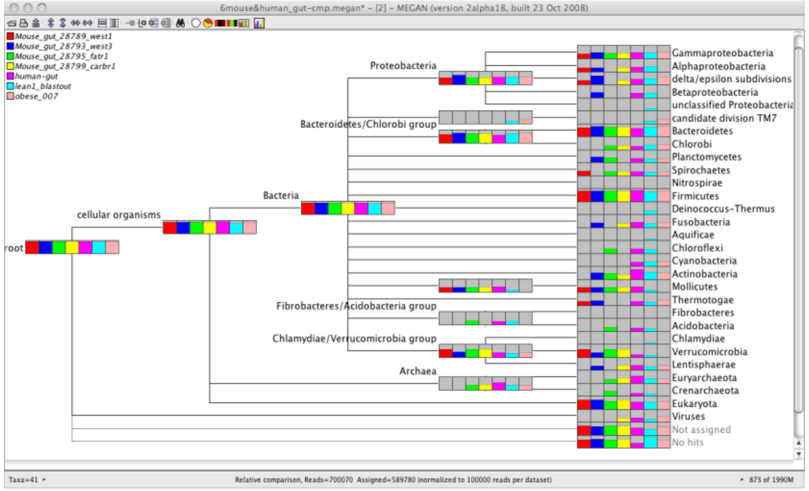
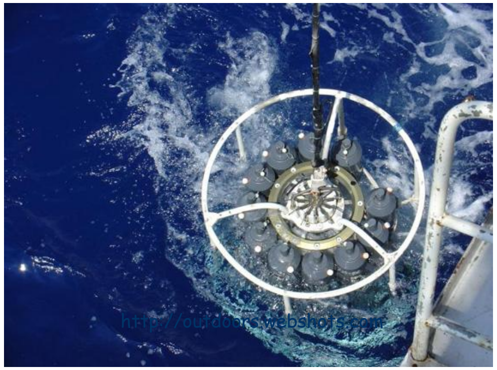
Computational Analysis of Metagenomes


6mouse&human_gut-cmp.megan* – [2] – MEGAN (version 2alpha18, built 23 Oct 2008)

Taxa=41 Relative comparison, Reads=700070 Assigned=589780 (normalized to 100000 reads per dataset) 873 of 1990M


http://outdoors.webshots.com

Daniel H. Huson

PRIB Nijmegen 2010

Eberhard Karls
Universität
Tübingen

Contents

- Genomics
- Sequencing
- Metagenomics
- Computational questions
- Outlook

Contents

- Genomics
- Sequencing
- Metagenomics
- Computational questions
- Outlook

Discovery of DNA

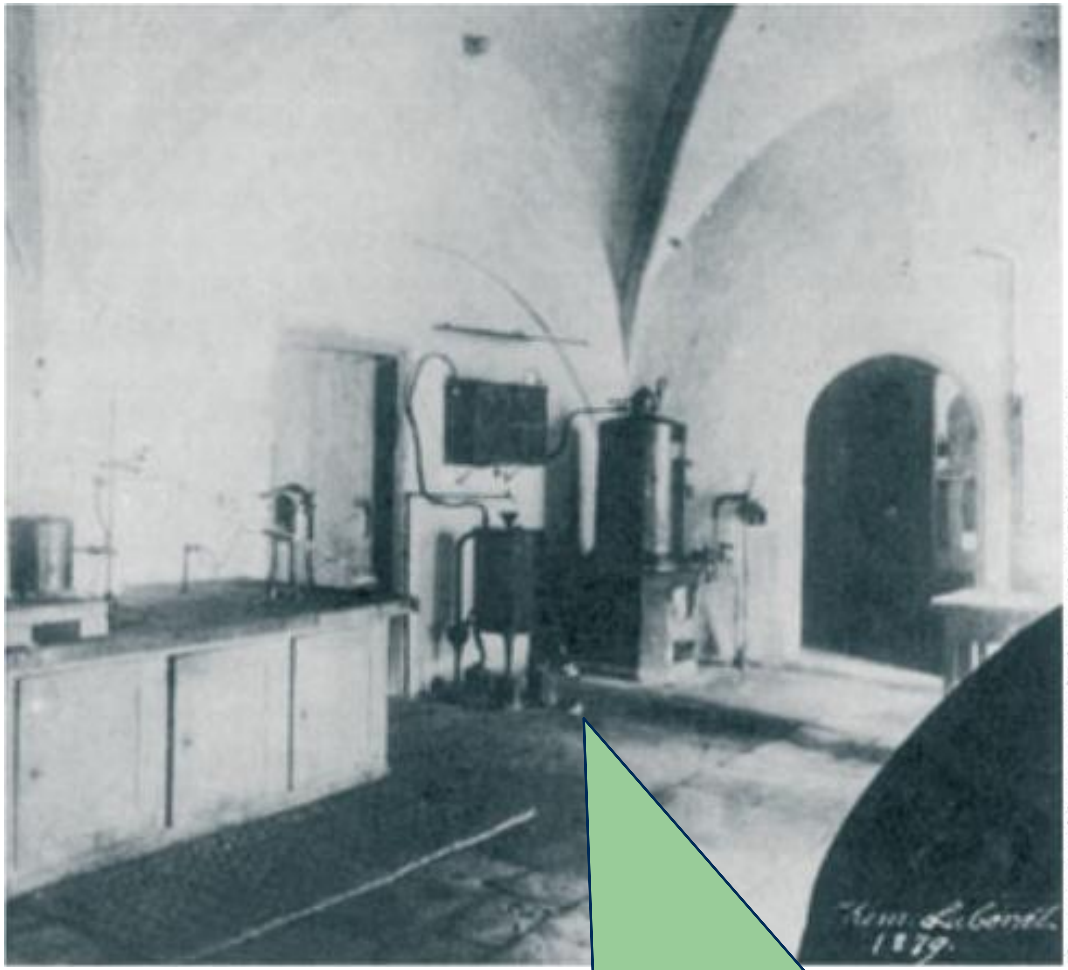
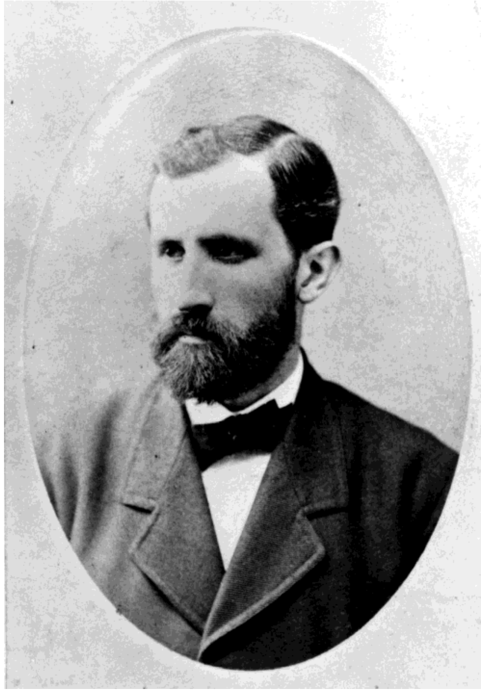


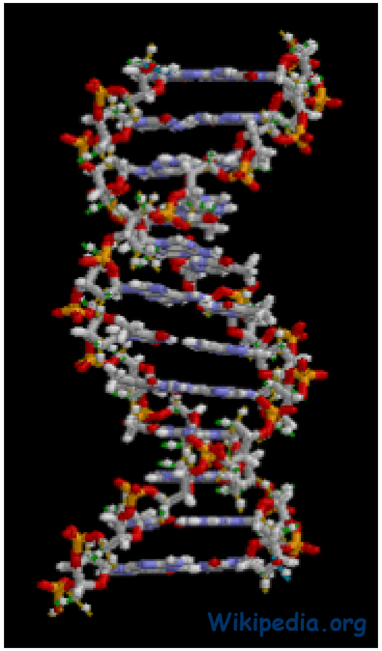
Photo courtesy of the University of Tübingen Library, Tübingen, Germany

1869: Miescher discovered DNA in the kitchen of Tübingen Castle



Friedrich Miescher
(1844-1895)

Role of DNA

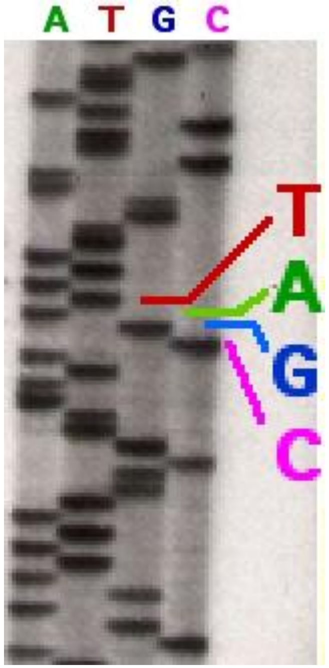


...
A – T
C – G
G – C
T – A
A – T
A – T
...

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



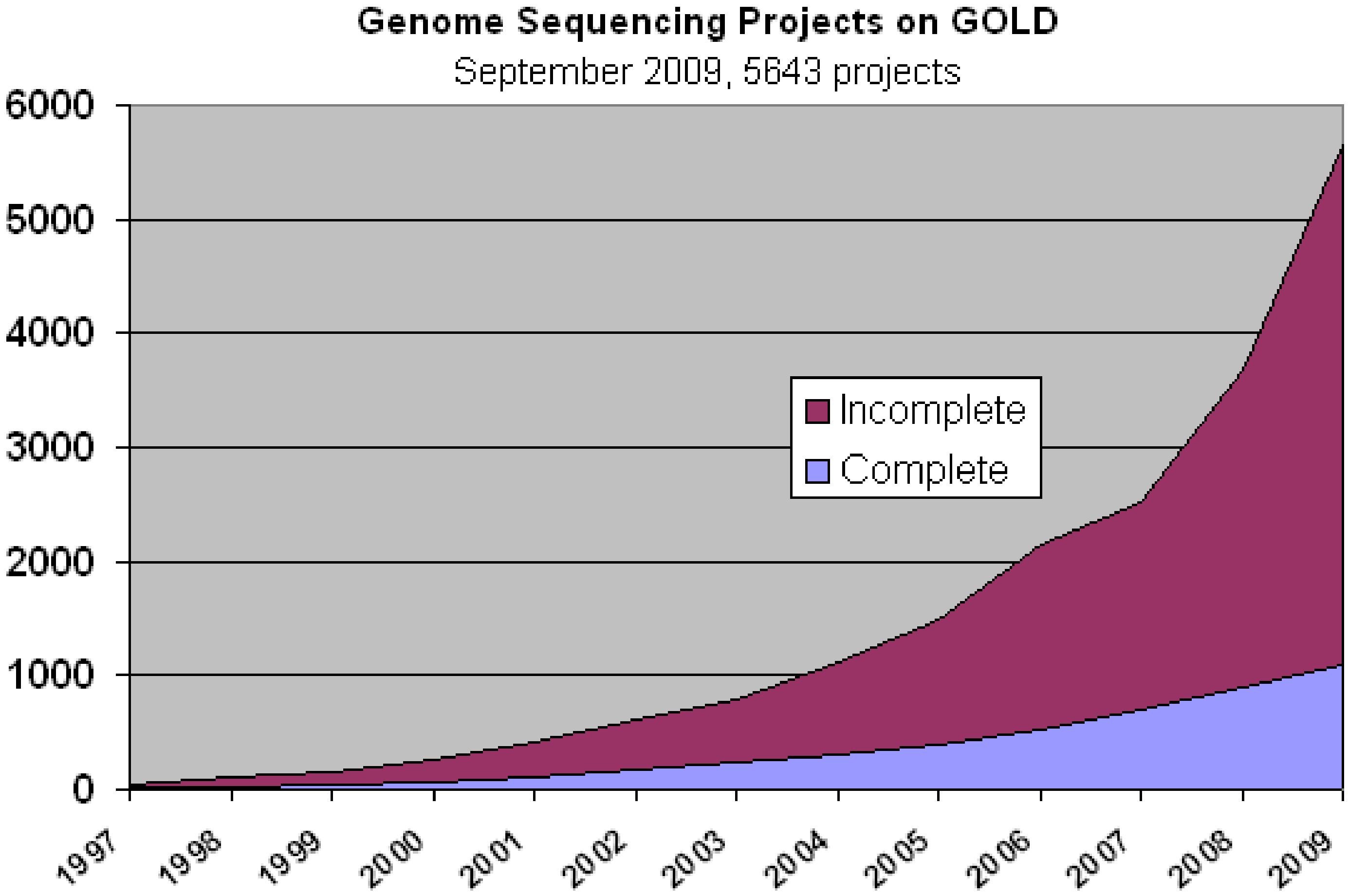
- Sanger sequencing basis of Genomics until 2005

Genomics

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

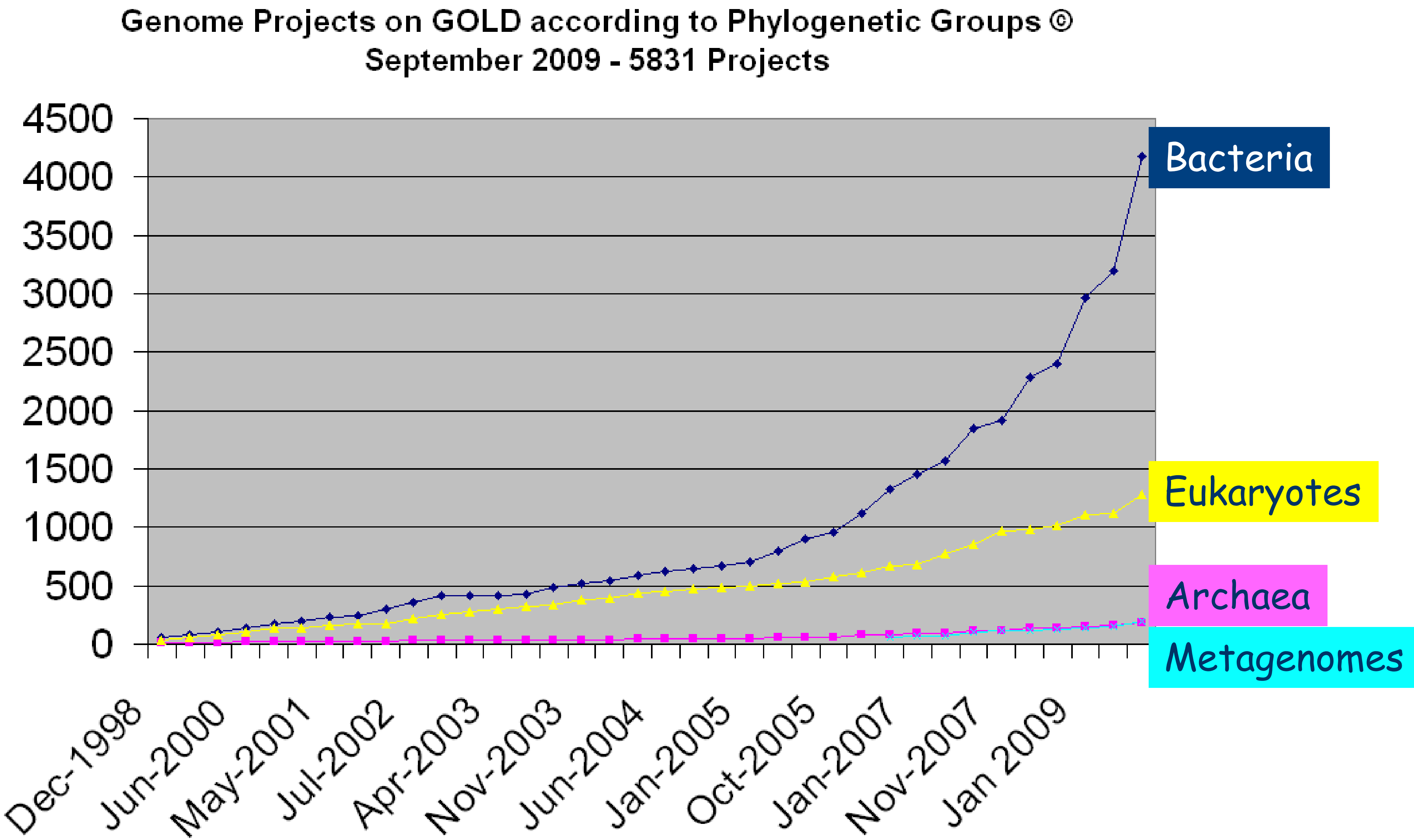
- Genome sizes:
  - Bacteria: 1-10 million bases (Mb)
  - Drosophila: 140Mb
  - Human: 3 billion bases (Gb)

Sequencing of Genomes



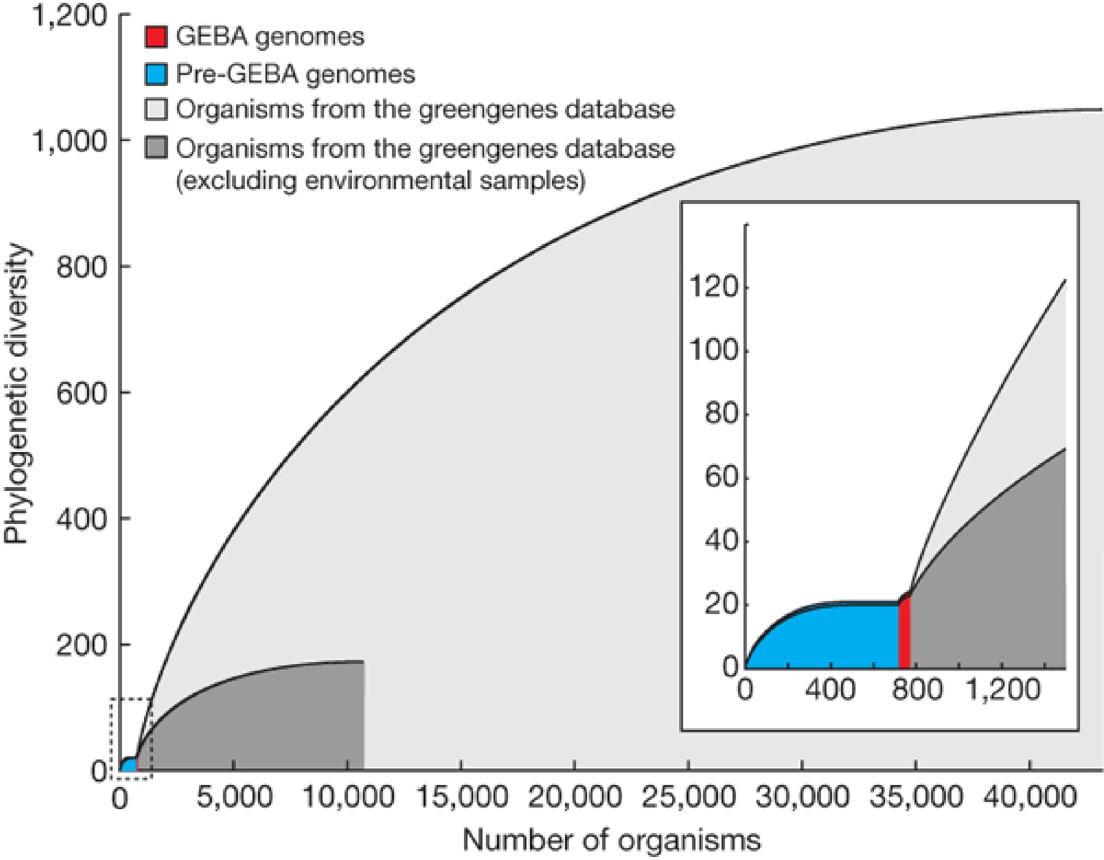
GOLD: Genomes online database
www.genomesonline.org

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/

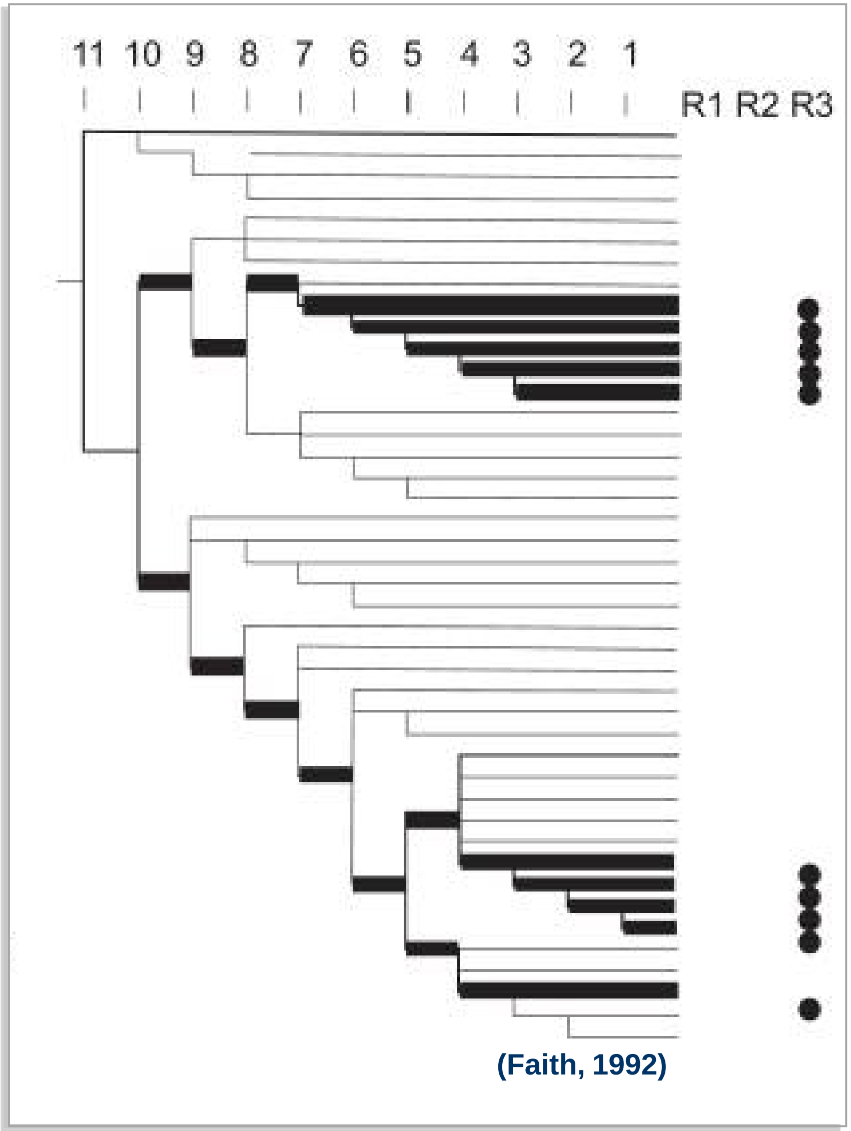
green genes
16S rRNA gene database and workbench compatible with ARB
greengenes.lbl.gov

Phylogenetic Diversity (PD)

PD of a set of taxa:

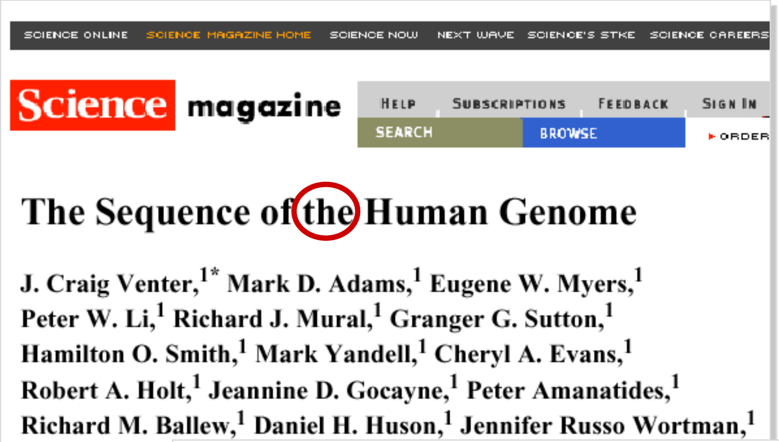
- The sum of all branches on the phylogenetic tree that spans the set

(Faith, 1992)



(Faith, 1992)

From One to Many…



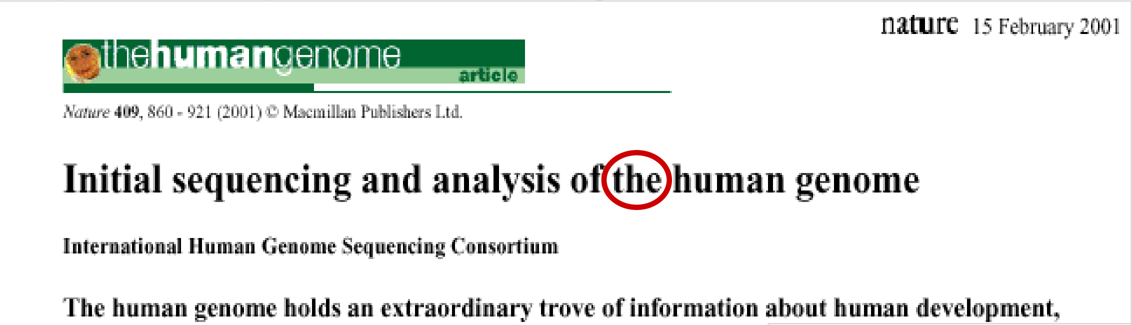
SCIENCE ONLINE SCIENCE MAGAZINE HOME SCIENCE NOW NEXT WAVE SCIENCE'S STKE SCIENCE CAREERS

