DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

Microbial diversity with NGS-PMGA

4. Data analysis

DEPARTMENT OF
Biochemistry &
Biotechnology
UNIVERSITY OF THESSALY

Microbial diversity with Illumina-PMGA

From DNA extracts to sequencing (hack the planet)

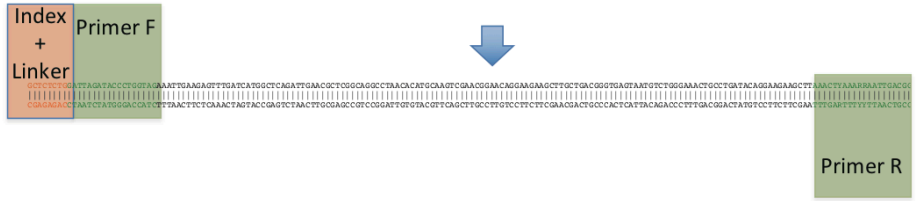
1) Low-cycle PCR (e.g. 28 cycles) with index-free primers on DNA extracts

```
Primer:                 TACGGRAGGCAGCAG
Matches:                ||||||||||||||||
Template: … NNNNNNNNN-VB-ATGCCYTCCGTCGTCNNNNNNNNNNNNNNN …
```

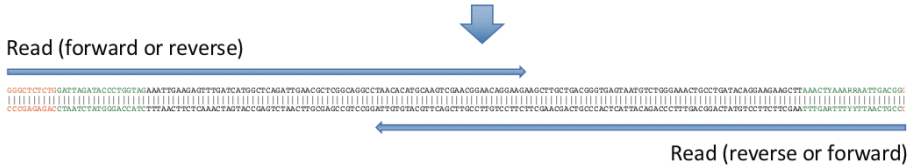
2) Sample-wise (e.g. 7-cycle) indexing on 1st PCR products on DNA extracts

```
Index:      NNNNNNNNN
                    \
Linker:              TA
                      \
Primer:                 TACGGRAGGCAGCAG
Matches:                ||||||||||||||||
Template:               ATGCCYTCCGTCGTCNNNNNNNNNNNNNNN …
```

3) Final, sample-wise indexed PCR amplicon construct



4) Multiplexing of indexed PCR products (in-house), ligation of sequencing adapters and sequencing (by vendor)



https://github.com/SotiriosVasileiadis/mconsort_tbz_degr#16s

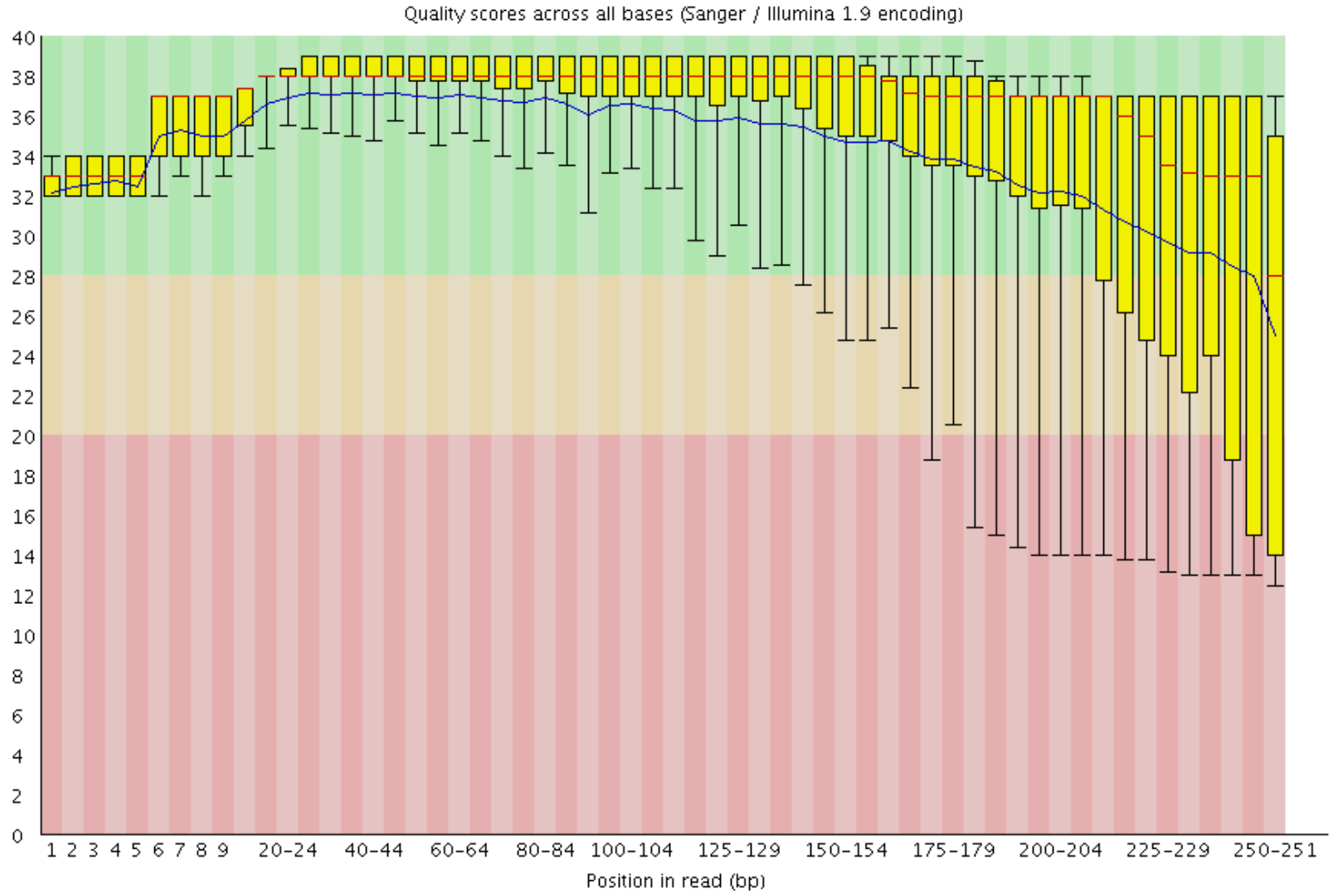
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa



DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

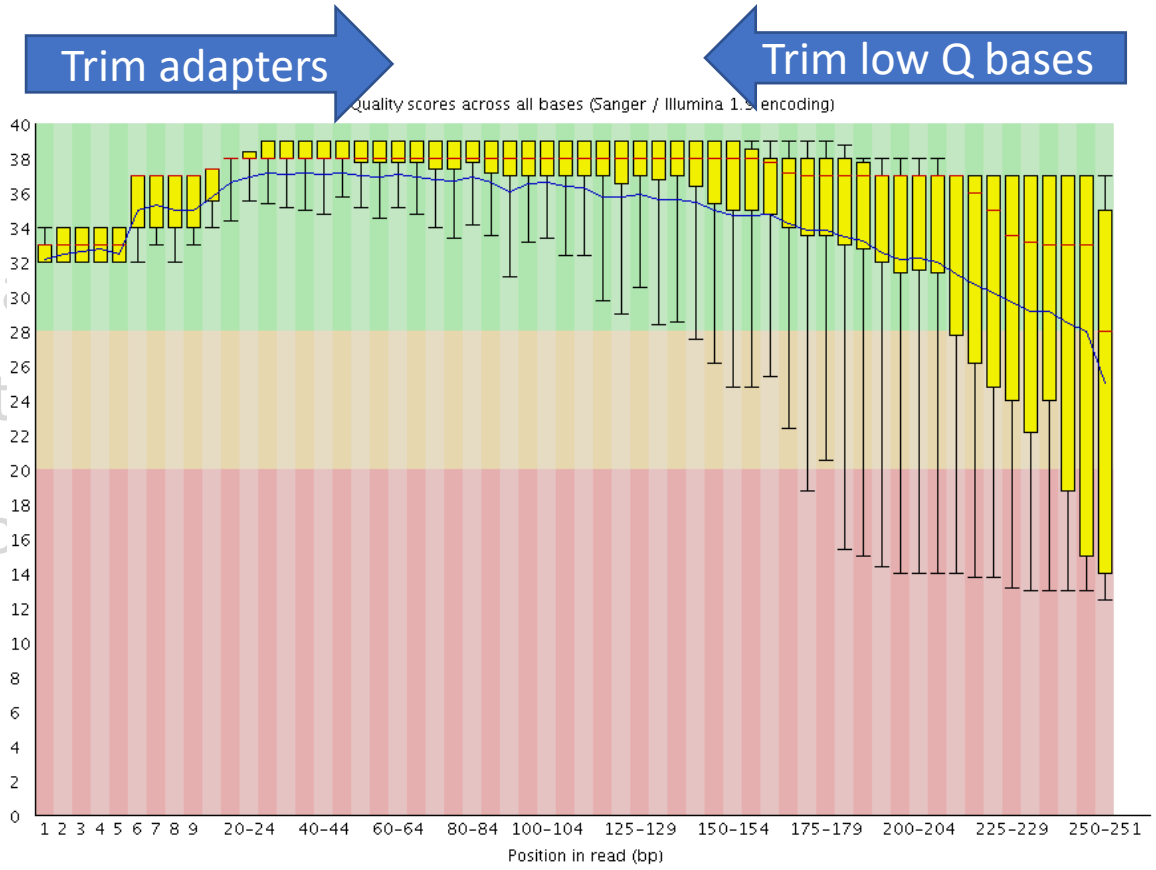
4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- **Read quality control**
- Read-pair assembly
- Formation of operational t
  amplicon sequence variant
- Removal of chimeras and c

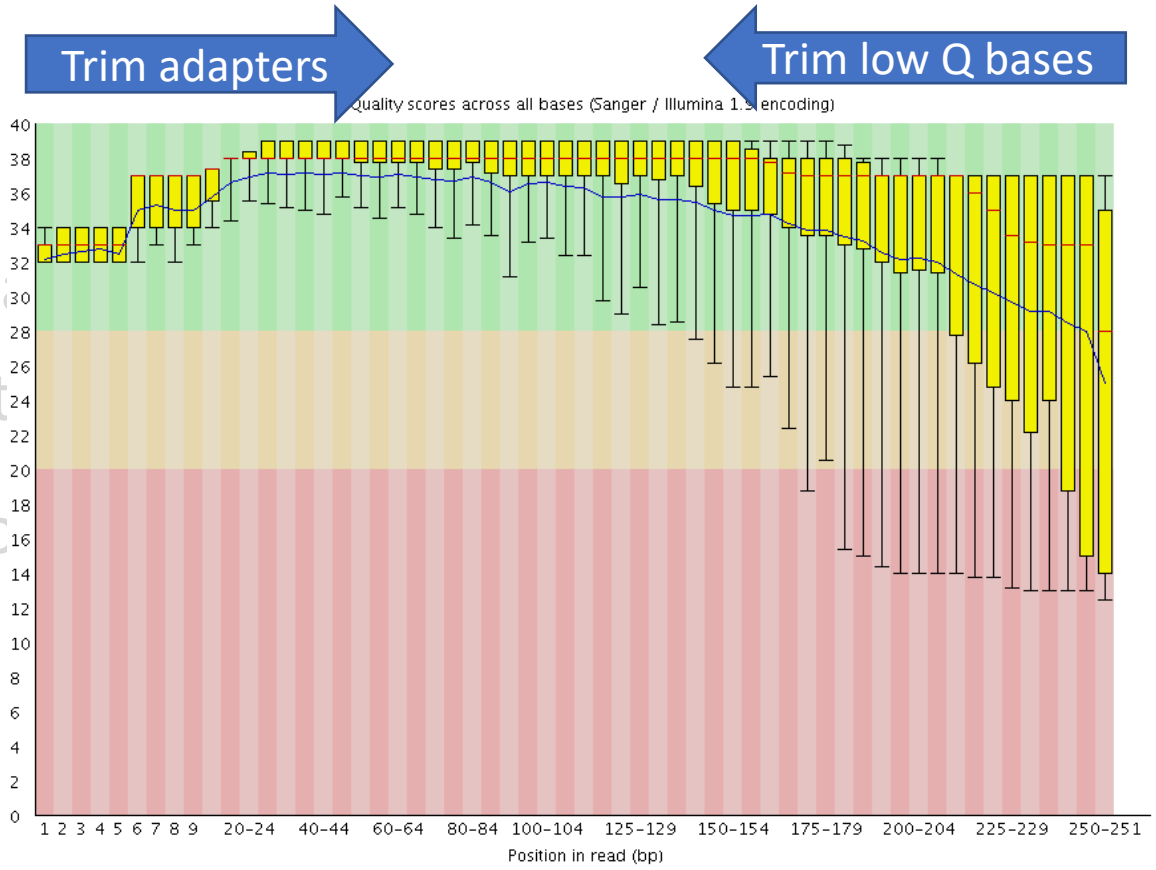


Reject low mean Q reads

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- **Read quality control**
- Read-pair assembly
- Formation of operational t
  amplicon sequence variant
- Removal of chimeras and c



