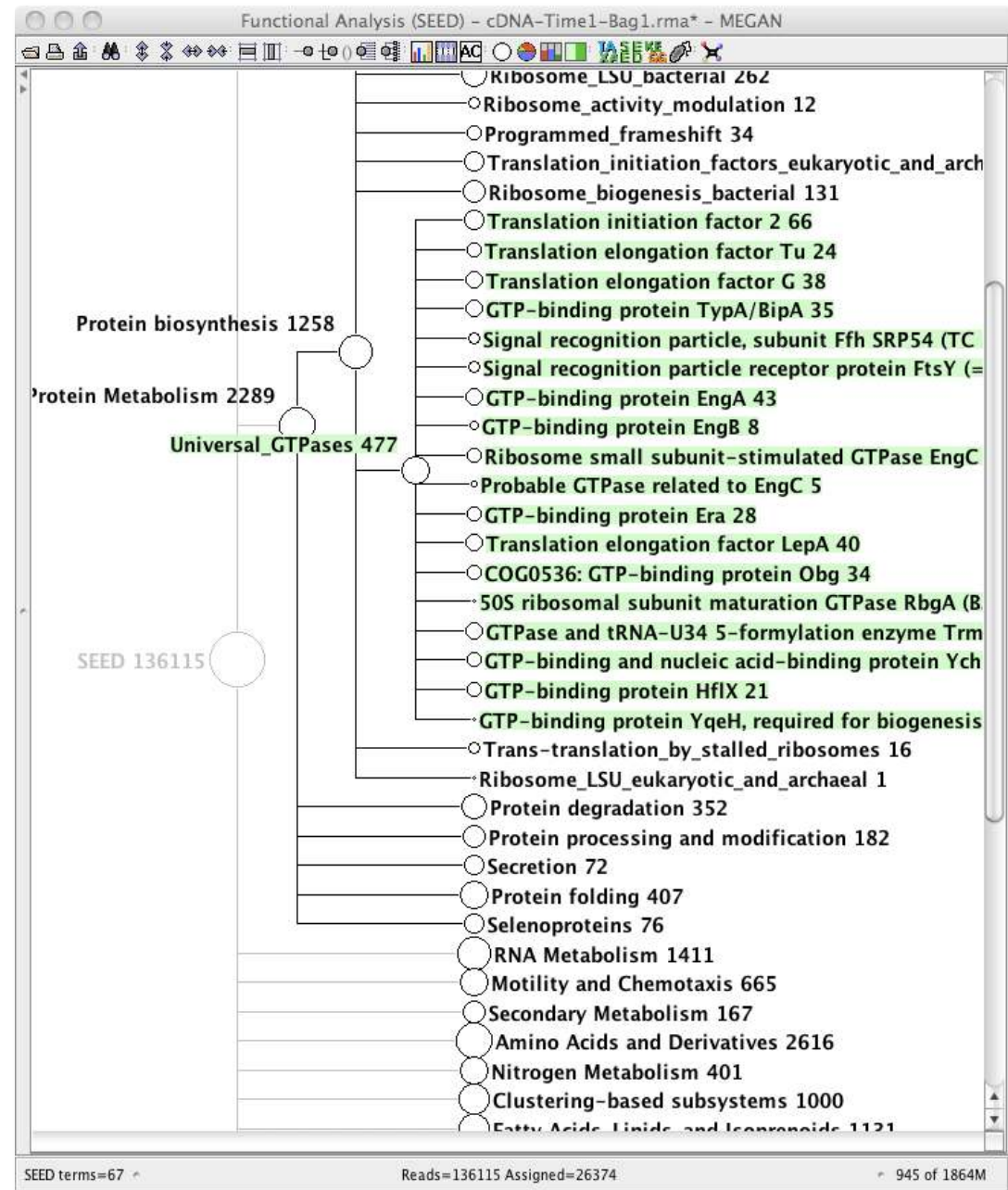
Details of
 Protein
 biosynthesis



SEED analysis of 136 000 reads

SEED Functional Analysis

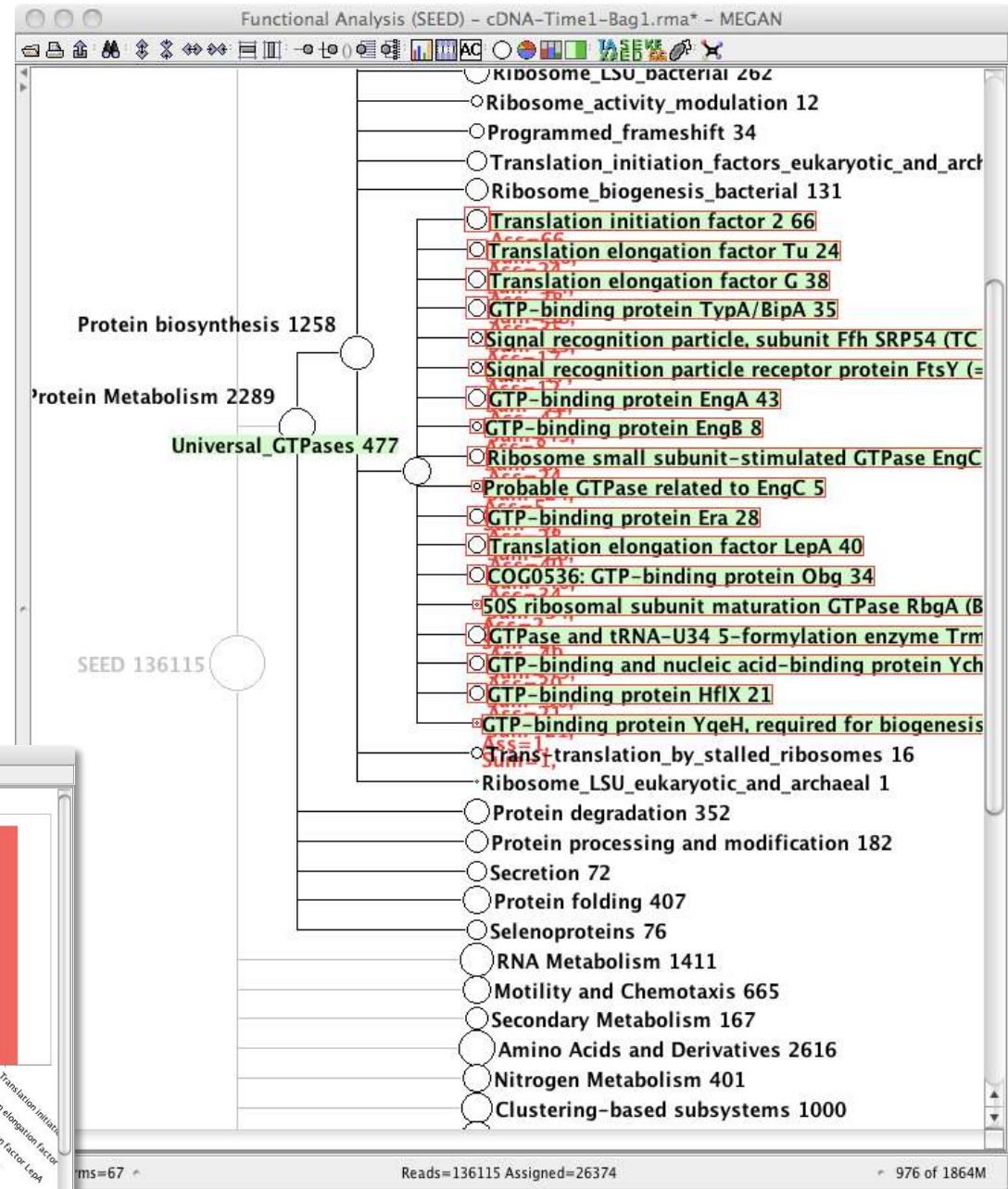
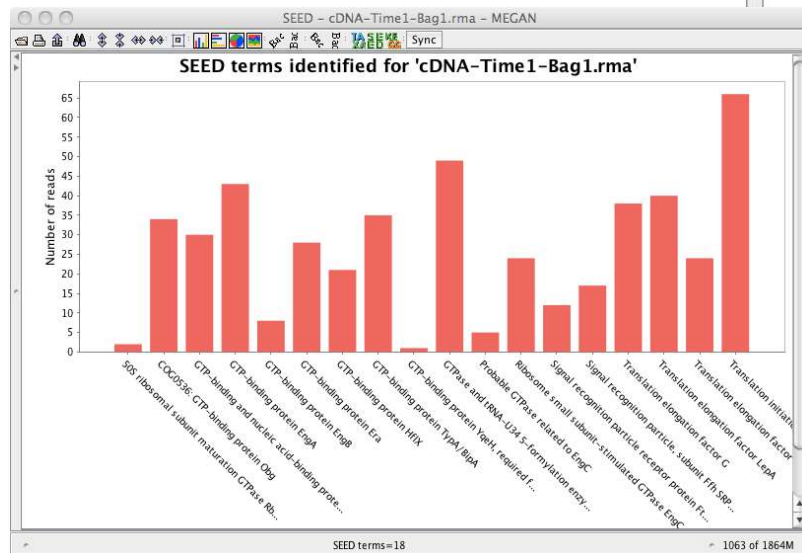
Details of Universal GTPases



SEED analysis of 136 000 reads

SEED Functional Analysis

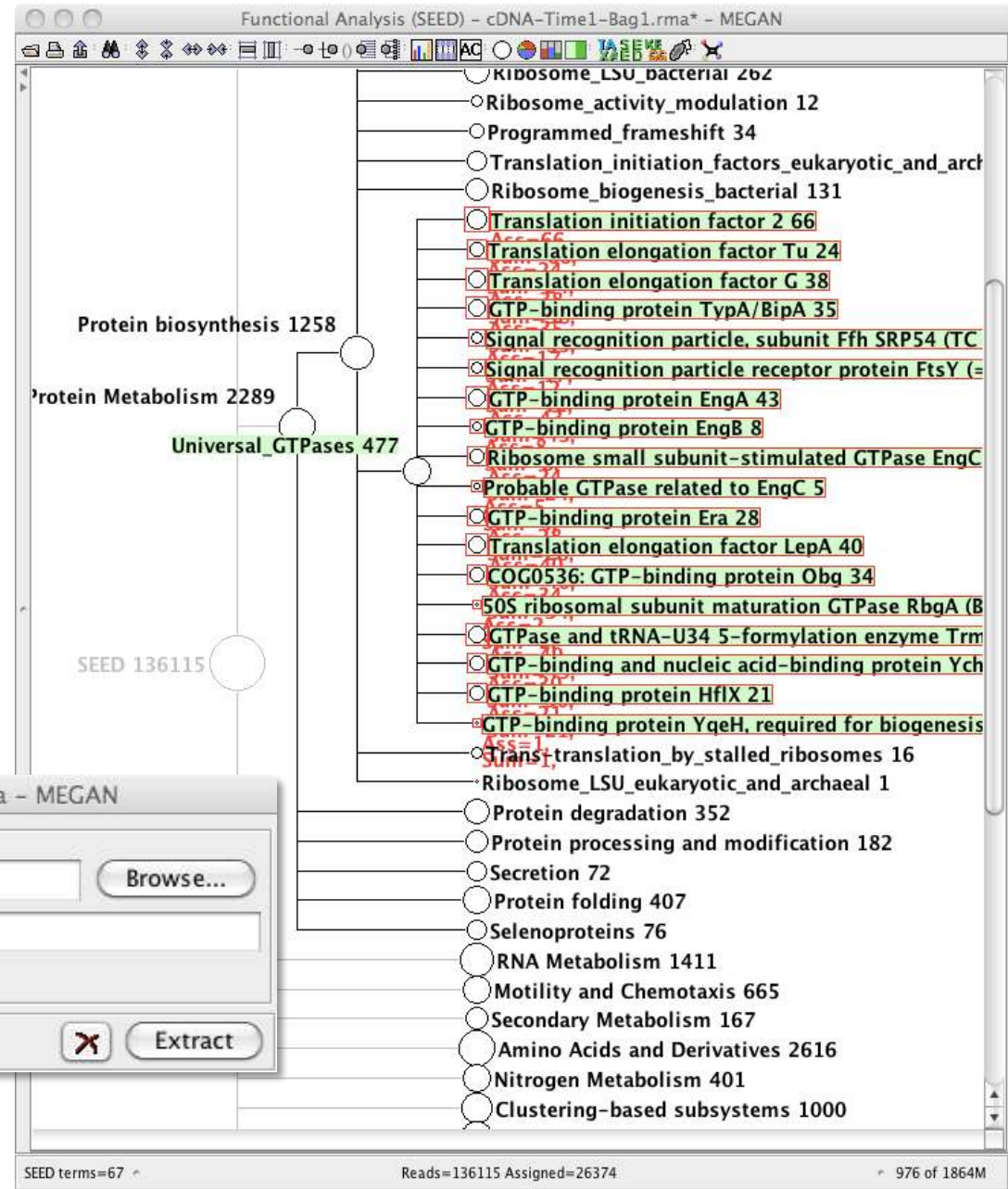
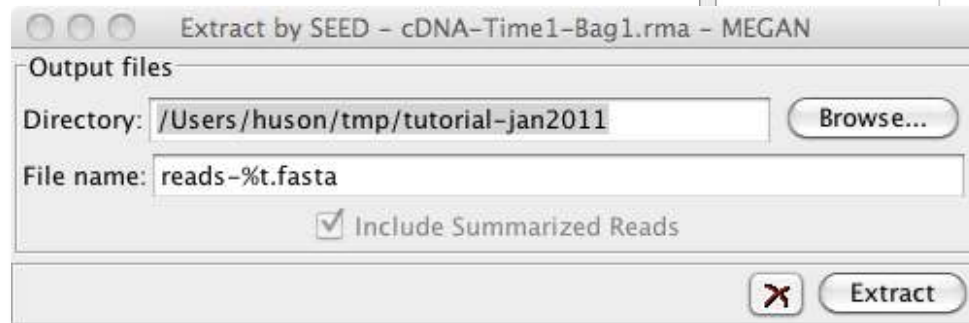
Chart selected nodes



