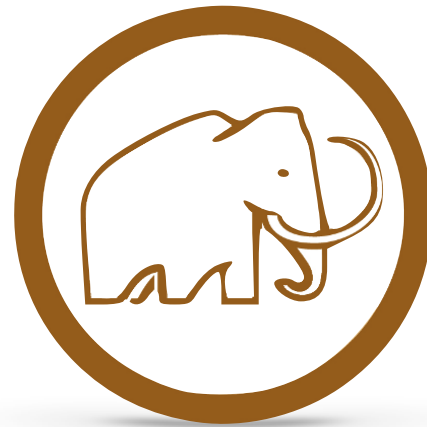


Introduction to Microbiome Analysis using **DIAMOND+MEGAN**



Daniel H. Huson



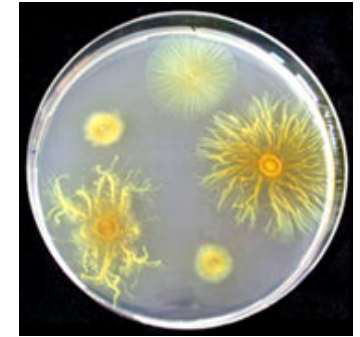
Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- MEGAN7
- Hands-on session

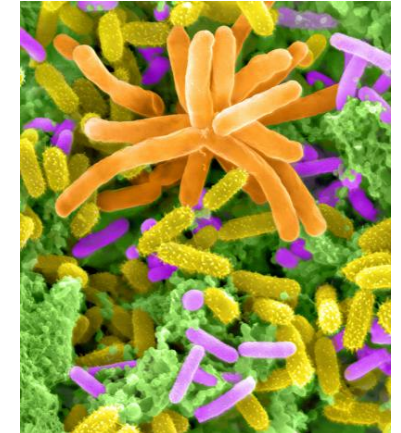
- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- MEGAN7
- Hands-on session

Microbiome

- 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.innovations-report.de



www.physorg.com

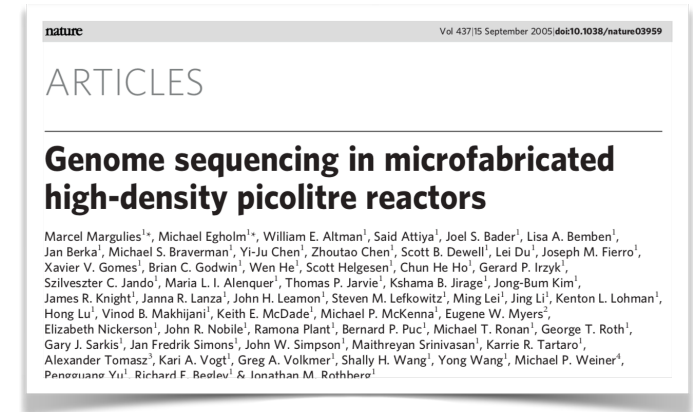
Sources of studied microbiomes

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





- First NGS technique 454 released
- Intended for genome sequencing...



★ Use NGS to sequence ancient DNA?

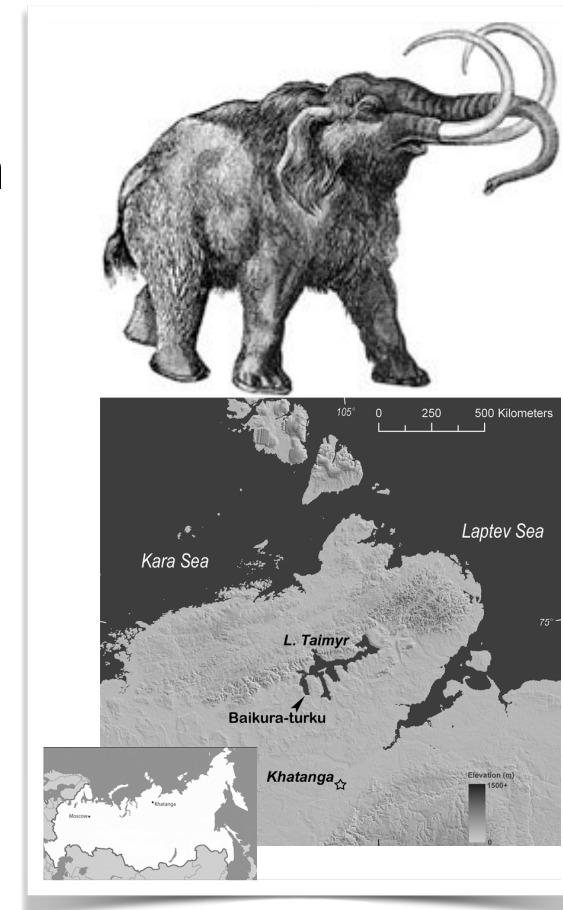


★ Use NGS to sequence metagenomic DNA?

NGS = next generation sequencing

Mammoth DNA & metagenome (2006)

- DNA collected from permafrost mammoth (28,000 years old)
 - DNA extracted from 1g bone
 - Sequenced using 454
 - ~302,000 reads, length ~95 bp
-
- ★ Can use NGS for ancient DNA
 - ★ First NGS metagenomics paper



REPORTS

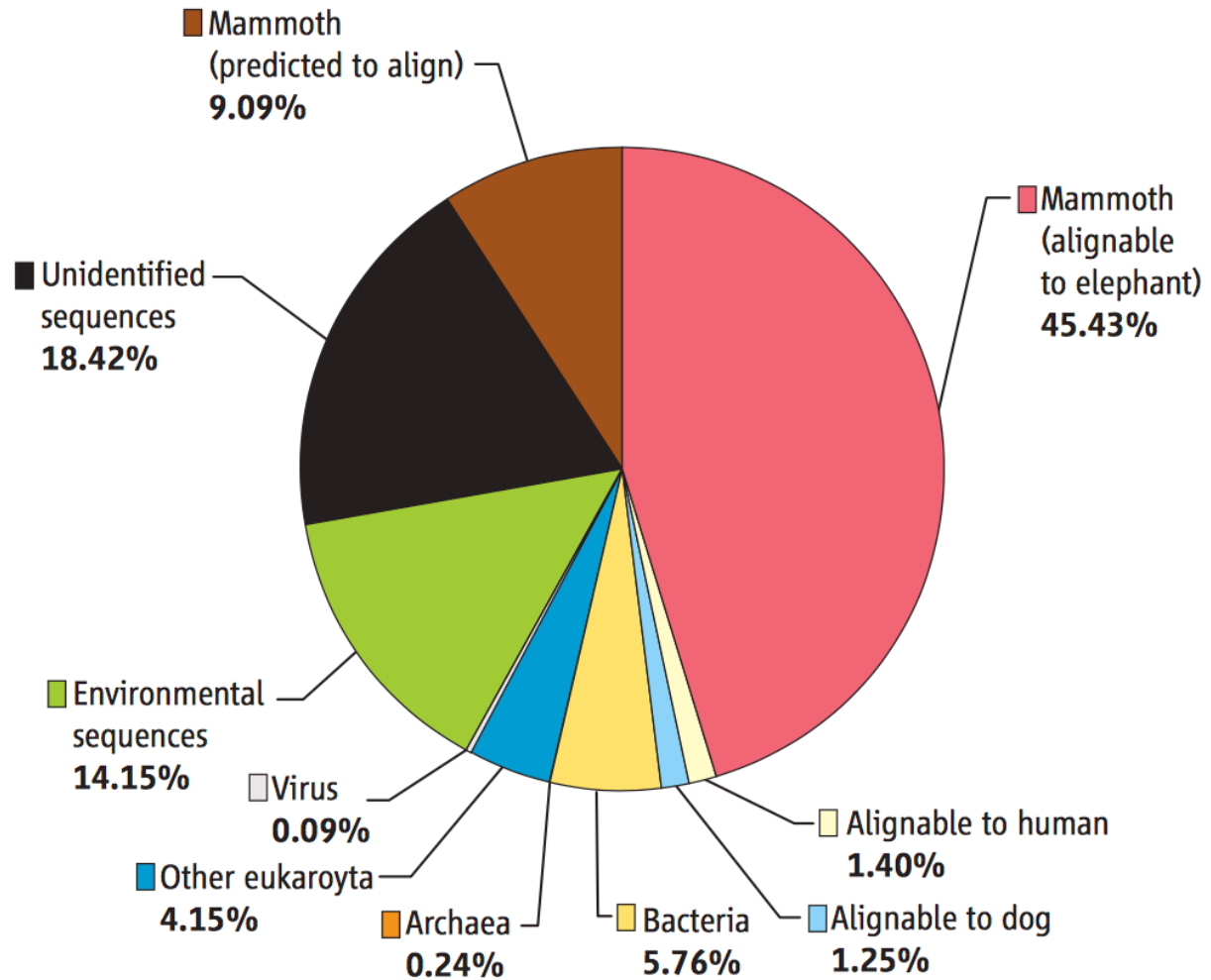
Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA

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

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.

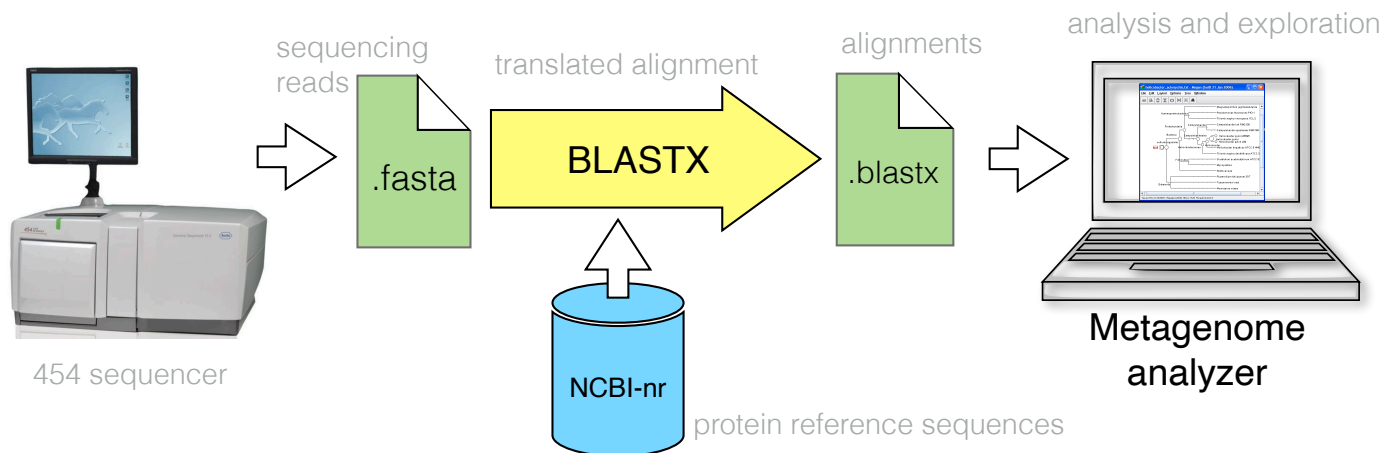


Poinar et al, Science 2006

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

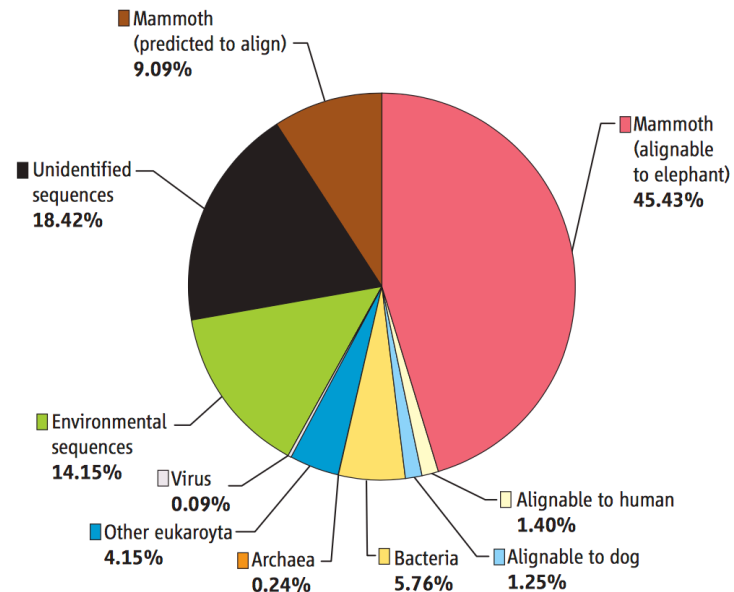


2006 MEGAN analysis pipeline

How to analyze metagenomic reads? (2006)

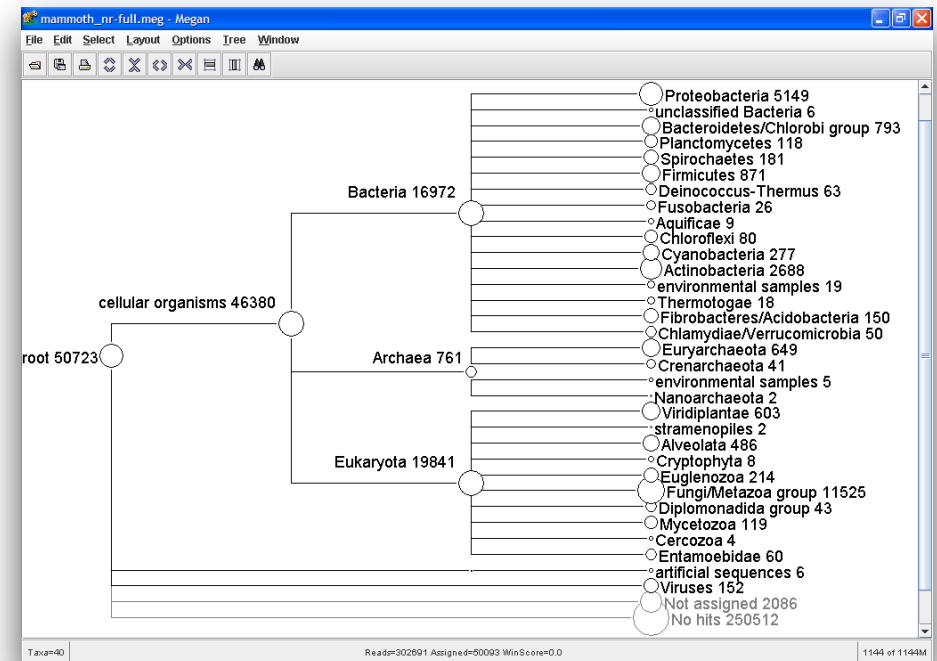
- MEGAN (MEtagenome ANalyzer 1.0)

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

MEGAN 1.0



H. et al, Genome Research, 2007



Computational bottleneck (2006)

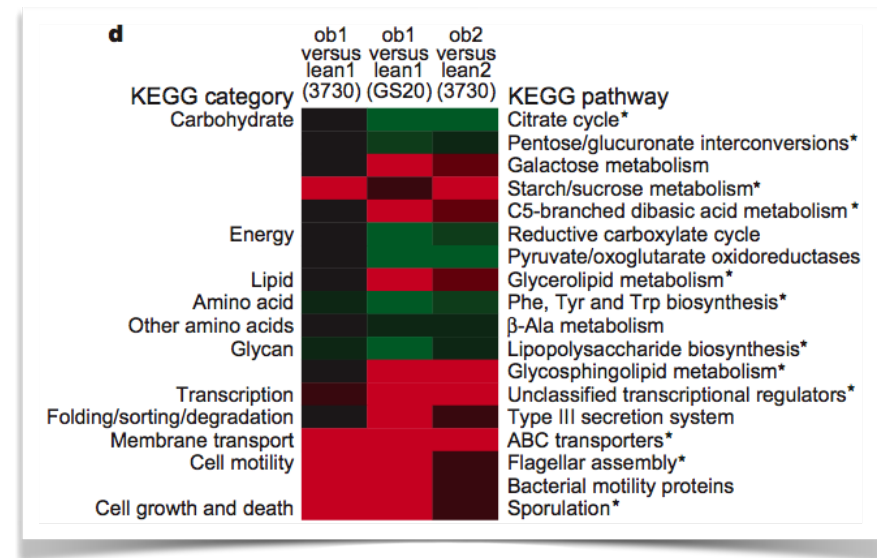
- Compare all reads against the NCBI-nr protein database
- Year 2006:
 - 300,000 reads of length ~ 100 bp
 - NCBI-nr: 3 million entries, ~ 1 billion letters
- ★ BLASTX took a couple of weeks on a small cluster

(NCBI-nr today: ~ 1.2 billion entries)

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)