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



nature 15 February 2001

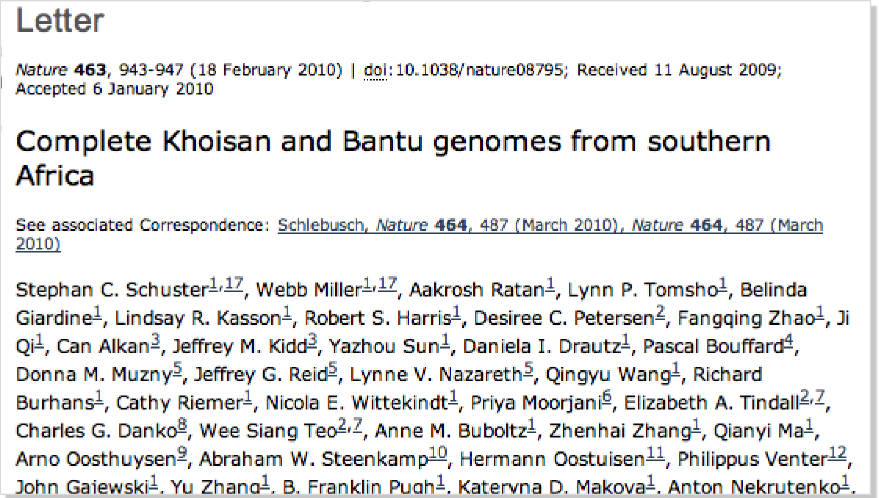
thehumangenome article

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

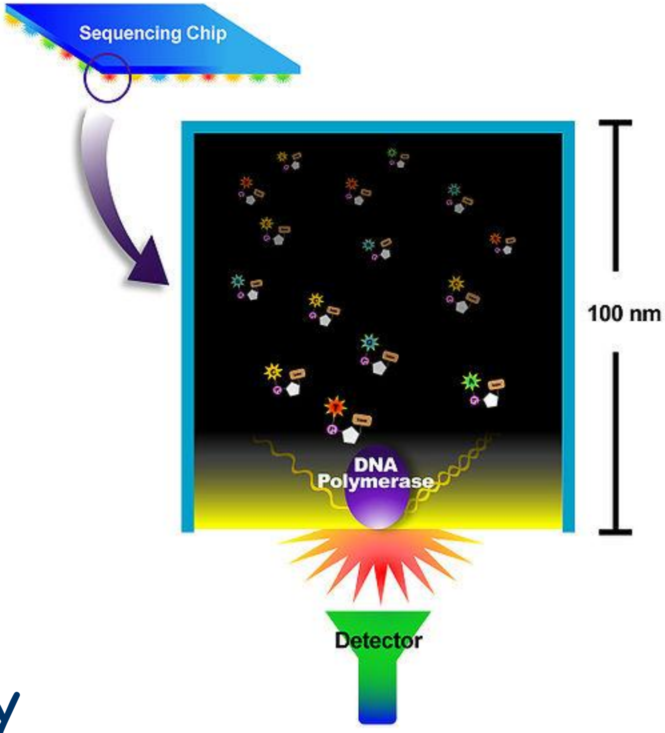
Advances in Sequencing Technologies

- 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$

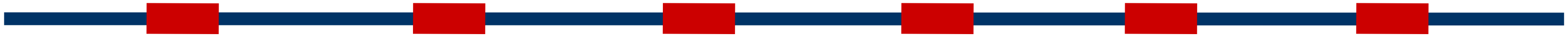
SMRT^{TM} Sequencing

Single Molecule Real Time Sequencing

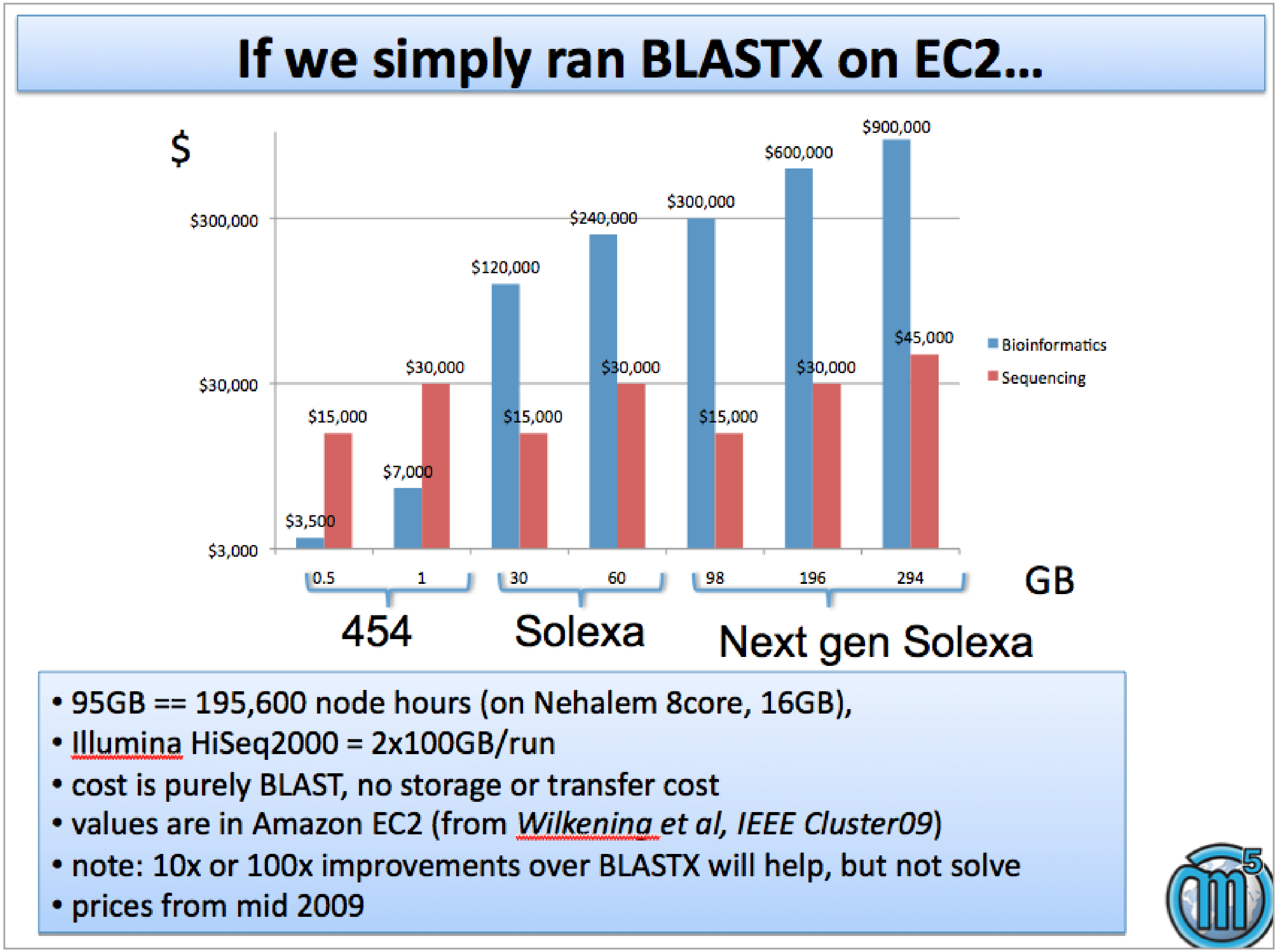
- Observes detached fluorescent dye molecules



- 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…



If we simply ran BLASTX on EC2…

- 95GB == 195,600 node hours (on Nehalem 8core, 16GB),
- Illumina HiSeq2000 = 2x100GB/run
- cost is purely BLAST, no storage or transfer cost
- values are in Amazon EC2 (from *Wilkening et al, IEEE Cluster09*)
- note: 10x or 100x improvements over BLASTX will help, but not solve
- prices from mid 2009

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.h

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

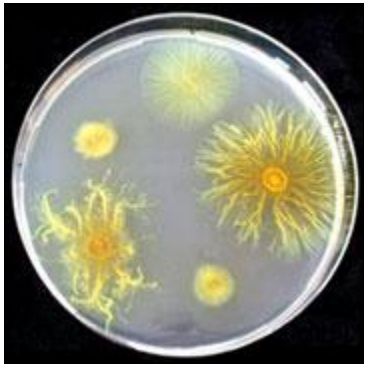
- ~1,250,000 animals, including:
  - 1,190,200 invertebrates:
    - 950,000 insects
    - 70,000 mollusks
    - 40,000 crustaceans
    - 130,200 others
  - 58,808 vertebrates:
    - 29,300 fish
    - 5,743 amphibians
    - 8,240 reptiles
    - 10,234 birds
    - 5,416 mammals

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

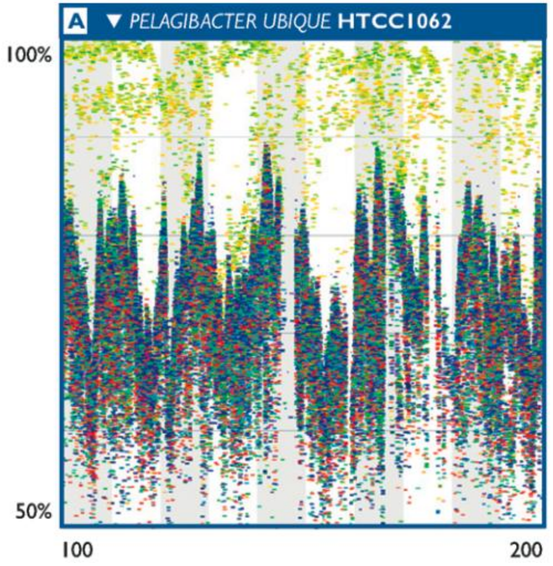
soils.usda.gov

http://outdoors.webshots.com

http://www.scienceimage.csiro.au/

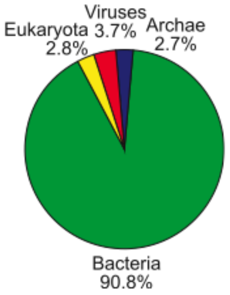
www.lanl.gov

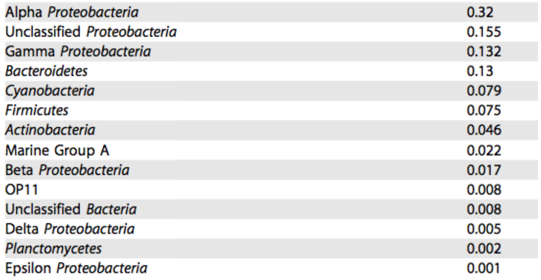
Global Ocean Sampling Expedition



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

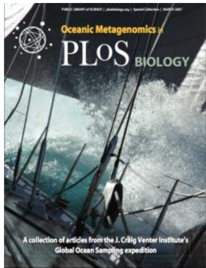




Alpha *Proteobacteria*	0.32
Unclassified *Proteobacteria*	0.155
Gamma *Proteobacteria*	0.132
Bacteroidetes	0.13
Cyanobacteria	0.079
Firmicutes	0.075
Actinobacteria	0.046
Marine Group A	0.022
Beta *Proteobacteria*	0.017
OP11	0.008
Unclassified *Bacteria*	0.008
Delta *Proteobacteria*	0.005
Planctomycetes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

Yooseph *et al.* (2007):

- 6 million proteins
  - linear rate of discovery



Oceanic Metagenomics

PLoS BIOLOGY

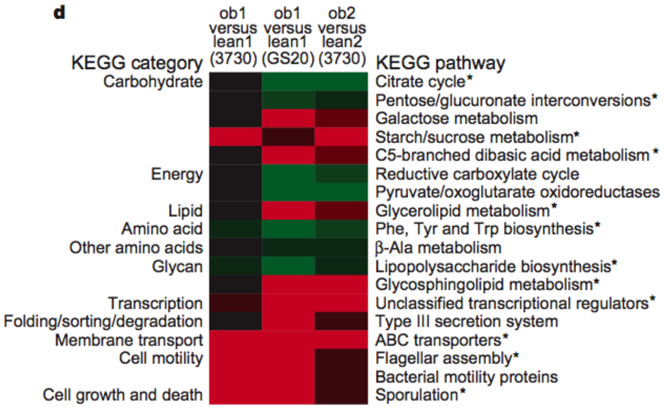
A collection of articles from the J. Craig Venter Institute's Global Ocean Sampling expedition

Gut Microbiota

http://en.wikipedia.org

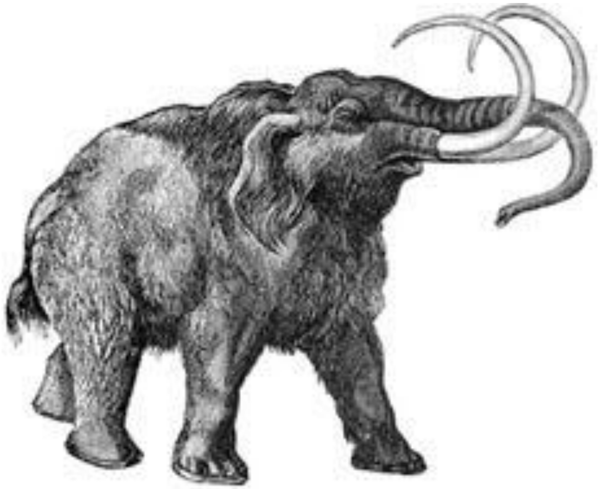
- Turnbaugh *et al* (2006)
- Caecal microbial DNA of *ob/ob*, *ob/+*, *+/+* mice
- Sanger sequencing:
  - 39.5 Mb
  - read length 750 bp
- 454 sequencing:
  - 160 Mb
  - read length 93 bp

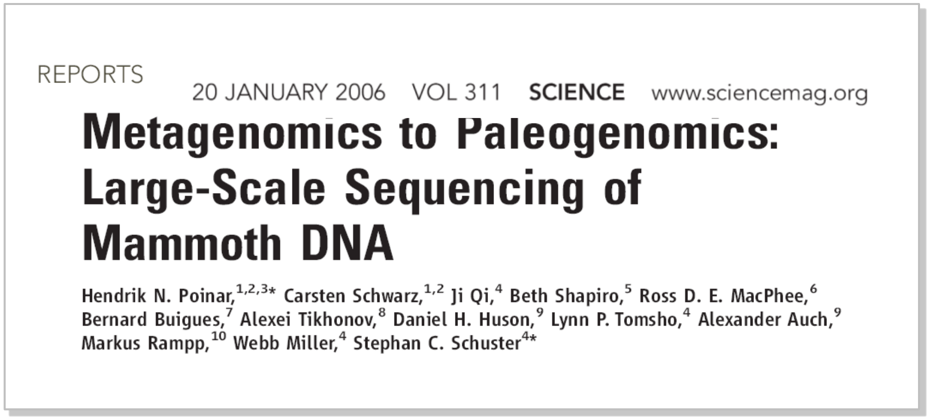
- Obesity-associated gut microbiome
  - Change in relative abundance of Bacteroidetes and Firmicutes
  - Change in functional capacity (toward energy harvesting)



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





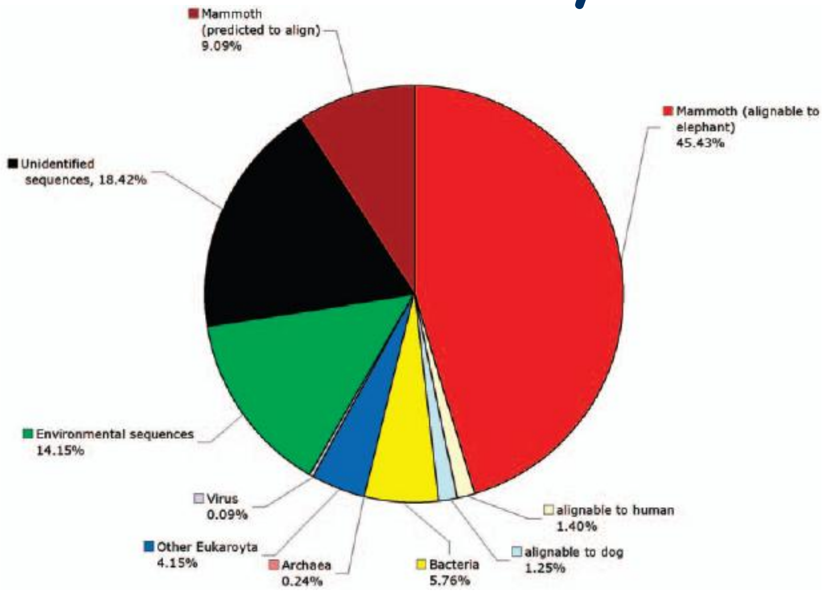
REPORTS

20 JANUARY 2006 VOL 311 SCIENCE www.sciencemag.org

Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA

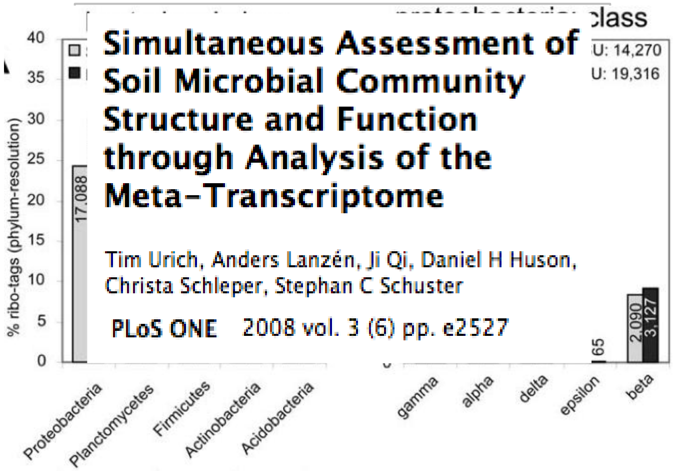
Hendrik N. Poinar,[1,2,3]* Carsten Schwarz,[1,2] Ji Qi,[4] Beth Shapiro,[5] Ross D. E. MacPhee,[6] Bernard Buigues,[7] Alexei Tikhonov,[8] Daniel H. Huson,[9] Lynn P. Tomsho,[4] Alexander Auch,[9] Markus Rampp,[10] Webb Miller,[4] Stephan C. Schuster[4]*

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

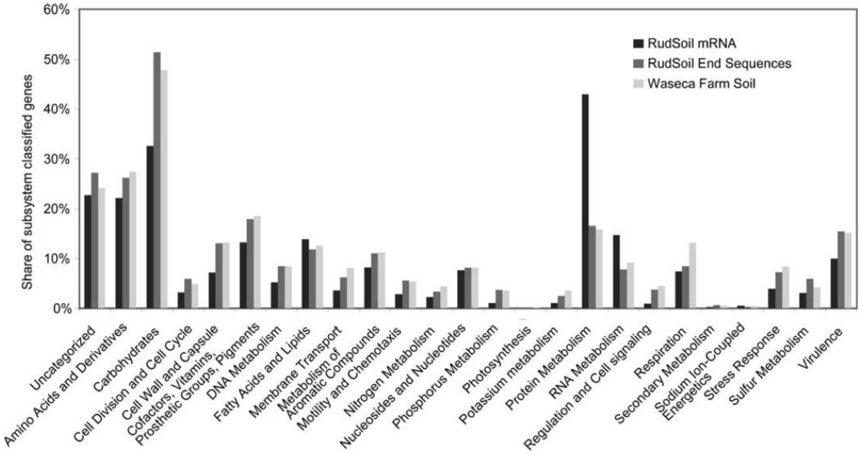


Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome

Tim Urich, Anders Lanzén, Ji Qi, Daniel H Huson, Christa Schleper, Stephan C Schuster

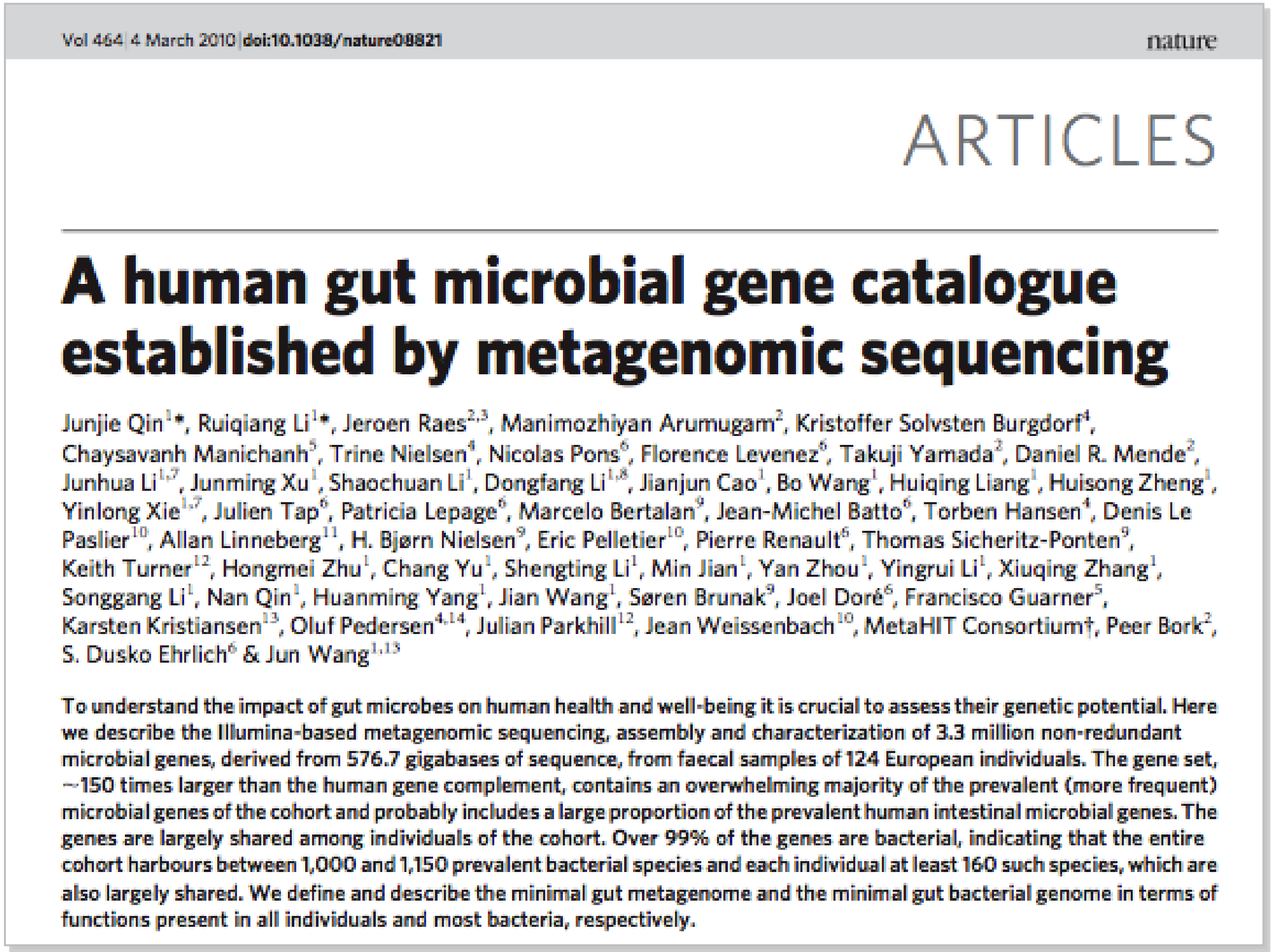
PLoS ONE 2008 vol. 3 (6) pp. e2527

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]

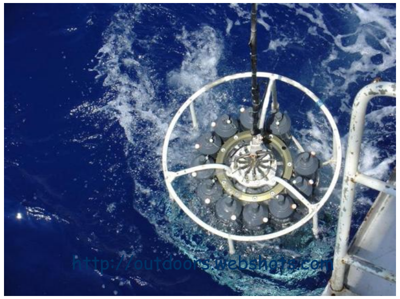
To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from 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

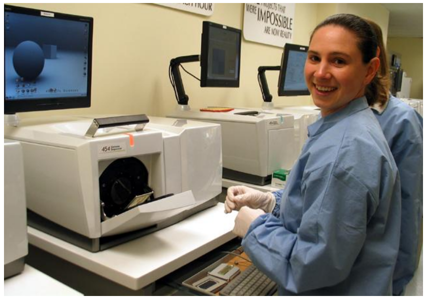
Contents

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

Metagenome Analysis



Environmental sample

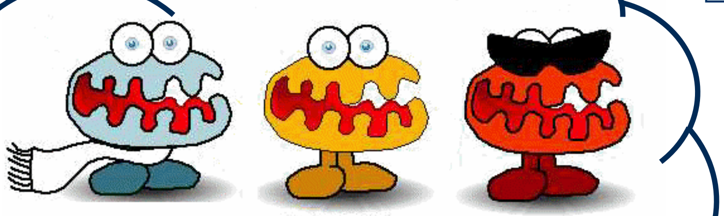


High-throughput DNA sequencing

10 million sequences

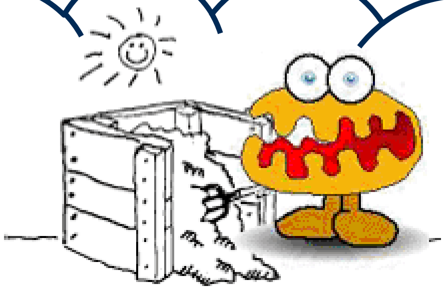
Basic computational analysis

10 000 hours



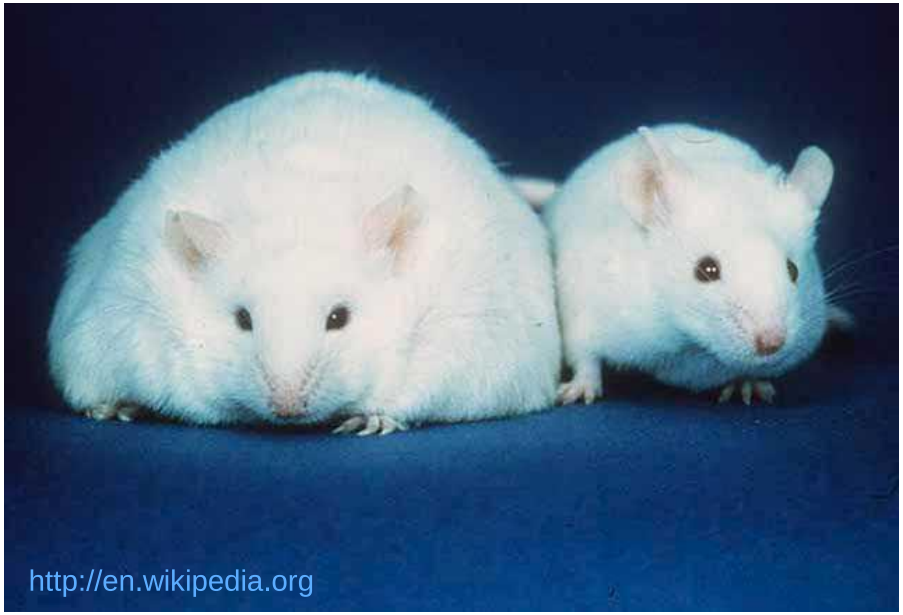
www.compostinfo.com/tutorial/microbes.htm

Q1: Who is out there?

