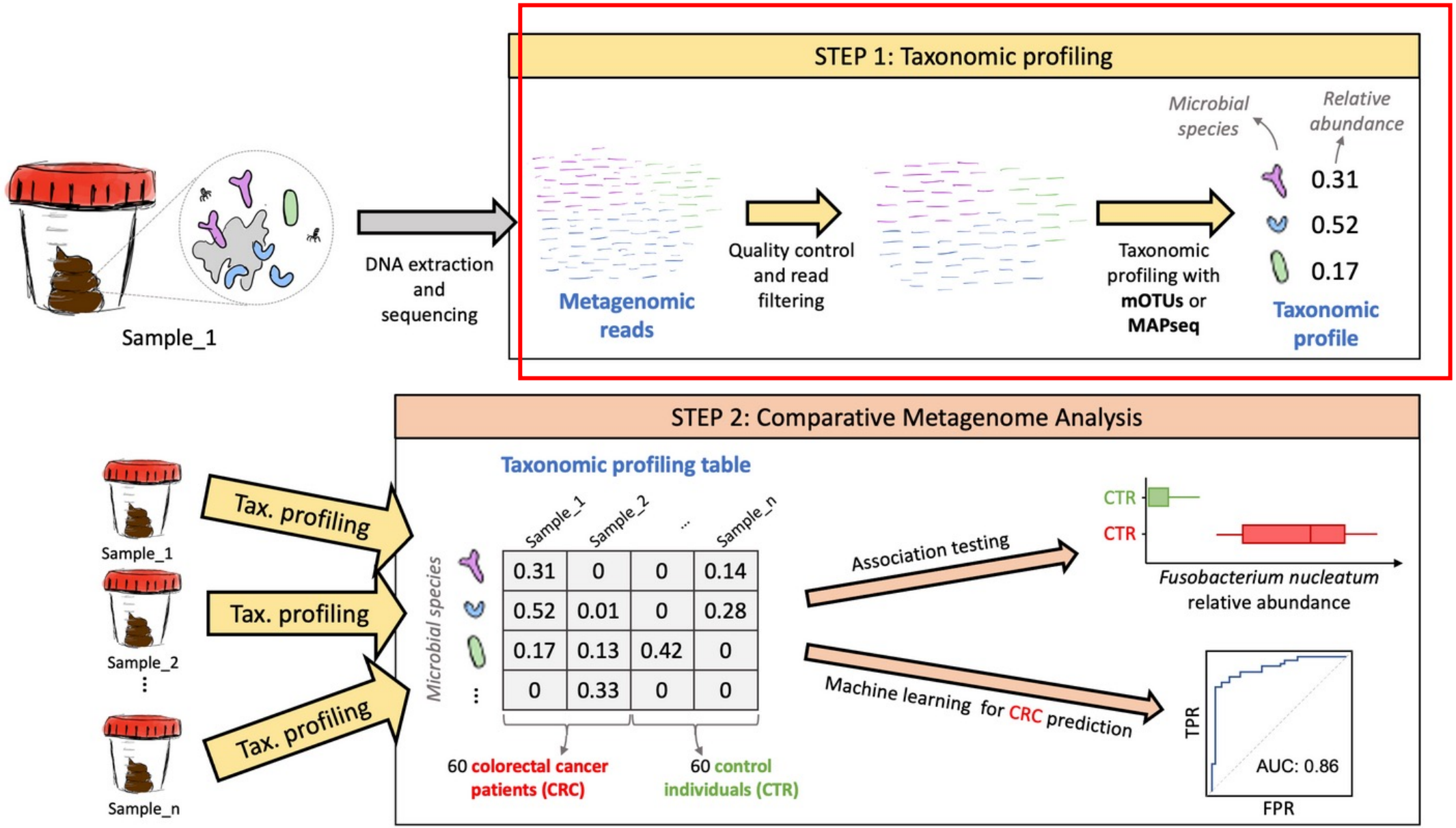
Profiling and modeling the colorectal cancer microbiome

Alessio Milanese, Lukas Malfertheiner

Project 3

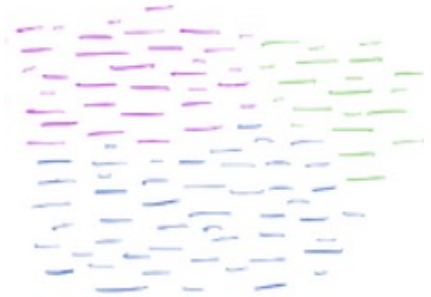
Spring School Bioinformatics and computational approaches in Microbiology


STEP 1: Taxonomic profiling
Sample_1
DNA extraction and sequencing
Metagenomic reads
Quality control and read filtering
Taxonomic profiling with mOTUs or MAPseq
Microbial species
Relative abundance
0.31
0.52
0.17
Taxonomic profile
STEP 2: Comparative Metagenome Analysis
Sample_1
Sample_2
Sample_n
Tax. profiling
Tax. profiling
Tax. profiling
Taxonomic profiling table
Microbial species
Sample_1
Sample_2
...
Sample_n
0.31 0 0 0.14
0.52 0.01 0 0.28
0.17 0.13 0.42 0
0 0.33 0 0
60 colorectal cancer patients (CRC)
60 control individuals (CTR)
Association testing
CTR
CTR
Fusobacterium nucleatum relative abundance
Machine learning for CRC prediction
TPR
AUC: 0.86
FPR

Part 1: Quality Filtering and Trimming

Sequencing by Synthesis

Not perfect!



Fastq Files

Adapter

Sequence- high quality

Sequence- poor quality

Fastq Files

```
@read98
CATCGACGACCTGGACGACCTGGACTTCATCGAGCGGGTGAAGATCCAGCAGAAGAACTGGATCGGCCGCTCCACCGGTGCCGAGGTCACCTTCAAGGCC
+
BBBFFFFFFFFFFBBFFFFIIFFFIIIIIIIIFBFIIFFFFFFFBBBFFFFBBFFFFFBBFFFBBBFBBBBFBBFBFFBBFFF0<B7BBFBB<BBFBBBF
@read169
GCGGTGGTTCGGATCTGATGTTCCCGCACCATGAATATCAAAATGGTTGCCCAACTGTTTGCAGTTCATCGCCGAACATTCGATGTGCC
+
BBBFB<F7B<BFBFFFBBFBFF<FFFFIFF0<BBFFB<<<BFFBBF7B0BFFBFBFF<BB7<BF'<7<BBBBBB<0<00<'0<B<''0<
@read221
GCCAATGGCACTGCCCGGTAGCTAAATGCGGAAGAGATAAGTGCTGAAAGCATCTAAGCACGAAACTTGCCCCGAGATGAGTTCTCCCTGACTCCTTGAG
+
BBBFFFFFFFFFFIIIFIBFFFIIIIIIIFIFFFFIIIIFFFFIFIIBFFBFFIIIBFIBBFF7BBBBFFFBBFB7BBBBFFFBFBBBBBFBFBFBBBBF
@read295
ATCATTCACAATGGGGGAAACCCTGATGGTGCGACGCCGCGTGGGGGAATGAAGGTCTTCGGATTGTAAACCCCTGTCATGTGGGAGCAAATTAAAANNN
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIFFFIIIIIIIIFFFFFFFFFFBF7BBBBBFFFFBFFFFFFBBFFFF<BBFFFFFFF7BBFBBBFFB<<(((
@read601
CCCACAAGCAGCGCGCCCACGAGCCCCGGTCCATAGGTAACCGCCACCGCCGTCATATCCTCTAGAGACATATCCGCTTCCTTTAACGCTTCCTCGATCA
+
BBBFFFFFFFFFFIIIIIIFFIFIIIIIIFFIIIFFFBBBFBFFFFFFBFFF7BBBFFFFFFFBFFFBFFFFFBFBBBFFFFFFBFFFBFFFFFF<<BBF
@read643
CTCATGGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACGTATTCACCGCAGCATGCTGATCTGCGATTACTAGCAATTCCGACTTCATACAGGCGAGTT
+
BBBFFFFFFFFFFIFIIIIFFIIIIIIFFFIIIIFFFFFFBBBBFFBFFFFFFBBFBFFFFFFFFFBBFFFFFFF<BBBFFFFFFFFFFBBFFBBBBFBB
```

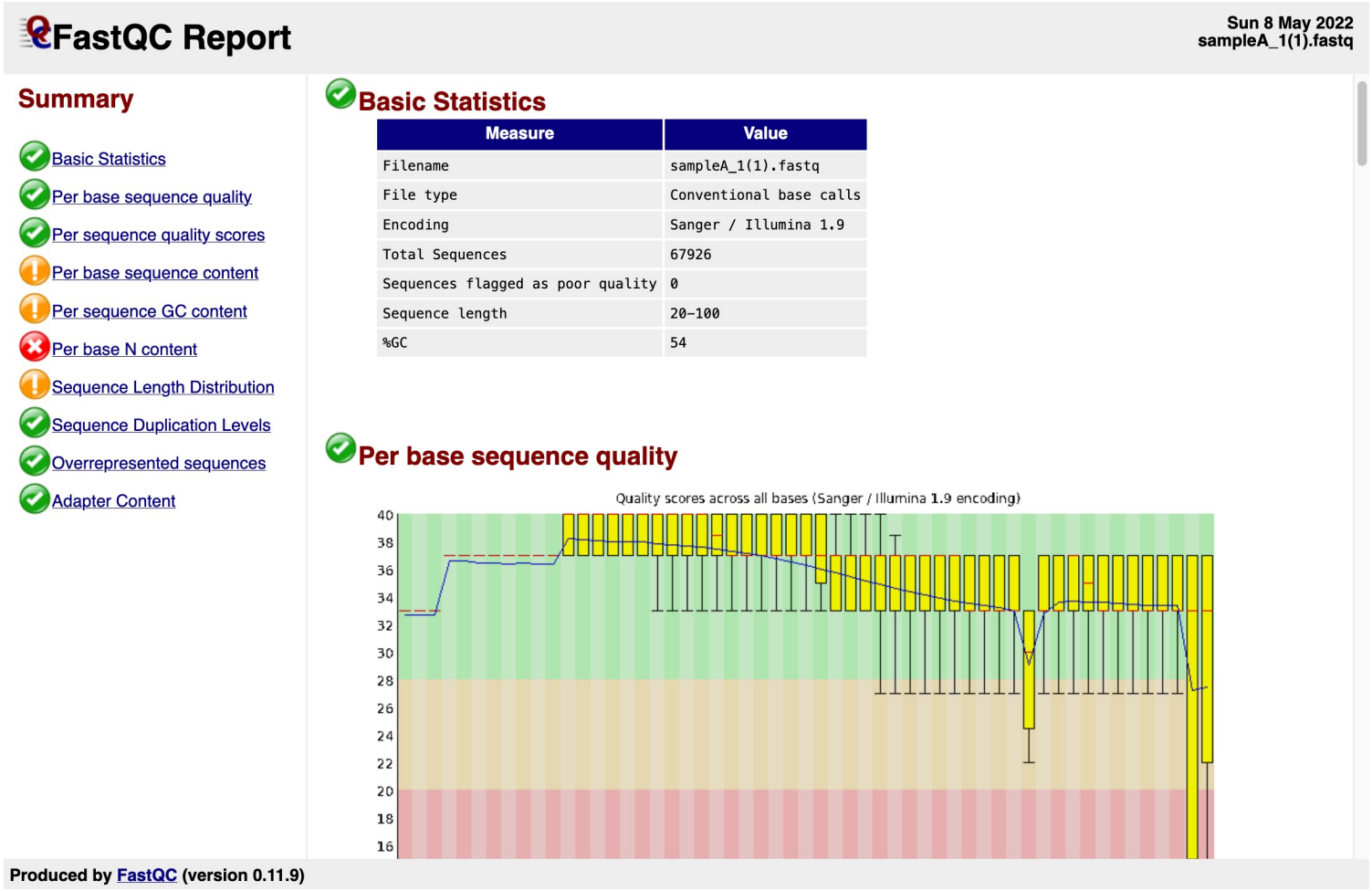
```
    Character: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
               |         |         |         |         |
Quality score: 0.........10........20........30........40
```

Phred Score

Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

Fastqc

- Program to asses quality of sequences with a convenient web interface



BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2114–2120
doi:10.1093/bioinformatics/btu170