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 Arumugam², Kristoffer Solvsten Burgdorf⁴, Chaysavanh Manichanh⁵, Trine Nielsen⁴, Nicolas Pons⁶, Florence Levenez⁶, Takuji Yamada², Daniel R. Mende², Junhua Li^{1,7}, Junming Xu¹, Shaochuan Li¹, Dongfang Li^{1,8}, Jianjun Cao¹, Bo Wang¹, Huiqing Liang¹, Huisong Zheng¹, Yinlong Xie^{1,7}, Julien Tap⁶, Patricia Lepage⁶, Marcelo Bertalan⁹, Jean-Michel Batto⁶, Torben Hansen⁴, Denis Le Paslier¹⁰, Allan Linneberg¹¹, H. Bjørn Nielsen⁹, Eric Pelletier¹⁰, Pierre Renault⁶, Thomas Sicheritz-Ponten⁹, Keith Turner¹², Hongmei Zhu¹, Chang Yu¹, Shengting Li¹, Min Jian¹, Yan Zhou¹, Yingrui Li¹, Xiuqing Zhang¹, Songgang Li¹, Nan Qin¹, Huanming Yang¹, Jian Wang¹, Søren Brunak⁹, Joel Doré⁶, Francisco Guarner⁵, Karsten Kristiansen¹³, Oluf Pedersen^{4,14}, Julian Parkhill¹², Jean Weissenbach¹⁰, MetaHIT Consortium†, Peer Bork², S. Dusko Ehrlich⁶ & Jun Wang^{1,15}

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

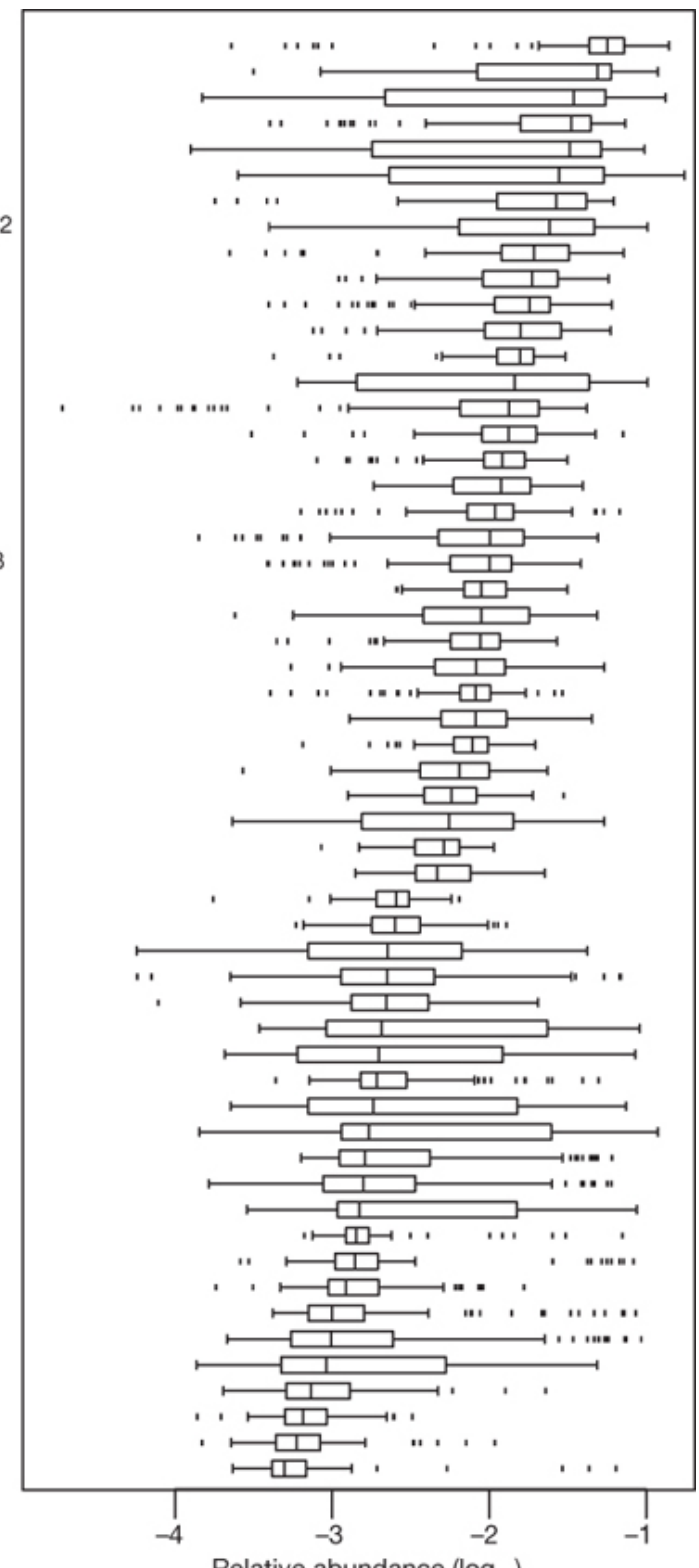
- 576Gb of sequence from 124 individuals

Core of human gut microbiome

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

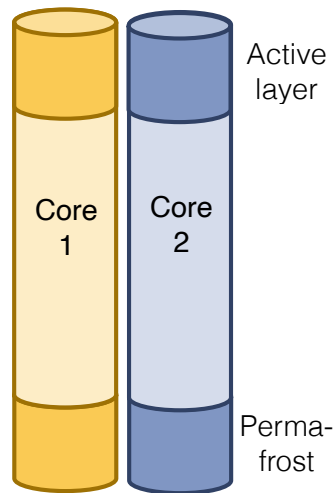
BLASTX at Super Computer
 Center in Barcelona, then
 MEGAN analysis

Bacteroides uniformis
Alistipes putredinis
Parabacteroides merdae
Dorea longicatena
Ruminococcus bromii L2-63
Bacteroides caccae
Clostridium sp. SS2-1
Bacteroides thetaiotaomicron VPI-5482
Eubacterium hallii
Ruminococcus torques L2-14
 Unknown sp. SS3 4
Ruminococcus sp. SR1 5
Faecalibacterium prausnitzii SL3 3
Ruminococcus lactaris
Collinsella aerofaciens
Dorea formicigenerans
Bacteroides vulgatus ATCC 8482
Roseburia intestinalis M50 1
Bacteroides sp. 2_1_7
Eubacterium siraeum 70 3
Parabacteroides distasonis ATCC 8503
Bacteroides sp. 9_1_42FAA
Bacteroides ovatus
Bacteroides sp. 4_3_47FAA
Bacteroides sp. 2_2_4
Eubacterium rectale M104 1
Bacteroides xylanisolvens XB1A
Coprococcus comes SL7 1
Bacteroides sp. D1
Bacteroides sp. D4
Eubacterium ventriosum
Bacteroides dorei
Ruminococcus obeum A2-162
Subdoligranulum variabile
Bacteroides capillosus
Streptococcus thermophilus LMD-9
Clostridium leptum
Holdemania filiformis
Bacteroides stercoris
Coprococcus eutactus
Clostridium sp. M62 1
Bacteroides eggerthii
Butyrivibrio crossotus
Bacteroides finegoldii
Parabacteroides johnsonii
Clostridium sp. L2-50
Clostridium nexile
Bacteroides pectinophilus
Anaerotruncus colihominis
Coprococcus gnauvus
Bacteroides intestinalis
Bacteroides fragilis 3_1_12
Clostridium asparagiforme
Enterococcus faecalis TX0104
Clostridium scindens
Blautia hansenii



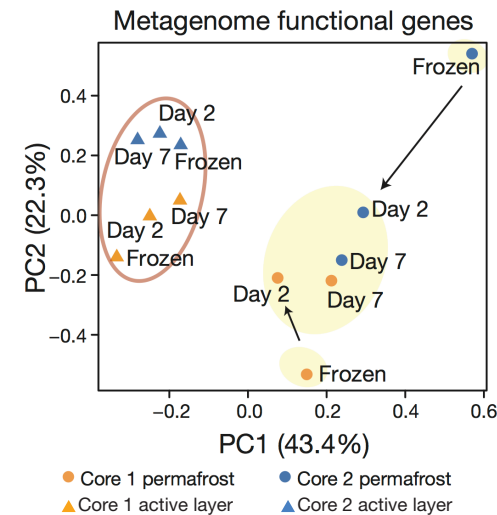
Permafrost study (2011)

(Mackelprang *et al*, Science 2011)



Their question:
Functional changes
during thawing?

Frozen, day 2, day 7



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



on 100 cores

Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- Hands-on session

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

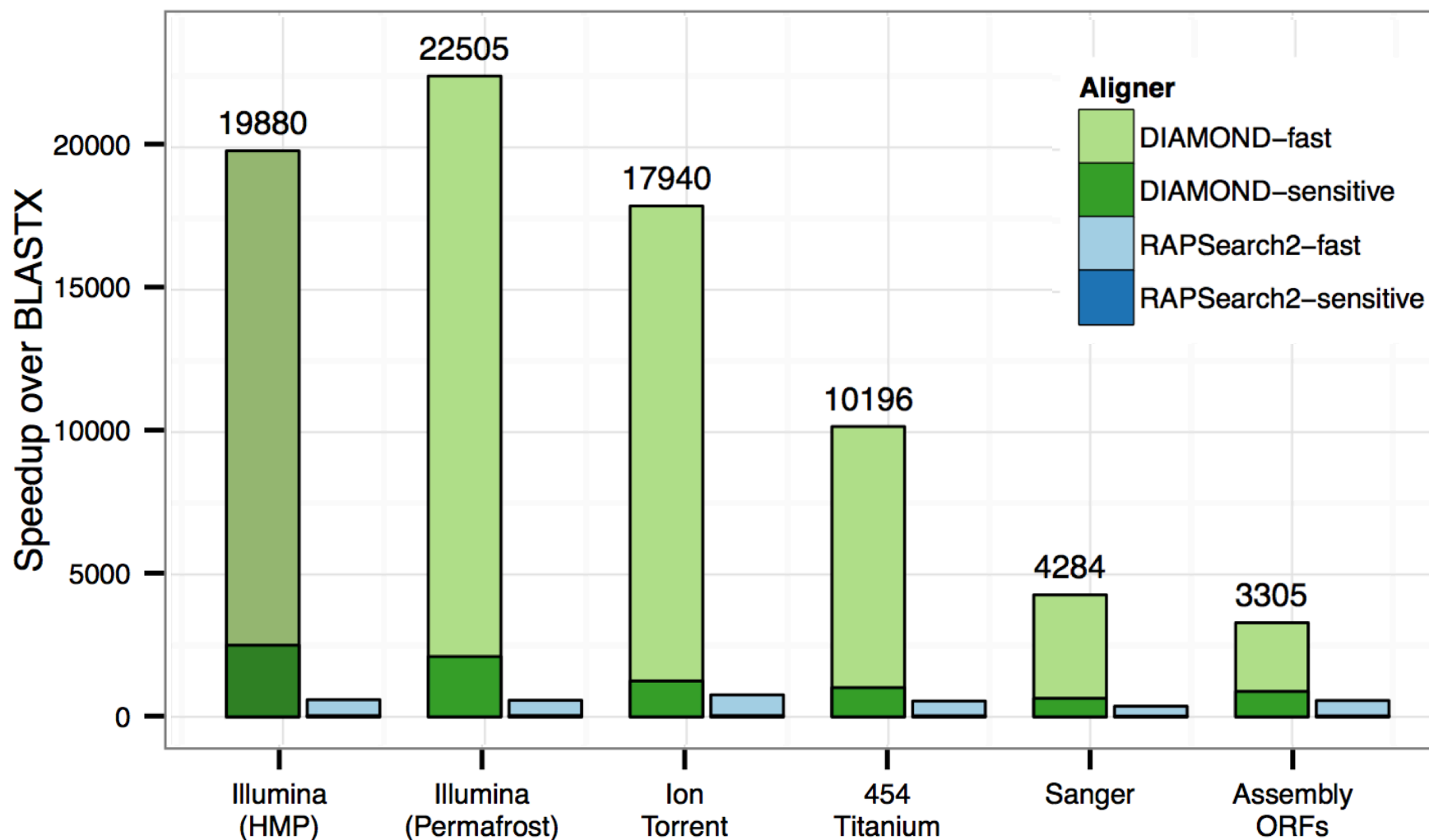
DIAMOND BLAST!

- Translated alignment tool DIAMOND
- DIAMOND replaces BLASTX on microbiome sequencing reads
- Very similar sensitivity to BLASTX on short reads
- Much, much faster...

**Fast and sensitive protein
alignment using DIAMOND**

Benjamin Buchfink¹, Chao Xie^{2,3} &
Daniel H Huson^{1,2} **NATURE METHODS** 2015

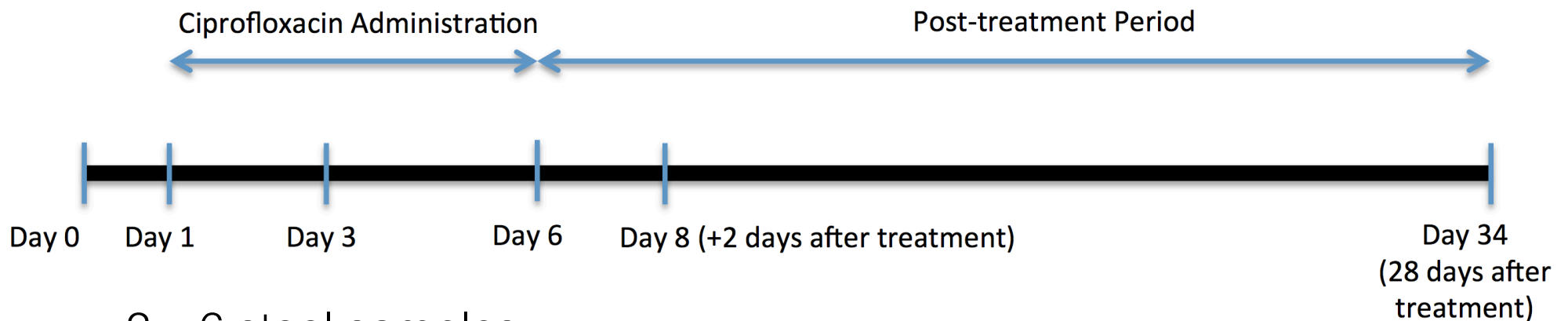
DIAMOND performance



ASARI- Antibiotic resistance pilot study



- 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

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	455 681 58	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





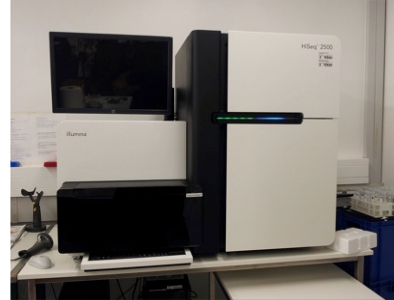
Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- MEGAN7
- Hands-on session

Three computational questions



Hundreds of Samples



High-throughput
DNA sequencing

Billions of sequences

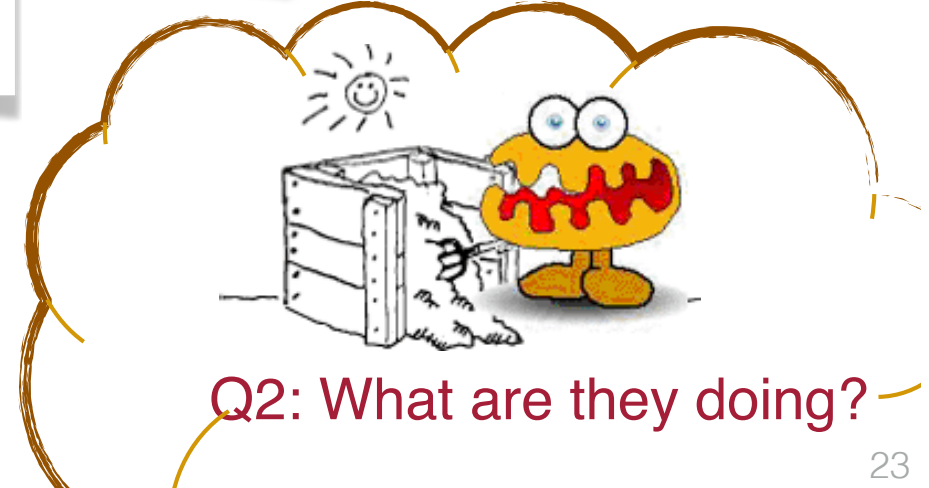


Basic computational
analysis

Many
CPU hours

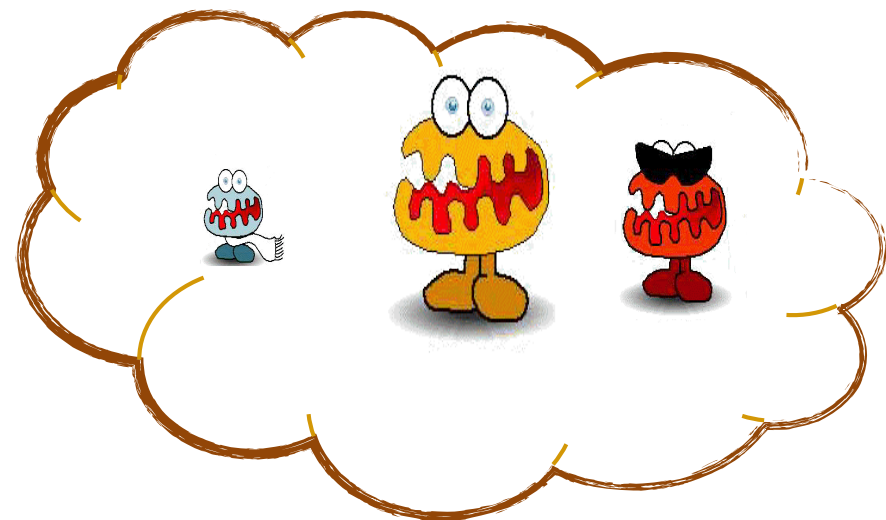


Q1: Who is out there?