SEED analysis of 136 000 reads

SEED Functional Analysis

Capture reads by assignment



SEED analysis of 136 000 reads

SEED Functional Analysis

Inspect assigned reads

Inspector - 155_PE_1_fixed.rma - [2] - MEGAN

155_PE_1_fixed.rma

- Translation elongation factor G [38]
 - GDEG1CX11G29DN.2 [25]
 - GDEG1CX11GY69S.2 [25]
 - GDEG1CX11GWE1K.2 [25]
 - DATA [length=111]
- Campylobacteriales bacterium GD 1 score=48.1

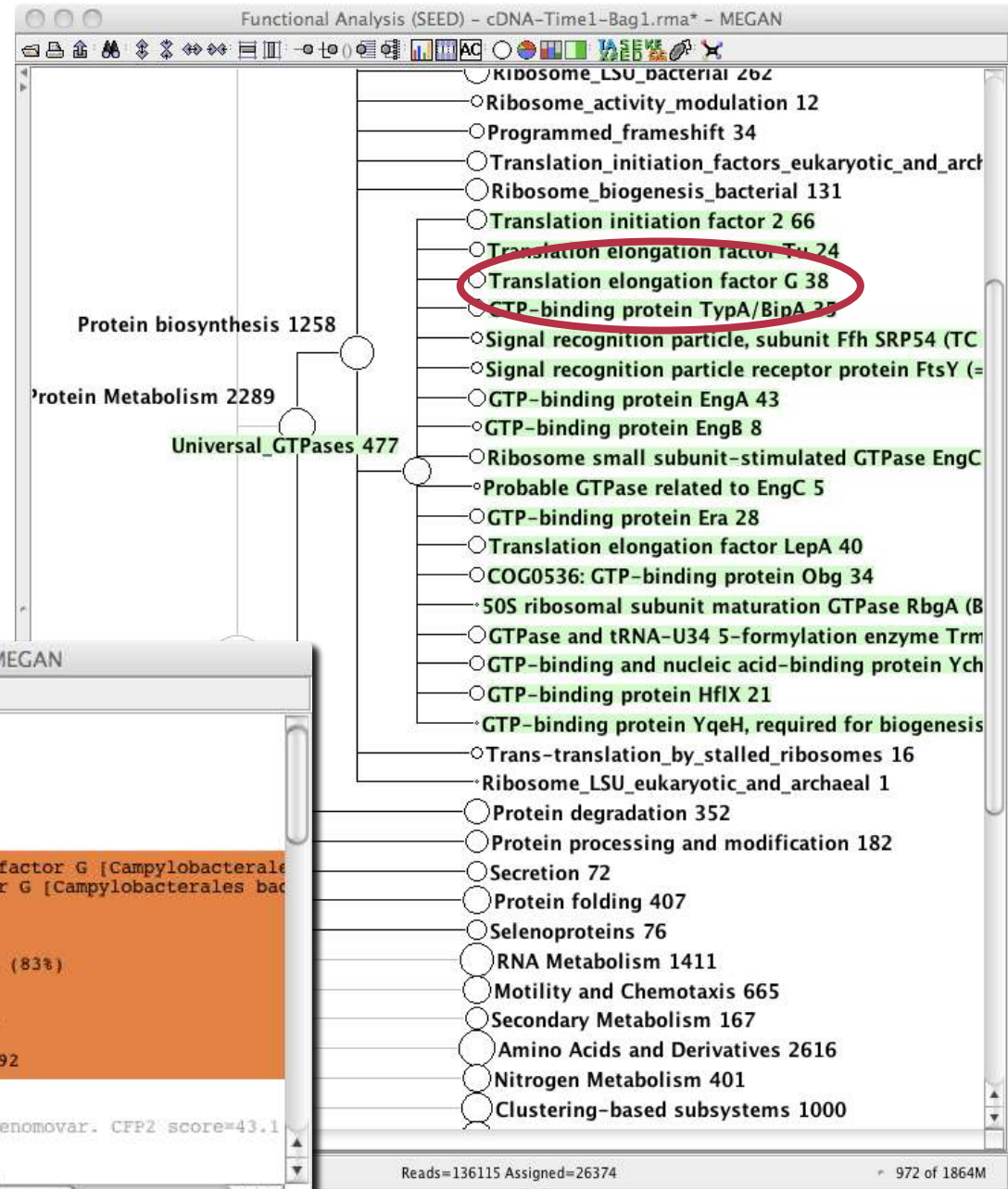

```
>ref|YP_002592669.1| translation elongation factor G [Campylobacteriales
gb|EDZ61772.1| translation elongation factor G [Campylobacteriales bac
Length = 695

Score = 48.1 bits (113), Expect = 3e-04
Identities = 23/31 (74%), Positives = 26/31 (83%)
Frame = -2

Query: 94 CRFCI+NKGVTLLDAVVAYLPSPVESAAIK 2
      C NKG+QTLDAVVAYLPSP+ESAAI+
Sbjct: 262 CGTAFKNKGIQTLDAVVAYLPSPMESAAIE 292

Sulfurimonas denitrificans DSM 1251 score=44.3
Sulfurovum sp. NBC37-1 score=43.5
Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2 score=43.1
Shewanella benthica KT99 score=43.1
Labrenzia alexandrii DFL-11 score=42.4
Caldicellulosiruptor horaei DSM 6725 score=42.4
```