Genome analysis

Advance Access publication April 1, 2014

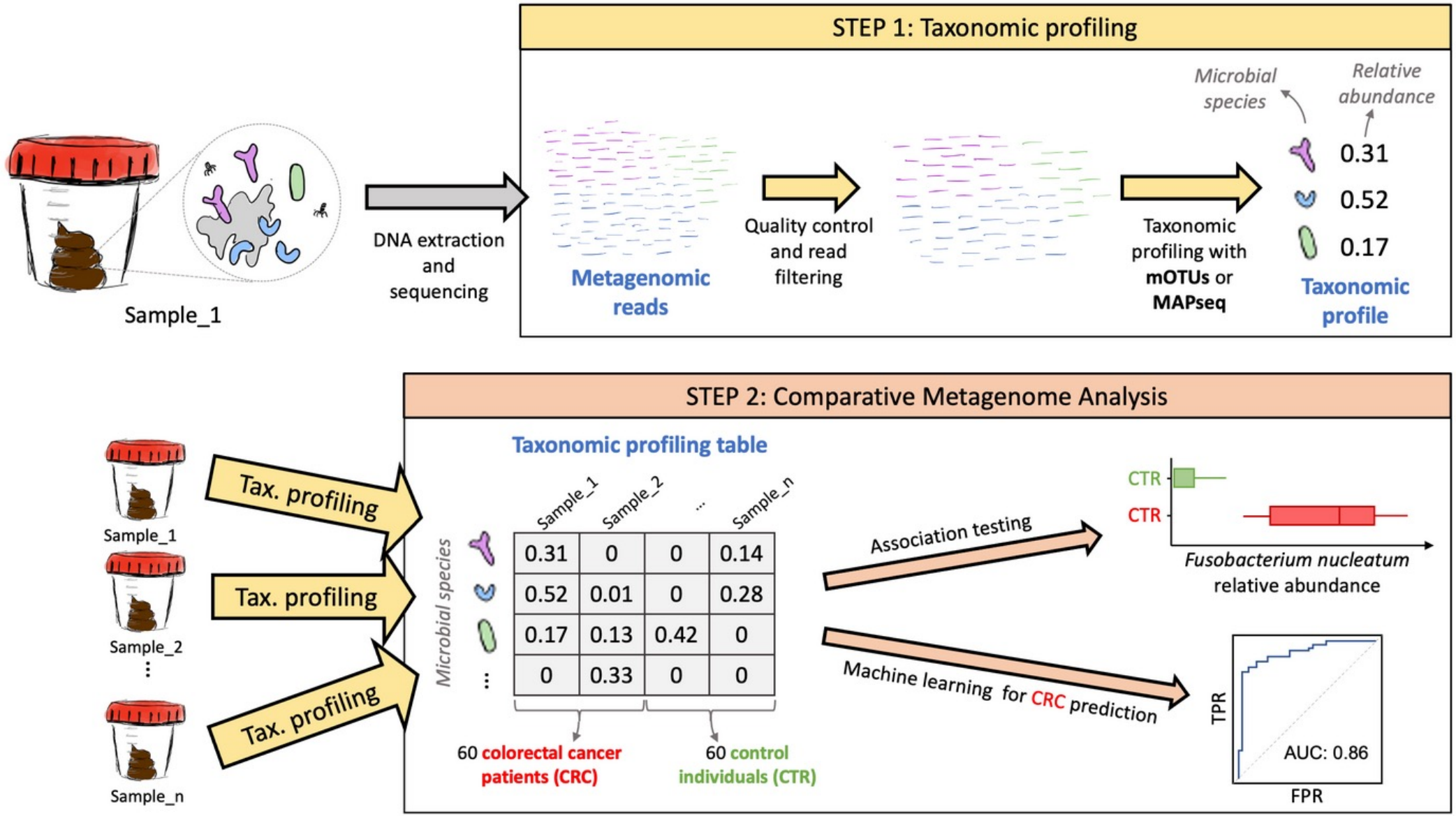
Trimmomatic: a flexible trimmer for Illumina sequence data

Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]

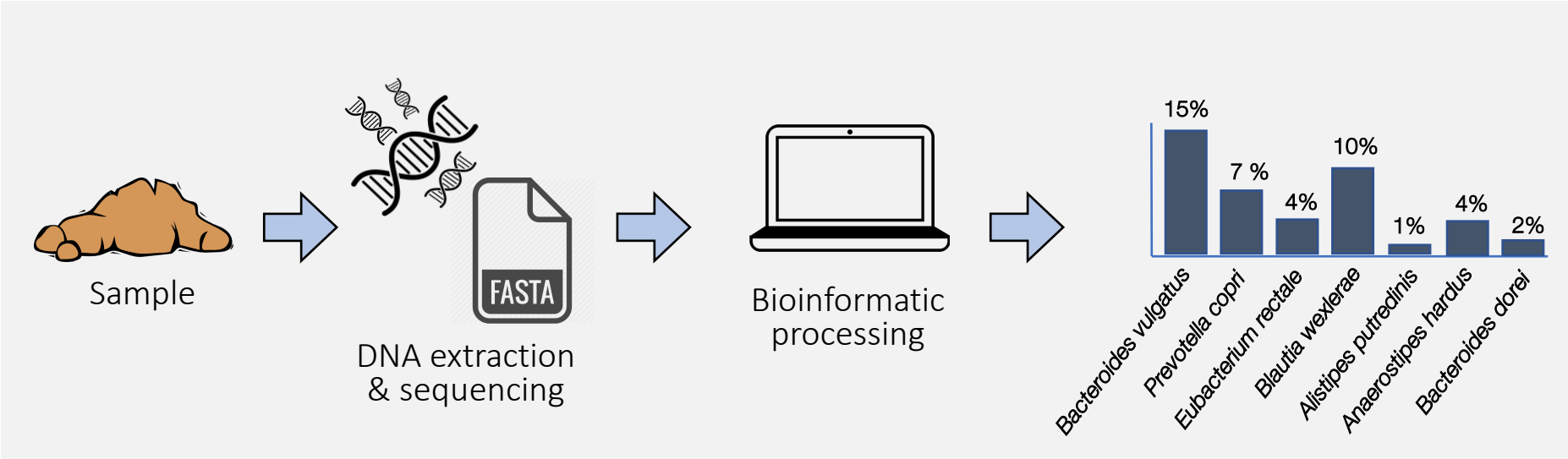
[1]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm,[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen and [3]Institute of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany

Associate Editor: Inanc Birol

- Use quality information obtained by fastqc in order to trim our fastq files
- Important to not get missleading results!


STEP 1: Taxonomic profiling
Sample_1
DNA extraction and sequencing
Metagenomic reads
Quality control and read filtering
Taxonomic profiling with mOTUs or MAPseq
Microbial species
Relative abundance
0.31
0.52
0.17
Taxonomic profile
STEP 2: Comparative Metagenome Analysis
Sample_1
Sample_2
Sample_n
Tax. profiling
Tax. profiling
Tax. profiling
Taxonomic profiling table
Microbial species
Sample_1
Sample_2
...
Sample_n
0.31 0 0 0.14
0.52 0.01 0 0.28
0.17 0.13 0.42 0
0 0.33 0 0
60 colorectal cancer patients (CRC)
60 control individuals (CTR)
Association testing
CTR
CTR
Fusobacterium nucleatum relative abundance
Machine learning for CRC prediction
TPR
AUC: 0.86
FPR

Taxonomic profiling – what is it?

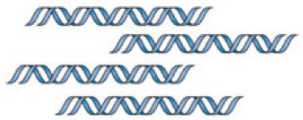


Taxonomic Profiling:
Estimate relative cell counts in a microbiome sample from metagenomic sequencing

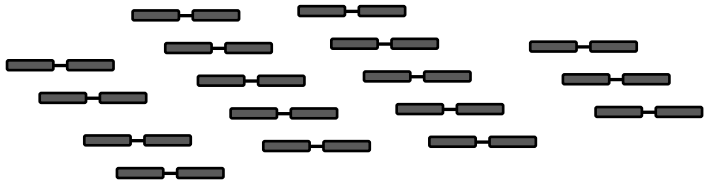
Taxonomic profiling – how it is done?