Q2: What are they doing?

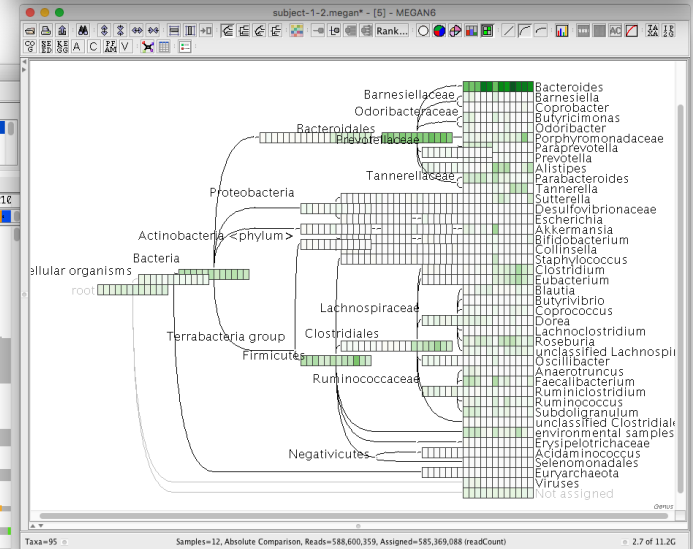
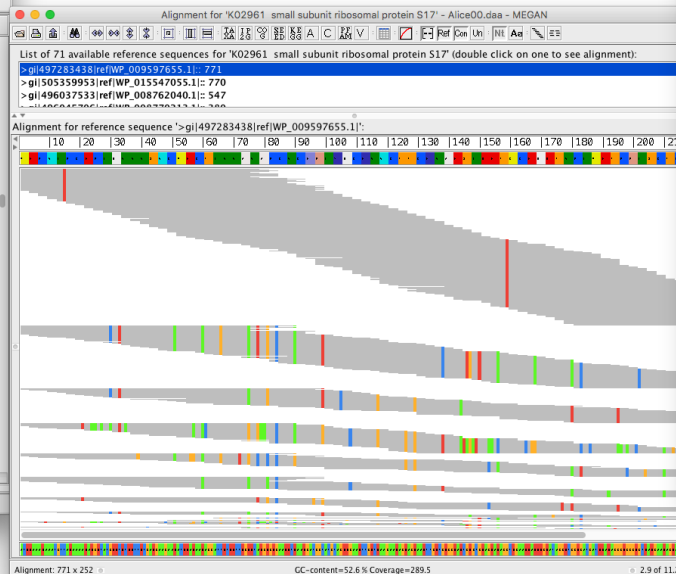
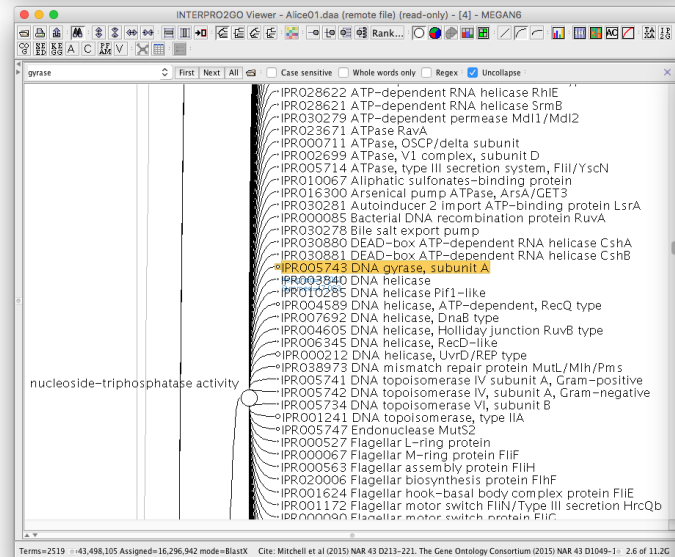
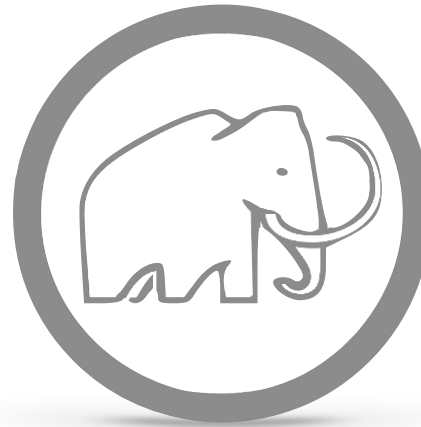
Q3: How do they compare?



Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- **MEGAN taxonomic and functional binning**
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- MEGAN7
- Hands-on session

PCoA analysis



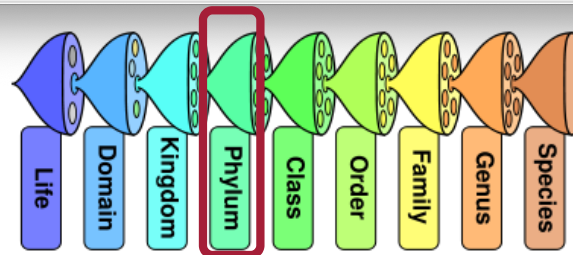
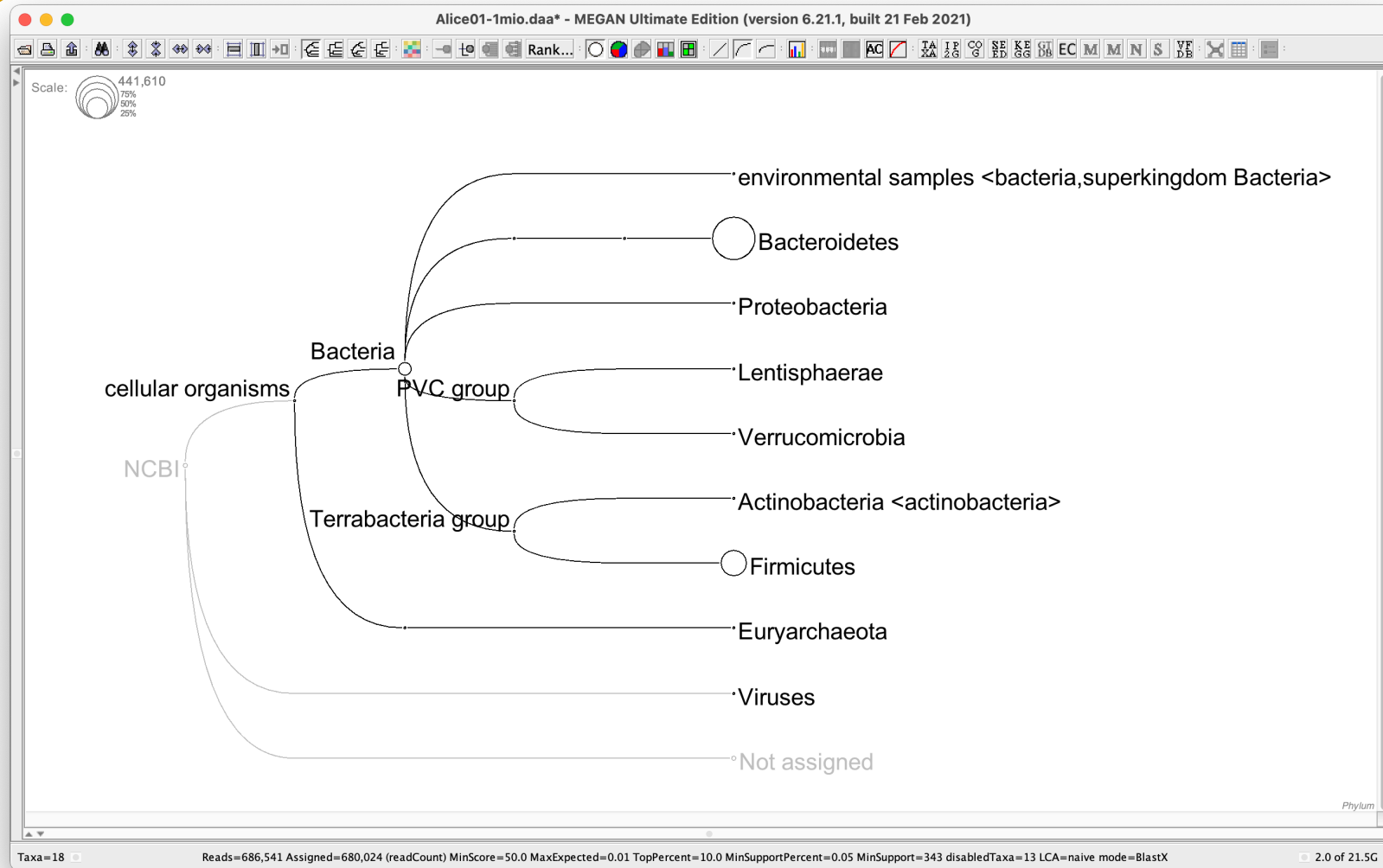
Comparative analysis

Gene-centric alignment and assembly

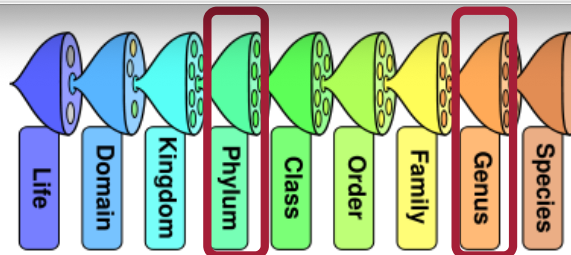
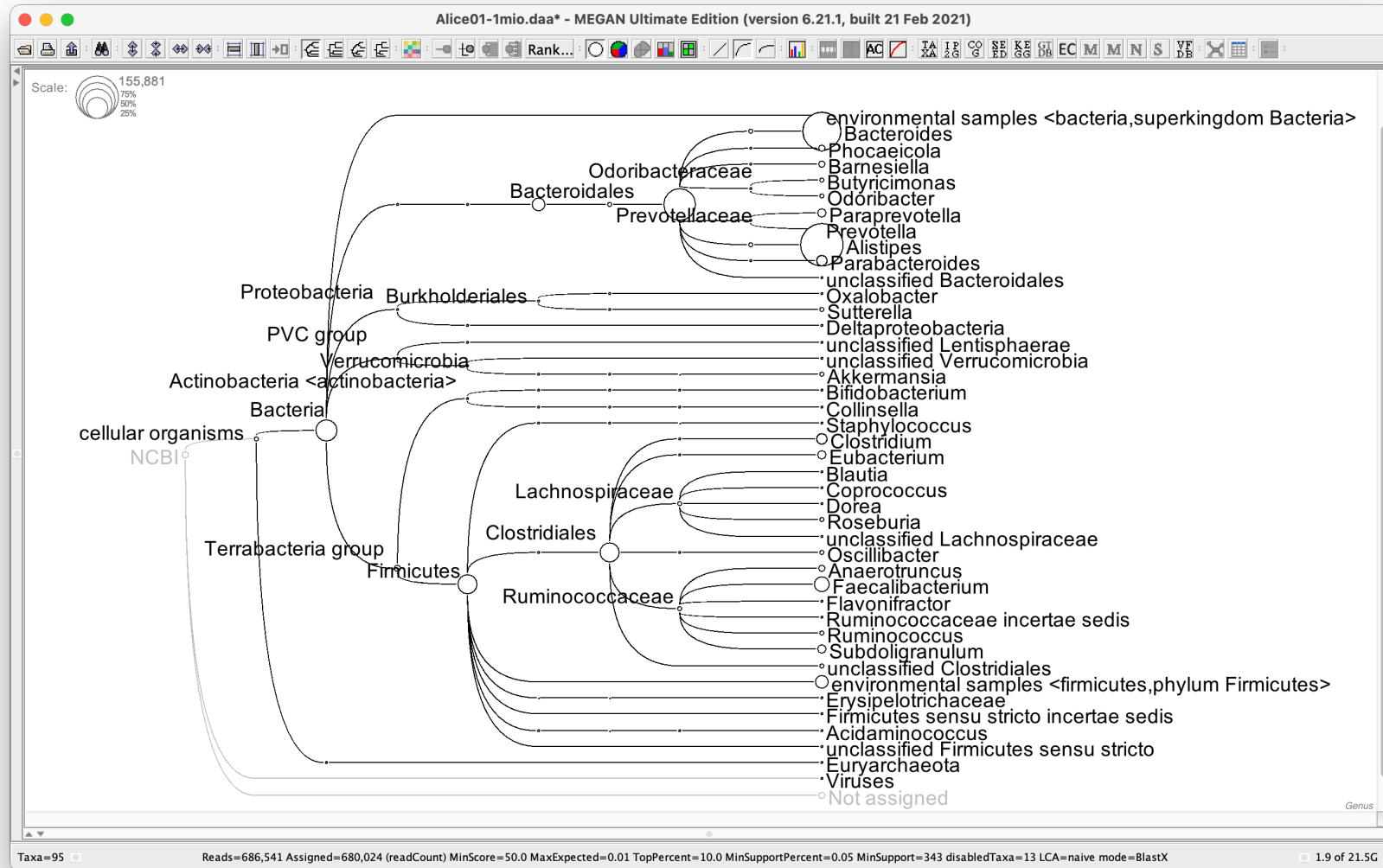


Taxonomic content

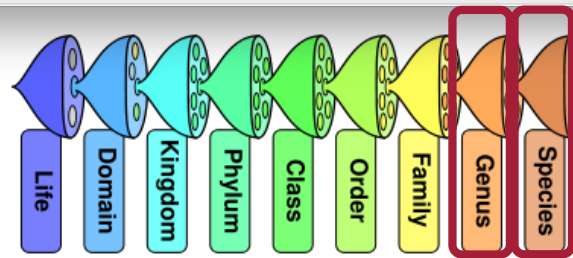
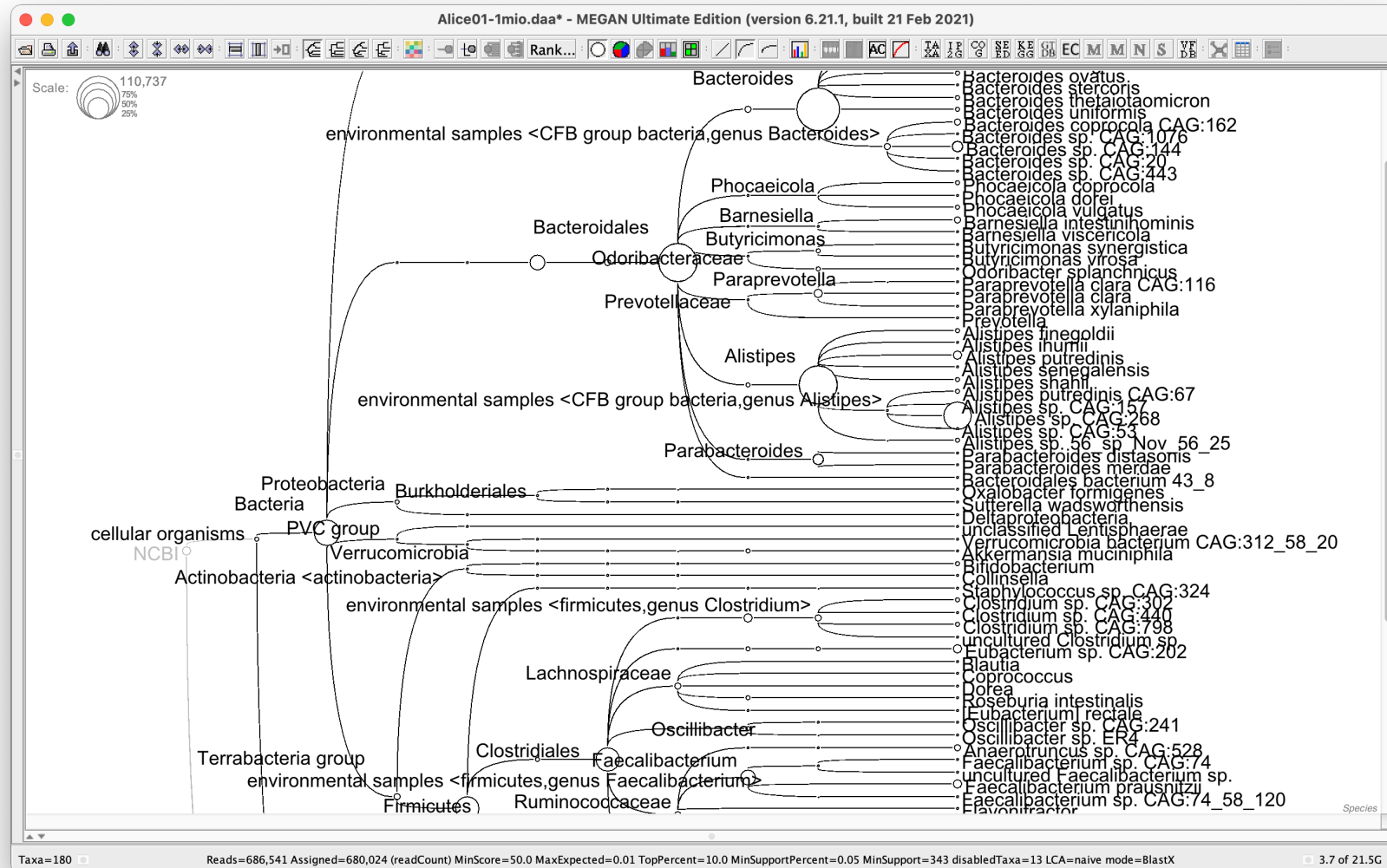
ASARI human gut microbiome



