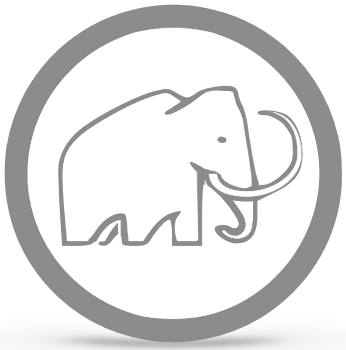
# Interactive Microbiome Analysis using DIAMOND+MEGAN



Daniel H. Huson
Anupam Gautam
Wenhuan Zeng

Tutorial web page

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Time table

| | |
|---|---|
| • Introduction to microbiome analysis | 9:00- 9:45 |
| • Part 0: Software setup | 9:45-10:00 |
| • Part 1: DIAMOND alignment | 10:00-10:45 |
| *Coffee* | 10:45-11:15 |
| • Part 2: Meganization | 11:15-11:45 |
| • Part 3: Interactive MEGAN | 11:45-13:00 |

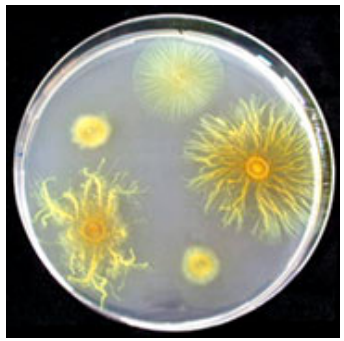
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

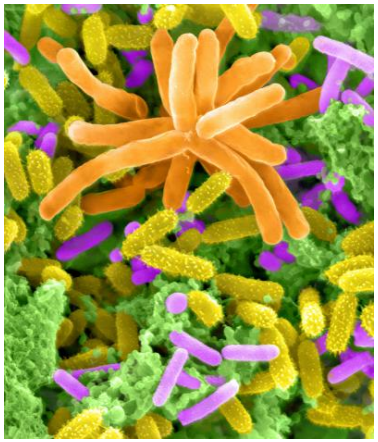
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Microbiome



www.innovations-report.de

- Traditionally, microbes are studied in pure culture
- Genome:
  - Entire DNA sequence of a single organism
- *But*: most microbes don't live in isolation and many can't be cultured
- Microbiome:
  - Collection of microbes in a specific theatre of activity
- Metagenome:
  - Entire DNA sequence of a microbiome

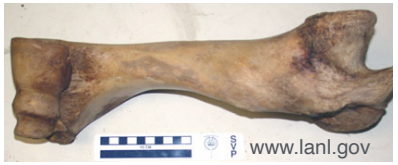


www.physorg.com

# Sources of studied microbiomes

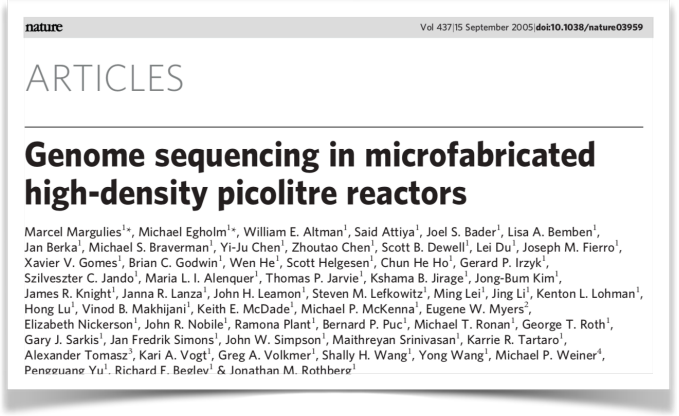
- Soil samples
- Water samples
- Seabed samples
- Air samples
- Ancient bones
- Host-associated samples
- Human microbiome
- …



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics


- First NGS technique *454* released


nature Vol 437|15 September 2005|doi:10.1038/nature03959

ARTICLES

**Genome sequencing in microfabricated high-density picolitre reactors**

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[3], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

- Intended for genome sequencing…

★Use NGS to sequence ancient DNA?

★Use NGS to sequence metagenomic DNA?

NGS = next generation sequencing

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

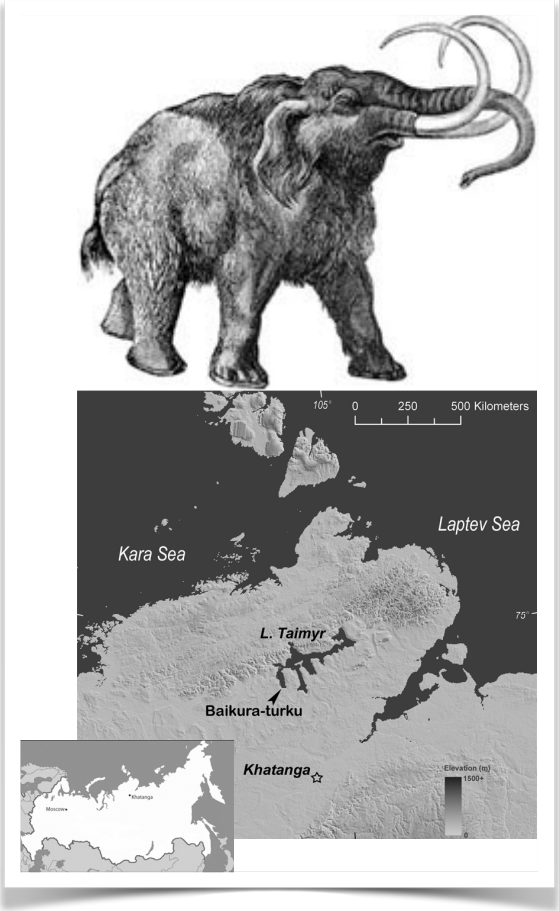
# Mammoth DNA & metagenome (2006)

- DNA collected from permafrost mammoth (28,000 years old)
- DNA extracted from 1g bone
- DNA sheared to 500-700 bp

- Sequenced using *454*
- ~302,000 reads, length ~95 bp

★ Can use NGS for ancient DNA

★ First NGS metagenomics paper




REPORTS

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

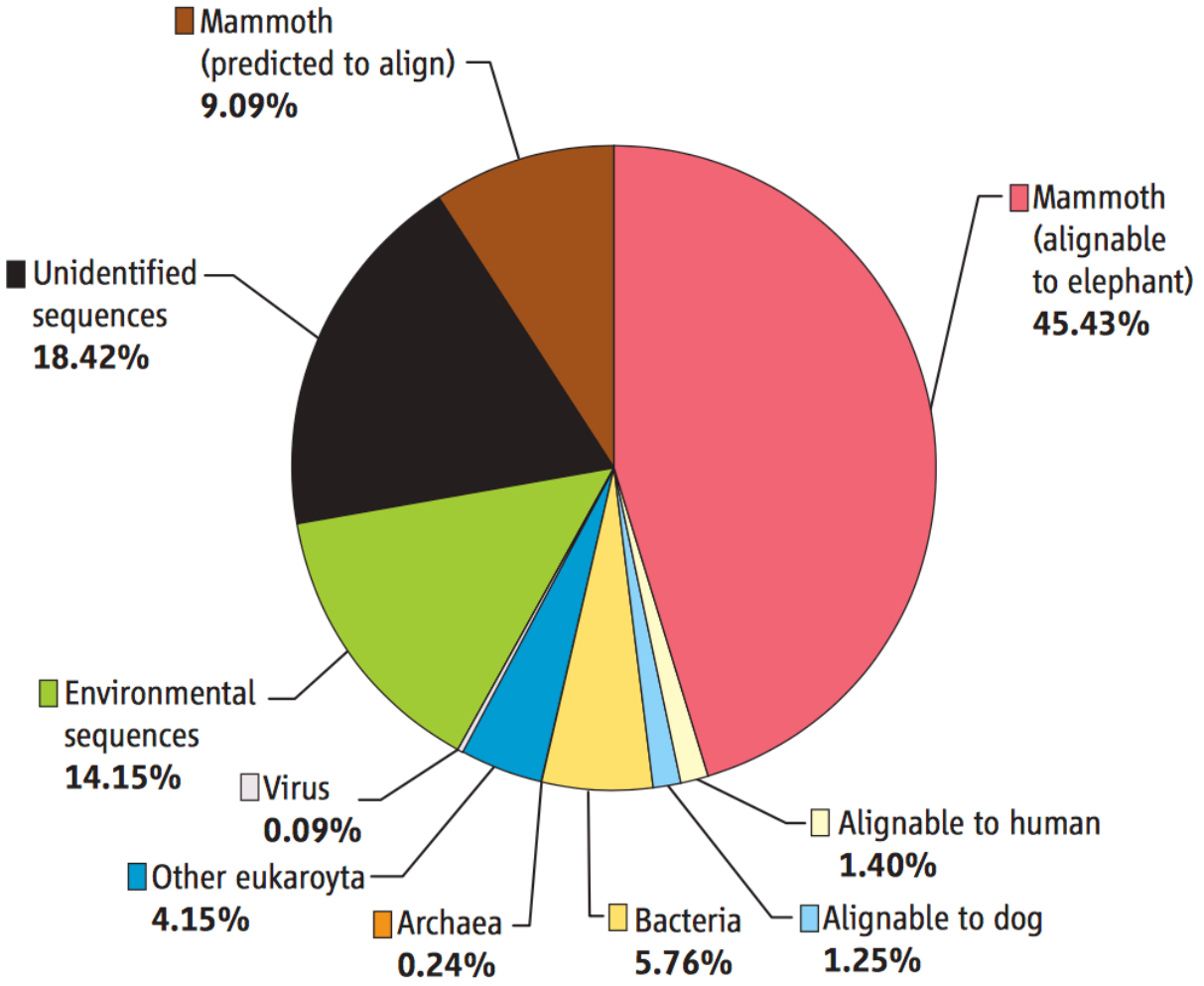
Hendrik N. Poinar,[1,2,3*] Carsten Schwarz,[1,2] Ji Qi,[4] Beth Shapiro,[5] Ross D. E. MacPhee,[6] Bernard Buigues,[7] Alexei Tikhonov,[8] Daniel H. Huson,[9] Lynn P. Tomsho,[4] Alexander Auch,[9] Markus Rampp,[10] Webb Miller,[4] Stephan C. Schuster[4*]

Science, 2006

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Mammoth bone metagenome (2006)

**Fig. 1.** Characterization of the mammoth metagenomic library, including percentage of read distributions to various taxa. Host organism prediction based on BLASTZ comparison against GenBank and environmental sequences database.



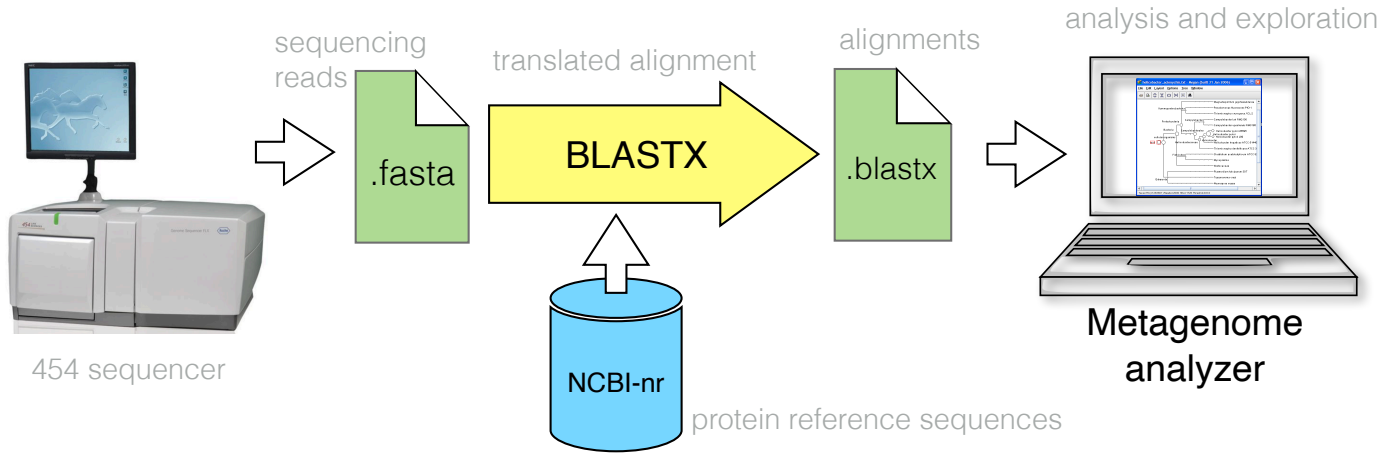

Poinar et al, Science 2006

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# How to analyze metagenomic reads? (2006)

Basic idea (with Stephan Schuster at Penn State):

- BLASTX non-host reads against NCBI-nr
- Assign reads to NCBI taxonomy using naive LCA (lowest common ancestor) approach
- Develop GUI to explore assignments and alignments

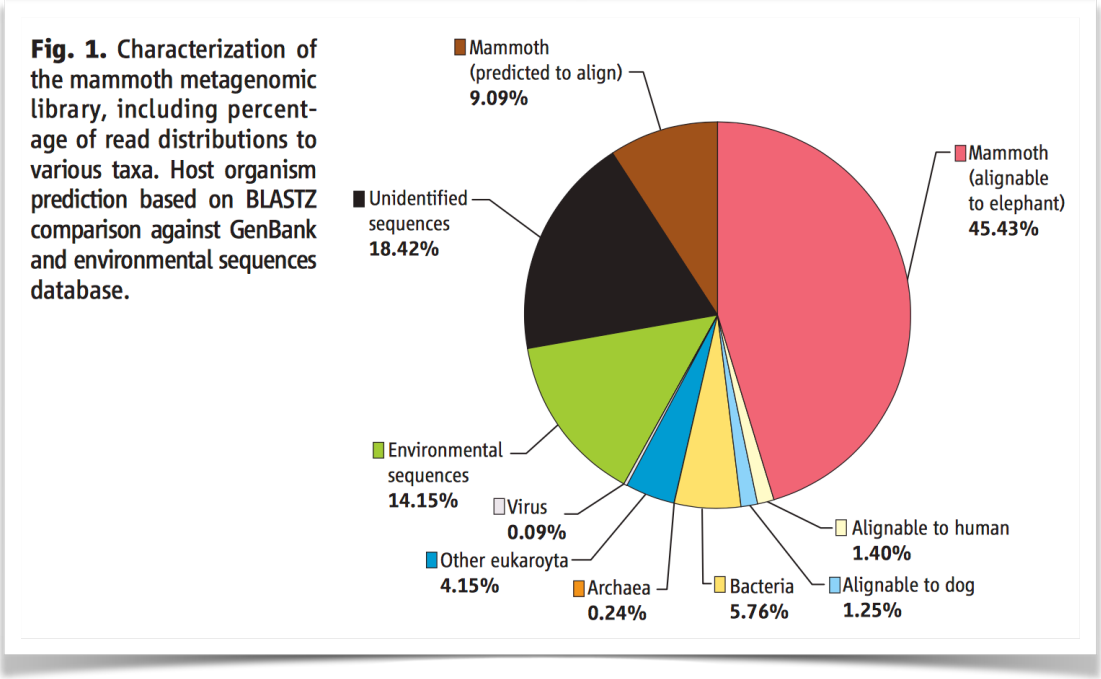



2006 MEGAN analysis pipeline

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

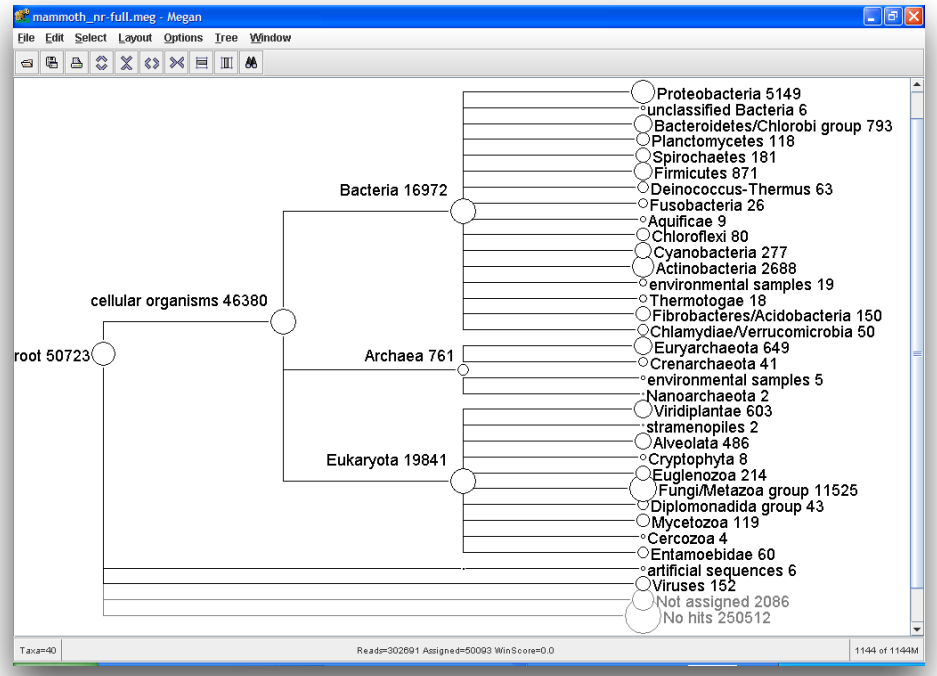
# How to analyze metagenomic reads? (2006)