Reject low mean Q reads

Reject short remains

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

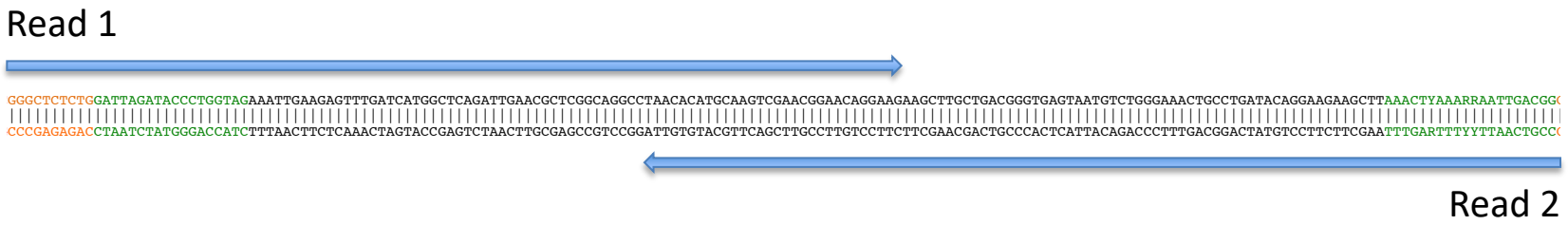
4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- **Read-pair assembly**
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences



Kozich, J.J., et al. (2013). Appl Environ Microbiol 79, 5112-5120, doi: doi:10.1128/AEM.01043-13

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

Reject (or retain) non assembled read-pairs

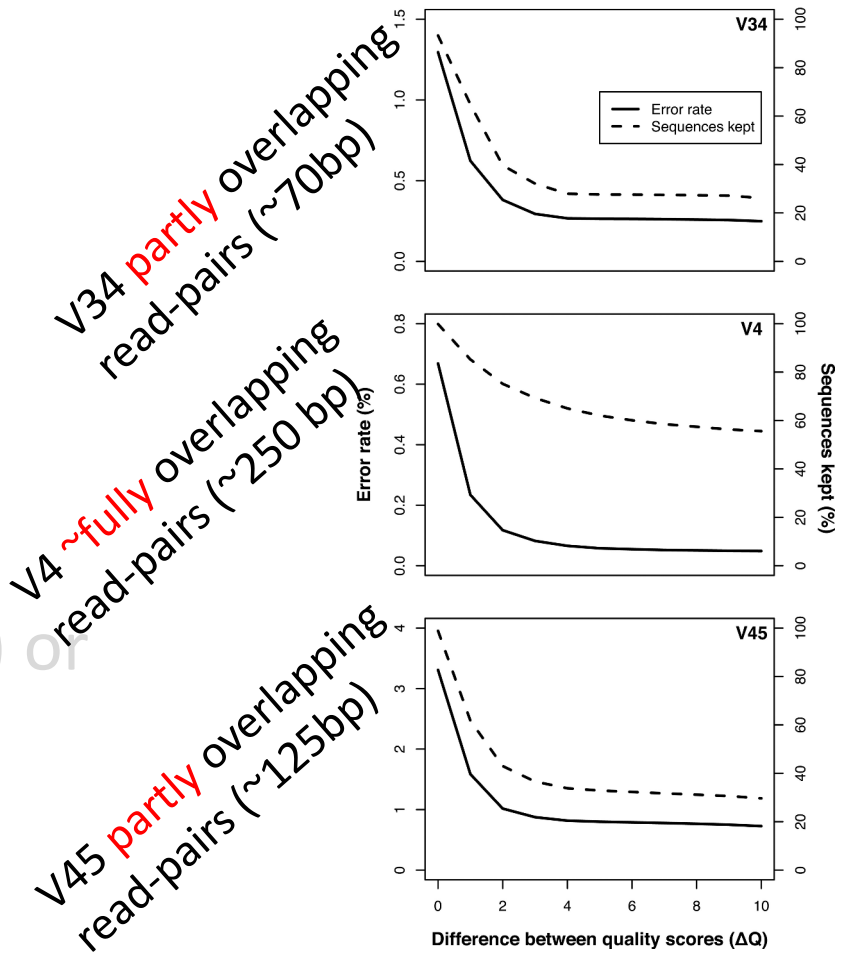
Kozich, J.J., et al. (2013). Appl Environ Microbiol 79, 5112-5120, doi: doi:10.1128/AEM.01043-13

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

>X10 better quality with full overlap



Reject (or retain) non assembled read-pairs

Kozich, J.J., et al. (2013). Appl Environ Microbiol 79, 5112-5120, doi: doi:10.1128/AEM.01043-13

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs), or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

Hughes et al 2001 AEM http://dx.doi.org/10.1128/AEM.67.10.4399-4406.2001
Konstantinidis et al 2006 http://dx.doi.org/10.1098/Rstb.2006.1920
Rodriguez-R et al 2018 http://dx.doi.org/10.1128/AEM.00014-18
Wang et al 2011 http://dx.doi.org/10.1038/ismej.2011.187
Schloss 2013 http://dx.doi.org/10.1038/ismej.2012.102

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating genomic units
- phylotypes: ASV or OTU representatives of amplicon classification in taxonomic bins

Bacterial 16S rRNA marker gene (convention)

if two sequences are 97% identical -> same OTU

Same OTU -> same species(?)

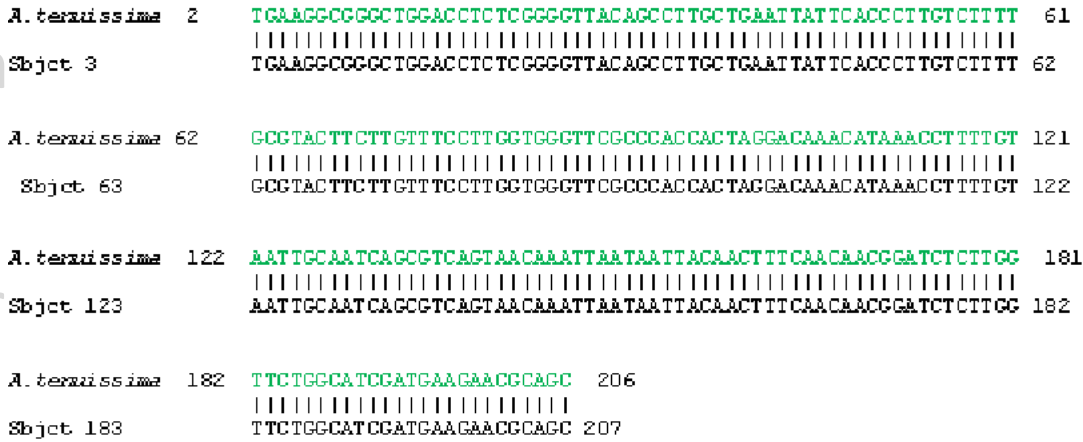
Hughes et al 2001 AEM http://dx.doi.org/10.1128/AEM.67.10.4399-4406.2001
Konstantinidis et al 2006 http://dx.doi.org/10.1098/Rstb.2006.1920
Rodriguez-R et al 2018 http://dx.doi.org/10.1128/AEM.00014-18
Wang et al 2011 http://dx.doi.org/10.1038/ismej.2011.187
Schloss 2013 http://dx.doi.org/10.1038/ismej.2012.102

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments

VS

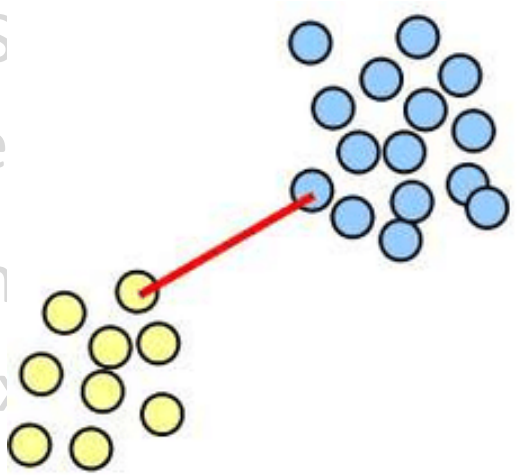


Hughes et al 2001 AEM http://dx.doi.org/10.1128/AEM.67.10.4399-4406.2001
Konstantinidis et al 2006 http://dx.doi.org/10.1098/Rstb.2006.1920
Rodriguez-R et al 2018 http://dx.doi.org/10.1128/AEM.00014-18
Wang et al 2011 http://dx.doi.org/10.1038/ismej.2011.187
Schloss 2013 http://dx.doi.org/10.1038/ismej.2012.102

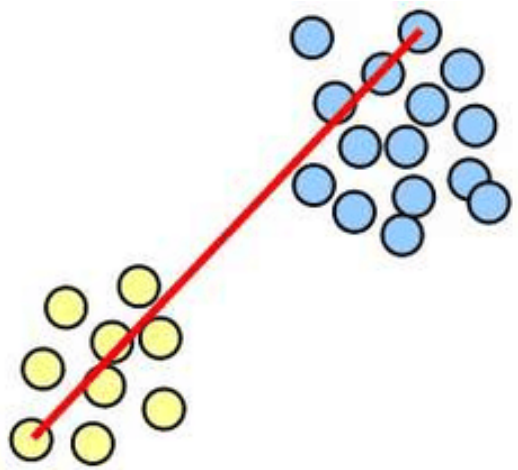
DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

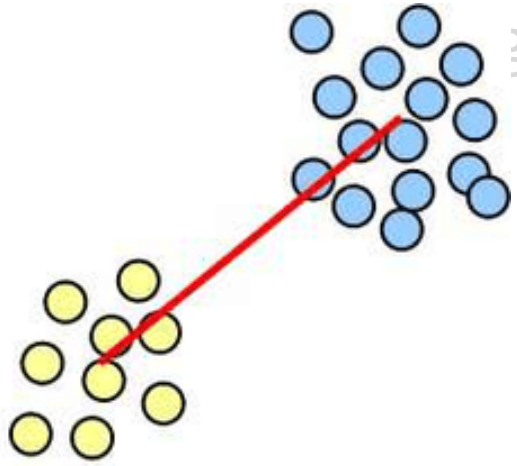
- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- AS
  ge
- ph
  ta
- ph



single-link



complete-link



average-link

Chen et al 2013 https://doi.org/10.1371/journal.pone.0070837
Wei et al 2021 https://doi.org/10.3389/fmicb.2021.644012
Lindahl et al 2017 https://doi.org/10.1111/nph.12243