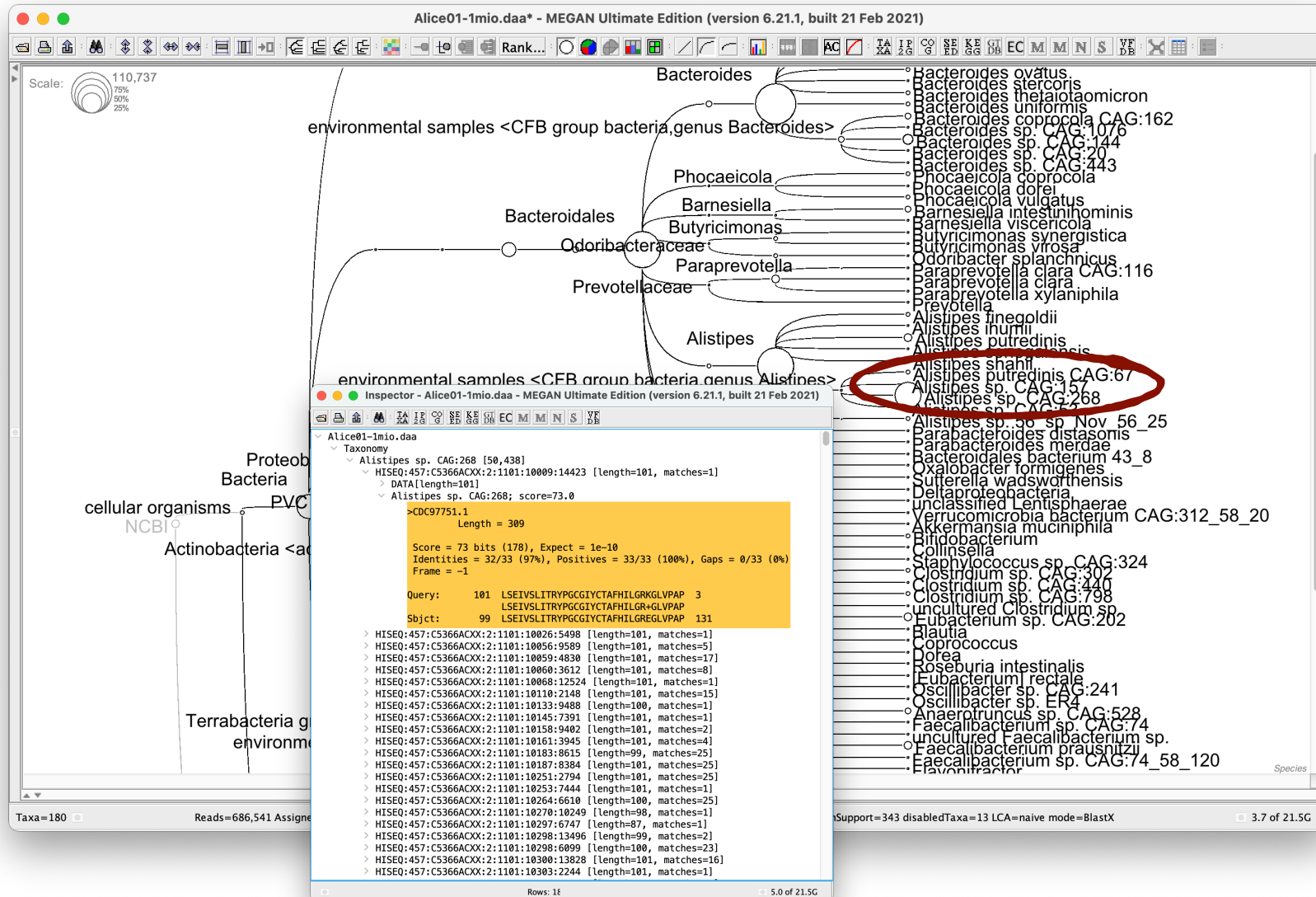
Taxonomic content



Taxonomic content



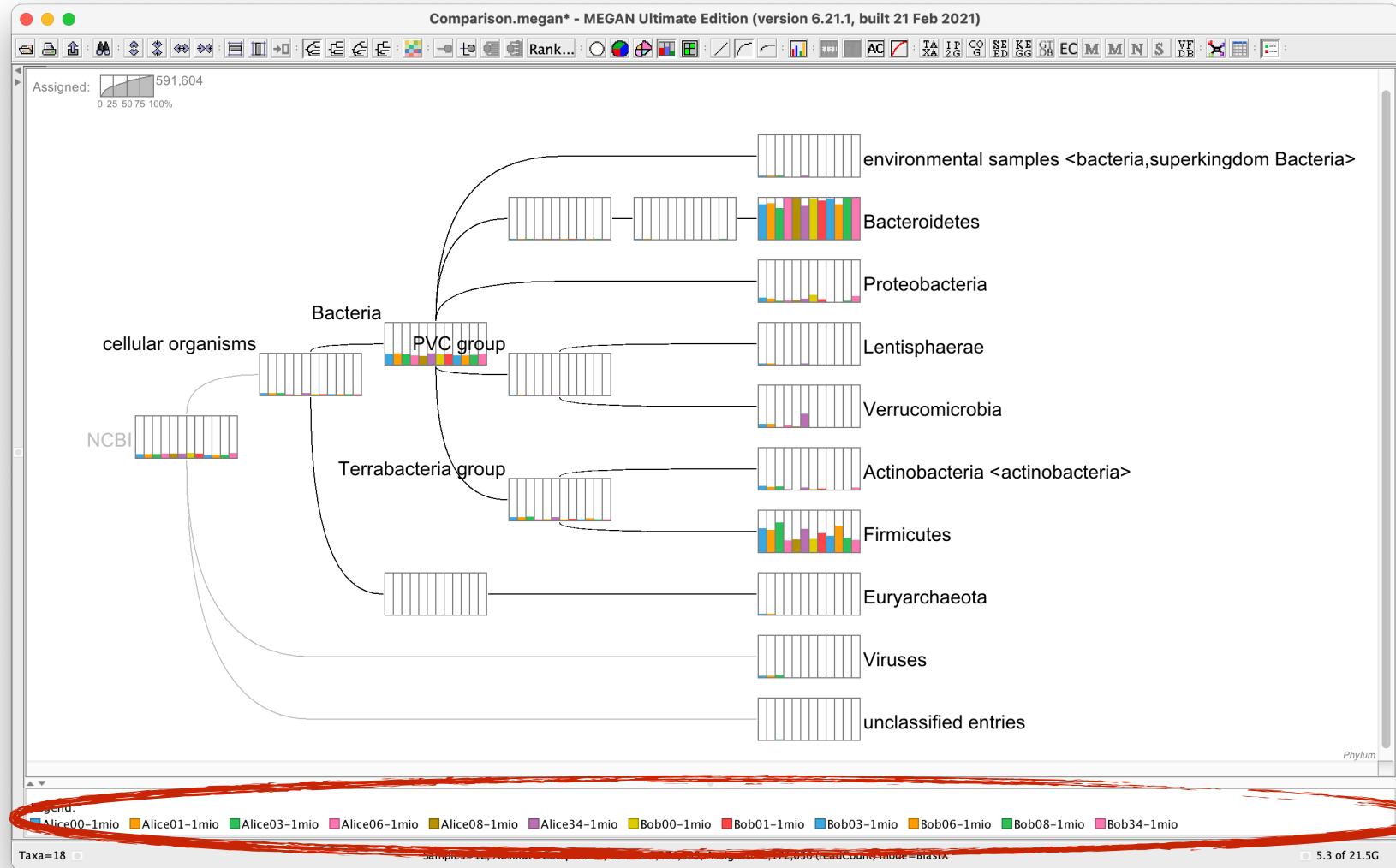
Drill down to details...



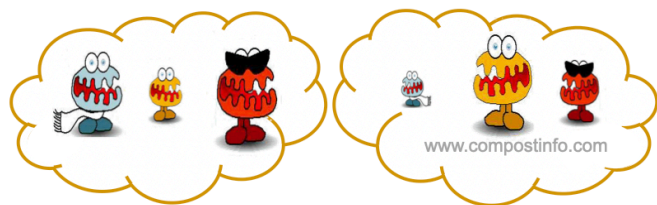


Comparison

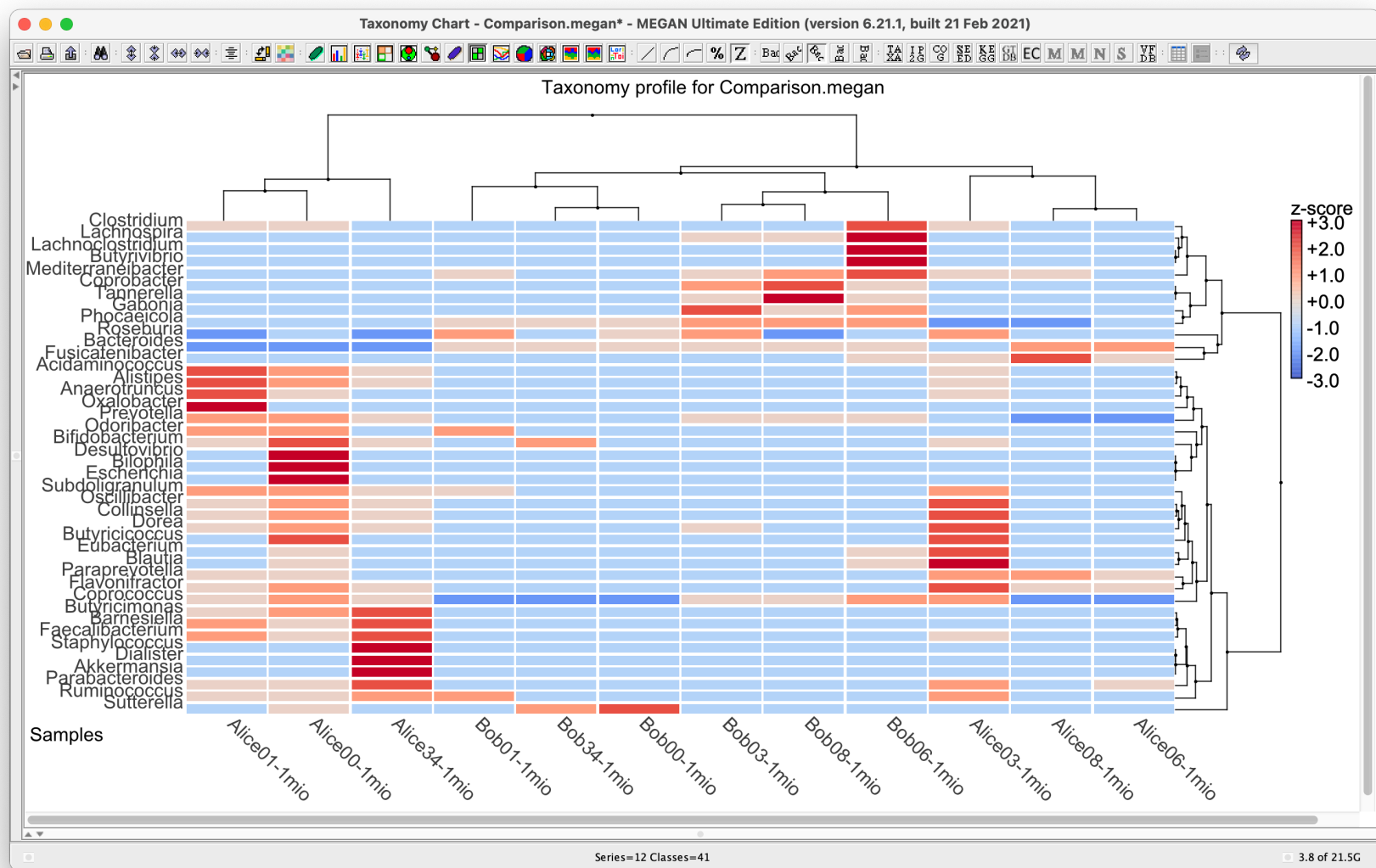
Q3: How do they compare?