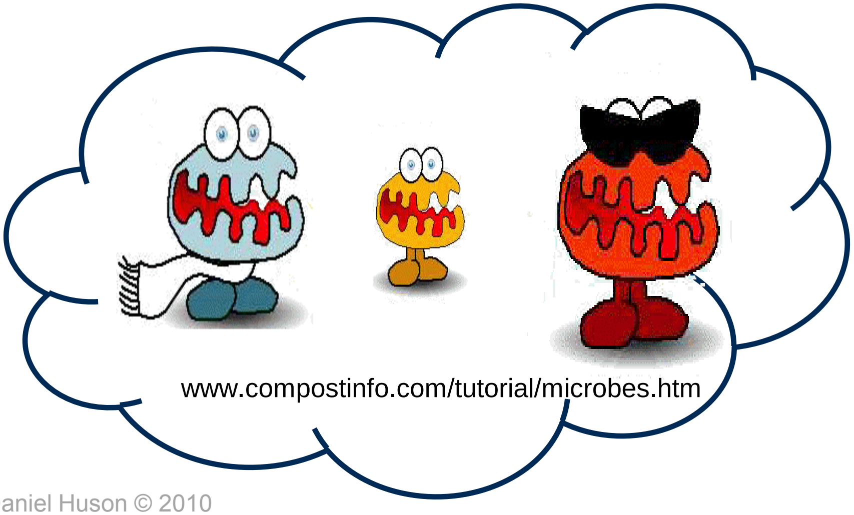


Q2: What are they doing?

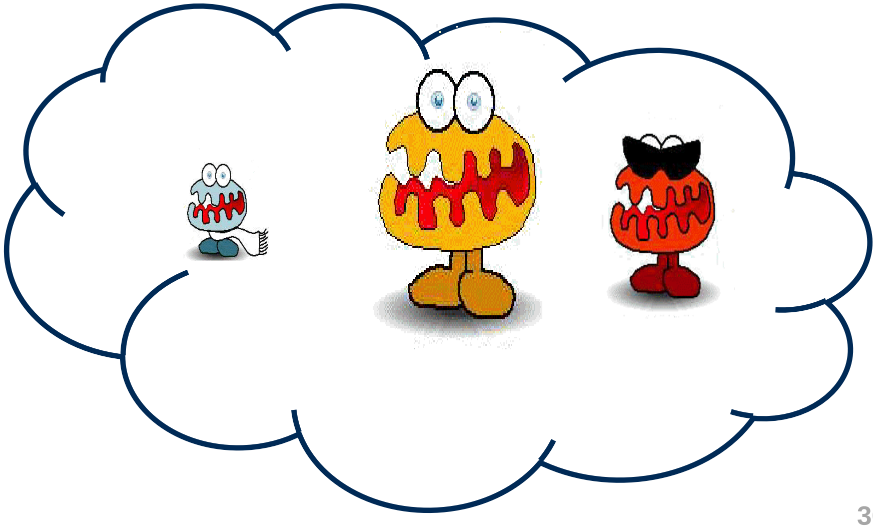
Q3: How Do They Compare?



http://en.wikipedia.org



www.compostinfo.com/tutorial/microbes.htm



Additional Questions

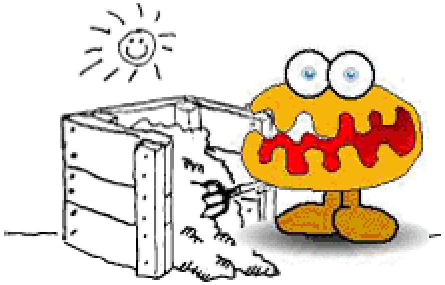
- How to cluster reads by relatedness (using machine learning techniques)?
- How to assemble metagenome data?
- Gene prediction?
- Faster sequence comparison
- etc

Three Basic Computational Questions

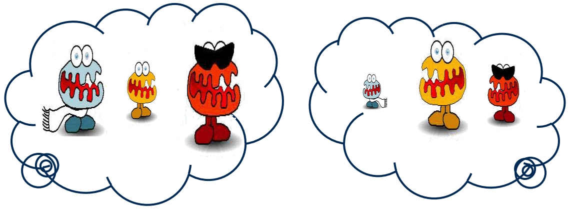
- Who is out there?
  - Types of organisms
  - In what proportions?

www.compostinfo.com/tutorial/microbes.htm

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



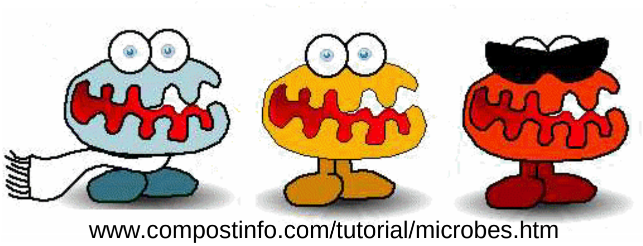
- How do different samples compare?
  - Pairwise and multiple comparisons
  - Correlations with environmental parameters?



- Serve to answer biological or medical questions

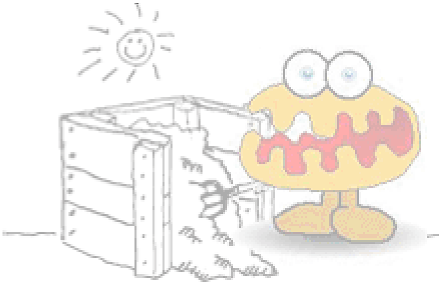
Three Basic Computational Questions

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

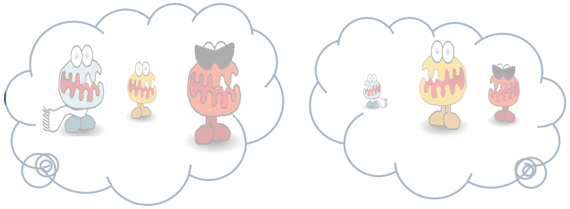


www.compostinfo.com/tutorial/microbes.htm

- 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?

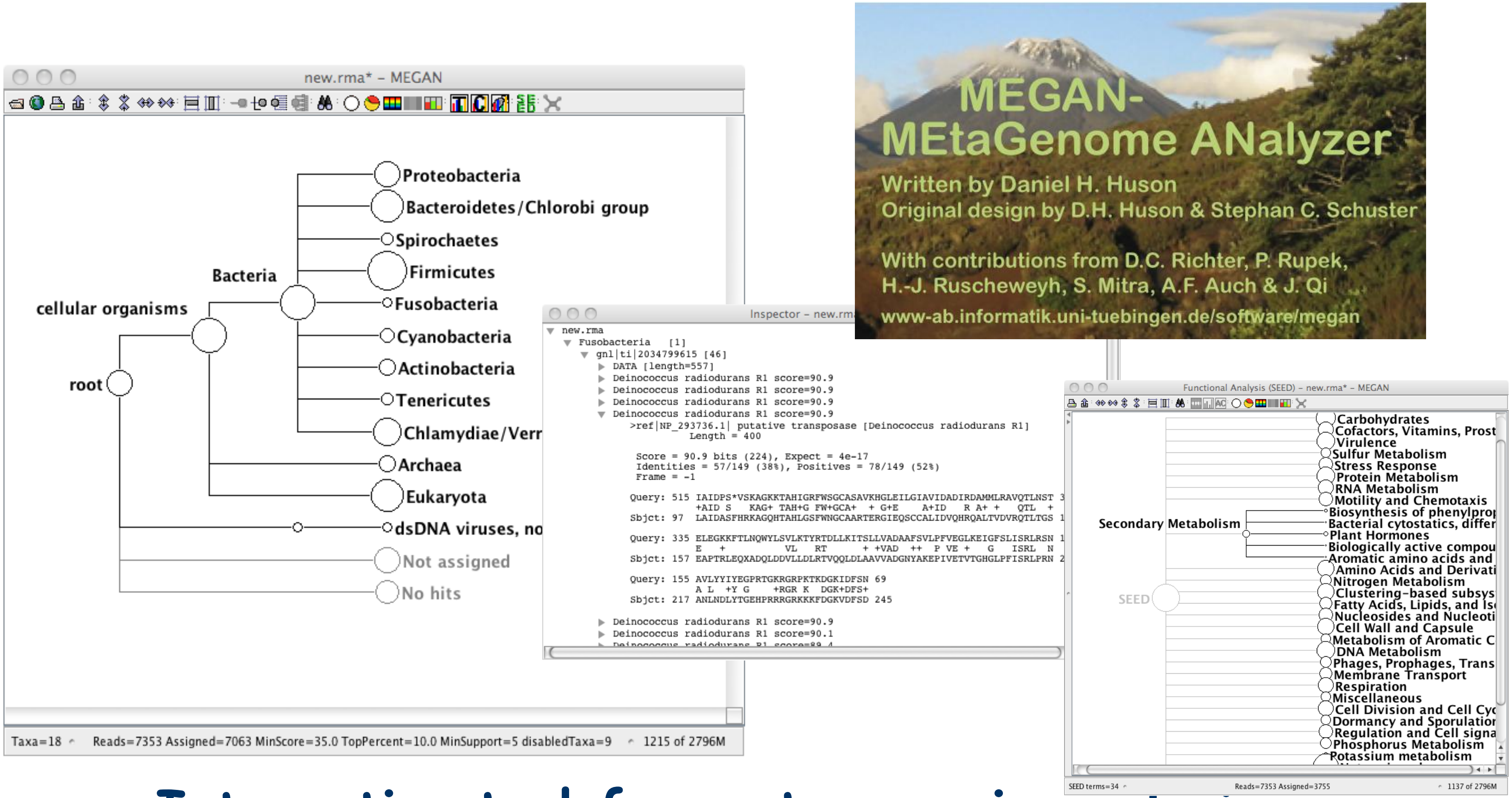
Main tool: Similarity search

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



Taxa=18 Reads=7353 Assigned=7063 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9 1215 of 2796M

```
▼ 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*VSKAGKKTAHIGRFWSGCASAVKHGLEILGIAVIDADIRDAMMLRAVQTLNST 3
                      +AID S   KAG+ TAH+G FW+GCA+   + G+E    A+ID    R A+ +    QTL  +
          Sbjct: 97  LAIDASFHRKAGQHTAHLGSFWNGCAARTERGIEQSCCALIDVQHRQALTVDVRQTLTGS 1

          Query: 335 ELEGKKFTLNQWYLSVLKTYRTDLLKITSLLVADAAFSVLPFVEGLKEIGFSLISRLRSN 1
                      E    +              VL    RT         + +VAD   ++   P VE +    G    ISRL   N
          Sbjct: 157 EAPTRLEQXADQLDDVLLDLRTVQQLDLAAVVADGNYAKEPIVETVTGHGLPFISRLPRN 2

          Query: 155 AVLYYIYEGPRTGKRGRPKTKDGKIDFSN 69
                      A L   +Y G     +RGR K   DGK+DFS+
          Sbjct: 217 ANLNDLYTGEHPRRRGRKKKFDGKVDFSD 245

      ▶ Deinococcus radiodurans R1 score=90.9
      ▶ Deinococcus radiodurans R1 score=90.1
```

MEGAN-
MEtaGenome ANalyzer

Written by Daniel H. Huson
Original design by D.H. Huson & Stephan C. Schuster

With contributions from D.C. Richter, P. Rupek,
H.-J. Ruscheweyh, S. Mitra, A.F. Auch & J. Qi
www-ab.informatik.uni-tuebingen.de/software/megan

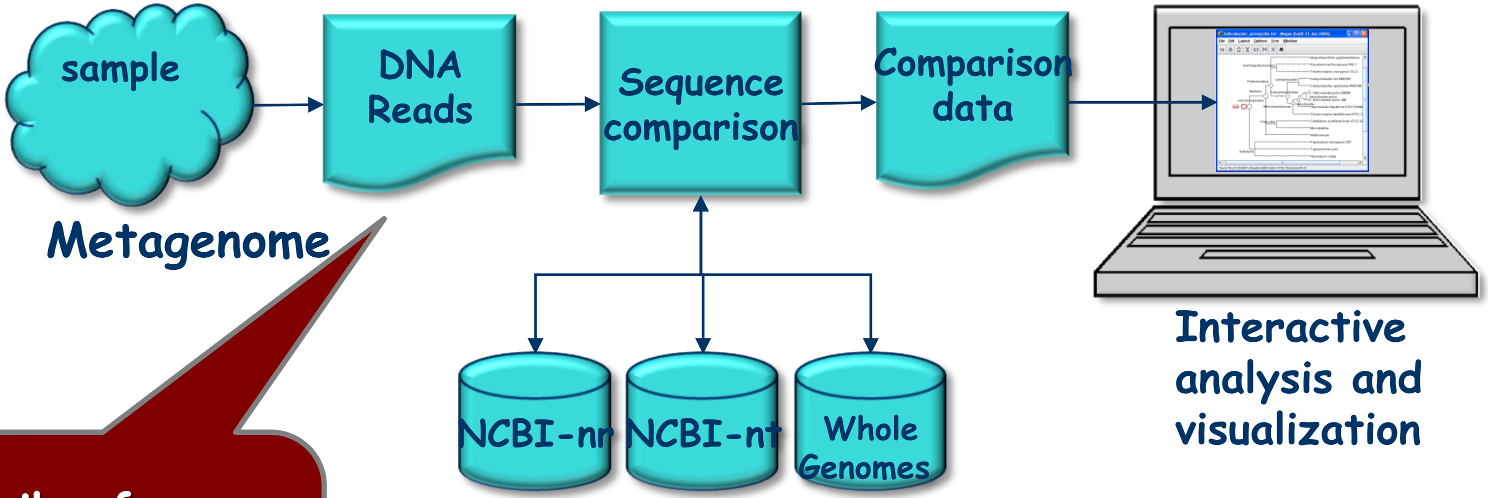
SEED terms=34 Reads=7353 Assigned=3755 1137 of 2796M

- **Interactive tool for metagenomic analysis**

  (Version 4, to be released Nov 2010)

www-ab.informatik.uni-tuebingen.de/software/megan

Metagenomics Pipeline



Metagenome

Interactive analysis and visualization

Reference databases

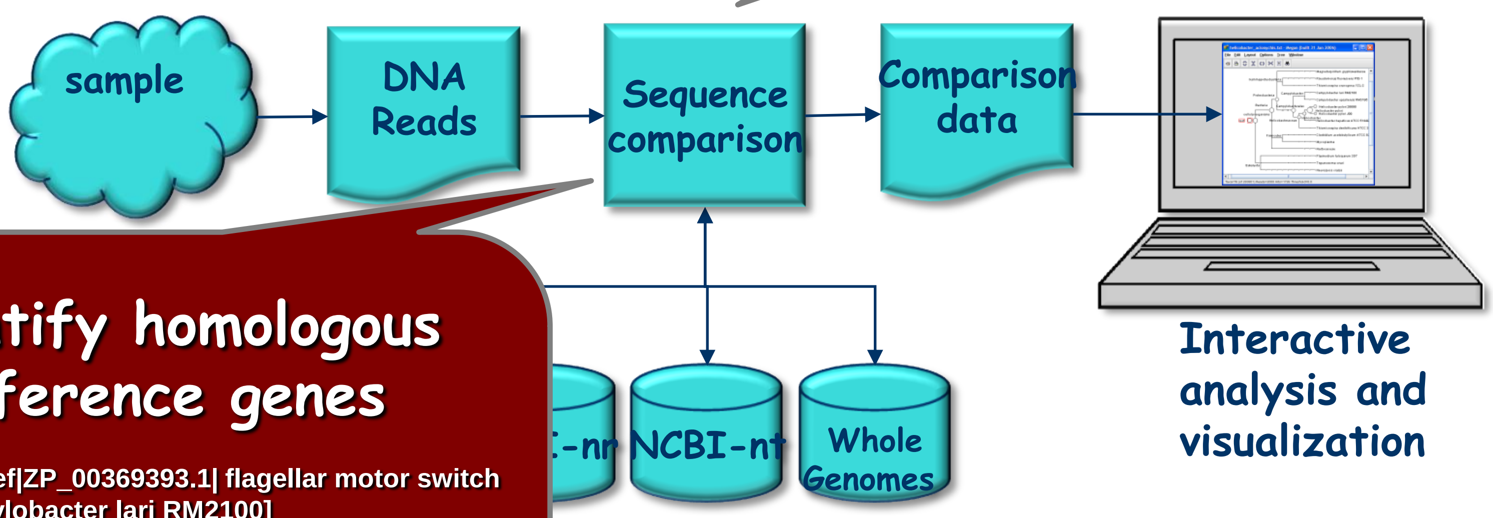
Similar for
- metatranscriptomics
- metaproteomics
- amplicon sequencing

Identifying Taxa and Genes

Metagenome analysis

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

BLASTX against NCBI-NR



Identify homologous reference genes

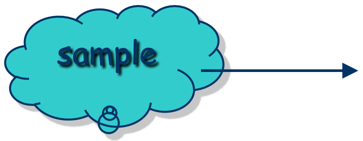
```
>gi|57241447|ref|ZP_00369393.1| flagellar motor switch
protein [Campylobacter lari RM2100]
 Score = 33.9 bits (76), Expect = 1.8
 Identities = 13/26 (50%), Positives = 19/26 (73%)

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

Interactive analysis and visualization

Reference databases

Sequence Comparison

- DNA Read



ACTGTGCACGTTGACGTAAGTTT...CGTGT

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

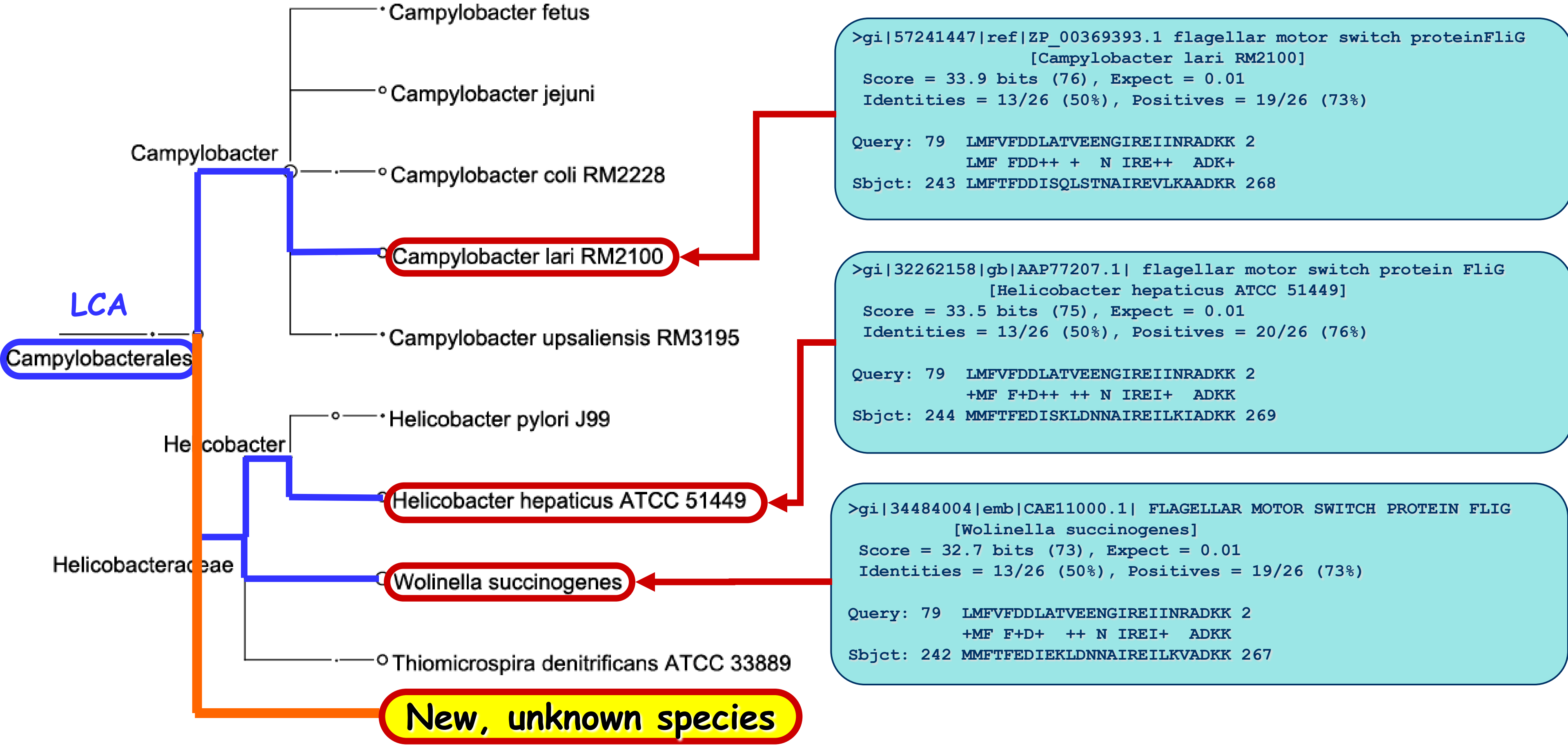
```
>gi|57241447|ref|ZP 00369393.1| flagellar motor switch protein
                  [Campylobacter lari RM2100]
 Score = 33.9 bits (76), Expect =0.01
 Identities = 13/26 (50%), Positives = 19/26 (73%)

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:



```
>gi|57241447|ref|ZP_00369393.1 flagellar motor switch proteinFliG
                    [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
```

```
>gi|32262158|gb|AAP77207.1| flagellar motor switch protein FliG
               [Helicobacter hepaticus ATCC 51449]
 Score = 33.5 bits (75), Expect = 0.01
 Identities = 13/26 (50%), Positives = 20/26 (76%)

Query:  79  LMFVFDDLATVEENGIREIINRADKK 2
            +MF F+D++ ++ N IREI  ADK
Sbjct: 244 MMFTFEDISKLDNNAIREILKIADKK 269
```

```
>gi|34484004|emb|CAE11000.1| FLAGELLAR MOTOR SWITCH PROTEIN FLIG
            [Wolinella succinogenes]
 Score = 32.7 bits (73), Expect = 0.01
 Identities = 13/26 (50%), Positives = 19/26 (73%)

Query:  79  LMFVFDDLATVEENGIREIINRADKK 2
            +MF F+D+   ++ N IREI+  ADK
Sbjct: 242 MMFTFEDIEKLDNNAIREILKVADKK 267
```