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments

OTU issues:

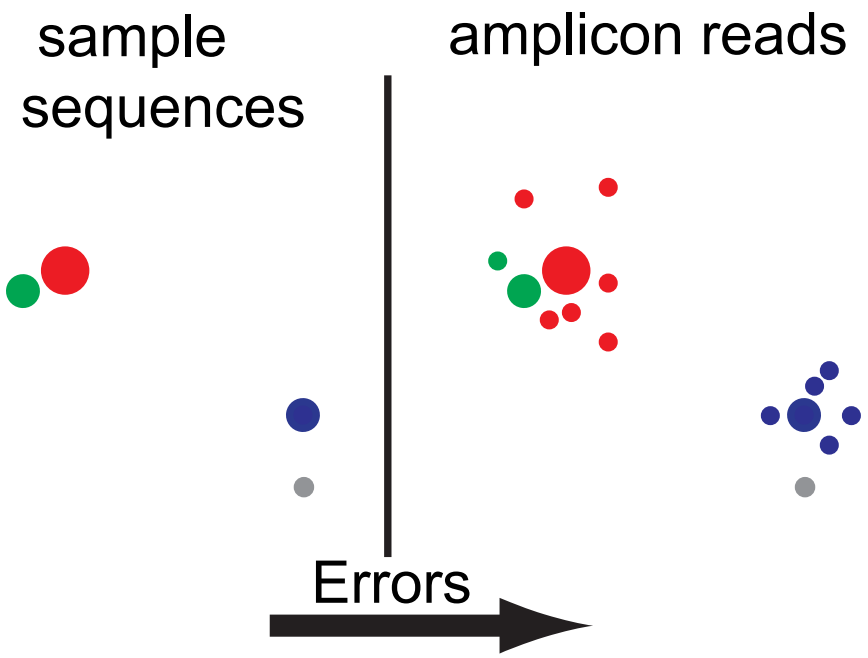
- Low resolution
  - 97% id or 0.03% distance used in the past for species approximation not necessarily the case
- Data dependent
  - Each data addition ---> re-run analysis
- Squencing error sensitive (particularly for singletons)
  - This particularly holds true for the phylogenetically more sound MSA based OTUs

Chen et al 2013 https://doi.org/10.1371/journal.pone.0070837
Wei et al 2021 https://doi.org/10.3389/fmicb.2021.644012
Lindahl et al 2017 https://doi.org/10.1111/nph.12243

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTU issues:

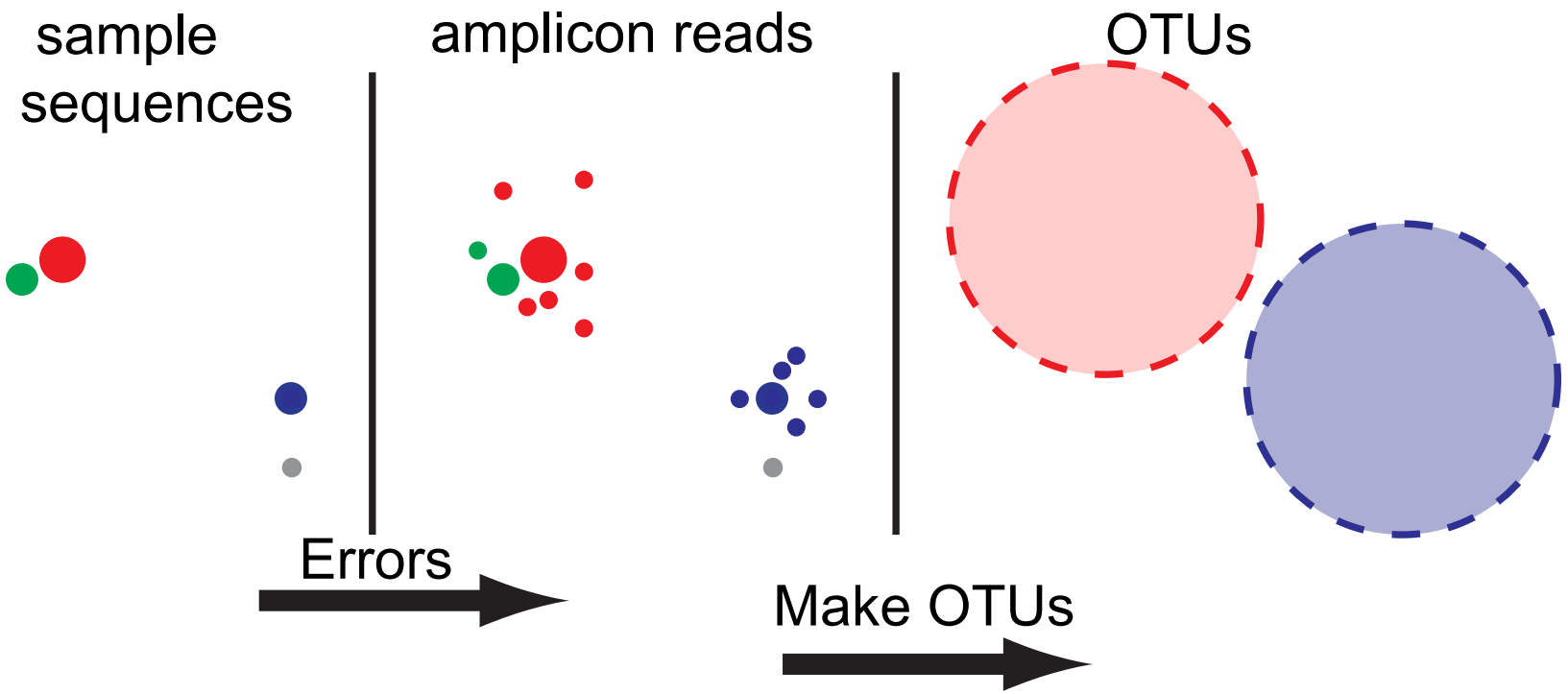


http://web.stanford.edu/class/bios221/Workshops/
https://benjjneb.github.io/dada2/SMBS_DADA2.pdf
Callahan et al 2016, Nat Methods, https://doi.org/10.1038/nmeth.3869
https://www.youtube.com/watch?v=wV5_z7rR6yw

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTU issues:



http://web.stanford.edu/class/bios221/Workshops/
https://benjjneb.github.io/dada2/SMBS_DADA2.pdf
Callahan et al 2016, Nat Methods, https://doi.org/10.1038/nmeth.3869
https://www.youtube.com/watch?v=wV5_z7rR6yw

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

ASVs can be OTUs at 0% distances of high quality sequences

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances acc
  - distances acc
- ASVs: unique se
  approximating
- phylotypes: AS
  classification in
- phylogenetic tr

benjamin.callahan Nov '16

> kmitchell:
>
> I'd bet its from the error correcting model dada2 uses. I haven't read their paper close enough to fully understand how it works. It's not going back to the truly raw data (the way that pyronoise error corrected), so they're error correcting off of the fastq?

You are right on both points. It's a bit unintuitive, since dada2 is resolving sequences exactly, but the removal of errors by dada2 almost always results in significantly fewer inferred sequence variants than the number of OTUs that are output by OTU pipelines like mothur/QIIME.

Error correction is based off the quality scores in the fastqs, combined with a statistical model of error abundances. And error-correction makes chimera removal an easier problem, which helps there as well.

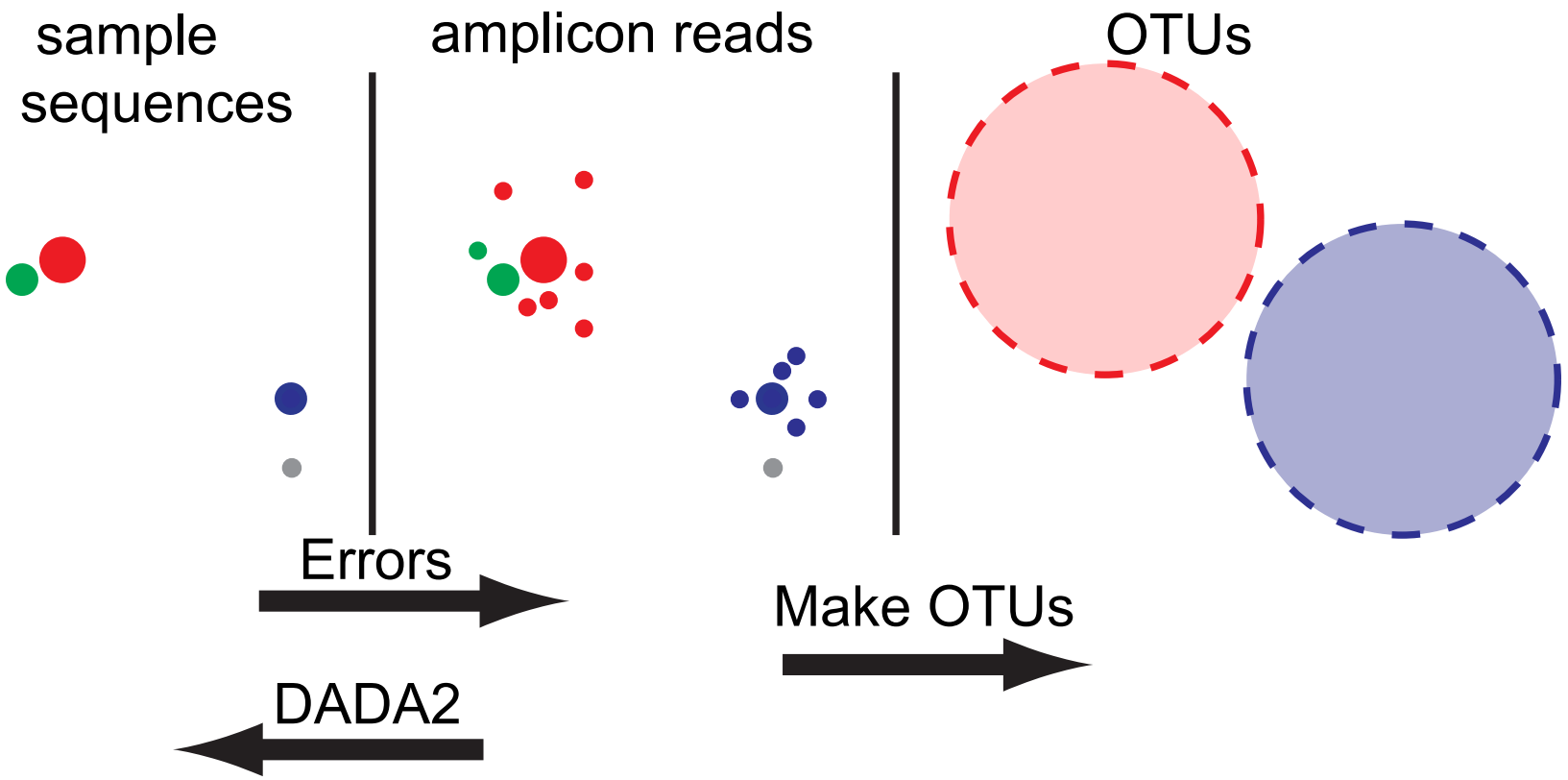
Dada2 ASVs

https://forum.mothur.org/t/otus-or-sequences/2975/6

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTU issues:



http://web.stanford.edu/class/bios221/Workshops/
https://benjjneb.github.io/dada2/SMBS_DADA2.pdf
Callahan et al 2016, Nat Methods, https://doi.org/10.1038/nmeth.3869
https://www.youtube.com/watch?v=wV5_z7rR6yw

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives of amplicon classification in taxonomic bins
- phylogenetic tree leaves

"ASVs are inferred by a *de novo* process in which biological sequences are discriminated from errors on the basis of, in part, the expectation **that biological sequences are more likely to be repeatedly observed** than are error-containing sequences. As a result, ASV inference cannot be performed independently on each read—the smallest unit of data from which ASVs can be inferred is a sample. However, unlike *de novo* OTUs, **ASVs are consistent labels** because ASVs **represent a biological reality** that exists outside of the data being analysed: **the DNA sequence of the assayed organism**. Thus, ASVs inferred independently from different studies or different samples can be validly compared."