- MEGAN (MEtagenome ANalyzer 1.0)




Poinar et al, Science 2006

MEGAN 1.0




H. et al, Genome Research, 2007

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Computational bottleneck (2006)

- Compare all reads against the NCBI-nr protein database

- Year 2006:
  - 300,000 reads of length ~100bp
  - NCBI-nr: 3 million entries, ~1 billion letters

★ BLASTX took a couple of weeks on a small cluster

(NCBI-nr today: ~ 550 million entries)

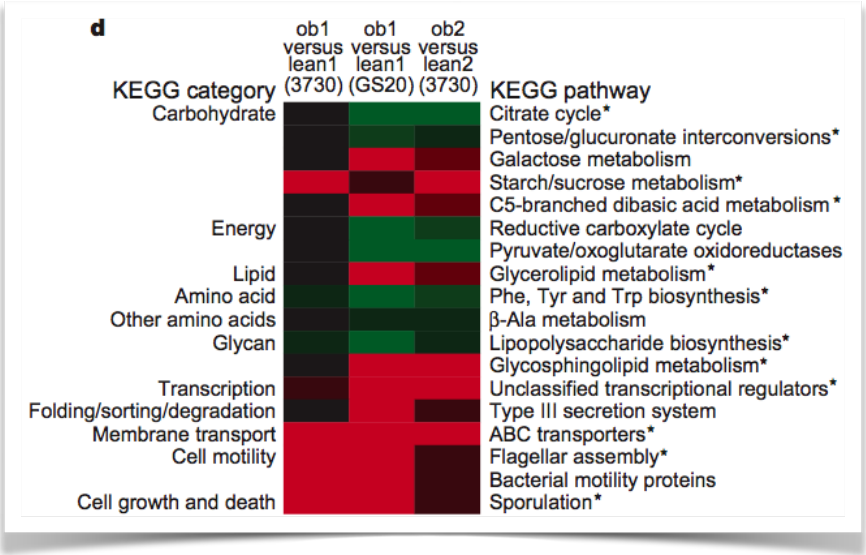
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Obesity-associated gut microbiome

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


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



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Large-scale human gut analysis

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

MetaHIT 2010

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.

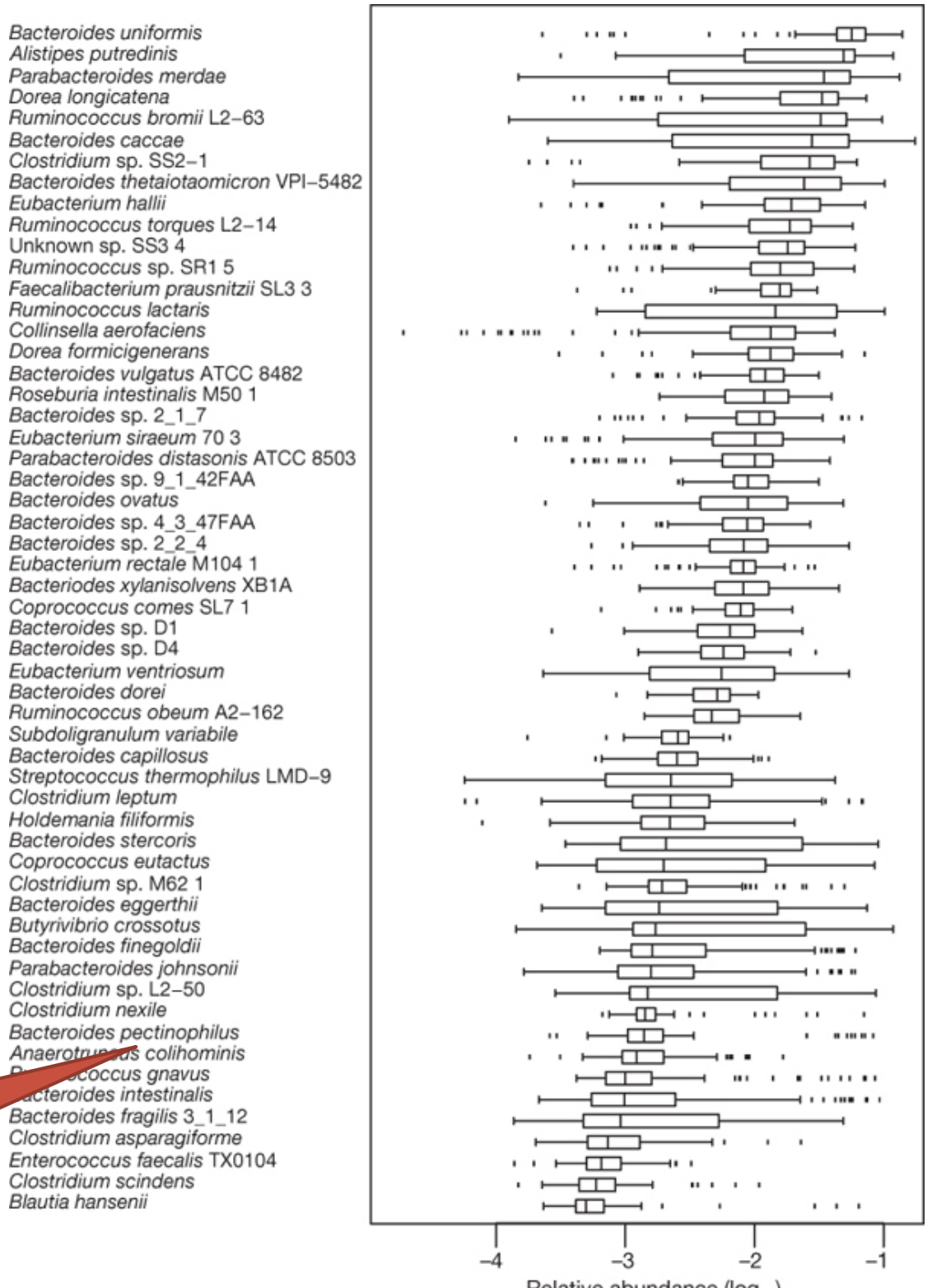
- 576Gb of sequence from 124 individuals

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Core of human gut microbiome

- 57 species present in ≥90% of individuals with coverage >1%
- High variability
- Bacteroidetes and Firmicutes most abundant

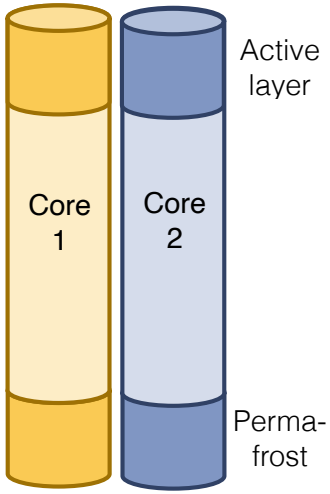
BLASTX at Super Computer Center in Barcelona, then MEGAN analysis



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

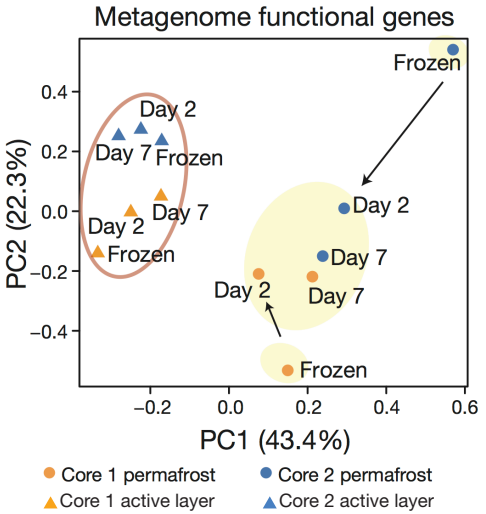
# Permafrost study (2011)

(Mackelprang *et al*, Science 2011)




Frozen, day 2, day 7

Their question:
Functional changes
during thawing?



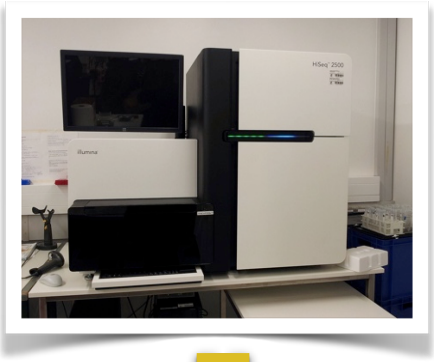

- Align ~250 million Illumina reads against KEGG
- 800,000 CPU hours at Super Computer Center in Berkeley


on 100 cores

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Three basic questions


Hundreds of Samples

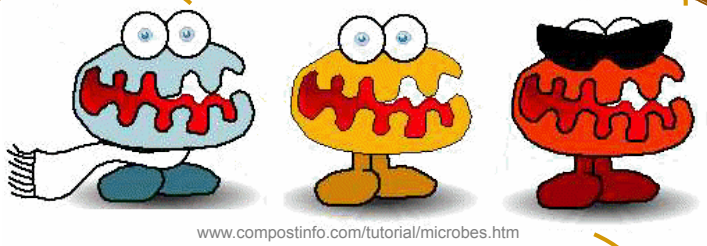


High-throughput
DNA sequencing

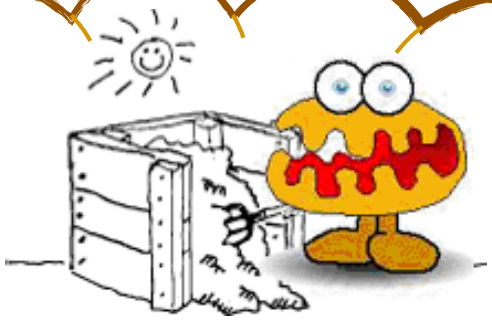
Billions of sequences

Basic computational
analysis

Many
CPU hours



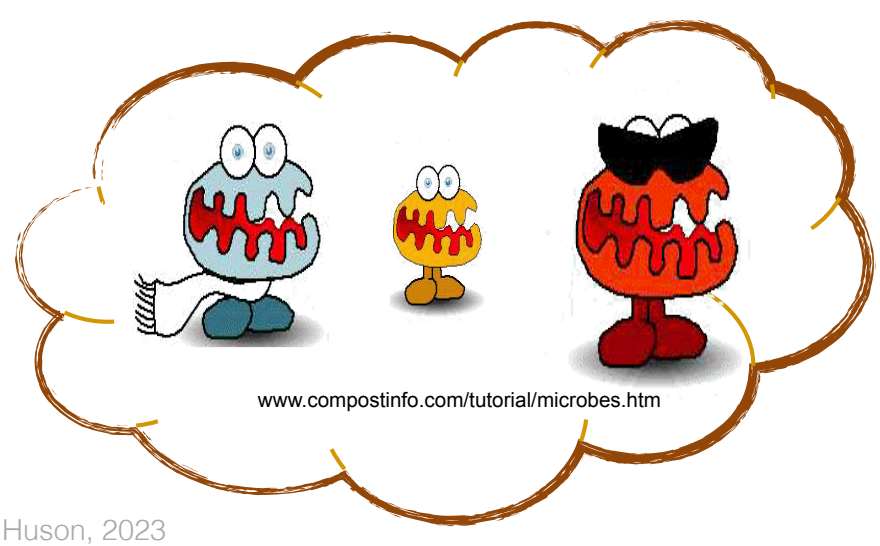

Q1: Who is out there?

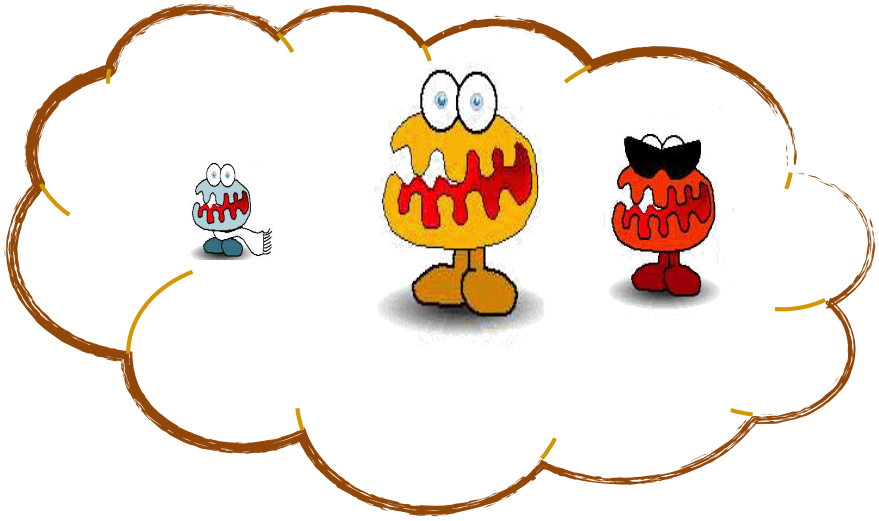


Q2: What are they doing?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Q3: How do they compare?