All 12 ASARI human gut samples together

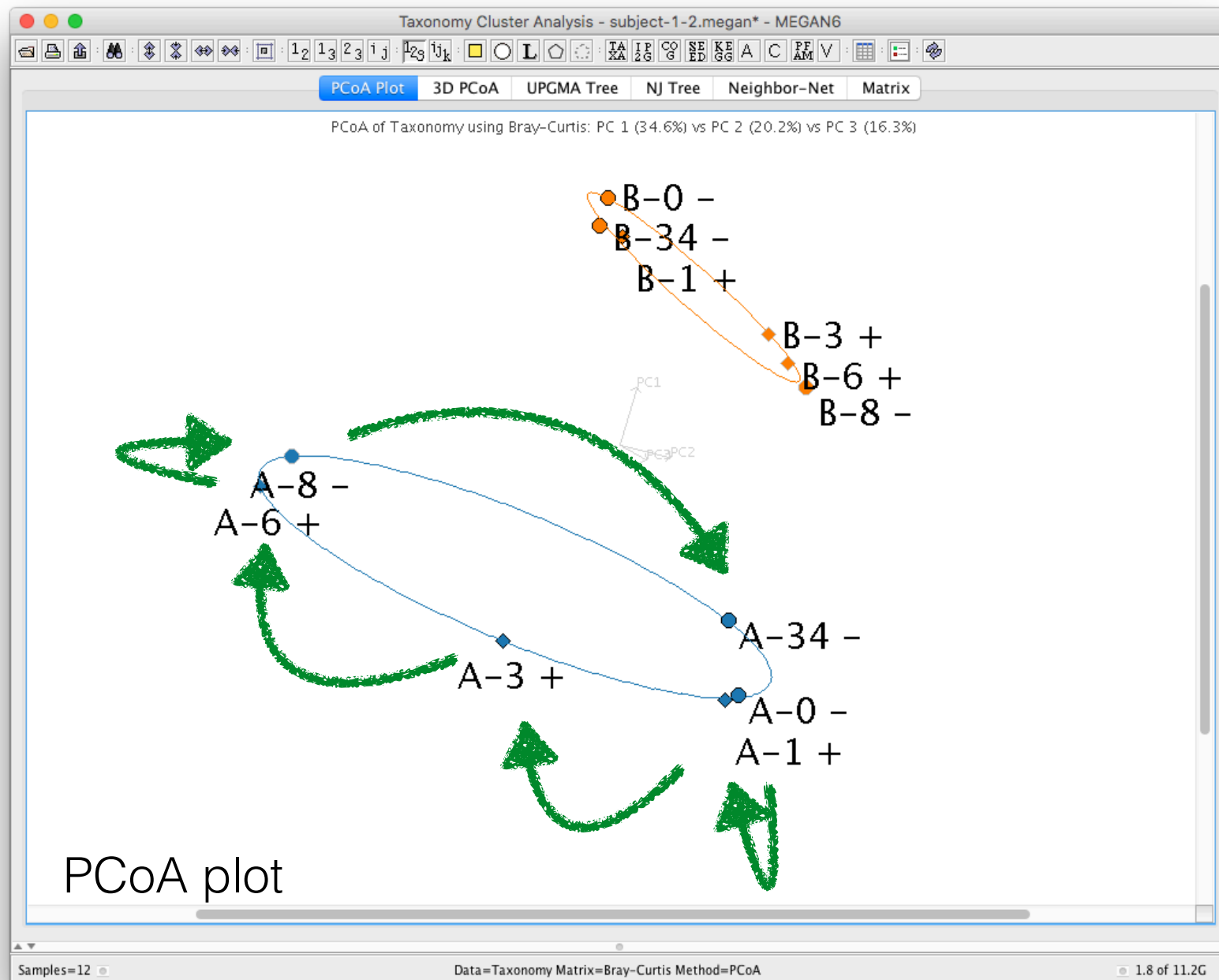


Comparison

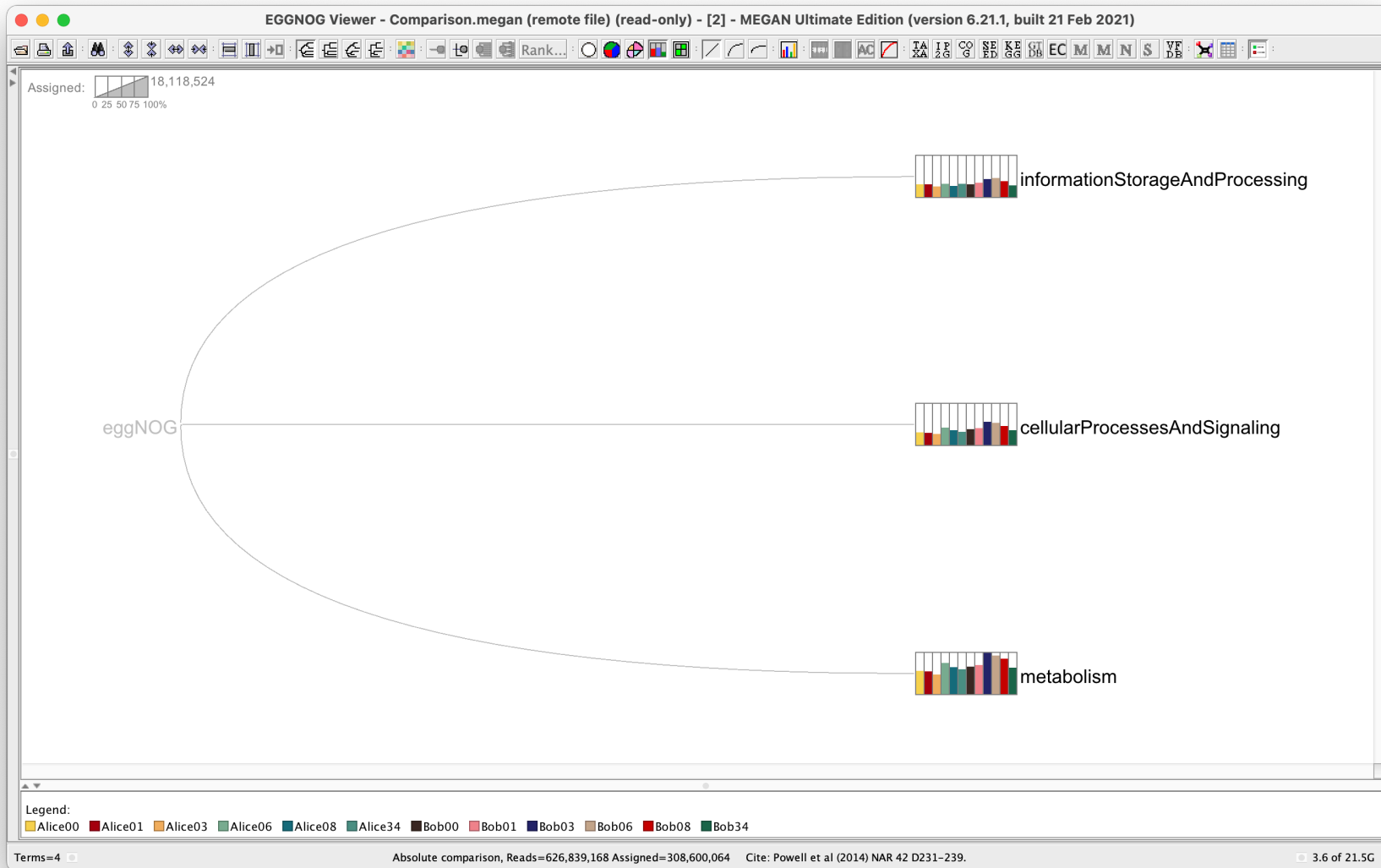


All 12 ASARI human gut samples together

E.g.: Does the microbiome rebound?



Functional content

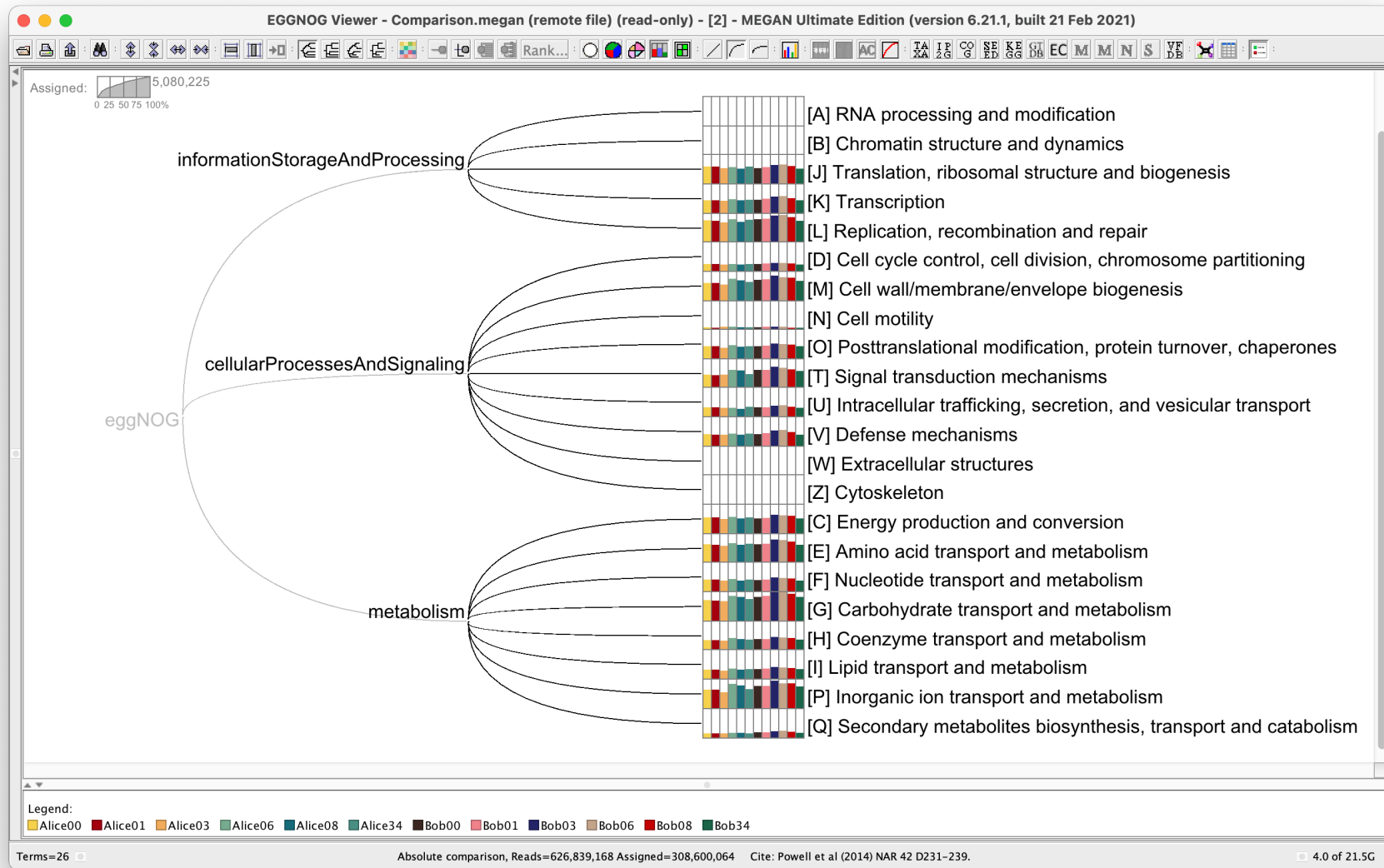


eggNOG classification
(Powell et al, 2014)



Q2: What are they doing?

Functional content

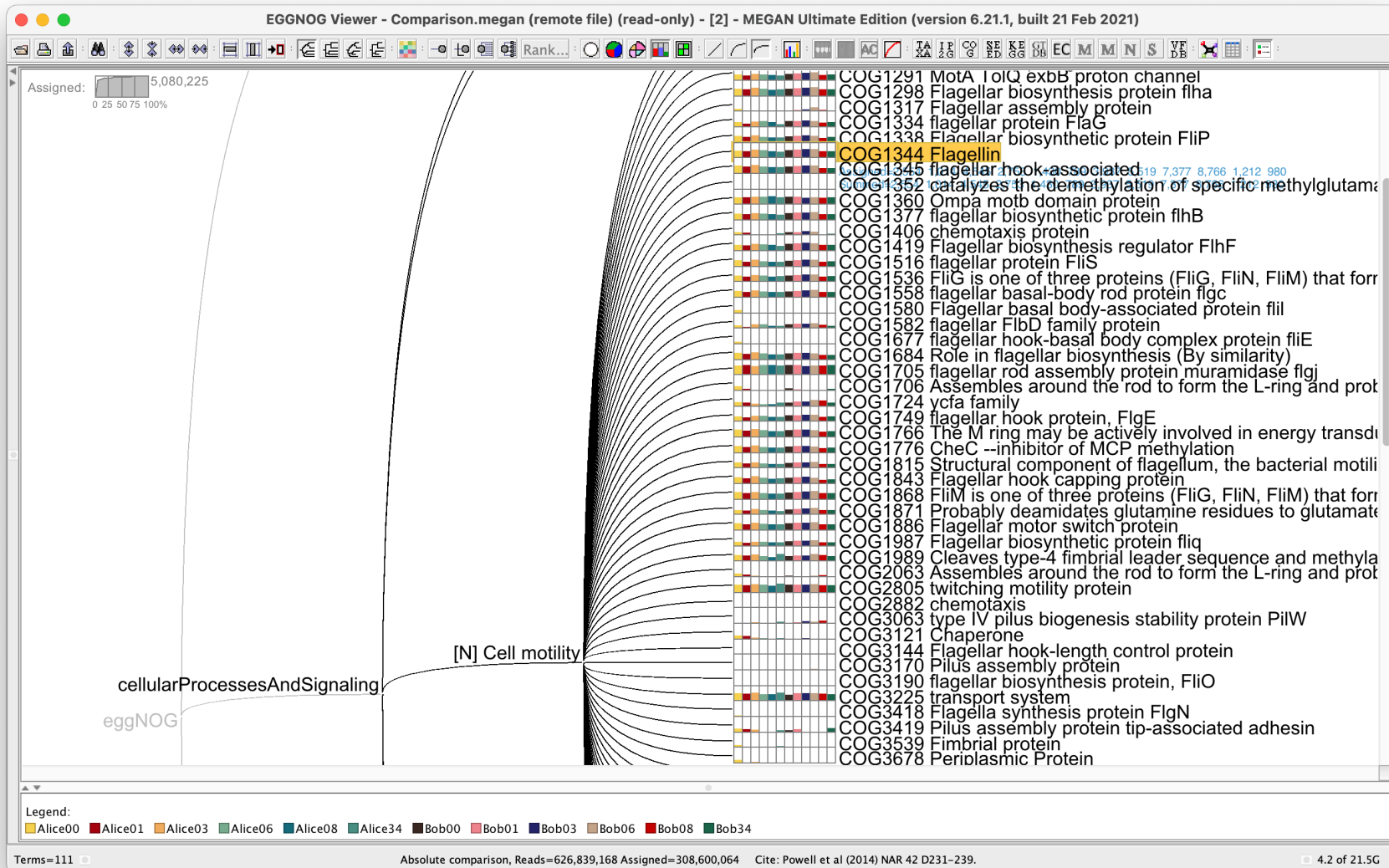


eggNOG classification
 (Powell et al, 2014)



Q2: What are they doing?

Functional content



EggNOG classification
(Powell et al, 2014)



MEGAN binning

- Taxonomic binning using: NCBI taxonomy or GTDB
- Functional binning using:
 - InterPro families (Mitchell et al, 2015)
 - eggNOG/COG (Powell et al, 2014)
 - SEED (Overbeek et al, 2014)
 - KEGG (license required) (Kanehisa M & Goto S, 2000)
 - EC numbers



Outline

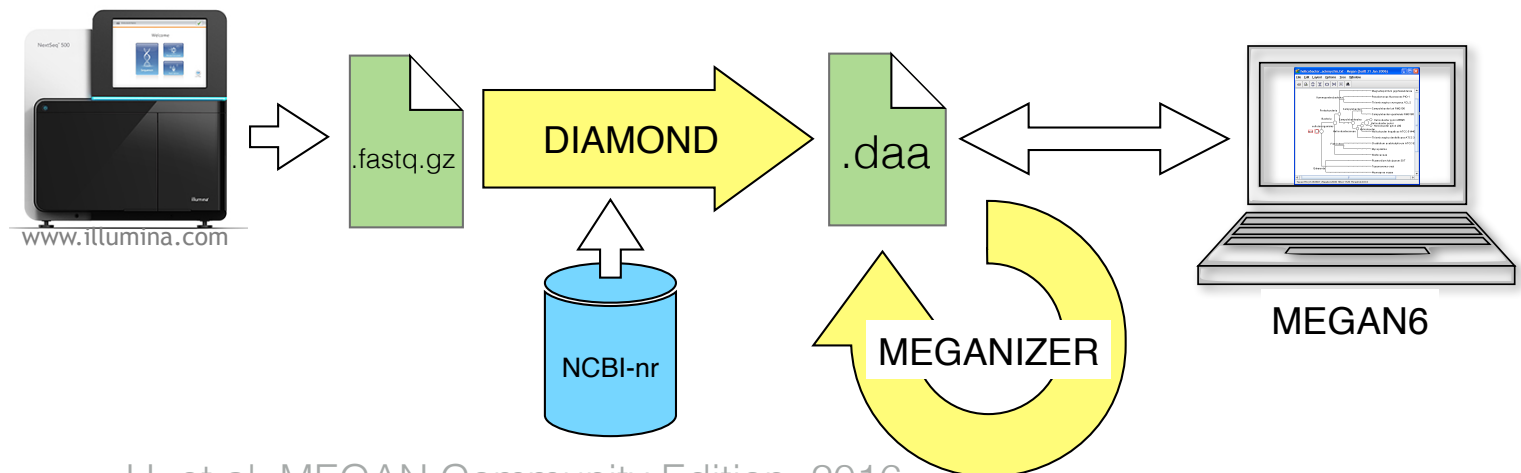
- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- Hands-on session

DIAMOND+MEGAN

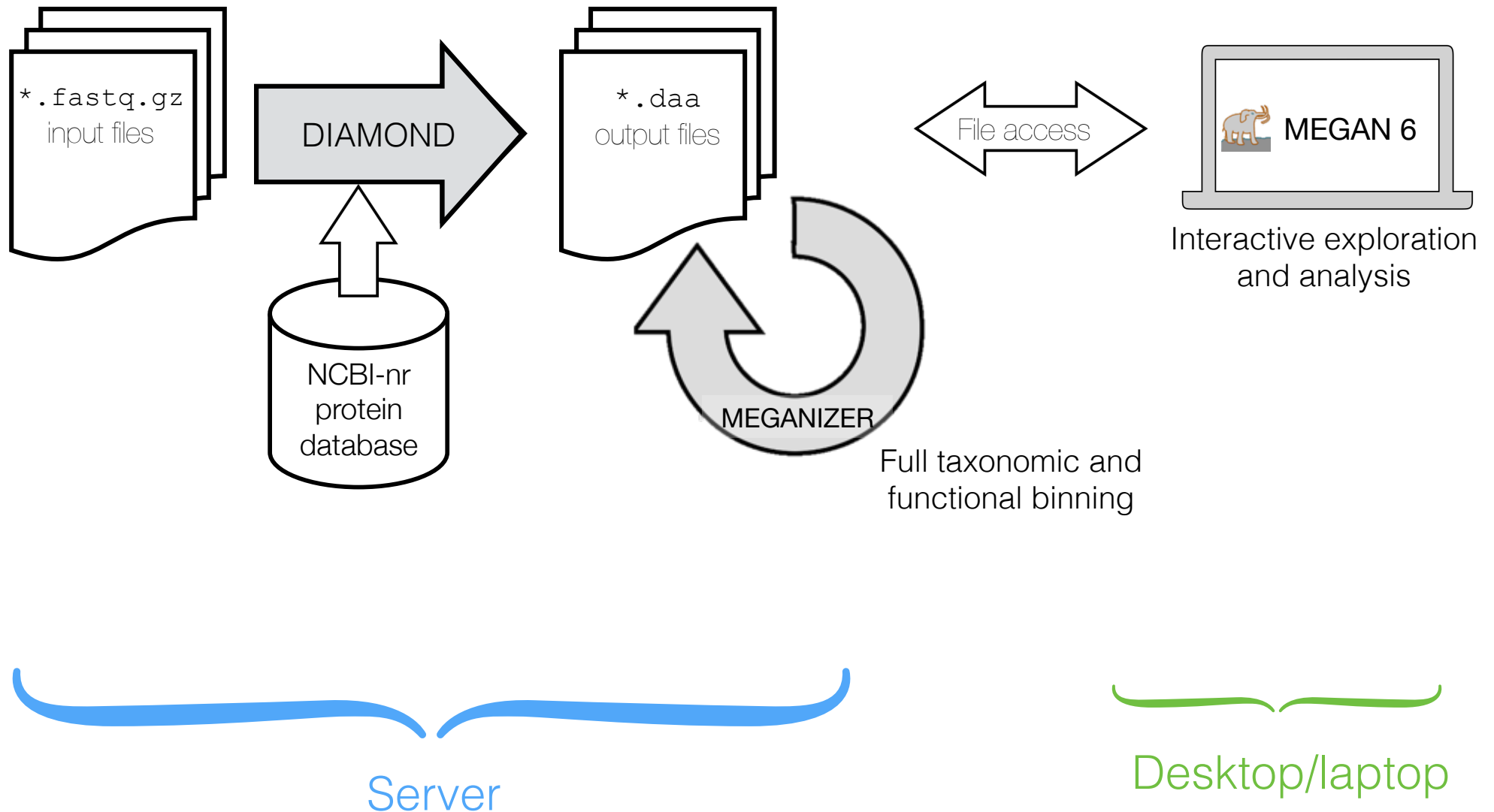
Meganizer program available with MEGAN

- Performs taxonomic and functional binning of reads
- Indexes all data
- Appends results to the DIAMOND output file
- Reduces the total number of files generated in a metagenome analysis to **2**

- Basic Pipeline:



DIAMOND+MEGAN pipeline



Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- Hands-on sessions

Microbiome read-length paradox

short reads

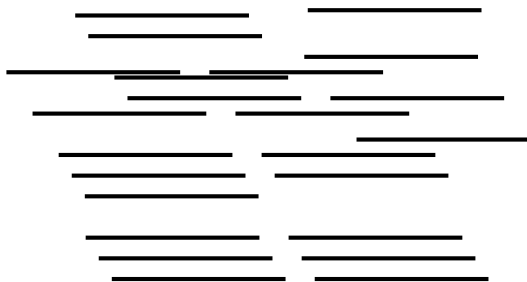


assembly

Input short, surely
should assemble?