Callahan et al 2017 https://doi.org/10.1038/ismej.2017.119

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Dada2 error correction during ASV generation:

```
s: ATTAACGAGATTATAACCAGAGTACGAATA...
     |                   |
r: ATCAACGAGATTATAACAAGAGTACGAATA...
```

$$p(r|s) = \prod_{i=1}^{L} p(r(i)|s(i), q_r(i), Z)$$

Error process is independent across nucleotides.

Per-nucleotide transition rate depends on:

- Sample nucleotide
- Read nucleotide
- Read quality at that position
- Batch effect (eg. run)

http://web.stanford.edu/class/bios221/Workshops/

Callahan et al 2016, Nat Methods, https://doi.org/10.1038/nmeth.3869

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

Genotypes?

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

ASV issues?

Schloss 2021 https://doi.org/10.1128/mSphere.00191-21
Callahan et al 2017 https://doi.org/10.1038/ismej.2017.119

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

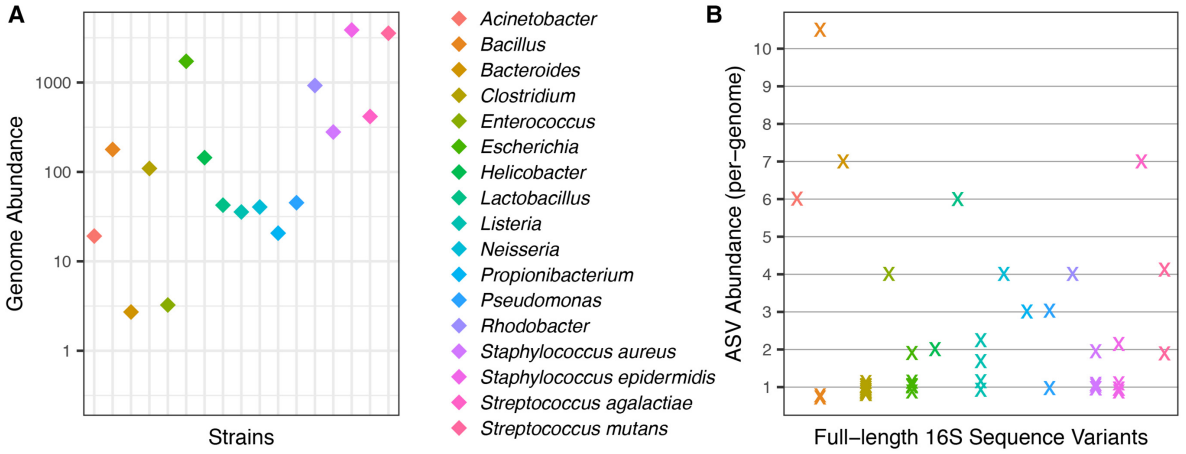


Figure 3. Abundances of genomes and ASVs recovered from the HMP mock community. (**A**) The abundances of each genome in the long-read amplicon sequencing data are plotted on the log-scaled y-axis (Methods). Observed genome abundances varied over three orders of magnitude. Significant counting noise exists for ASVs from genomes with abundances below 100. (**B**) The abundance of each ASV divided by the genomic abundance of the mock community strain from which it originated is plotted on the y-axis. Integral values are indicated by horizontal grid lines. No other ASVs were detected.

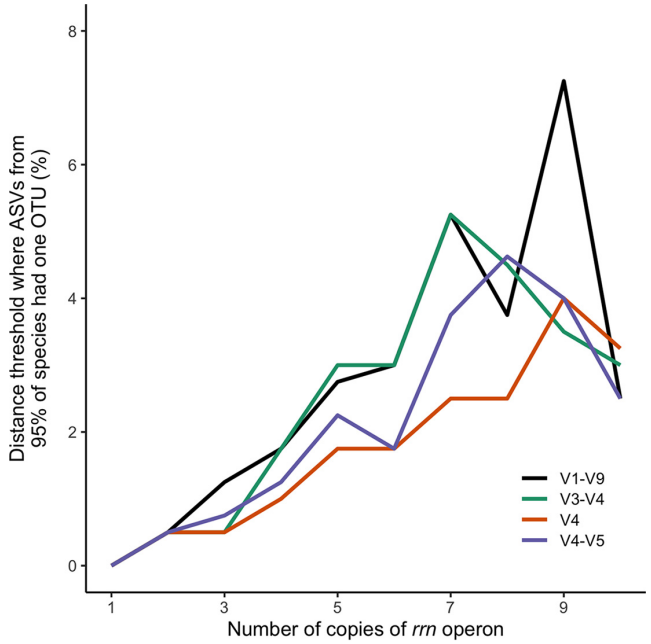


FIG 1 The distance threshold required to prevent the splitting of genomes into multiple OTUs increased as the number of *rrn* operons in the genome increased. Each line represents the median distance threshold for each region of the 16S rRNA gene that is required for 95% of the genomes with the indicated number of *rrn* operons to cluster their ASVs to a single OTU. The median distance threshold was calculated across 100 randomizations in which one genome was sampled from each species. Only those numbers of *rrn* operons that were found in more than 100 species are included.

ASVs

- With the increase of *rrn* operons (e.g. copiotrophic behaviours/environments), the necessary distance for a species-specific OTU increases…
- Hence, ASVs not easy to be used for inferring taxonomy levels like strain/species.
- Frequently, they represent allelic copy variants

Callahan et al 2019, Nucleic Acid Research, https://doi.org/10.1093/nar/gkz569

Schloss 2021, *mSphere*, https://doi.org/10.1128/msphere.00191-21
Stevenson and Schmit 1998, J Bacteriol, https://doi.org/10.1128/JB.180.7.1970-1972.1998
Klappenbach et al 2000, Appl Environ Microbiol, https://doi.org/10.1128/AEM.66.4.1328-1333.2000
Roller et al 2016, Nat Microbiol, https://doi.org/10.1038/nmicrobiol.2016.160

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

- Pairwise comparisons (e.g. BLAST or alike)

```
A.terrissima   2    TGAAGGCGGGCTGGACCTCTCGGGGTTACAGCCTTGCTGAATTATTCACCCTTGTCTTTT  61
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct 3             TGAAGGCGGGCTGGACCTCTCGGGGTTACAGCCTTGCTGAATTATTCACCCTTGTCTTTT 62

A.terrissima  62    GCGTACTTCTTGTTTCCTTGGTGGGTTCGCCCACCACTAGGACAAACATAAACCTTTTGT 121
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 Sbjct 63           GCGTACTTCTTGTTTCCTTGGTGGGTTCGCCCACCACTAGGACAAACATAAACCTTTTGT 122

A.terrissima  122   AATTGCAATCAGCGTCAGTAACAAATTAATAATTACAACTTTCAACAACGGATCTCTTGG  181
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct 123           AATTGCAATCAGCGTCAGTAACAAATTAATAATTACAACTTTCAACAACGGATCTCTTGG 182

A.terrissima  182   TTCTGGCATCGATGAAGAACGCAGC  206
                    |||||||||||||||||||||||||
Sbjct 183           TTCTGGCATCGATGAAGAACGCAGC 207
```

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

- Pairwise comparisons (e.g. BLAST or alike)
- Naïve Bayesian classifier:

For *seq* = $Kmer_1$, $Kmer_2$,$Kmer_3$, $Kmer_4$ … and $taxon_x$ for *x* in *1, 2, 3… n*

Wang, Q., et al. (2007). Appl Environ Microbiol 73, 5261-5267, doi: AEM.00062-07 [pii] 10.1128/AEM.00062-07

Also e.g. https://www.youtube.com/watch?app=desktop&v=lFJbZ6LVxN8

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

- Pairwise comparisons (e.g. BLAST or alike)
- Naïve Bayesian classifier:

For *seq* = $Kmer_1$, $Kmer_2$,$Kmer_3$, $Kmer_4$ … and $taxon_x$ for *x* in *1, 2, 3… n*

Posterior Likelihood Prior

$$P(taxon_x|seq) = \frac{P(seq|taxon_x)P(taxon_x)}{P(seq)}$$

Evidence ($P(seq) = P(seq|taxon_x) + P(seq|!\,taxon_x)$)

, with $Kmer_1$, $Kmer_2$,$Kmer_3$, $Kmer_4$ … in seq being independent (Naïve)

Wang, Q., et al. (2007). Appl Environ Microbiol 73, 5261-5267, doi: AEM.00062-07 [pii] 10.1128/AEM.00062-07
Also e.g. https://www.youtube.com/watch?app=desktop&v=lFJbZ6LVxN8

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

- Pairwise comparisons (e.g. BLAST or alike)
- Naïve Bayesian classifier:

For *seq* = $Kmer_1$, $Kmer_2$,$Kmer_3$, $Kmer_4$ … and $taxon_x$ for *x* in *1, 2, 3… n*

Posterior Likelihood Prior

$$P(taxon_x|seq) = \frac{P(seq|taxon_x)P(taxon_x)}{P(seq)}$$

Evidence ($P(seq) = P(seq|taxon_x) + P(seq|!\,taxon_x)$)

, with $Kmer_1$, $Kmer_2$,$Kmer_3$, $Kmer_4$ … in seq being independent (Naïve)

- Evolutionary placement (e.g. EPA-ng, pplacer)

Wang, Q., et al. (2007). Appl Environ Microbiol 73, 5261-5267, doi: AEM.00062-07 [pii] 10.1128/AEM.00062-07

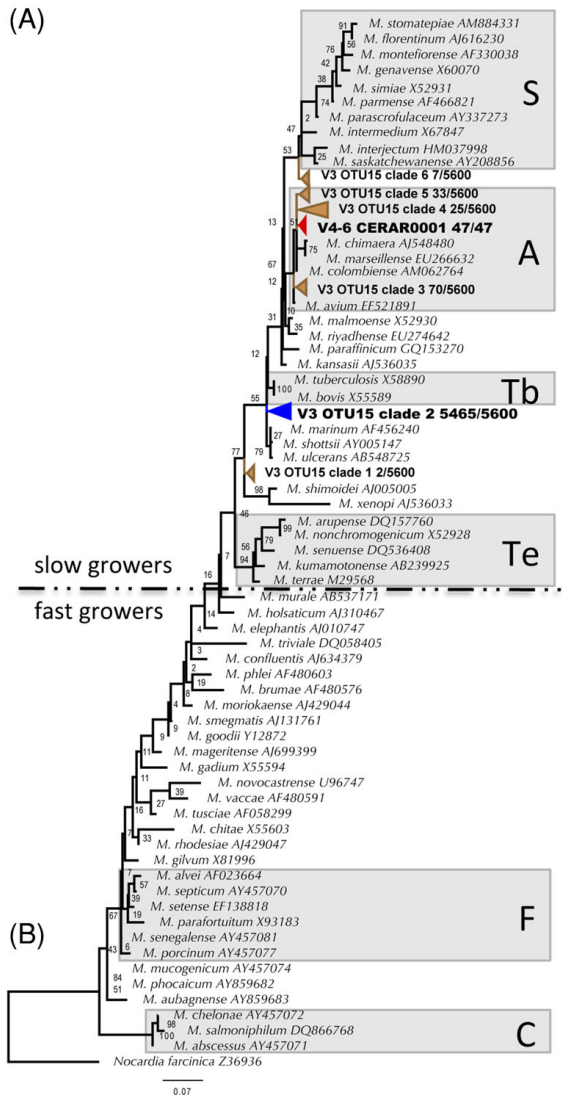
Also e.g. https://www.youtube.com/watch?app=desktop&v=lFJbZ6LVxN8