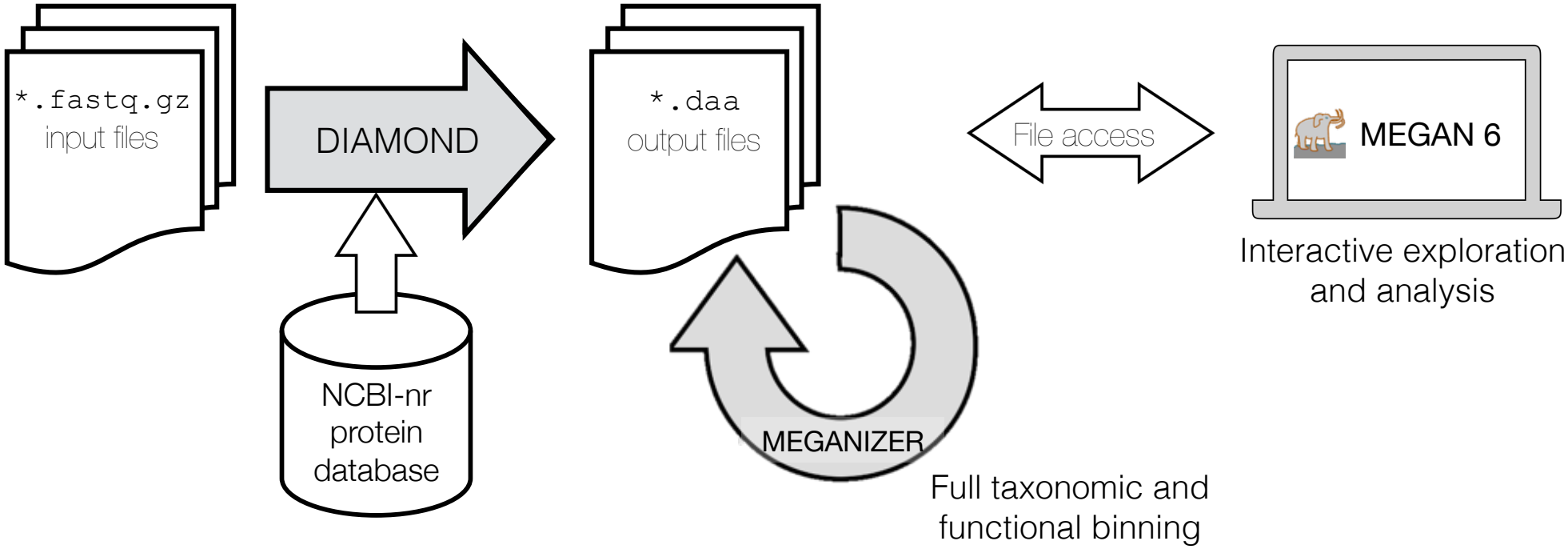
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Three-part approach

- Part 1: Align reads or assembled contigs against protein reference sequences - DIAMOND
- Part 2: Analyze alignments to assign sequences to taxonomic and functional classes - MEGANIZER
- Part 3: Interactively explore, analyze and compare samples - MEGAN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# DIAMOND+MEGAN pipeline




Server

Desktop/laptop

# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part 0 - Software setup

- Tutorial webpage, available from:

  https://github.com/husonlab/tutorials/wiki/Tutorial

You should have the following files and software installed:

- DIAMOND, from:

  https://github.com/bbuchfink/diamond

- MEGAN, from:

  https://software-ab.cs.uni-tuebingen.de/download/megan6

- Data files, tutorial reference and mapping files, from:

  https://software-ab.cs.uni-tuebingen.de/download/megan6/tutorial/tutorial.zip

# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

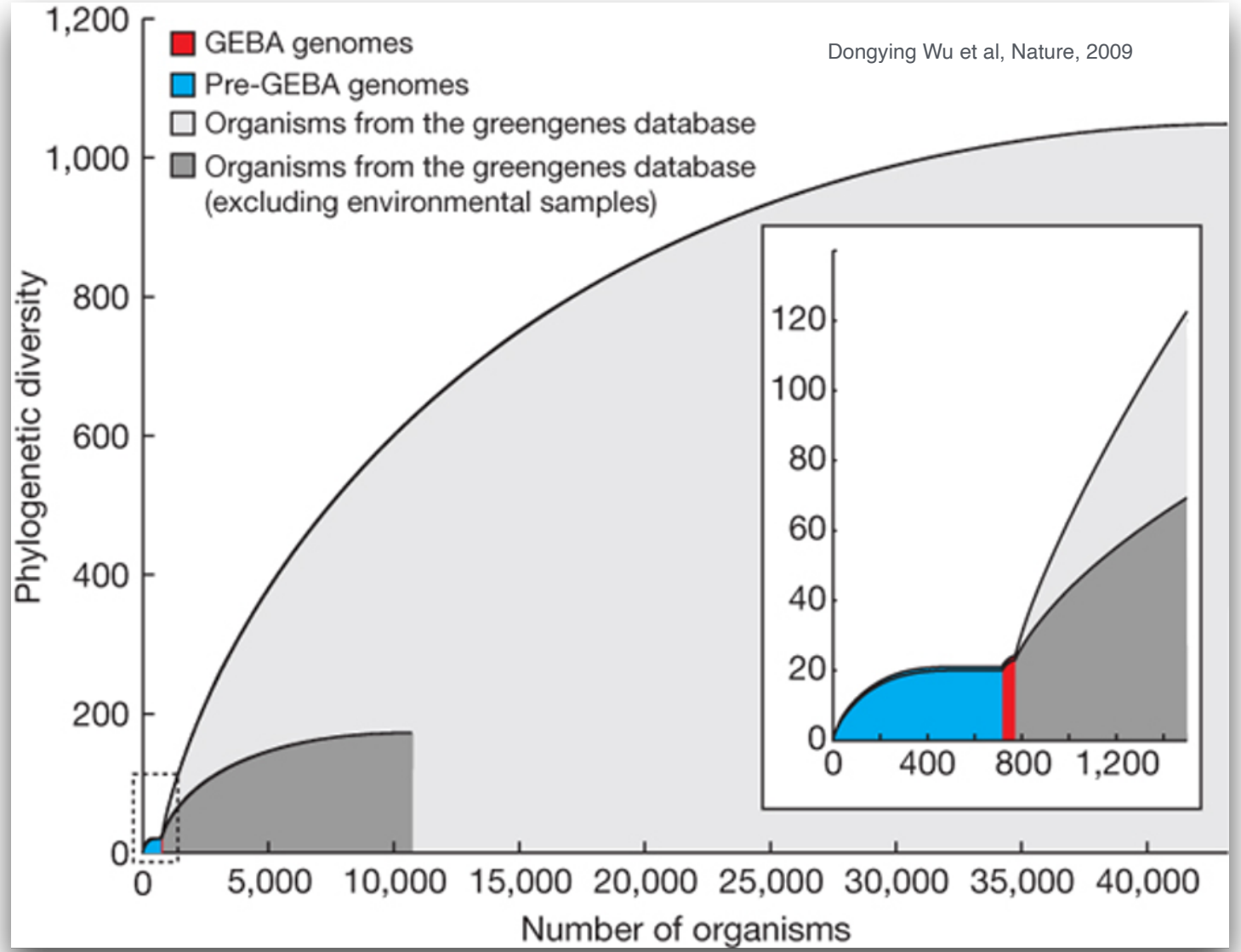
# Part I - Protein alignment

- Key idea: To perform metagenome analysis, align against a protein reference database

Why protein alignment?

- To *identify known* genomes in a sample, use DNA alignment (e.g., pathogen detection, human gut)
- To analyze *unknown* organisms, protein alignment is more suitable due to higher level of conservation

# Genome databases don't cover enough diversity



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Translated alignment

- Read:

  >HISEQ:457:C5366ACXX:2:1101:5937:60460 (101 bases)
  **TTATATTAATTAGAAAACCAATTAAAAATACGAACGTTATGAAGAAGTACATTTGC**...

- Translation (frame +3):

  **..I L I R K P I K N T N V M K K Y I C** ...

- Translated alignment:

```
>EEC52678.1 Length = 65

 Score = 56 bits (135), Expect = 1e-05
 Identities = 22/33 (67%), Positives = 27/33 (82%), Gaps = 0/33 (0%)
 Frame = +3

Query:          3  ILIRKPIKNTNVMKKYICTVCEYIYDPEQGDPE  101
                   +L +K  K   VM+KYICT+CEY+YDPEQGDPE
Sbjct:          1  MLSKKKFKQKRVMEKYICTICEYVYDPEQGDPE  33
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Comprehensive database: NCBI-nr

- NCBI-nr database of non-redundant protein sequences has over 500M entries

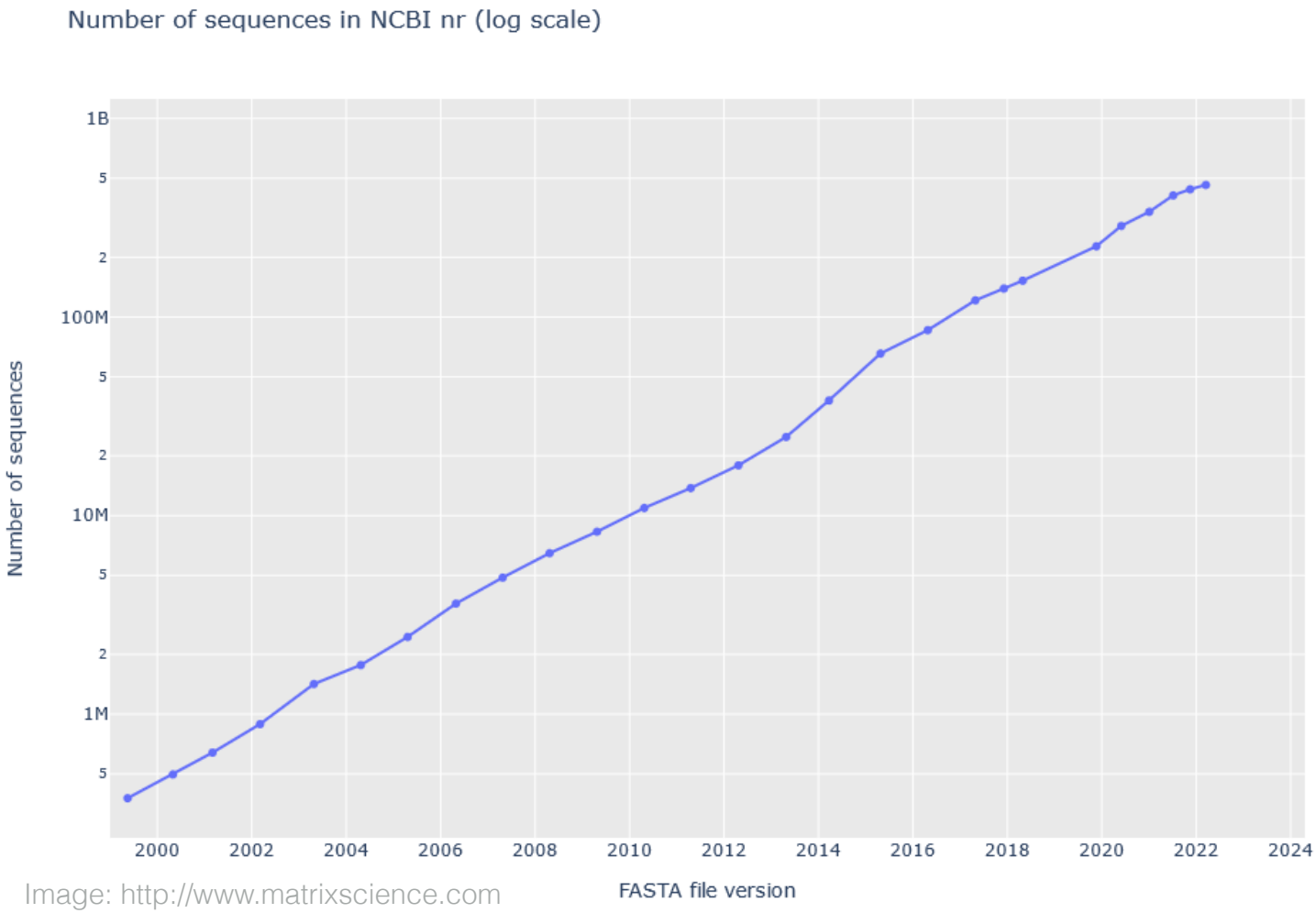
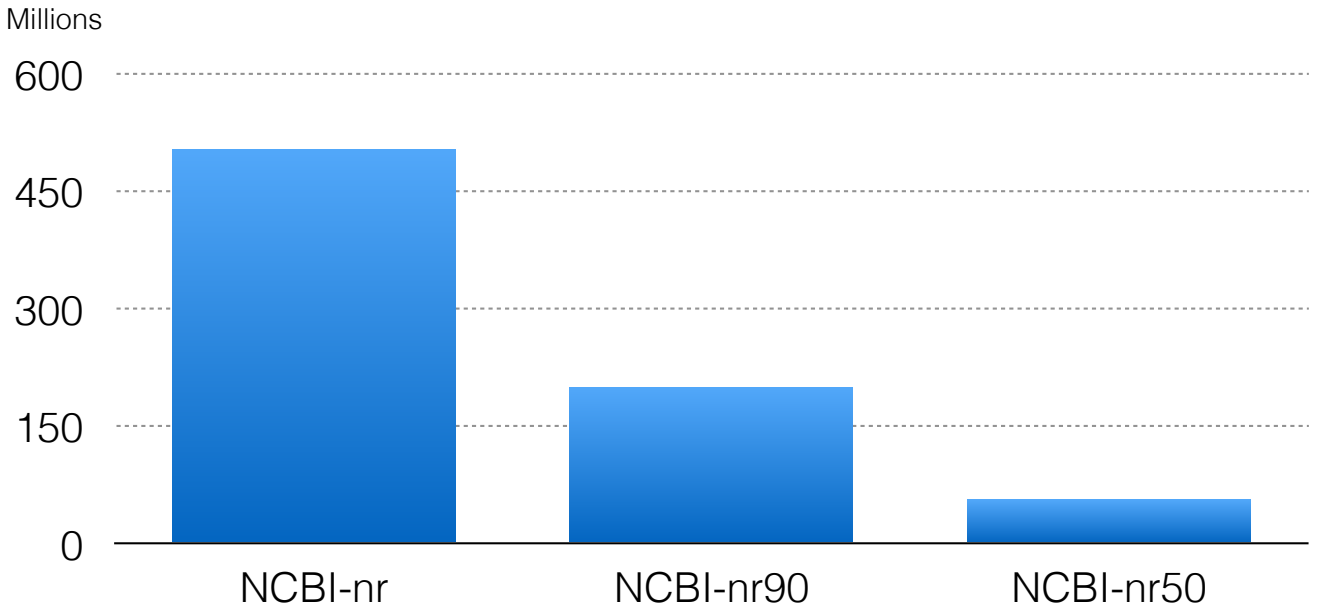



Image: http://www.matrixscience.com

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Smaller alternative?

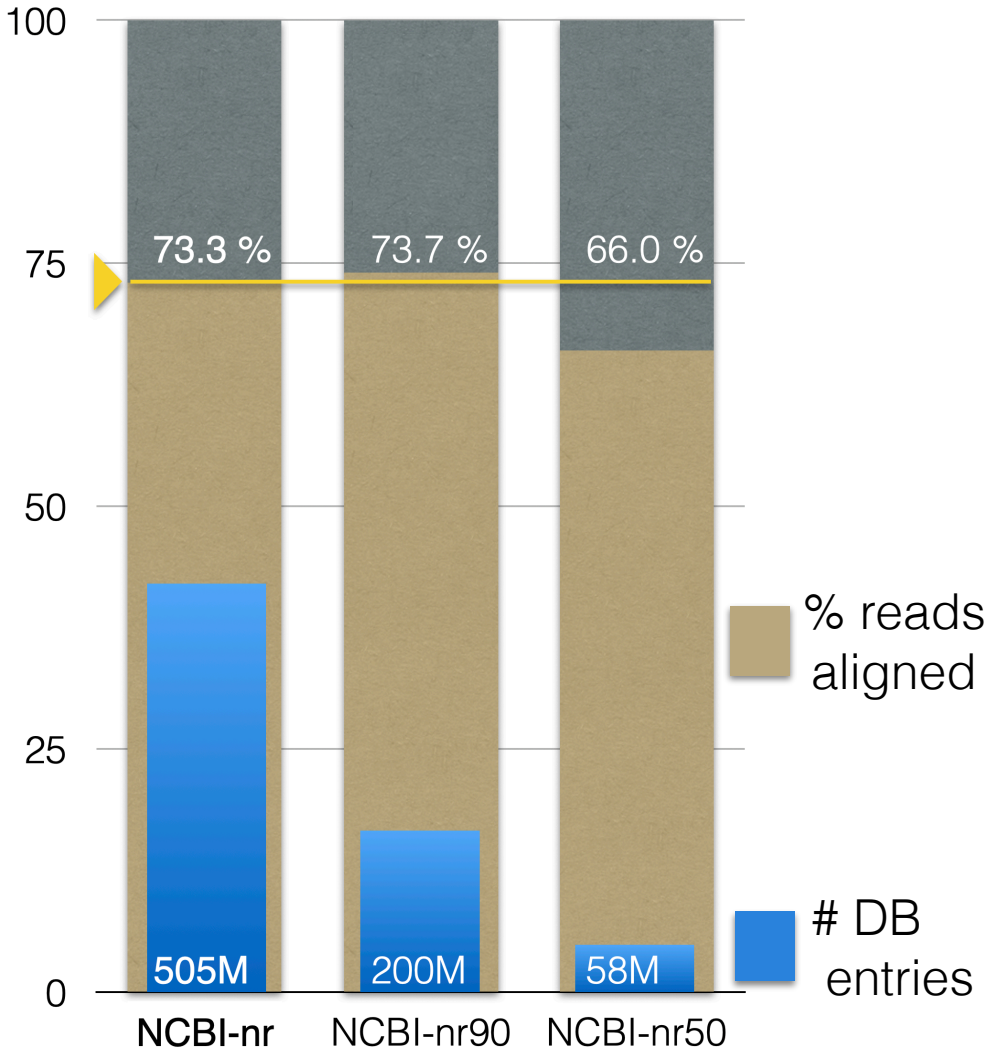
- Latest release of DIAMOND allows clustering of NCBI-nr (Buchfink et al, submitted)
- Similarity: 90 and 50%
- NCBI-nr90 and NCBI-nr50



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Comparison NCBI-nr vs nr50

- 201M reads aligned by DIAMOND & analyzed using MEGAN




| Classification | NCBI-nr | NCBI-90 | NCBI-50 |
|---|---|---|---|
| NCBI Taxonomy | 1.0 | 1.01 | 0.85 |
| GTDB Taxonomy | 1.0 | 1.03 | 0.88 |
| INTERPRO | 1.0 | 1.34 | 1.39 |
| SEED | 1.0 | 1.07 | 1.23 |
| EC | 1.0 | 1.12 | 1.17 |
| EGGNOG | 1.0 | 1.55 | 1.87 |
| Speed-up | 1.0 | 3.0 | 34.6 |

# Tutorial-nr file

- While NCBI-nr50 is small enough to be used on a high-end laptop, it is too big for this tutorial
- For the tutorial, we provide `tutorial-nr.gz`
  - This is a small subset of nr50.gz
  - It only contains accessions relevant for the provided short-read datasets
  - It can not be used for analysis of real data

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part I - Protein alignment

- Will use DIAMOND
  - designed for metagenomics (Buchfink et al, 2015)

- Need to:
  - Install DIAMOND
  - Download reference sequence
  - Run DIAMOND to build index
  - Run DIAMOND on fastq.gz files

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part I - DIAMOND index

- Build a DIAMOND index:

```
diamond makedb --in tutorial-nr.gz -d tutorial-nr
```

- Note: Using `tutorial-nr.gz`, due to time restrictions

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part 1 - Run DIAMOND

- Run DIAMOND on one input FASTQ file:

```
diamond blastx -d tutorial-nr \
-q data/Alice00-1mio.fa.gz \
-o out/Alice00-1mio.daa \
-f 100 --masking 0
```

- Run DIAMOND on *all* input files in the directory:

```
for file in data/*.fa.gz
do
ofile="out/$(basename "${file%.*}").daa"
diamond blastx --db tutorial-nr \
-q $file -o $ofile -f 100 --masking 0
done
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part 1 - Run DIAMOND

- For full size datasets, DIAMOND alignment (and subsequent meganization) is run on a server

- The 12 small datasets against `tutorial-nr.gz` will take less than 20 minutes on a modern laptop

- If you failed to run DIAMOND on the data, you can download the resulting files here:

`https://software-ab.cs.uni-tuebingen.de/download/megan6/tutorial/diamond-out.zip`

# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

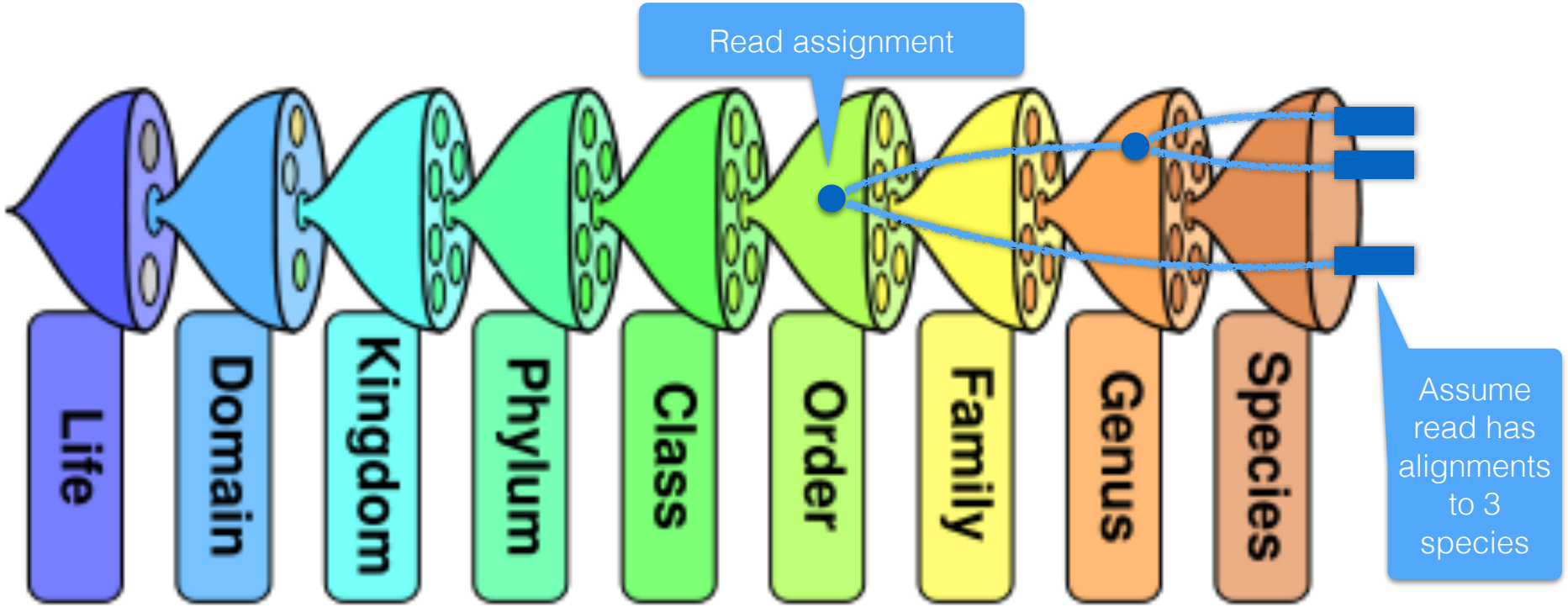
# Part 2: Meganization

- Run DIAMOND with option -f 100, so that
  - the output is a “DAA” file, a binary file containing all aligned sequences and reported alignments.
- Then run tool daa-meganizer (or MEGAN)
  - to “meganize” the DAA file; performing taxonomic and functional analysis of all aligned sequences, and
  - the result of meganization is appended to the DAA file; no new file is created.
- A meganized DAA file can be opened in MEGAN.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Taxonomy meganization

- Taxonomic binning uses
  - the NCBI taxonomy (Benson et al, 2005),
  - the GTDB taxonomy (Parks et al, 2018),
  - naive LCA algorithm for short reads (Huson et al, 2007),
  - interval-union LCA for long reads (Huson et al, 2018).



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Function meganization

- Functional binning uses e.g.
  - EggNOG (Powell et al, NAR 2014)
  - InterPro (Mitchell et al, NAR 2015)
  - SEED (Overbeek et al, NAR 2014)
  - KEGG (MEGAN UE only, Kanehisa & Goto, NAR 2000)
- Assignment uses the top-hit strategy (Huson, 2011)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Meganization database

- A DAA file contains reference sequences and their accessions
- Meganization requires a mapping of accessions to taxonomic and functional classes
- Provided as a "MEGAN mapping database" `megan-map-tutorial.db`
- Here is the SQLITE schema:

```
CREATE TABLE mappings (Accession PRIMARY KEY, Taxonomy INT, GTDB
INT, EGGNOG INT, INTERPRO2GO INT, SEED INT, EC INT);
```

- A typical entry:

```
EKP93748|867903||253||22932|501010007
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part 2: Meganization

- Meganize one DIAMOND file:

```
~/megan/tools/daa-meganizer \
-i out/Alice00-1mio.daa   \
-mdb megan-map-tutorial.db
```

- Meganize *all* DIAMOND files:

```
~megan/tools/daa-meganizer \
-i out/*.daa -mdb megan-map-tutorial.db
```

# Part 2: Meganization

- If you failed to meganize the 12 files, you can download the meganized files here:

```
https://software-ab.cs.uni-tuebingen.de/download/
megan6/tutorial/meganizer-out.zip
```

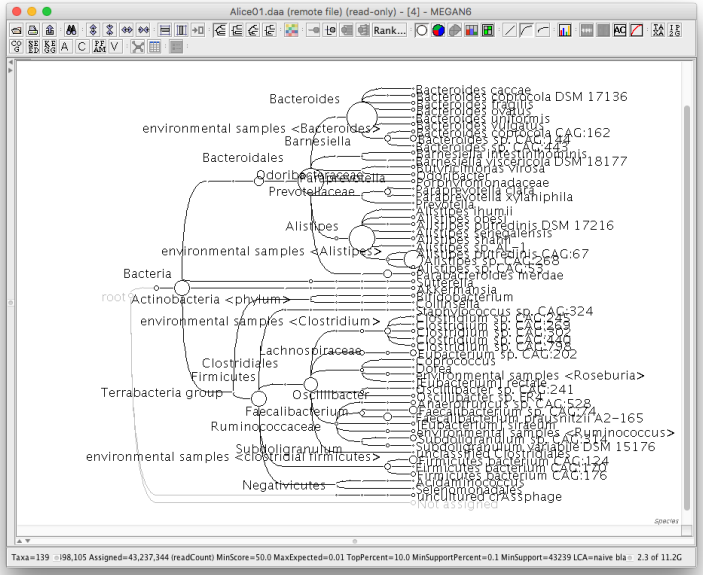
# Outline

- Introduction to microbiome analysis
- Part 0: Software setup
- Part 1: DIAMOND alignment against protein database
- Part 2: Meganization of reads and alignments
- Part 3: Interactive exploration using MEGAN

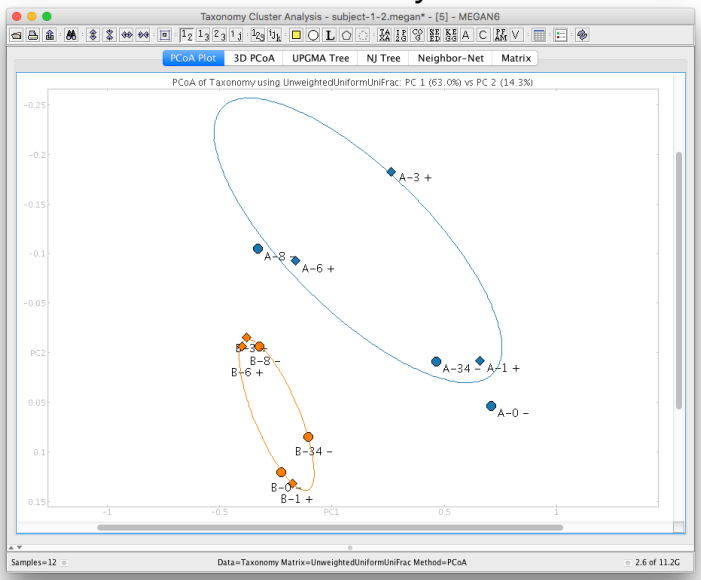
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

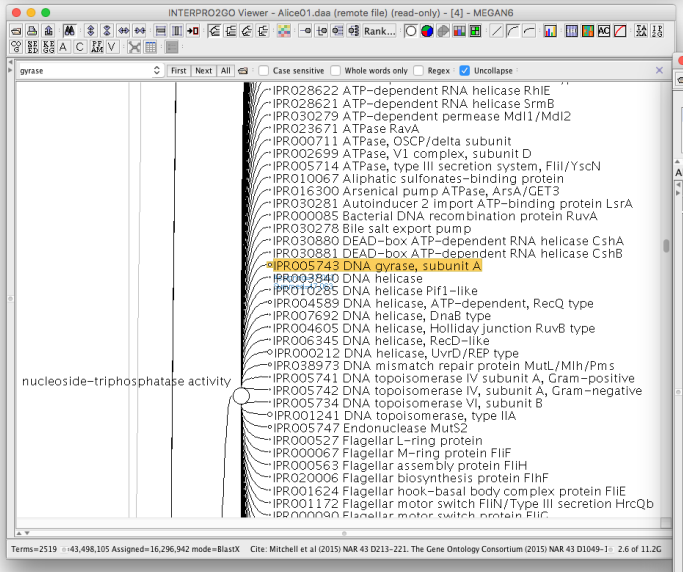
# Interactive MEGAN analysis

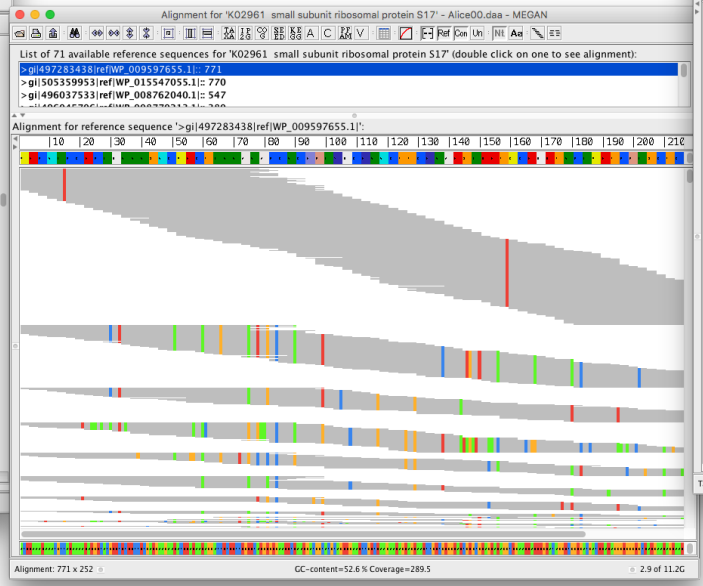
Taxonomic content

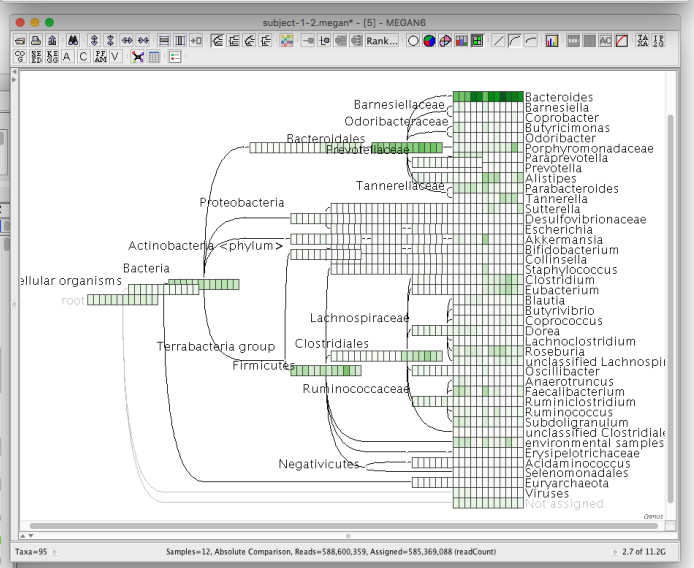



PCoA analysis




Functional content

Gene-centric alignment and assembly

Comparative analysis

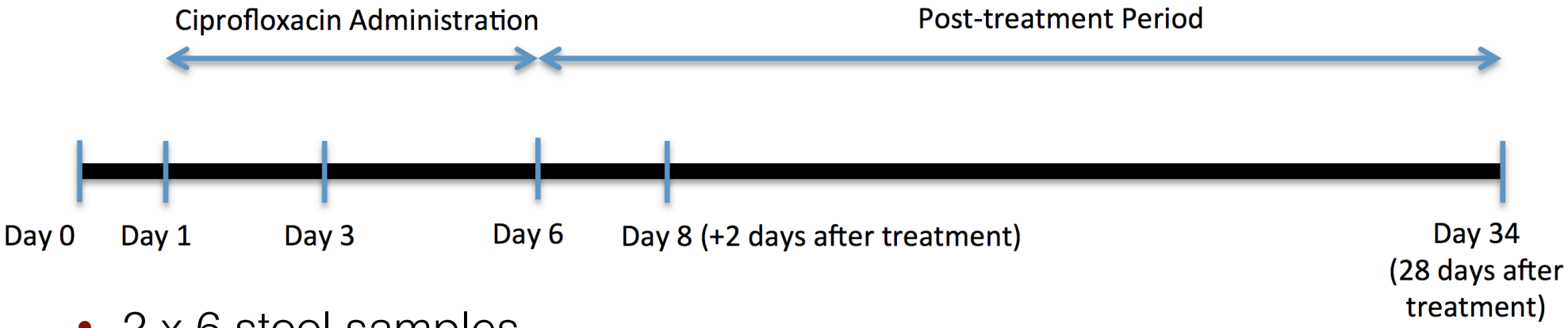
H. et al, MEGAN Community Edition, 2016

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# ASARI- Antibiotic resistance pilot study

UNIVERSITÄTS
KLINIKUM
TÜBINGEN

- Two volunteers, subject 1 and subject 2




- 2 x 6 stool samples
- Shotgun sequencing
  - ~60 million reads per sample (101 bp per read)
  - ~800 million reads in total
- Initial analysis: compare against NCBI-nr protein database

Willmann et al (2015) J. Antimicrobial Agents and Chemotherapy

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Performance of DIAMOND+MEGAN

- 12 human gut samples, total 816 million HiSeq reads

| Sample | Reads | DIAMOND (s) | Alignments | Aligned reads | Meganizer (s) |
|---|---|---|---|---|---|
| Alice 0 | 66 393 401 | 19 062 | 627 405 772 | 44 900 227 | 9 299 |
| Alice 1 | 64 923 975 | 15 771 | 595 715 349 | 43 498 105 | 11 338 |
| Alice 3 | 55 092 349 | 13 435 | 515 249 349 | 37 675 494 | 8 621 |
| Alice 6 | 66 289 376 | 16 801 | 910 892 059 | 52 627 776 | 11 771 |
| Alice 8 | 57 957 661 | 14 134 | 790 946 244 | 45 358 448 | 13 911 |
| Alice 34 | 64 380 386 | 15 615 | 608 114 143 | 44 741 897 | 11 962 |
| Bob 0 | 61 232 588 | 14 573 | 825 213 917 | 48 882 884 | 12 058 |
| Bob 1 | 65 763 766 | 16 203 | 841 038 616 | 51 408 892 | 12 270 |
| Bob 3 | 89 034 641 | 34 598 | 1 233 571 041 | 72 017 720 | 15 789 |
| Bob 6 | 89 339 172 | 27 333 | 1 138 796 522 | 70 344 161 | 15 507 |
| Bob 8 | 78 001 118 | 19 734 | 1 049 831 855 | 63 336 241 | 13 423 |
| Bob 34 | 57 627 119 | 15 406 | 780 844 319 | 45568158 | 11 433 |