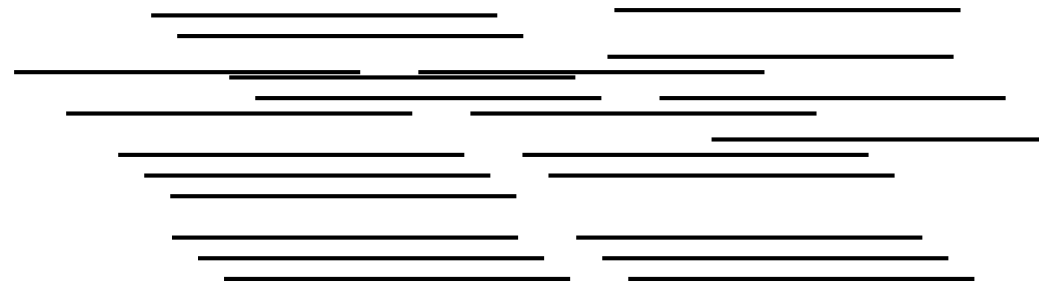
✗ No -

Contigs usually too short



short contigs

long reads

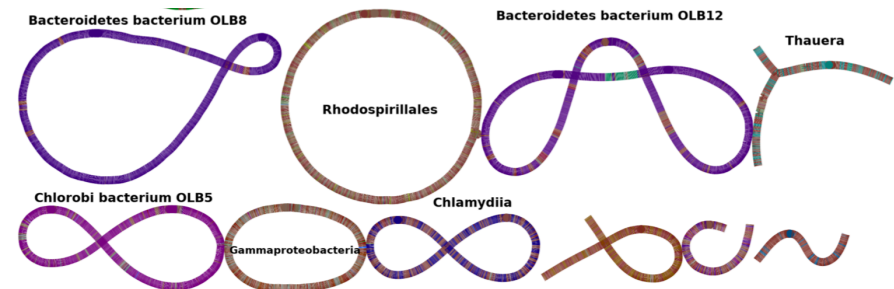


assembly

Input long, do we
need to assemble?

✓ Yes -

Complete chromosomes out!

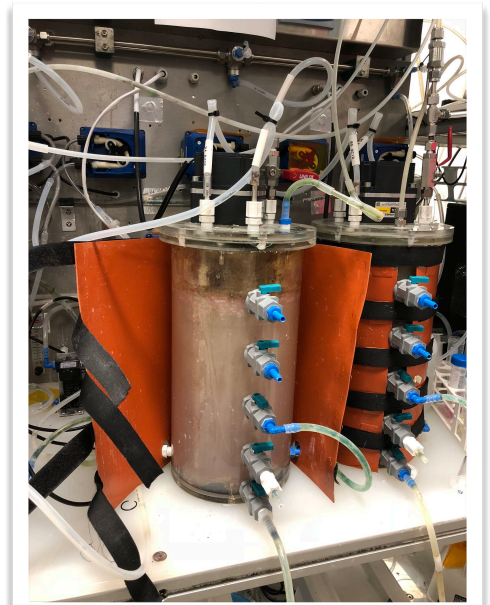


chromosomes

Long-read metagenomics

- EBPR waste-water bio-reactor
- MinION sequencing 2018
 - Reads: ~695,000 (~ 6 Gb)
 - Length: ~9 kb mean (2 bp - 66 kb)
 - Short Read Archive SRX5120474

Joint work with: Rohan Williams,
Krithika Arumugam, Irina Bessarab
and others at NUS and SCELSE



Krithika Arumugam



Short report | [Open Access](#) | Published: 16 April 2019

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

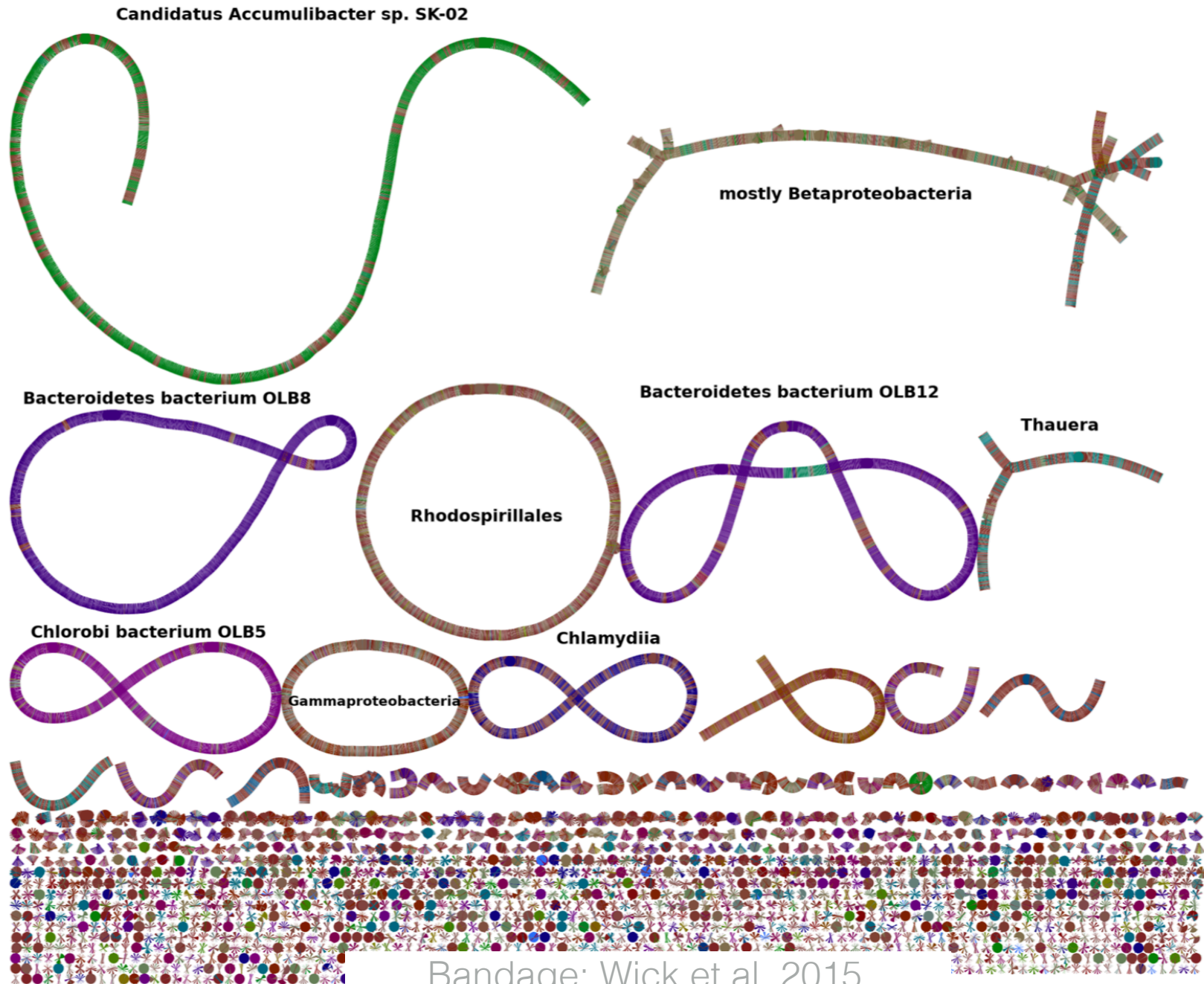
[Krithika Arumugam](#), [Caner Bağcı](#), [Irina Bessarab](#), [Sina Beier](#), [Benjamin Buchfink](#), [Anna Górska](#), [Guanglei Qiu](#), [Daniel H. Huson](#) & [Rohan B. H. Williams](#) 

[Microbiome](#) 7, Article number: 61 (2019) | [Cite this article](#)

Long-read metagenome assembly

- Input:
 - Reads: ~695,000 (~ 6 Gb)
 - Length: ~9 kb mean (2 bp - 66 kb)
- Assembly using Unicycler (miniasm and racon)
(Li 2016, Vaser *et al* 2017, Wick *et al*, 2017)
- Output:
 - Contigs: ~1,700 (~ 104 Mb)
 - Length: ~ 61 kb mean (1.3 kb - 5.2 Mb)

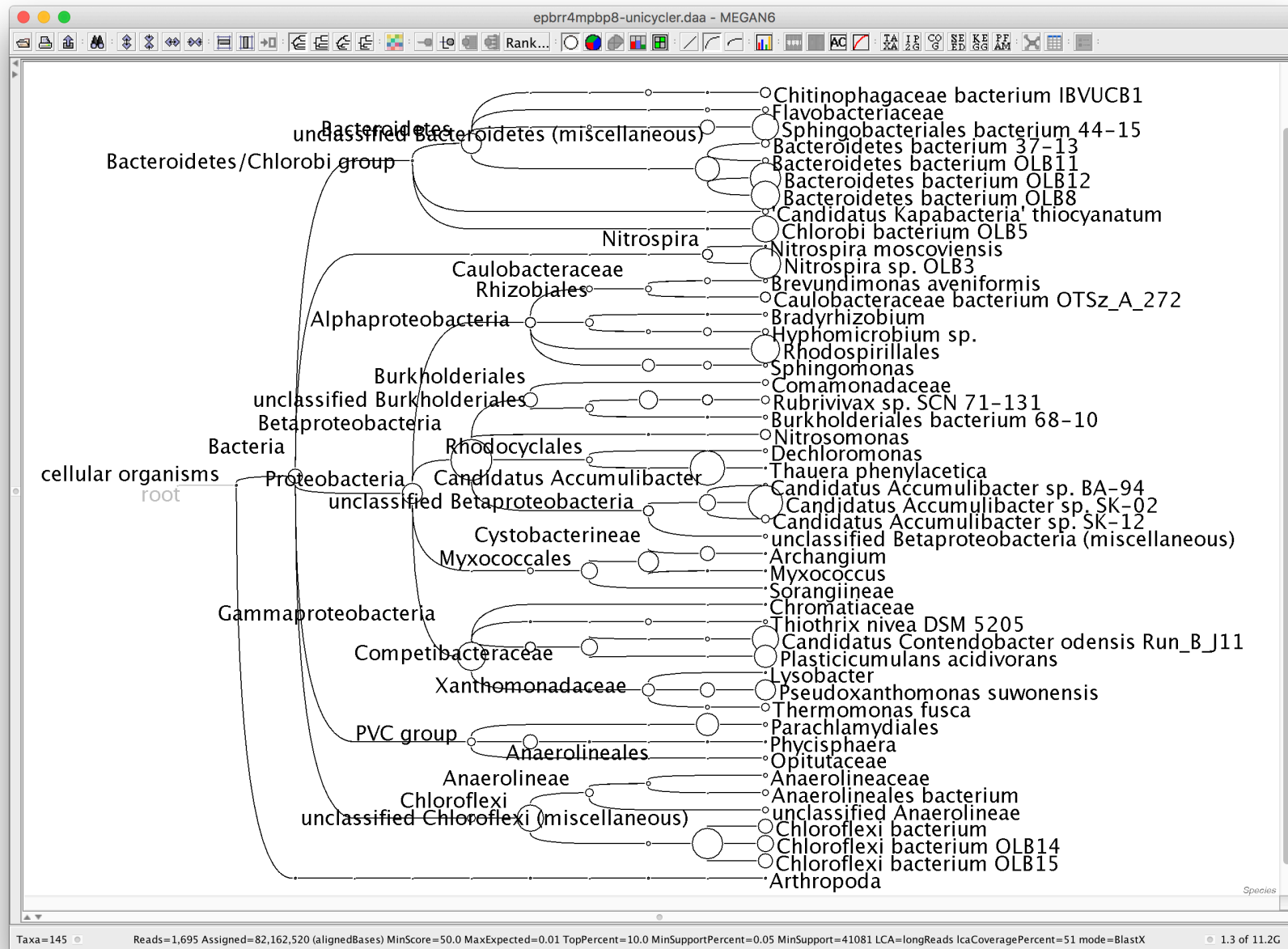
Bandage visualization of assembly graph



Bandage: Wick et al, 2015

Layout: Hachul S., Jünger M., 2007

Taxonomic binning of contigs



DIAMOND+MEGAN

Taxonomic bins $\geq 50\%$ complete

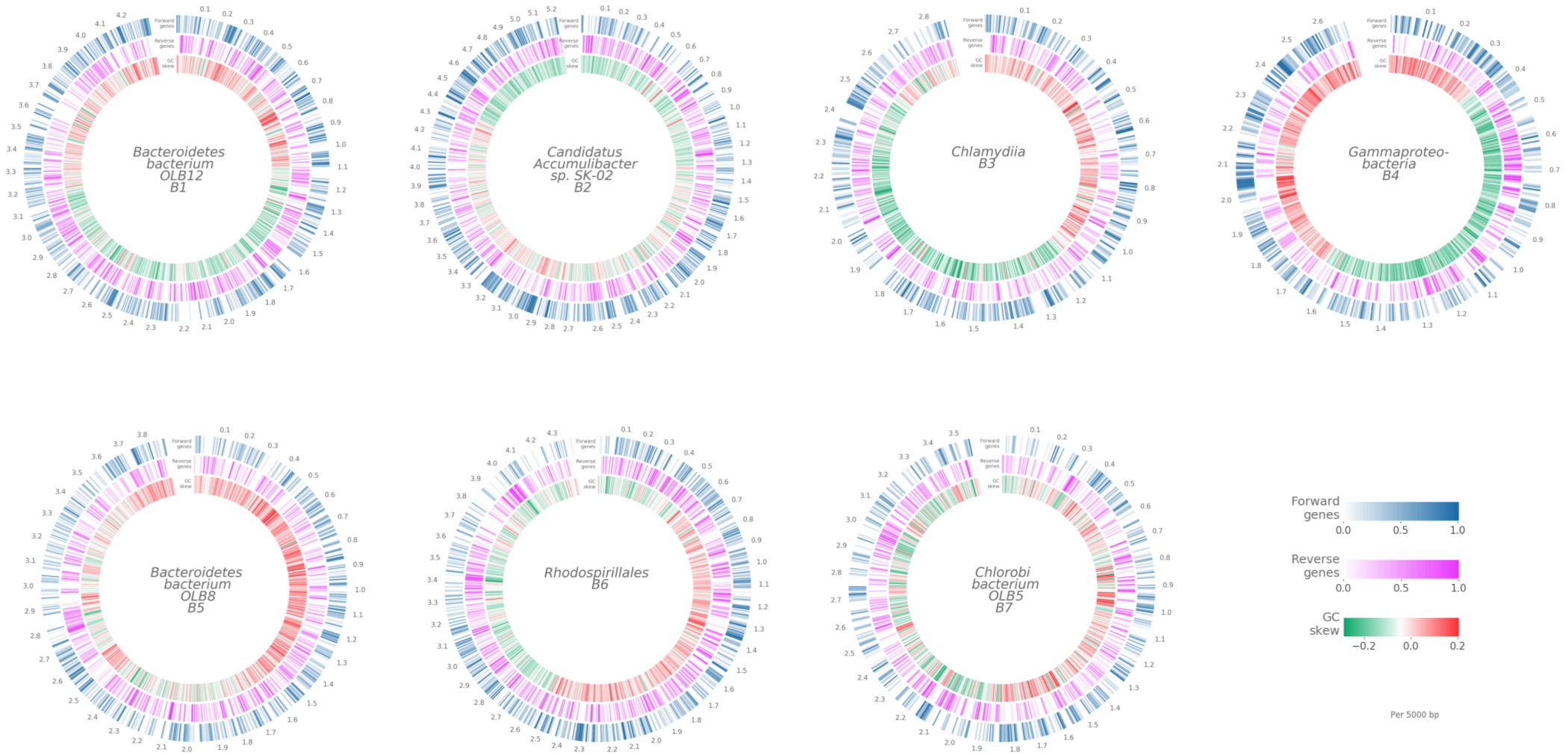
DIAMOND+MEGAN taxonomic bin		Unicycler contigs	Total (Mb)	Aligned (Mb)	Average coverage	CheckM		Prokka		
						Complete.	Contam.	rRNA	tRNA	CDS
High quality draft genomes:										
B1	<i>Bacteroidetes bacterium</i> OLB12	1	4.2	3.5	57.3	95%	0.1%	6	39	4,163
B2	<i>Candidatus Accumolibacter</i> SK-02	1	5.2	4.1	384.2	94%	0.6%	4	53	4,915
B3	<i>Chlamydiia</i> (class)	1	2.8	1.8	48.8	94%	2%	6	39	3,387
B4	<i>Gammaproteobacteria</i> (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	<i>Bacteroidetes bacterium</i> OLB8	1	3.8	3.0	52.1	93%	1%	6	37	3,394
B6	<i>Rhodospirillales</i> (order)	1	4.4	3.0	29.5	92%	0.5%	3	47	4,015
B7	<i>Chlorobi bacterium</i> OLB5	1	3.5	2.5	38.7	88%	1%	3	41	4,131
Medium quality draft genomes:										
B8	<i>Thauera</i> (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	<i>Sphingobacteriales bacterium</i> 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	<i>Bacteroidetes</i> (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	<i>Candidatus Contendobacter</i> 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	<i>Betaproteobacteria</i> (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	<i>Nitrospira</i> (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	<i>Chloroflexi</i> (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