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

Common classification methods

- Evolutionary placement (e.g. EPA-ng, pplacer)



Barbera, P., et al. (2018). Syst Biol 68, 365-369, doi: 10.1093/sysbio/syy054
Matsen, F., et al. (2010). BMC Bioinformatics 11, 538, doi: 10.1186/1471-2105-11-538
Vasileiadis, S., et al. (2015). FEMS Microbiol Ecol 91, fiv114, doi: 10.1093/femsec/fiv114

DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

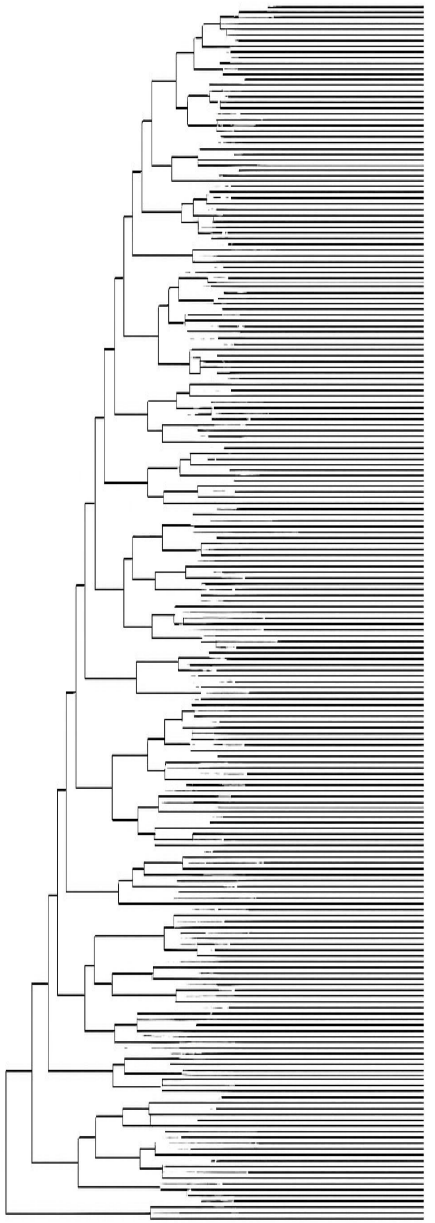
4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- OTUs: clusters of sequences based on:
  - distances according to multiple sequence alignments
  - distances according to pairwise sequence alignments
- ASVs: unique sequences equal to amplicon variants approximating gene variants
- phylotypes: ASV or OTU representatives or amplicon classification in taxonomic bins
- phylogenetic tree leaves



DEPARTMENT OF Biochemistry & Biotechnology UNIVERSITY OF THESSALY

4a. Read quality assessment, quality control, OTUs/ASVs/phylotypes/taxa

- Read stats and sequencing quality assessment
- Read quality control
- Read-pair assembly
- Formation of operational taxonomic units (OTUs) or amplicon sequence variants (ASVs)
- Removal of chimeras and off-target sequences

sample	raw	post trimming	post assembly	Analyzed: post chimera and specificity check	Good's coverage
D1	52211	49321	48150	41947	1.000
D2	48090	45522	44260	39244	1.000
D3	30969	29485	28679	26163	0.999
M01	54529	51853	50924	45471	1.000
M02	32242	30720	30116	27542	1.000
M03	29110	27636	27096	24177	1.000
1MO1	28234	26861	26331	23135	1.000
1MO2	66358	62527	61334	53552	1.000
1MO3	43473	41473	40668	36576	1.000
2MO2	20418	19466	19063	17891	1.000
2MO3	27732	26307	25777	23990	1.000
2MO7	16075	15372	15065	14103	0.999
2MO8	31558	29946	29423	27614	1.000
3MO1	33863	32143	31605	29606	1.000
3MO2	4063	3860	3789	3565	0.991
3MO3	18565	17659	17320	15723	0.999
3MO7	37993	36034	35466	32919	1.000
3MO8	45630	43437	42745	39740	1.000
4MO1	97897	93063	91498	82601	1.000
4MO2	84060	80106	78854	73560	1.000
4MO3	52472	49859	48920	46218	1.000
total	946358	898292	881164	803585	
remaining % of total		94.92%	93.11%	84.91%	

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis

- **Ecological indicators**
  - ❖α-diversity indices (within sample diversity)
  - ❖β-diversity (between sample diversity)

- **Common questions (not exhaustive list)**
  - ❖α-diversity indices
    - ❑ α-diversity differences between treatments/environments
    - ❑ α-diversity relations with (a-)biotic parameters
  - ❖β-diversity
    - ❑ Compositional differences between samples
    - ❑ Compositional differences between treatments/environments/sample-groups
    - ❑ “Core” community OTUs/ASVs/phylotypes/taxa
    - ❑ Unique OTUs/ASVs/phylotypes/taxa for treatments/environments
    - ❑ Co-occurrence/correlation between OTUs/ASVs/phylotypes/taxa and (a-)biotic parameters or other OTUs/ASVs/phylotypes/taxa

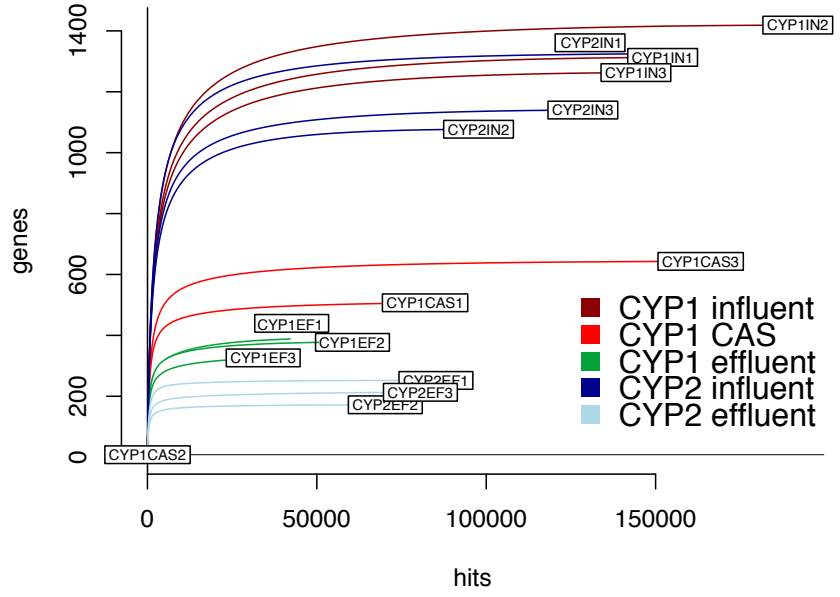
DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

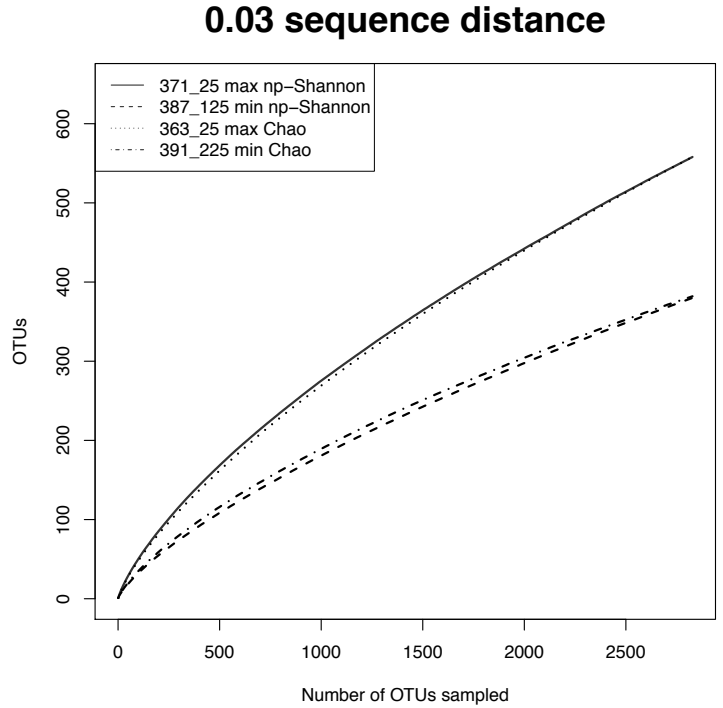
4b. Diversity analysis (α-diversity)

- **α-diversity indices (coverage assessment)**
  - ❖ Good's Coverage variants: based on rare event counts

$$C = 1 - F1/N$$, F1 = singletons, N = total events

  - ❖ Rarefaction curves: aka subsampling with replacement





Good 1953, Biometrika, https://doi.org/10.1093/biomet/40.3-4.237

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (α-diversity)

- **α-diversity indices (in terms of relative abundance p; non exhaustive)**
  - ❖ Richness: number of different OTUs/ASVs/phylotypes/taxa "X" or

$$\sum_{i=1}^{S} p_i^0$$

  - ❖ Shannon: the entropy of a system in terms of OTUs/ASVs/phylotype/taxon

$$-\sum_{i=1}^{S} p_i \ln p_i$$

  - ❖ Simpson variants: the probability of resampling with replacement the same OTUs/ASVs/phylotype/taxon

$$1/\sum_{i=1}^{S} p_i^2$$

  - ❖ Shannon equitability/evenness: Shannon / max-Shannon (or richness)

$$E_q = \frac{S_q}{S_q^{max}}$$

Joost 2006 http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x
Mendes et al 2008 http://dx.doi.org/10.1111/j.0906-7590.2008.05469.x

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (α-diversity)

- **α-diversity indices (in terms of relative abundance p; non exhaustive)**
  - ❖ Richness: number of different OTUs/ASVs/phylotypes/taxa "X" or

    $$\sum_{i=1}^{S} p_i^0$$

    q = 0: 0 order (all members are represented)
  - ❖ Shannon: the entropy of a system in terms of OTUs/ASVs/phylotype/taxon

    $$\sum_{i=1}^{S} p_i \ln p_i$$

    q = 1: 1st order (most members are represented)
  - ❖ Simpson variants: the probability of resampling with replacement the same OTUs/ASVs/phylotype/taxon

    $$1/\sum_{i=1}^{S} p_i^2$$

    q = 2: 2nd order (more dominant members are represented)
  - ❖ Shannon equitability/evenness: Shannon / max-Shannon (or richness)

    $$E_q = \frac{S_q}{S_q^{max}}$$

$$^{q}D \equiv \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}$$

Joost 2006 http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x
Mendes et al 2008 http://dx.doi.org/10.1111/j.0906-7590.2008.05469.x

4b. Diversity analysis (α-diversity)

- **α-diversity indices: common strategies in statistics**
  - ❖Student's *t* tests
  - ❖Analysis of variance (ANOVA) with *post hoc* Student's *t and* alike tests

If conditions are not met (e.g. normal data distribution, homogeneity of variance; independence of trials is assumed to exist) examine normalization possibility or performance of non-parametric equivalents

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics… focus?**
  - ❖ Explore the overall dataset
    - ❑ Just don't be selective
  - ❖ Abundance based feature selection (discard all rare events)… hmm
    - ❑ …
  - ❖ Hypothesis testing selection of responsive features to focus on
    - ❑ pime

Roesch et al 2020, Molecular Ecology Resources https://doi.org/10.1111/1755-0998.13116