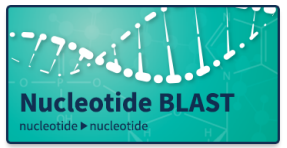
sample

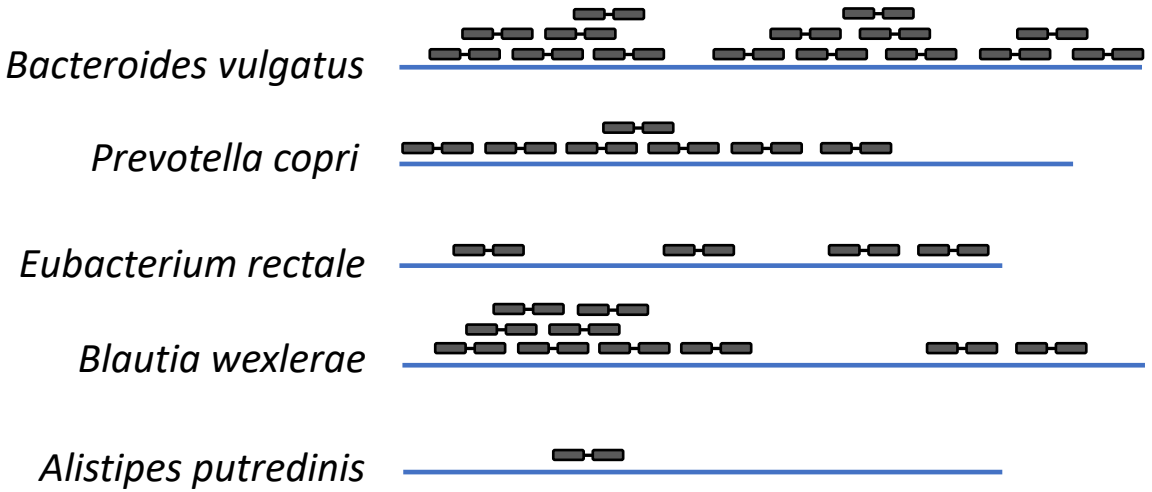
DNA purification

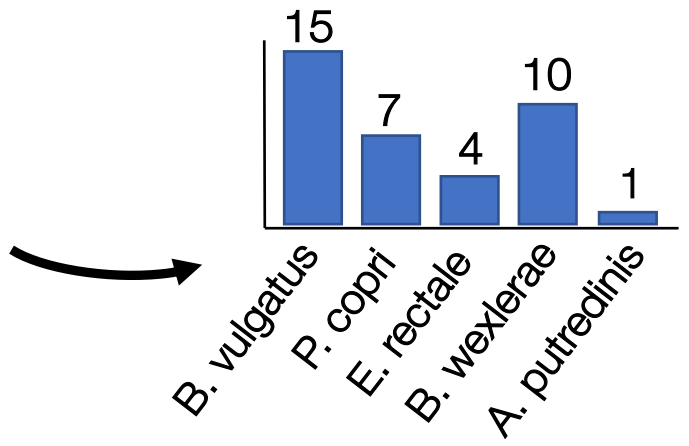


Shotgun sequencing





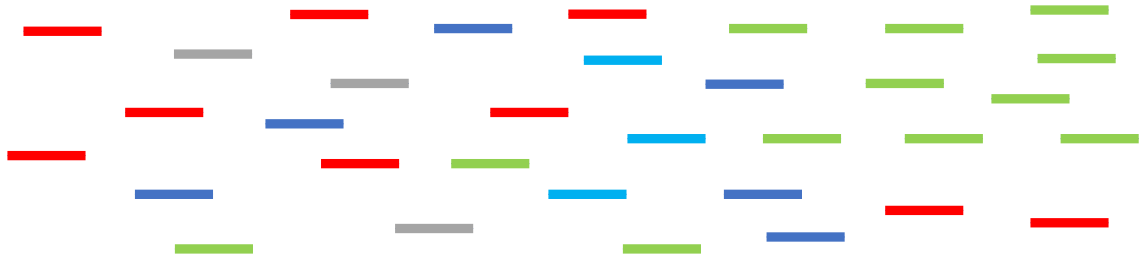




Taxonomic profiling approaches – whole-genome mapping

Environmental sample

Shotgun sequencing



- DNA extraction bias
- sequencing biases
- sampling noise

Taxonomic profiling approaches – whole-genome mapping

Environmental sample

Shotgun sequencing

True taxonomic annotation

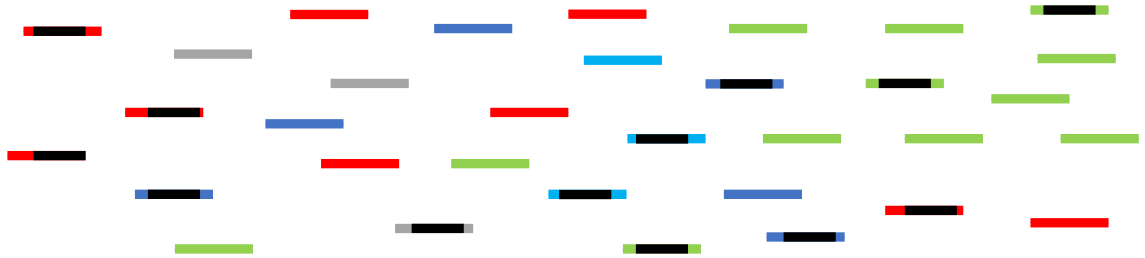
Estimated by whole-genome mapping

- genome size issue

Taxonomic profiling approaches – marker gene mapping

Environmental sample

Shotgun sequencing



Taxonomic profiling approaches – marker gene mapping

Environmental sample

Shotgun sequencing

Taxonomic profiling approaches – marker gene mapping

Environmental sample

Shotgun sequencing

Taxonomic profiling approaches – marker gene mapping

Environmental sample

Shotgun sequencing

True taxonomic annotation

Estimated by whole-genome mapping

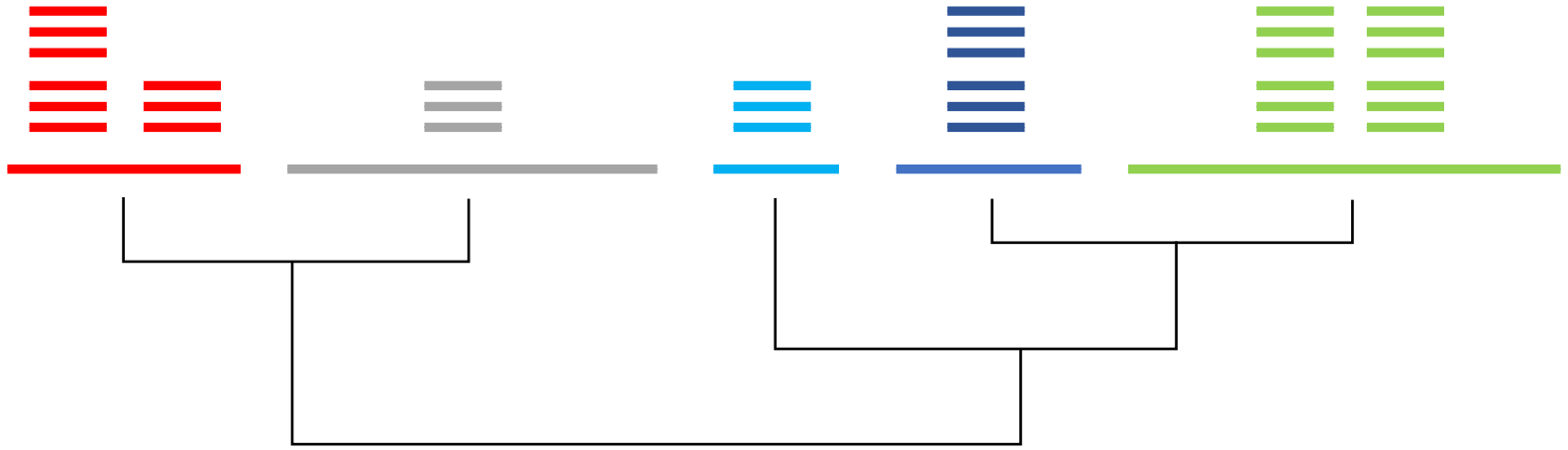
- genome size issue

Estimated by universal marker

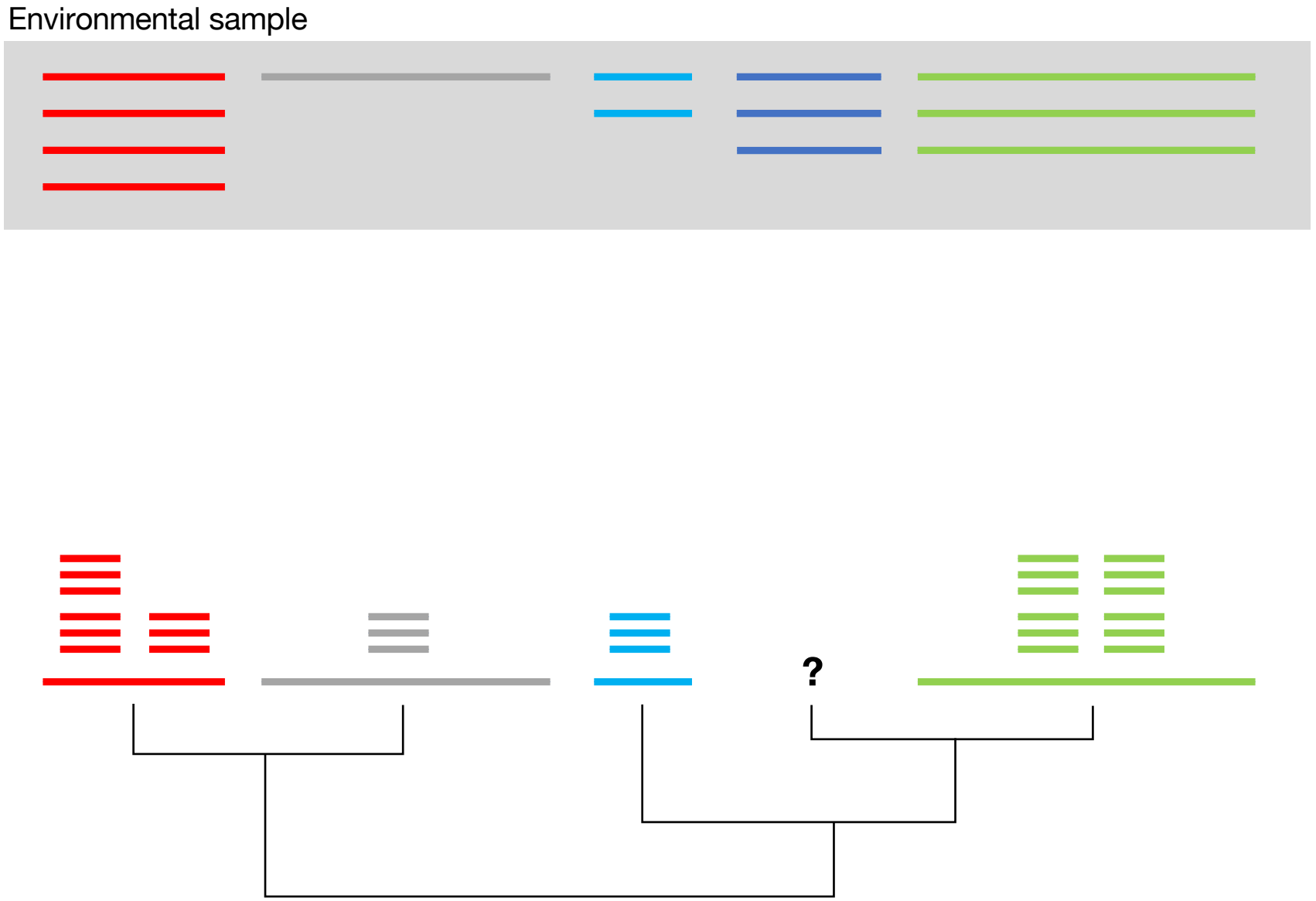
Taxonomic profiling – mapping reads to genomes

Environmental sample

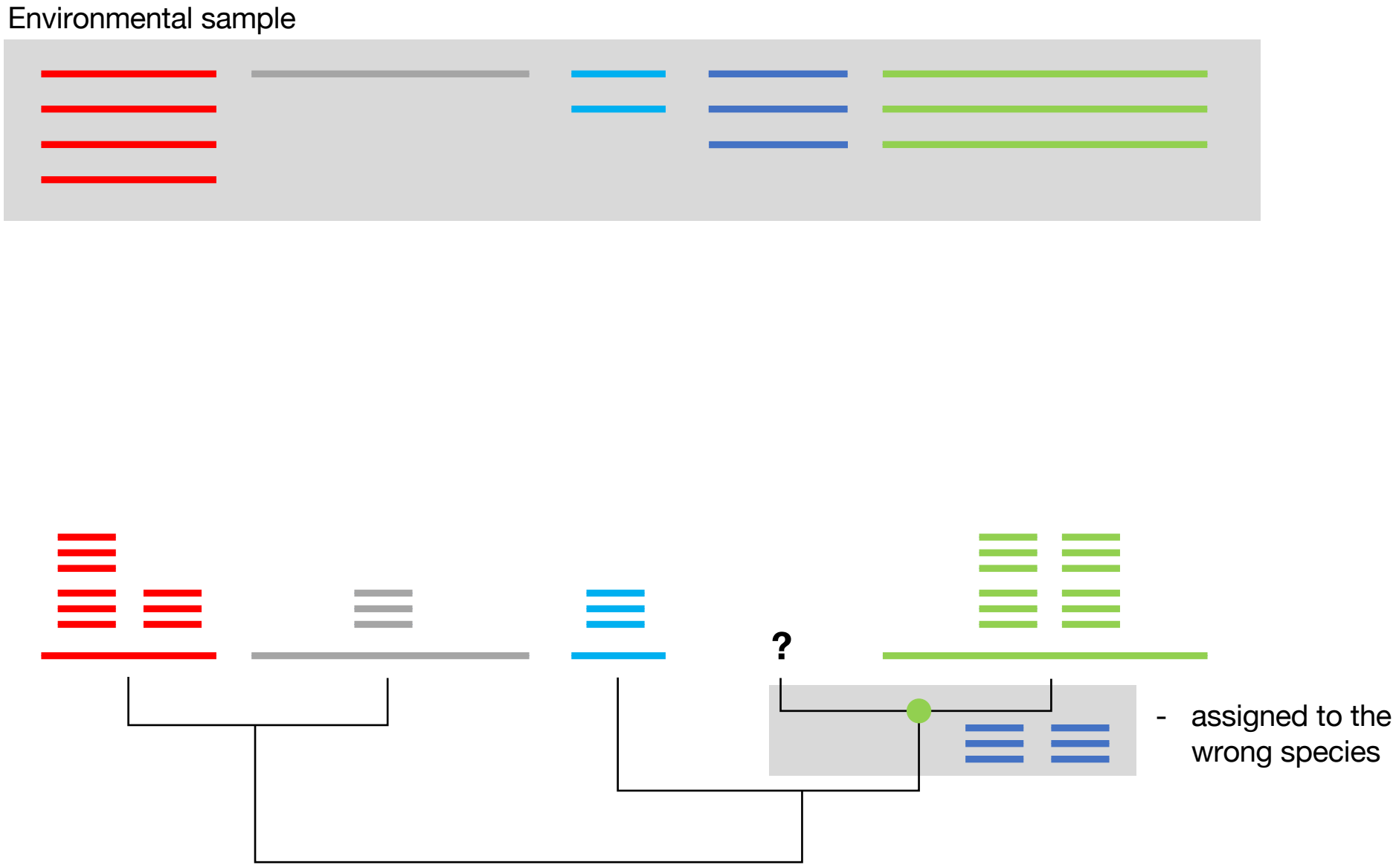
Shotgun sequencing



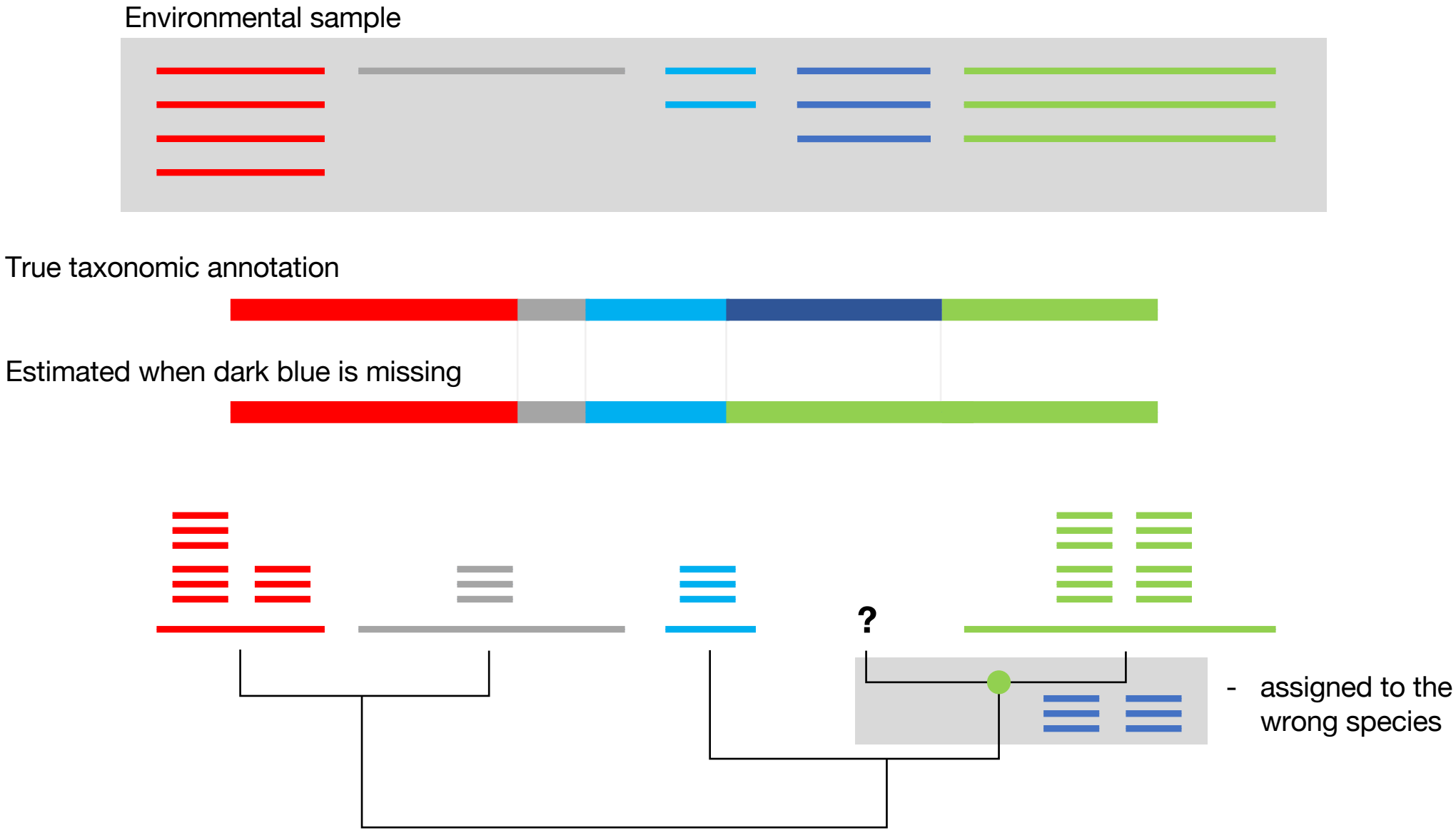
Taxonomic profiling – incomplete reference databases



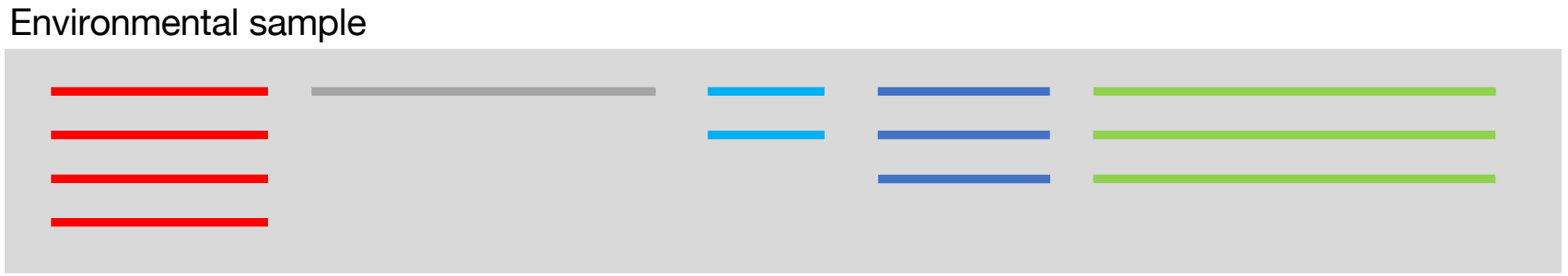
Taxonomic profiling – incomplete reference databases