| Total | 816 035 552 | 222 665 | 9 917 619 186 | 620 360 003 | Max: 15 789 |
| Time | | ≈ 62 h | | | ≈ 5 h |

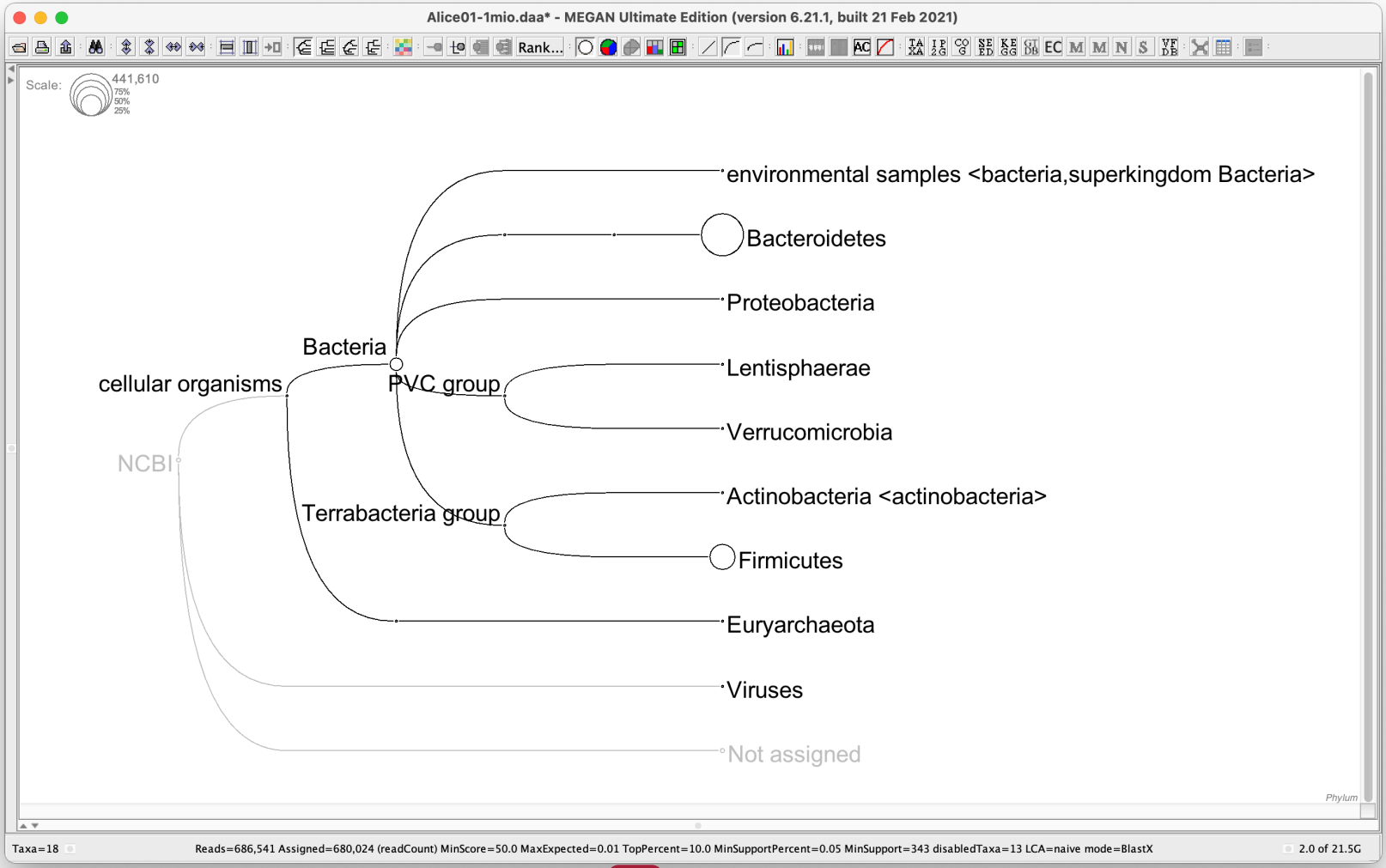
doi:10.1371/journal.pcbi.1004957.t001

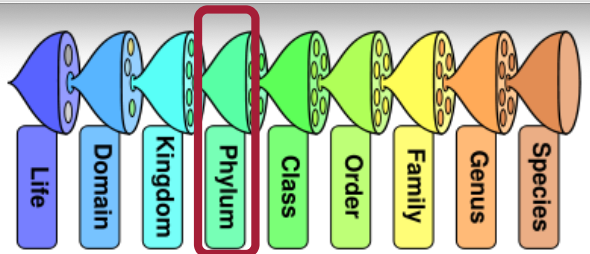
- Complete analysis in 62+5 hours on a single server


H. et al, MEGAN Community Edition, 2016


# Taxonomic content

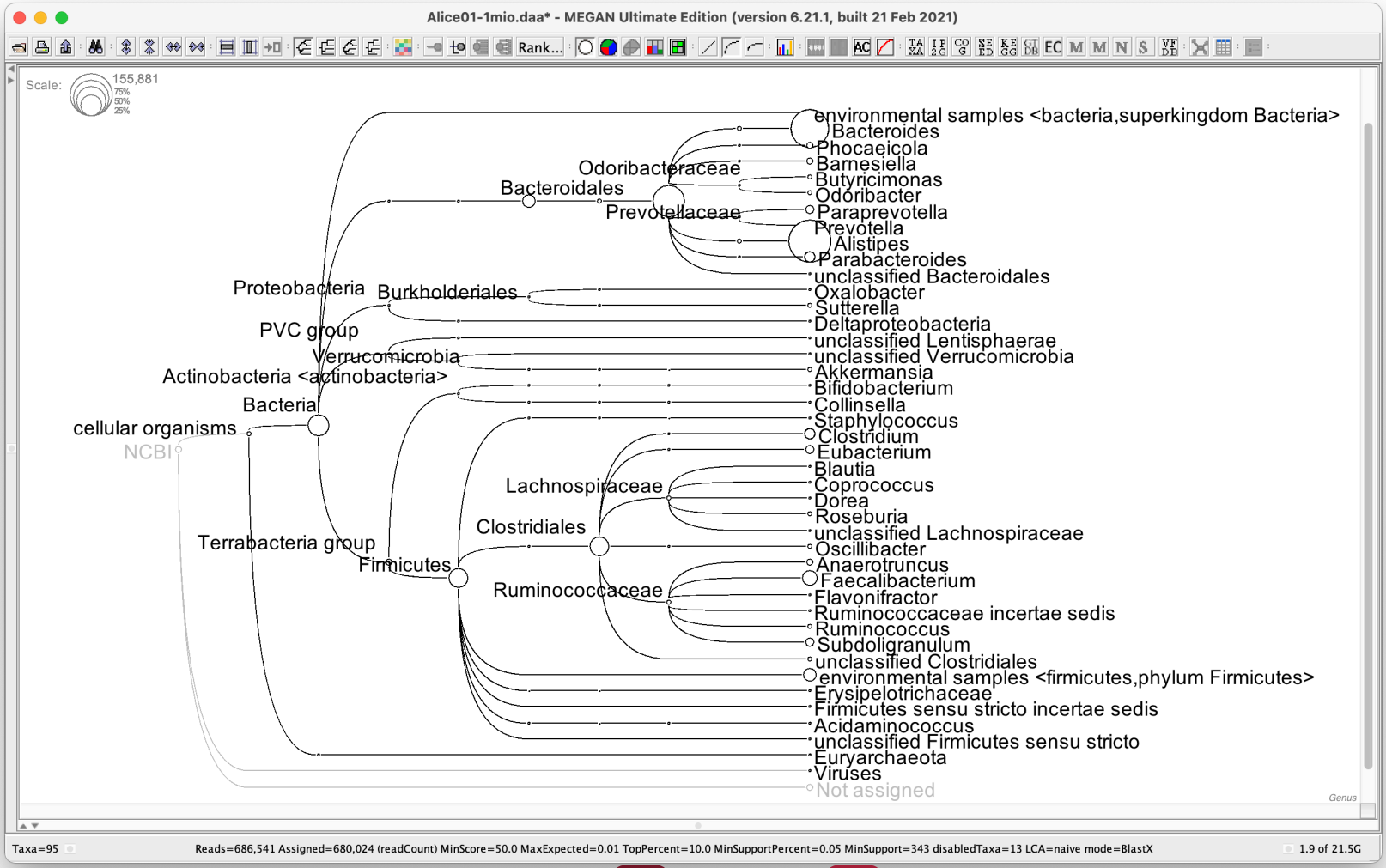
## ASARI human gut microbiome

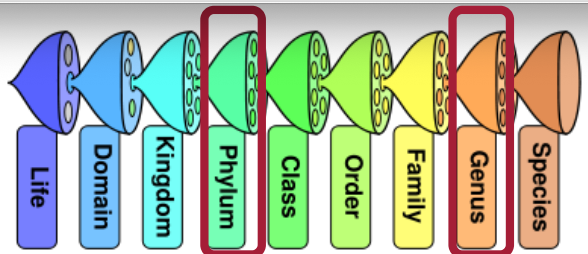


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

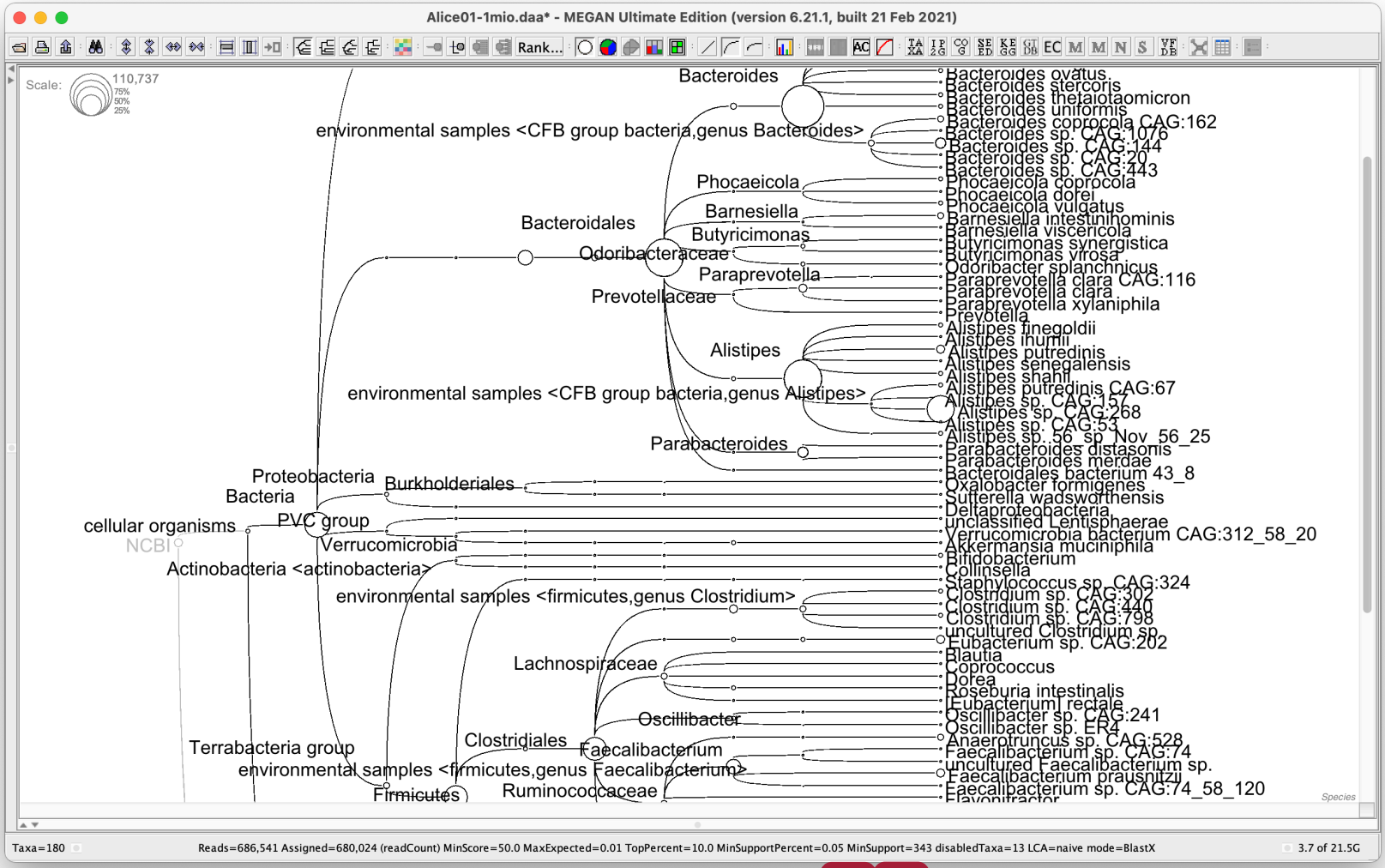
# Taxonomic content

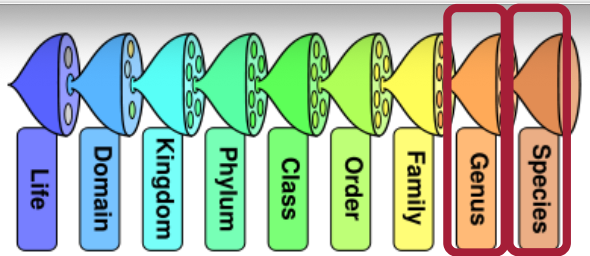


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

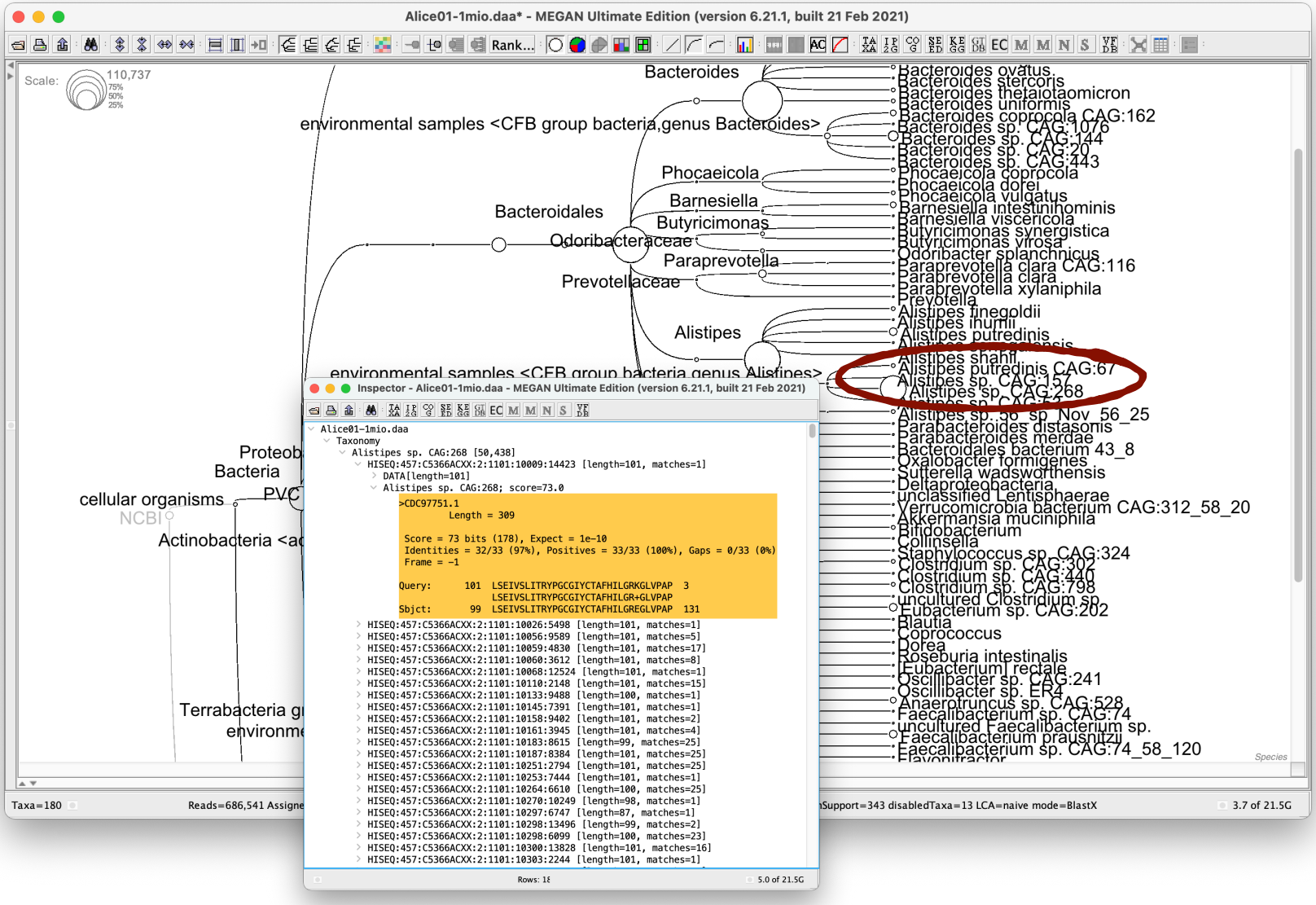
# Taxonomic content



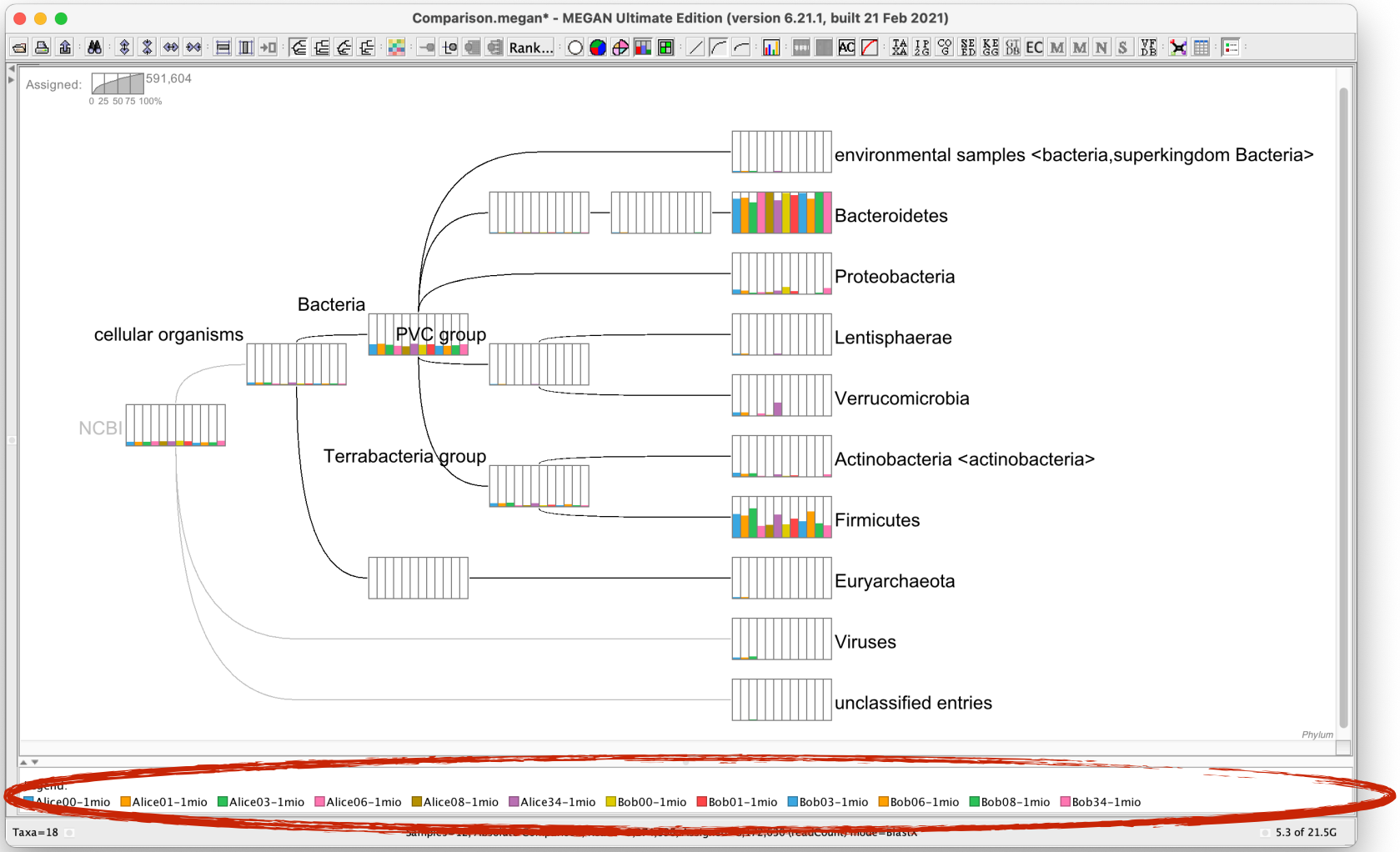
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Drill down to details...

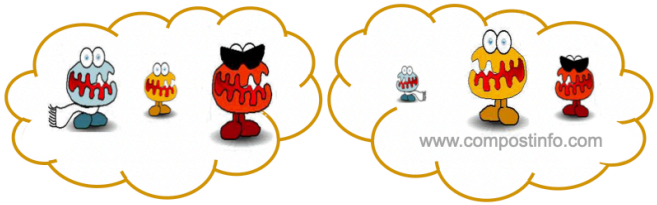



# Comparison

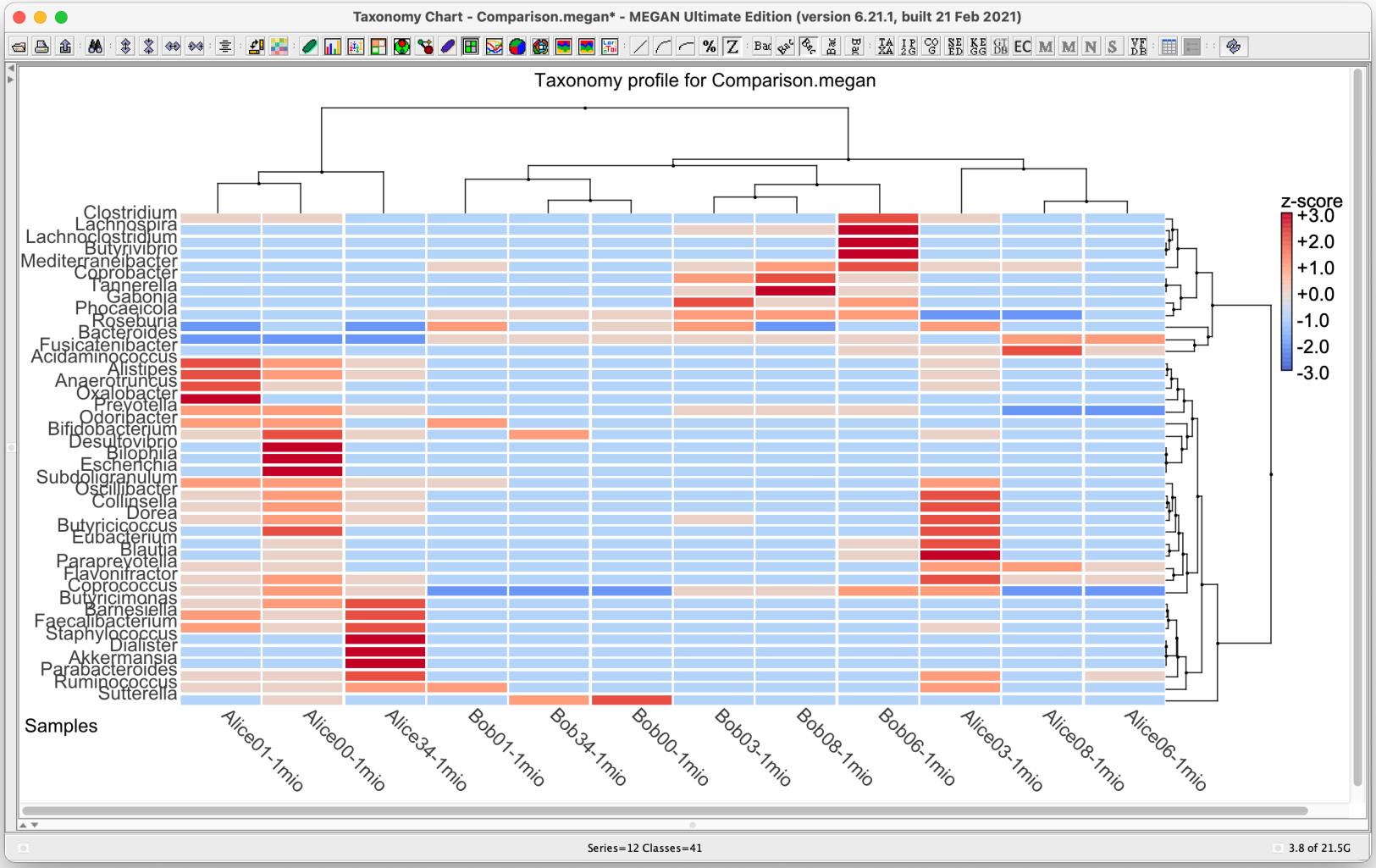
Q3: How do they compare?




All 12 ASARI human gut samples together



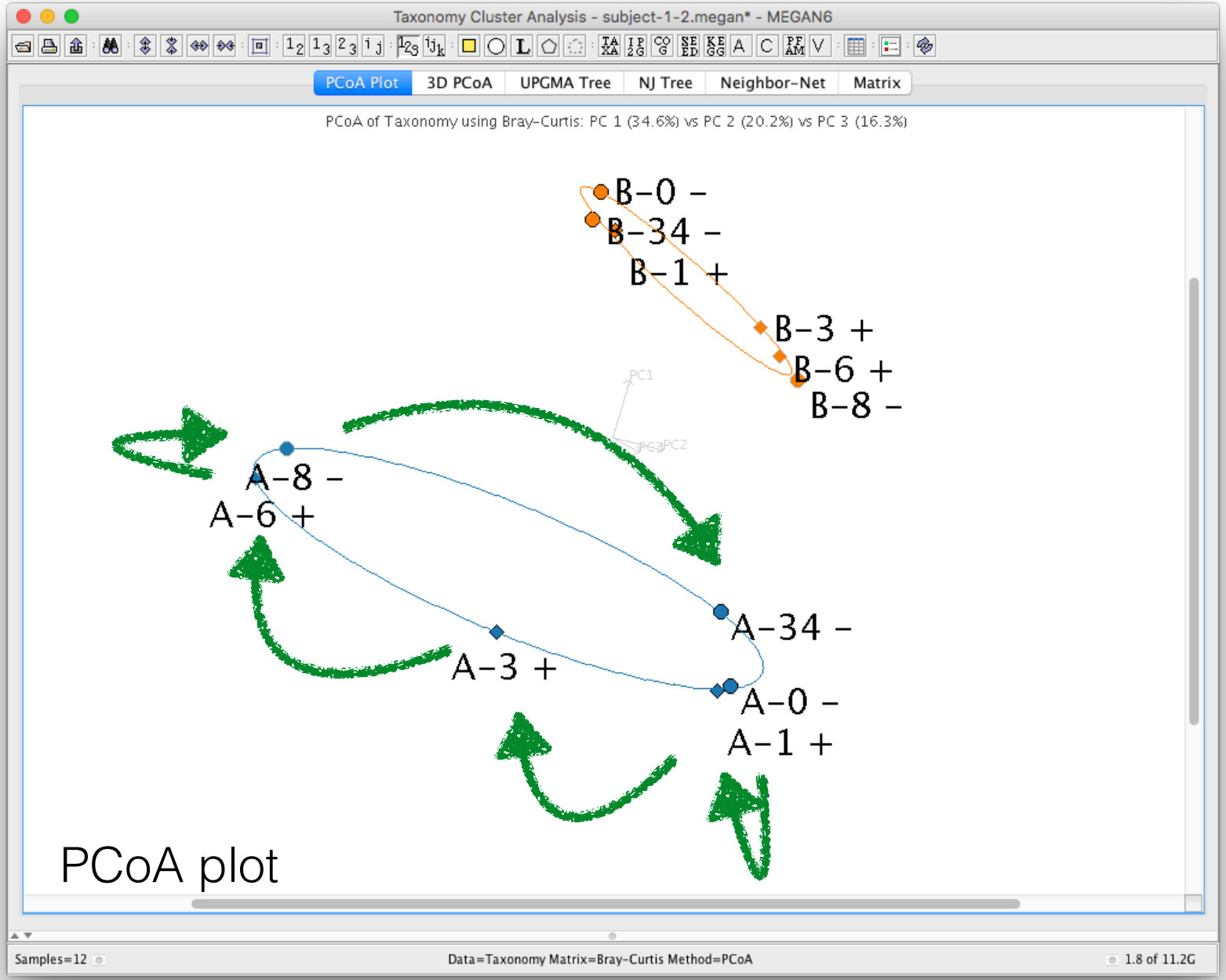

# Comparison




All 12 ASARI human gut samples together

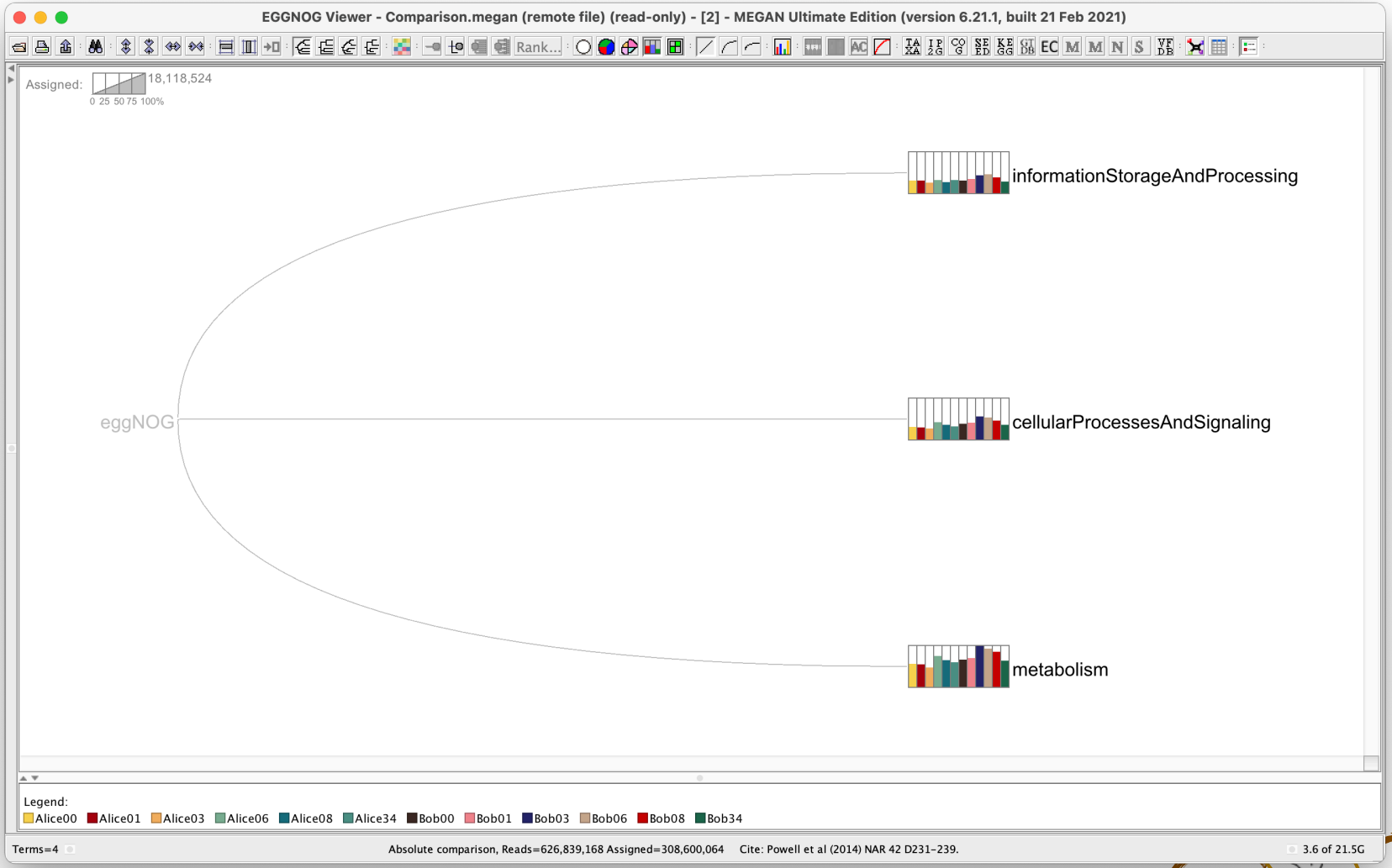
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# E.g.: Does the microbiome rebound?

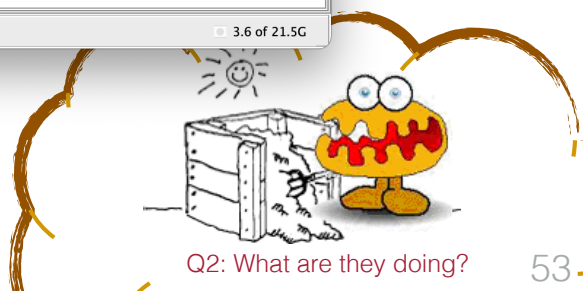


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Functional content



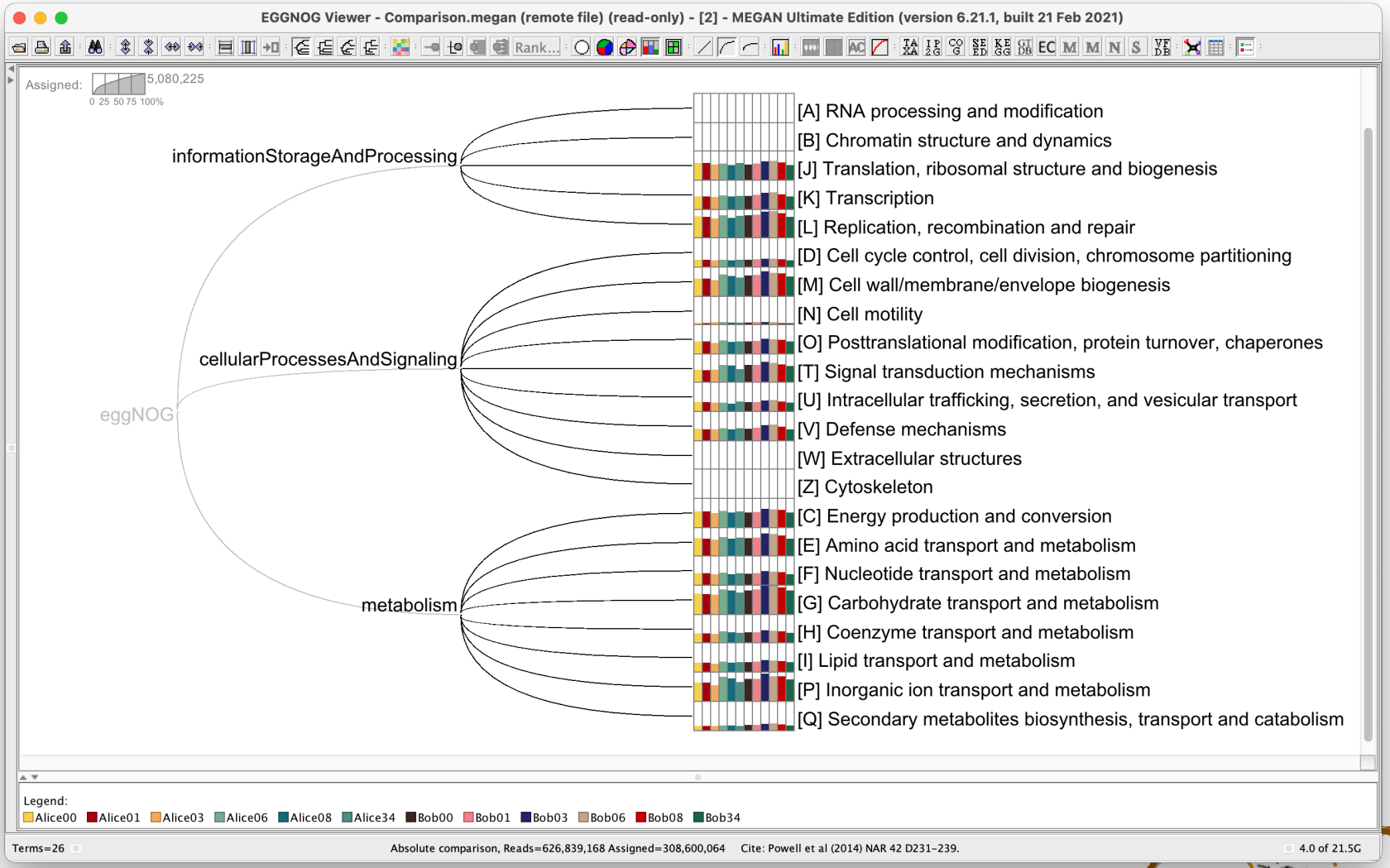

eggNOG classification
(Powell et al, 2014)



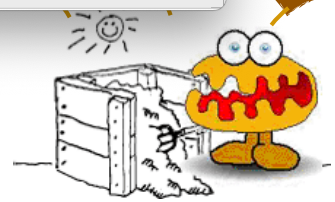
Q2: What are they doing?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Functional content



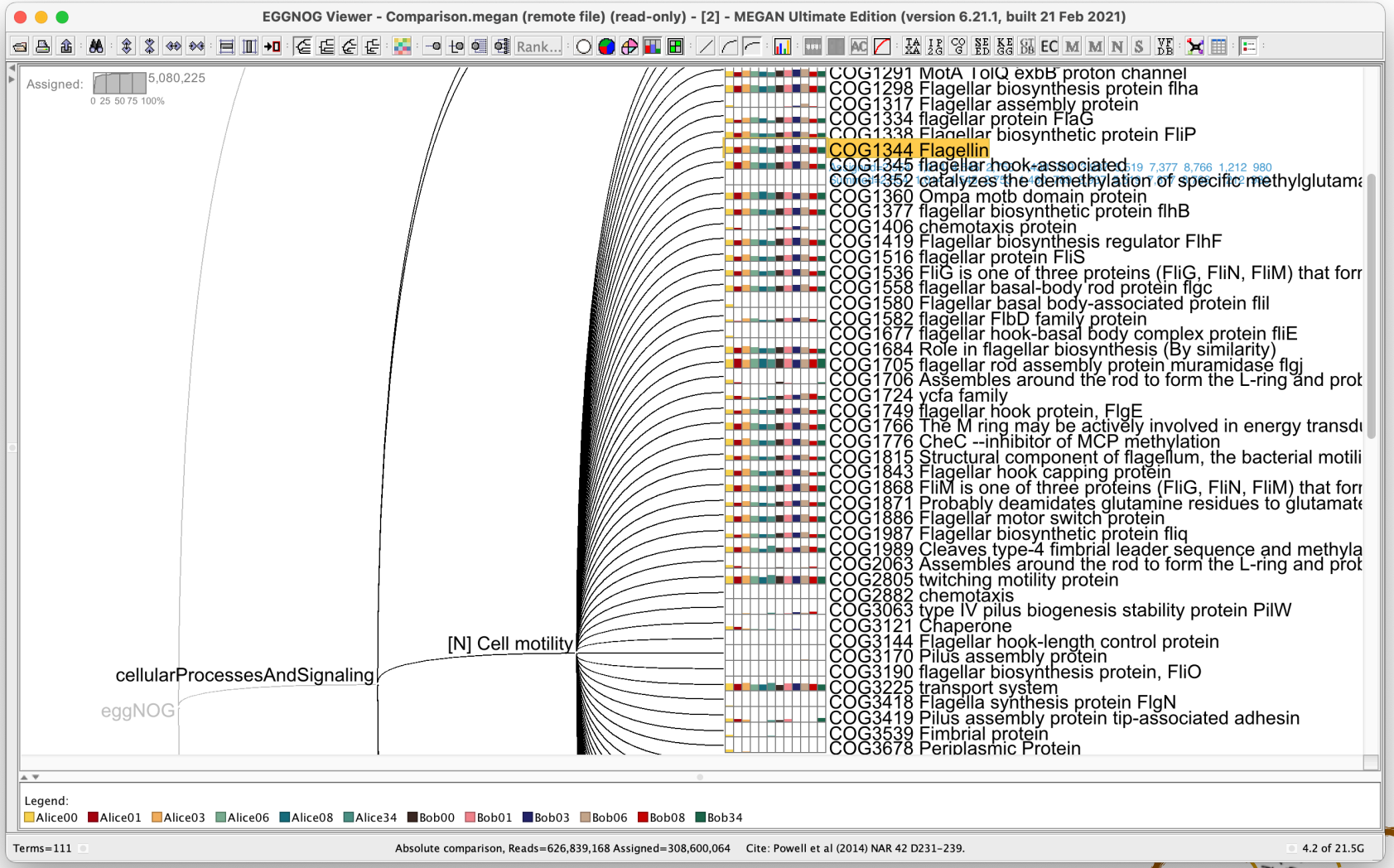

eggNOG classification
(Powell et al, 2014)



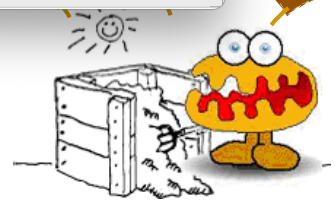
Q2: What are they doing?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Functional content




eggNOG classification
(Powell et al, 2014)



Q2: What are they doing?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Part 3: MEGAN analysis

- Launch MEGAN by typing:

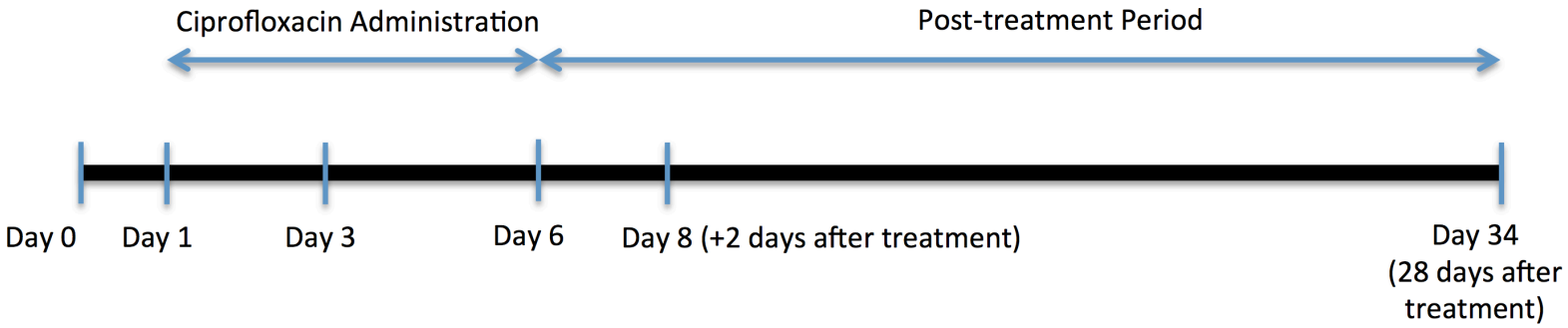
  `~/megan/MEGAN`

- Open individual files with the File->Open... item
- Compare files using the File->Compare... item

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Alice and Bob- short reads

Alice and Bob, 6 time points each