LCA approach: Assign read to LCA of hits in taxonomy

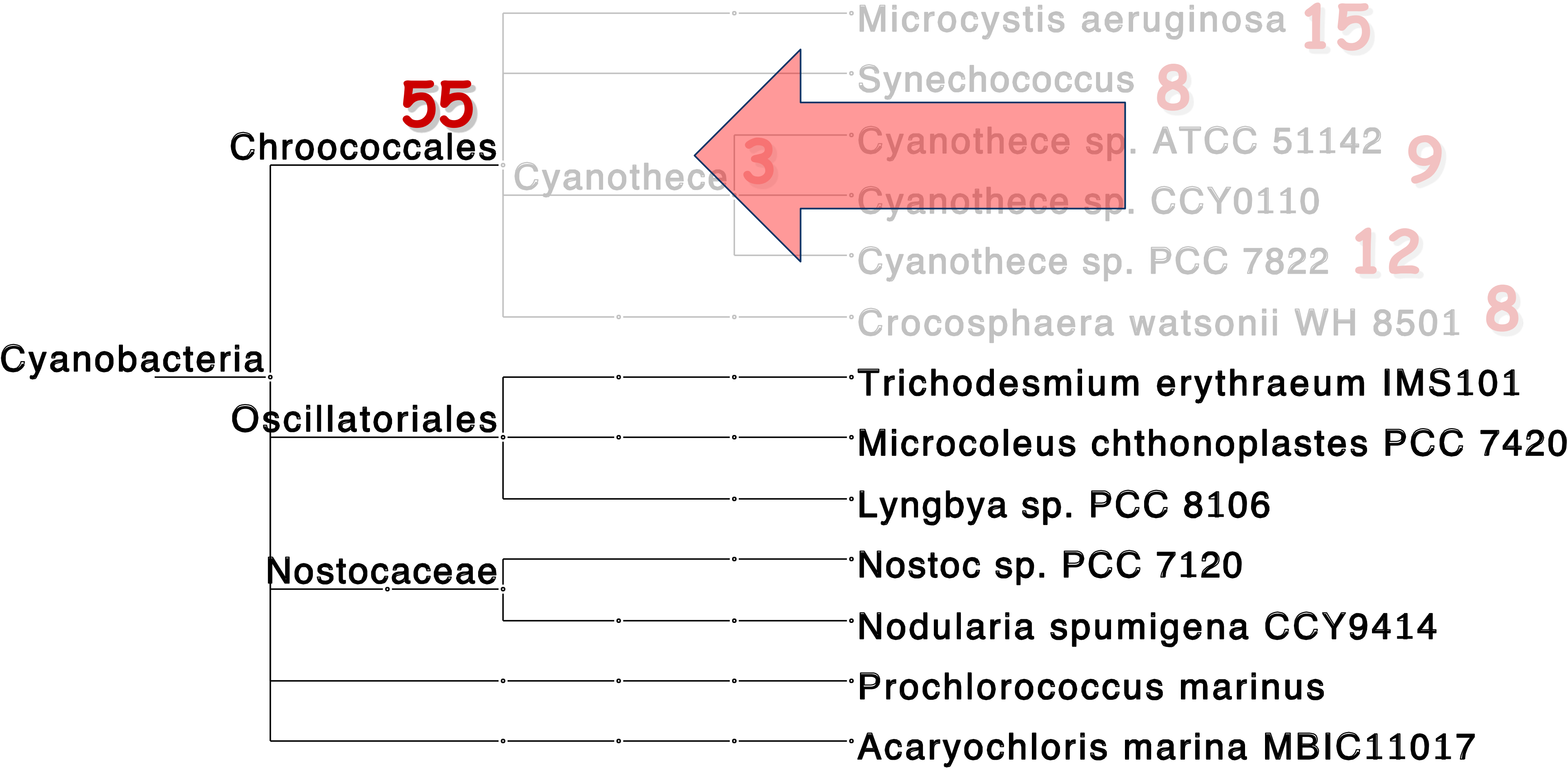
Taxonomic Placement Using LCA

- 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


55
Chroococcales
Cyanothece
3
Microcystis aeruginosa 15
Synechococcus 8
Cyanothece sp. ATCC 51142
9
Cyanothece sp. CCY0110
Cyanothece sp. PCC 7822 12
Crocosphaera watsonii WH 8501 8
Cyanobacteria
Oscillatoriales
Trichodesmium erythraeum IMS101
Microcoleus chthonoplastes PCC 7420
Lyngbya sp. PCC 8106
Nostocaceae
Nostoc sp. PCC 7120
Nodularia spumigena CCY9414
Prochlorococcus marinus
Acaryochloris marina MBIC11017

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

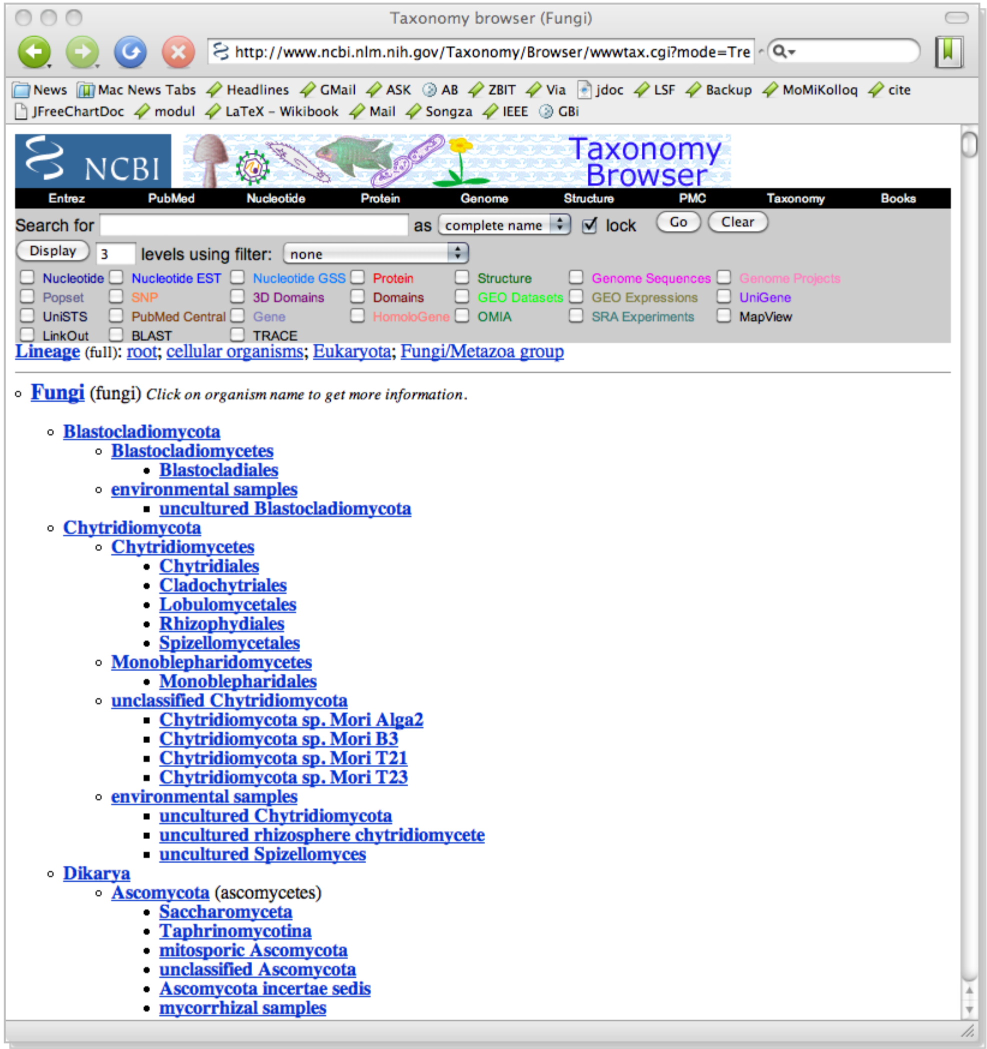
Taxonomic Analysis using NCBI Taxonomy

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)



Search for ______ as complete name ☑ lock Go Clear

Display 3 levels using filter: none

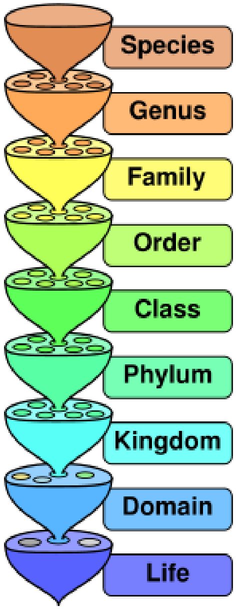
□ Nucleotide □ Nucleotide EST □ Nucleotide GSS □ Protein □ Structure □ Genome Sequences □ Genome Projects
□ Popset □ SNP □ 3D Domains □ Domains □ GEO Datasets □ GEO Expressions □ UniGene
□ UniSTS □ PubMed Central □ Gene □ HomoloGene □ OMIA □ SRA Experiments □ MapView
□ LinkOut □ BLAST □ TRACE

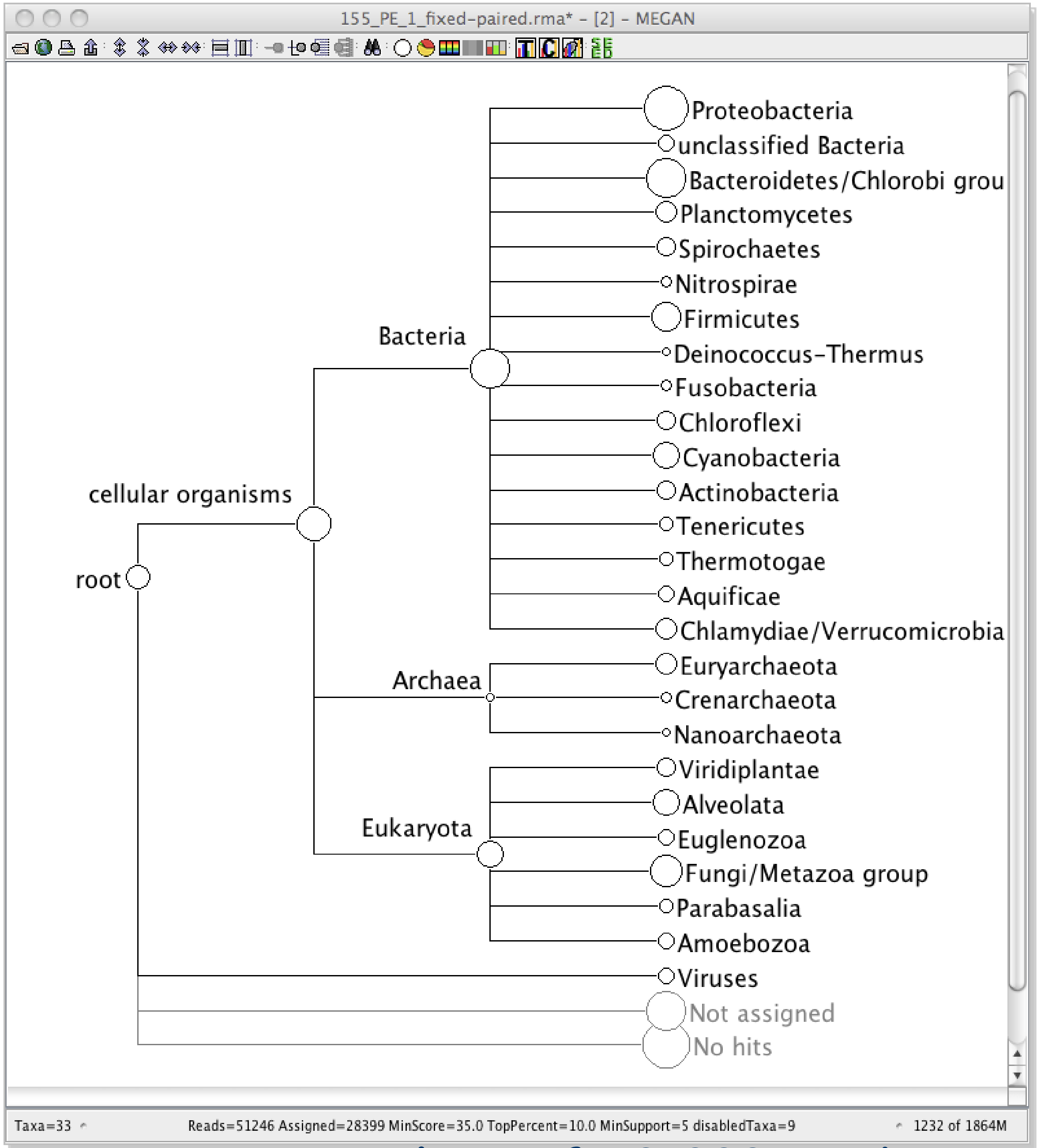
Lineage (full): root; cellular organisms; Eukaryota; Fungi/Metazoa group

- Fungi (fungi) *Click on organism name to get more information.*
  - Blastocladiomycota
    - Blastocladiomycetes
      - Blastocladiales
    - environmental samples
      - uncultured Blastocladiomycota
  - Chytridiomycota
    - Chytridiomycetes
      - Chytridiales
      - Cladochytriales
      - Lobulomycetales
      - Rhizophydiales
      - Spizellomycetales
    - Monoblepharidomycetes
      - Monoblepharidales
    - unclassified Chytridiomycota
      - Chytridiomycota sp. Mori Alga2
      - Chytridiomycota sp. Mori B3
      - Chytridiomycota sp. Mori T21
      - Chytridiomycota sp. Mori T23
    - environmental samples
      - uncultured Chytridiomycota
      - uncultured rhizosphere chytridiomycete
      - uncultured Spizellomyces
  - Dikarya
    - Ascomycota (ascomycetes)
      - Saccharomyceta
      - Taphrinomycotina
      - mitosporic Ascomycota
      - unclassified Ascomycota
      - Ascomycota incertae sedis
      - mycorrhizal samples

Organize and Visualize

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



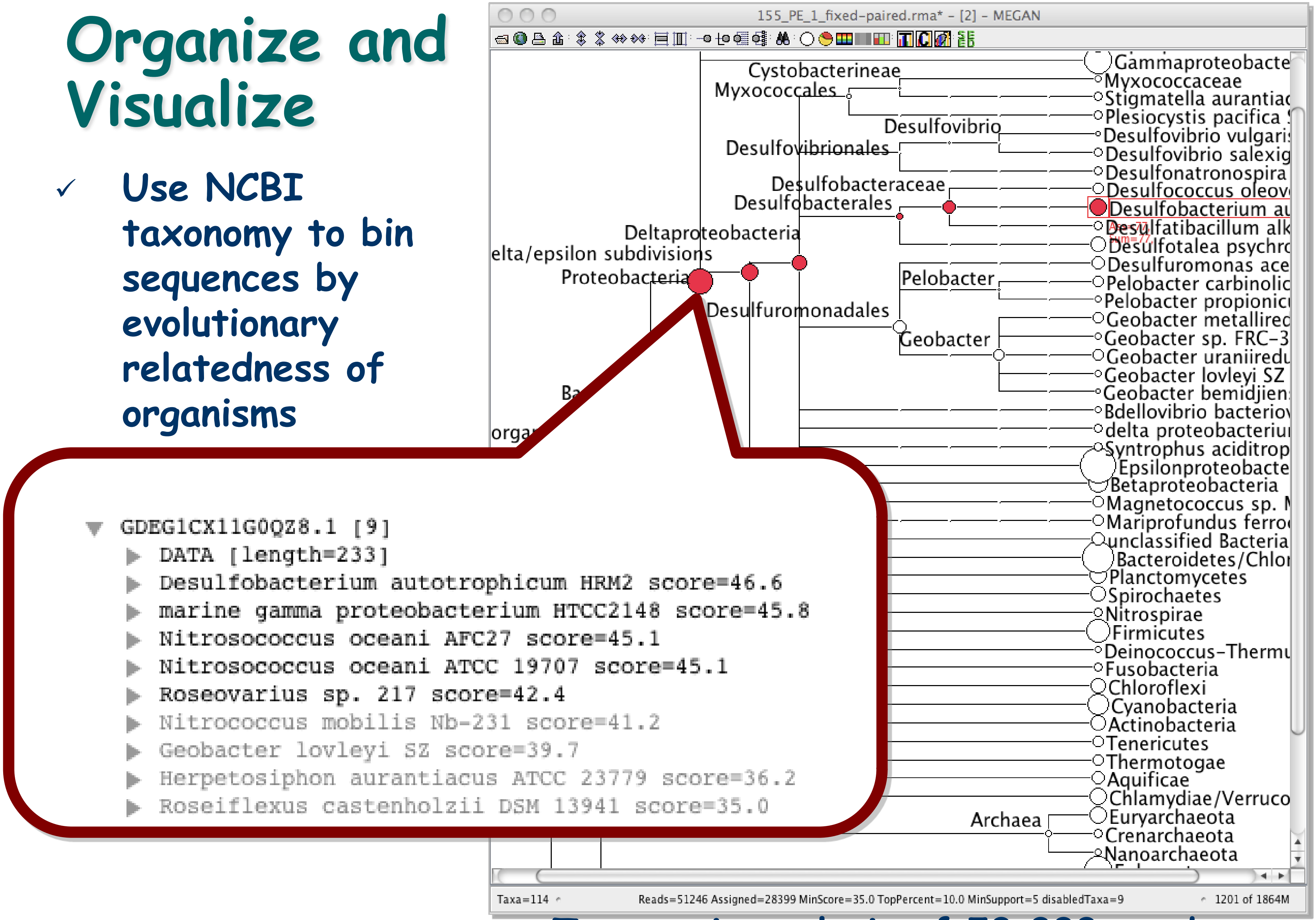

155_PE_1_fixed-paired.rma* – [2] – MEGAN

```
Taxa=33     Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9     1232 of 1864M
```

Taxonomic analysis of 50,000 reads

Organize and Visualize

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



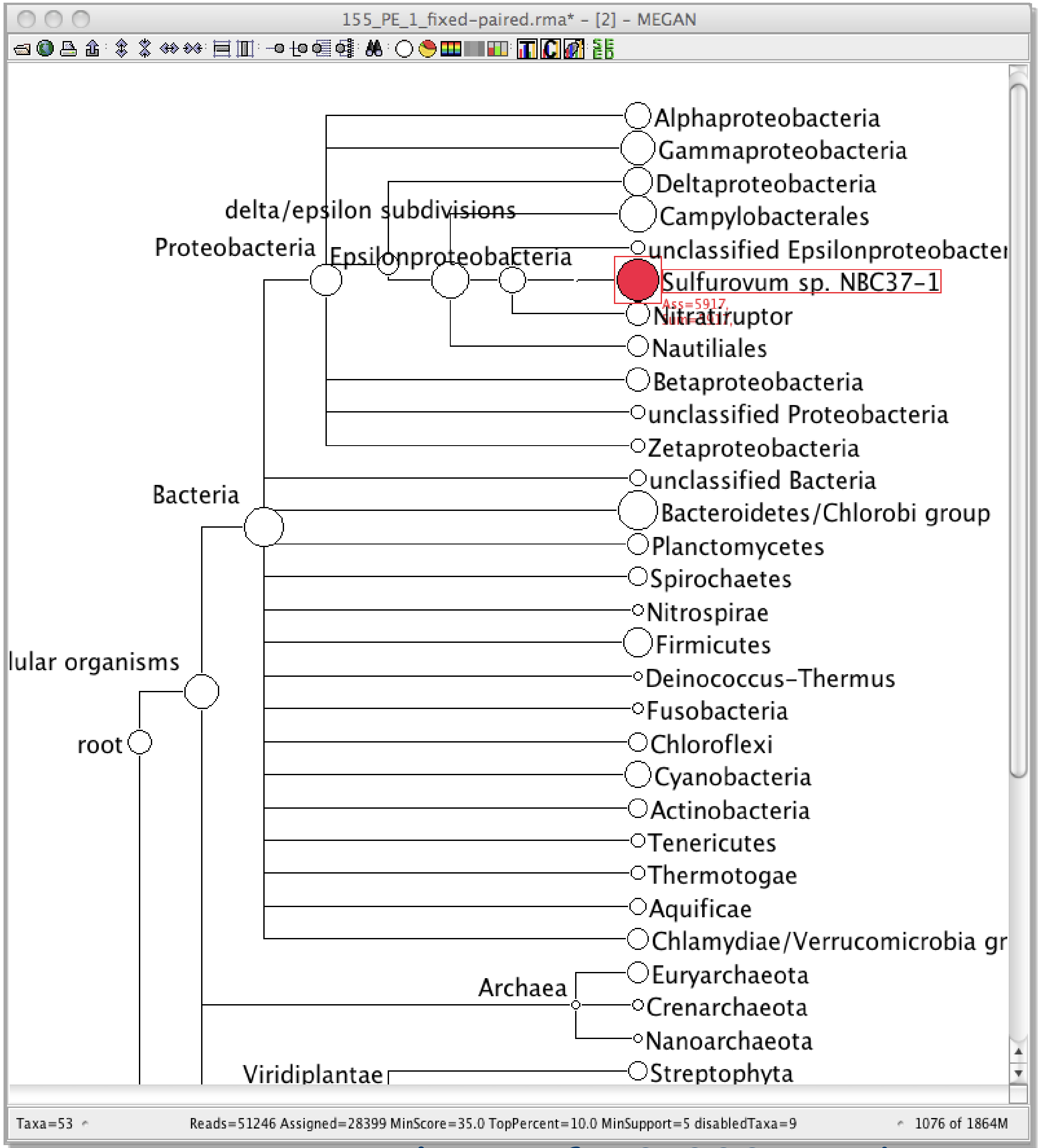
```
▼ GDEG1CX11G0QZ8.1 [9]
   ▶ DATA [length=233]
   ▶ Desulfobacterium autotrophicum HRM2 score=46.6
   ▶ marine gamma proteobacterium HTCC2148 score=45.8
   ▶ Nitrosococcus oceani AFC27 score=45.1
   ▶ Nitrosococcus oceani ATCC 19707 score=45.1
   ▶ Roseovarius sp. 217 score=42.4
   ▶ Nitrococcus mobilis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

```
Taxa=114     Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9     1201 of 1864M
```

Taxonomic analysis of 50,000 reads

Interact and Summarize

- ✓ Search for nodes of interest



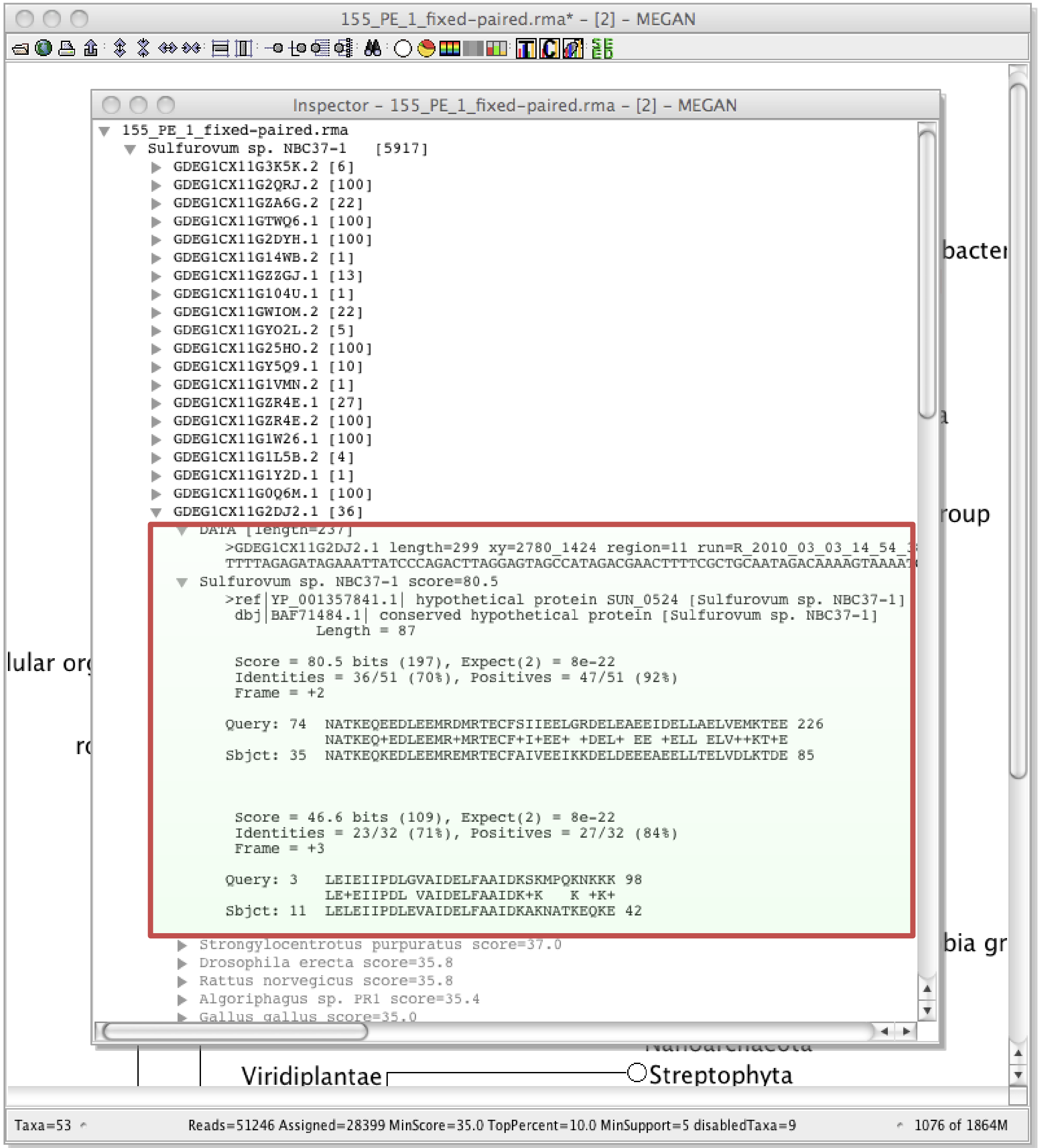
155_PE_1_fixed-paired.rma* – [2] – MEGAN

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


155_PE_1_fixed-paired.rma* – [2] – MEGAN
Inspector – 155_PE_1_fixed-paired.rma – [2] – MEGAN
bacte
a
roup
lular or
r
bia gr
Nanoarchaeota
Viridiplantae
Streptophyta

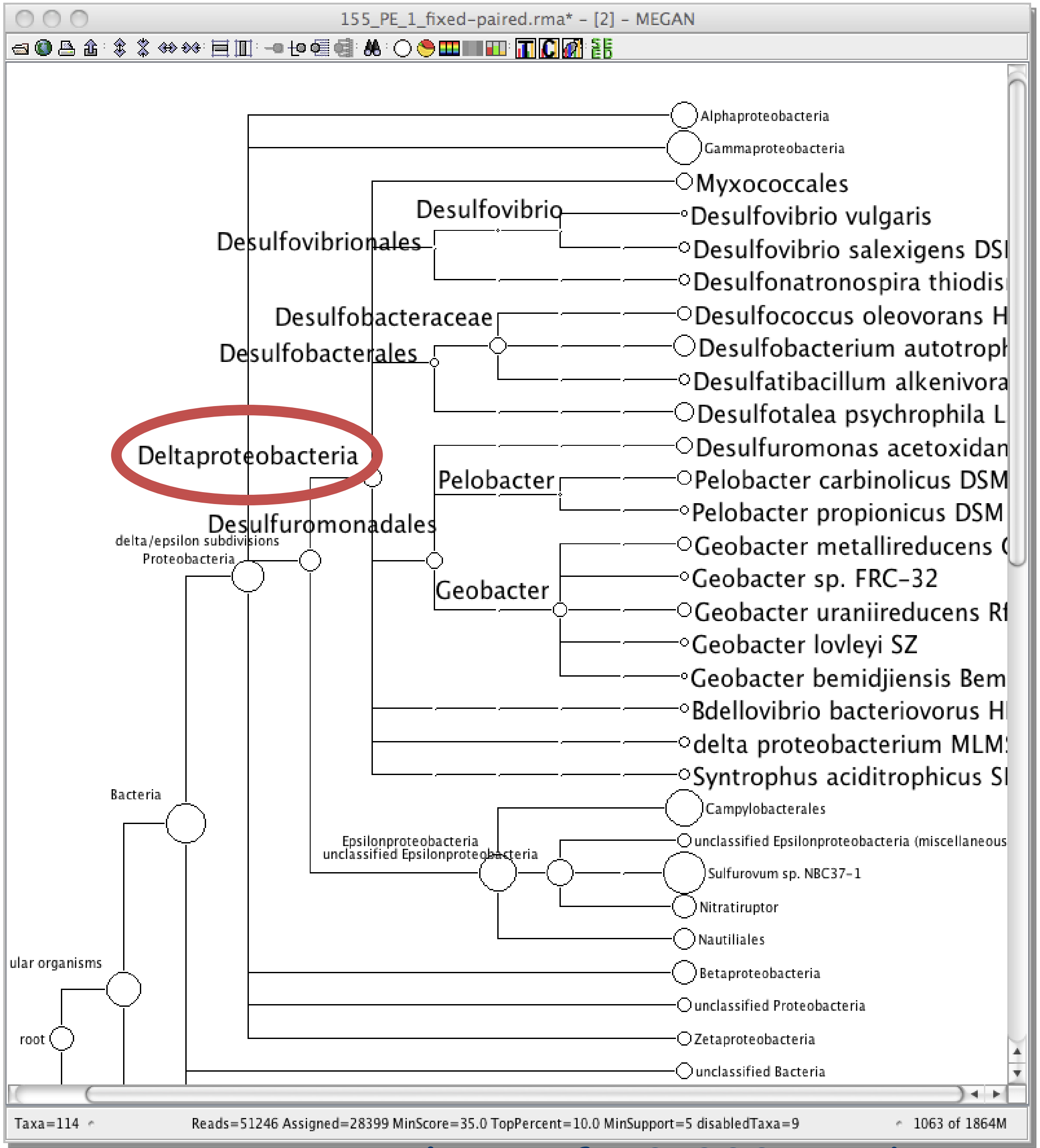
```
▼ 155_PE_1_fixed-paired.rma
  ▼ Sulfurovum sp. NBC37-1   [5917]
    ▶ GDEG1CX11G3K5K.2 [6]
    ▶ GDEG1CX11G2QRJ.2 [100]
    ▶ GDEG1CX11GZA6G.2 [22]
    ▶ GDEG1CX11GTWQ6.1 [100]
    ▶ GDEG1CX11G2DYH.1 [100]
    ▶ GDEG1CX11G14WB.2 [1]
    ▶ GDEG1CX11GZZGJ.1 [13]
    ▶ GDEG1CX11G104U.1 [1]
    ▶ GDEG1CX11GWIOM.2 [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_3
          TTTTAGAGATAGAAATTATCCCAGACTTAGGAGTAGCCATAGACGAACTTTTCGCTGCAATAGACAAAAGTAAAAT
      ▼ 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  NATKEQKEDLEEMREMRTECFAIVEEIKKDELDEEEAEELLTELVDLKTDE 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  LELEIIPDLEVAIDELFAAIDKAKNATKEQKE 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
```

