Taxonomic assignment

Workshop in marine metagenomics

Tromsø November 2018

www.elixir-europe.org

Overview of this talk

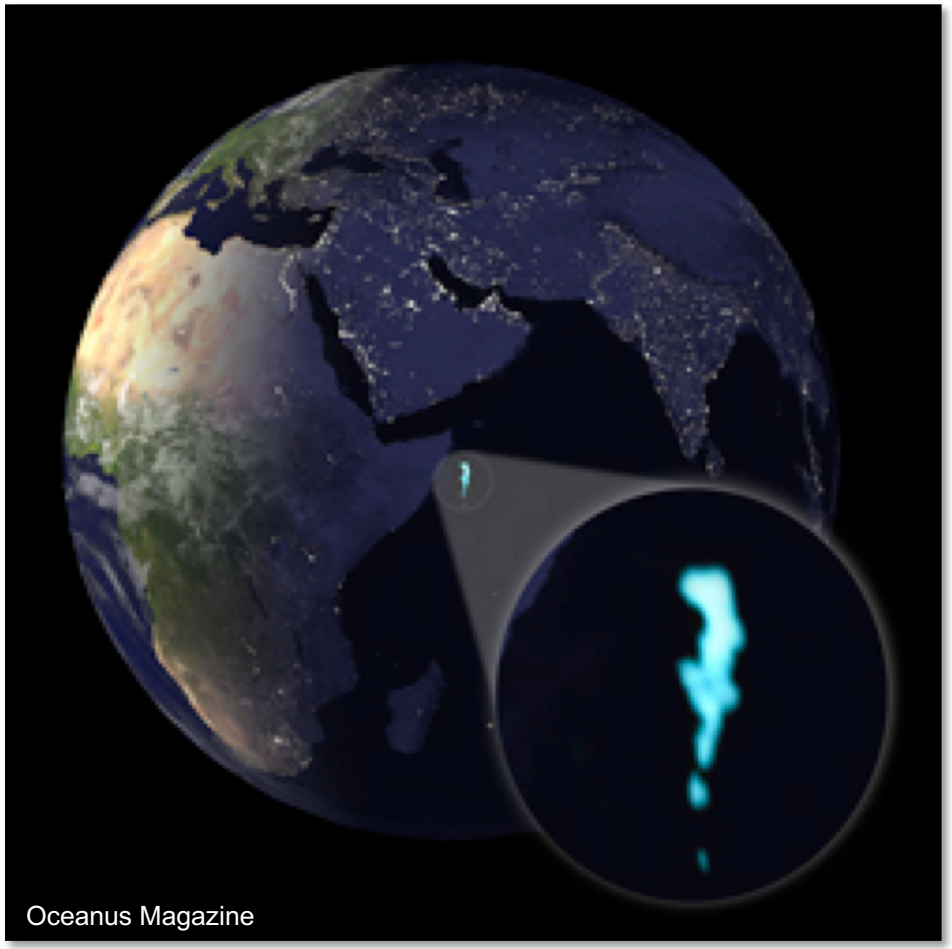
Microorganisms

Microbiome

Metagenome

Taxonomic analysis of metagenomes

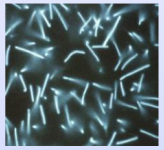
What's in the databases



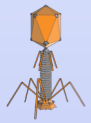
Microorganisms – too small to see

Domain (marine environment)

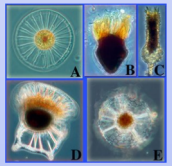
Bacteria (archaea/eubacteria)
Species
–>10^9 (billion)?
Genome size
0,6 - 12 Mb (5 Mb)



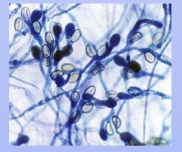
Virus (bacteriophages)
species:
–> 10^6 (million)?
genome size
3 200 bp - 1.2 Mb (50 Kbp)

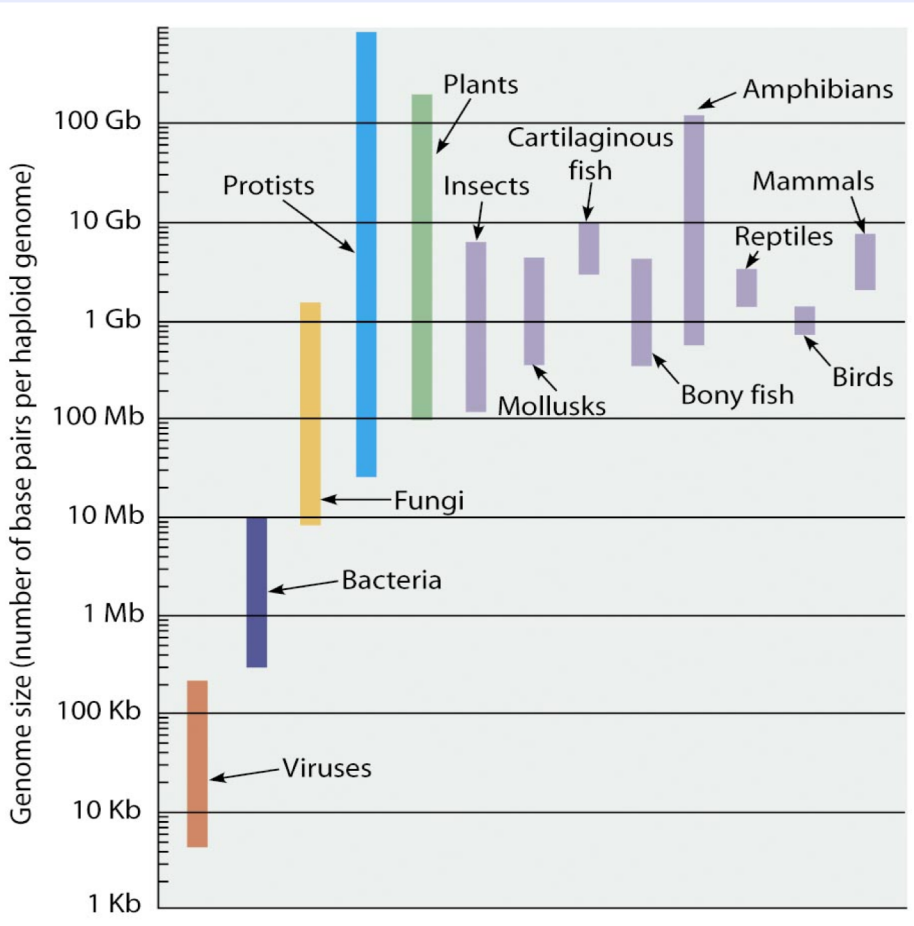


Protist (eukaryote microbes)
Species
Estimate >> 300 000?
Genome size
>>35 Mb - 215 Gb (10 000 Mb)



Fungi (yeast/mold)
Species
estimate >150 000?
Genome size
> 10Mb - 1,2 Gb (100 Mb)





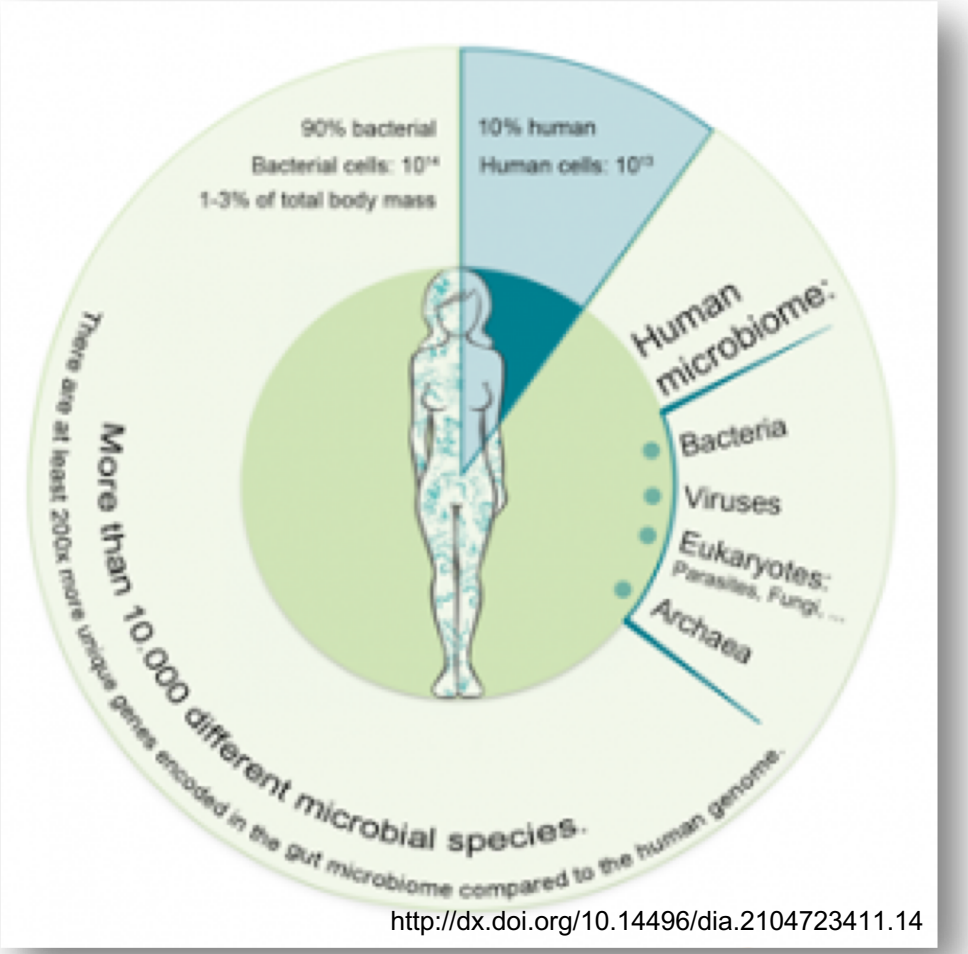
Enormous variety in appearance and capabilities

Microbes and host

Bacterial cells in the human body outnumber human cells 10 to one

Human genome = ~20 000 genes

Bacterial genome on human = ~2 - 20 000 000 genes

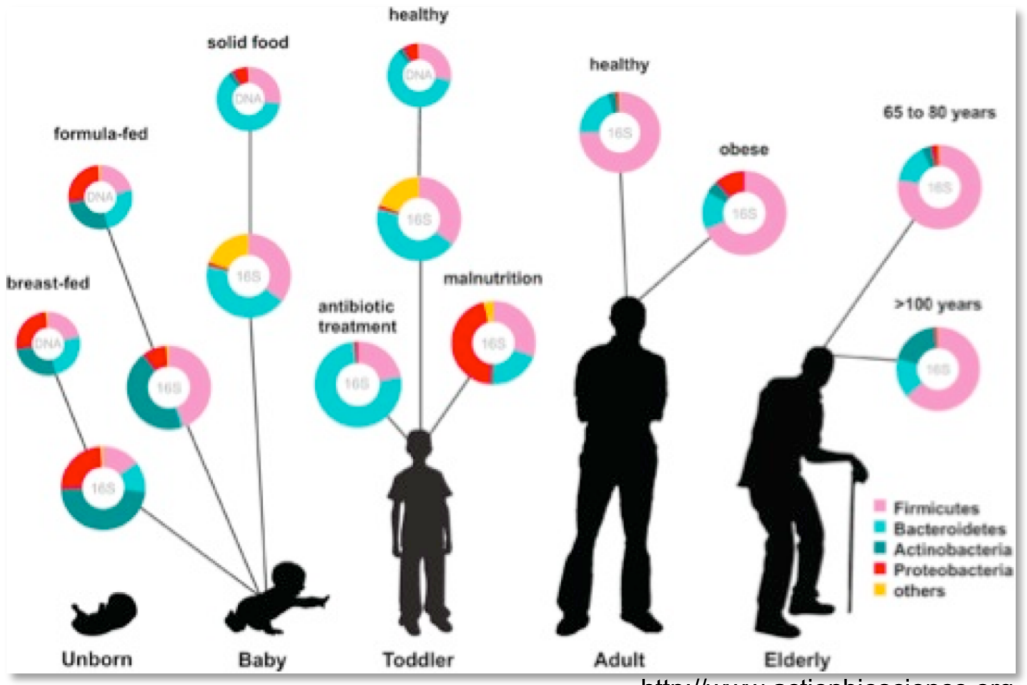


http://dx.doi.org/10.14496/dia.2104723411.14

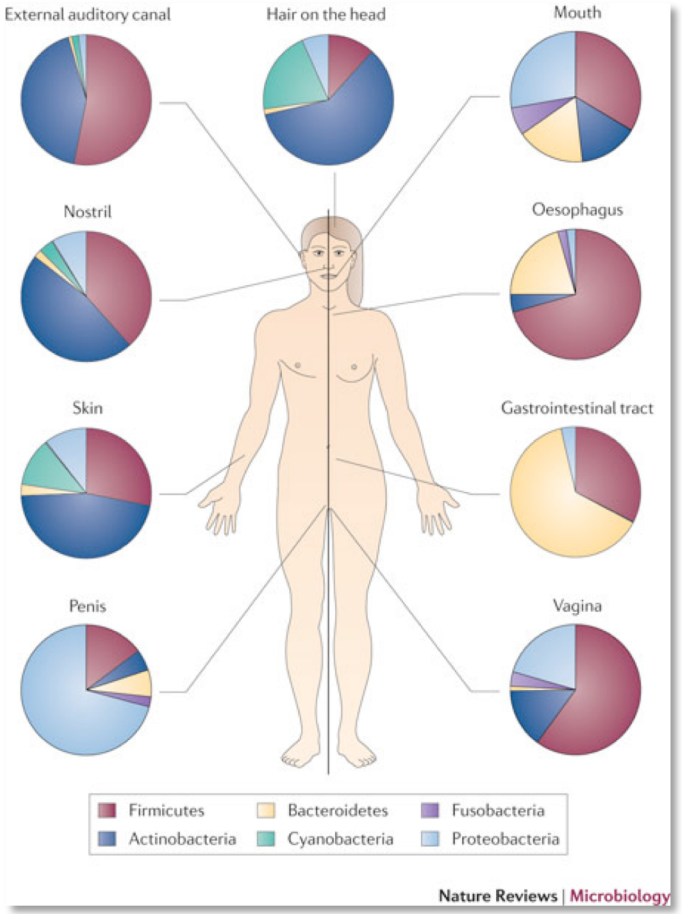
Microbiome are the microorganisms in a particular environment