Rows: 67 993 of 1864M



Analysis of 136 000 reads

KEGG Functional Analysis

Graph with
13,000 nodes

Overview

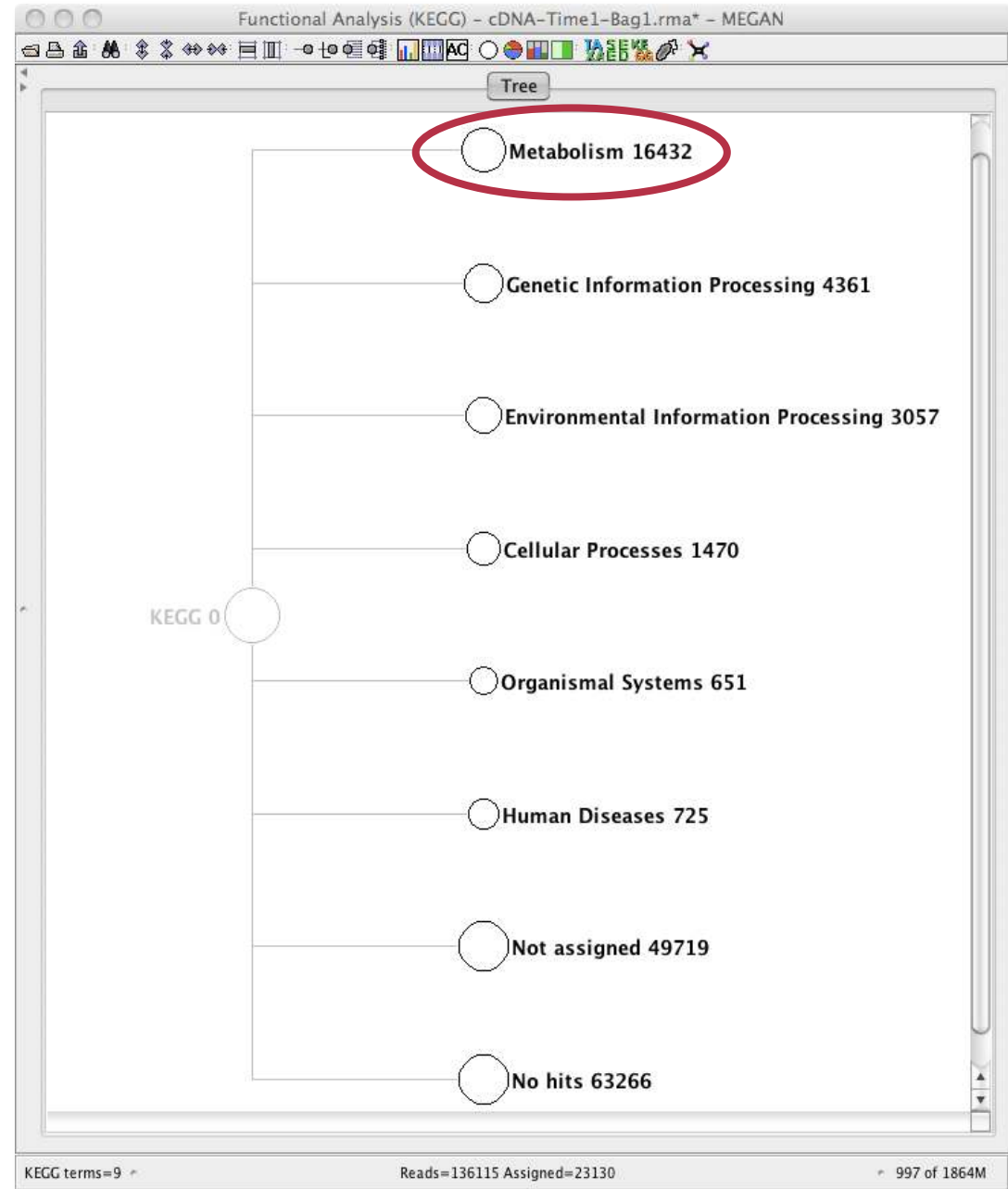
23 000 assigned



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

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

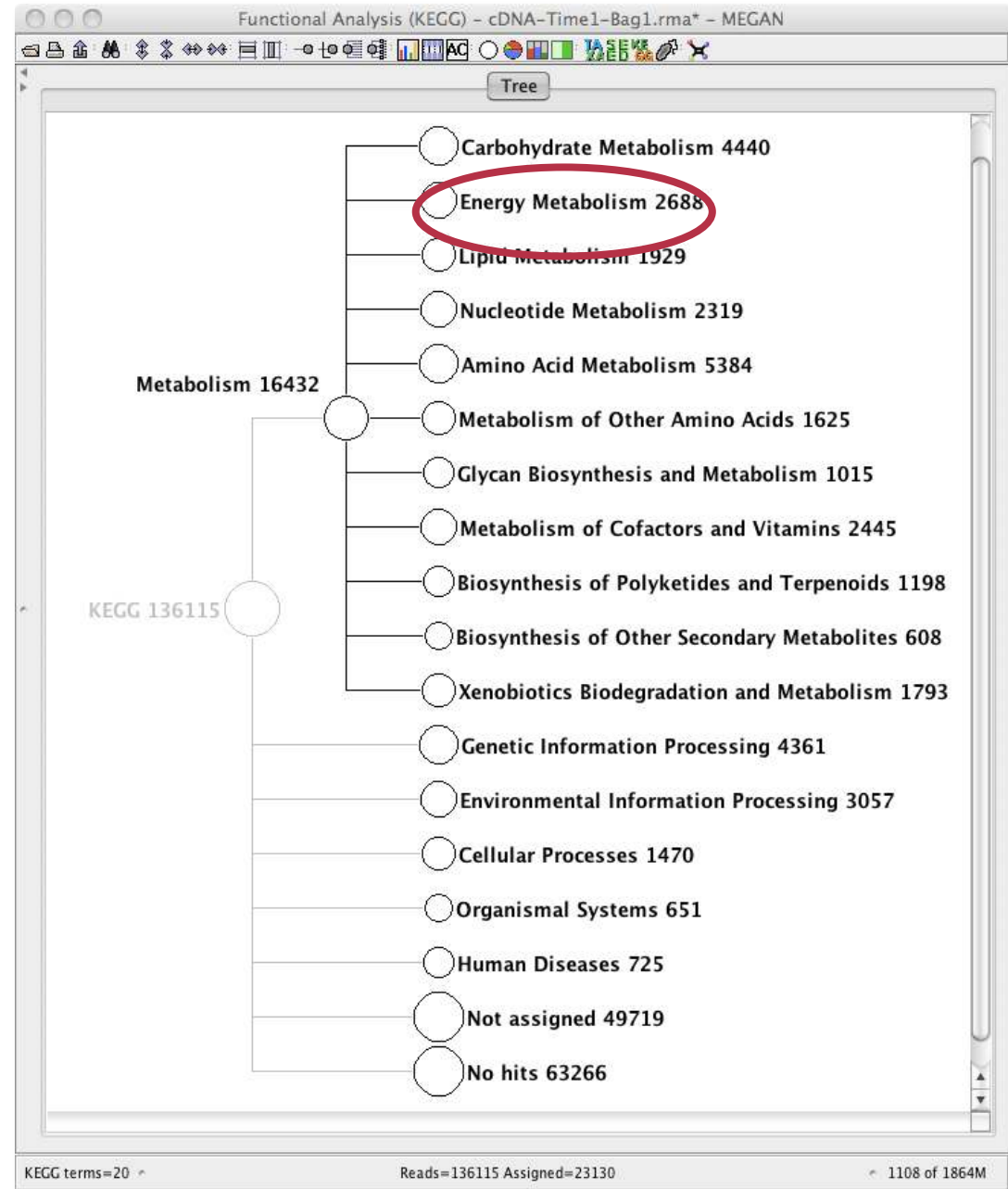
Daniel Huson © 2010



KEGG analysis of 136 000 reads

KEGG Functional Analysis

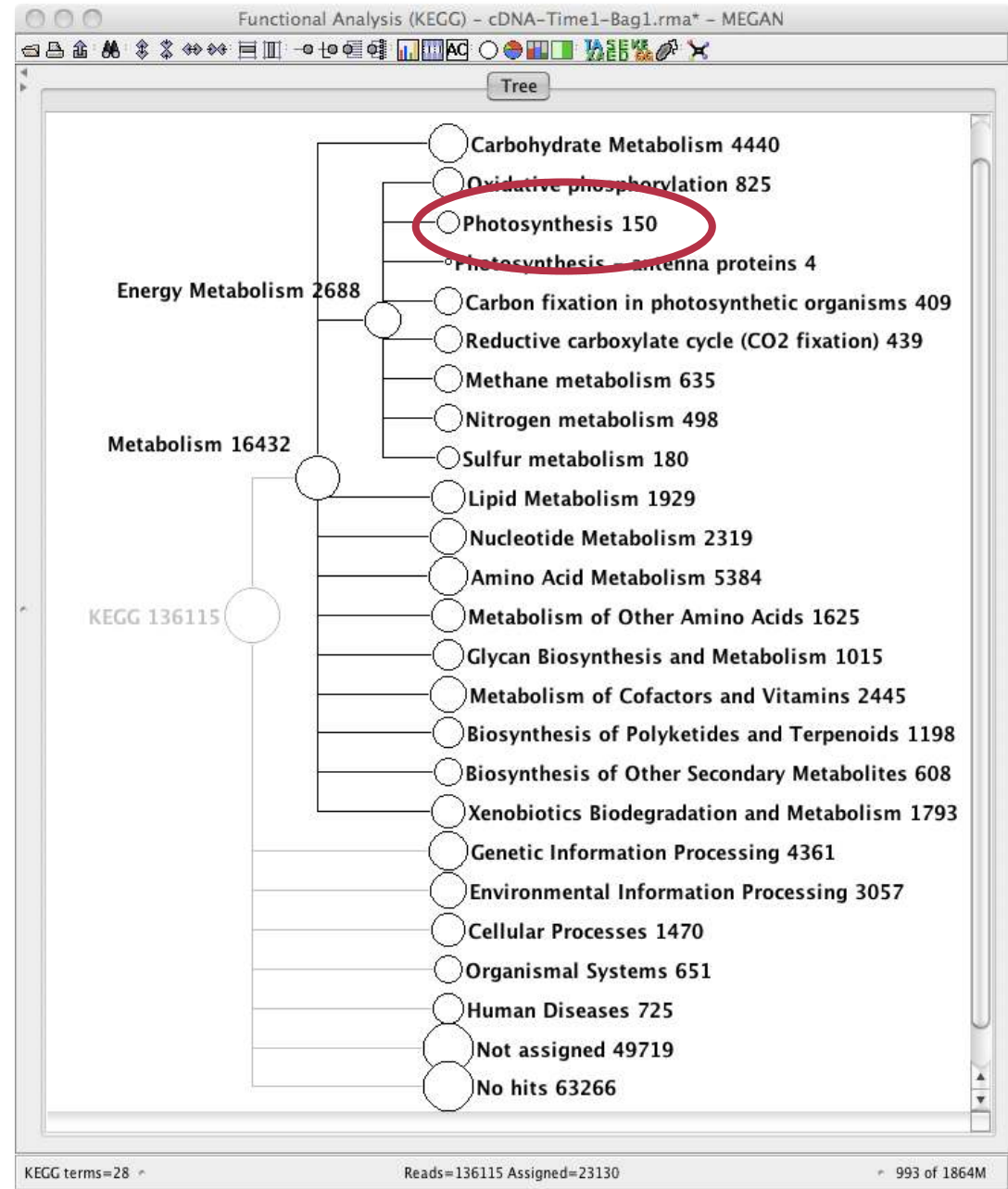
Details of Metabolism



KEGG analysis of 136 000 reads

KEGG Functional Analysis

Details of Energy Metabolism



KEGG analysis of 136 000 reads

KEGG Functional Analysis

Inspect or capture reads by assignment

Inspector - 155_PE_1_fixed.rma - [2] - MEGAN

155_PE_1_fixed.rma

- K02111 [30]
 - GDEG1CX11GVW2.1 [25]
 - GDEG1CX11GZ031.2 [25]
 - DATA [length=120]
 - Fusobacterium nucleatum subsp. polymorphum ATCC 10953 score=35.0
 - Fusobacterium nucleatum subsp. nucleatum ATCC 25586 score=35.0

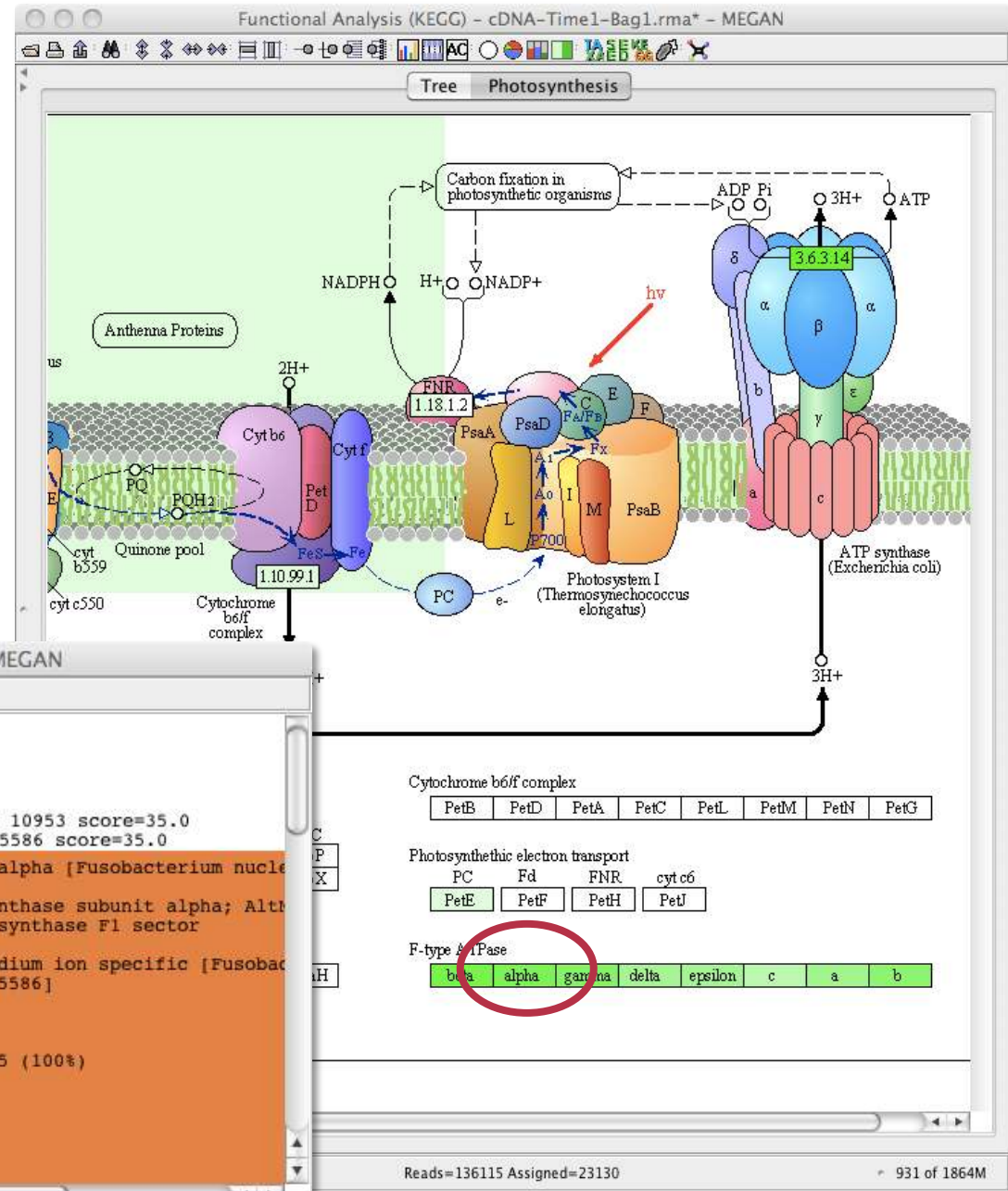
```

>ref|NP_603264.1| F0F1 ATP synthase subunit alpha [Fusobacterium nucleatum ATCC 25586]
sp|Q8RGE0.1|ATPA_FUSNN RecName: Full=ATP synthase subunit alpha; AltName: Full=ATP synthase F1 sector subunit alpha
gb|AAL94563.1| ATP synthase alpha chain, sodium ion specific [Fusobacterium nucleatum subsp. nucleatum ATCC 25586]
Length = 500

Score = 35.0 bits (79), Expect(2) = 0.023
Identities = 15/15 (100%), Positives = 15/15 (100%)
Frame = +1

Query: 40 KGQDVKCIYVAIGQK 84
      KGQDVKCIYVAIGQK
Sbjct: 187 KGQDVKCIYVAIGQK 201
  
```

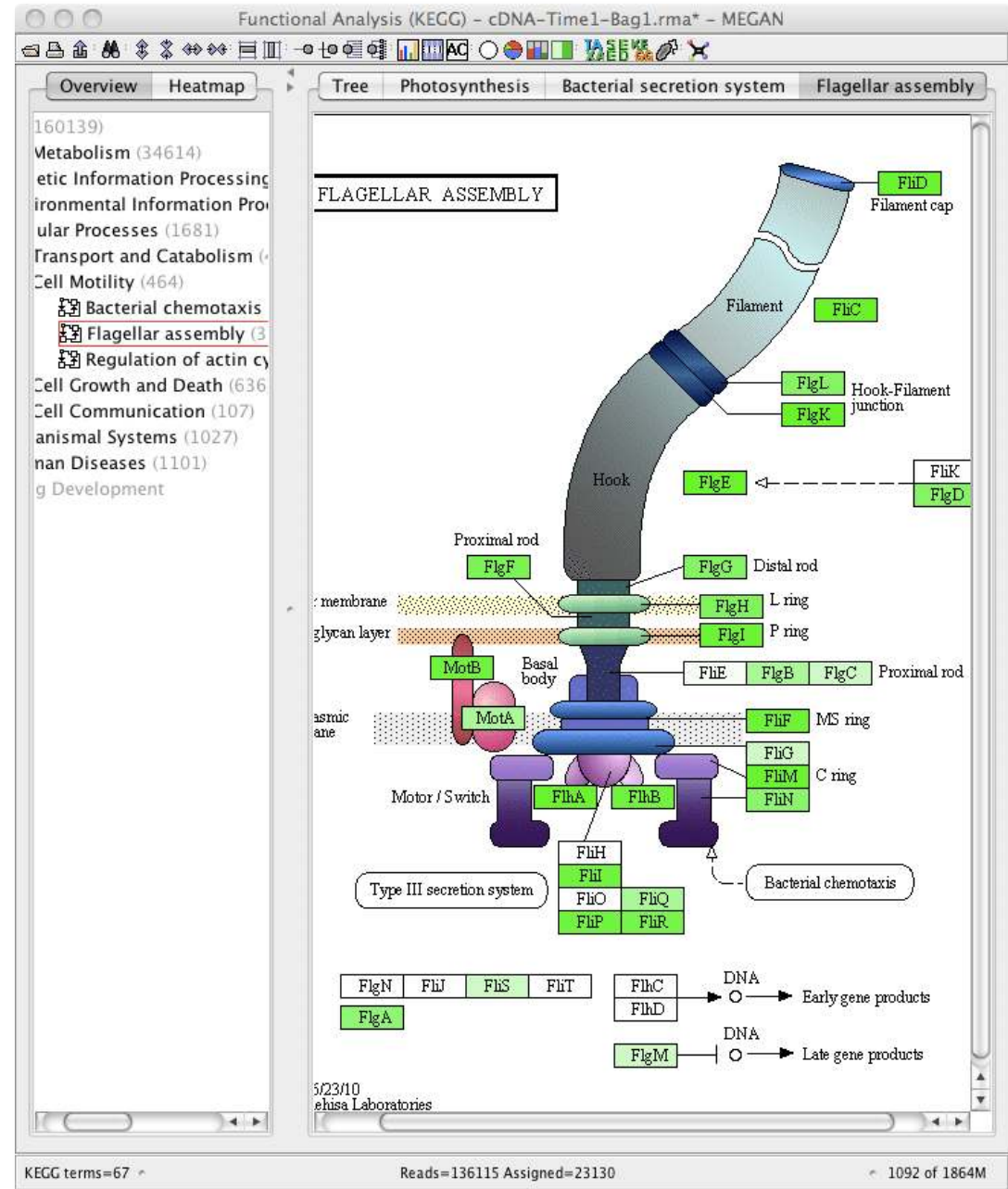
Rows: 59 1068 of 1864M



Analysis of 136 000 reads

KEGG Functional Analysis

Another example:
 Details of
 Flagellar assembly



KEGG analysis of 136 000 reads



Mapping Reads To Functional Roles

Use RefSeq-to-SEED mapping file (1.4 million entries)

Example:

Lookup
ref2seed.map

```
>ref|YP_585064.1| molecular chaperone DnaK [Ralstonia metallidurans CH34]
sp|O33522.2|DNAK_RALME RecName: Full=Chaperone protein dnaK;
AltName: Full=Heat shock
protein 70; AltName: Full=Heat shock 70 kDa protein; AltName:
Full=HSP70
gb|ABF09795.1| Chaperone DnaK [Ralstonia metallidurans CH34]
Length=648

Score = 34.3 bits (77), Expect = 2.3, Method: Composition-based
stats.
Identities = 14/15 (93%), Positives = 14/15 (93%), Gaps = 0/15 (0%)

Query 1      MLGTFNLEGIAPAPR 15
      MLG FNLEGIAPAPR
Sbjct 454    MLGEFNLEGIAPAPR 468
```

BLAST match

→ “Chaperone protein DnaK”



Mapping Reads To KEGG Orthologs

Use RefSeq-to-KEGG mapping file (2.1 million entries)

Example:

Lookup
ref2kegg.map

```
>ref|YP_585064.1| molecular chaperone DnaK [Ralstonia metallidurans CH34]
sp|O33522.2|DNAK_RALME RecName: Full=Chaperone protein dnaK;
AltName: Full=Heat shock
protein 70; AltName: Full=Heat shock 70 kDa protein; AltName:
Full=HSP70
gb|ABF09795.1| Chaperone DnaK [Ralstonia metallidurans CH34]
Length=648

Score = 34.3 bits (77), Expect = 2.3, Method: Composition-based
stats.
Identities = 14/15 (93%), Positives = 14/15 (93%), Gaps = 0/15 (0%)

Query 1      MLGTFNLEGIAPAPR 15
      MLG FNLEGIAPAPR
Sbjct 454    MLGEFNLEGIAPAPR 468
```

BLAST match

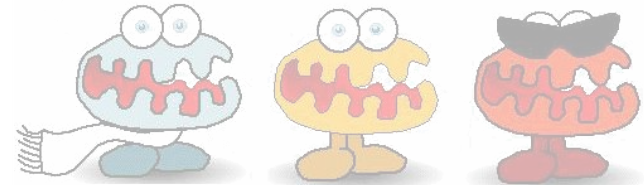
→ **K04043 = “molecular chaperone DnaK”**



Three Basic Computational Questions

Who is out there?