```
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

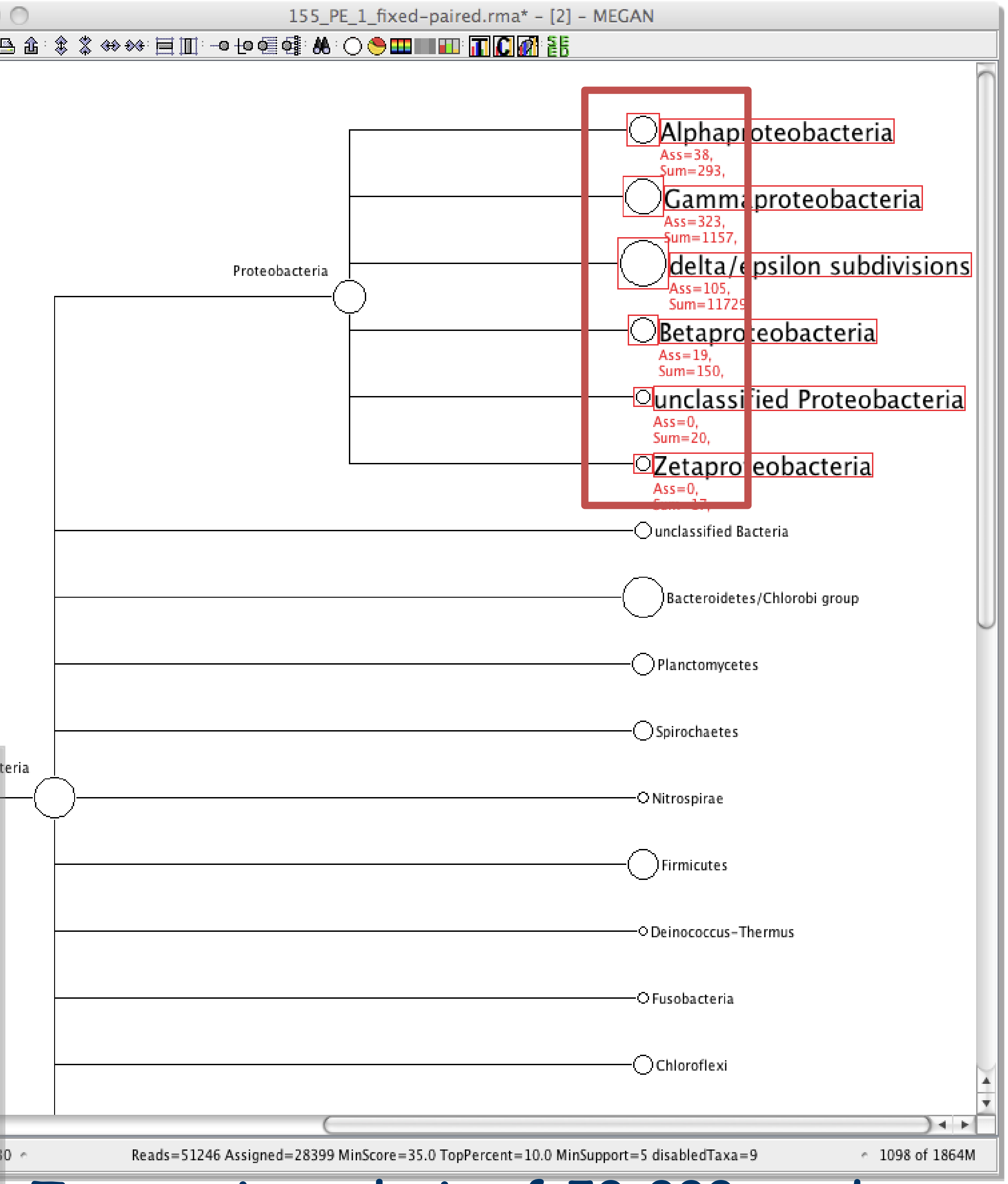

155_PE_1_fixed-paired.rma* - [2] - MEGAN
Alphaproteobacteria
Gammaproteobacteria
Myxococcales
Desulfovibrio
Desulfovibrio vulgaris
Desulfovibrionales
Desulfovibrio salexigens DSI
Desulfonatronospira thiodis
Desulfobacteraceae
Desulfococcus oleovorans H
Desulfobacterales
Desulfobacterium autotropl
Desulfatibacillum alkenivora
Desulfotalea psychrophila L
Deltaproteobacteria
Desulfuromonas acetoxidan
Pelobacter
Pelobacter carbinolicus DSM
Pelobacter propionicus DSM
Desulfuromonadales
delta/epsilon subdivisions
Proteobacteria
Geobacter metallireducens (
Geobacter sp. FRC-32
Geobacter
Geobacter uraniireducens Rf
Geobacter lovleyi SZ
Geobacter bemidjiensis Bem
Bdellovibrio bacteriovorus H
delta proteobacterium MLM
Syntrophus aciditrophicus SI
Bacteria
Campylobacterales
Epsilonproteobacteria
unclassified Epsilonproteobacteria
unclassified Epsilonproteobacteria (miscellaneous
Sulfurovum sp. NBC37-1
Nitratiruptor
Nautiliales
ular organisms
Betaproteobacteria
unclassified Proteobacteria
root
Zetaproteobacteria
unclassified Bacteria

```
Taxa=114     Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9     1063 of 1864M
```

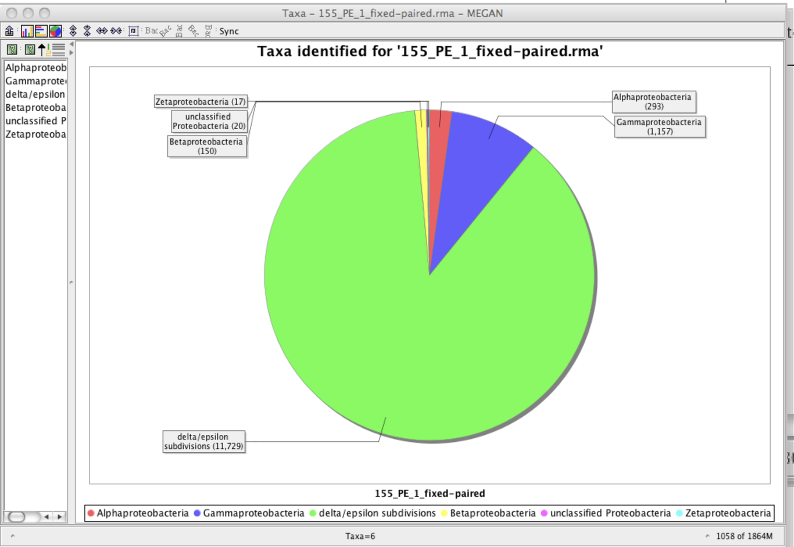
Taxonomic analysis of 50,000 reads

Interact and Summarize

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



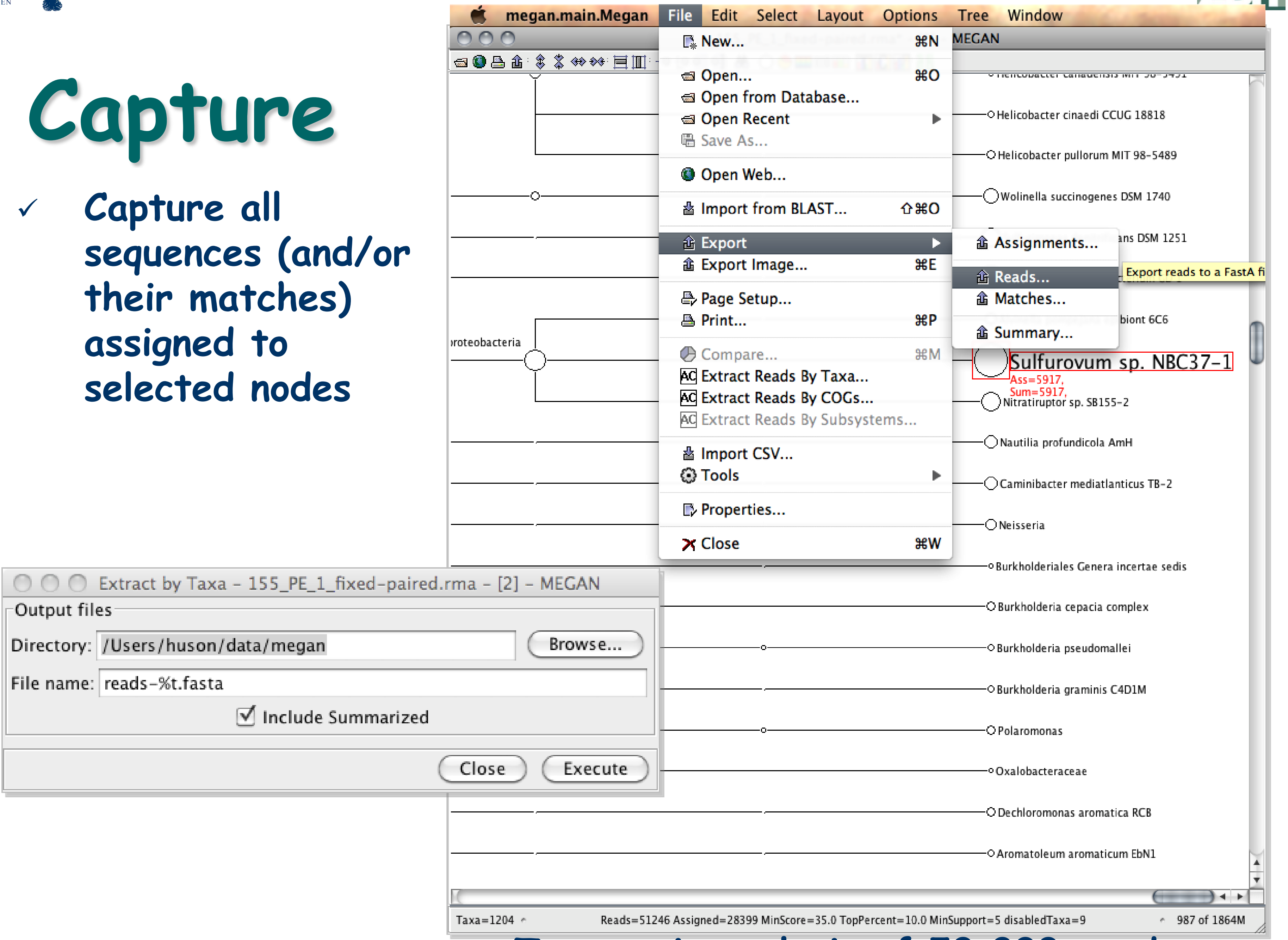
Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9


Taxa – 155_PE_1_fixed-paired.rma – MEGAN
Sync
Alphaproteob
Gammaprote
delta/epsilon
Betaproteoba
unclassified P
Zetaproteoba
Taxa identified for '155_PE_1_fixed-paired.rma'
Zetaproteobacteria (17)
unclassified Proteobacteria (20)
Betaproteobacteria (150)
Alphaproteobacteria (293)
Gammaproteobacteria (1,157)
delta/epsilon subdivisions (11,729)
155_PE_1_fixed-paired
● Alphaproteobacteria ● Gammaproteobacteria ● delta/epsilon subdivisions ● Betaproteobacteria ● unclassified Proteobacteria ● Zetaproteobacteria
Taxa=6
1058 of 1864M

Taxonomic analysis of 50,000 reads

Capture

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



Taxa=1204

Reads=51246 Assigned=28399 MinScore=35.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9

987 of 1864M

Extract by Taxa - 155_PE_1_fixed-paired.rma - [2] - MEGAN

Output files

Directory: /Users/huson/data/megan

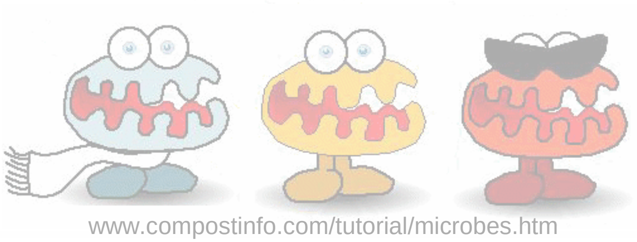
File name: reads-%t.fasta

☑ Include Summarized

Taxonomic analysis of 50,000 reads

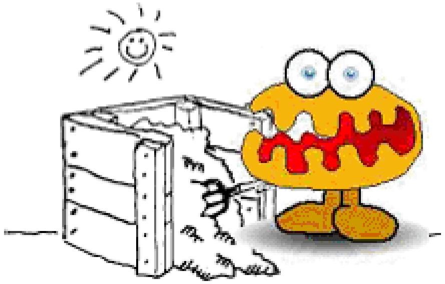
Three Basic Computational Questions

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

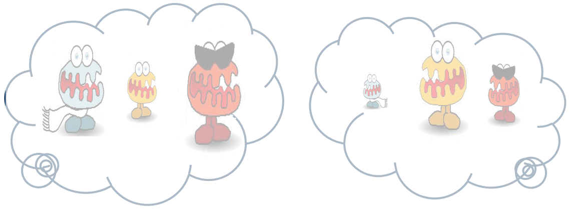


www.compostinfo.com/tutorial/microbes.htm

- 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?



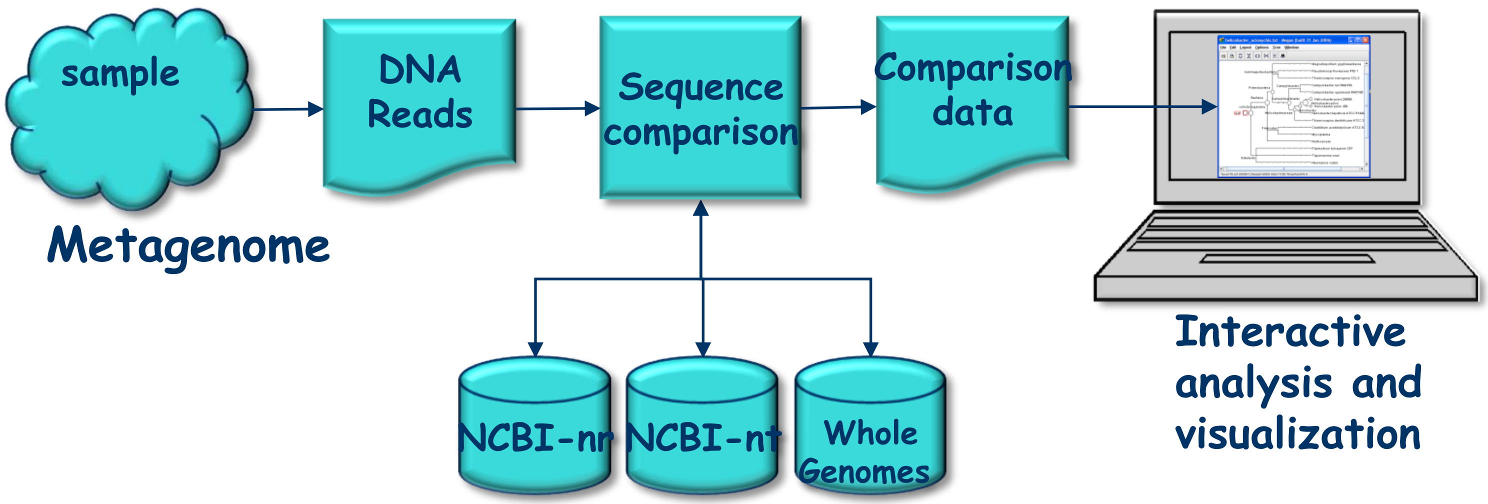
- Serve to answer biological or medical questions