Including the body or a part of the body

The human microbiome change over time

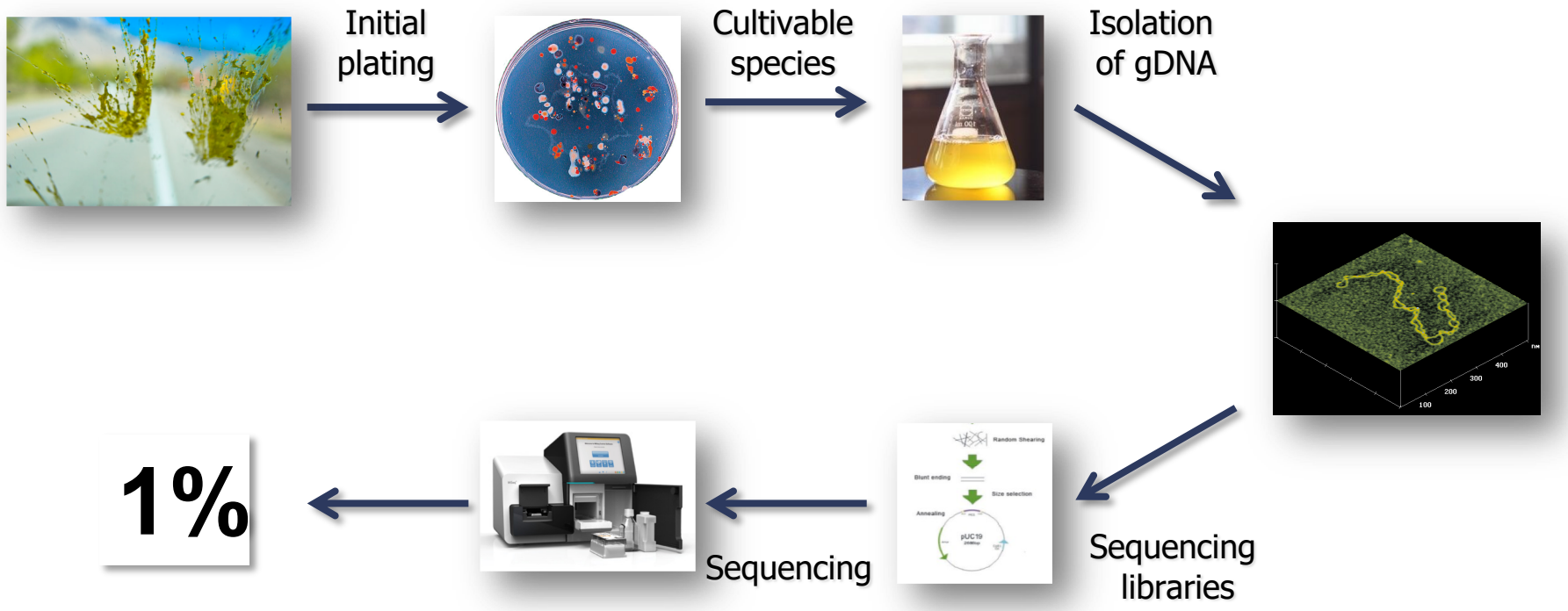


http://www.actionbioscience.org



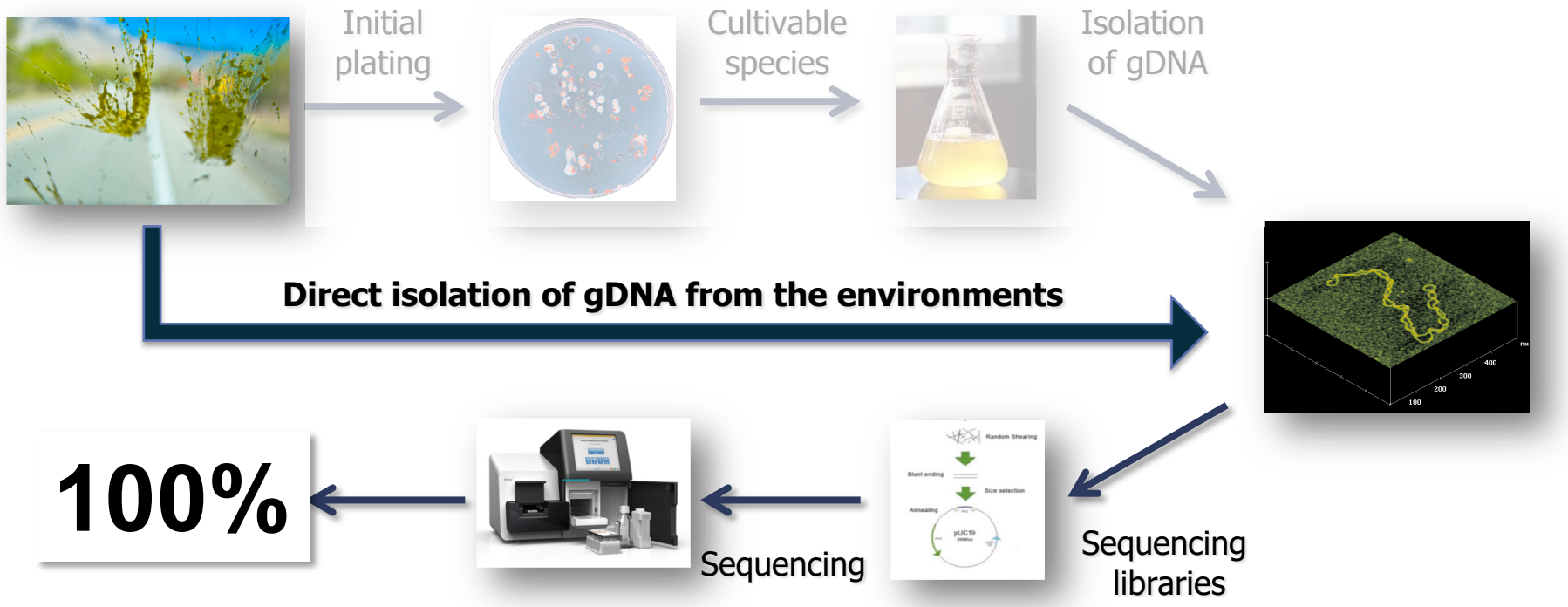
Recap - How do we study microbiomes?

Cultivation: Only 1% in most environmental samples



Recap - How do we study microbiomes?

Cultivation: Only 1% in most environmental samples



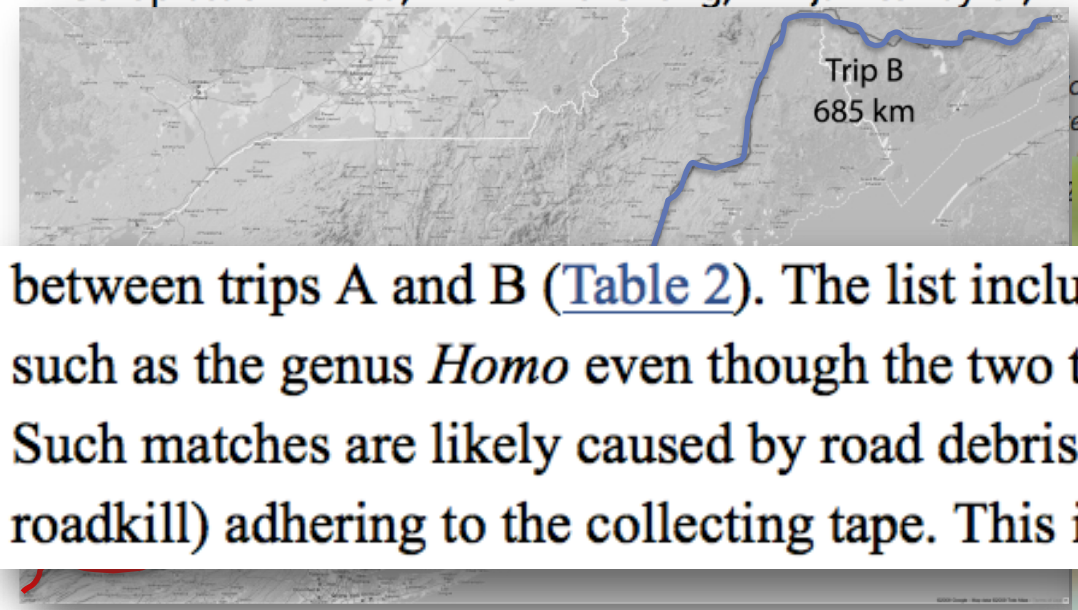
A "typical"☺ metagenomic study

Resource

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond,[1,2,6,9] Samir Wadhawan,[3,6,7] Francesca Chiaromonte,[4] Guruprasad Ananda,[1,3] Wen-Yu Chung,[1,3,8] James Taylor,[1,5,9] Anton Nekrutenko,[1,3,9]

cs, School of Medicine University of
es, Penn State University, University Park,



between trips A and B (Table 2). The list included unexpected entries such as the genus *Homo* even though the two trips were uneventful. Such matches are likely caused by road debris (which often includes roadkill) adhering to the collecting tape. This illustrates, at least at genus