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

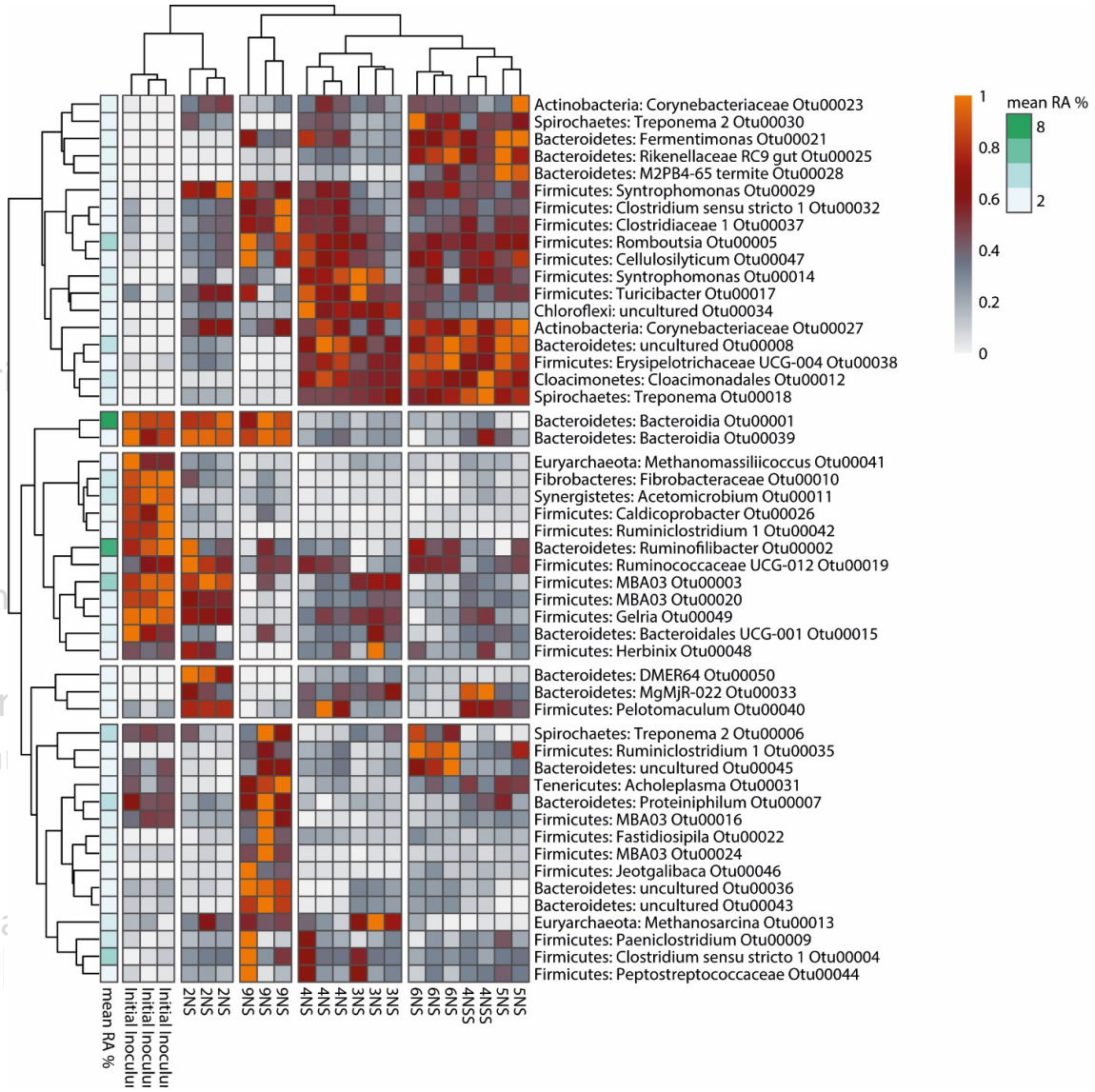
- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learning (e.g. random forests)
  - ❖ Differential abundance tests (normalization usually required)
    - ❑ Non-parametric, Fisher's exact test, generalized linear models and Student's *t* test variants… normalization?
  - ❖ Correlations of OTUs/ASVs/phylotypes/taxa with (a-)biotic parameters and OTUs/ASVs/phylotypes/taxa
    - ❑ Co-occurrence/correlations with suitable algorithms accounting for multiple zeros, combined with network proximity algorithms or not.

https://mb3is.megx.net/gustame
Weiss et al 2017 https://doi.org/10.1186/s40168-017-0237-y
Roesch 2019 https://doi.org/10.1111/1755-0998.13116

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivar
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learnin
  - ❖ Differential abundance tests (normalization usually r
    - ❑ Non-parametric, Fisher's exact test, generalized linea
  - ❖ Correlations of OTUs/ASVs/phylotypes/taxa with
    - ❑ Co-occurrence/correlations with suitable algorithms a algorithms or not.

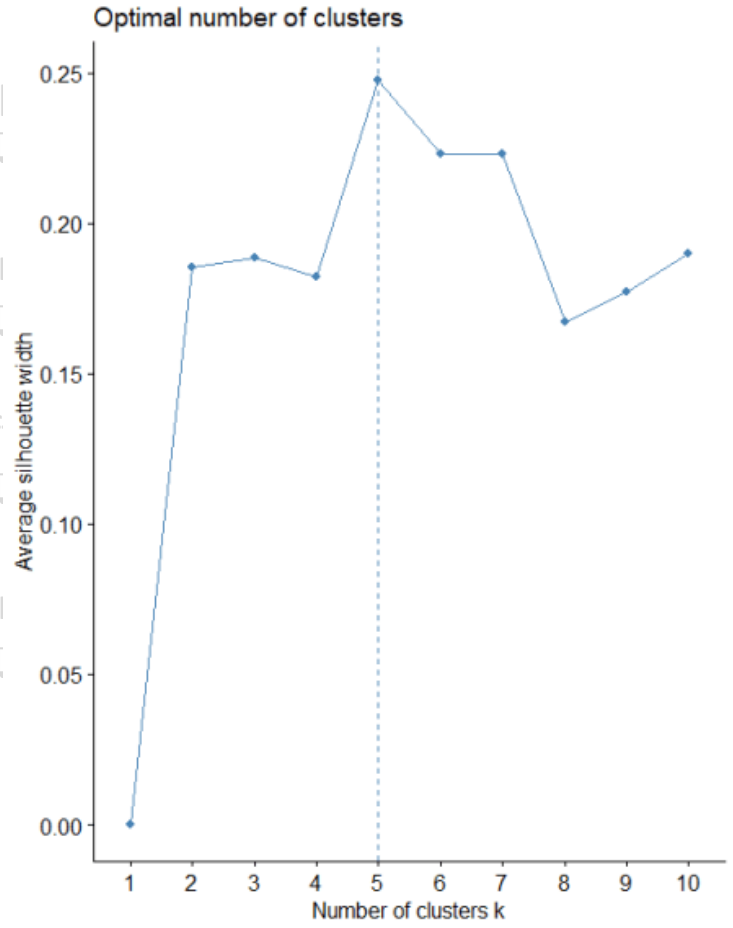


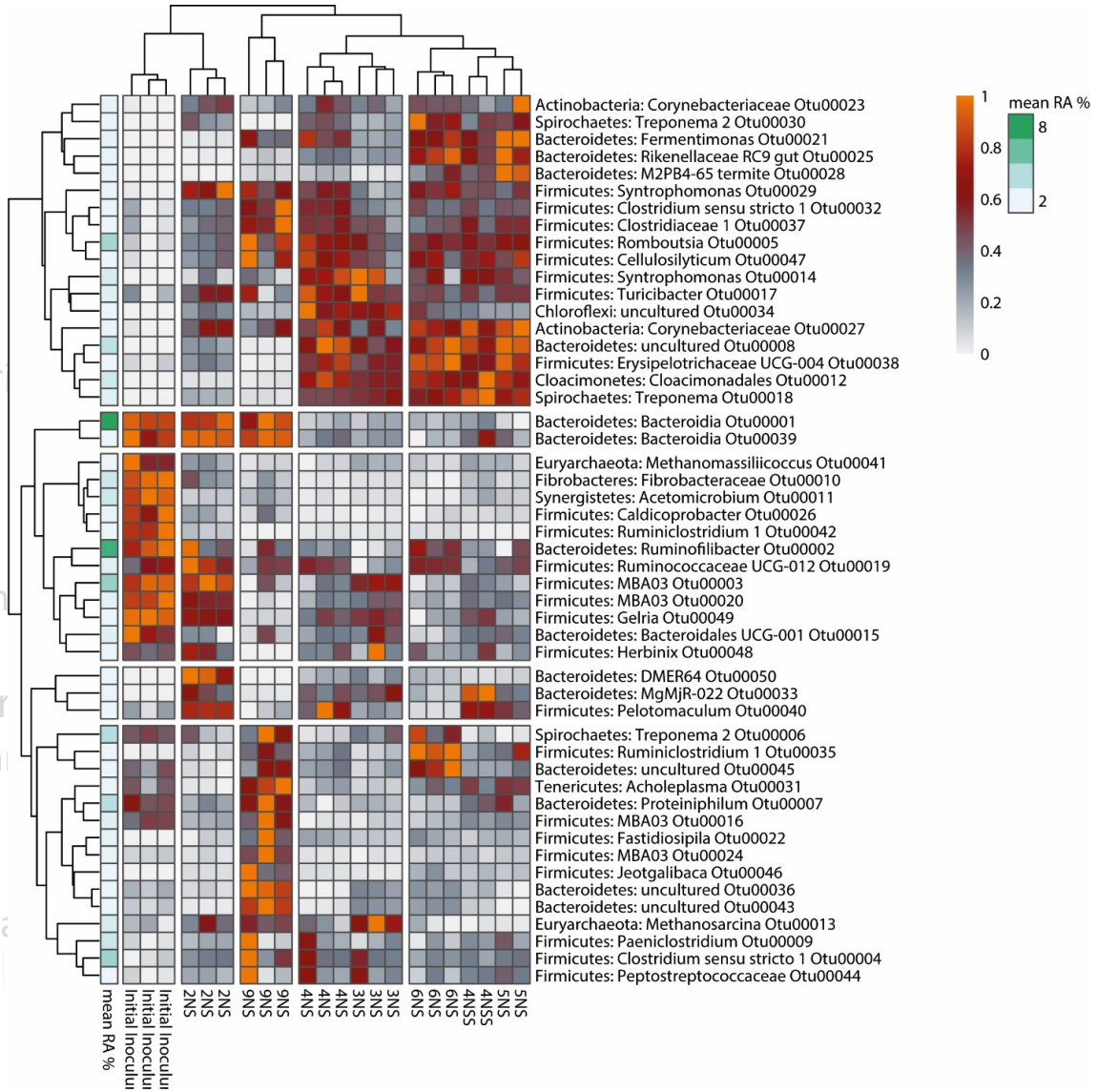
Kalamaras, S.D., et al. (2020). doi: 10.1002/jctb.6385