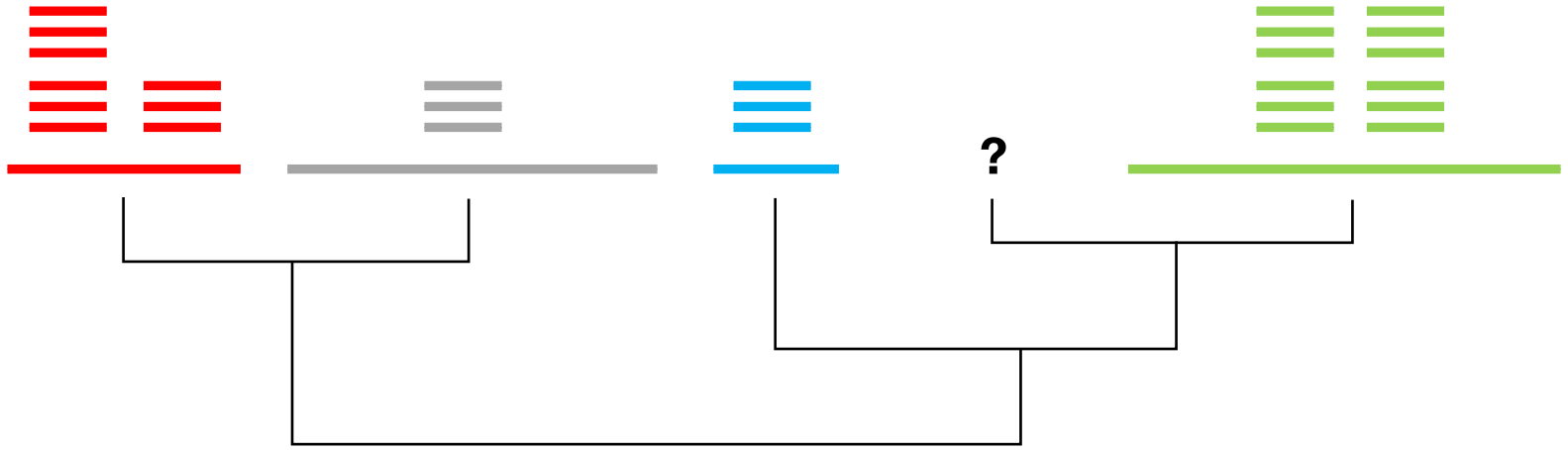
Taxonomic profiling – incomplete reference databases



Taxonomic profiling – incomplete reference databases



- ignore the reads



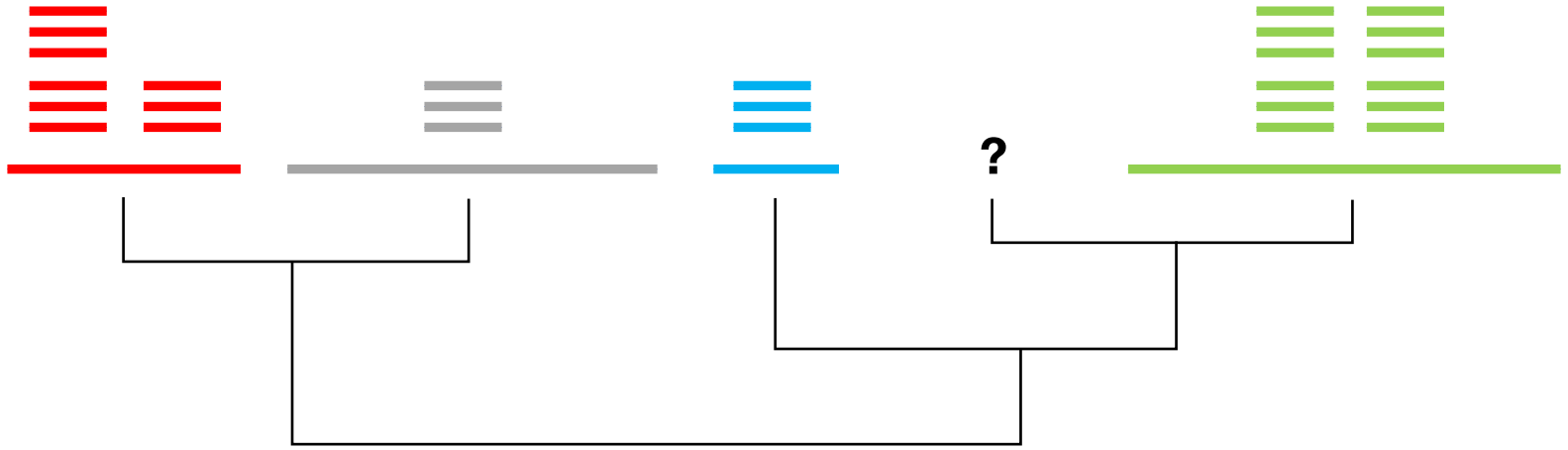
Taxonomic profiling – incomplete reference databases

Environmental sample

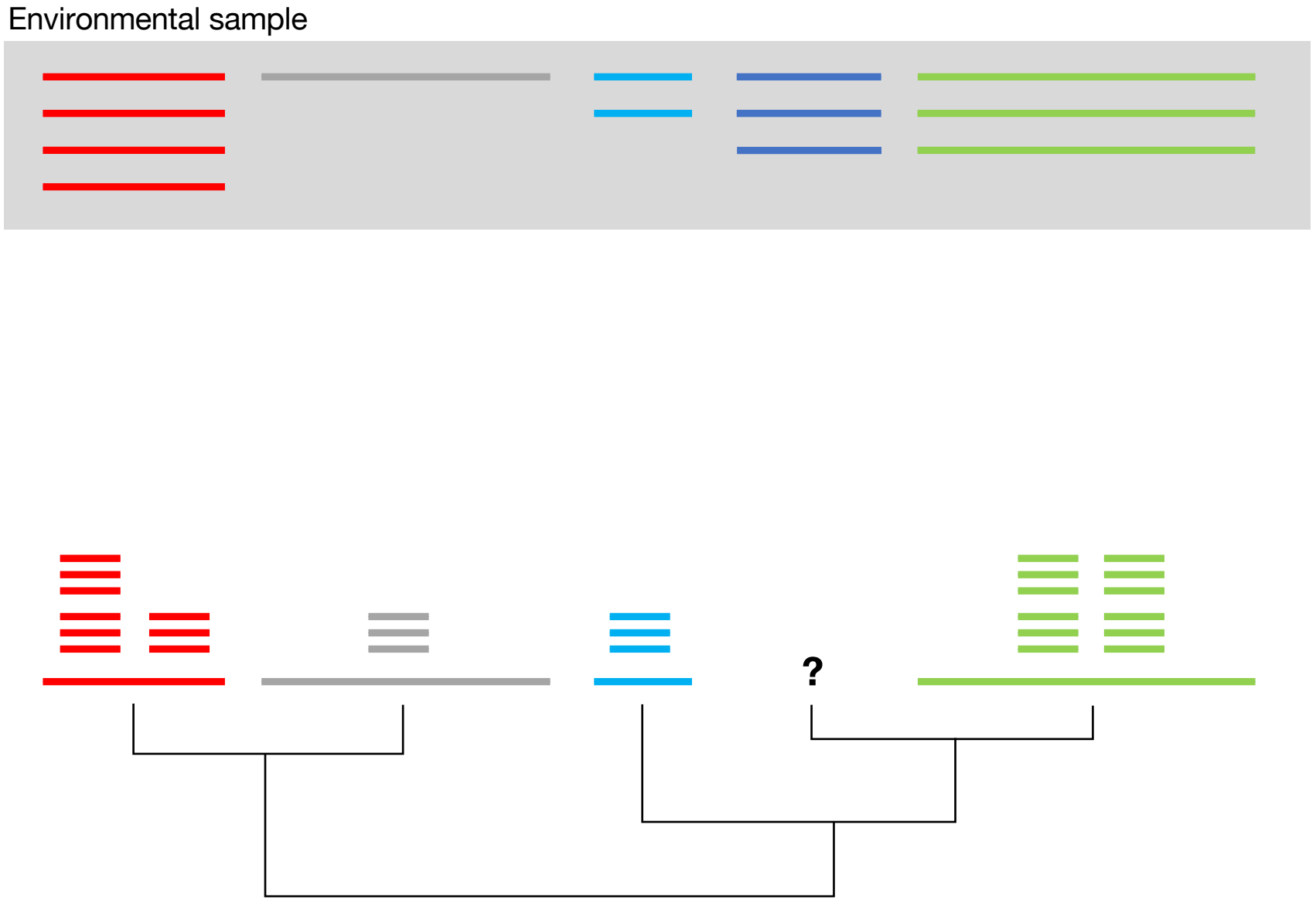
True taxonomic annotation

Estimated when dark green is missing

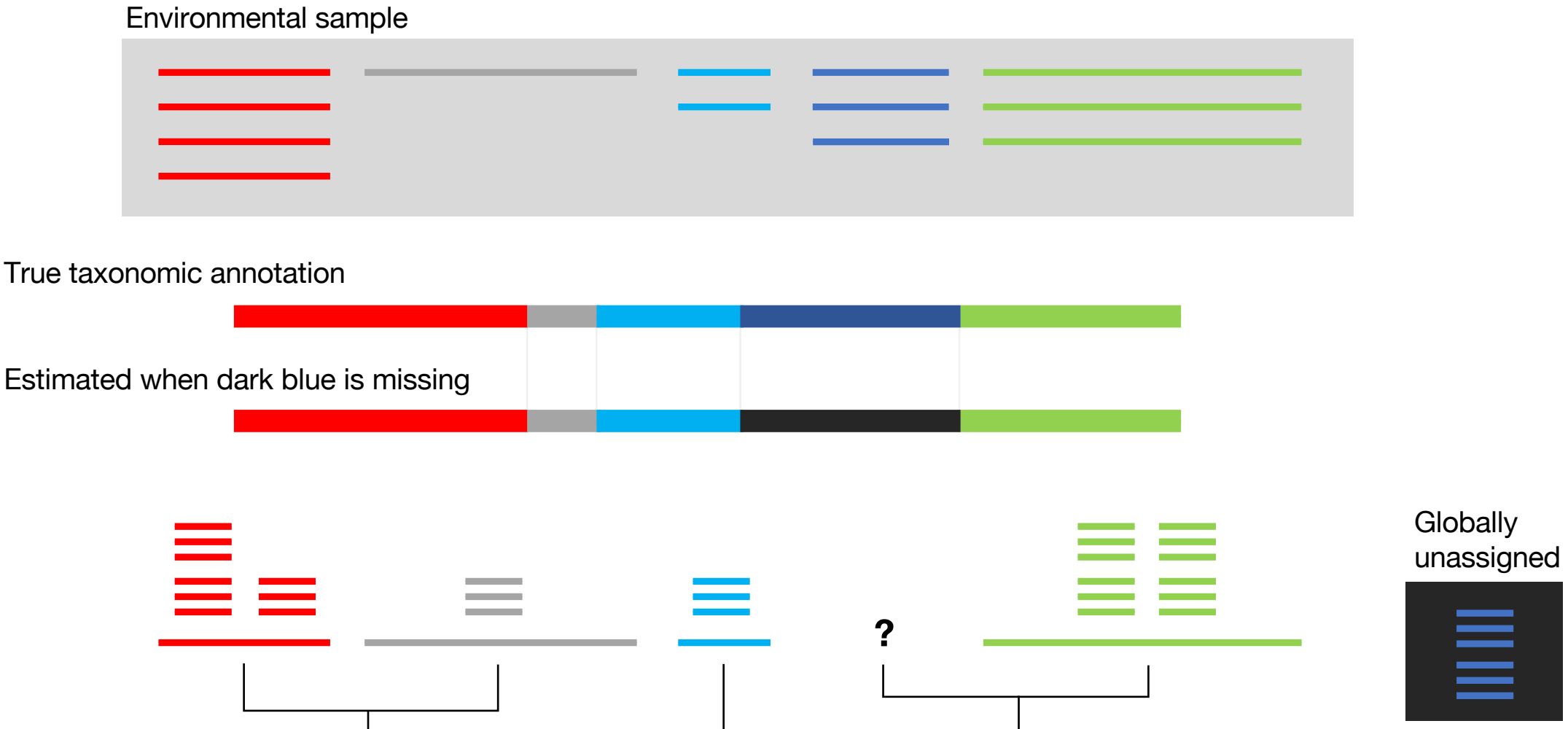
- ignore the reads



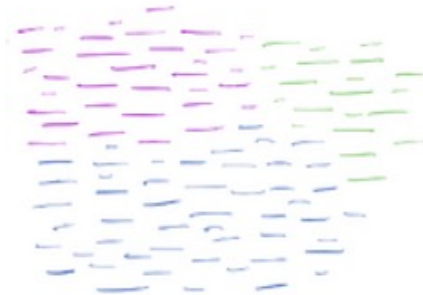
Taxonomic profiling – incomplete reference databases



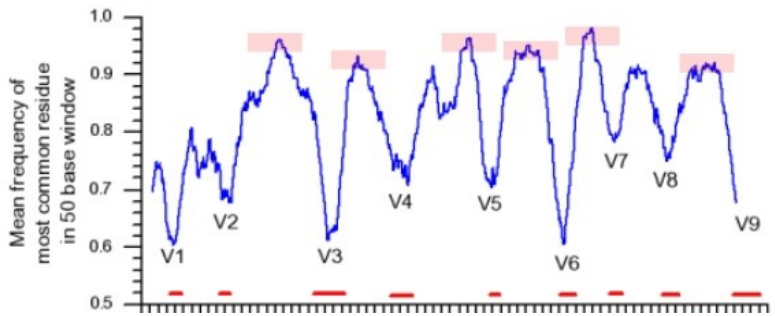
Taxonomic profiling – incomplete reference databases



Trimmed, filtered reads



16S rRNA



› Bioinformatics. 2017 Dec 1;33(23):3808-3810. doi: 10.1093/bioinformatics/btx517.

MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis

João F Matias Rodrigues [1], Thomas S B Schmidt [1], Janko Tackmann [1], Christian von Mering [1]

Affiliations + expand

PMID: 28961926 PMCID: PMC5860325 DOI: 10.1093/bioinformatics/btx517

Free PMC article

10 universal protein marker genes

nature › nature communications › articles › article

Article | Open Access | Published: 04 March 2019

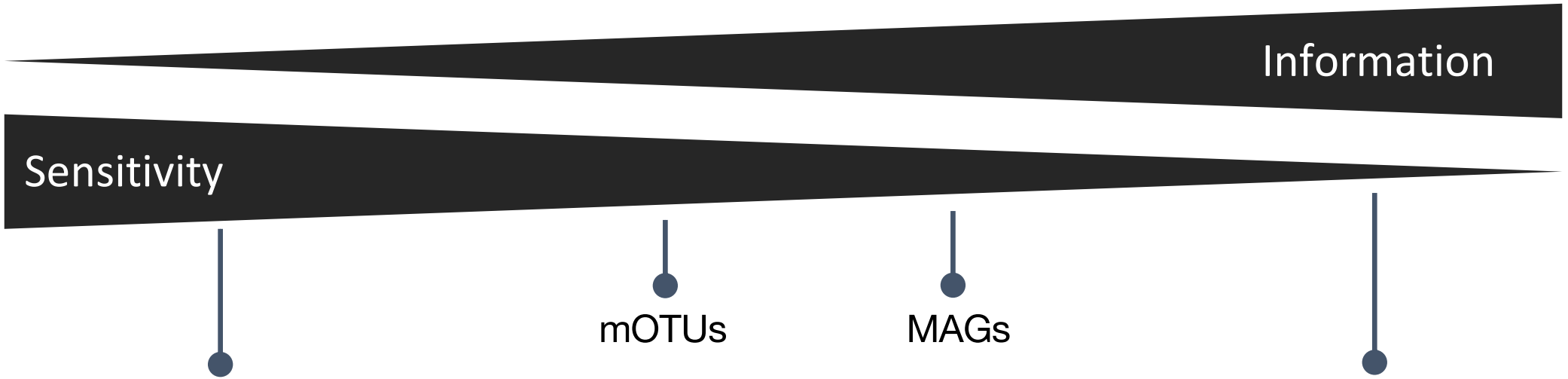
Microbial abundance, activity and population genomic profiling with mOTUs2

Alessio Milanese, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, Renato Alves, Paul I Costea, Luis Pedro Coelho, Thomas S. B. Schmidt, Alexandre Almeida, Alex L Mitchell, Robert D. Finn, Jaime Huerta-Cepas, Peer Bork, Georg Zeller ✉ & Shinichi Sunagawa ✉

Nature Communications **10**, Article number: 1014 (2019) | Cite this article

25k Accesses | **107** Citations | **78** Altmetric | Metrics

Strengths and weaknesses of different approaches



16S amplicon profiling

- Taxonomic profiling
- High sensitivity (discover the unknown)
- Largest available databases
- No functional profiling
- Sequencing errors
- PCR amplification errors
- Copy number bias
- Lower resolution

whole-genome profiling

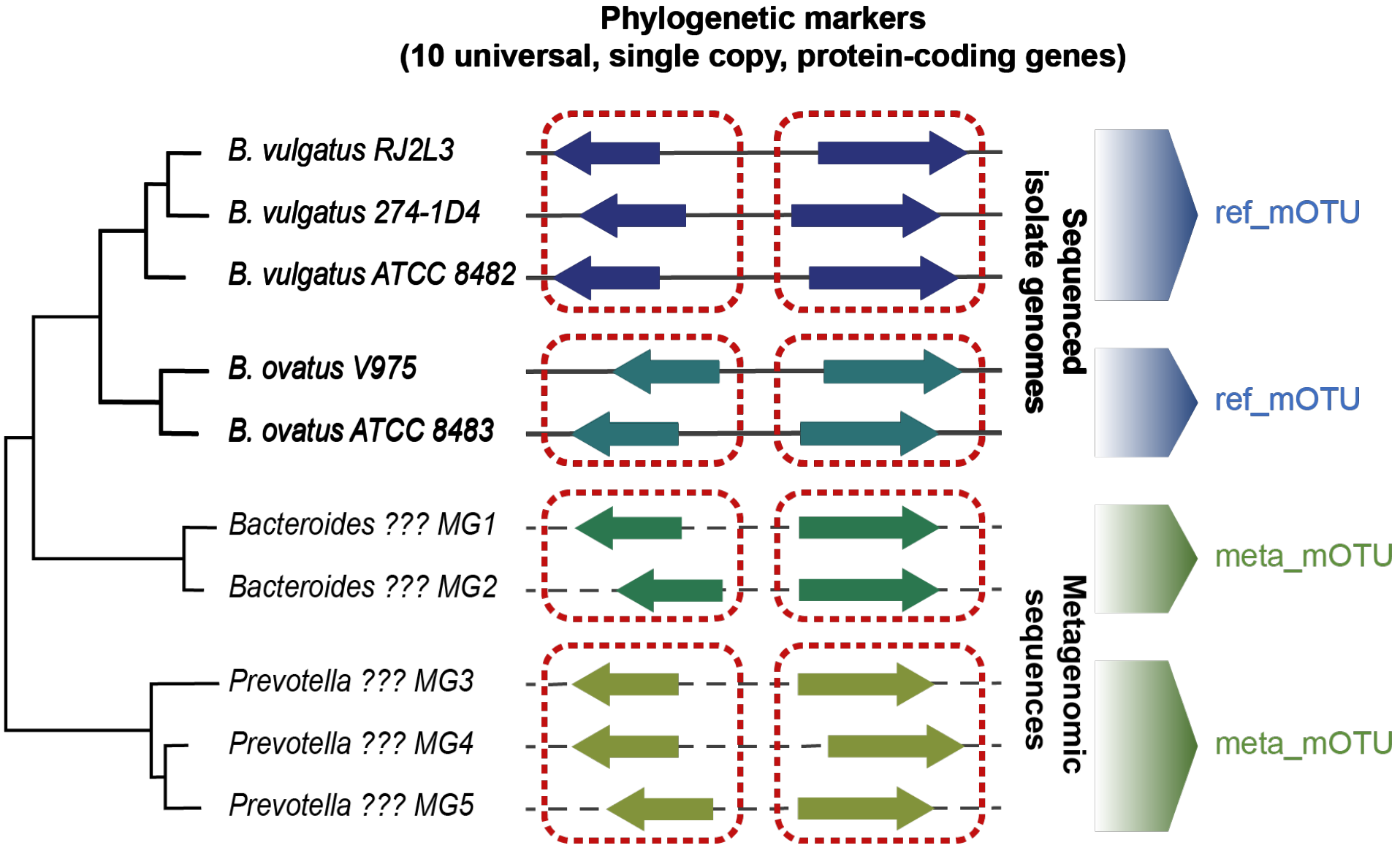
- Taxonomic profiling
- Need to estimate unknown species
- Functional profiling (potential)
- Sequencing errors
- High resolution

culture-based analysis

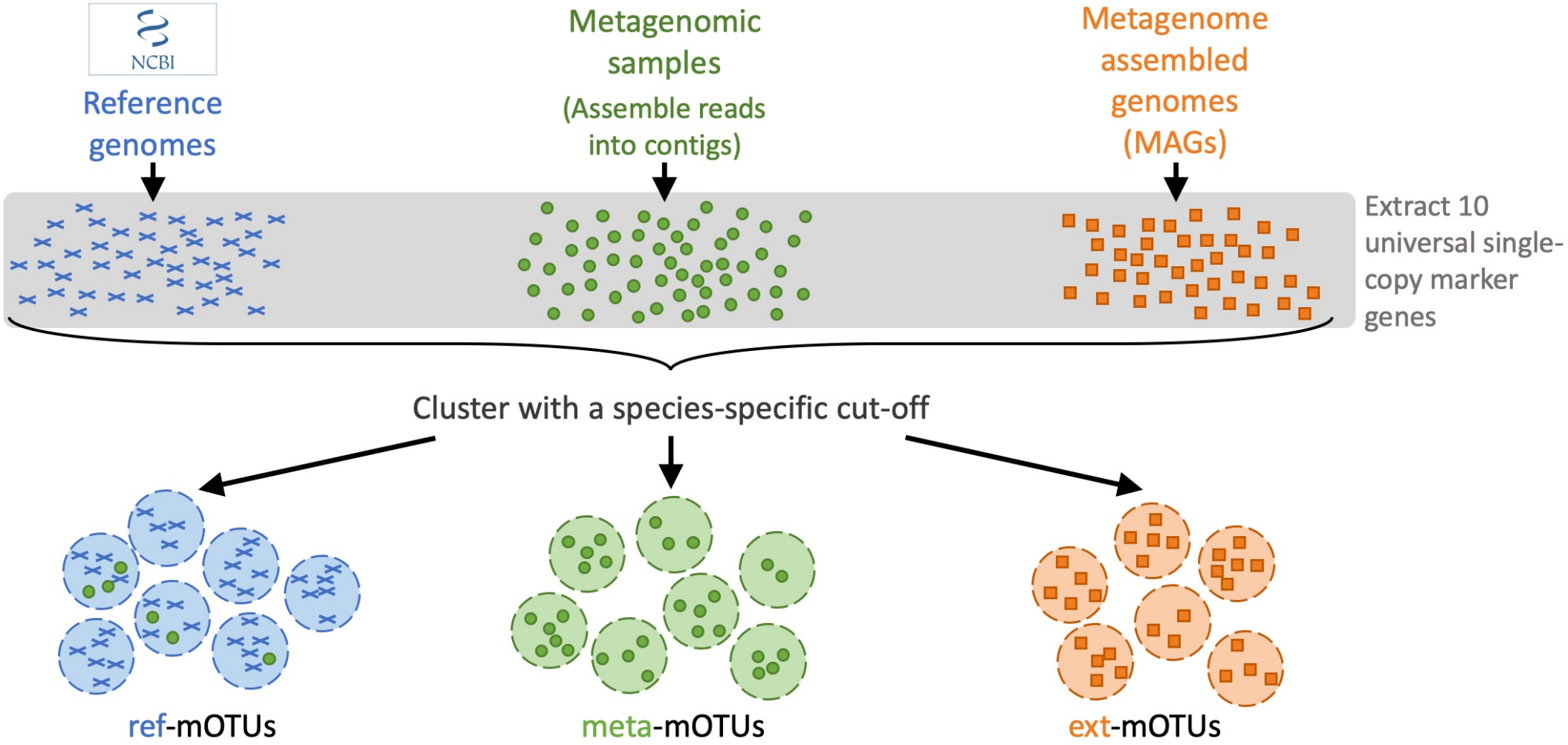
- Taxonomic profiling is limited to cultivable species
- Full potential for functional activity that can be tested
- Best resolution

The mOTUs framework – DB construction

[Ciccarelli et al. *Science* 2006]
[Sunagawa et al. *Nat. Methods* 2013]
[Milanese et al. *Nat. Commun.* 2019]