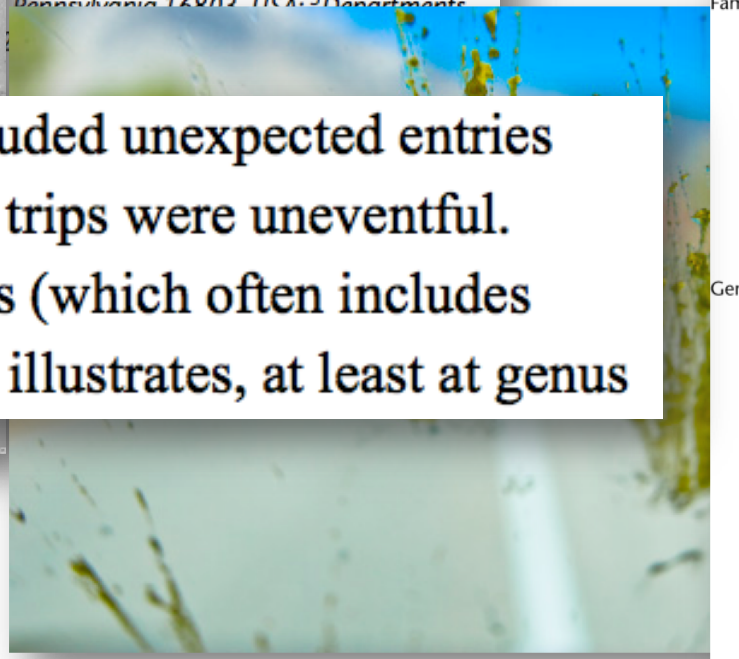


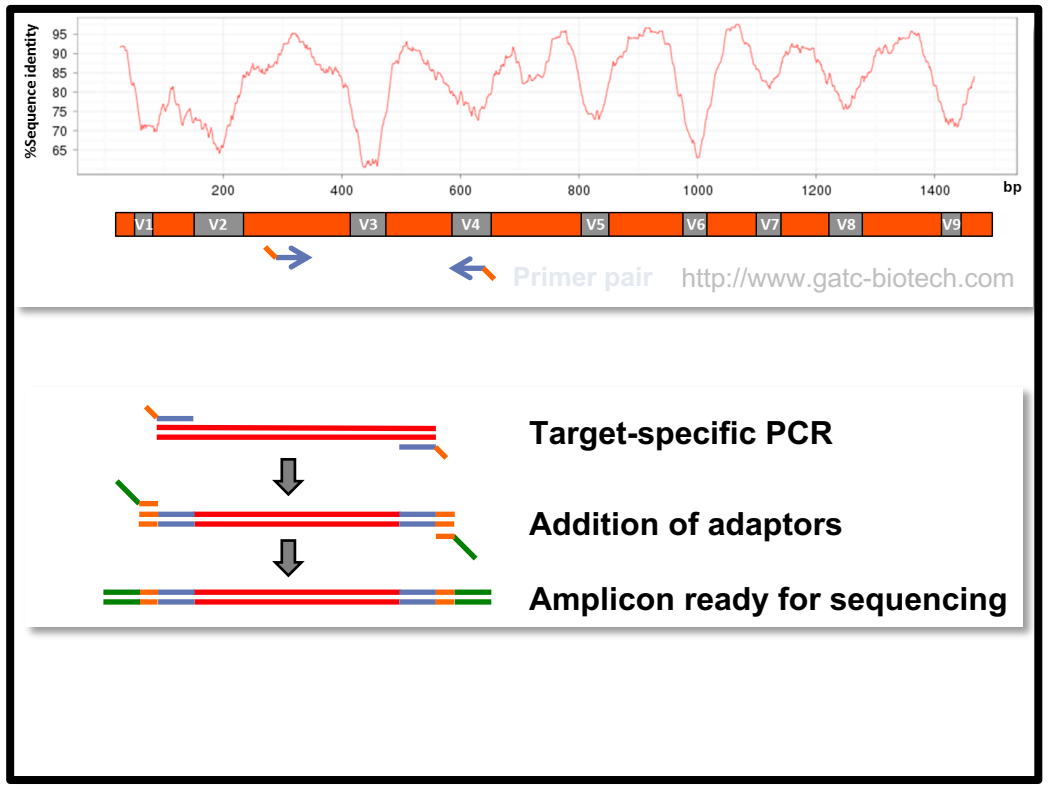
Table 2. Taxa with significant (at 1% level) differences in read abundance between trip A and trip B

Rank	Name	Trip A	Trip B
Phylum	Arthropoda	711	1531
	Chordata	300	272
	Cnidaria	10	87
	Firmicutes	12,927	5623
	Proteobacteria	45,946	24,663
Class	Bacilli	10,748	4004
	Betaproteobacteria	228	45
	Clostridia	2178	1616
	Gammaproteobacteria	44,934	24,413
	Hydrozoa	10	87
	Insecta	711	1516
	Mammalia	294	256
Order	Aeromonadales	540	21
	Bacillales	83	58
	Clostridiales	2178	1615
	Diptera	296	350
	Enterobacteriales	41,174	23,729
	Hemiptera	383	1027
	Hydroida	10	87
	Lactobacillales	10,643	3943
	Primates	112	10
	Pseudomonadales	1792	408
	Rhodospirillales	56	1
Family	Aeromonadaceae	540	21
	Aphididae	382	1016
	Clostridiaceae	2170	1608
	Culicidae	86	64
	Drosophilidae	32	95
	Enterobacteriaceae	41,172	23,729
	Enterococcaceae	706	1512
	Hominidae	97	6
	Hydridae	10	87
	Lactobacillaceae	5837	209
	Leuconostocaceae	2978	1498
	Pseudomonadaceae	1703	391
	Streptococcaceae	928	545
Genus	*Acyrthosiphon*	381	995
	Aeromonas	540	21
	Anopheles	80	45
	Anopheles	80	1
	Buchnera	9	59
	Clostridium	2170	1607
	Drosophila	31	94
	Enterobacter	4142	5507
	Enterococcus	706	1511
	Erwinia	2	240
	Homo	96	4
	Hydra	10	87
	Klebsiella	15,169	1695
	Lactobacillus	5740	167
	Lactococcus	809	509
	Leuconostoc	2971	1496
	Photorhabdus	57	1
	Providencia	123	3
	Pseudomonas	1648	390

Two methods for performing metagenomic studies

Amplicon sequencing (16S rRNA) and random sequencing

16S rRNA amplicon sequencing



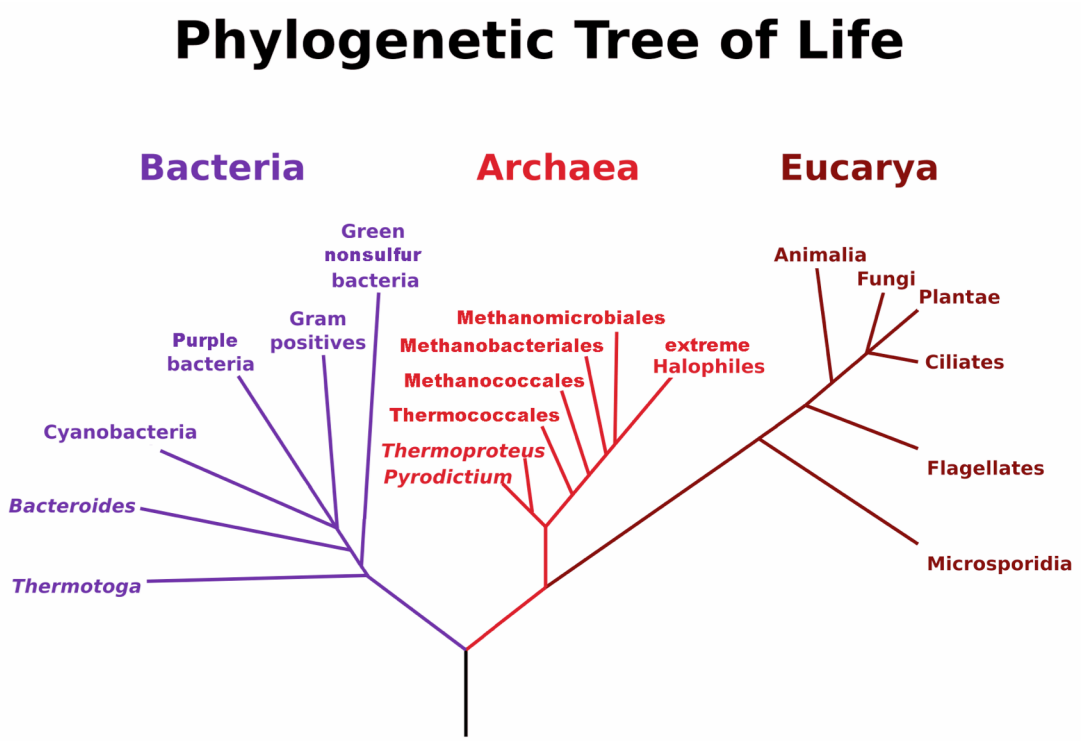
Random sequencing

Amplicon vs random sequencing

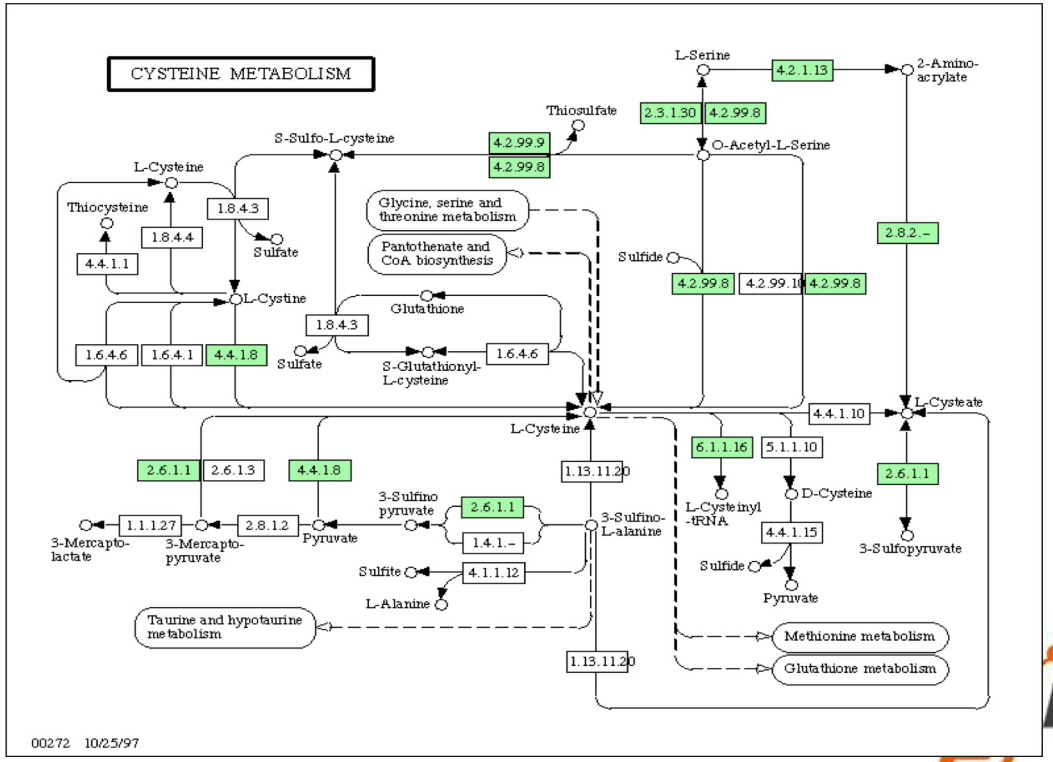
It depends on what you want to know

Main difference: taxonomic profile vs taxonomic and functional profile

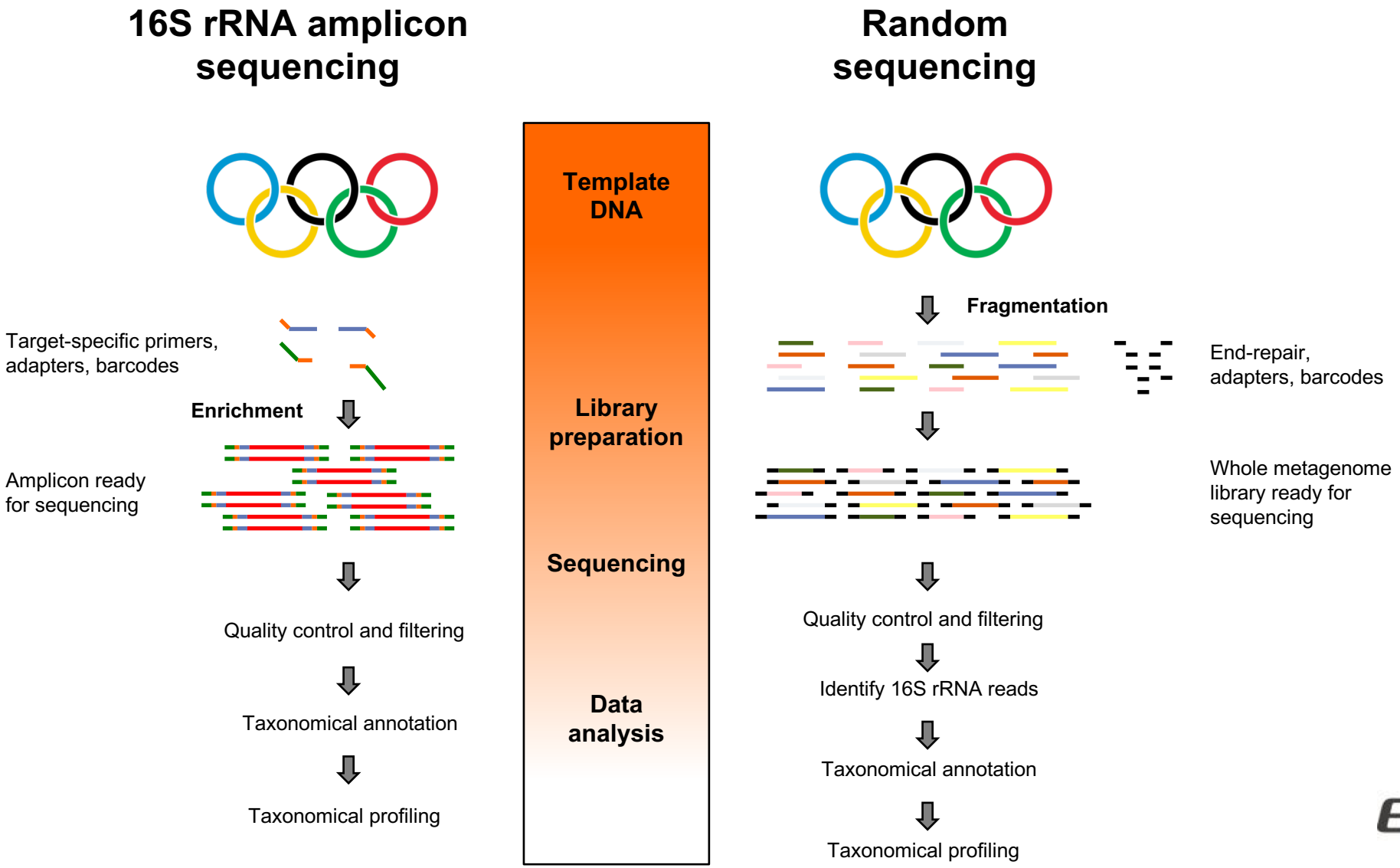
16S rRNA amplicon sequencing



Random sequencing



Amplicon vs random sequencing



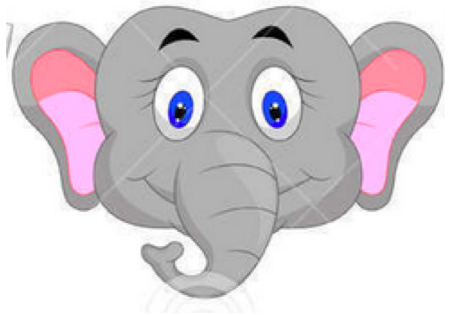
Amplicon vs random sequencing – pros & cons

	16S amplicon	Random
Analysis of large number of samples	pro	con
Depth - resolution	pro	con
Computational resources (and skills)	pro	con
Expenses	pro	con
PCR amplification bias	con	pro
Discovery of new bacterial genes and genomes	con	pro
Simultaneous study of several domains	con	pro

How is taxonomic classification done?