Identifying Taxa and Genes

Metagenome analysis

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

BLASTX against NCBI-NR



Metagenome

Interactive analysis and visualization

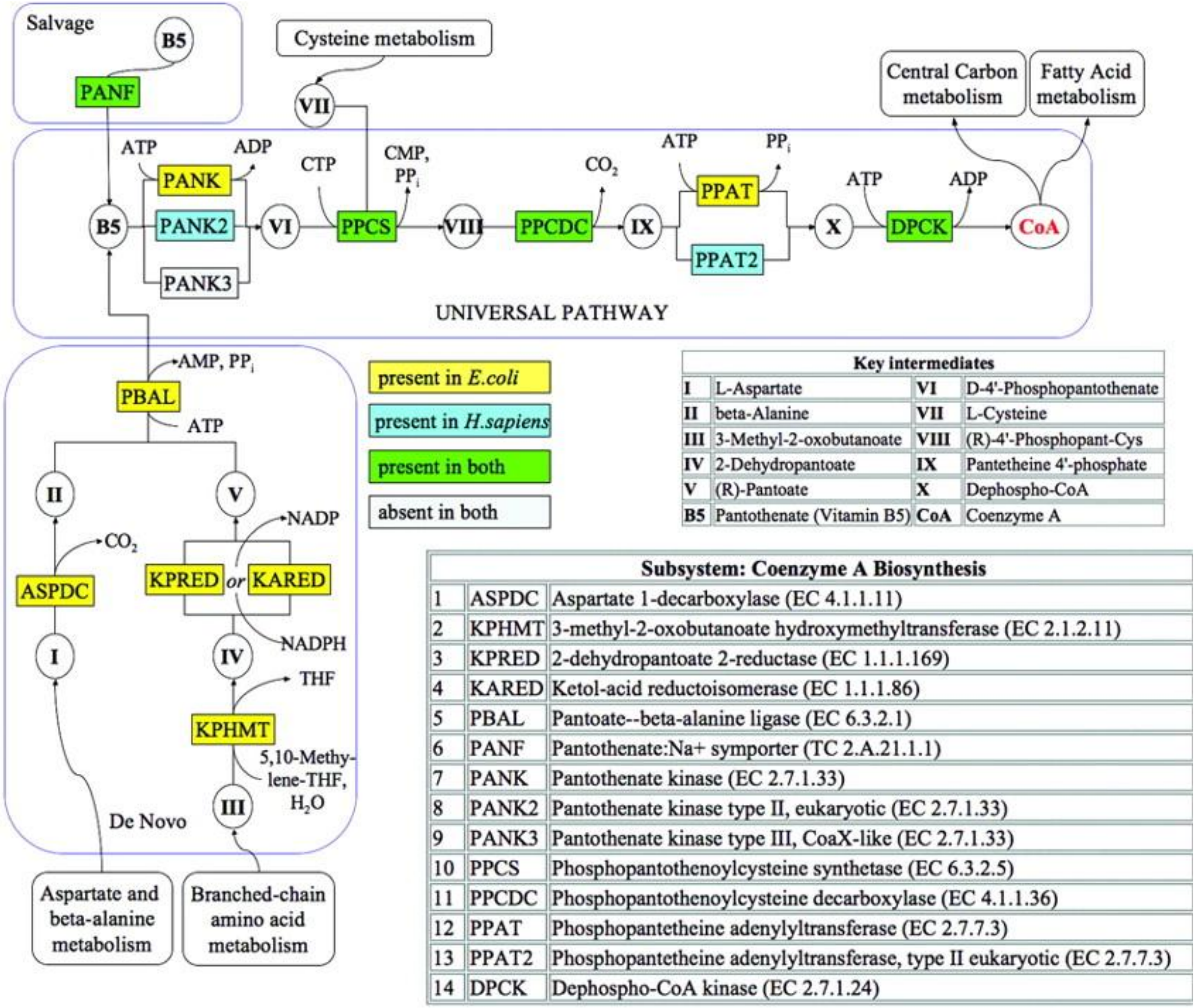
Reference databases

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

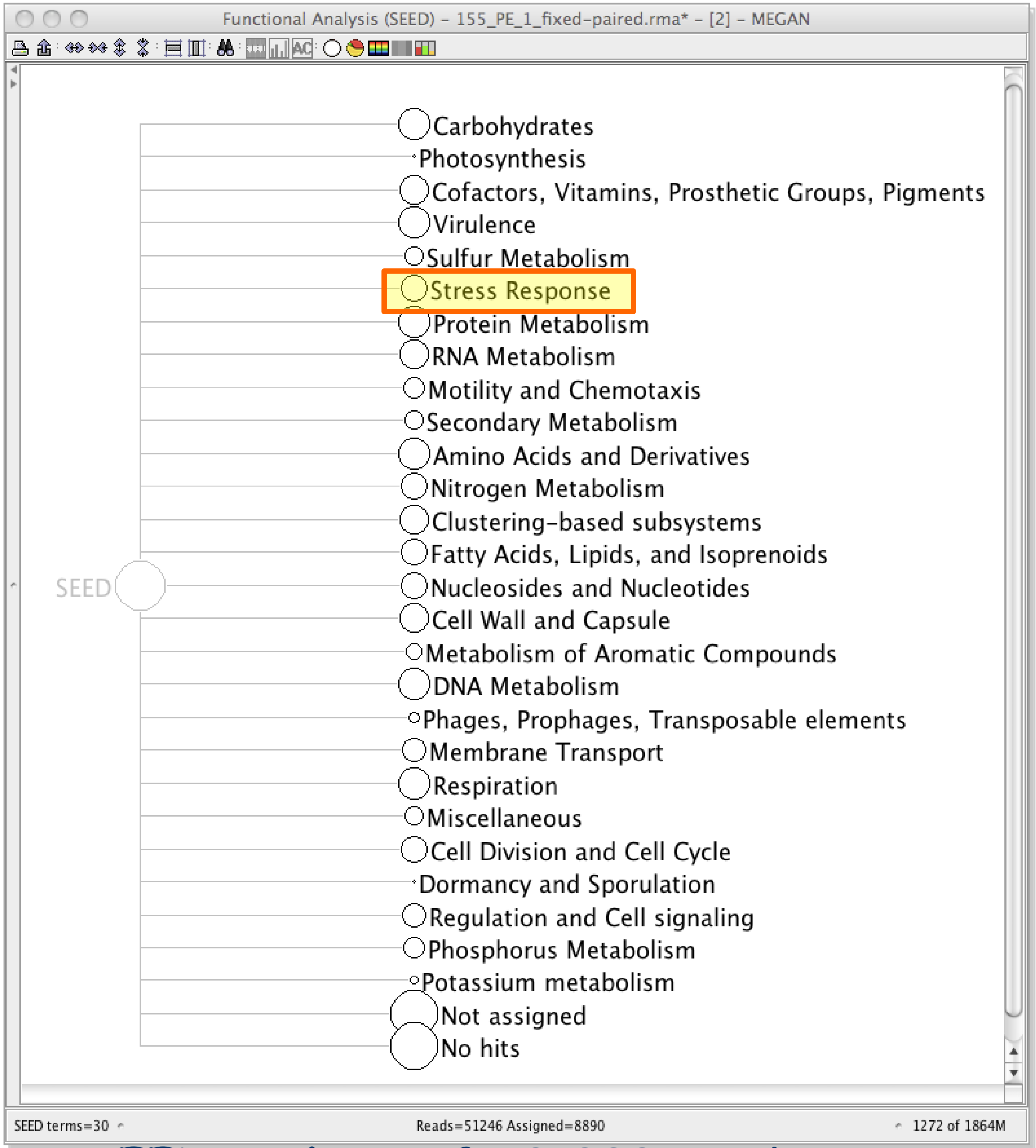
Organize and Visualize

Functional analysis

- ✓ Use SEED classification to bin sequences by subsystems

www.theseed.org

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


Functional Analysis (SEED) – 155_PE_1_fixed-paired.rma* – [2] – MEGAN

SEED terms=30 Reads=51246 Assigned=8890 1272 of 1864M

SEED analysis of 50,000 reads

Organize and Visualize

Functional analysis

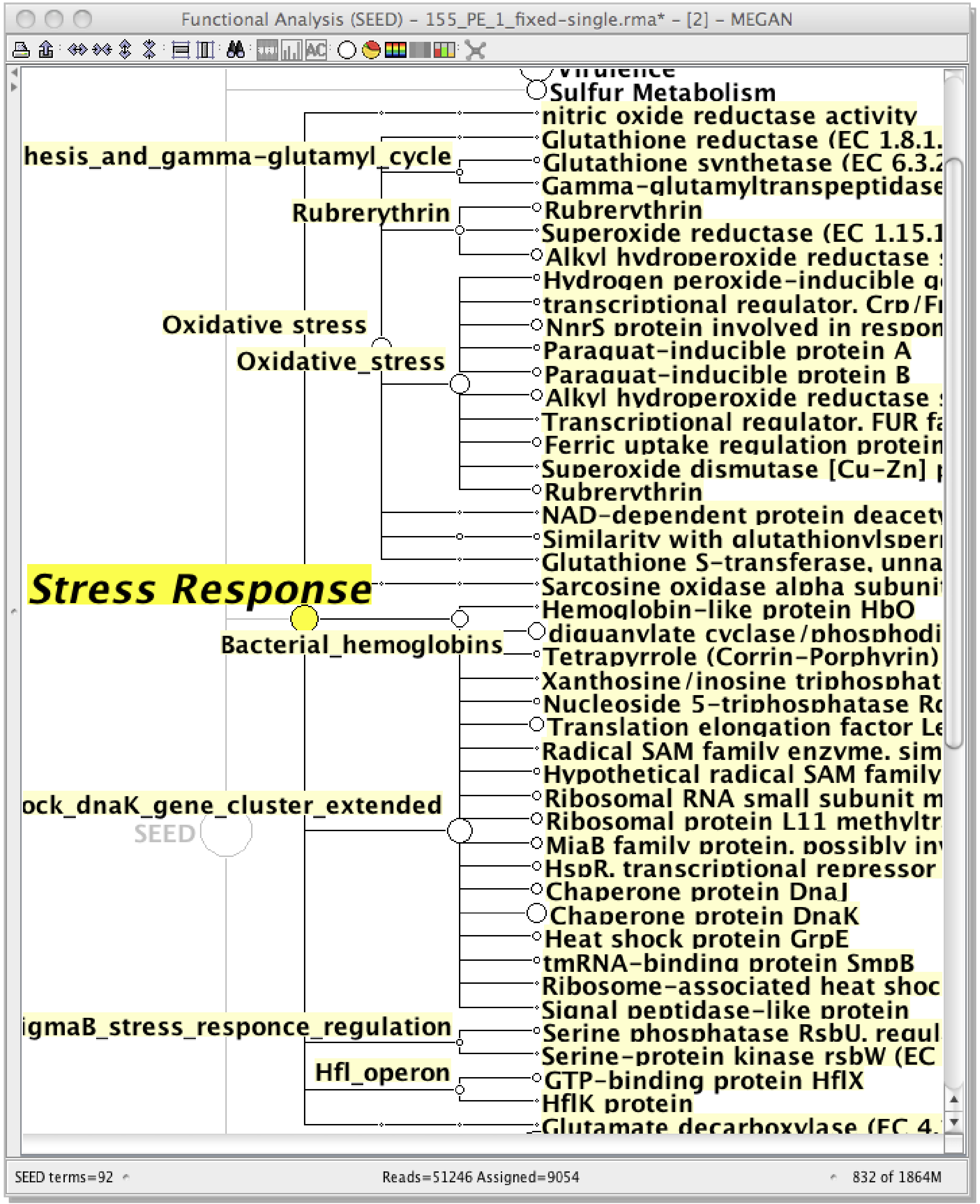
✓Use SEED classification to bin sequences by subsystems

✓… and by functional roles

The SEED
Home of the SEED.

www.theseed.org

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



Functional Analysis (SEED) – 155_PE_1_fixed-single.rma* – [2] – MEGAN

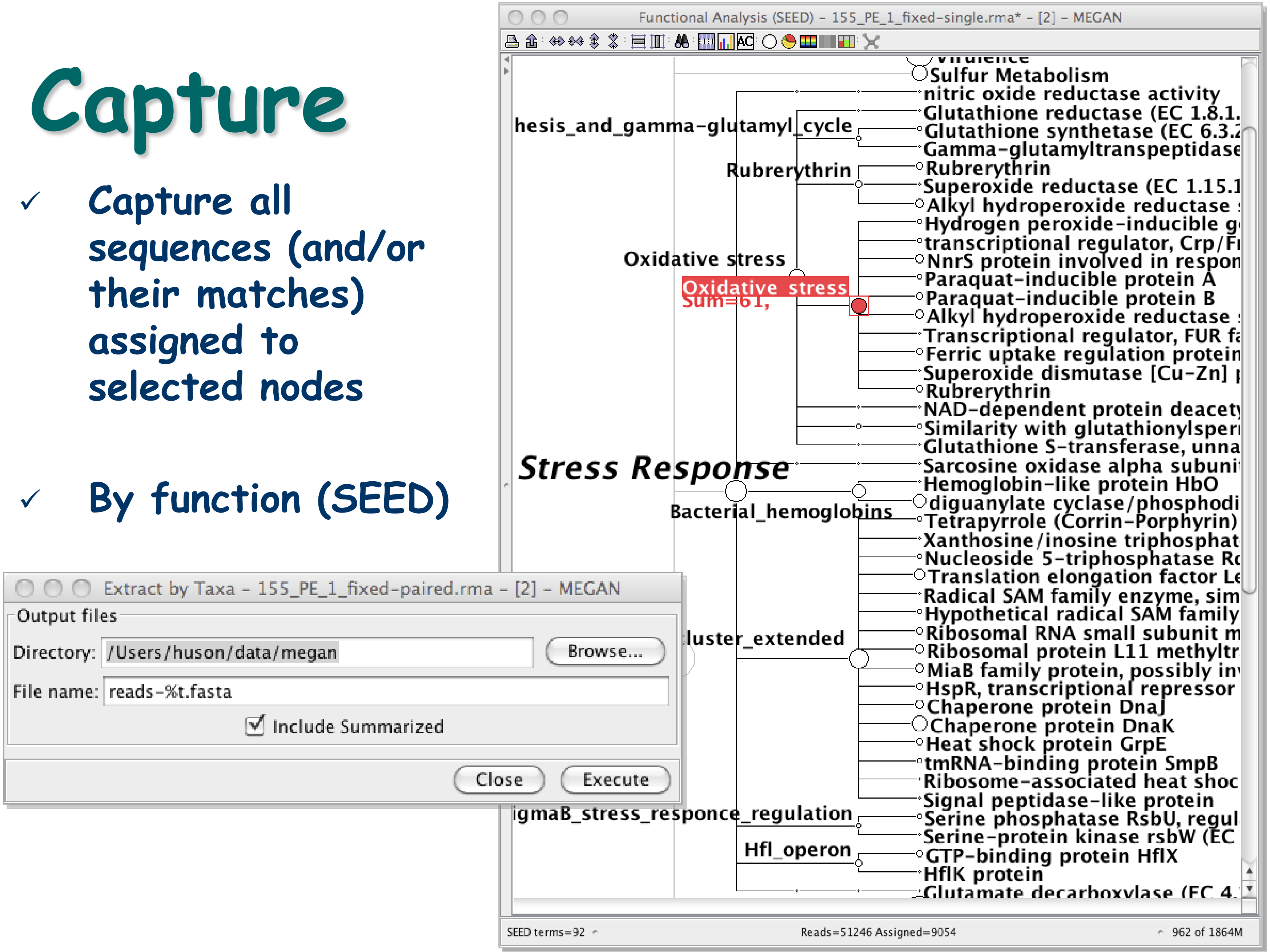
Virulence
Sulfur Metabolism
nitric oxide reductase activity
hesis_and_gamma-glutamyl_cycle
Glutathione reductase (EC 1.8.1.
Glutathione synthetase (EC 6.3.2
Gamma-glutamyltranspeptidase
Rubrerythrin
Rubrerythrin
Superoxide reductase (EC 1.15.1
Alkyl hydroperoxide reductase
Hydrogen peroxide-inducible g
Oxidative stress
transcriptional regulator, Crp/Fi
NnrS protein involved in respon
Oxidative_stress
Paraquat-inducible protein A
Paraquat-inducible protein B
Alkyl hydroperoxide reductase
Transcriptional regulator, FUR fa
Ferric uptake regulation protein
Superoxide dismutase [Cu-Zn] p
Rubrerythrin
NAD-dependent protein deacety
Similarity with glutathionylsper
Glutathione S-transferase, unna
Stress Response
Sarcosine oxidase alpha subuni
Hemoglobin-like protein HbO
Bacterial_hemoglobins
diguanylate cyclase/phosphodi
Tetrapyrrole (Corrin-Porphyrin)
Xanthosine/inosine triphosphat
Nucleoside 5-triphosphatase Rc
Translation elongation factor Le
Radical SAM family enzyme, sim
Hypothetical radical SAM family
Ribosomal RNA small subunit m
ock_dnaK_gene_cluster_extended
Ribosomal protein L11 methyltr
SEED
MiaB family protein, possibly inv
HspR, transcriptional repressor
Chaperone protein DnaJ
Chaperone protein DnaK
Heat shock protein GrpE
tmRNA-binding protein SmpB
Ribosome-associated heat shoc
Signal peptidase-like protein
igmaB_stress_responce_regulation
Serine phosphatase RsbU, regul
Serine-protein kinase rsbW (EC
Hfl_operon
GTP-binding protein HflX
HflK protein
Glutamate decarboxylase (EC 4.

SEED terms=92   Reads=51246 Assigned=9054   832 of 1864M

SEED analysis of 50,000 reads

Capture

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



Functional Analysis (SEED) – 155_PE_1_fixed-single.rma* – [2] – MEGAN
Virulence
Sulfur Metabolism
nitric oxide reductase activity
hesis_and_gamma-glutamyl_cycle
Glutathione reductase (EC 1.8.1.
Glutathione synthetase (EC 6.3.2
Gamma-glutamyltranspeptidase
Rubrerythrin
Rubrerythrin
Superoxide reductase (EC 1.15.1
Alkyl hydroperoxide reductase
Hydrogen peroxide-inducible g
transcriptional regulator, Crp/Fi
Oxidative stress
NnrS protein involved in respon
Paraquat-inducible protein A
Oxidative stress
Sum=61,
Paraquat-inducible protein B
Alkyl hydroperoxide reductase
Transcriptional regulator, FUR fa
Ferric uptake regulation protein
Superoxide dismutase [Cu-Zn] p
Rubrerythrin
NAD-dependent protein deacety
Similarity with glutathionylsperi
Glutathione S-transferase, unna
Stress Response
Sarcosine oxidase alpha subuni
Hemoglobin-like protein HbO
diguanylate cyclase/phosphodi
Bacterial_hemoglobins
Tetrapyrrole (Corrin-Porphyrin)
Xanthosine/inosine triphosphat
Nucleoside 5-triphosphatase Rd
Translation elongation factor Le
Radical SAM family enzyme, sim
Hypothetical radical SAM family
luster_extended
Ribosomal RNA small subunit m
Ribosomal protein L11 methyltr
MiaB family protein, possibly inv
HspR, transcriptional repressor
Chaperone protein DnaJ
Chaperone protein DnaK
Heat shock protein GrpE
tmRNA-binding protein SmpB
Ribosome-associated heat shoc
Signal peptidase-like protein
igmaB_stress_responce_regulation
Serine phosphatase RsbU, regul
Serine-protein kinase rsbW (EC
Hfl_operon
GTP-binding protein HflX
HflK protein
Glutamate decarboxylase (EC 4.
Extract by Taxa – 155_PE_1_fixed-paired.rma – [2] – MEGAN
Output files
Browse...
Close
Execute

Directory: /Users/huson/data/megan

File name: reads-%t.fasta

☑ Include Summarized

SEED terms=92 Reads=51246 Assigned=9054 962 of 1864M

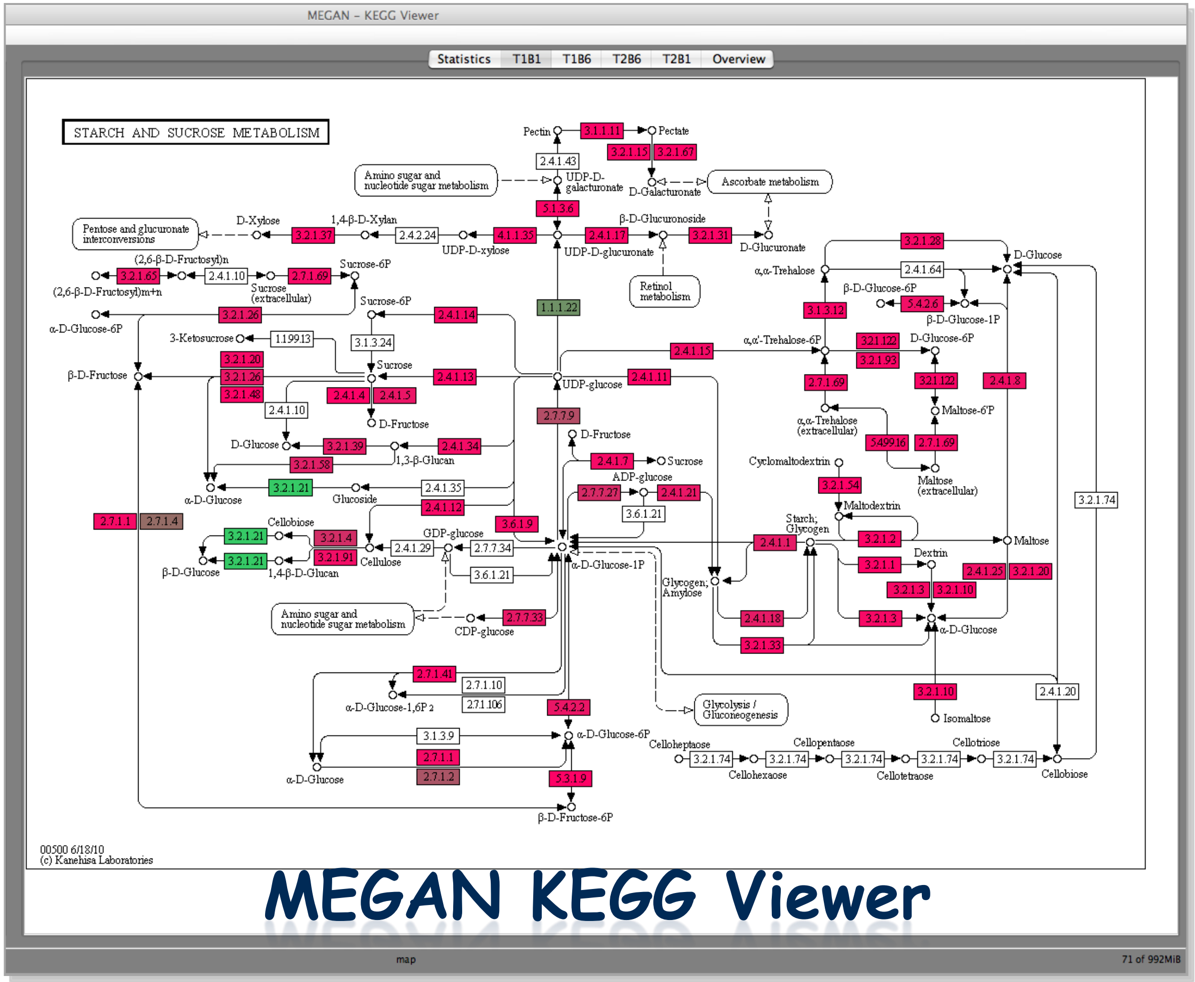
SEED analysis of 50,000 reads

Pathway Analysis

Functional analysis

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

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

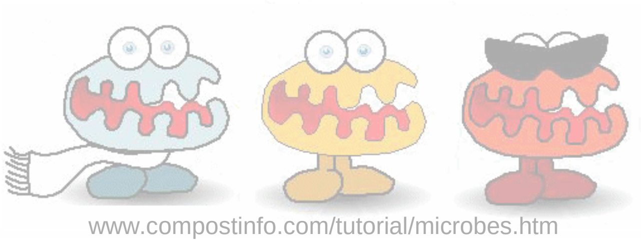


MEGAN KEGG Viewer

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

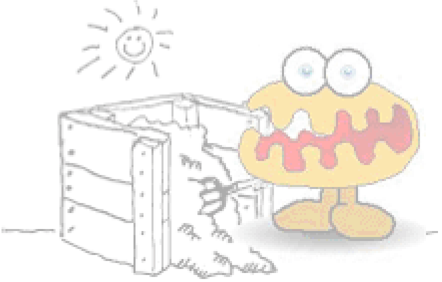
Three Basic Computational Questions

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

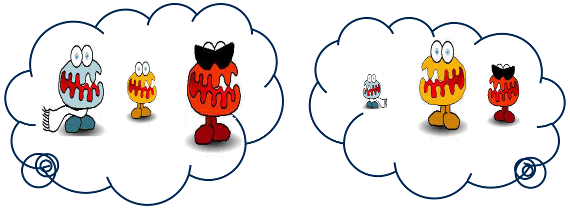


www.compostinfo.com/tutorial/microbes.htm

- 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?

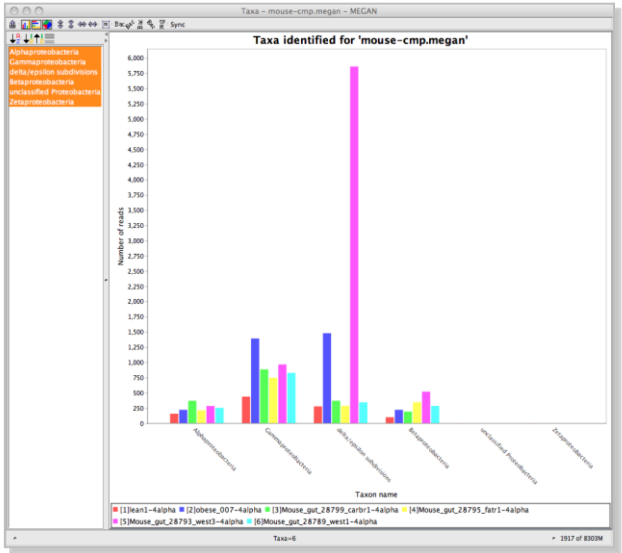


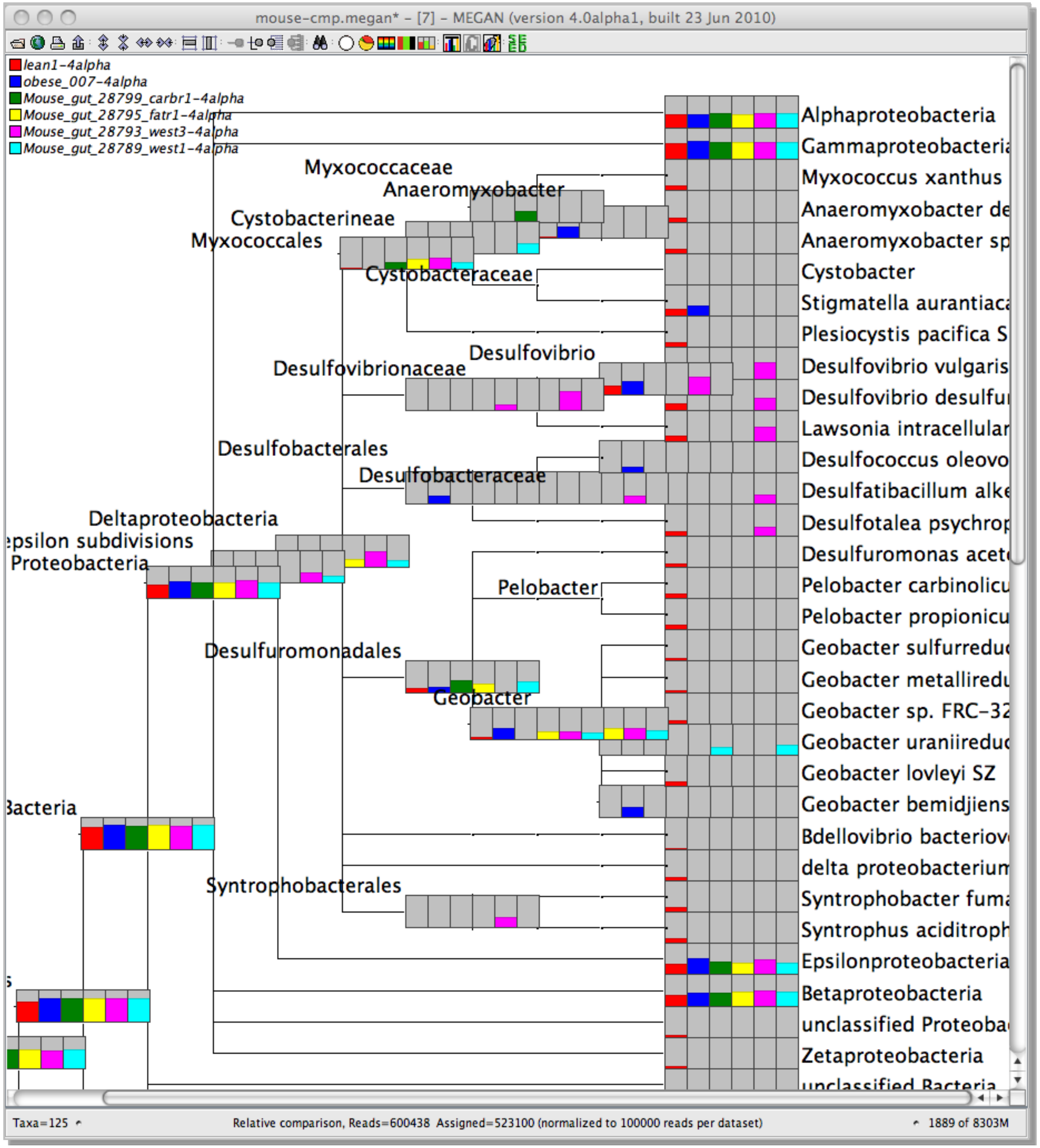
- Serve to answer biological or medical questions

Compare

Display multiple datasets simultaneously

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


Taxa – mouse-cmp.megan – MEGAN
Sync
Alphaproteobacteria
Gammaproteobacteria
delta/epsilon subdivisions
Betaproteobacteria
unclassified Proteobacteria
Zetaproteobacteria
Taxa identified for 'mouse-cmp.megan'
Number of reads
6,000
5,750
5,500
5,250
5,000
4,750
4,500
4,250
4,000
3,750
3,500
3,250
3,000
2,750
2,500
2,250
2,000
1,750
1,500
1,250
1,000
750
500
250
0
Alphaproteobacteria
Gammaproteobacteria
delta/epsilon subdivisions
Betaproteobacteria
unclassified Proteobacteria
Zetaproteobacteria
Taxon name
🟥 [1]lean1-4alpha 🟦 [2]obese_007-4alpha 🟩 [3]Mouse_gut_28799_carbr1-4alpha 🟨 [4]Mouse_gut_28795_fatr1-4alpha
🟪 [5]Mouse_gut_28793_west3-4alpha 🟦 [6]Mouse_gut_28789_west1-4alpha
Taxa=6
1917 of 8303M


mouse-cmp.megan* – [7] – MEGAN (version 4.0alpha1, built 23 Jun 2010)

🟥 lean1-4alpha
🟦 obese_007-4alpha
🟩 Mouse_gut_28799_carbr1-4alpha
🟨 Mouse_gut_28795_fatr1-4alpha
🟪 Mouse_gut_28793_west3-4alpha
🟦 Mouse_gut_28789_west1-4alpha

Taxa=125

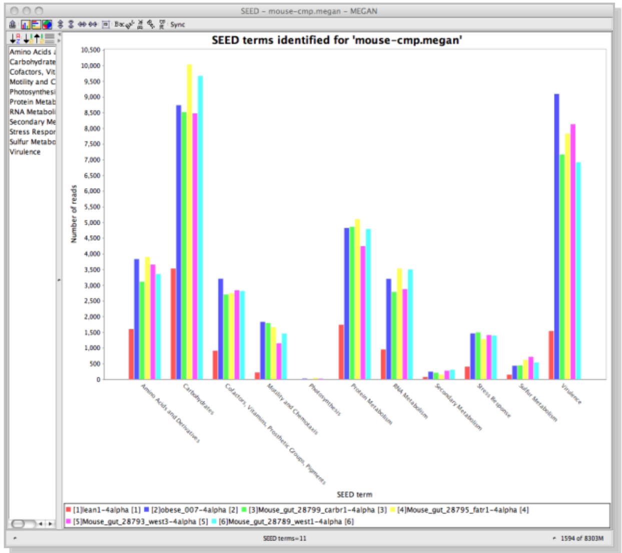
Relative comparison, Reads=600438 Assigned=523100 (normalized to 100000 reads per dataset)