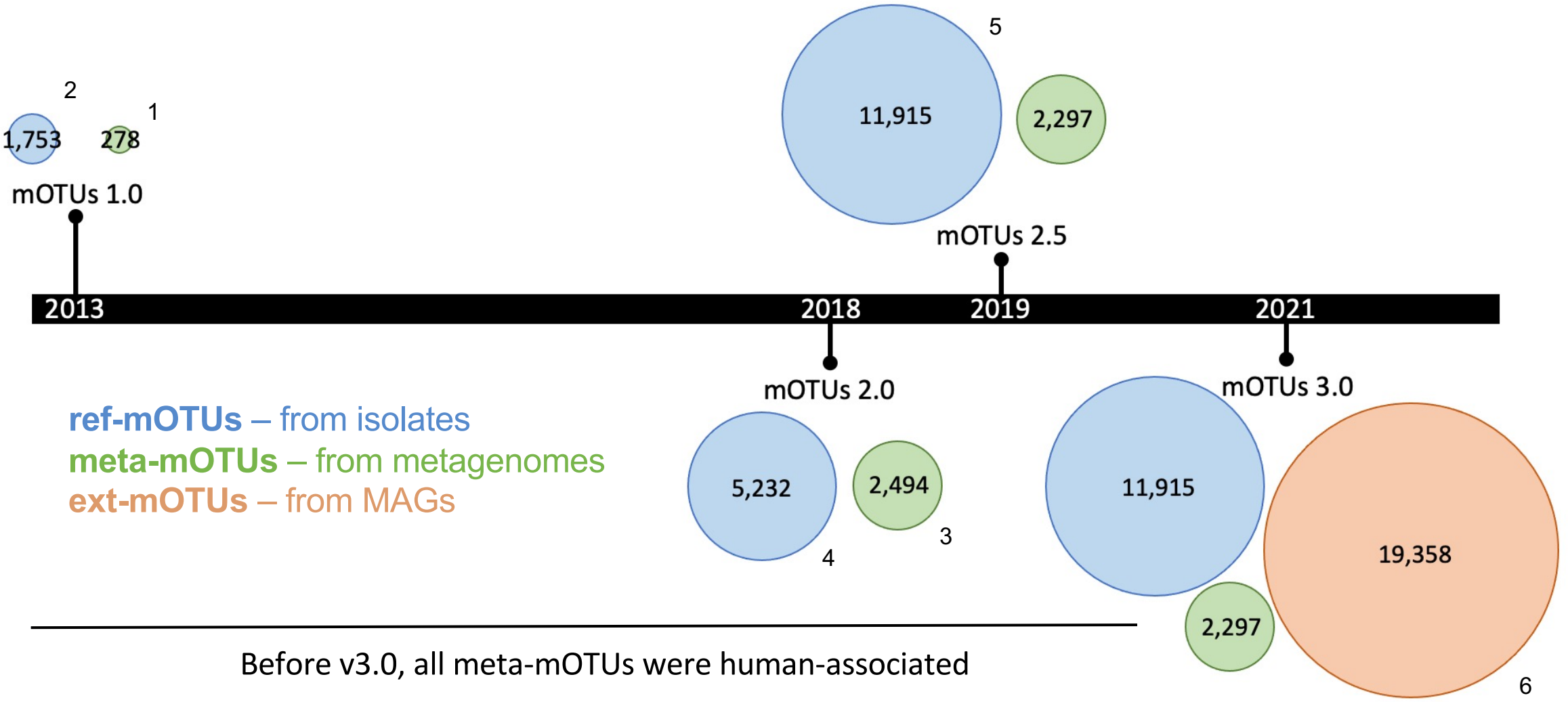


Incorporation of MAGs into the mOTUs3 database



MAG-derived mOTUs are called ext_mOTUs

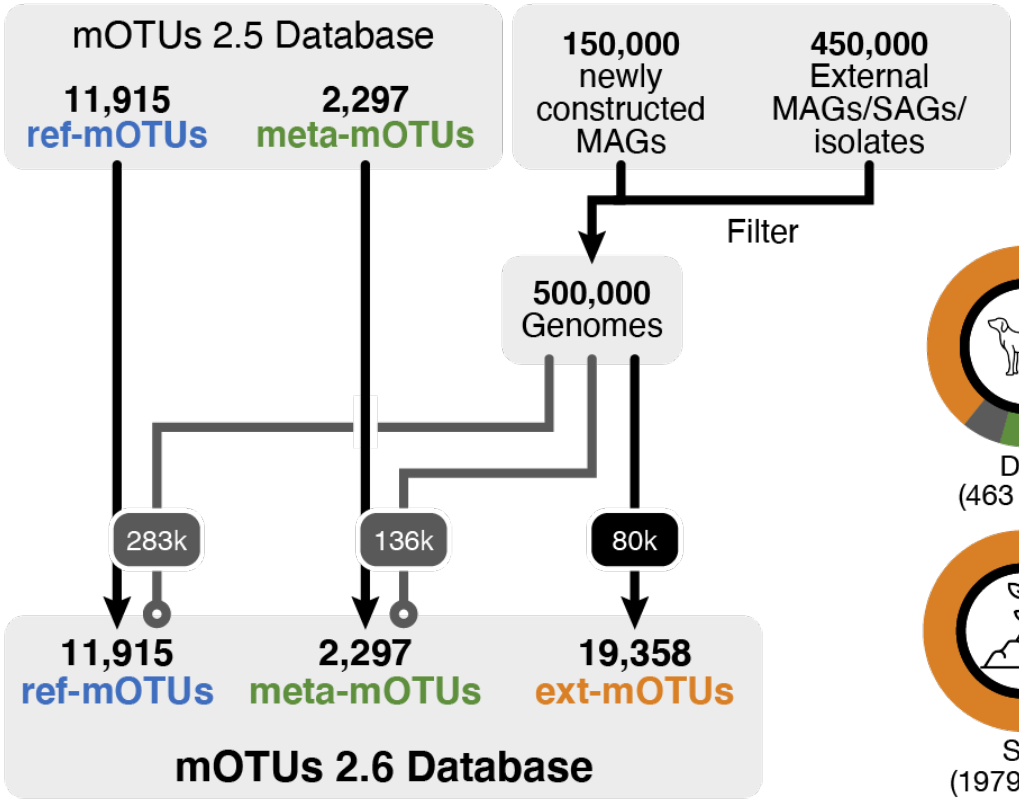
Improvement of scope in mOTUs since first version

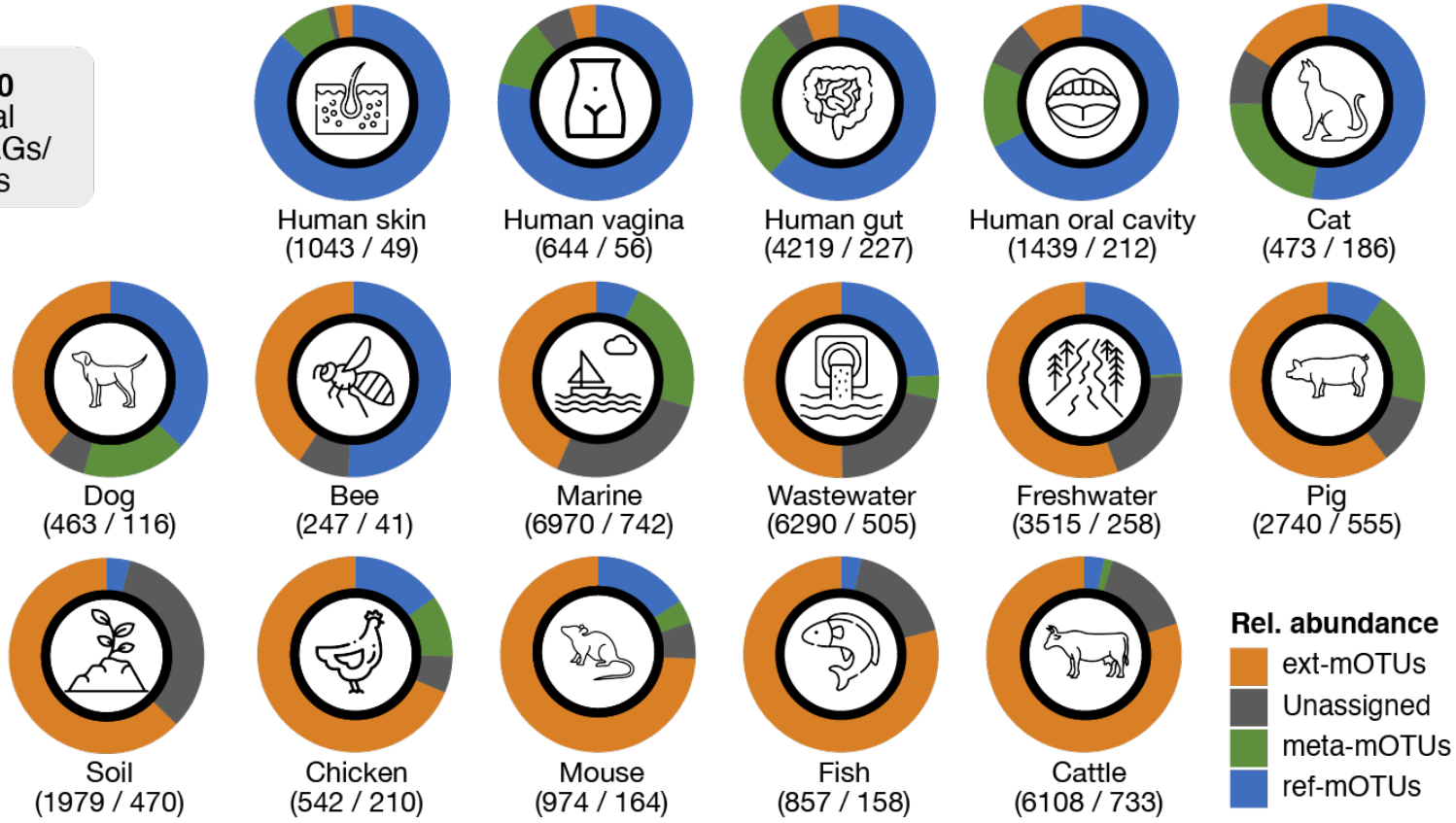


Before v3.0, all meta-mOTUs were human-associated

1. [Sunagawa et al., *Nat. Methods* 2013]
2. [Mende et al. *Nat. Methods* 2013]
3. [Milanese et al., *Nat. Commun.* 2019]
4. [Mende et al., Nucleic Acids Res. 2017]
5. [Mende et al., Nucleic Acids Res. 2020]
6. [Ruscheweyh, Milanese et al. bioRxiv 2021]

mOTUs3 – database extension by marker genes from metagenome-assembled genomes (>500,000 MAGs)



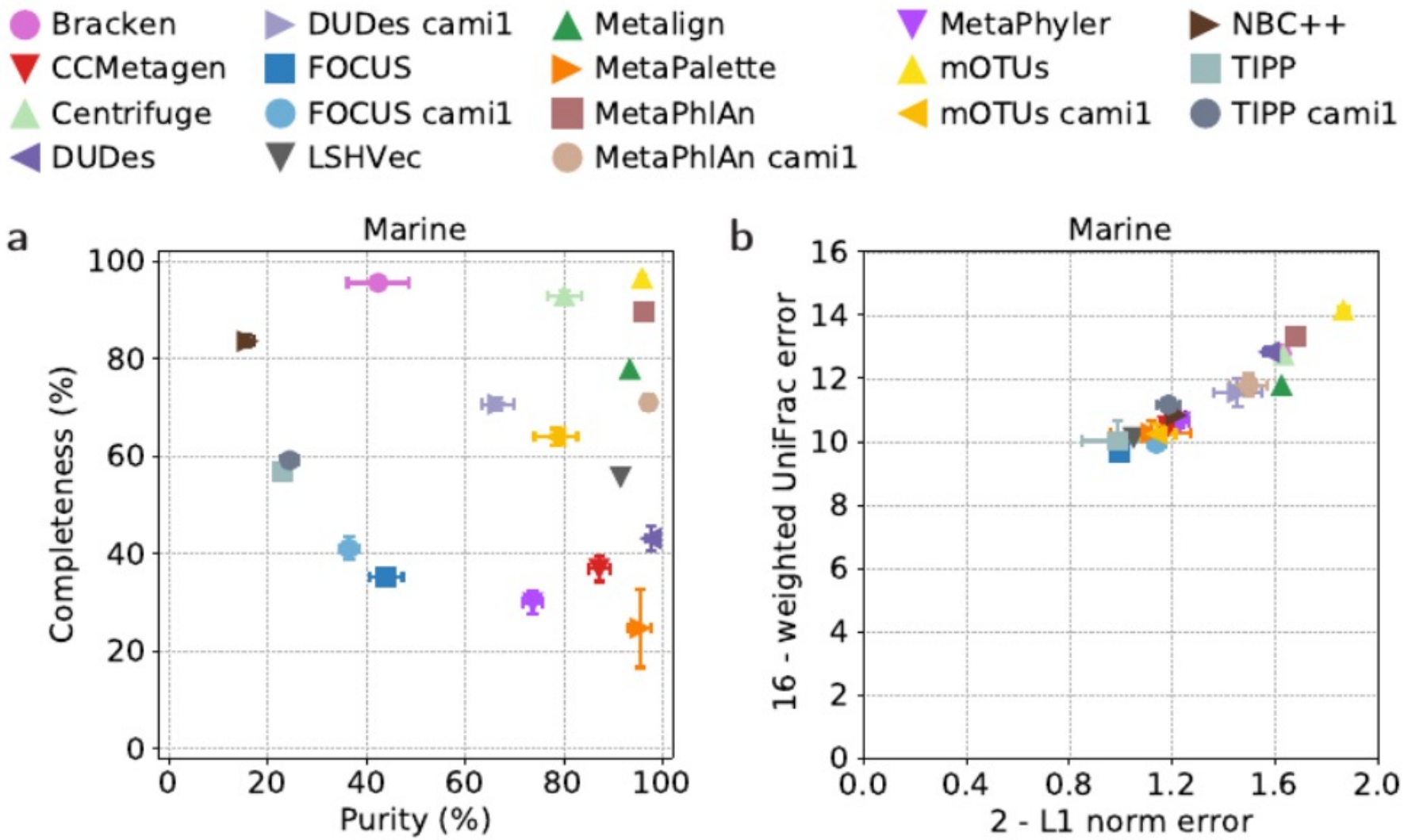


Enables profiling an unprecedented diversity of prokaryotes (33,570 species) across many environments.

motu-tool.org

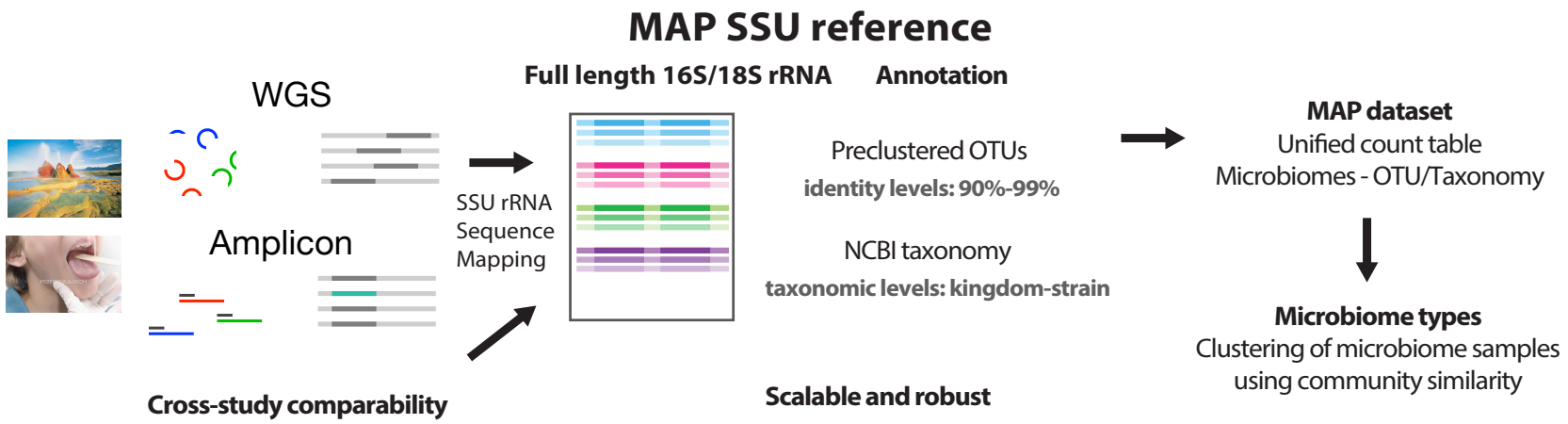
[Ruschewey, Milanese et al., bioRxiv 2021]

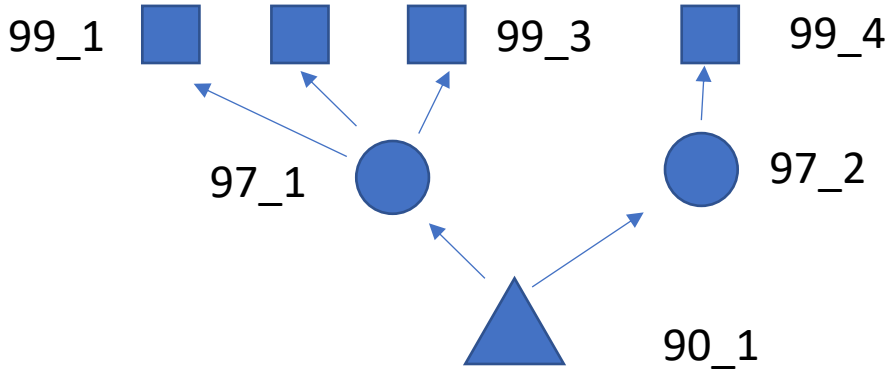
High-accuracy profiling as evaluated by an independent benchmark - CAMI