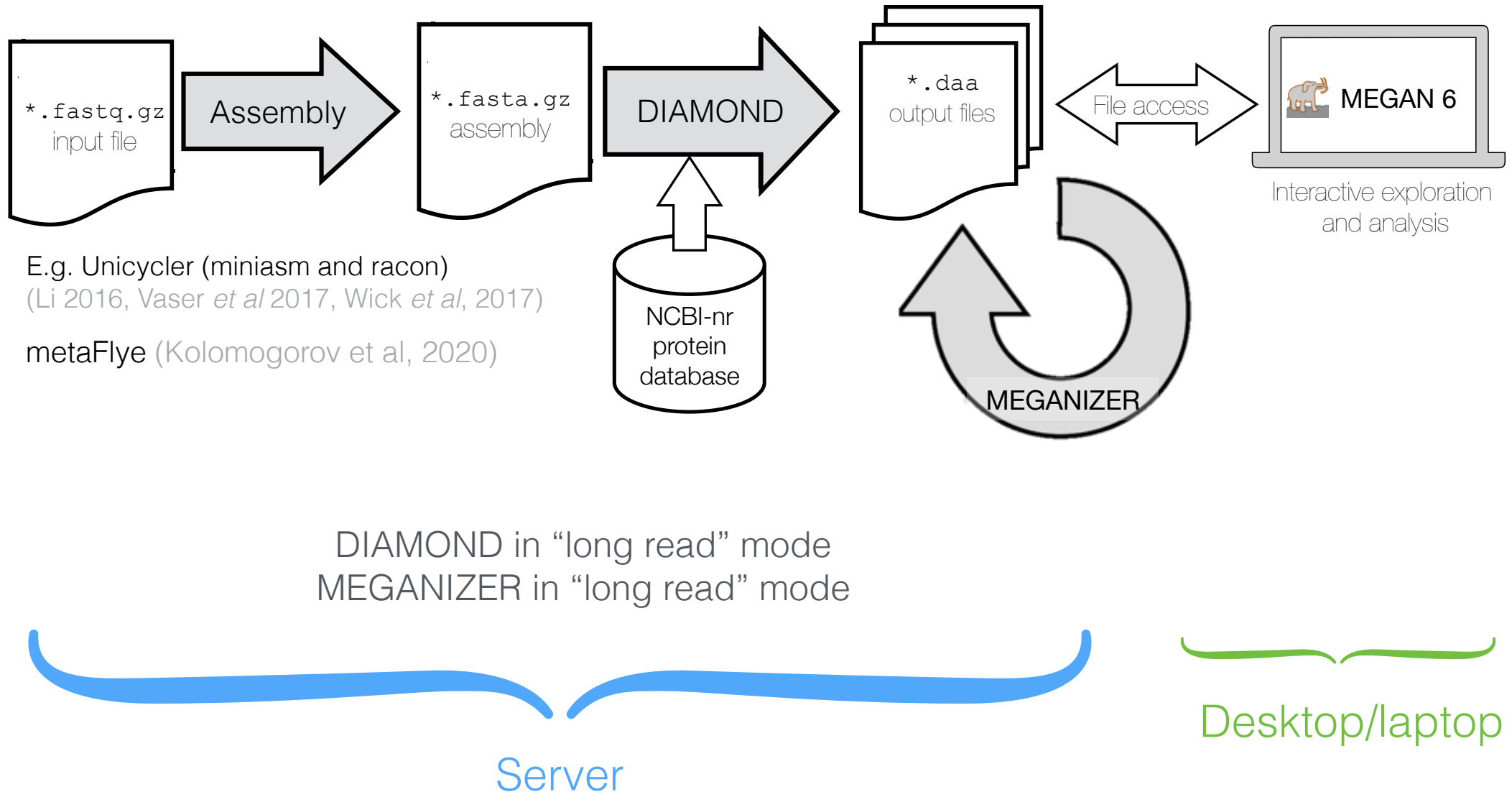
CheckM (Parks et al. 2014)

Prokka (Seemann, 2014)

Assembled chromosomes



Long-read analysis pipeline



Detailed protocols



PROTOCOL | Open Access | CC BY-NC-ND

DIAMOND+MEGAN: Fast and Easy Taxonomic and Functional Analysis of Short and Long Microbiome Sequences

Caner Bağcı, Sascha Patz, Daniel H. Huson

First published: 03 March 2021 | <https://doi.org/10.1002/cpz1.59>

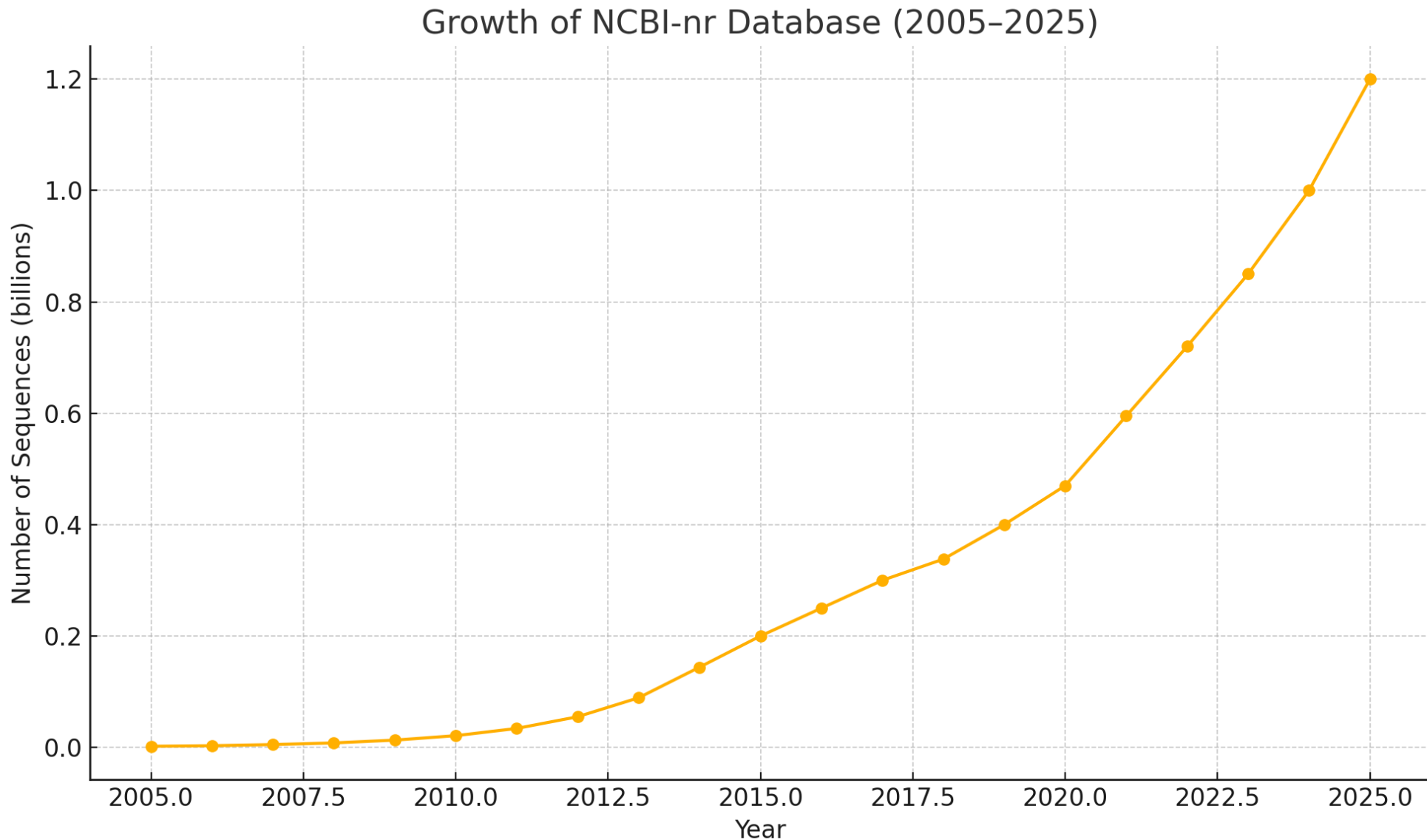
<https://doi.org/10.1002/cpz1.59>



Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- **Megan7**
- Hands-on session

NCBI-nr growing exponentially

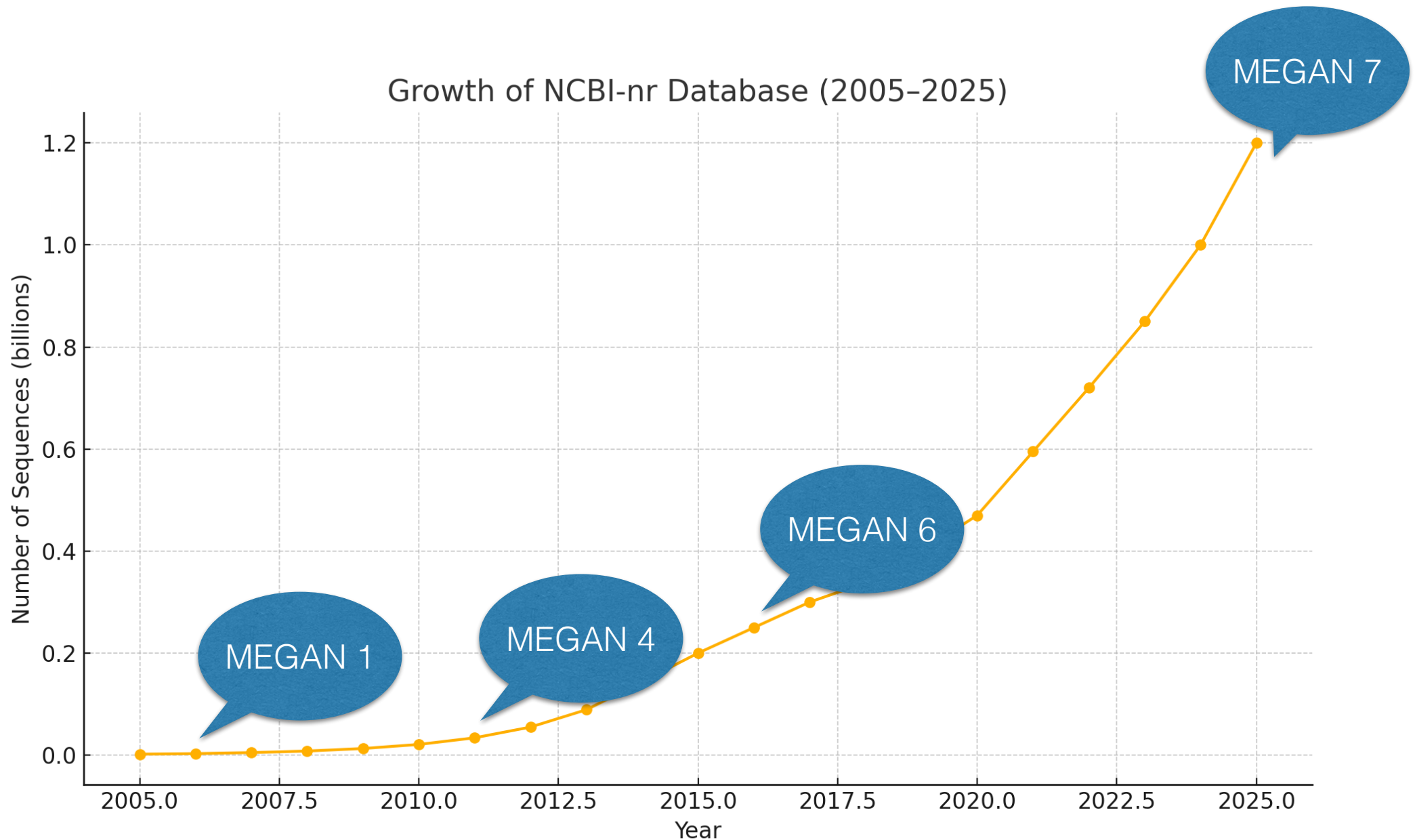




How to address this?

- Taxonomic analysis
 - Alignment-free (e.g. Kraken2):
 - k-mer based
- Functional analysis
 - Tiered approach (e.g. HUMAnN3):
 - Taxonomic pre-screening
 - Reduced databases
- MEGAN7 - use clustered versions of NCBI-nr and Uniref

Exponential DB increase



Replace NCBI-nr by:

- nr90 & nr50 - clustered at 90% & 50% sequence identity
- UniRef100, 90 and 50 - clustered UniRef
(high annotation quality)

More assignments, faster...