1889 of 8303M

Compare

Display multiple datasets simultaneously

- ✓ Taxonomical comparison
- ✓ Interact
- ✓ … and summarize

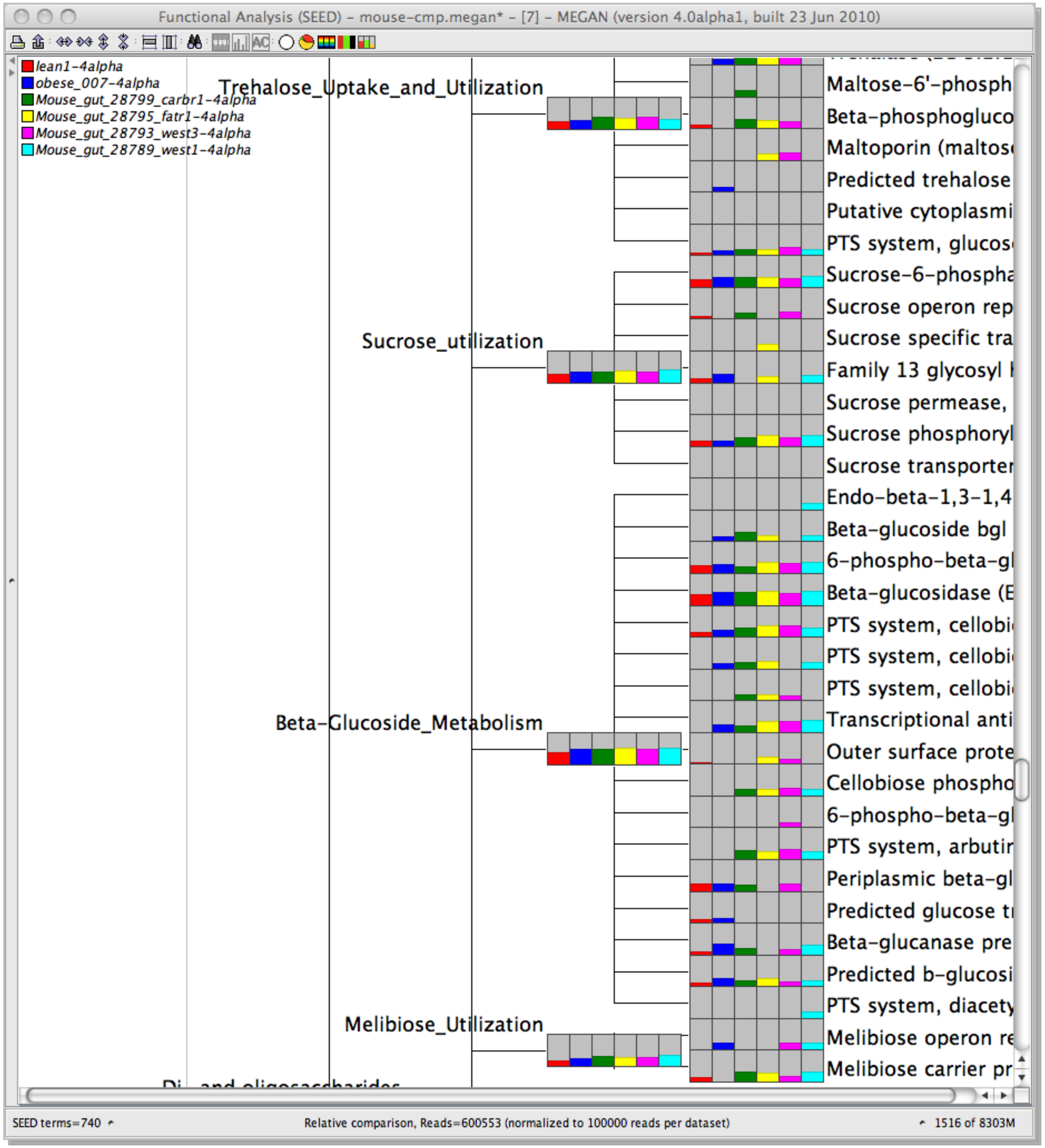


SEED – mouse-cmp.megan – MEGAN

SEED terms identified for 'mouse-cmp.megan'

SEED terms=11

1594 of 8303M



Functional Analysis (SEED) – mouse-cmp.megan* – [7] – MEGAN (version 4.0alpha1, built 23 Jun 2010)

SEED terms=740

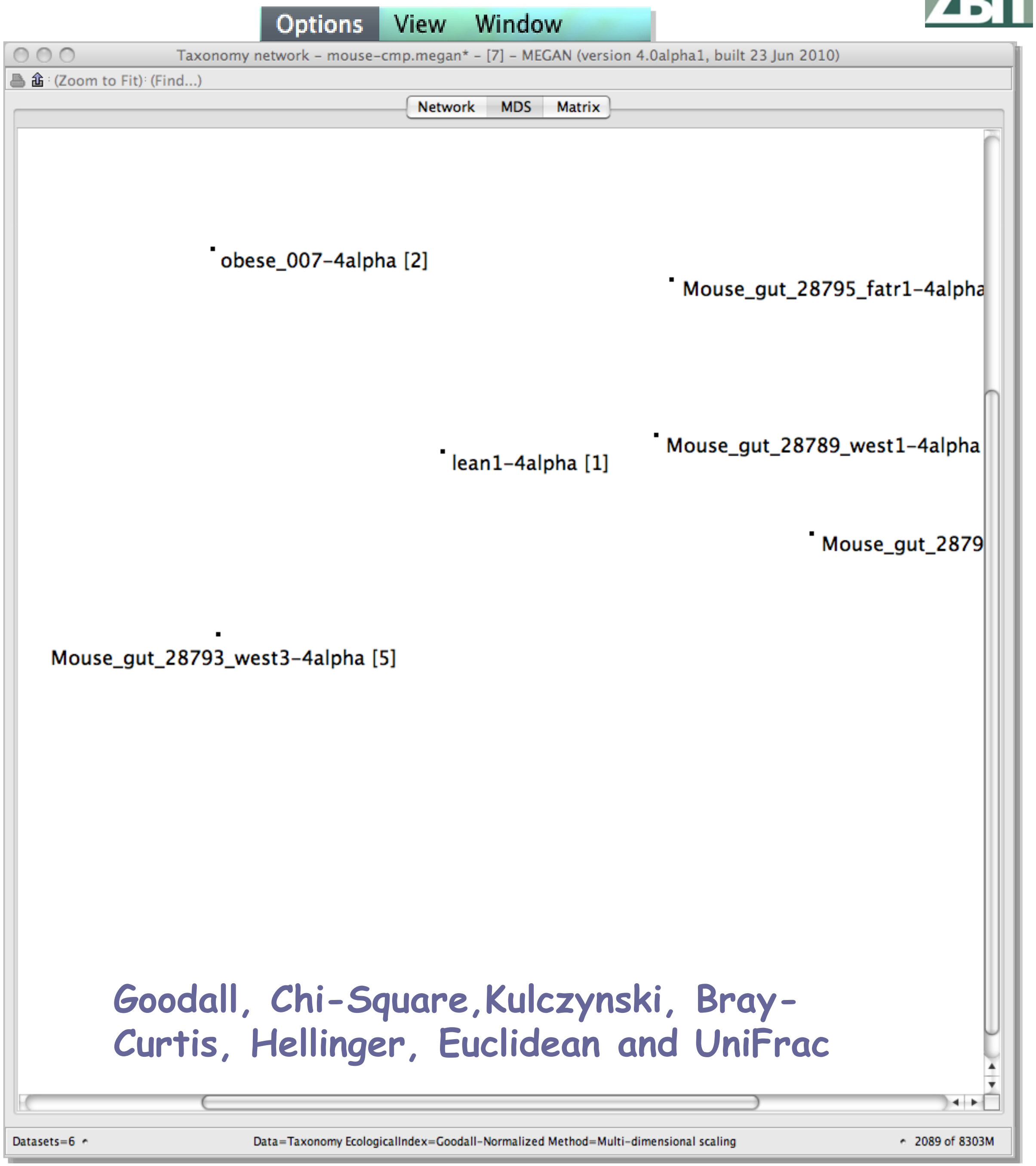
Relative comparison, Reads=600553 (normalized to 100000 reads per dataset)

1516 of 8303M

Compare

High-level comparison:

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



Taxonomy network - mouse-cmp.megan* - [7] - MEGAN (version 4.0alpha1, built 23 Jun 2010)

Datasets=6 Data=Taxonomy EcologicalIndex=Goodall-Normalized Method=Multi-dimensional scaling 2089 of 8303M

Goodall, Chi-Square,Kulczynski, Bray-Curtis, Hellinger, Euclidean and UniFrac

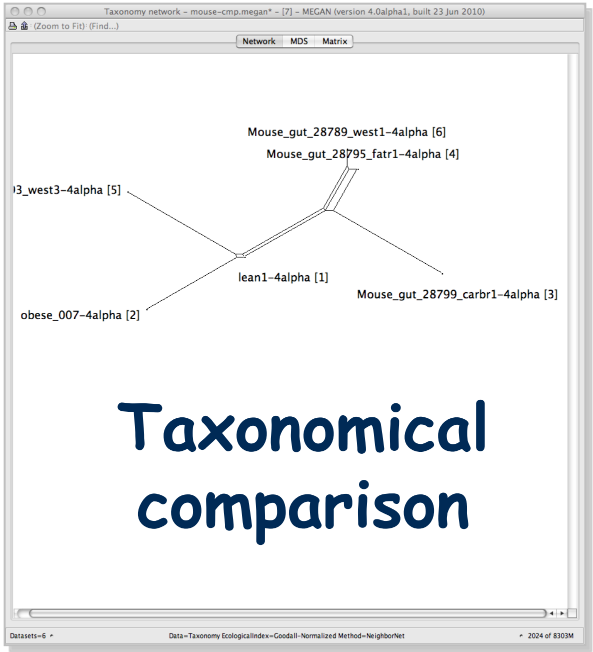
Mitra, Gilbert, Field and Huson, ISME J, 2010

Neighbor-net: Bryant and Moulton, 2003

Compare

High-level comparison:

- ✓ Select taxa



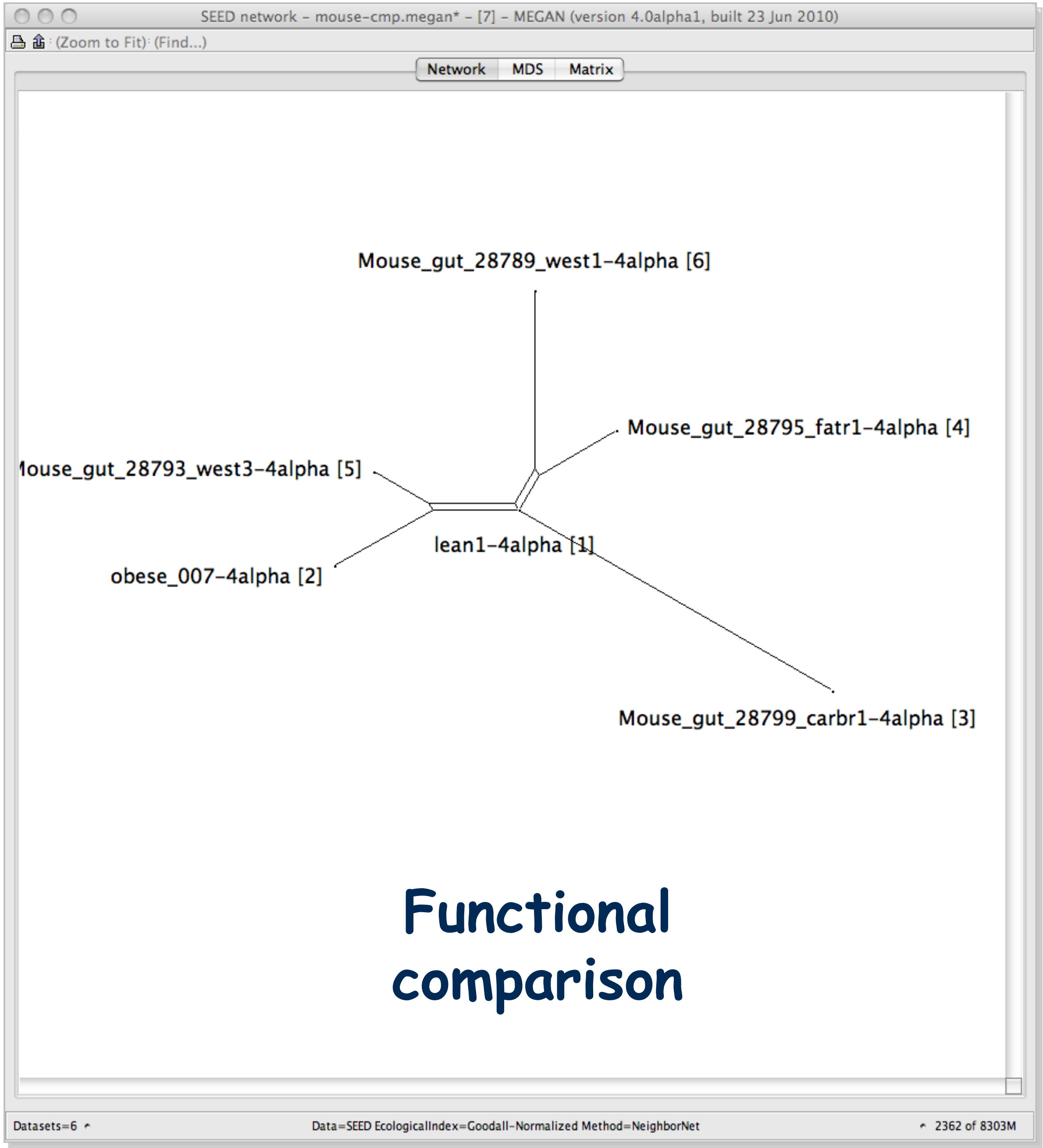
(Zoom to Fit) (Find...)
Network
MDS
Matrix
Mouse_gut_28789_west1–4alpha [6]
Mouse_gut_28795_fatr1–4alpha [4]
)3_west3–4alpha [5]
lean1–4alpha [1]
Mouse_gut_28799_carbr1–4alpha [3]
obese_007–4alpha [2]

Taxonomy network – mouse-cmp.megan* – [7] – MEGAN (version 4.0alpha1, built 23 Jun 2010)

Datasets=6 Data=Taxonomy EcologicalIndex=Goodall-Normalized Method=NeighborNet 2024 of 8303M

Taxonomical comparison

- ✓ Select functions…



(Zoom to Fit) (Find...)
Network
MDS
Matrix
Mouse_gut_28789_west1–4alpha [6]
Mouse_gut_28795_fatr1–4alpha [4]
/ouse_gut_28793_west3–4alpha [5]
lean1–4alpha [1]
obese_007–4alpha [2]
Mouse_gut_28799_carbr1–4alpha [3]

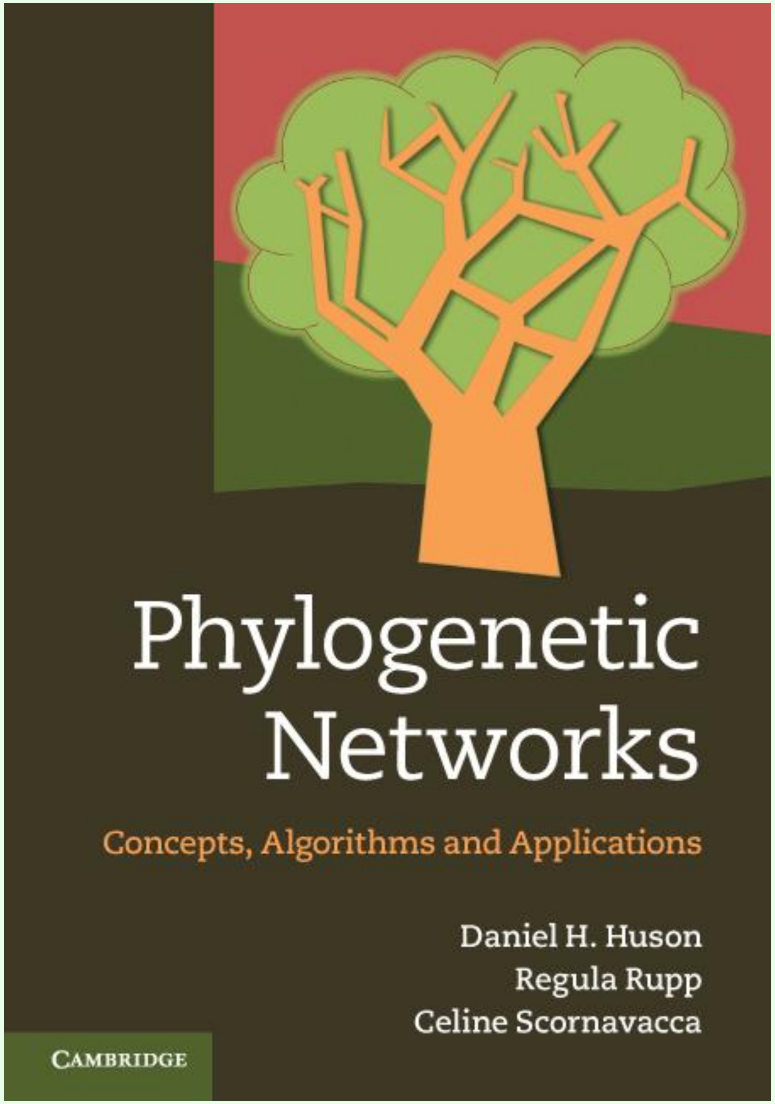
SEED network – mouse-cmp.megan* – [7] – MEGAN (version 4.0alpha1, built 23 Jun 2010)

Datasets=6 Data=SEED EcologicalIndex=Goodall-Normalized Method=NeighborNet 2362 of 8303M

Functional comparison

Phylogenetic Networks

Concepts, Algorithms and Applications



Phylogenetic Networks

Concepts, Algorithms and Applications

Daniel H. Huson
Regula Rupp
Celine Scornavacca

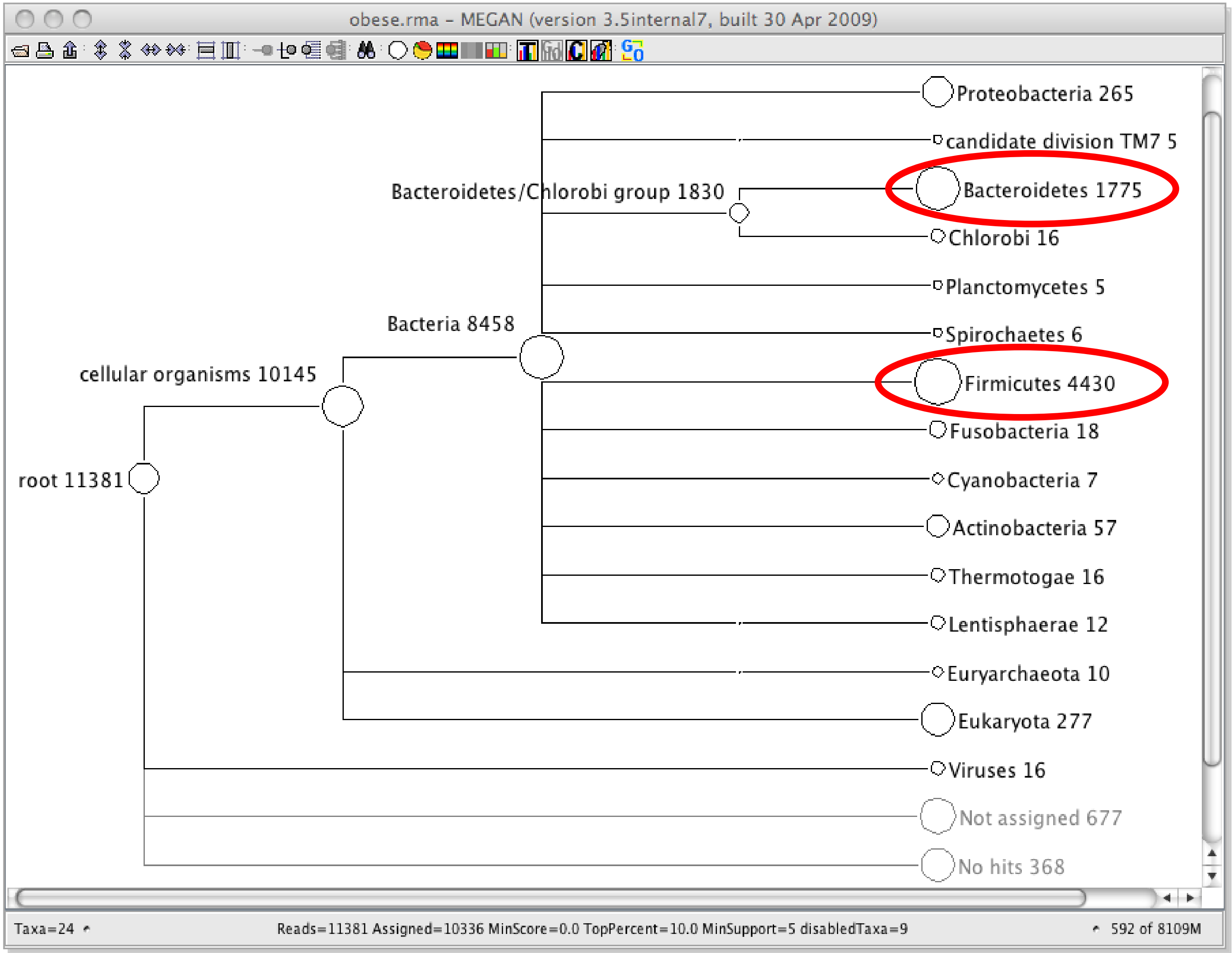
CAMBRIDGE

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

~ 40 EUR

Dec 2010

Example: Mouse Gut Microbiome



obese.rma – MEGAN (version 3.5internal7, built 30 Apr 2009)

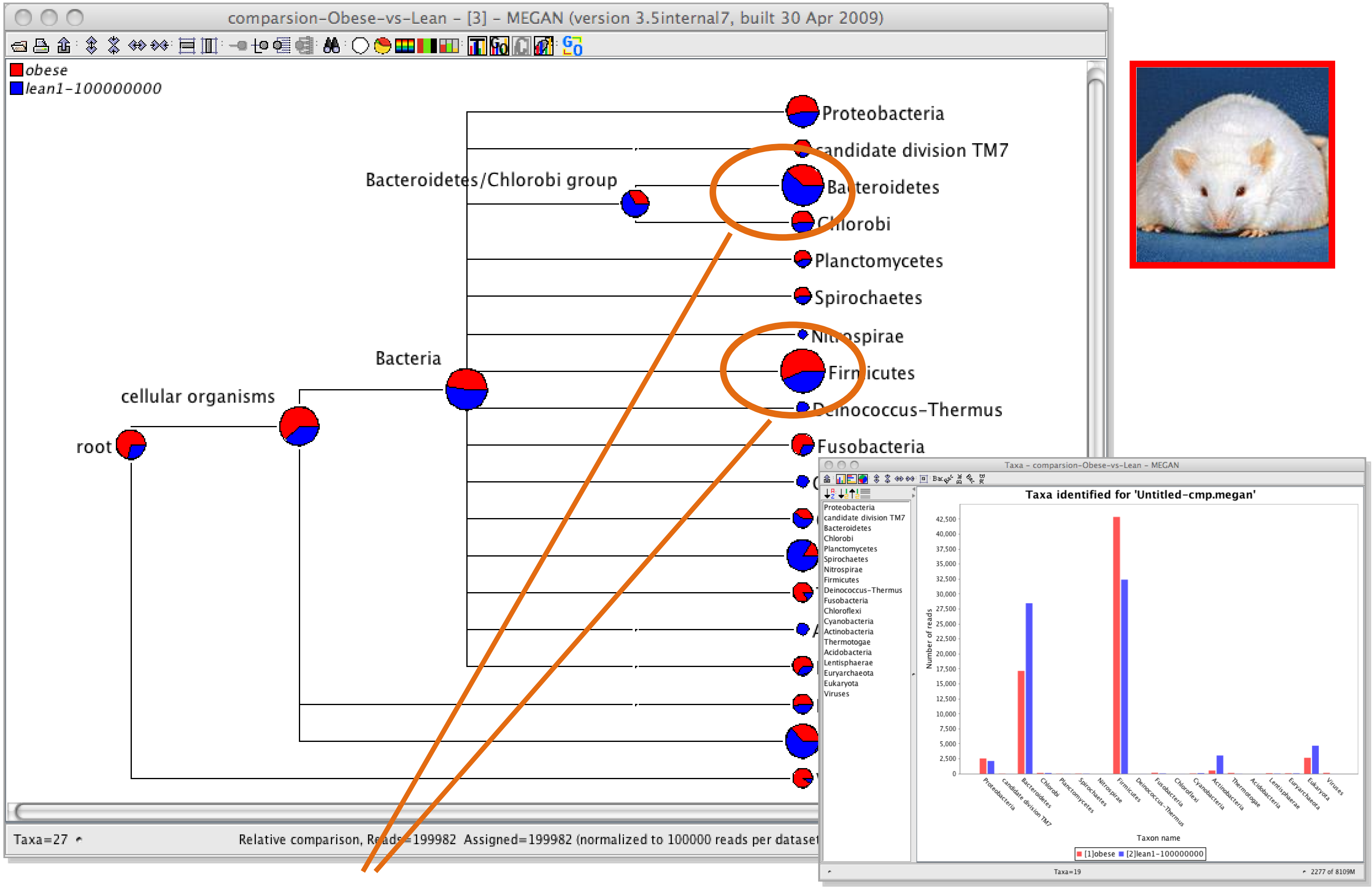
Taxa=24

Reads=11381 Assigned=10336 MinScore=0.0 TopPercent=10.0 MinSupport=5 disabledTaxa=9

592 of 8109M

Dominant gut microbes: Bacteroidetes and Firmicutes

Comparative Analysis



Relative comparison, Reads=199982 Assigned=199982 (normalized to 100000 reads per dataset

Change in proportions of these two phyla

Basic Principles

How to analyze a meta genome?