Types of organisms
In what proportions?



www.compostinfo.com/tutorial/microbes.htm

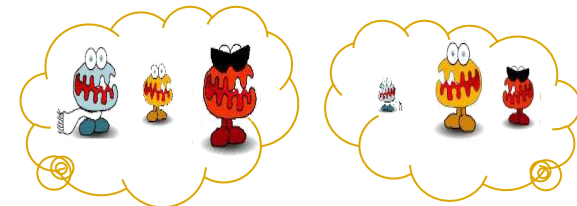
What are they doing?

Types of genes
Which metabolic pathways?
In what proportions?



How do different samples compare?

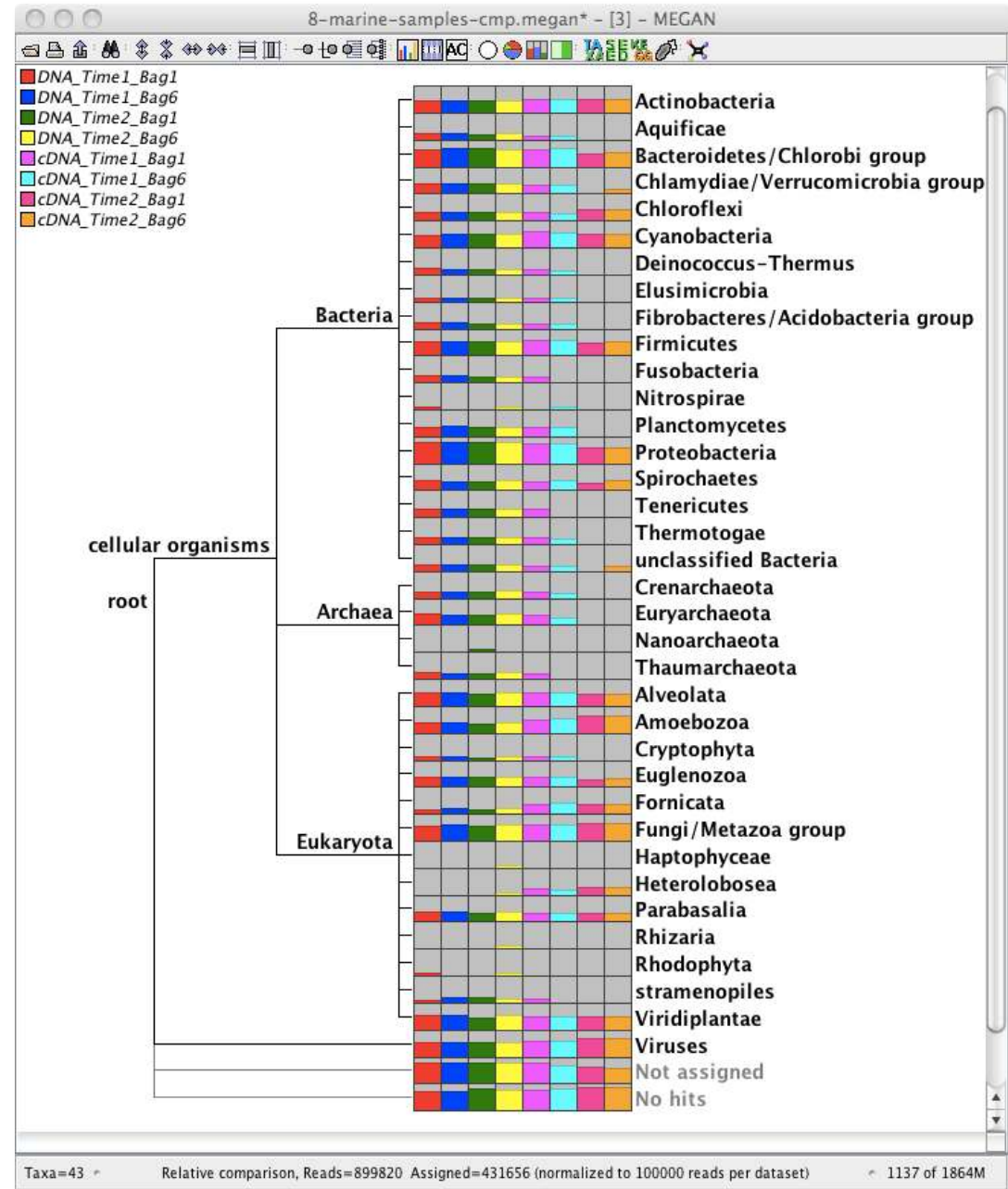
Pairwise and multiple comparisons
Correlations with environmental
parameters?