A sequence is basically representing a specie (taxa)

Sample



AGTCCAGGTAACGTTACAACG

How is taxonomic classification done?

Compare your sample against a database of known species

Sample

Compare

Database

How is taxonomic classification done?

Compare your sample against a database of known species

Sample

AGTCCAGGTAACGTTACAACG
GTTACAACAGCCTGAAGCCAC
CCAATTTTCGTGCAATTTACAA
GAAGCCACAGCAGTGCAGTTA

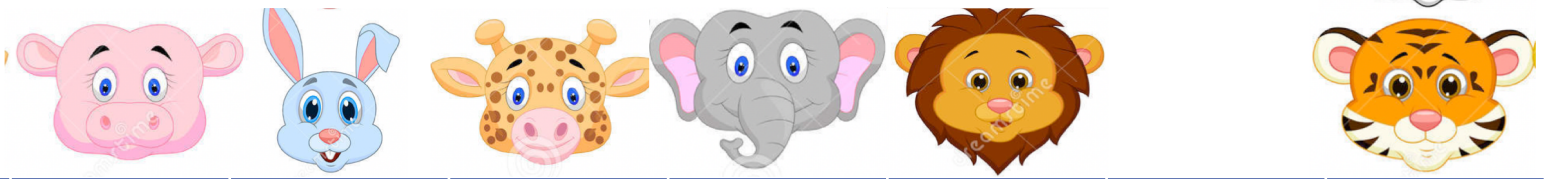
Compare

Database

Create a taxonomic profile

Quantify occurrences

Unknown



Sample	4	2	5	5	6		2

Compare taxonomic profiles

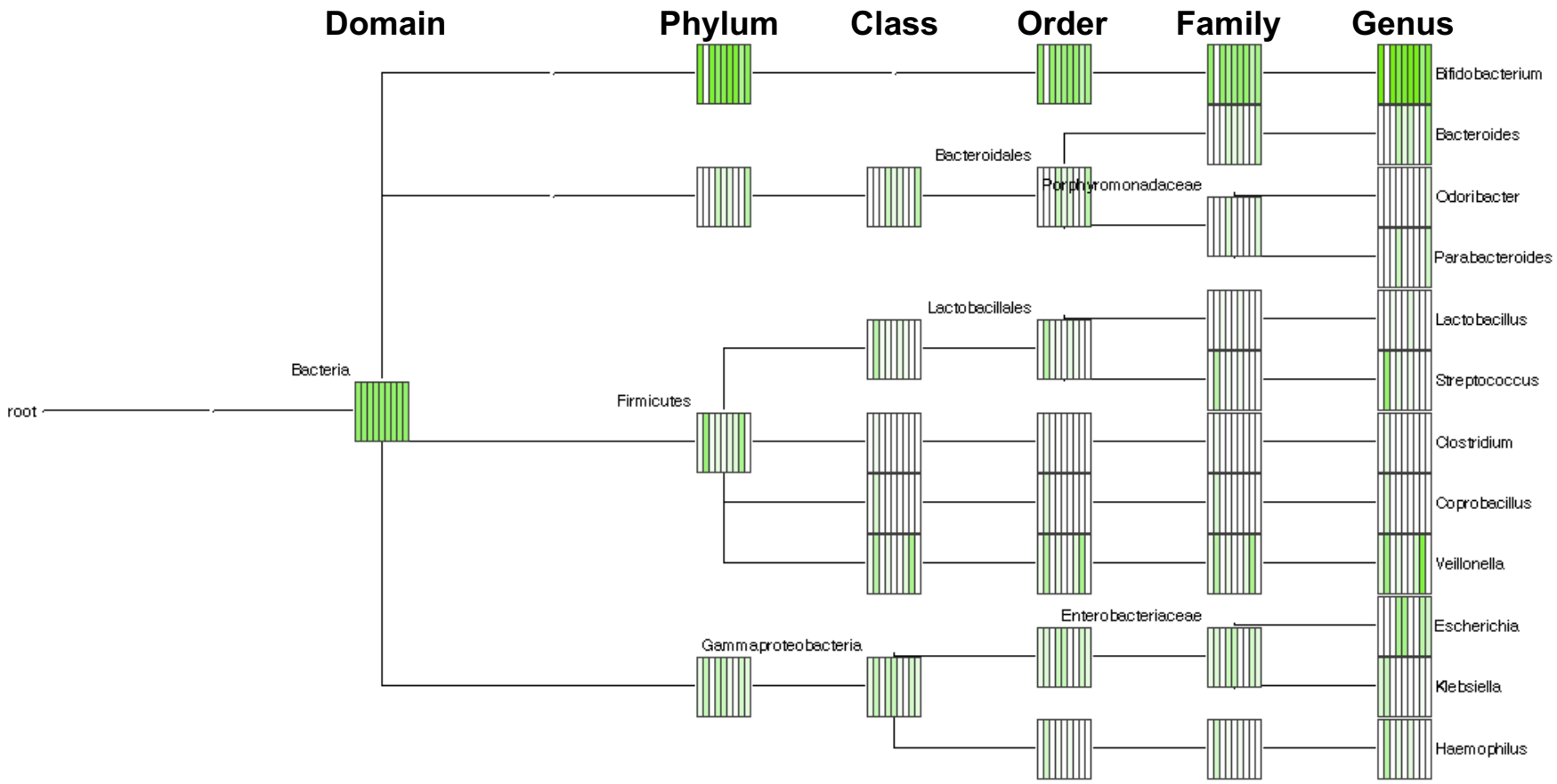
Compare two or more samples

Unknown

Sample	4	2	5	5	6		2
Sample 2	1	10	0	2	0		3

The taxonomy of species that contain highly similar sequences will be more difficult to resolve

When reads are too similar, they are assigned at higher levels of the taxonomy tree



Comparison of methods and tools

16S rRNA amplification differences lead to biased estimates of relative abundance

This can give an over-representation or under-representation of sequences in the some genera

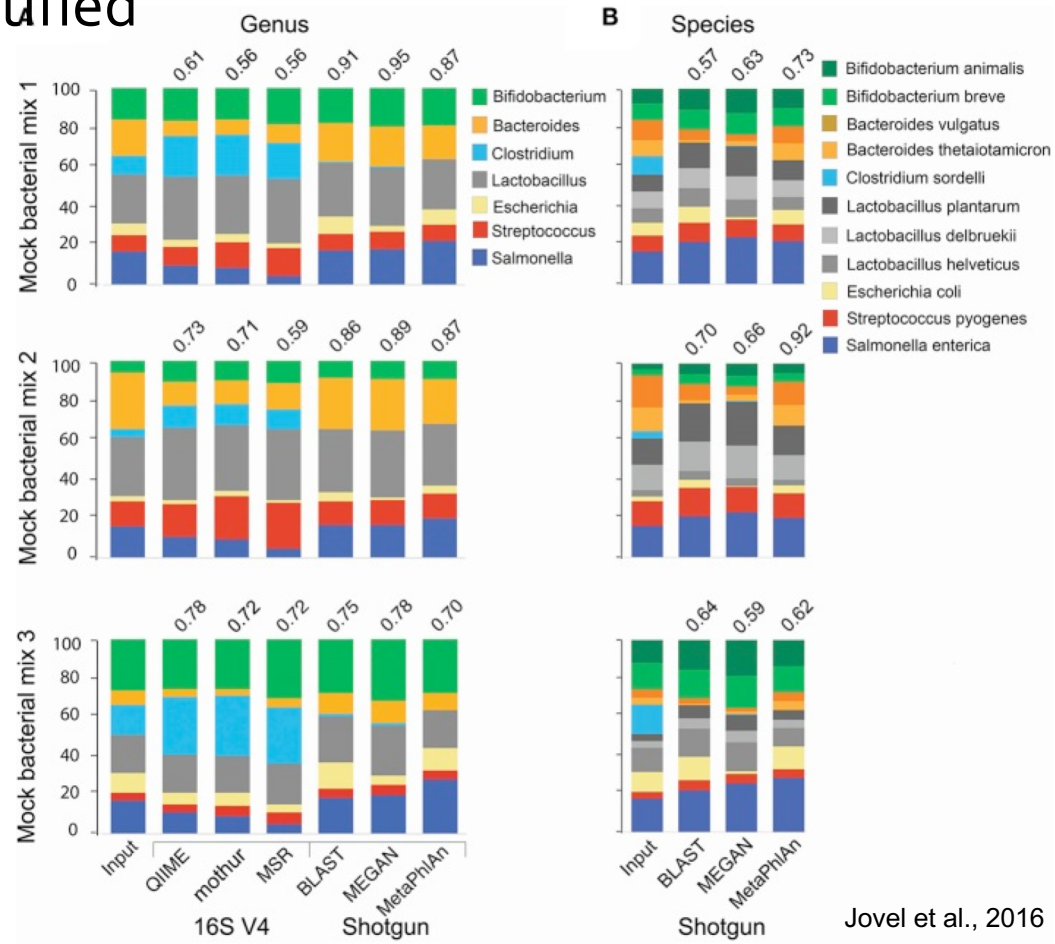
Eg. Clostridium and Lactobacillus contain sequences that are perfectly complementary to the primers used for amplification

Sequences in the Enterobacteriaceae family and the Clostridiales order poorly resolves using the 16S V4 or V3-V4 regions

Jovel et al., 2016

Random sequencing may identify taxa not amplified by 16S primers

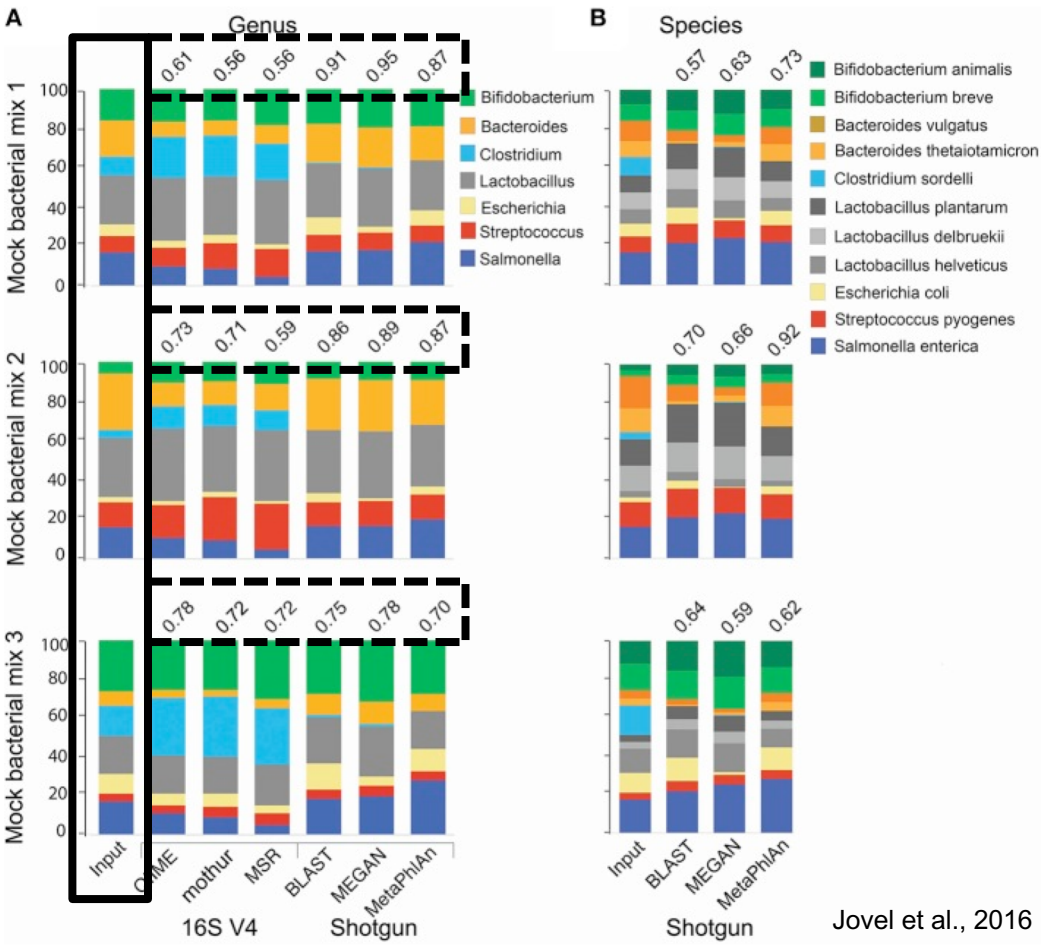
16S primers are not universal, in metagenomic data taxa not amplified by 16S primers may be identified