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS



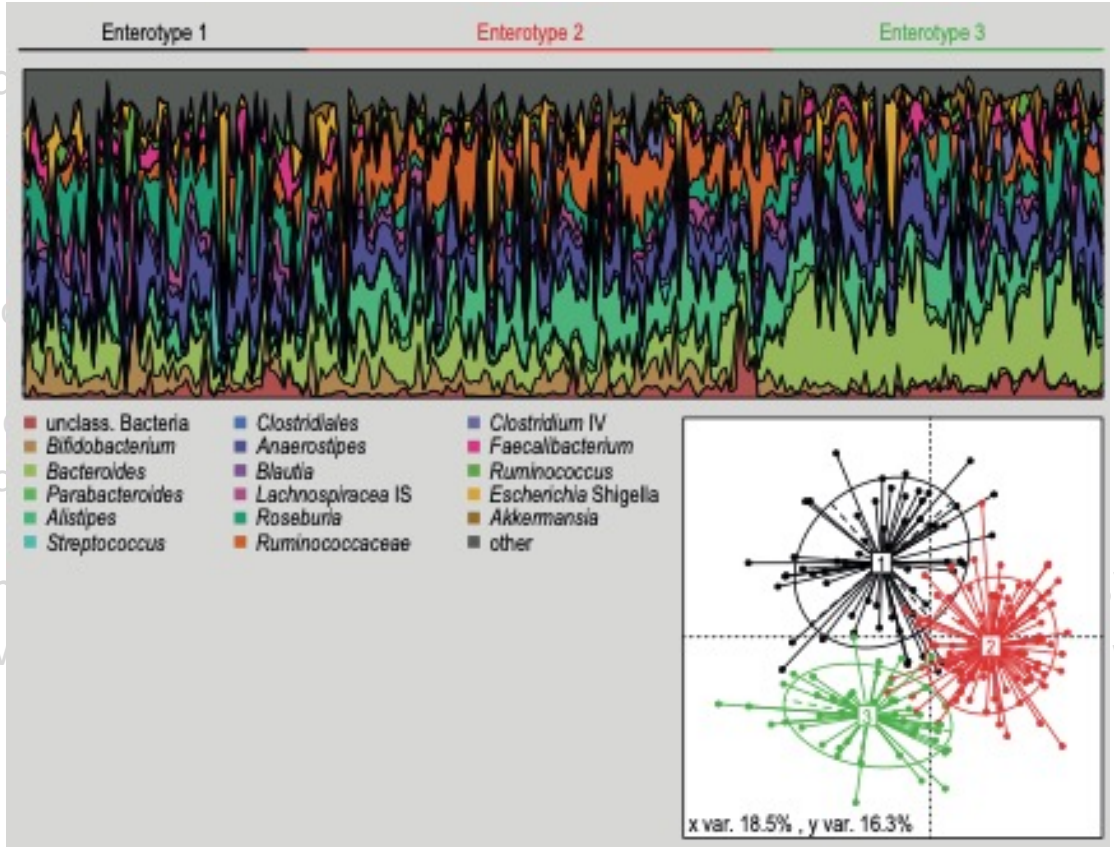


Kalamaras, S.D., et al. (2020). doi: 10.1002/jctb.6385

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence
  - ❖ Differential abundance tests (n
    - ❑ Non-parametric, Fisher's exac
  - ❖ Correlations of OTUs/ASVs/ph Vs/phylotypes/taxa
    - ❑ Co-occurrence/correlations w with network proximity algorithms or not.



Soldi, S., et al. (2019). Beneficial Microbes 0, 1-11, doi: 10.3920/bm2018.0116

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learning (e.g. random forests)
  - ❖ Differential abundance tests (normalization usually required)
    - ❑ Non-parametric, Fisher's exact test, generalized linear models and Student's *t* test variants
  - ❖ Correlations of OTUs/ASVs/phylotypes/taxa with (a-)biotic parameters and OTUs/ASVs/phylotypes/taxa
    - ❑ Co-occurrence/correlations with suitable algorithms accounting for multiple zeros, combined with network proximity algorithms or not.

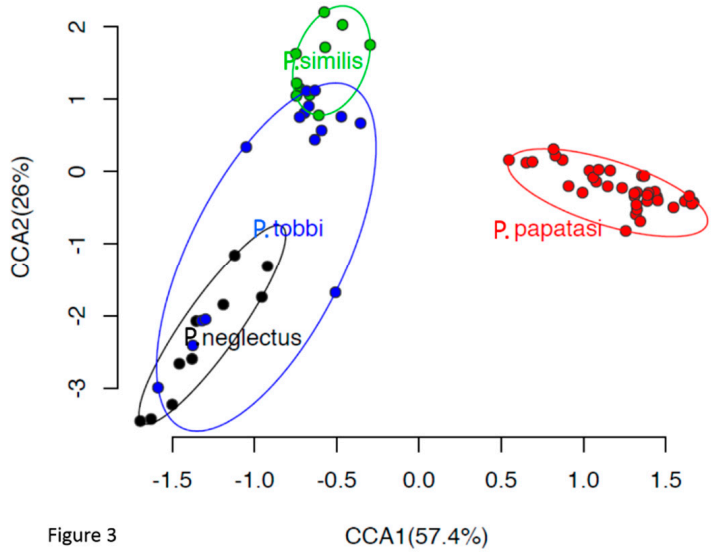


Figure 3

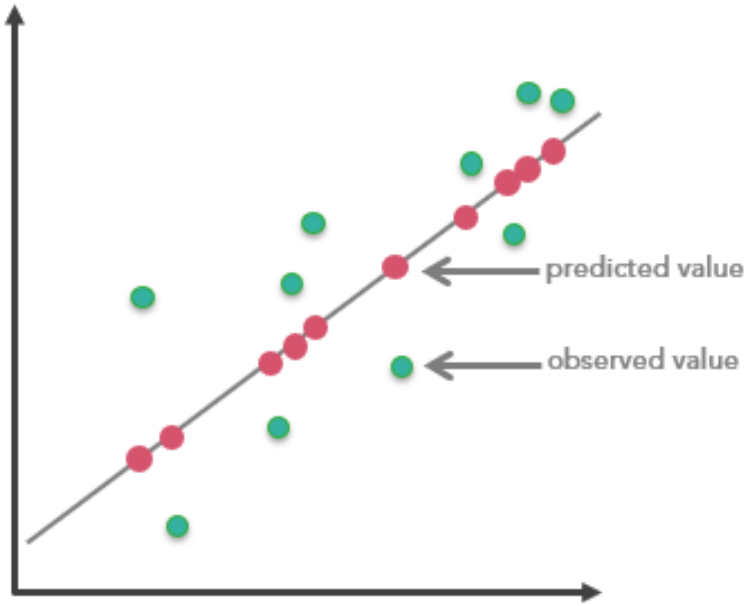
Figure 3. Canonical correspondence analysis (CCA) of the bacterial community in the gut of *P. papatasi, P. tobbi, P. similis* and *P. neglectus*. Ellipses encompass all samples of the same *Phlebotomus* species. The model testing the host genotype effect on the bacterial community structure was significant ($p < 0.001$) and explained 45.1% of the total variance.

Papadopoulos, et al. (2020). Pathogens doi: 10.3390/pathogens9060428

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_n X_n + \varepsilon$
  - ❖ ine learning (e.g. random forests)
  - ❖ n usually required)
  - ❖ alized linear models and Student's *t* test variants
  - ❖ taxa with (a-)biotic parameters and OTUs/ASVs/phylotypes/taxa
  - ❖ lgorithms accounting for multiple zeros, combined with network proximity



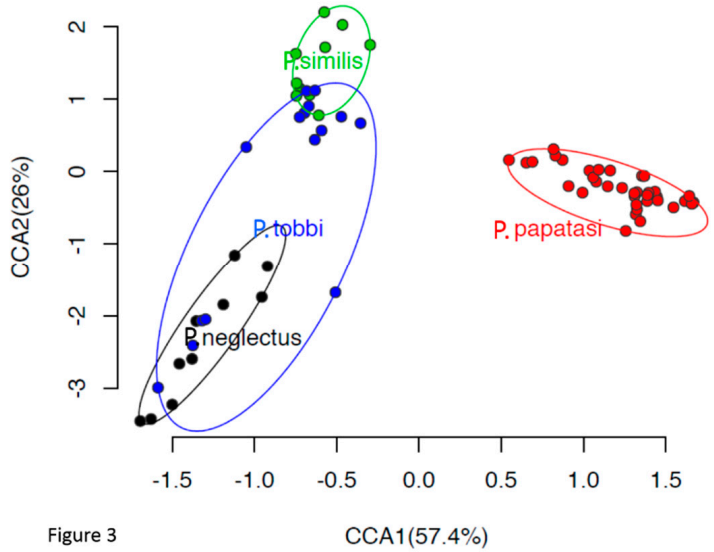


Figure 3

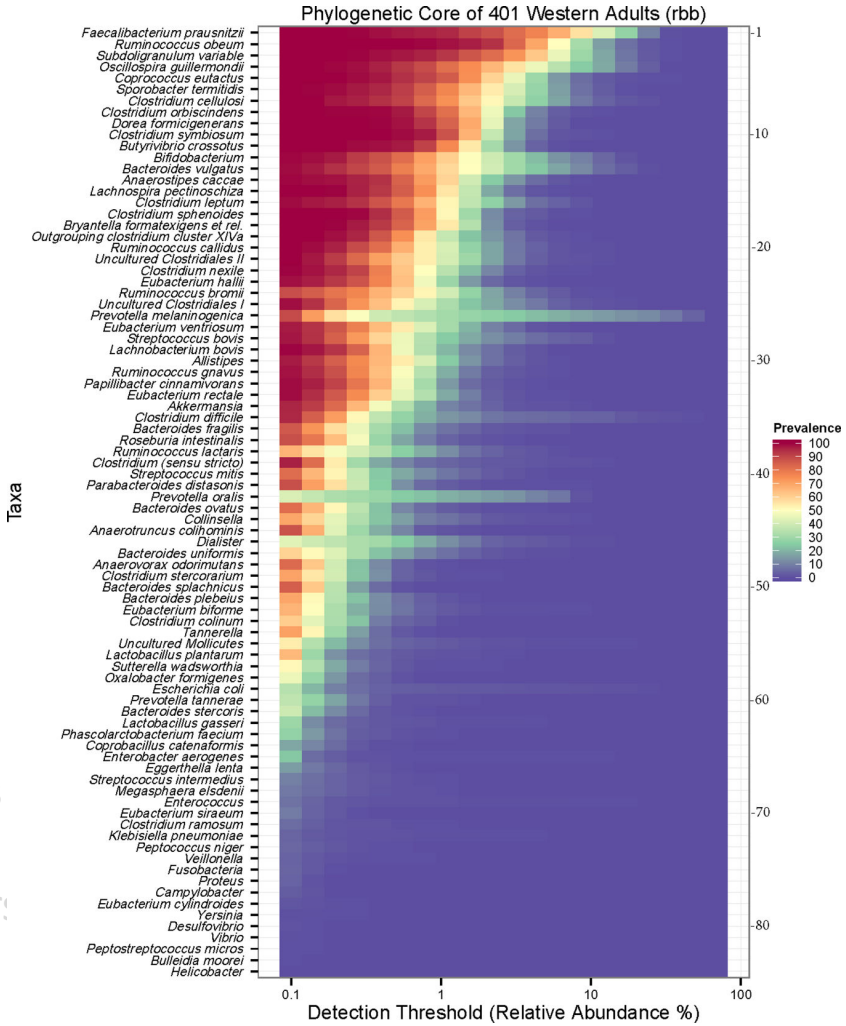
Figure 3. Canonical correspondence analysis (CCA) of the bacterial community in the gut of *P. papatasi, P. tobbi, P. similis* and *P. neglectus*. Ellipses encompass all samples of the same *Phlebotomus* species. The model testing the host genotype effect on the bacterial community structure was significant ($p < 0.001$) and explained 45.1% of the total variance.

Papadopoulos, et al. (2020). Pathogens doi: 10.3390/pathogens9060428

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learning (e.g. random forests)
  - ❖ Differential abundance tests (normalization usually required)
    - ❑ Non-parametric, Fisher's exact test, generalized linear models and Student's
  - ❖ Correlations of OTUs/ASVs/phylotypes/taxa with (a-)biotic parameters
    - ❑ Co-occurrence/correlations with suitable algorithms accounting for multiple algorithms or not.