Serve to answer biological or medical questions

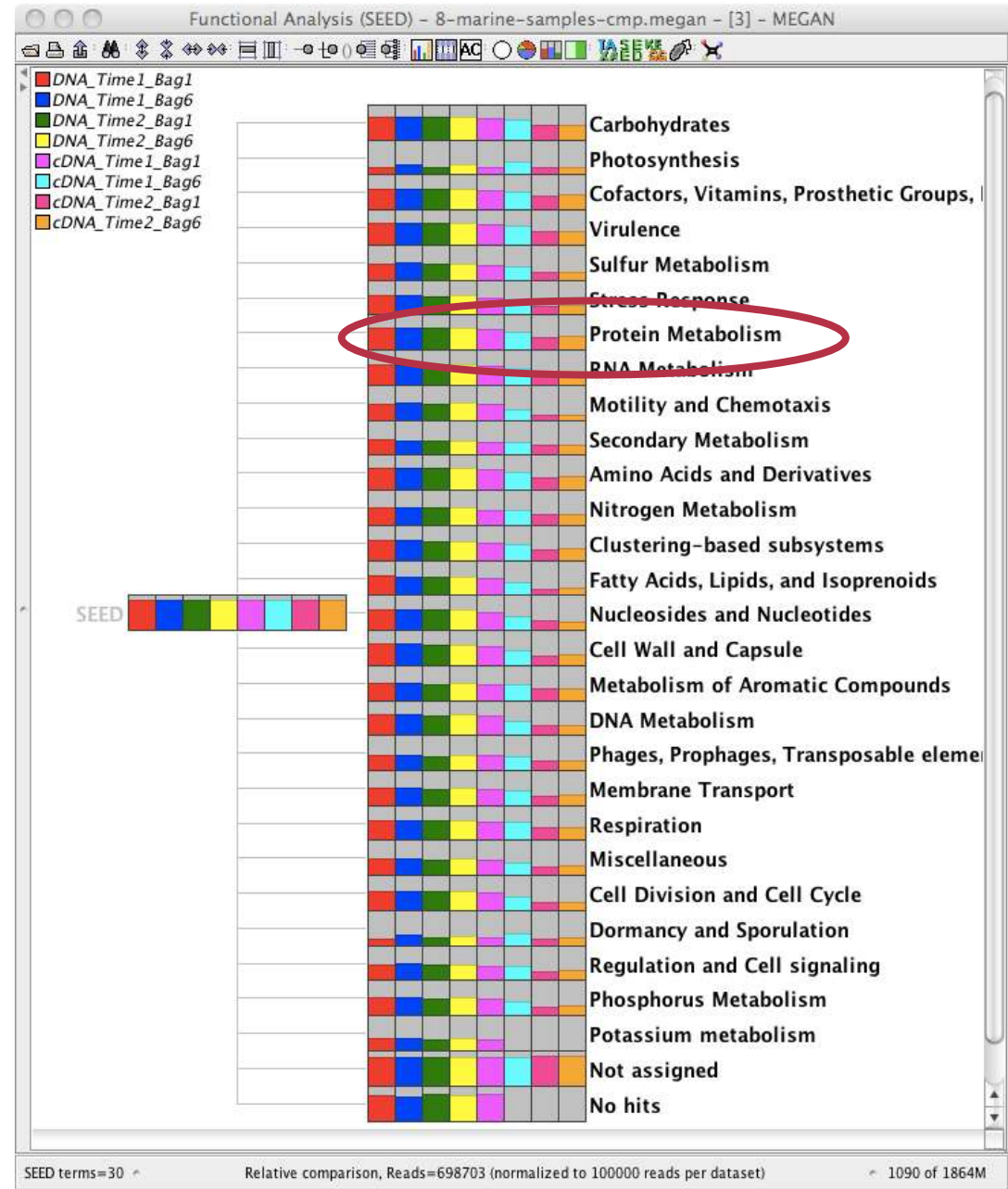
Taxonomic Comparison

Simultaneously
analyze multiple
datasets



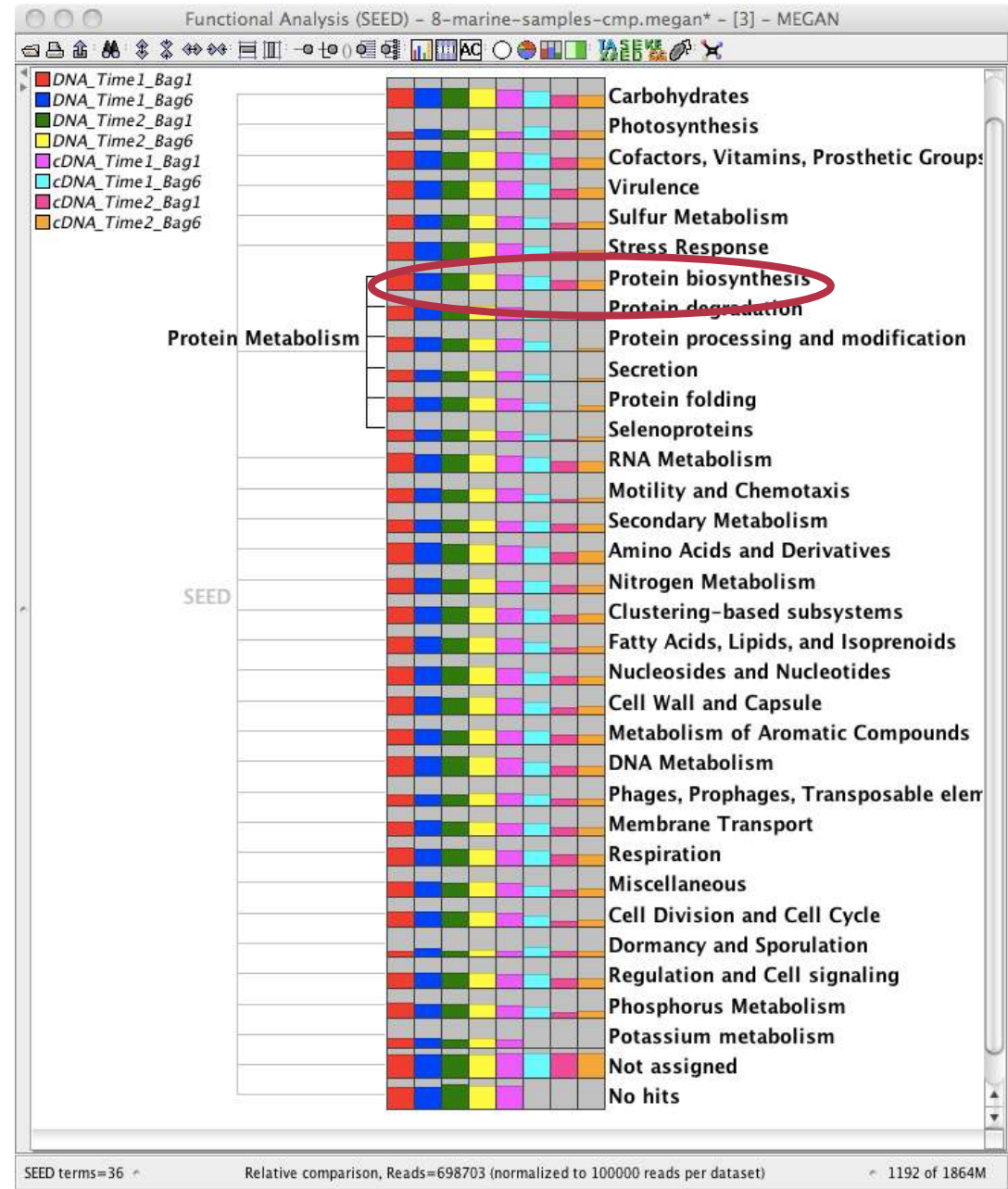
SEED-based Comparison

Simultaneously
analyze multiple
datasets



SEED-based Comparison

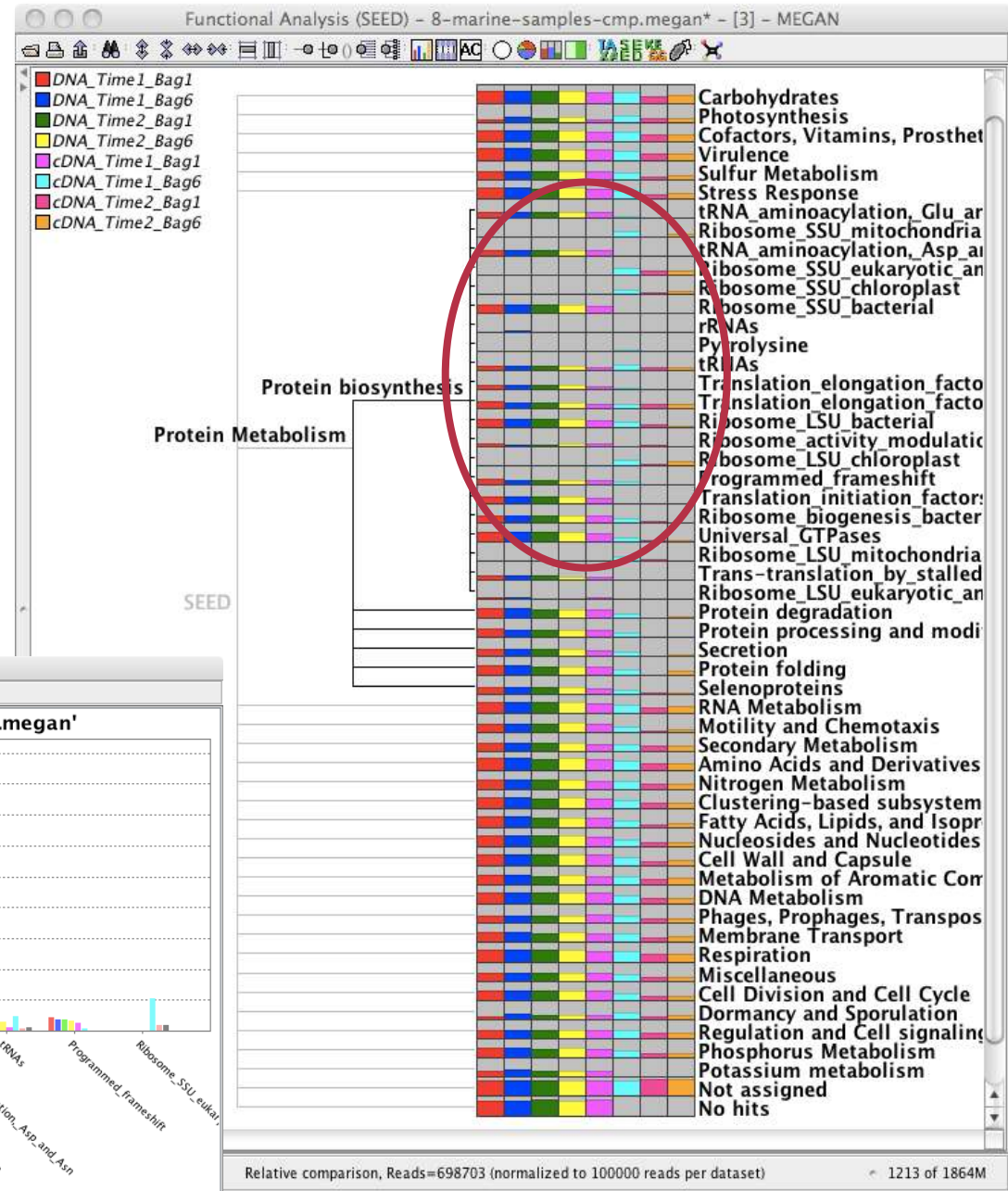
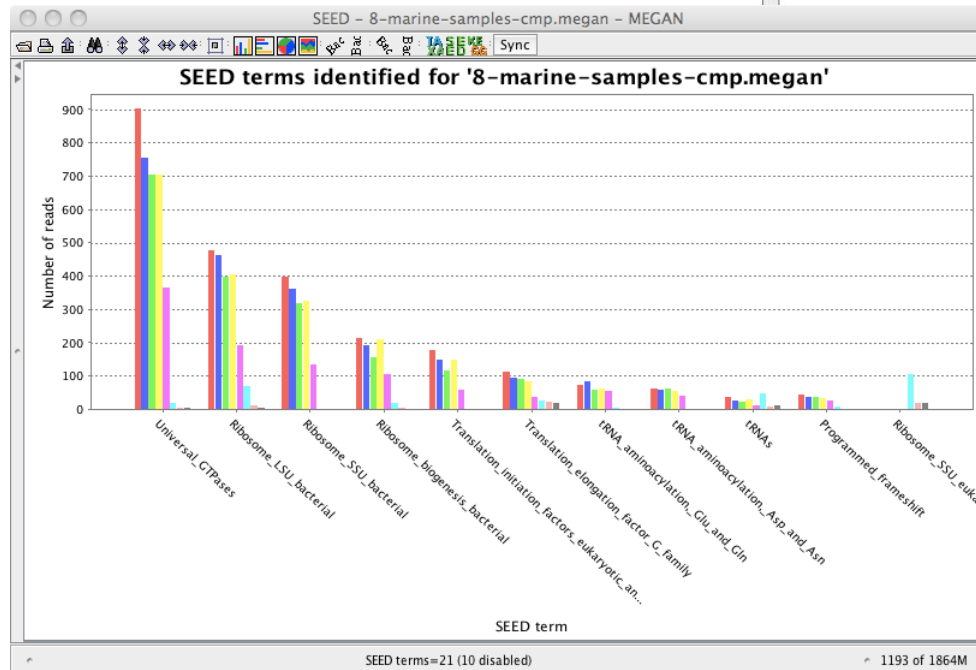
Details of Protein Metabolism