Jovel et al., 2016

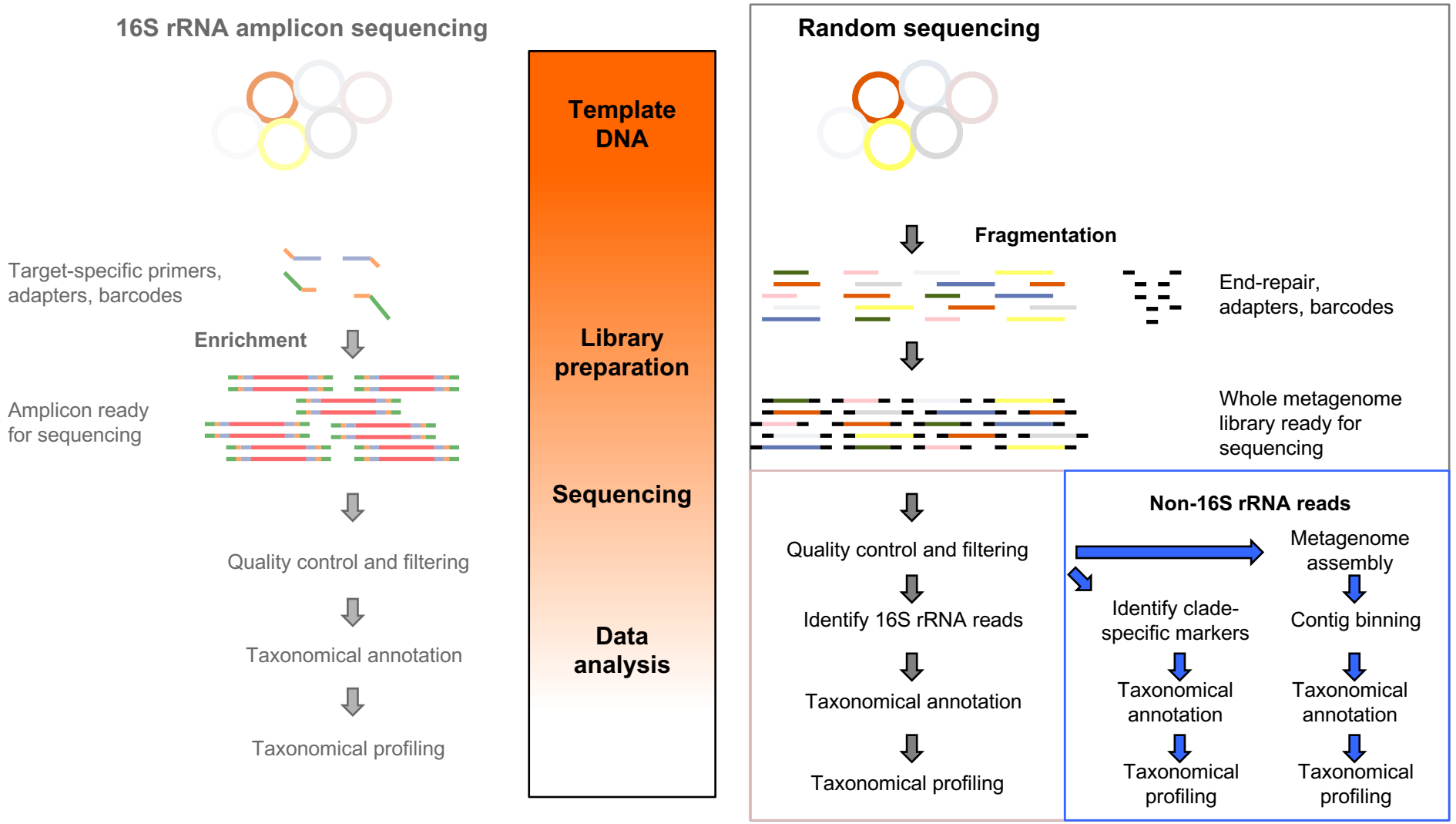
Random sequencing is more accurate to to reconstruct the taxonomic profile

The higher number = the better correlation (Pearson correlation coefficient)



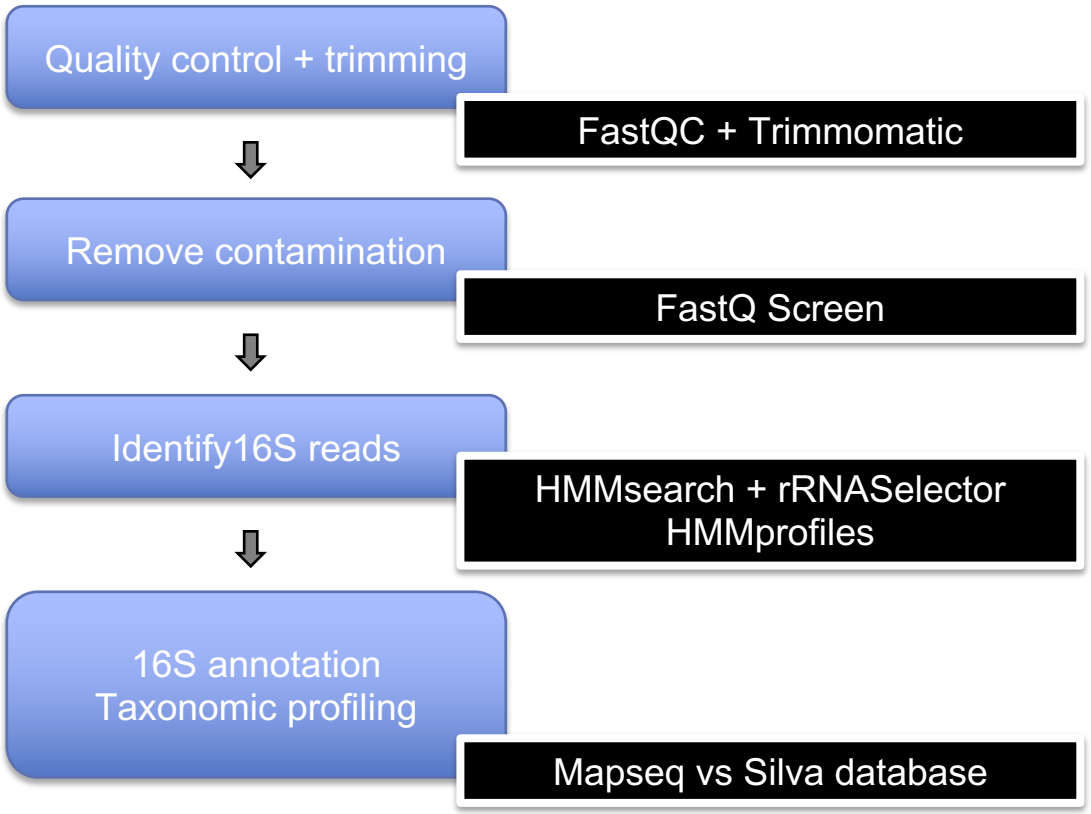
Jovel et al., 2016

Taxonomic profiling with random sequencing data



Random sequencing – Identifying 16S rRNA

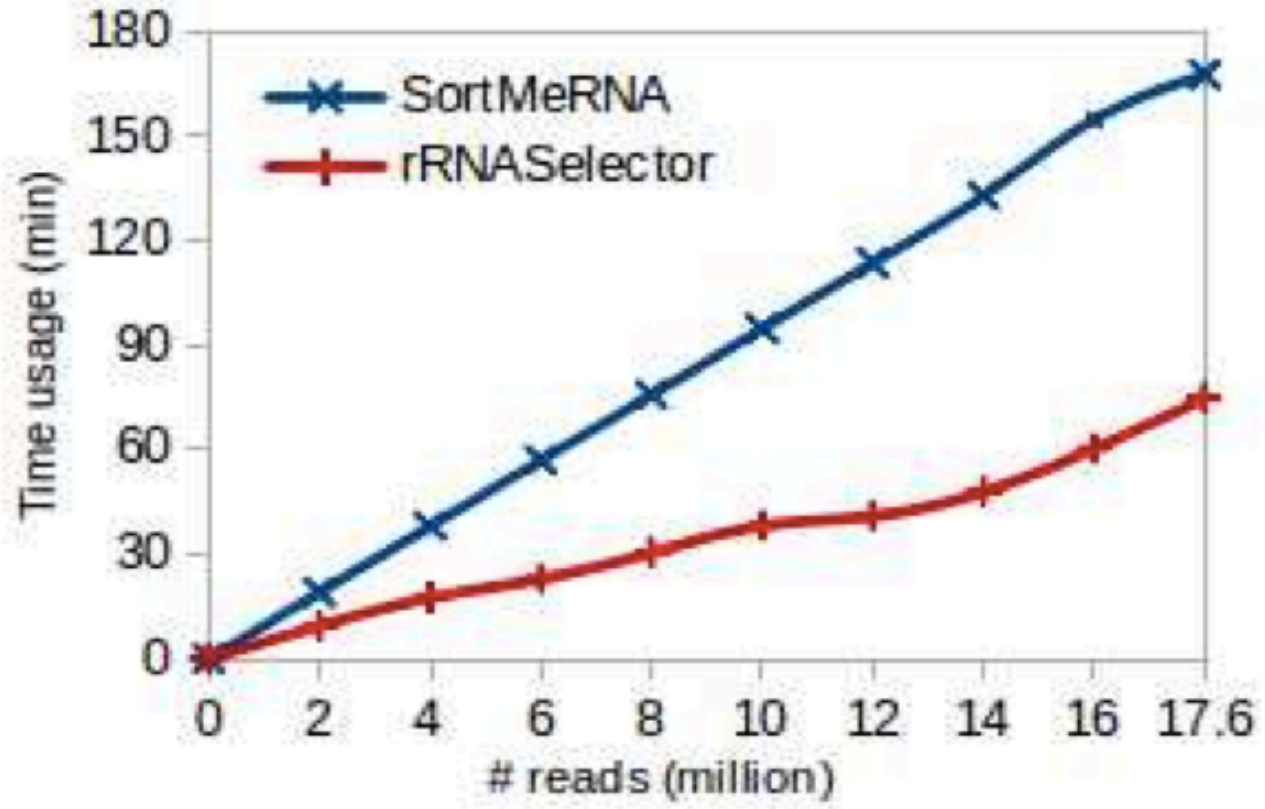
Taxonomic profiling based on identified 16S rRNA reads in the sample