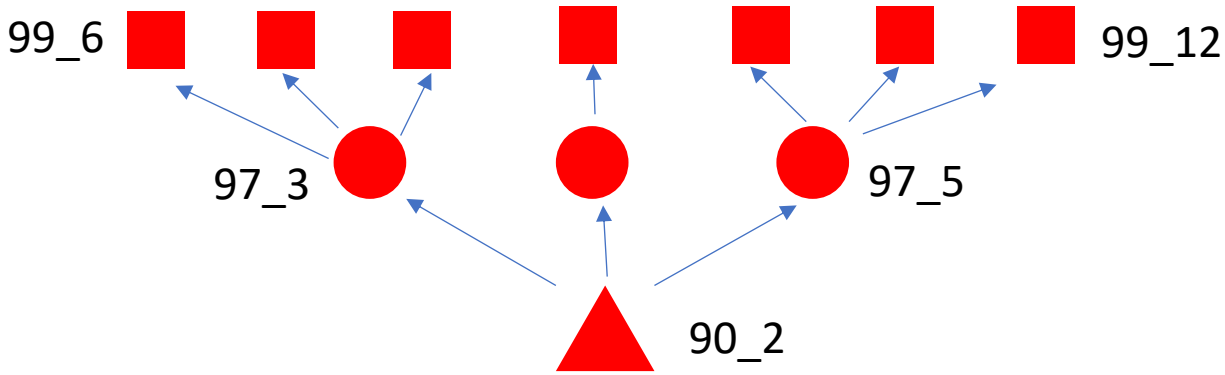


[Meyer et al., *Accepted in Nature Methods*, 2022]

MAPseq


MAP SSU reference
Full length 16S/18S rRNA
Annotation
WGS
Amplicon
SSU rRNA
Sequence
Mapping
Preclustered OTUs
identity levels: 90%-99%
NCBI taxonomy
taxonomic levels: kingdom-strain
MAP dataset
Unified count table
Microbiomes - OTU/Taxonomy
Microbiome types
Clustering of microbiome samples
using community similarity
Cross-study comparability
Scalable and robust

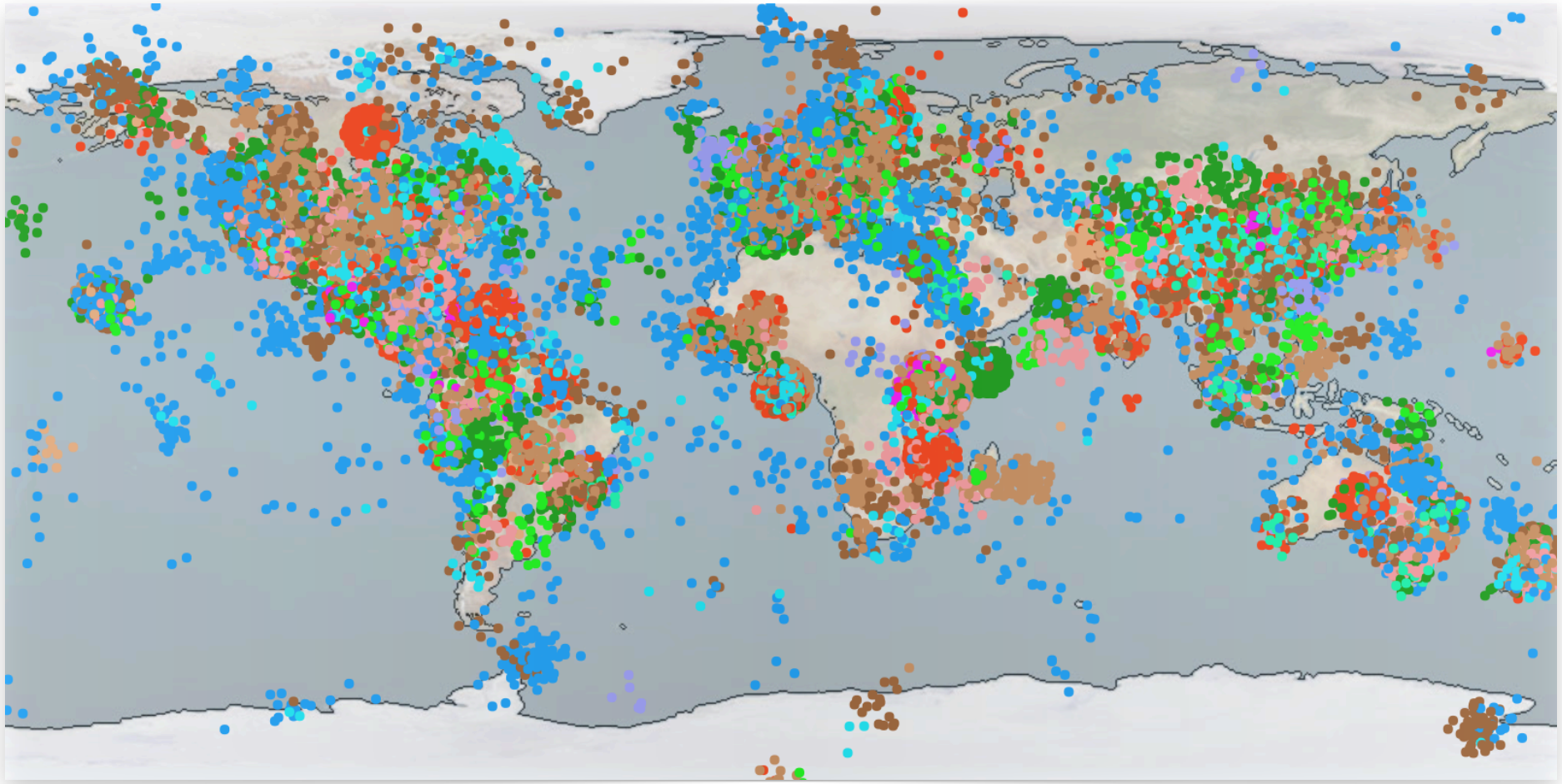

99_1
99_3
99_4
97_1
97_2
90_1


99_6
99_12
97_3
97_5
90_2

Microbe Atlas Project

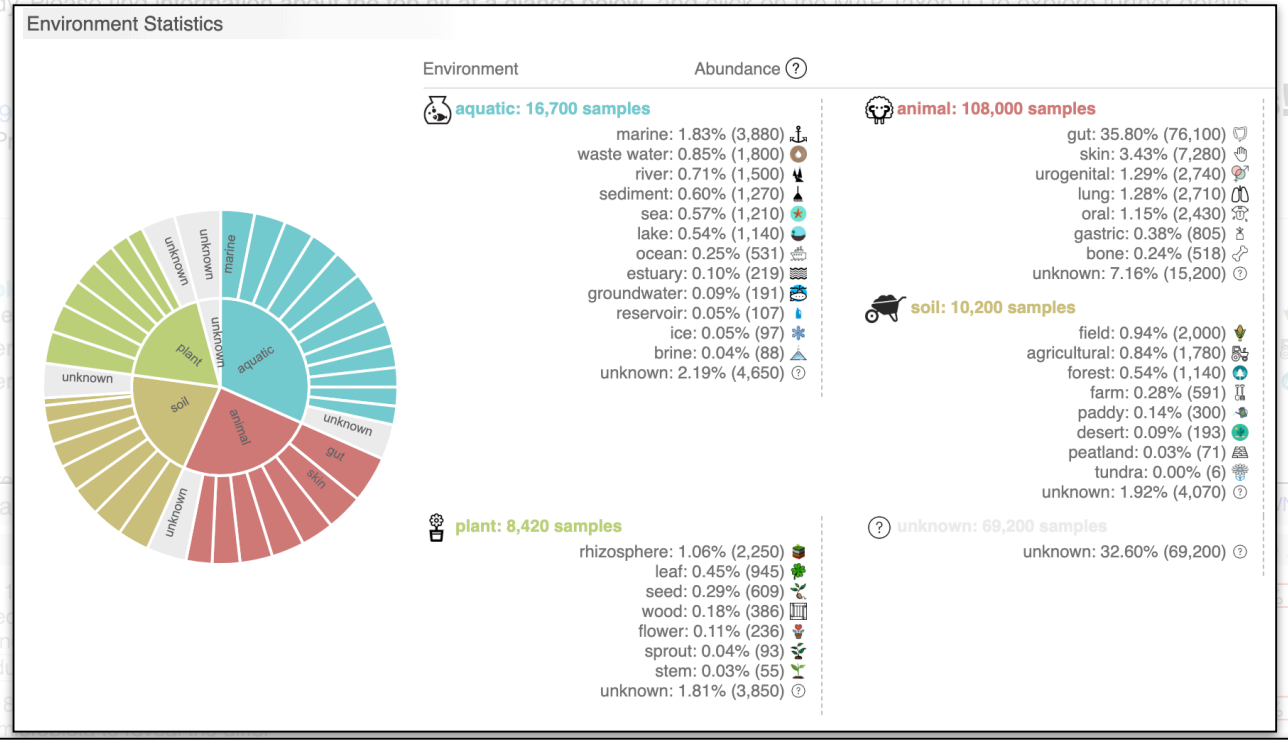
microbeatlas.org

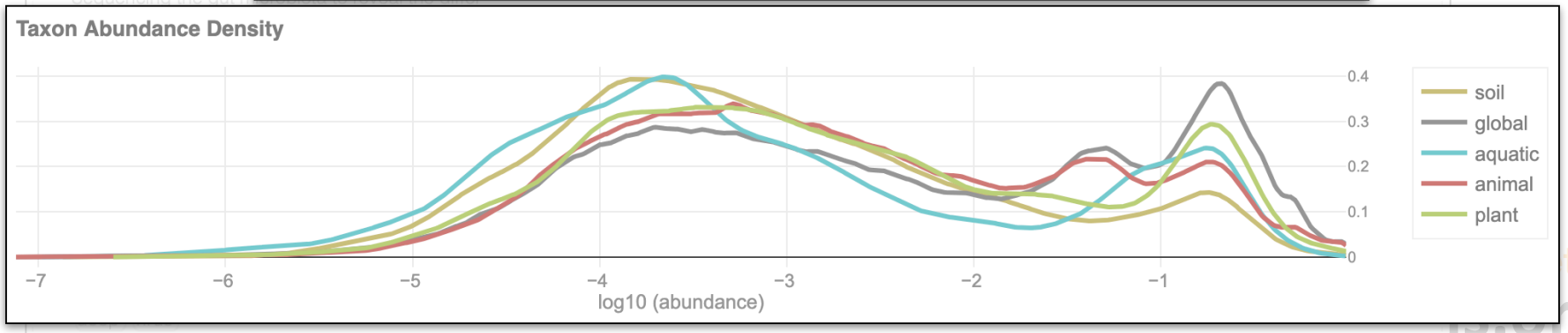
compare your metagenomic data to a global reference set of a million microbiome samples