2 1

Classification rate compared to NCBI-nr on 10 datasets

Databases:	NCBI-nr	UniRef100	UniRef90	UniRef50	nr-90	nr-50
NCBI taxonomy	1.0	1.04	1.04	0.95	1.01	0.87
GTDB taxonomy	1.0	0.96	0.97	0.89	1.00	0.86
EggNOG	1.0	1.15	1.43	1.57	1.33	1.39
SEED	1.0	0.99	1.09	1.35	1.08	1.26
Speedup	1.0	1.4	4.0	16.8	3.1	17.7
Size	1.0	0.6	0.3	0.1	0.4	0.1

Outline

- Introduction to microbiome analysis
- Protein alignment against the NCBI-nr database
- Who is out there, what are they doing, how do they compare?
- MEGAN taxonomic and functional binning
- The DIAMOND+MEGAN pipeline
- Long-read metagenomics
- MEGAN7
- Hands-on session



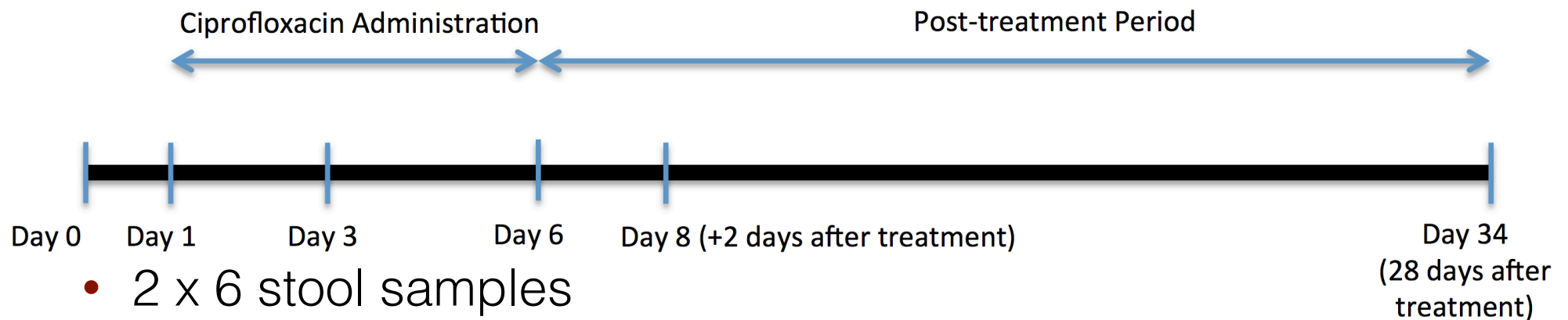
Hands-on session

- Run DIAMOND
- Run MEGANIZATION
- Explore using MEGAN

ASARI- Antibiotic resistance pilot study



- 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

Tutorial data

Subset data:

- Alice00, Alice01, Alice03, Alice06, Alice08, Alice34
- Bob00, Bob01, Bob03, Bob06, Bob08, Bob34
- 1 million reads each

Subset references:

- `tutorial-nr.gz` - nr50 proteins found in 12 files
(Works *only* for tutorial, *not* for other data)



Tutorial-nr file

- While NCBI-nr50 is small enough to be used on a good 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
 - MEGAN also requires a “mapping db file”
 - `megan-map-tutorial.db`

(Works *only* for tutorial, *not* for other data)



Step I - Protein alignment

- Will use DIAMOND (other options: LAST, mmseq)
 - designed for metagenomics (Buchfink et al, 2015)
- Need to:
 - Install DIAMOND (already done )
 - Download reference sequences (already done )
 - Run DIAMOND to build index
 - Run DIAMOND on fastq.gz files

Tutorial instructions

- Open a terminal and type:

```
cp -r /mnt/data/Huson_data/tutorial-DM ~  
cd tutorial-DM  
mkdir out  
ls  
data megan-map-tutorial.db out tutorial-nr.gz
```

terminal

Step I - DIAMOND index

- Build a DIAMOND index:

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

terminal

- Note: Using `tutorial-nr.gz`, due to time restrictions
(Works *only* for tutorial, *not* for other data)

```
ls  
data megan-map-tutorial.db out tutorial-nr.dmnd tutorial-nr.gz
```

terminal

Step I - Run DIAMOND

- Run DIAMOND on one input FASTQ file:

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

terminal

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

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

terminal



Step I - Run DIAMOND

- For full size datasets, DIAMOND alignment (and subsequent meganization) is run on a server
- The 12 small datasets against `tutorial-nr.gz` should take less than 10 minutes
- 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>

Step I - Run DIAMOND

- This should produce 12 files:

```
training26@VCTL-MET-U20-83:~/tutorial-DM$ ls out
Alice00-1mio.fq.daa  Alice06-1mio.fq.daa  Bob00-1mio.fq.daa  Bob06-1mio.fq.daa
Alice01-1mio.fq.daa  Alice08-1mio.fq.daa  Bob01-1mio.fq.daa  Bob08-1mio.fq.daa
Alice03-1mio.fq.daa  Alice34-1mio.fq.daa  Bob03-1mio.fq.daa  Bob34-1mio.fq.daa
training26@VCTL-MET-U20-83:~/tutorial-DM$
```

Outline

- Introduction to microbiome analysis
- Step 0: Installation
- Step 1: DIAMOND alignment against protein database
- Step 2: MEGANization of reads and alignments
- Step 3: MEGAN interactive analysis






Step 2: Meganization

- Ran 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.



Step 2: Meganization

- Requires MEGAN
 - metagenome analyzer 6 (Huson et al, 2016)
- Need to:
 - Install MEGAN (already done )
 - Download reference file (already done )
 - Download mapping db file (already done )

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
```


Step 2: Meganization

- Meganize one DIAMOND file:

```
export megan=/mnt/data/Huson_data/megan  
  
$megan/tools/daa-meganizer \  
-i out/Alice00-1mio.daa \  
-mdb megan-map-tutorial.db
```

terminal

- Meganize *all* DIAMOND files:

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

terminal



Step 2: Meganization

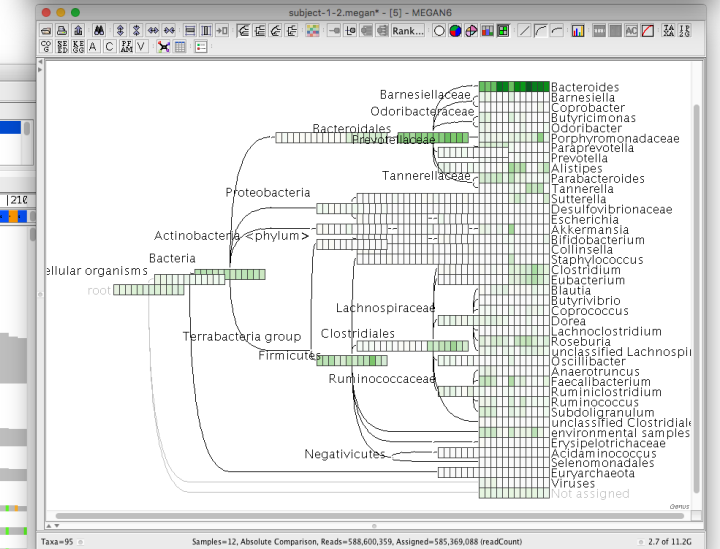
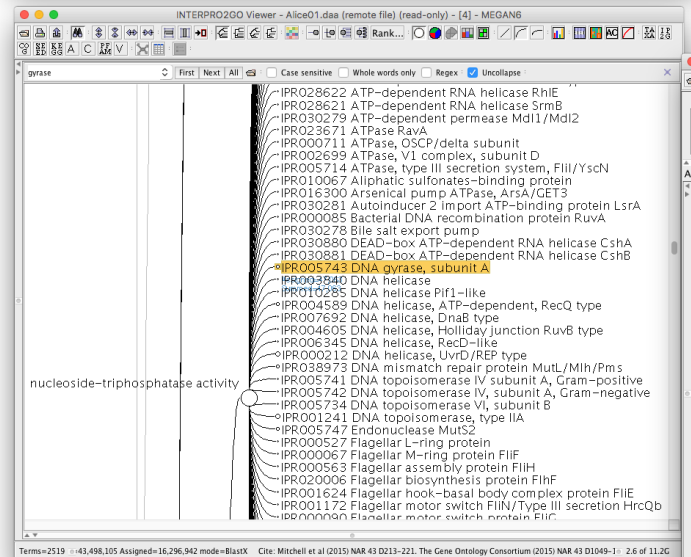
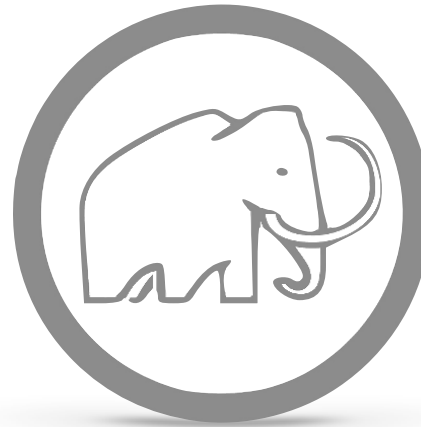
- Meganization does not produce any new files. It prepares the .daa files for opening in MEGAN
- If you failed to meganize the 12 files, you can download the meganized files here:

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

Outline

- Introduction to microbiome analysis
- Step 0: Software setup
- Step 1: DIAMOND alignment against protein database
- Step 2: MEGANization of reads and alignments
- Step 3: MEGAN interactive analysis

PCoA analysis



Comparative analysis

n, 2025

Step 3: MEGAN analysis

- Launch MEGAN by typing:

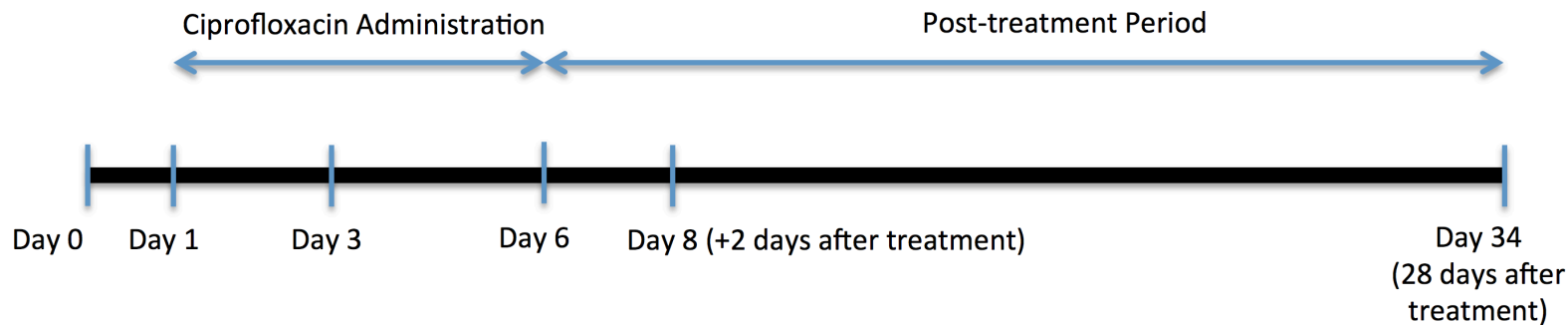
```
export megan=/mnt/data/Huson_data/megan  
$megan/MEGAN
```

terminal

- Open individual files with the File->Open... item
 - Find them in directory `tutorial-DM/out`
- Compare files using the File->Compare... item

Alice and Bob- short reads

Alice and Bob, 6 time points each



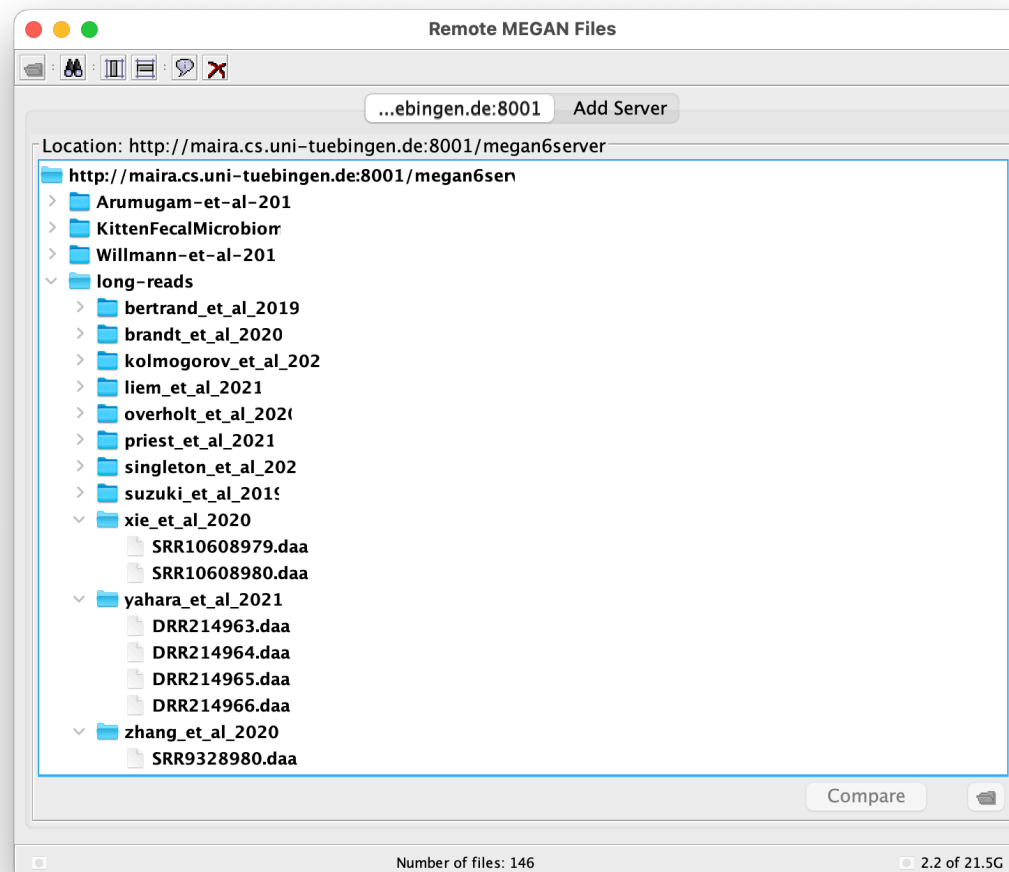
- Each subsampled to 1 mio reads.
- `data/Alice00-1mio.fq.gz` etc



Alice and Bob- short reads

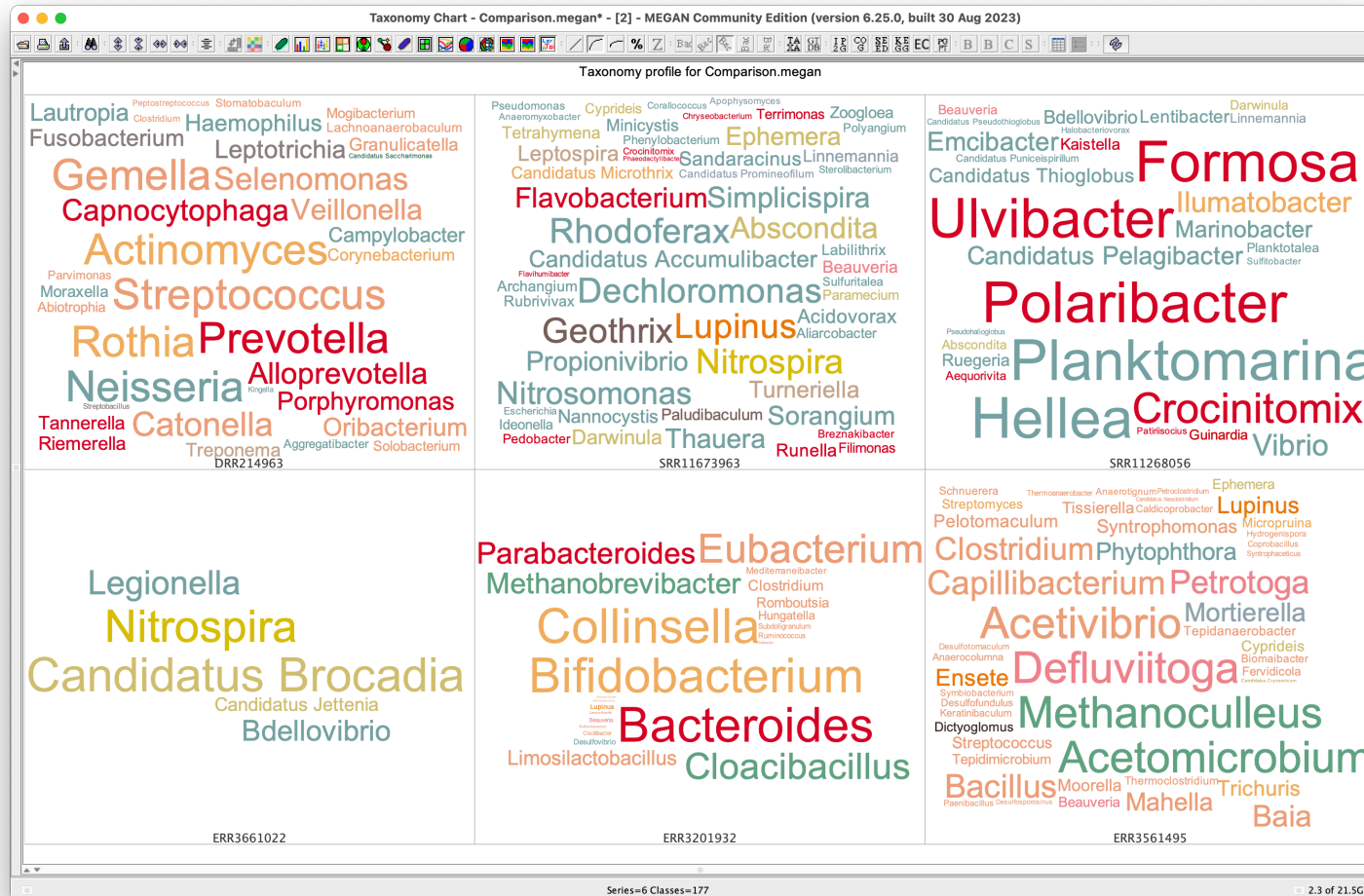
- Tutorial tasks:
 - confirm that these are gut samples - should be dominated by Bacteroidota, Firmicutes and Proteobacteria.
 - Open all twelve files together in a comparison document and add the provided metadata.
 - Confirm that the taxonomic profiles of either subject changes during the course of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

- MeganServer serves MEGAN files to the web
- The default server provides access to a number of published metagenome datasets:



MeganServer

- Here are six long-read datasets:



- Can you match them to: biogas plant, ground water, human gut, oral, sea water, waste water ?



Thank you!

Joint work with:

- Caner, Baci, Anupam Gautam, Timo Lucas, Sascha Patz, Wenhuan Zeng
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