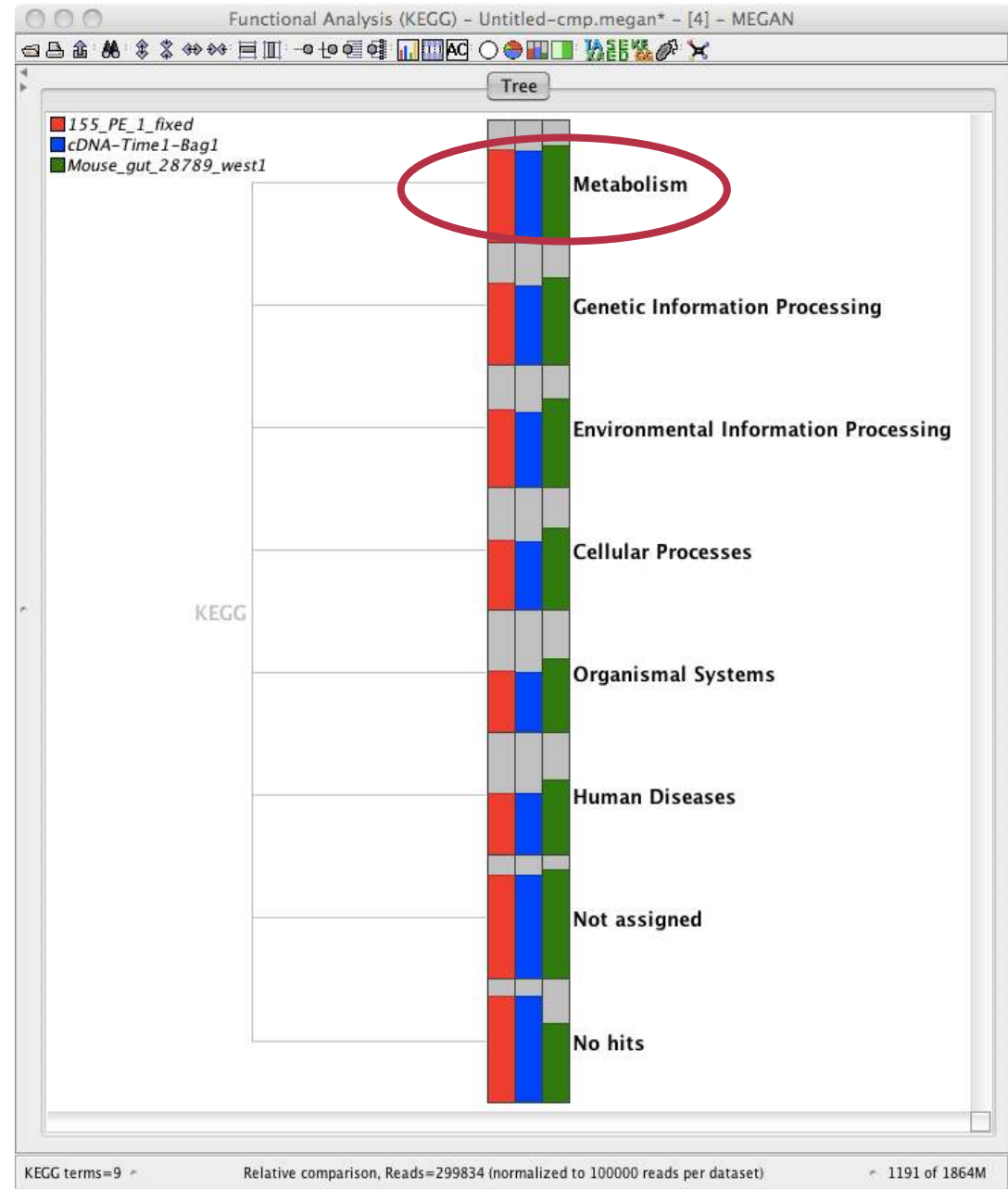
SEED-based Comparison

Simultaneously
analyze multiple
datasets



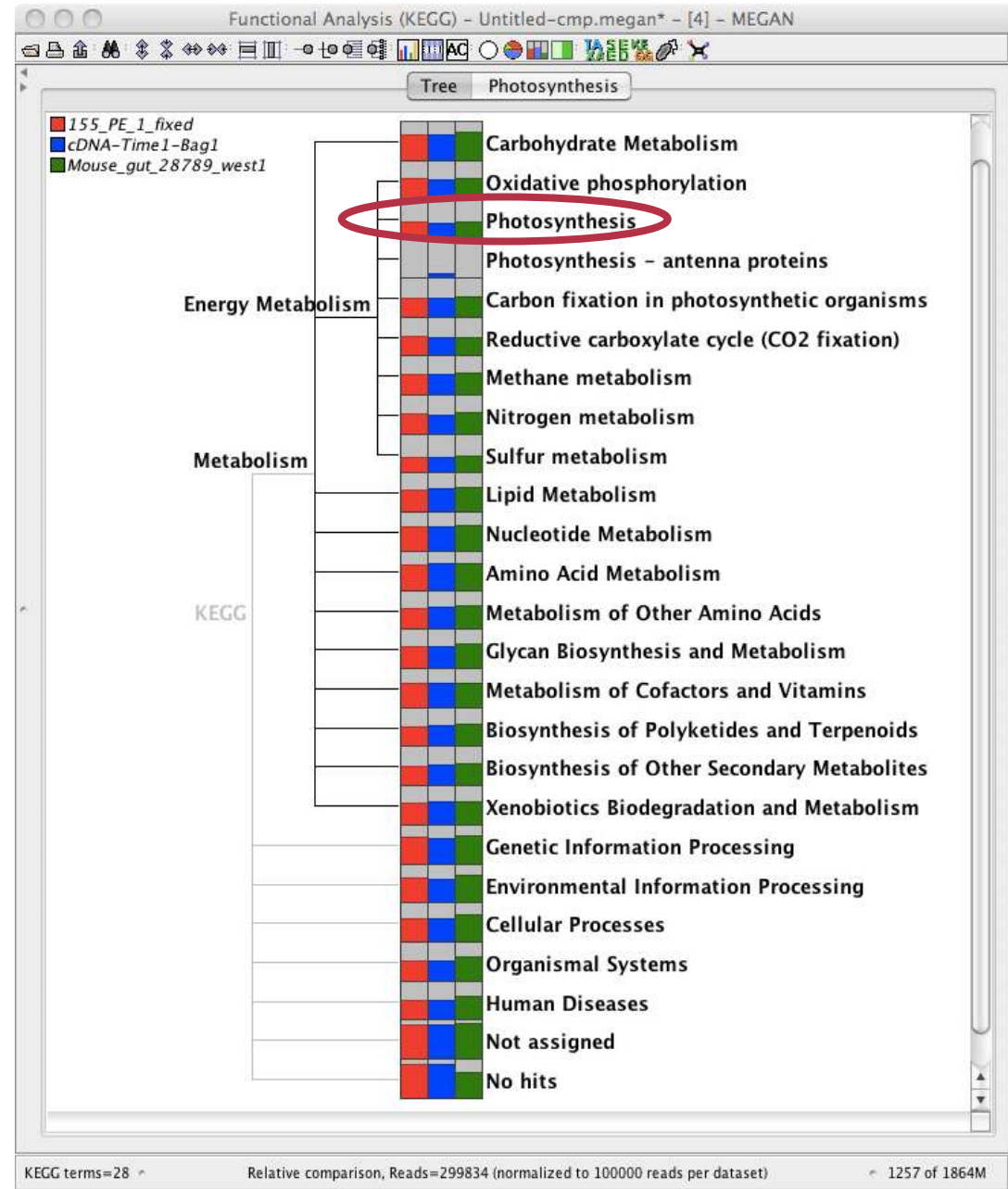
KEGG-based Comparison

Simultaneously
analyze multiple
datasets



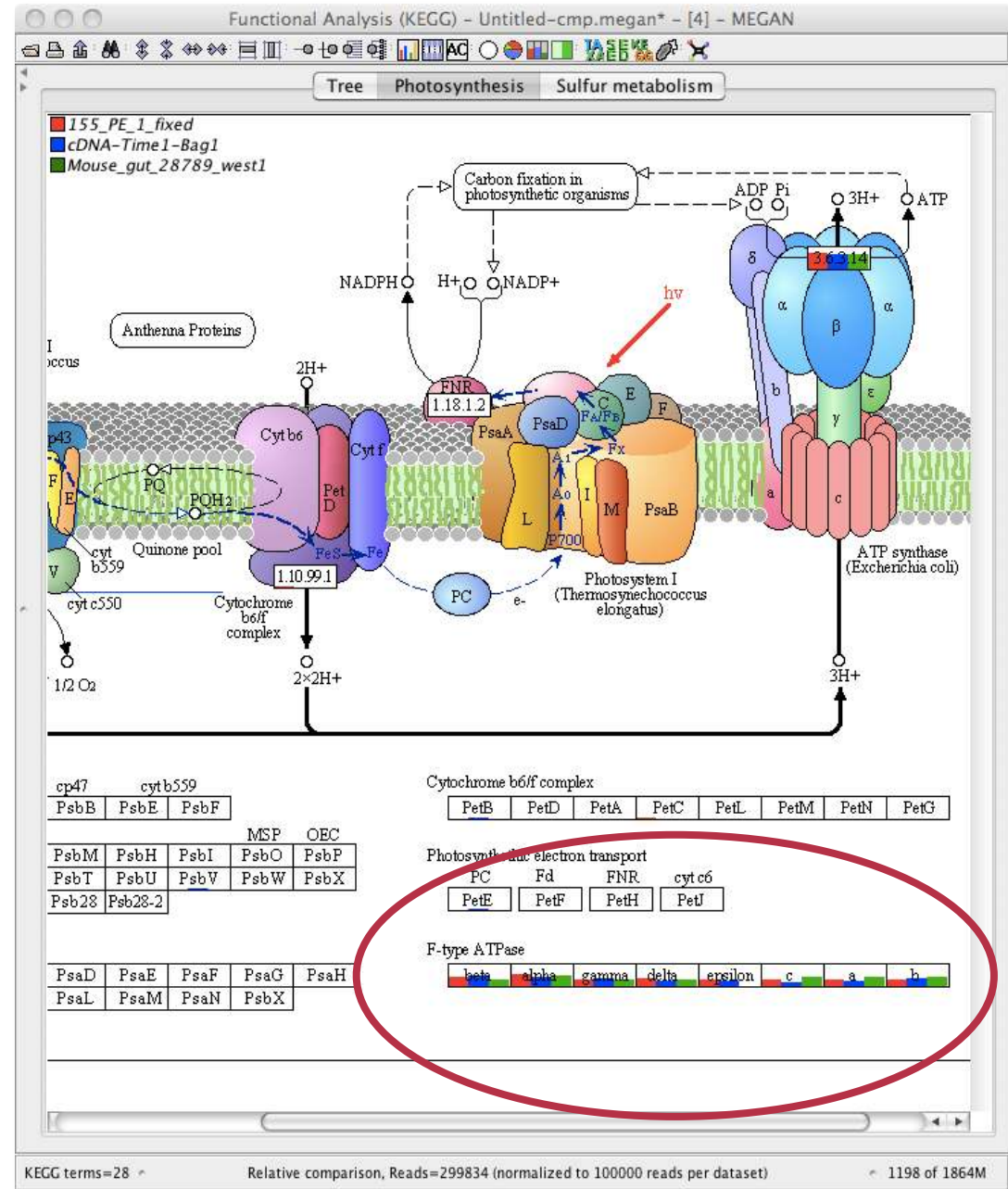
KEGG-based Comparison

Details of Energy Metabolism



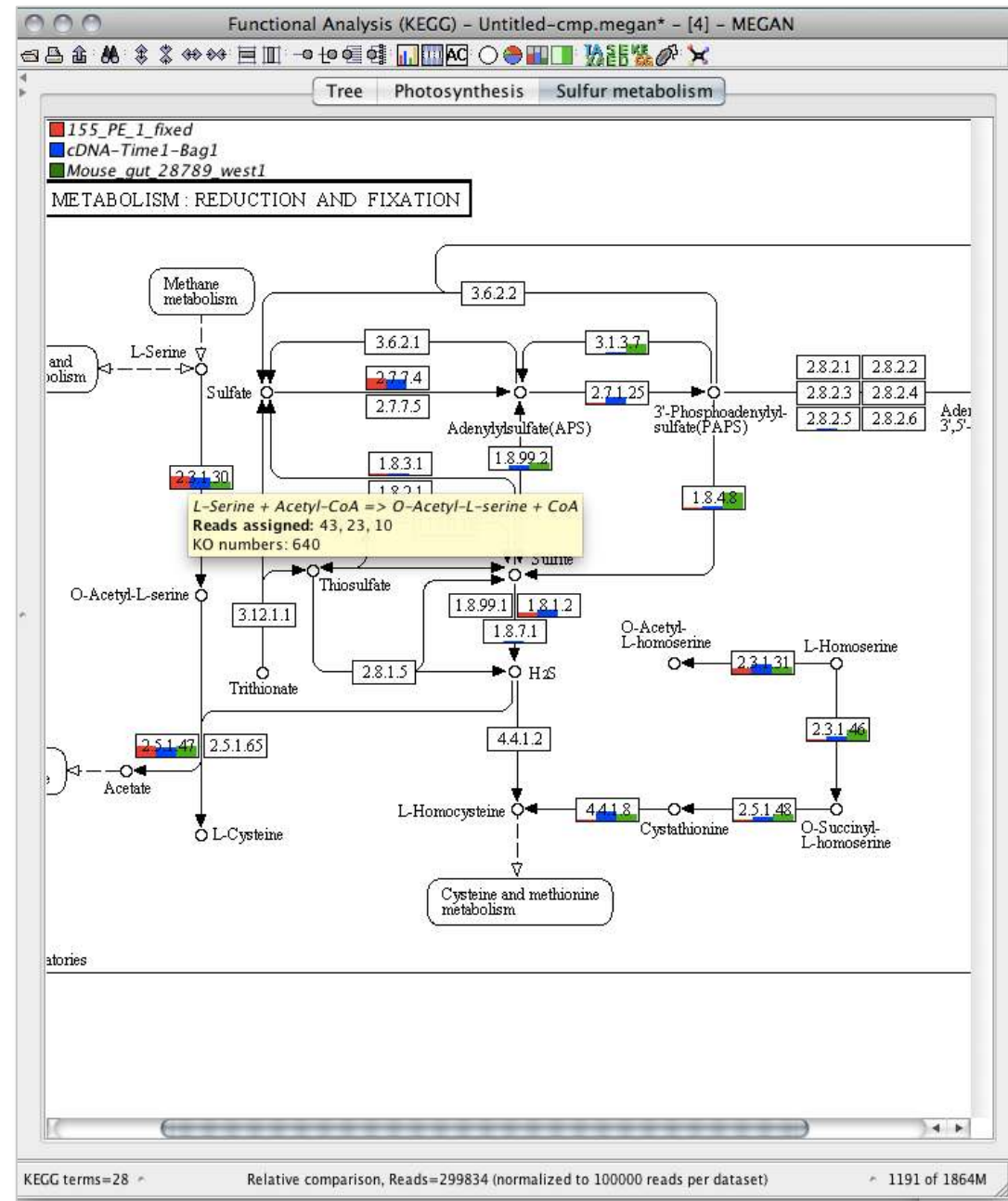
KEGG-based Comparison

Details of
Photosynthesis
KEGG pathway



KEGG-based Comparison

Details of Sulfur metabolism pathway





Compare

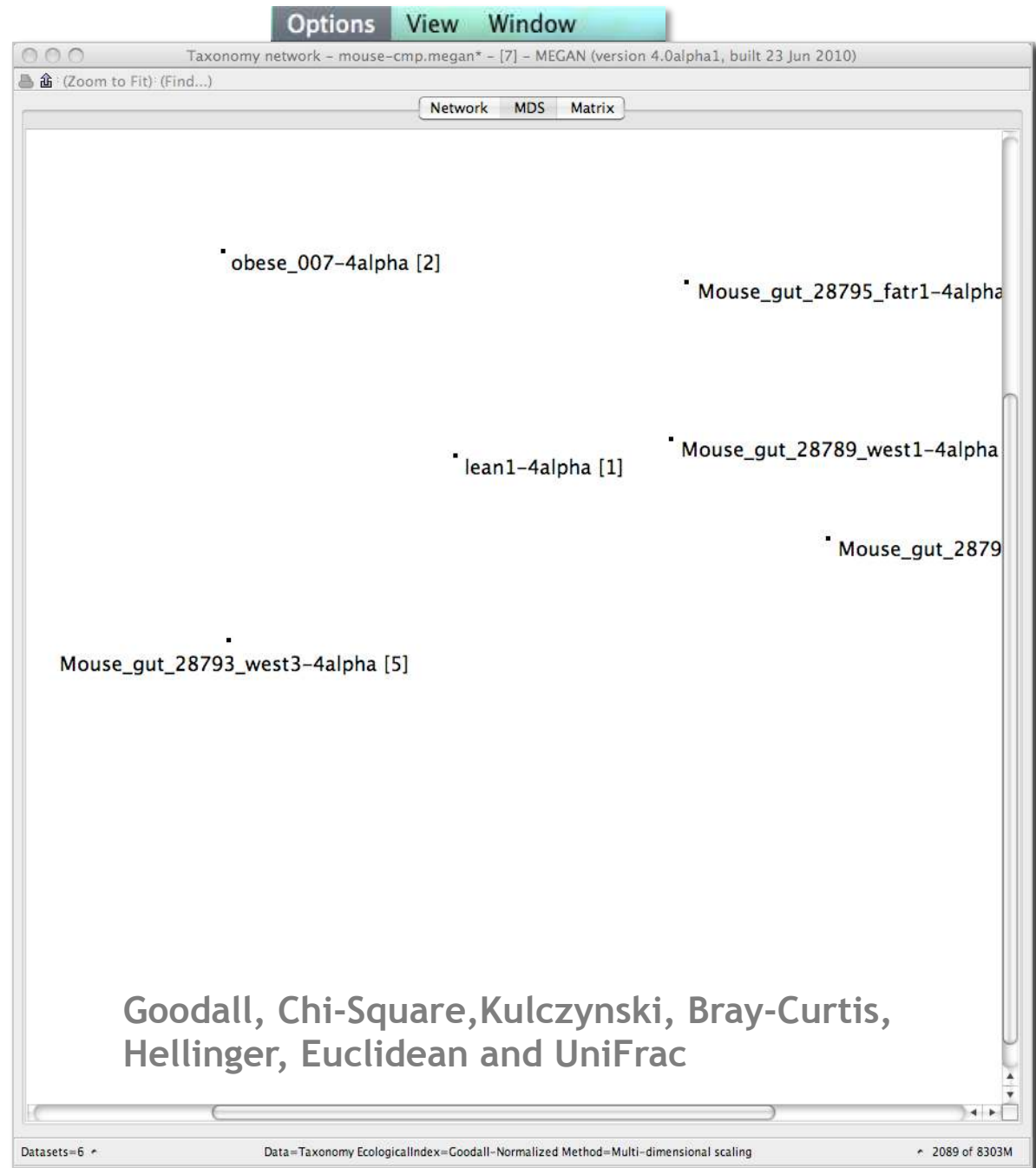
High-level comparison:

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

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

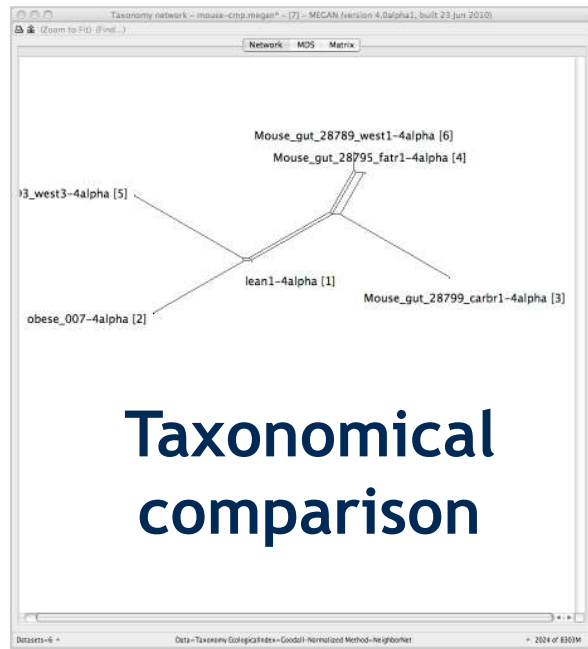
Neighbor-net:
Bryant and Moulton, 2003

Daniel Huson © 2010



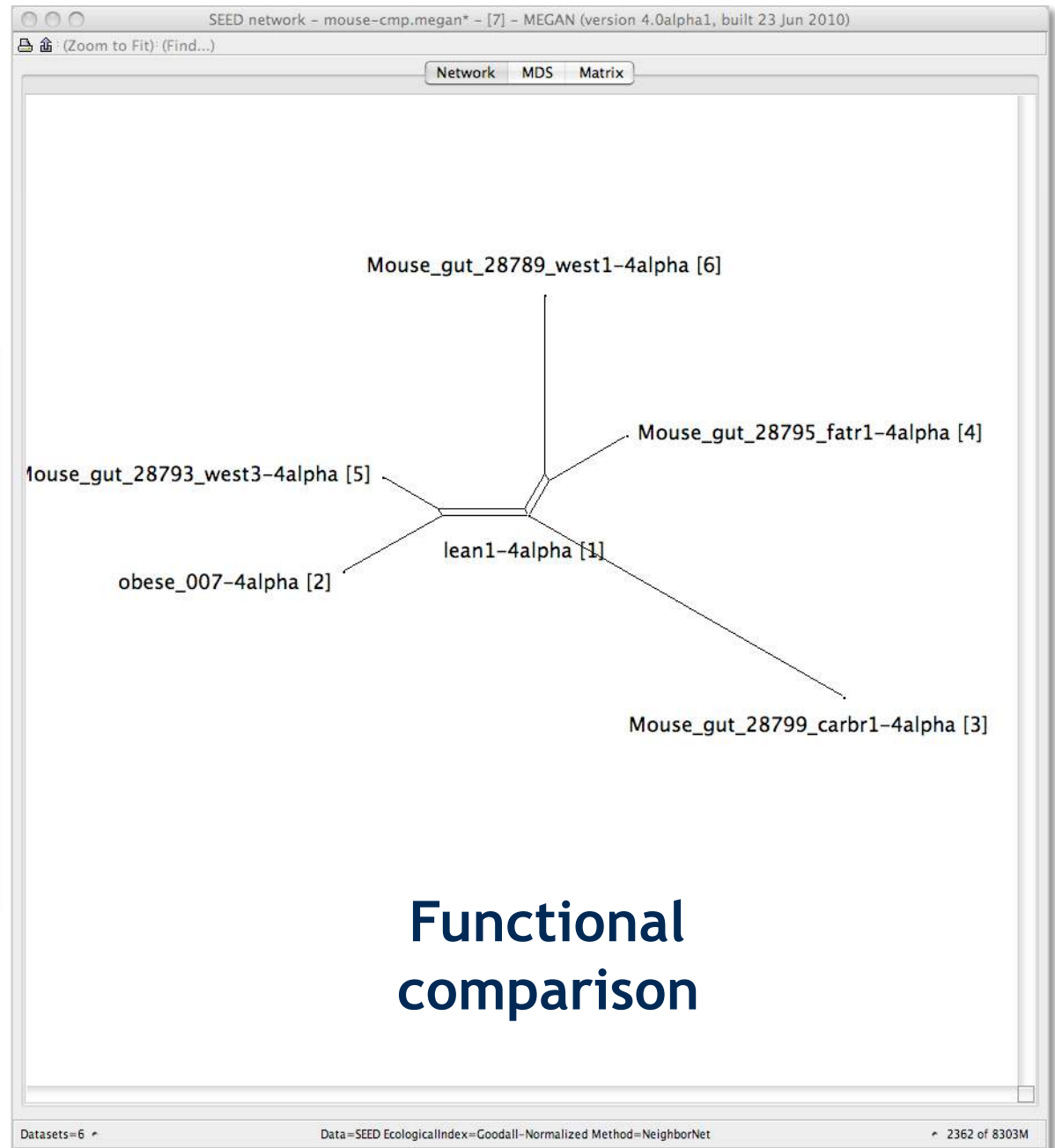
Compare

High-level comparison:



**Taxonomical
comparison**

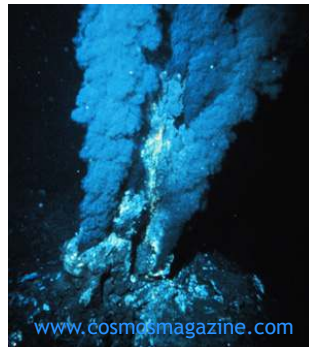
✓ Select functions...