Random sequencing – Identifying 16S rRNA

16S rRNA prediction tools using rRNA HMM profiles

- SortMeRNA
- rRNASelector



Kjærner-Semb Master Thesis 2016

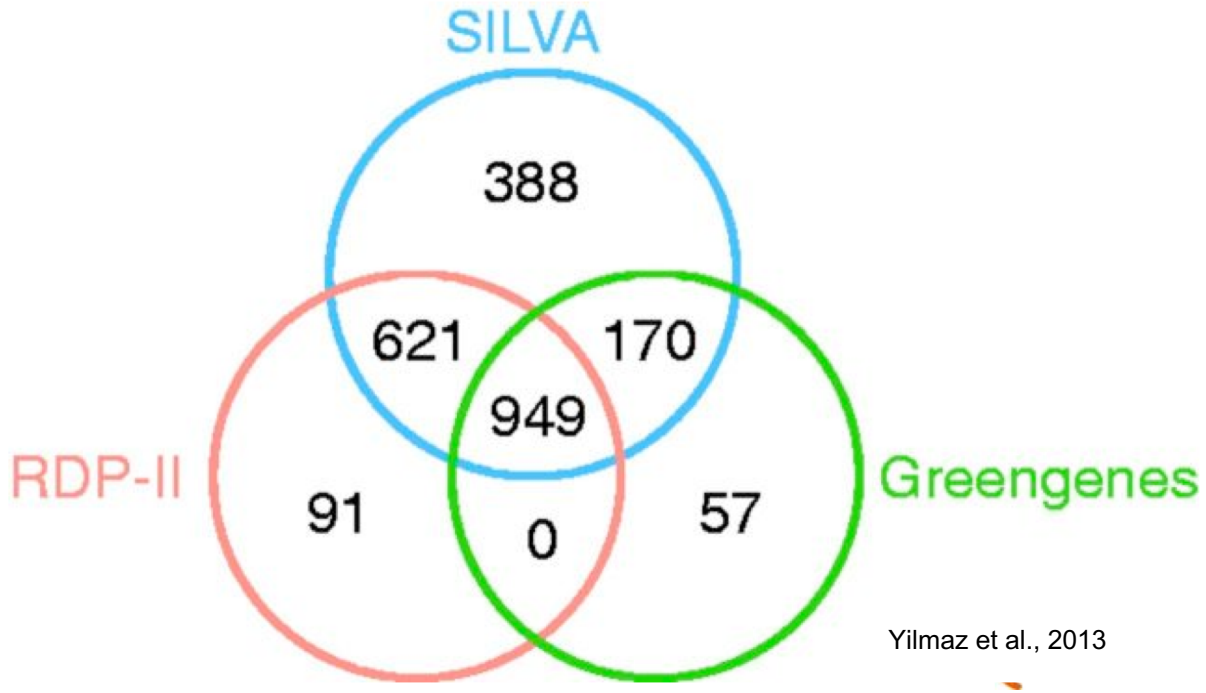
Random sequencing – Identifying 16S rRNA

Alignment/homology tools

- Megablast
- MAPseq

Database - 16S rRNA is the most widely used taxonomic marker gene

- Ribosomal Database Project
- SILVA SSU database
- Greengenes database
- Specialized databases
- HGM



Yilmaz et al., 2013

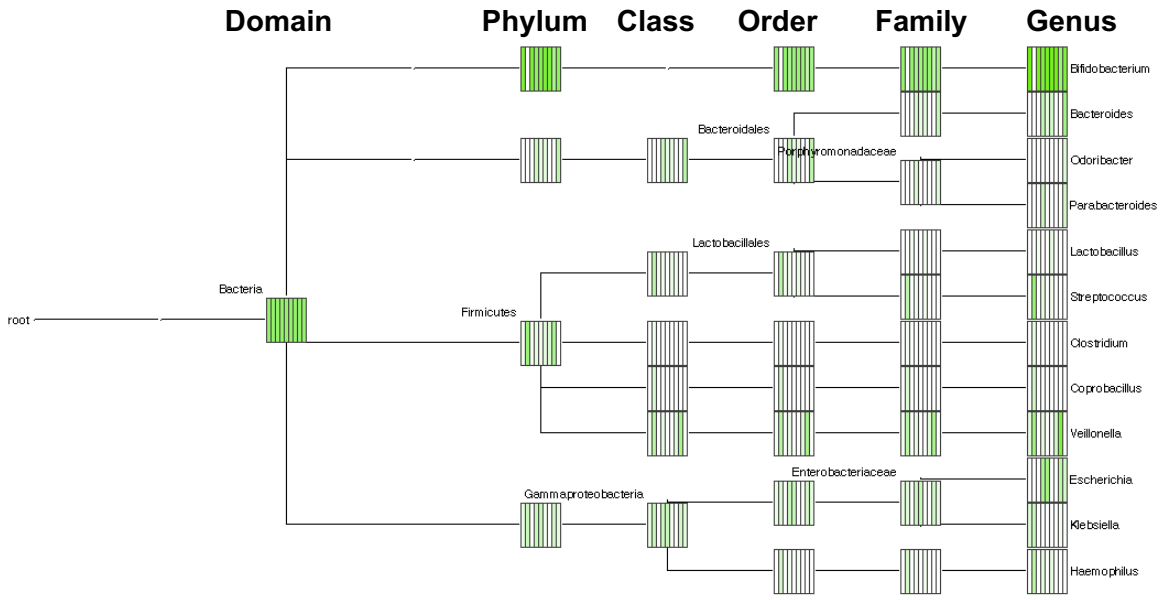
Random sequencing – Taxonomic profiling

Tools to hierarchically classify pre-aligned sequences on a taxonomy tree using an LCA algorithm

- Qiime
- LCAClassifyer
- Megan

Lots of other

- Motur



Taxonomic profiling – Alignment based methods

The most basic method is to use BLAST

Search for the best hit in a database of sequences with known origin

It is very compute intensive and slow!!!

Clostridium novyi fliC gene for flagellin, complete cds, strain: ATCC 25758 Length=864

Score = 580 bits (314), Expect = 1e-162

Identities = 360/382 (94%), Gaps = 4/382 (1%)

```
Query  1    AAAAATGAGAGGACAAATTAGAGGATTAAACCCAAGC-TCAAGAAATGCTCAAGATGGTA  59
            |||||||||||||||||| ||||||||||  ||||| ||||||||||||||||||||||
Sbjct  172  AAAAATGAGAGGACAAATCAGAGGATTAAA-TCAAGCATCAAGAAATGCTCAAGATGGTA  230

Query  60   TCTCTTTAATCCAAACAGCTGAAGGAGCTGTAAACGAAACACACGCAATACTTCAAAGAA  119
            |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  231  TCTCTTTAATCCAAACAGCTGAAGGAGCTTTAAACGAAACACACGCAATACTTCAAAGAA  290

Query  120  TGAGAGAATTATCAGTACAAGCTGCTAATGATACAAACAAAACAGAAGATAGAGCAATGA  179
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  291  TGAGAGAATTATCAGTACAAGCTGCTAATGATACAAACAAAACAGAAGATAGAGCAATGA  350

Query  180  TACAAAAAGAATTCTCACAATTACAAACAGAAATCACAAAAATTGGAAAAGACACTCAAT  239
            |||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  351  TACAAAAAGAATTCTCACAATTACAAACAGAAATCACAAGAATTGGAAAAGACACTCAAT  410

Query  240  TCAATAAACAAAACCTATTAACAGGATCAGCTTCAAGCAT-AGACTTCCAAGTAGGAGCT  298
            |||||||||||||||||||||||||||||||||  || | | ||||||||||||||||||
Sbjct  411  TCAATAAACAAAACCTATTAACAGGATCAGCTA-AATCTTTAGACTTCCAAGTAGGAGCT  469

Query  299  AATGAAAAACAAGTTATAAATGTTAAAATTGGTGATATGAGAGCCACTGCTTTAAATGTT  358
            |||| |  ||||||||||||||||||||||  |||||||||||| ||||||||||  |
Sbjct  470  AATGCAGGACAAGTTATAAATGTTAAAATTAATGATATGAGAGCTACTGCTTTAAAAATA  529

Query  359  GGCGCAGCTAATGTTAGCATAA  380
            | ||||||||| ||||||||||
Sbjct  530  GACGCAGCTAAAGTTAGCATAA  551
```

Taxonomic profiling – Alignment based methods

HMMER using probabilistic models - profile hidden Markov models

Searching HMM profile databases for sequence homologs

Much lower false positive (FP) rates

	Order-filtered				Species-filtered			
	$C3_{BLASTx}$		$C3_{HMMER3}$		$C3_{BLASTx}$		$C3_{HMMER3}$	
	TP	FP	TP	FP	TP	FP	TP	FP
Superkingdom	12282	799	6668	660	20059	113	9563	516
Phylum	8532	1094	4194	657	18968	183	8065	377
Class	3700	1257	1983	721	15793	274	6329	322
Order	–	2019	–	1158	14829	275	5084	367
Family	–	926	–	531	11126	239	3400	324
Genus	–	144	–	175	6897	427	1852	517
Species	–	9	–	25	–	142	–	214

	CARMA3	SOrt-ITEMS	MEGAN
BLASTx	54 h 15 m	54 h 15 m	54 h 15 m
-classification	52 m 22 s	12 m 36 s	3 m 4 s
HMMER3	6 h 20 m	–	–
-classification	41 m 8 s	–	–

Nucleic Acids Res. 2011 Aug; 39(14): e91.

Taxonomic profiling – K-mer based search

Kraken is a taxonomic sequence classifier that assigns taxonomic labels to short DNA reads

- Using exact alignments of k-mers
- Kraken's default database contains just under 14 billion distinct k-mers, and requires at least 500GB of disk space (Oct 2017).
- Kraken requires enough free memory to hold the database in RAM. The default database size is 174GB (Oct 2017), and so you will need at least that much RAM if you want to build or run with the default database.
- When Kraken is run with a reduced database, it is called MiniKraken

Taxonomic profiling – Search against protein databases

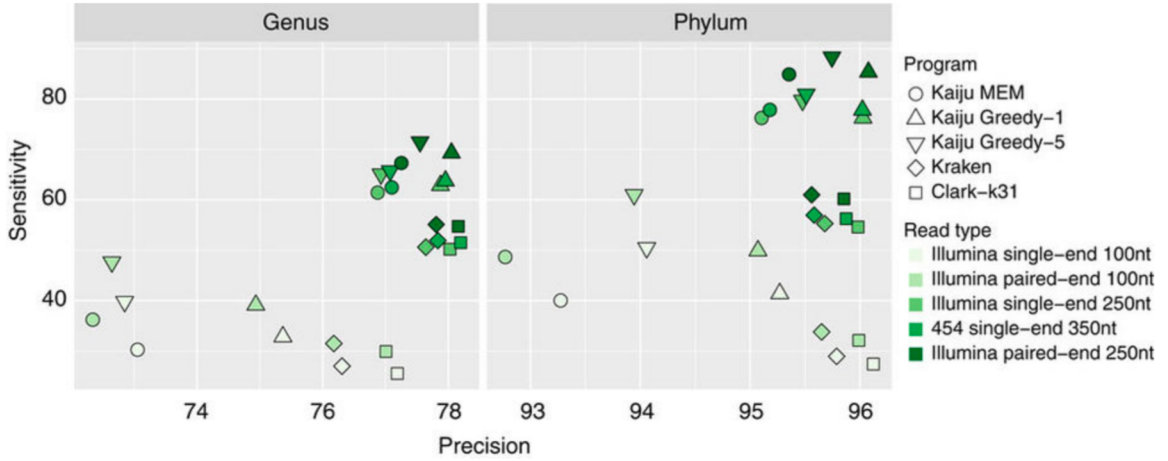
Kaiju is a taxonomic sequence classifier that use a reference database of protein sequences

- Finds maximum matches on the protein-level using the Burrows–Wheeler transform
- Reads are directly assigned to taxa using the NCBI taxonomy and a reference database of protein sequences from microbial and viral genomes
- Kaiju can be installed locally or used via a web server
- Can be run against various databases (eg. NCBI RefSeq)
- It can also be run against the Mar databases from the Marine Metagenomics Portal

Taxonomic profiling – Search against protein databases

Kaiju is a taxonomic sequence classifier that use a reference database of protein sequences

Claim to be faster and more sensitive than K-mer based methods



Peter Menzel Nature Communications 7, Article number: 11257 (2016)

Taxonomic profiling - Clade-specific markers

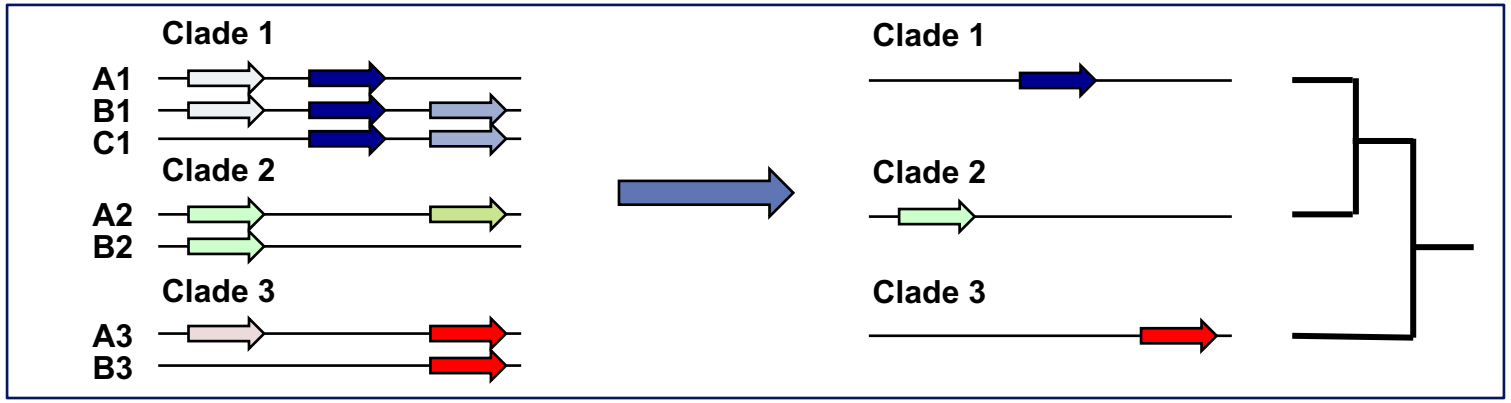
MetaPhlAn2 is a taxonomic sequence classifier that use a clade-specific marker database