- Each subsampled to 1 mio reads.
- `data/Alice00-1mio.fq.gz` etc

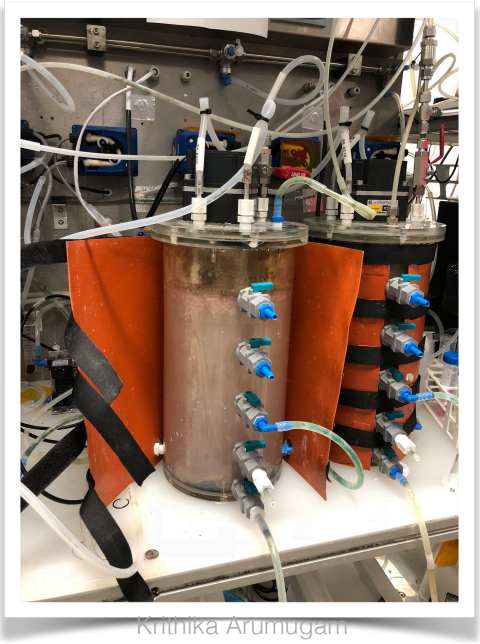
# Alice and Bob- short reads

- Tutorial tasks:
  - confirm that these are gut samples - should be dominated by Bacteroidota, Firmicutes and Proteobacteria.
  - Open all twelve files together in a comparison document and add the provided metadata.
  - Confirm that the taxonomic profiles of either subject changes during the course of antibiotics and then returns to a similar state after treatment.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Enrichment reactor - long reads

- Nanopore reads from enrichment reactor:



Krithika Arumugam

Short report | Open Access | Published: 16 April 2019

**Annotated bacterial chromosomes from frame-shift-corrected long-read metagenomic data**

Krithika Arumugam, Caner Bağcı, Irina Bessarab, Sina Beier, Benjamin Buchfink, Anna Górska, Guanglei Qiu, Daniel H. Huson & Rohan B. H. Williams

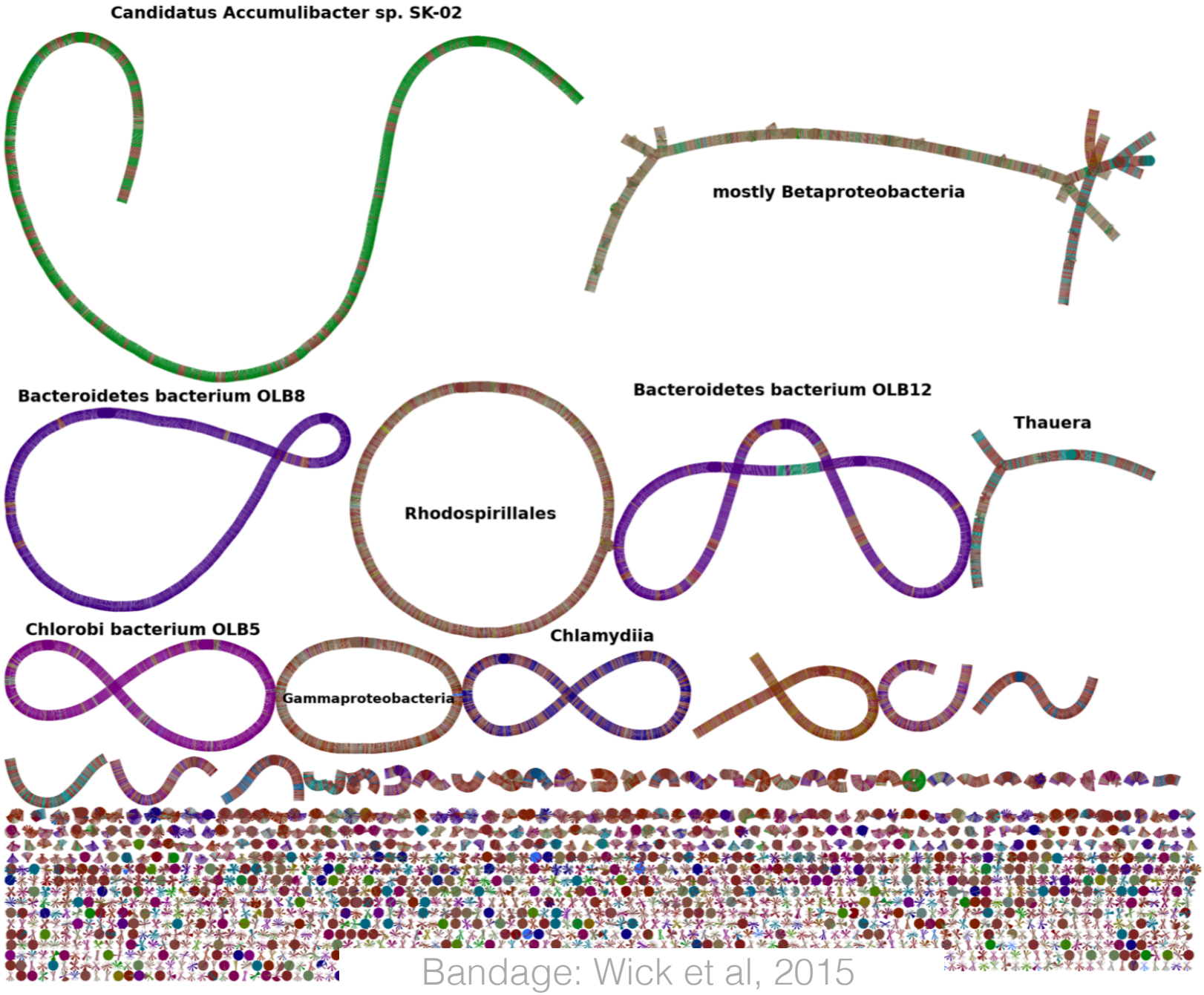
*Microbiome* **7**, Article number: 61 (2019) | Cite this article

- Reads ~695,000, length ~9kb, total ~6Gb

- Unicycler assembly:
  `long-reads/assembly.fa.gz`

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Bandage visualization of assembly graph




Bandage: Wick et al, 2015
Layout: Hachul S., Jünger M., 2007

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Taxonomic bins ≥ 50% complete

| DIAMOND+MEGAN taxonomic bin | | Unicycler contigs | Total (Mb) | Aligned (Mb) | Average coverage | CheckM Complete. | CheckM Contam. | Prokka rRNA | Prokka tRNA | Prokka CDS |
|---|---|---|---|---|---|---|---|---|---|---|
| High quality draft genomes: | | | | | | | | | | |
| B1 | *Bacteroidetes bacterium* OLB12 | 1 | 4.2 | 3.5 | 57.3 | 95% | 0.1% | 6 | 39 | 4,163 |
| B2 | *Candidatus Accumulibacter* SK-02 | 1 | 5.2 | 4.1 | 384.2 | 94% | 0.6% | 4 | 53 | 4,915 |
| B3 | *Chlamydiia* (class) | 1 | 2.8 | 1.8 | 48.8 | 94% | 2% | 6 | 39 | 3,387 |
| B4 | *Gammaproteobacteria* (class) | 43 | 4.7 | 3.0 | | 93% | 2% | 6 | 52 | 4,833 |
| | -longest contig | | 2.7 | 1.6 | 25.1 | 93% | 0.2% | 3 | 40 | 3,359 |
| B5 | *Bacteroidetes bacterium* OLB8 | 1 | 3.8 | 3.0 | 52.1 | 93% | 1% | 6 | 37 | 3,394 |
| B6 | *Rhodospirillales* (order) | 1 | 4.4 | 3.0 | 29.5 | 92% | 0.5% | 3 | 47 | 4,015 |