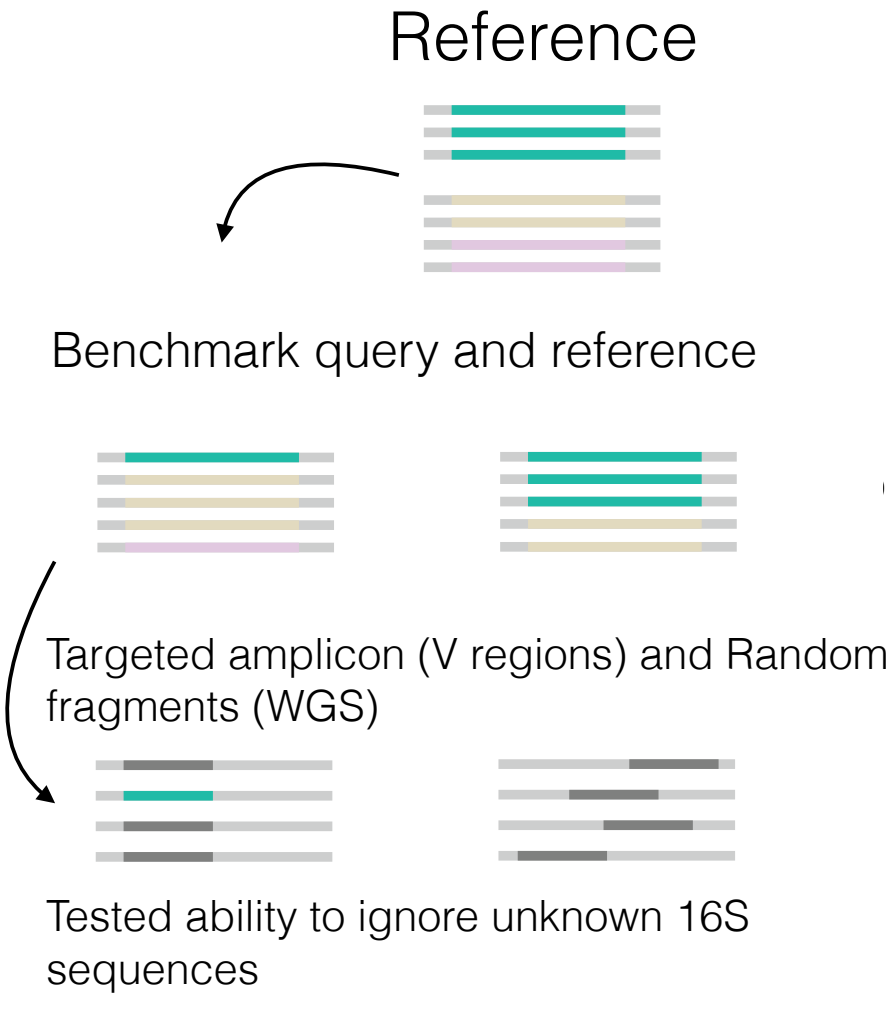
Microbe Atlas Project Website

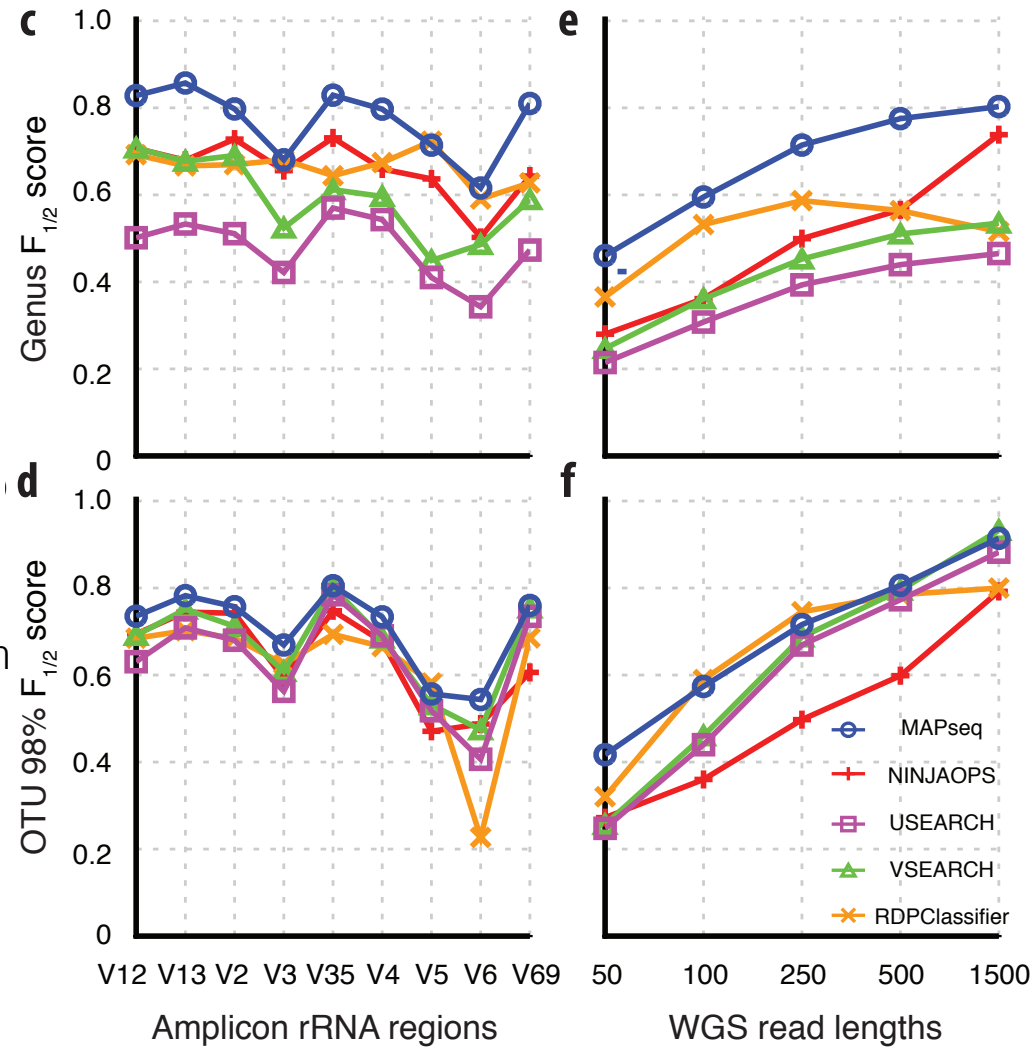
Search for any microbial taxa by sequence or name





Benchmark of accuracy on known taxonomy

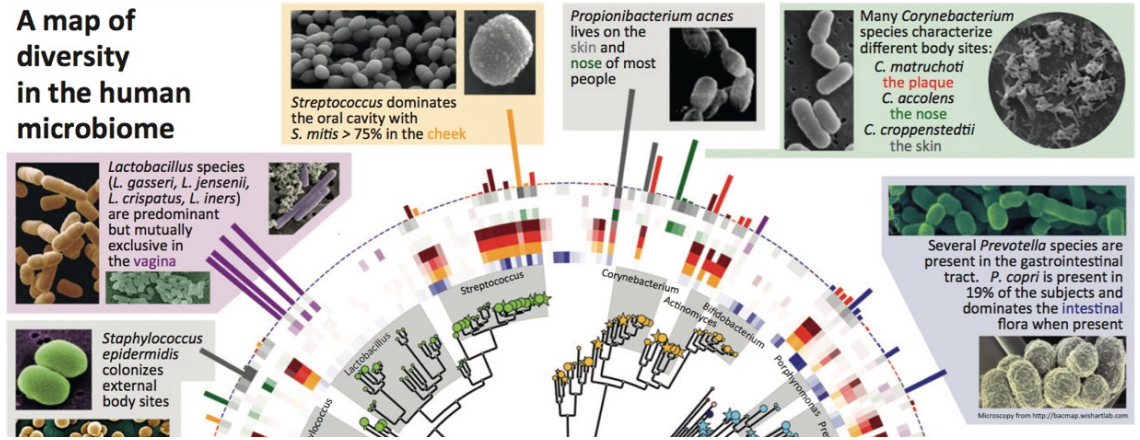




Taxonomic profiling – why it is important?

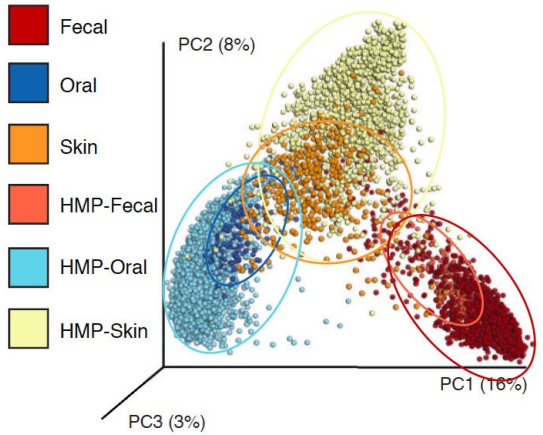
Taxonomic analysis is fundamental to the analysis of microbial communities

Describing the microbial community under study



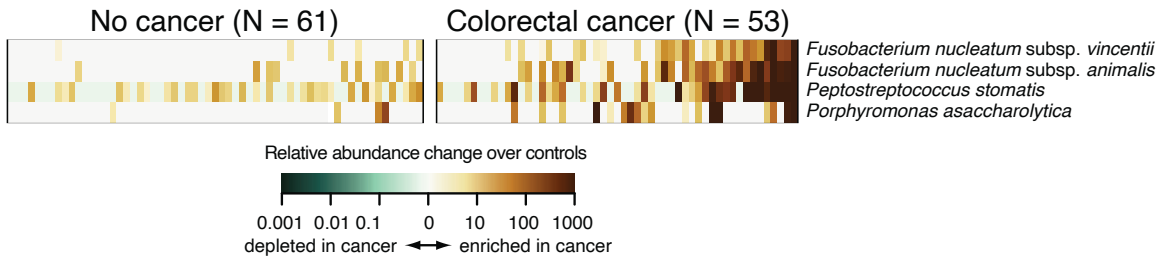
[Morgan et al., *Trends in Genetics*, 2013]

Comparing different microbial communities



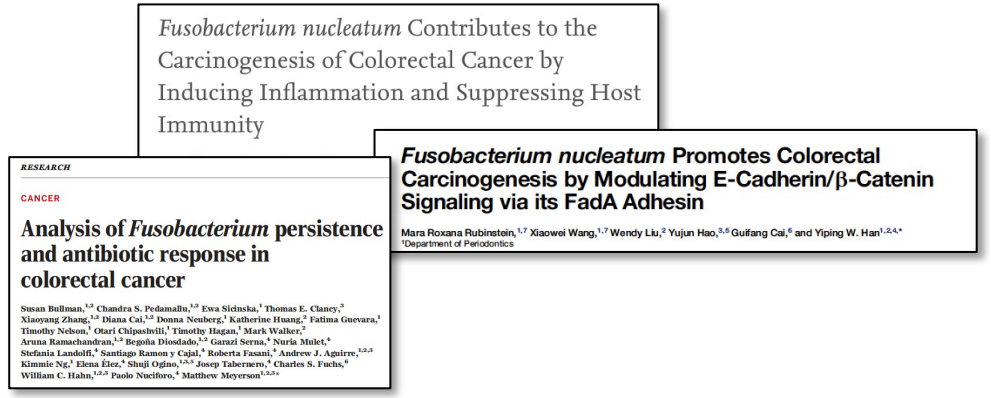
[McDonald et al., mSystems, 2018]

Correlating environm. or host features to microbes



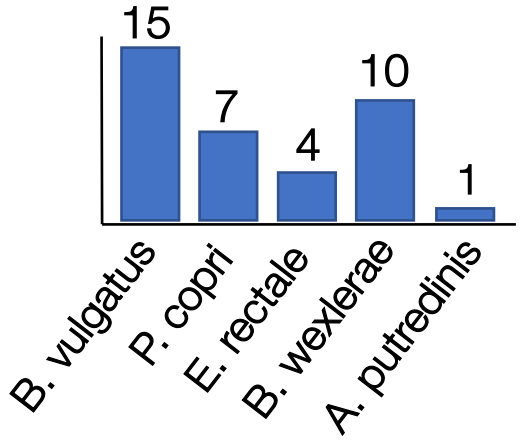
[Zeller et al., MSB, 2014]

Comparing findings to literature



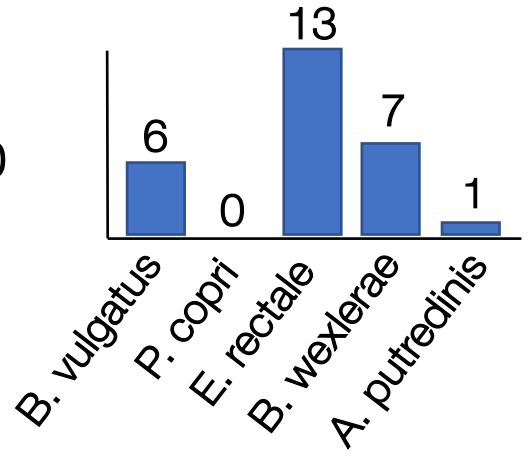
Profiling multiple samples

Sample 1



⋮

Sample 10



	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	15	0	9	6	9	21	3	0	45	6
P. copri	7	11	0	0	12	0	6	0	0	0
E. rectale	4	4	0	4	0	7	0	0	0	13
B. wexlerae	10	0	2	0	0	5	0	0	4	7
A. putredinis	1	0	0	0	0	3	0	0	0	1
E. coli	0	3	12	0	0	5	0	4	1	0
C. innocuum	0	2	0	0	0	1	2	8	0	6
R. intestinalis	12	0	0	6	4	0	5	2	0	0
A. finegoldii	6	1	1	0	0	0	2	0	0	23

Profiling multiple samples – Library size

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	4	6	12	2	4	0	2	3	0	9
P. copri	6	2	4	1	4	8	6	1	5	0
E. rectale	3	0	0	8	1	2	0	0	3	3
B. wexlerae	0	3	6	0	8	4	4	3	4	0
A. putredinis	0	0	0	6	0	14	1	0	0	7
E. coli	0	0	0	0	0	12	6	8	21	4
C. innocuum	0	1	2	0	4	2	1	1	0	5
R. intestinalis	5	1	2	0	2	3	9	0	2	0
A. finegoldii	0	0	0	1	0	0	0	0	0	0

Profiling multiple samples – Library size

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	4	6	12	2	4	0	2	3	0	9
P. copri	6	2	4	1	4	8	6	1	5	0
E. rectale	3	0	0	8	1	2	0	0	3	3
B. wexlerae	0	3	6	0	8	4	4	3	4	0
A. putredinis	0	0	0	6	0	14	1	0	0	7
E. coli	0	0	0	0	0	12	6	8	21	4
C. innocuum	0	1	2	0	4	2	1	1	0	5
R. intestinalis	5	1	2	0	2	3	9	0	2	0
A. finegoldii	0	0	0	1	0	0	0	0	0	0
SUM	18	13	26	18	23	45	29	16	35	28

Profiling multiple samples – Library size

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	4	6	12	2	4	0	2	3	0	9
P. copri	6	2	4	1	4	8	6	1	5	0
E. rectale	3	0	0	8	1	2	0	0	3	3
B. wexlerae	0	3	6	0	8	4	4	3	4	0
A. putredinis	0	0	0	6	0	14	1	0	0	7
E. coli	0	0	0	0	0	12	6	8	21	4
C. innocuum	0	1	2	0	4	2	1	1	0	5
R. intestinalis	5	1	2	0	2	3	9	0	2	0
A. finegoldii	0	0	0	1	0	0	0	0	0	0
SUM	18	13	26	18	23	45	29	16	35	28

Profiling multiple samples – Relative abundance

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0

Profiling multiple samples – Richness

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0
	↓					↓				
	4					5				

- The richness is calculated per sample
- It represents the total number of species observed in a sample

Profiling multiple samples – Prevalence

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10	
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3	→ 8
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0	
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1	
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0	
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3	
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1	
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2	
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0	
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0	→ 1

- The prevalence is calculated per species
- It measure the number of sample where the species is detected

Profiling multiple samples – Prevalence

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6	Sample 7	Sample 8	Sample 9	Sample 10
B. vulgatus	0.2	0.5	0.5	0.1	0.2	0	0.1	0.2	0	0.3
P. copri	0.3	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0
E. rectale	0.2	0	0	0.4	0	0	0	0	0.1	0.1
B. wexlerae	0	0.2	0.2	0	0.3	0.1	0.1	0.2	0.1	0
A. putredinis	0	0	0	0.3	0	0.3	0	0	0	0.3
E. coli	0	0	0	0	0	0.3	0.2	0.5	0.6	0.1
C. innocuum	0	0.1	0.1	0	0.2	0	0	0.1	0	0.2
R. intestinalis	0.3	0.1	0.1	0	0.1	0.1	0.3	0	0.1	0
A. finegoldii	0	0	0	0.1	0	0	0	0	0	0

→ 8 (0.8)

→ 1 (0.1)

- The prevalence is calculated per species
- It measure the number of sample where the species is detected
- It can also be represented as fraction of the total amount of samples