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

Contents

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

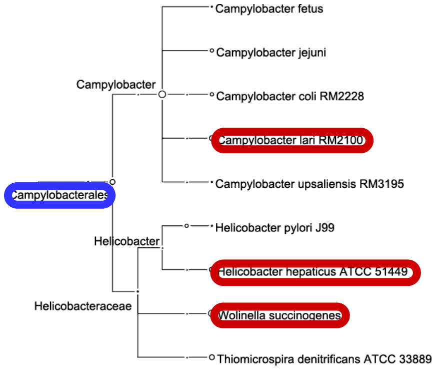
Single Reads vs Paired Reads

In metagenomics:

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

Taxonomical Analysis Based on Gene Content

Set of all species containing gene


Campylobacter fetus
Campylobacter jejuni
Campylobacter
Campylobacter coli RM2228
Campylobacter lari RM2100
Campylobacter upsaliensis RM3195
Campylobacterales
Helicobacter pylori J99
Helicobacter
Helicobacter hepaticus ATCC 51449
Helicobacteraceae
Wolinella succinogenes
Thiomicrospira denitrificans ATCC 33889

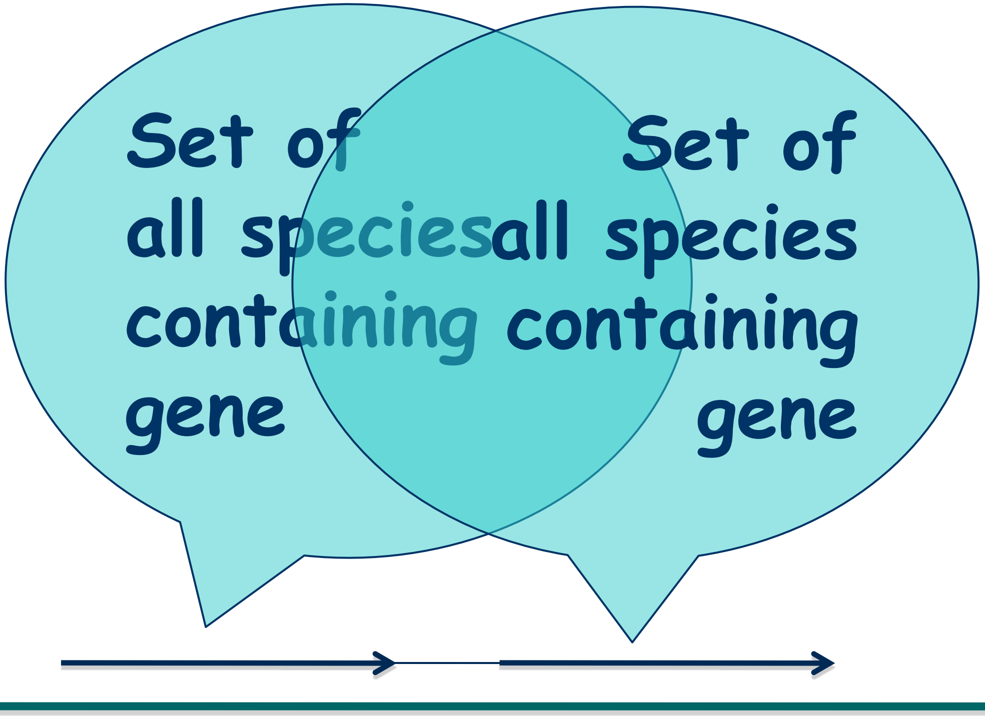
read →

Unknown source genome

Use LCA to assign to (higher-rank) taxon

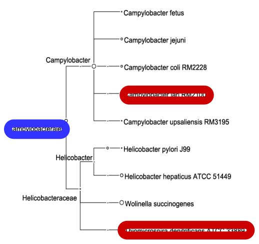
Taxonomical Analysis Based on Gene Content

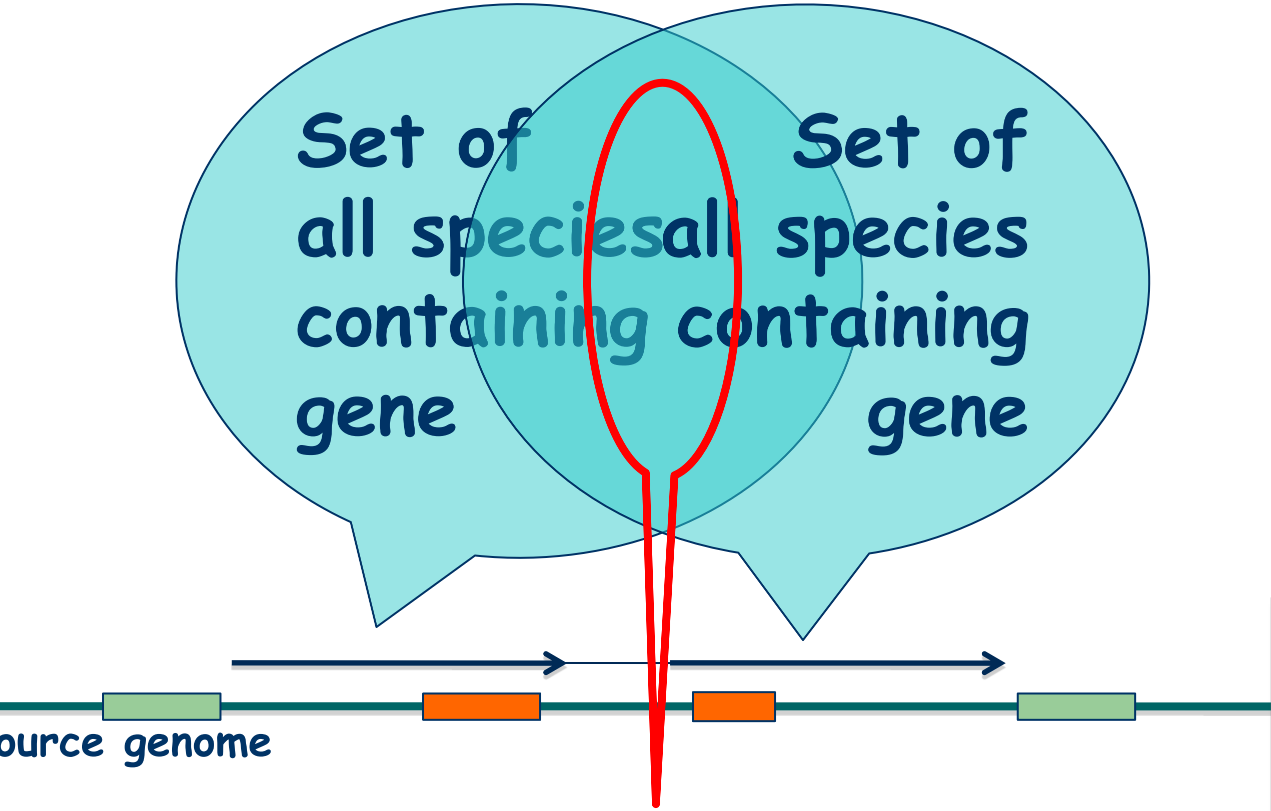
Short clones or long clones?


Set of all species containing gene
Set of all species containing gene
Unknown source genome

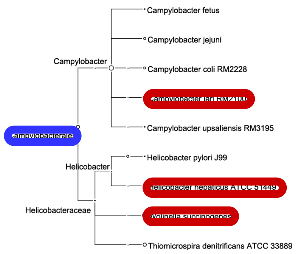
Taxonomical Analysis Based on Gene Content

Claim: long clones are more specific

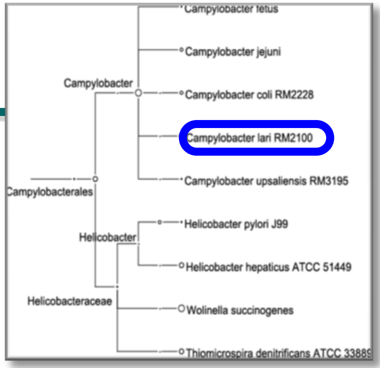



Set of all species containing gene

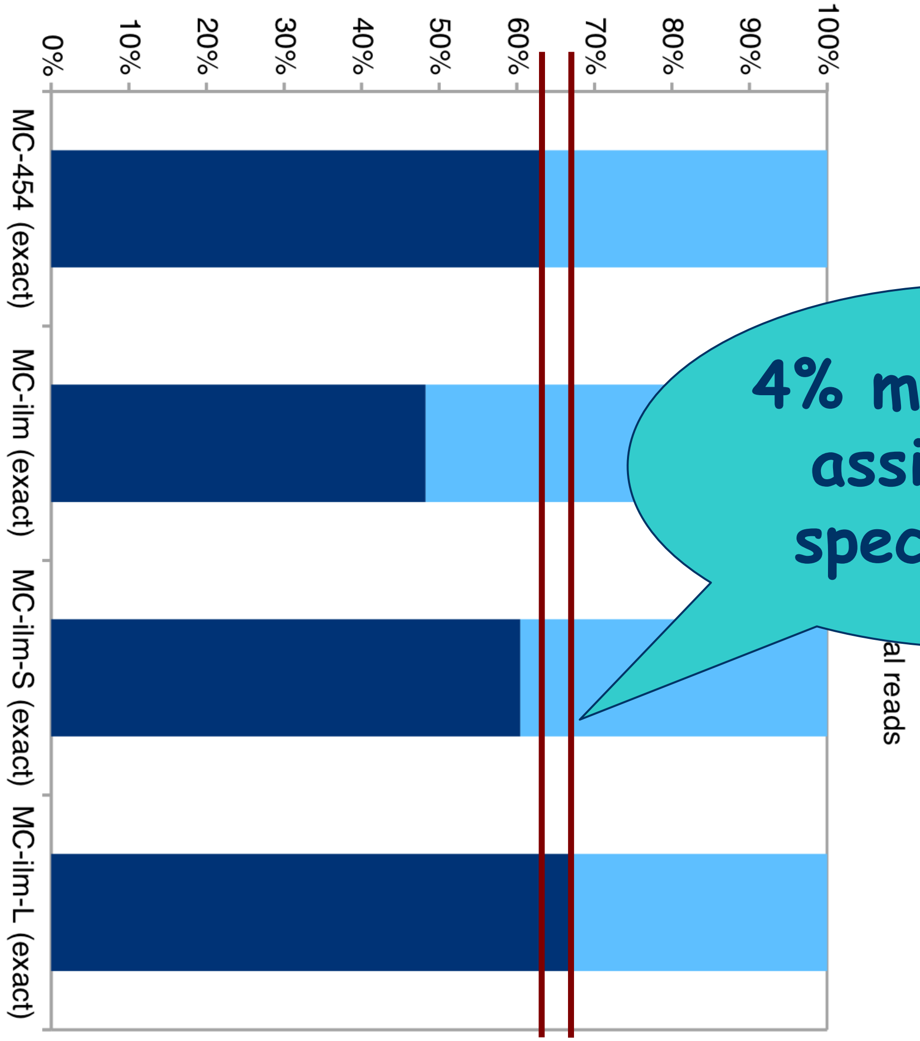
Set of all species containing gene



LCA of intersection of species:


Campylobacter fetus
Campylobacter jejuni
Campylobacter
Campylobacter coli RM2228
Campylobacter lari RM2100
Campylobacter upsaliensis RM3195
Campylobacterales
Helicobacter pylori J99
Helicobacter
Helicobacter hepaticus ATCC 51449
Helicobacteraceae
Wolinella succinogenes
Thiomicrospira denitrificans ATCC 3388

Simulated Performance 454 vs Illumina



250bp, single

75bp, single

75bp,short clones

75bp, long clones

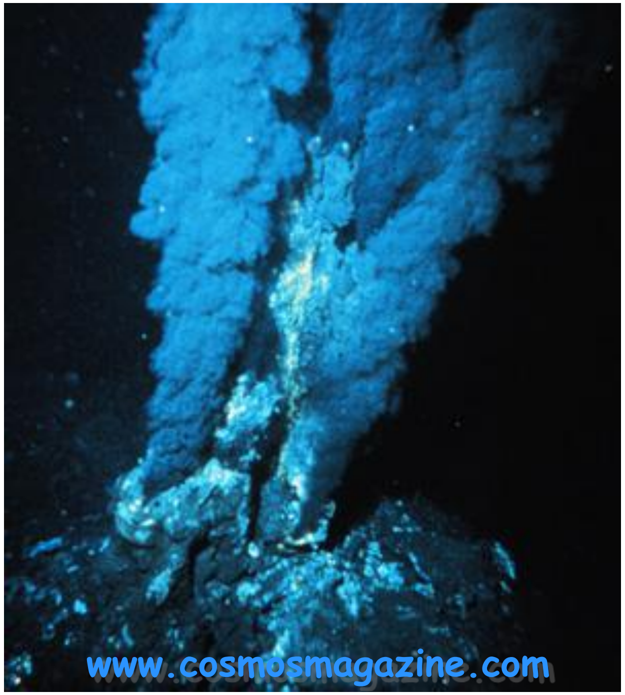
4% more reads assigned at species level

% reads (correctly) assigned at species level

MetaSim – simulator (Richter et al, 2008)

Ocean Seabed Sample

- Joint work with Ida Steen (Bergen)



www.cosmosmagazine.com

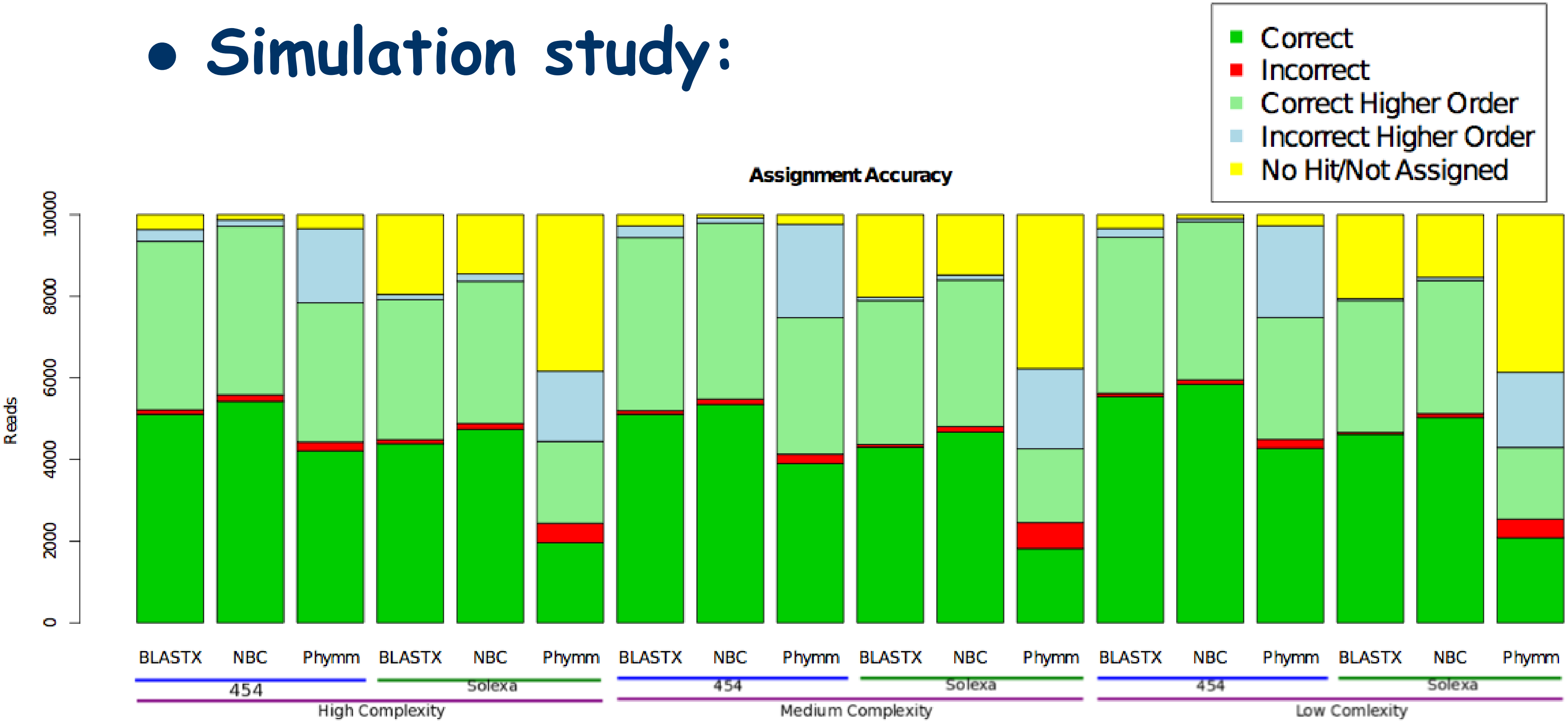
- Currently analyzing large set of 454 paired reads (7kb clones) collected from an Arctic hot vent at 3km depth

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:

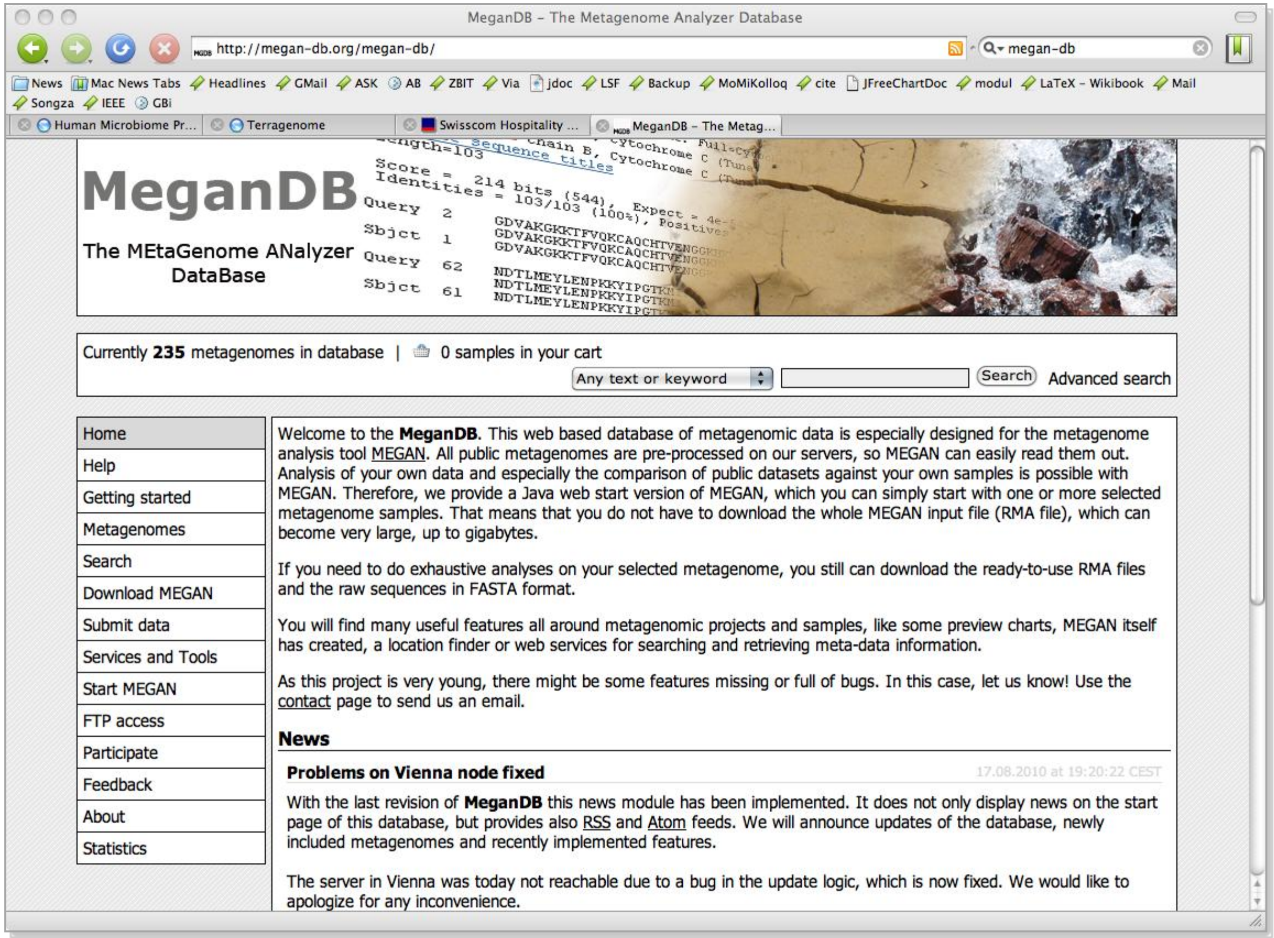


- ~10x speed-up over full BLASTX
- Increased accuracy using NCB
- Decreased accuracy using Phymm

Weber et al, submitted

MEGAN-DB

- Database of precomputed MEGAN files


MeganDB - The Metagenome Analyzer Database
http://megan-db.org/megan-db/
megan-db
News
Mac News Tabs
Headlines
GMail
ASK
AB
ZBIT
Via
jdoc
LSF
Backup
MoMiKolloq
cite
JFreeChartDoc
modul
LaTeX - Wikibook
Mail
Songza
IEEE
GBi
Human Microbiome Pr...
Terragenome
Swisscom Hospitality ...
MeganDB - The Metag...
Score = 214 bits (544), Expect = 4e-
Identities = 103/103 (100%), Positives
Query 2 GDVAKGKKTFVQKCAQCHTVENGG
Sbjct 1 GDVAKGKKTFVQKCAQCHTVENGG
Query 62 NDTLMEYLENPKKYIPGTK
Sbjct 61 NDTLMEYLENPKKYIPGTK
Any text or keyword
Search
Advanced search

MeganDB

The MEtaGenome ANalyzer DataBase

Currently **235** metagenomes in database | 0 samples in your cart

- Home
- Help
- Getting started
- Metagenomes
- Search
- Download MEGAN
- Submit data
- Services and Tools
- Start MEGAN
- FTP access
- Participate
- Feedback
- About
- Statistics

Welcome to the **MeganDB**. This web based database of metagenomic data is especially designed for the metagenome analysis tool MEGAN. All public metagenomes are pre-processed on our servers, so MEGAN can easily read them out. Analysis of your own data and especially the comparison of public datasets against your own samples is possible with MEGAN. Therefore, we provide a Java web start version of MEGAN, which you can simply start with one or more selected metagenome samples. That means that you do not have to download the whole MEGAN input file (RMA file), which can become very large, up to gigabytes.

If you need to do exhaustive analyses on your selected metagenome, you still can download the ready-to-use RMA files and the raw sequences in FASTA format.

You will find many useful features all around metagenomic projects and samples, like some preview charts, MEGAN itself has created, a location finder or web services for searching and retrieving meta-data information.

As this project is very young, there might be some features missing or full of bugs. In this case, let us know! Use the contact page to send us an email.

News

Problems on Vienna node fixed 17.08.2010 at 19:20:22 CEST

With the last revision of **MeganDB** this news module has been implemented. It does not only display news on the start page of this database, but provides also RSS and Atom feeds. We will announce updates of the database, newly included metagenomes and recently implemented features.

The server in Vienna was today not reachable due to a bug in the update logic, which is now fixed. We would like to apologize for any inconvenience.

- www.megan-db.org

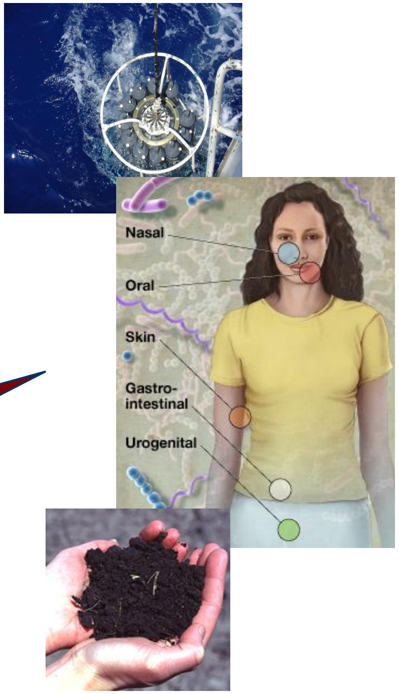
Joint work with Thomas Rattei and Simon Domke

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

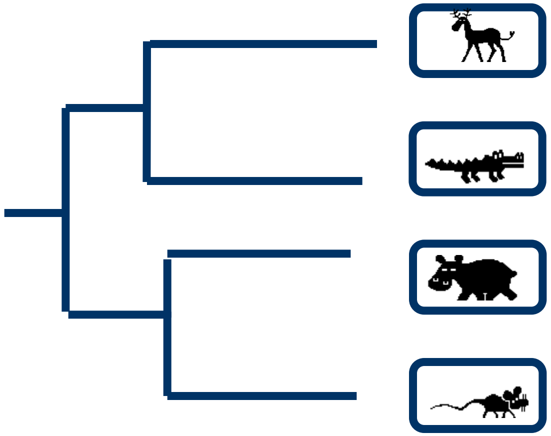
(mega, giga, tera, peta, exa…)

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