- Using read coverage of clade-specific markers to detect the taxonomic clades present in a microbiome sample and estimate their relative abundance
- Map reads against clade-specific marker sequences that are pre-selected from coding sequences that identify specific microbial clades at the species or higher taxonomic levels
- The clade-specific markers cover all main functional categories
- MetaPhlAn2 includes ~1 million markers from >7,500 species

Taxonomic profiling - Clade-specific markers

MetaPhlAn2 is a taxonomic sequence classifier that use a clade-specific marker database

- Dark blue is restricted yet universal across Clade 1
- Green genes are restricted to Clade 2, red genes to Clade 3



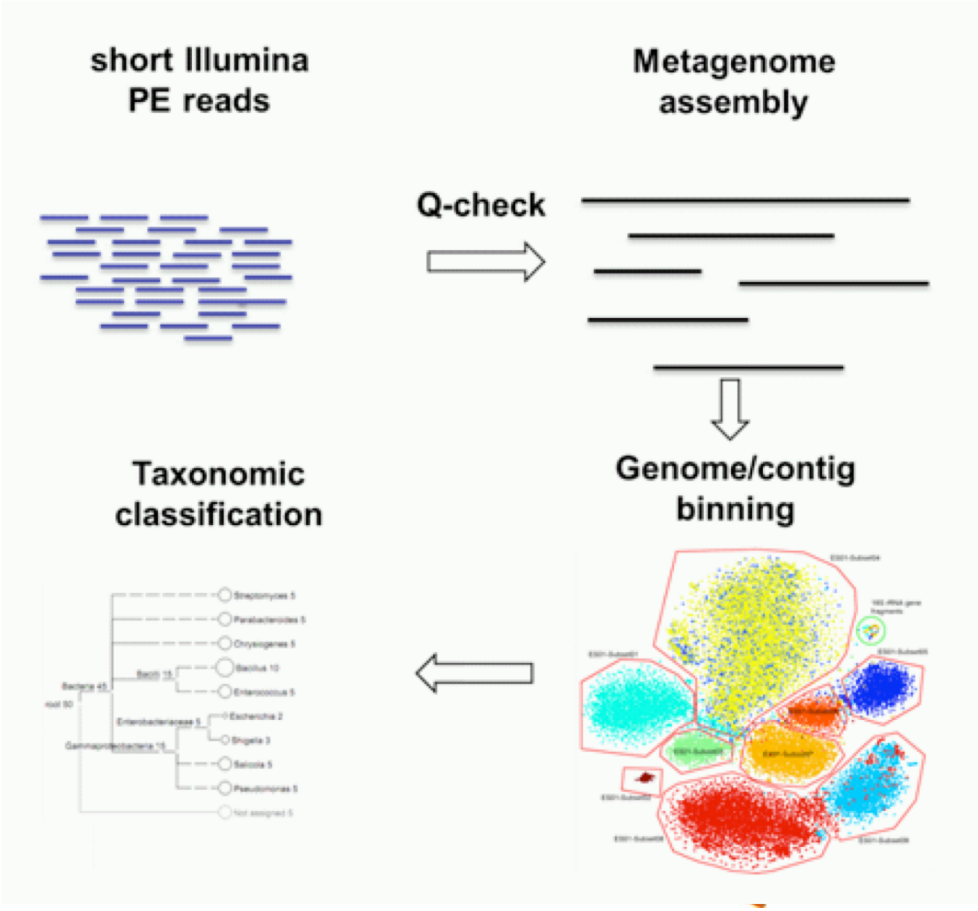
Taxonomic binning – More tomorrow

Clustering of assembled contigs that apparently originate from the same source population

- Assign to the closest possible taxonomy
- Enables the discovery of new microbial of new organisms

Tools for binning of contigs

- MaxBin
- MyCC
- Metawatt
- MetaBAT



CAMI - Compared taxonomic profilers – not binning

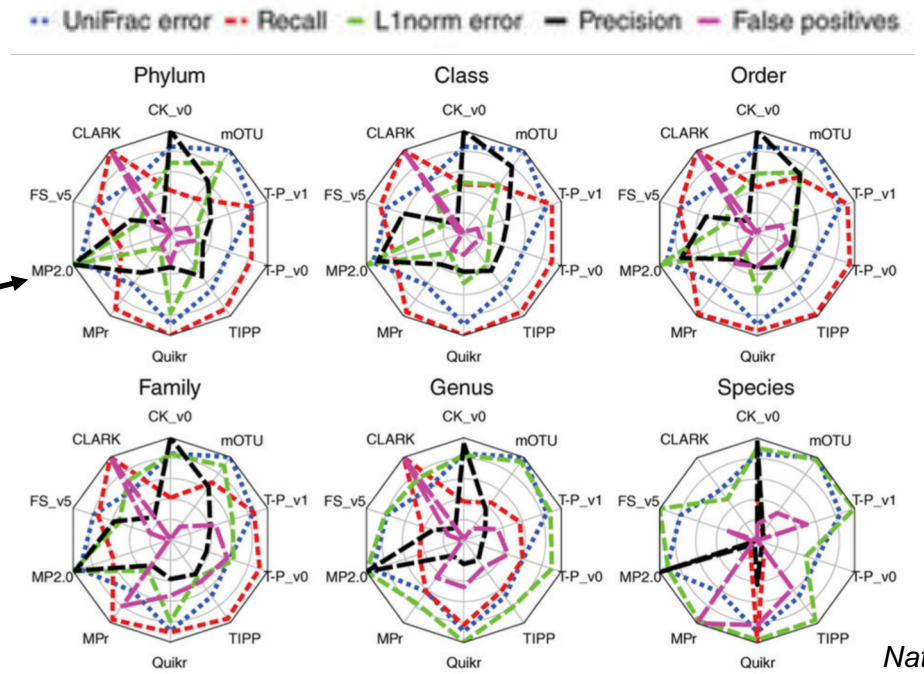
Profilers fell into three categories:

(i) profilers that correctly predicted relative abundances

(ii) precise profilers

(iii) profilers with high recall

Highest precision



Nature Methods **14**, 1063–1071 (2017)

Technical variations influence results

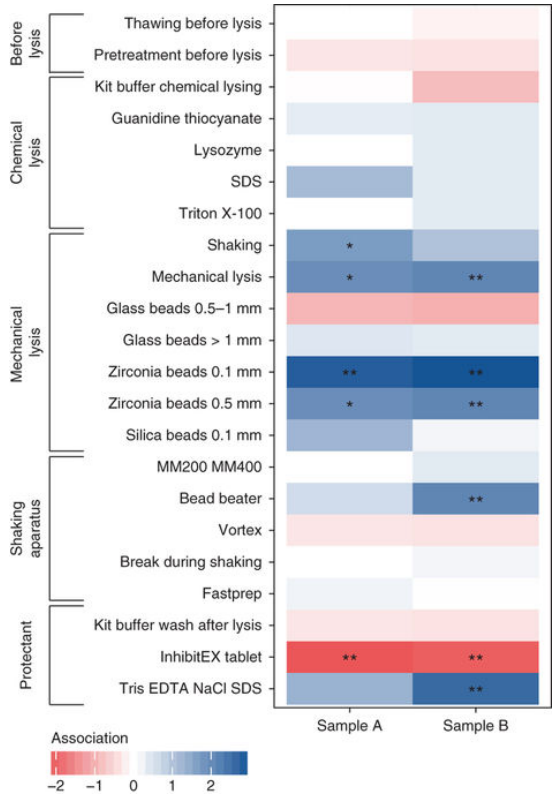
DNA extraction had the largest effect on the outcome of metagenomic analysis

Effects of protocol manipulations on sample composition

SAMPLE 1 EXTRACTION A

SAMPLE 1 EXTRACTION B

amazon.com



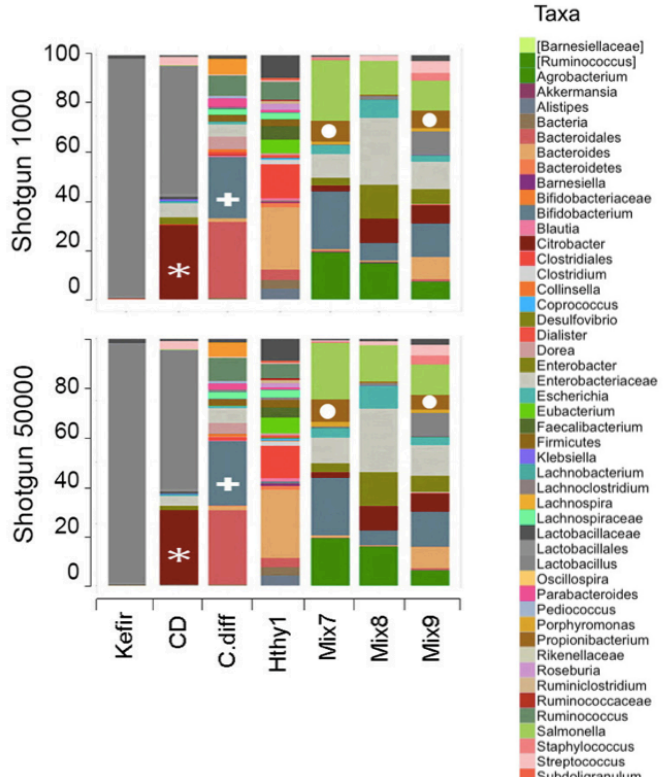
Costea et al, Nature Biotechnology 35, 1069–1076 (2017)

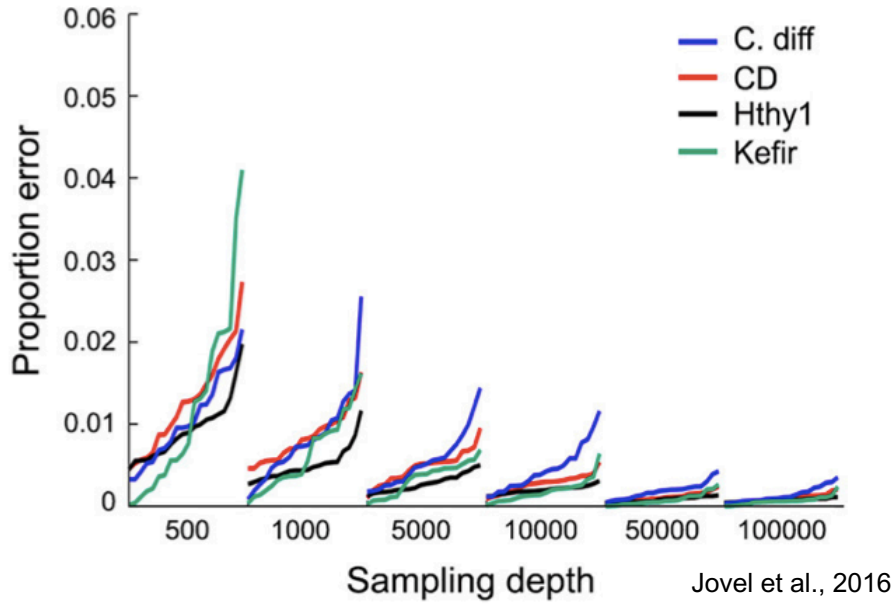
Sequencing depth influence results

Increasing sampling depth = increased detection of taxa

Taxonomic classification for the same library at different sequencing depths is surprisingly consistent (Jovel et al., 2016)

The proportion error and its variance decrease with increasing sampling depth





Jovel et al., 2016

Number of species on earth

We know very few…

Earth contains 10^11 to 10^12 species of microbes (some estimate 10^19)

The total number of described bacterial species is very low 10^4

NCBI list of taxonomically approved names contain 17.989 bacterial species

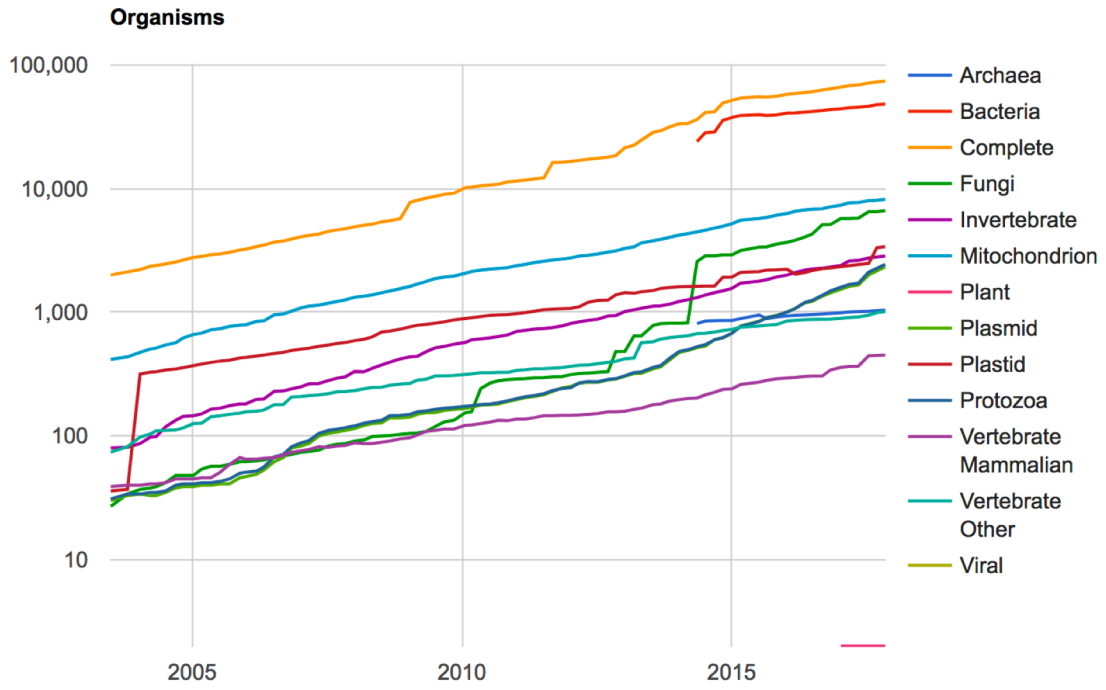
= **510 100 000 km²**

= **5,1 km²**

You only find what is in the database…

What is in the databases - for example RefSeq?