| B7 | *Chlorobi bacterium* OLB5 | 1 | 3.5 | 2.5 | 38.7 | 88% | 1% | 3 | 41 | 4,131 |
| Medium quality draft genomes: | | | | | | | | | | |
| B8 | *Thauera* (genus) | 25 | 4.6 | 4.0 | | 89% | 4% | 12 | 64 | 4,040 |
| | -longest contig | | 0.8 | 0.7 | 32.7 | 14% | 0% | 0 | 5 | 672 |
| B9 | *Sphingobacteriales bacterium* 44-15 | 59 | 3.2 | 2.8 | | 76% | 1% | 2 | 17 | 2,953 |
| | -longest contig | | 0.2 | 0.1 | 10.2 | 0% | 0% | 0 | 0 | 172 |
| B10 | *Bacteroidetes* (phylum) | 43 | 3.9 | 2.6 | | 72% | 7% | 1 | 12 | 1,997 |
| | -longest contig | | 1.2 | 0.8 | 14.1 | 32% | 0% | 0 | 3 | 807 |
| B11 | *Candidatus Contendobacter* B J11 | 39 | 2.5 | 2.0 | | 59% | 9% | 2 | 37 | 2,668 |
| | -longest contig | | 0.3 | 0.3 | 15.4 | 19% | 0% | 0 | 7 | 295 |
| Low quality draft genomes: | | | | | | | | | | |
| B12 | *Betaproteobacteria* (class) | 111 | 6.6 | 5.5 | | 89% | 79% | 6 | 71 | 4,655 |
| | -longest contig | | 0.4 | 0.3 | 37.1 | 10% | 0% | 0 | 1 | 372 |
| B13 | *Nitrospira* (genus) | 34 | 4.2 | 3.7 | | 83% | 13% | 0 | 6 | 563 |
| | -longest contig | | 1.1 | 0.9 | 17.6 | 27% | 0% | 0 | 2 | 99 |
| B14 | *Chloroflexi* (phylum) | 151 | 5.4 | 4.3 | | 71% | 29% | 0 | 11 | 3,565 |
| | -longest contig | | 0.2 | 0.2 | 13.3 | 8% | 0% | 0 | 1 | 86 |

Arumugam et al, 2019

CheckM (Parks et al. 2014)

Prokka (Seemann, 2014)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Enrichment reactor - long reads

- Run DIAMOND in long-read mode:

```
diamond blastx -d nr \
-q data2/assembly.fa.gz \
-o data/assembly.daa -f 100 \
--range-culling -k 25 -F 15
```

- Run MEGANIZER in long-read mode:

```
$MEGAN/tools/daa-meganizer \
-i data2/assembly.daa -lg  \
-mdb megan-map-Feb2022.db
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Enrichment reactor - long reads

- Investigate the long-read inspector dialog
- Discuss frame-shift correction

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
Institute for Bioinformatics
and Medical Informatics

# Thank You!

Please provide tutorial feedback to ISMB using this QR-code:

Feedback webpage

**Joint work with:**

- Caner, Baci, Timo Lucas, Sascha Patz, Patrick Wörz — Tübingen
- Krithika Arumugam, Irina Bessarab & Rohan Williams — SCELSE/NUS Singapore

**Funding:**

- Deutsche Forschungsgemeinschaft (MAIRA & BinAC)
- Life Sciences Institute at NUS
- NRF/MOE and NRF-EW, Singapore