**Functional
comparison**

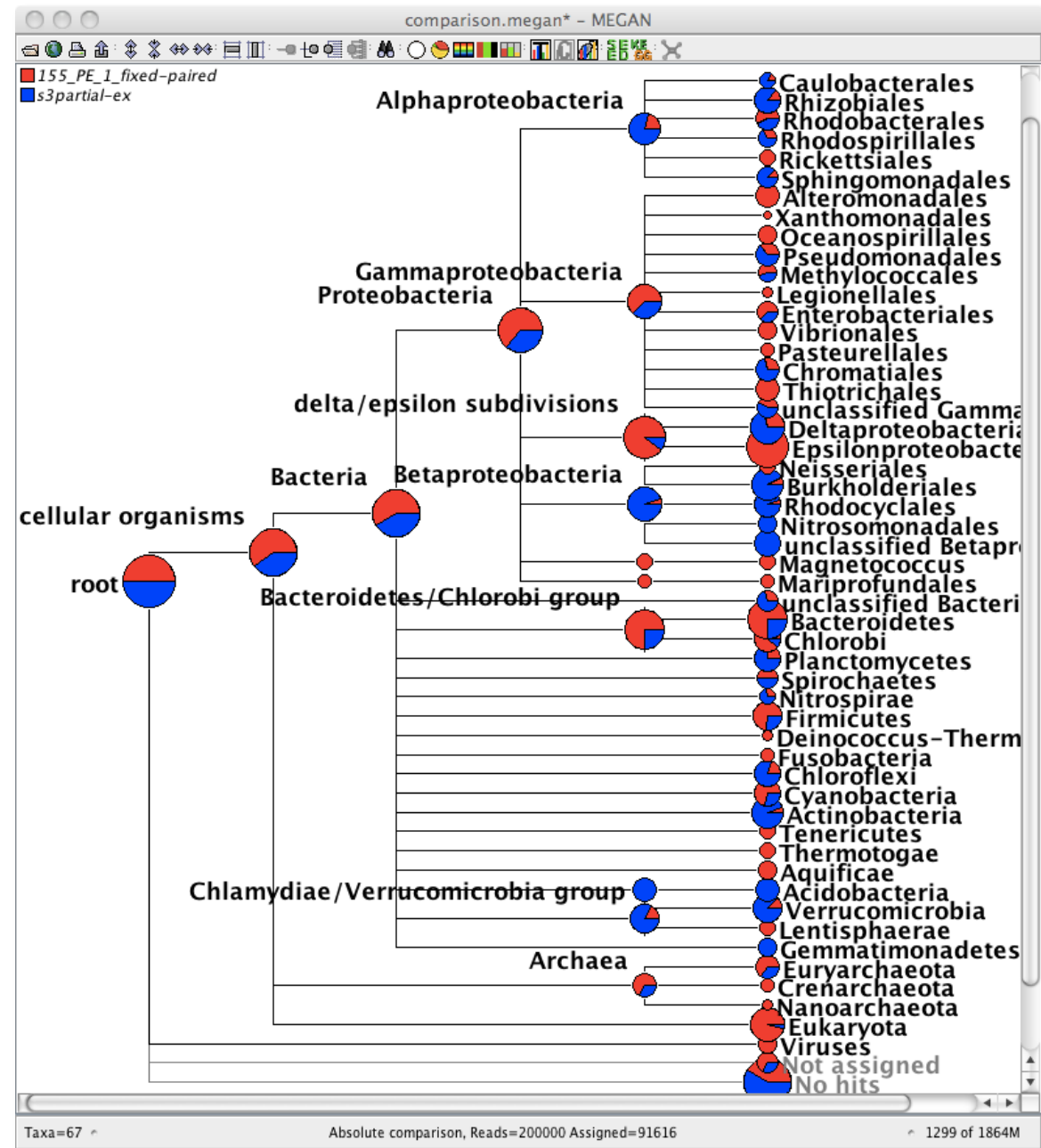
Example

Deep sea hot vent
dataset



VS

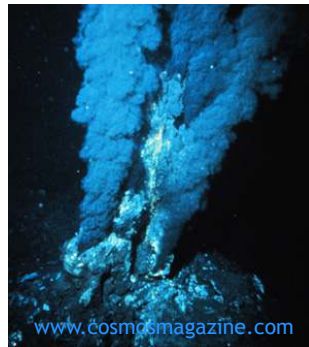
SCElse sludge



Taxonomic analysis

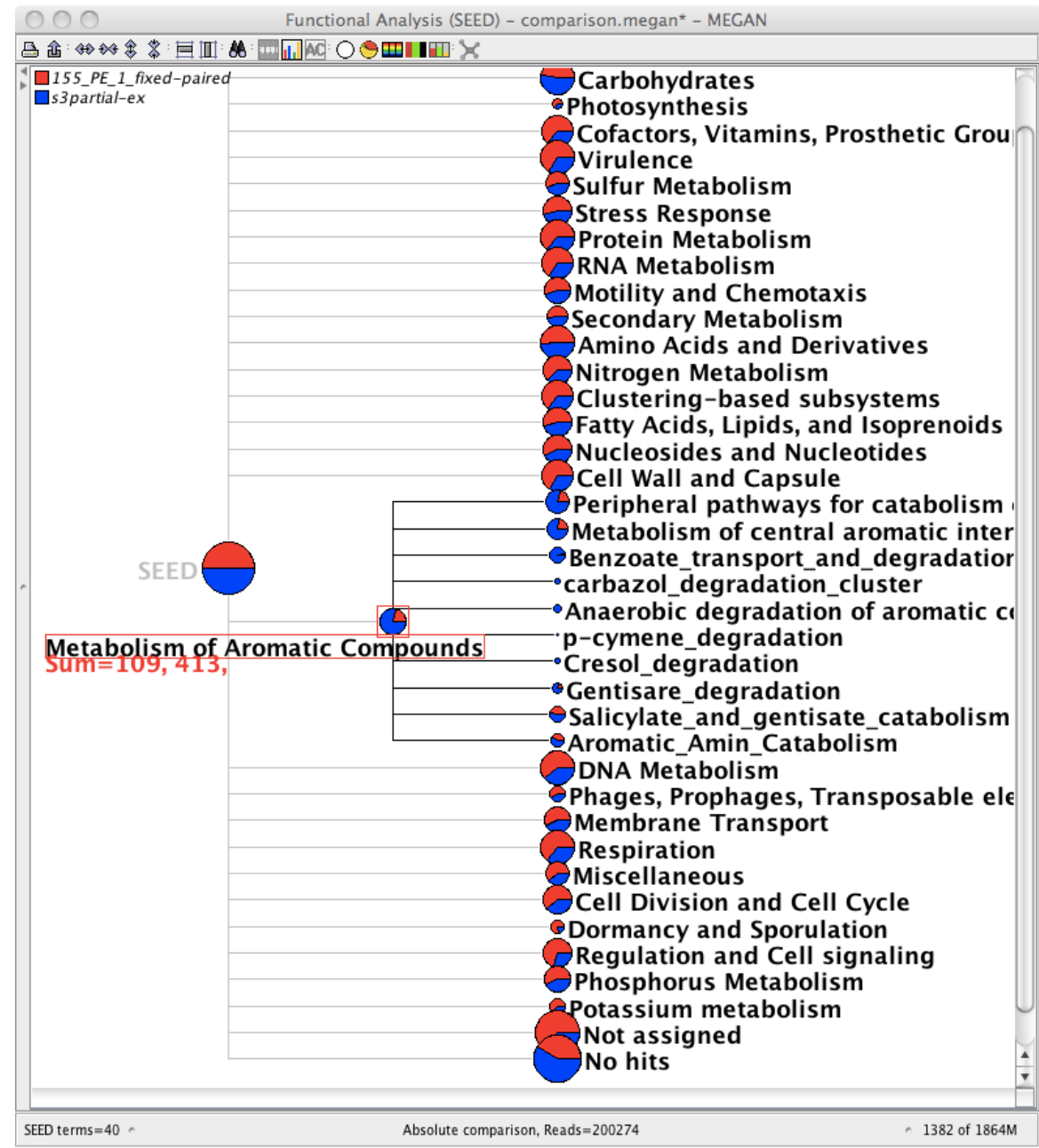
Example

Deep sea hot vent
dataset



VS

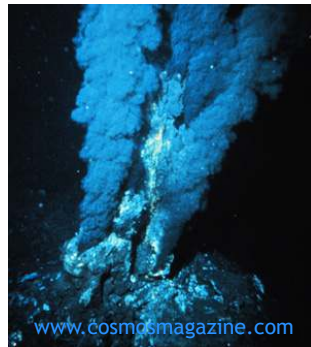
SCElse sludge



SEED analysis

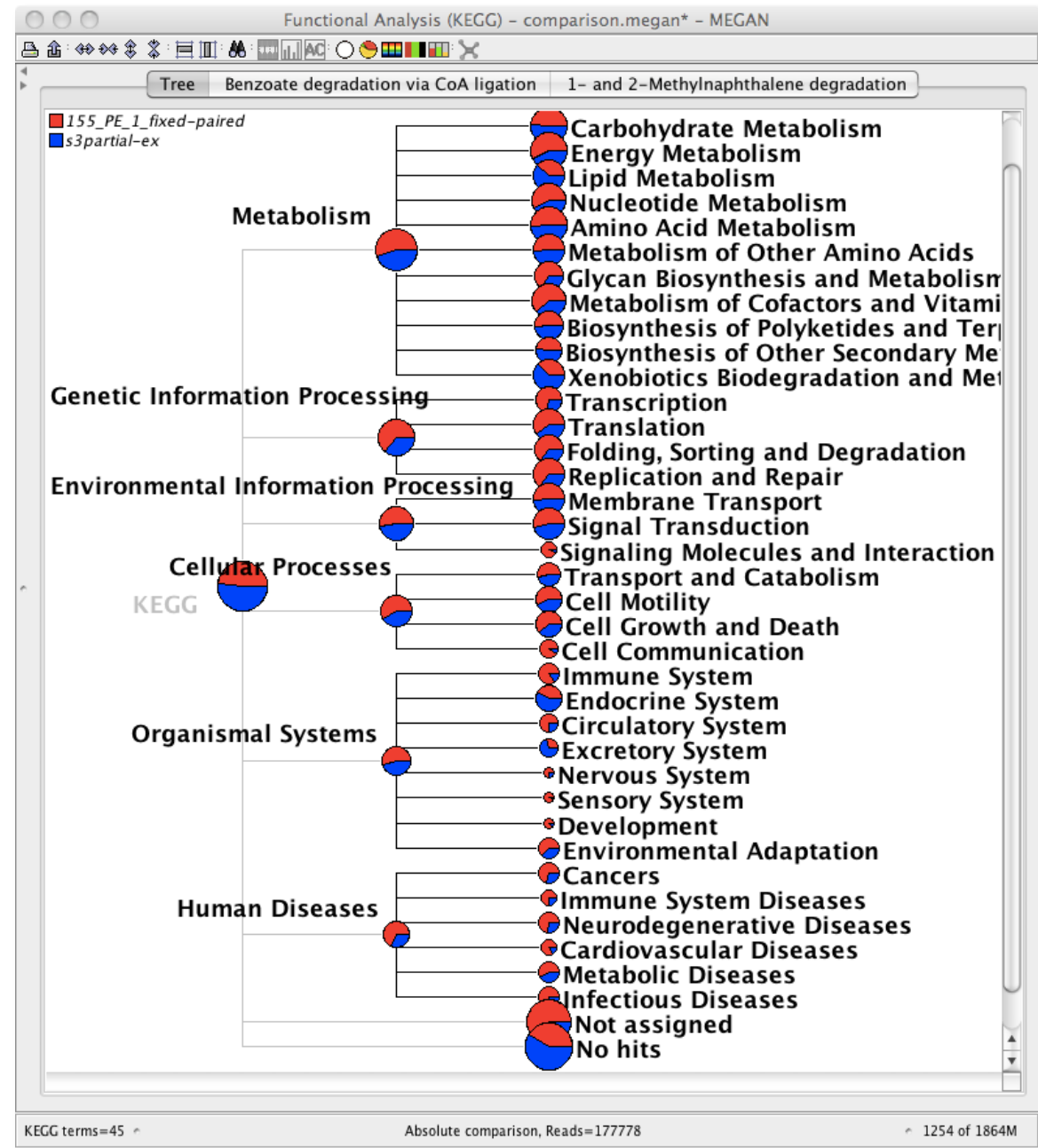
Example

Deep sea hot vent
dataset



VS

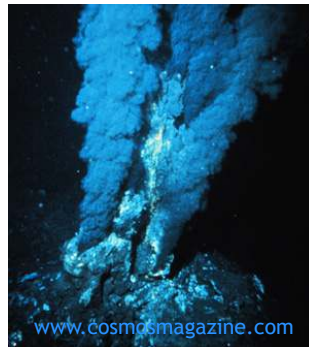
SCElse sludge



KEGG analysis

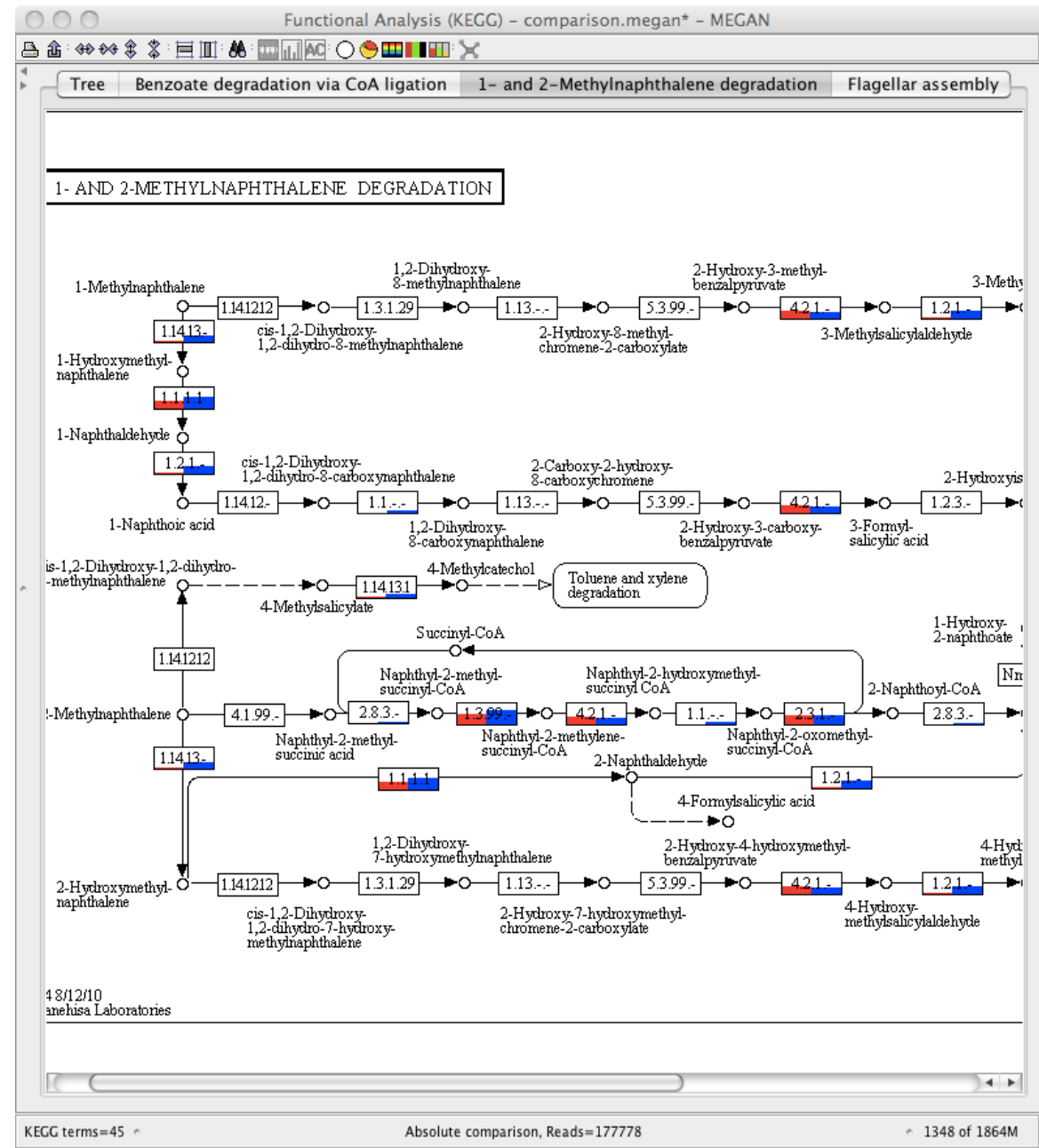
Example

Deep sea hot vent
dataset



VS

SCElse sludge



KEGG pathway analysis



Contents

Genomics

Sequencing

Metagenomics

Computational questions

Outlook



Computational Challenges...

Global Ocean Sampling

www.jcvi.org/cms/research/projects/gos

Human Microbiome Project

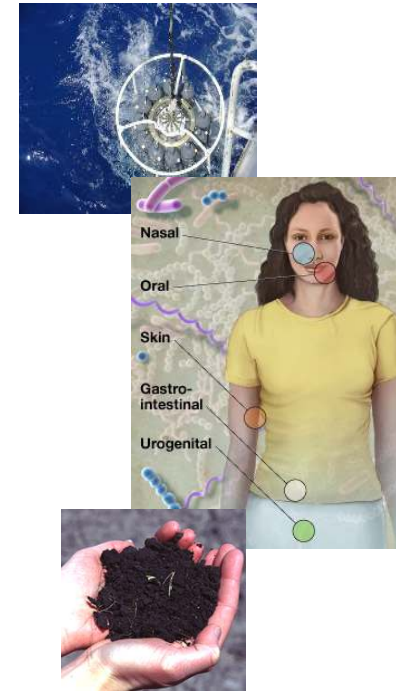
nihroadmap.nih.gov/hmp/

Terragenome Consortium

www.terragenome.org

SCELS

Singapore centre for life science engineering



Terabases 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

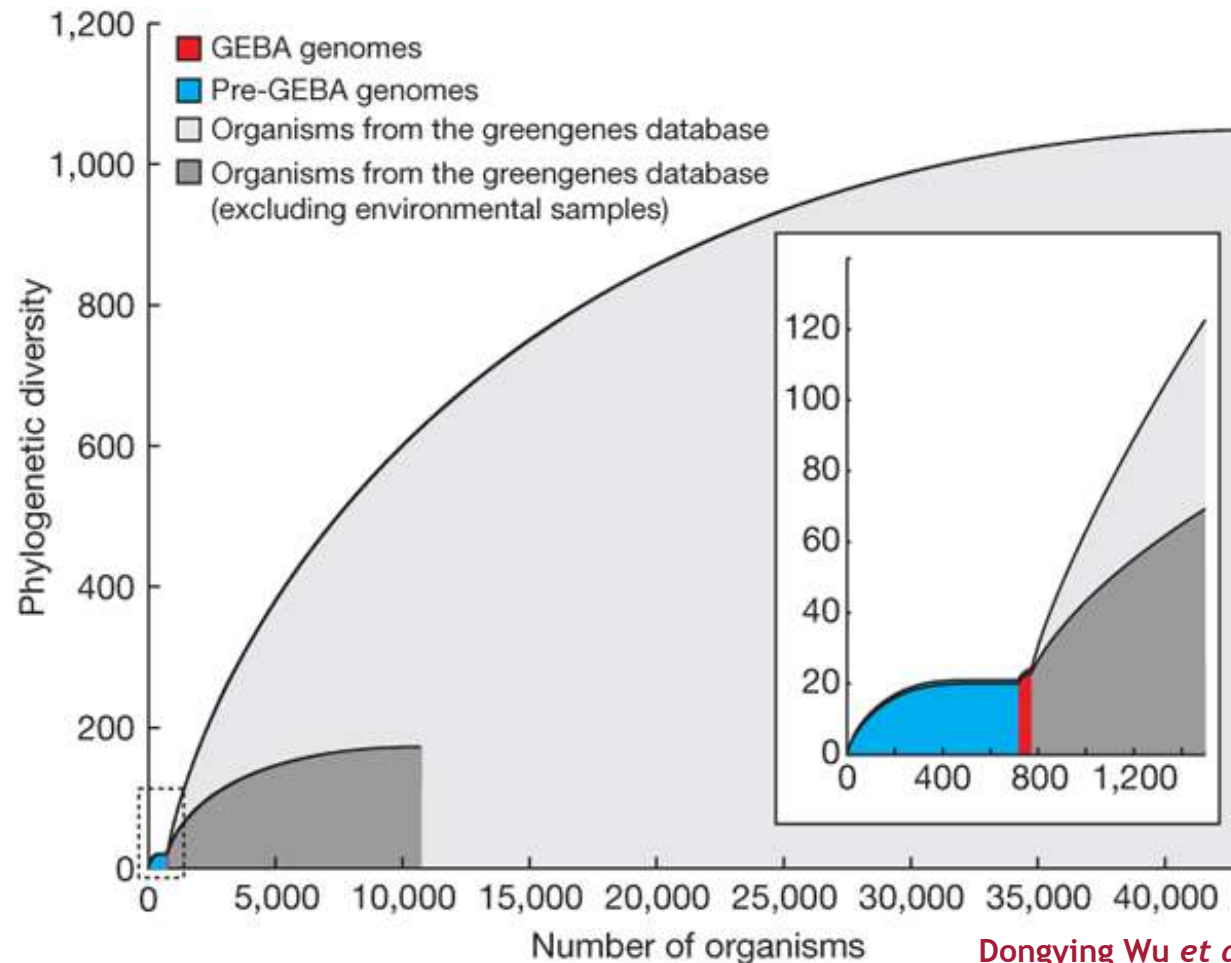
Ever faster analysis methods

How to learn across multiple datasets?



Challenges

Need substantially more reference genomes
and much better understanding of diversity





Joint Work With:

Suparna Mitra, Daniel Richter, Hans Ruscheweyh,

Paul Rupek, Nico Weber & Max Schubach

Stephan Schuster, Fangqing Zhao & Qi Ji

Christa Schleper and Tim Urich

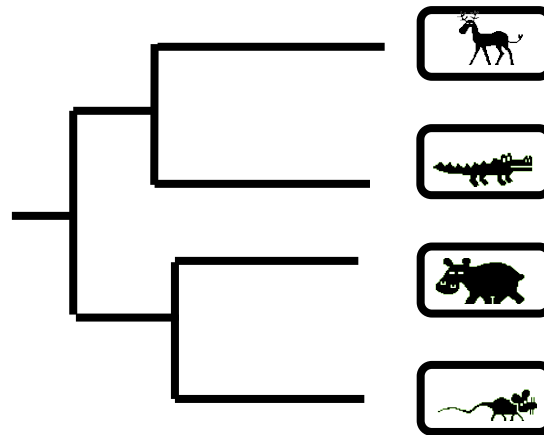
Ida Steen and Anders Lanzen

Tübingen

Penn State

Vienna

Bergen



www-ab.informatik.uni-tuebingen.de