The Reference Sequence (RefSeq) collection is a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA



Organism group
Animals (1,618)
Plants (584)
Fungi (3,102)
Protists (669)
Bacteria (131,896)
Archaea (2,362)
Viruses (14,001)
Customize ...

Status
Latest (151,417)
Latest GenBank (151,444)
Latest RefSeq (113,005)
Replaced (5,953)

Assembly level
Complete genome (23,742)
Chromosome (3,097)
Scaffold (65,086)
Contig (65,445)

Organism group
✓ **Bacteria** (131,896)
Customize ...

Status
Latest (126,962)
Latest GenBank (126,965)
Latest RefSeq (103,882)
Replaced (4,934)

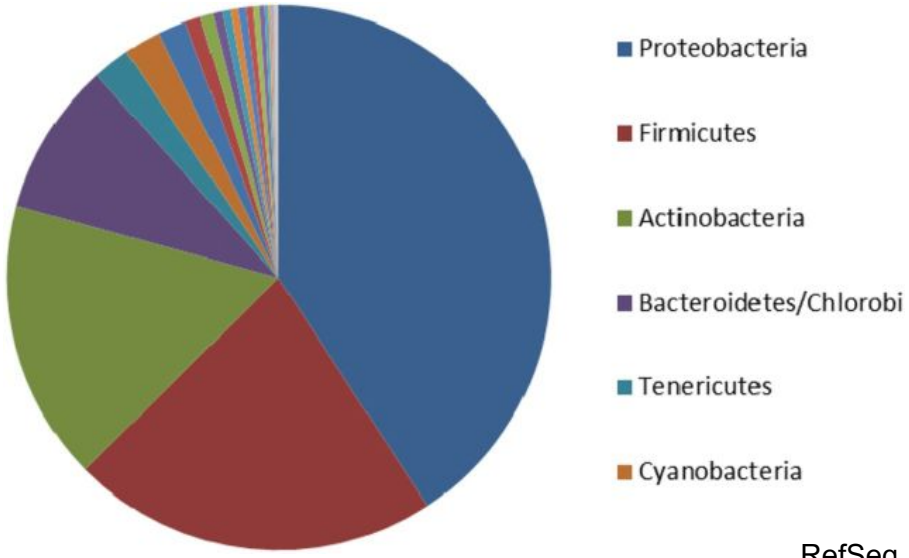
Assembly level
Complete genome (9,496)
Chromosome (1,863)
Scaffold (59,894)
Contig (60,643)

You only find what is in the database…

What is in the databases - for example RefSeq?

Large fraction of Proteobacteria

Host-associated are overrepresented



RefSeq

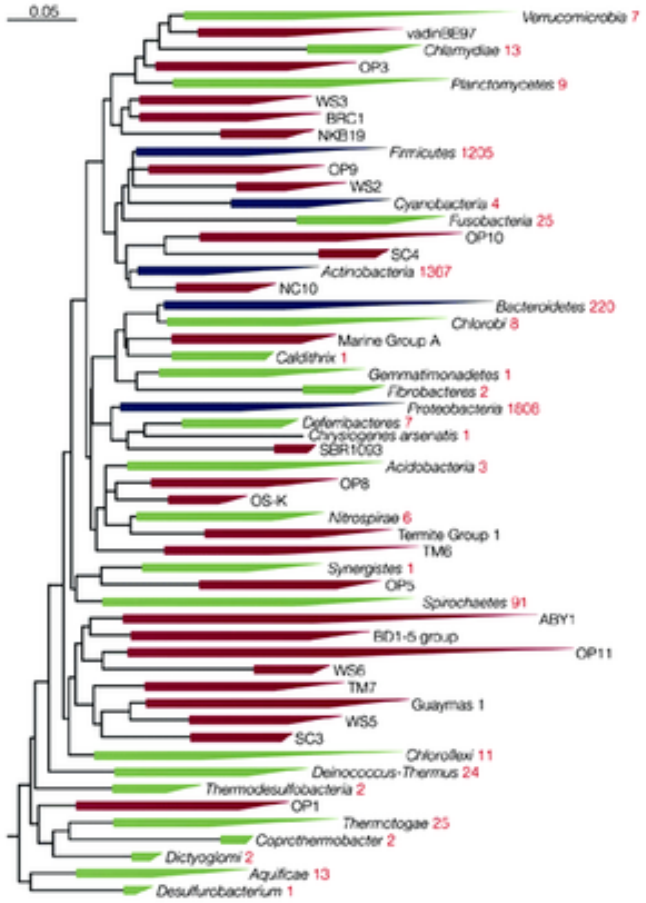
Ecosystem	Total
Host-associated	**11,816**
Humans	4973
Animal	1804
Plants	1410
Mammals	867
Other	2762
Environmental	6774
Aquatic	4559
Terrestrial	2057
Other	158
Engineered systems	**1658**
Food production	440
Wastewater	410
Lab synthesis	387
Other	418
Total	**20,248**

GOLD database

You only find what is in the database…

92 named bacterial phyla – but constantly changing

The total number has been estimated to exceed 1,000 bacterial phyla



nature microbiology

A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

Nature Microbiology **1**, Article number: 16048 (2016)
doi:10.1038/nmicrobiol.2016.48

Received: 25 January 2016
Accepted: 10 March 2016
Published online: 11 April 2016

Martin Keller & Karsten Zengler
Nature Reviews Microbiology volume 2, pages 141–150 (2004)

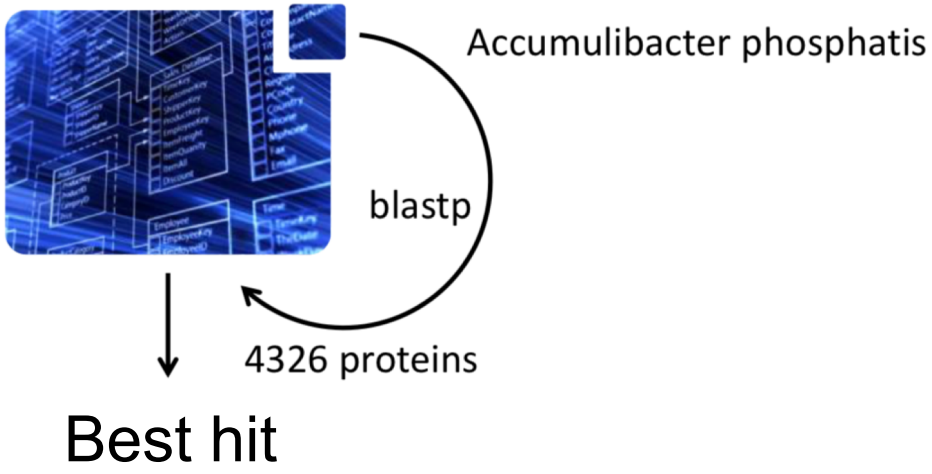
Effect of missing genome

What is the effect of not having closely related genomes in the database?

1. Remove a genome from the database

2. Search the removed genome against the database

Effect of missing genome



Effect of missing genome

blastp

4326 proteins

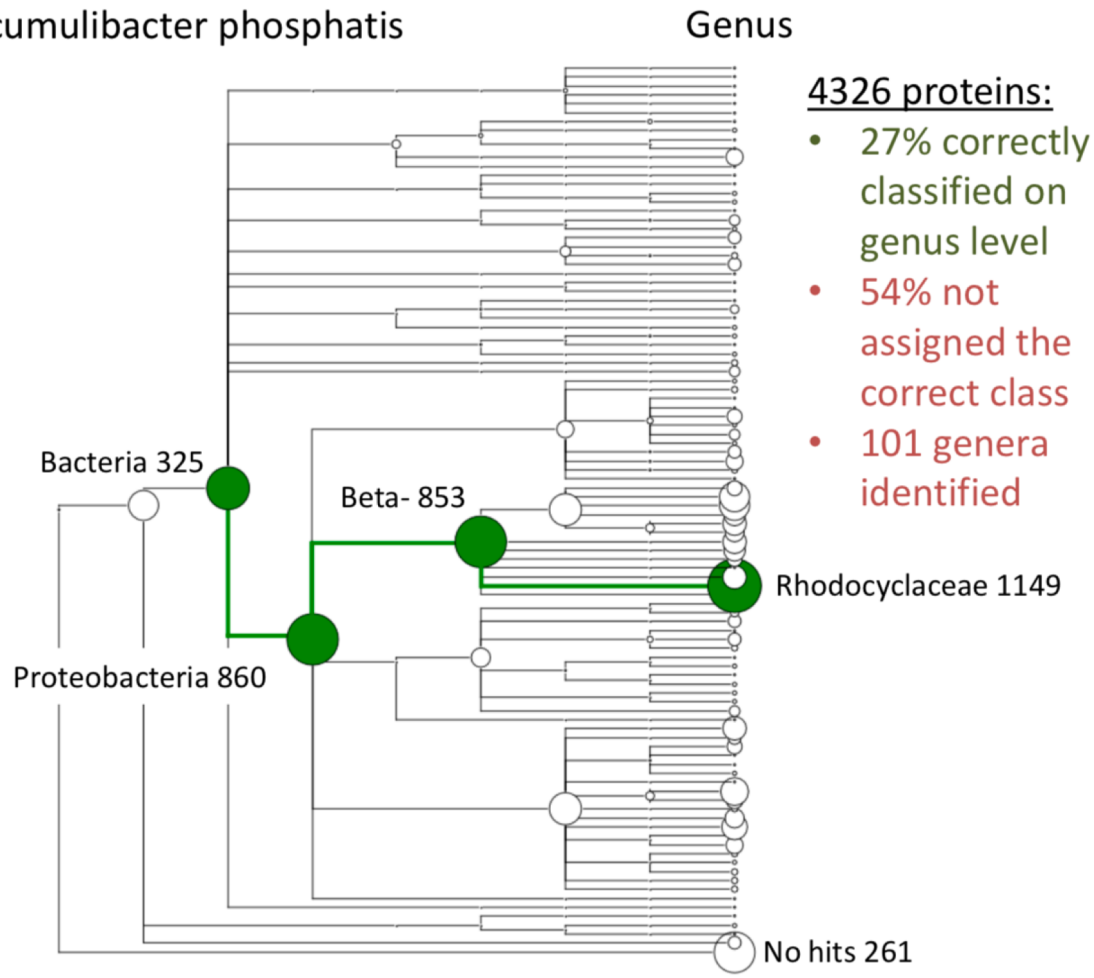
MEGAN LCA

Lowest common ancester (LCA) approach:
Hit 1: Beta-proteobacteria 80% ID
Hit 2: Gamma-proteobacteria 79% ID
Hit 3: Actinobacteria 59% ID

Assigned to Proteobacteria

Related genomes

Bacteria	1268
Proteobacteria	564
Betaproteobacteria	84
Rhodocyclales	5
Rhodocyclaceae	5



4326 proteins:

- 27% correctly classified on genus level
- 54% not assigned the correct class
- 101 genera identified

EXERCISE – day 2

From raw reads to a classification of organisms present in the dataset

- Taxonomic classification using reads with 16S rRNA
- Taxonomic classification using protein and k-mer based databases