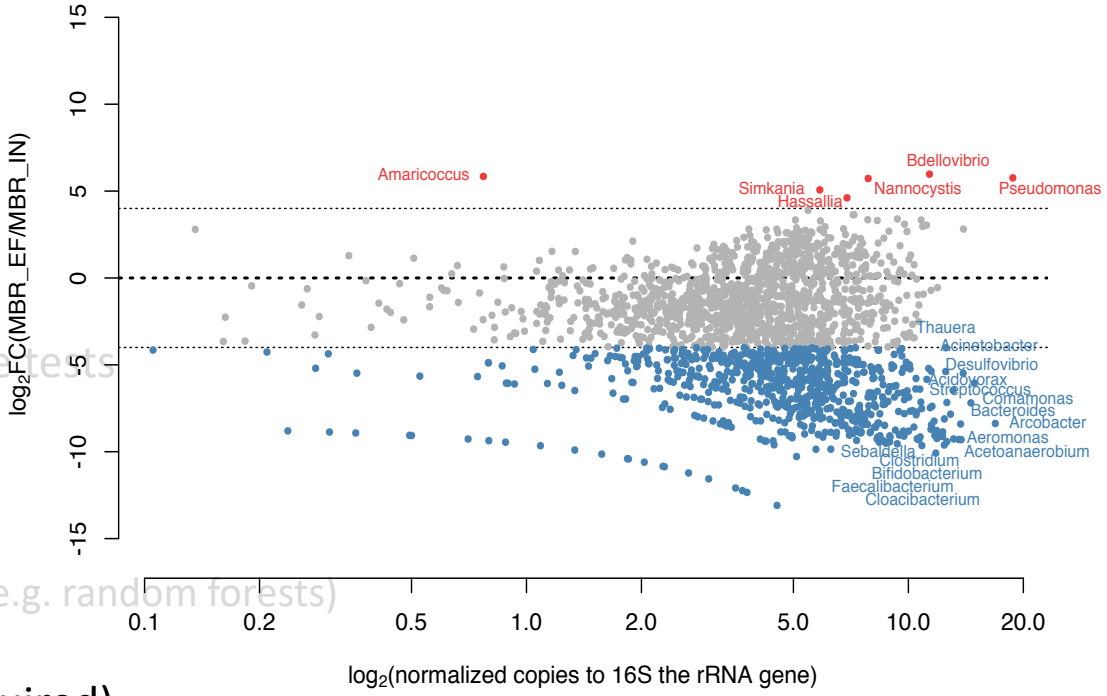


Sudarshan et al 2017; doi: 10.1093/femsre/fuw045

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learning (e.g. random forests)
  - ❖ Differential abundance tests (normalization usually required)
    - ❑ Non-parametric, Fisher's exact test, generalized linear models and Student's *t* test variants
  - ❖ Correlations of OTUs/ASVs/phylotypes/taxa with (a-)biotic parameters and OTUs/ASVs/phylotypes/taxa
    - ❑ Co-occurrence/correlations with suitable algorithms accounting for multiple zeros, combined with network proximity algorithms or not.

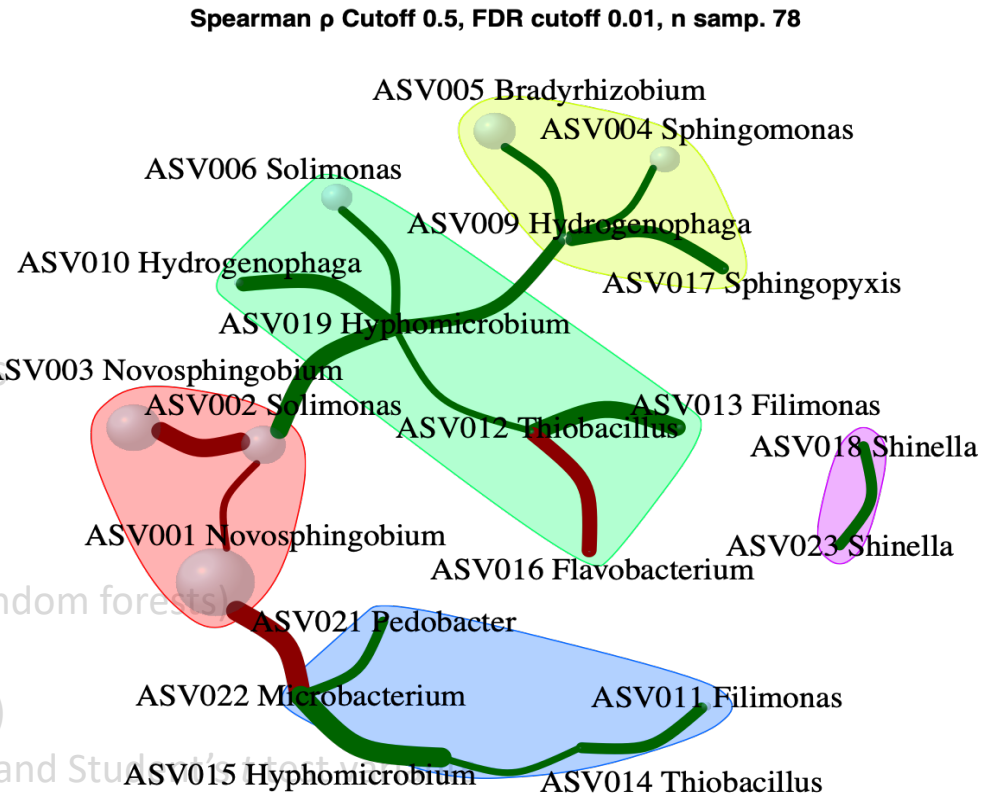


Sudarshan et al 2017; doi: 10.1093/femsre/fuw045

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

- **β-diversity: common strategies in statistics**
  - ❖ Descriptive multivariate analyses
    - ❑ cluster analysis, PCA, CA, PCoA, NMDS
  - ❖ Hypothesis testing (constrained, canonical) multivariate tests
    - ❑ RDA, CCA, PERMANOVA
  - ❖ Core microbial communities
    - ❑ Participation and prevalence cutoffs, machine learning (e.g. random forests)
  - ❖ Differential abundance tests (normalization usually required)
    - ❑ Non-parametric, Fisher's exact test, generalized linear models and Student's t-test



❖ Correlations of OTUs/ASVs/phylotypes/taxa with (a-)biotic parameters and OTUs/ASVs/phylotypes/taxa

❑ Co-occurrence/correlations with suitable algorithms accounting for multiple zeros, combined with network proximity algorithms or not.

Hirano and Takemoto 2019 https://doi.org/10.1186/s12859-019-2915-1
https://github.com/SotiriosVasileiadis/mconsort_tbz_degr

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

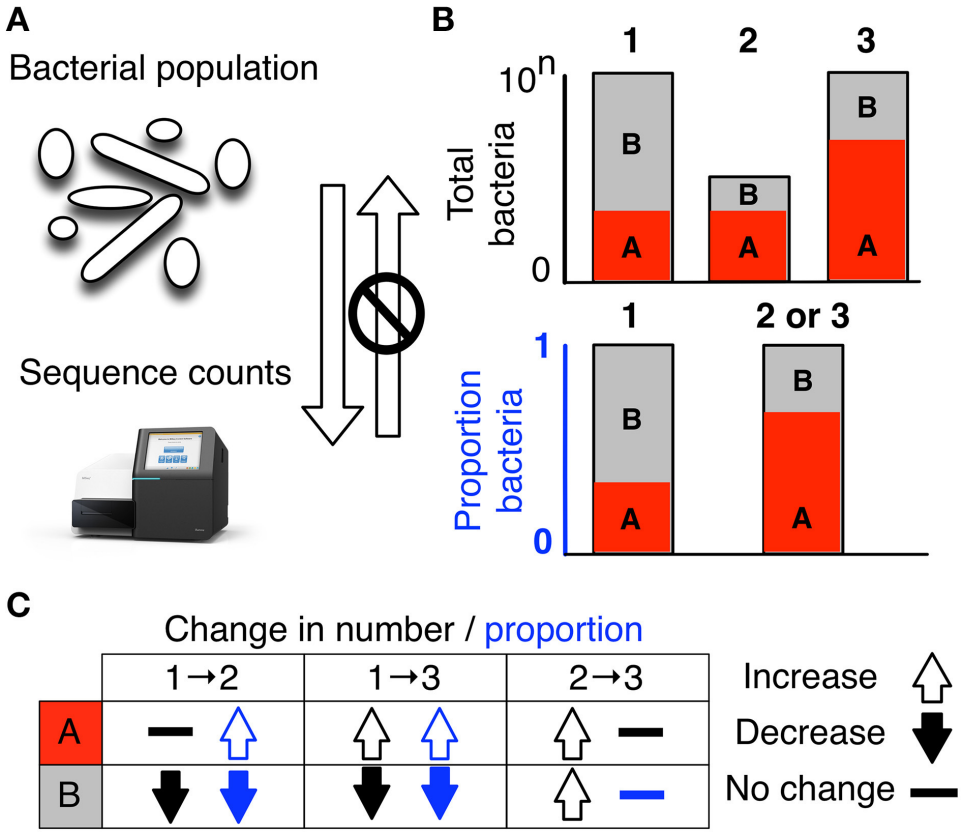
Major issue to consider!!!

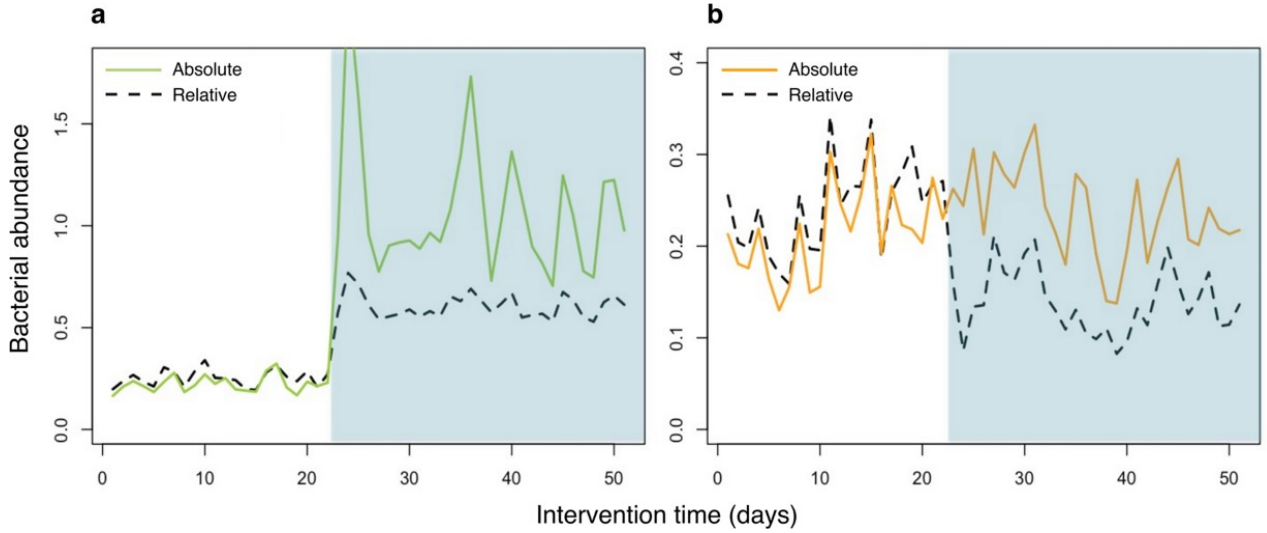
Data compositionality

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

Data compositionality examples





Gloor et al 2017, Frontier Microbiol, https://doi.org/10.3389/fmicb.2017.02224

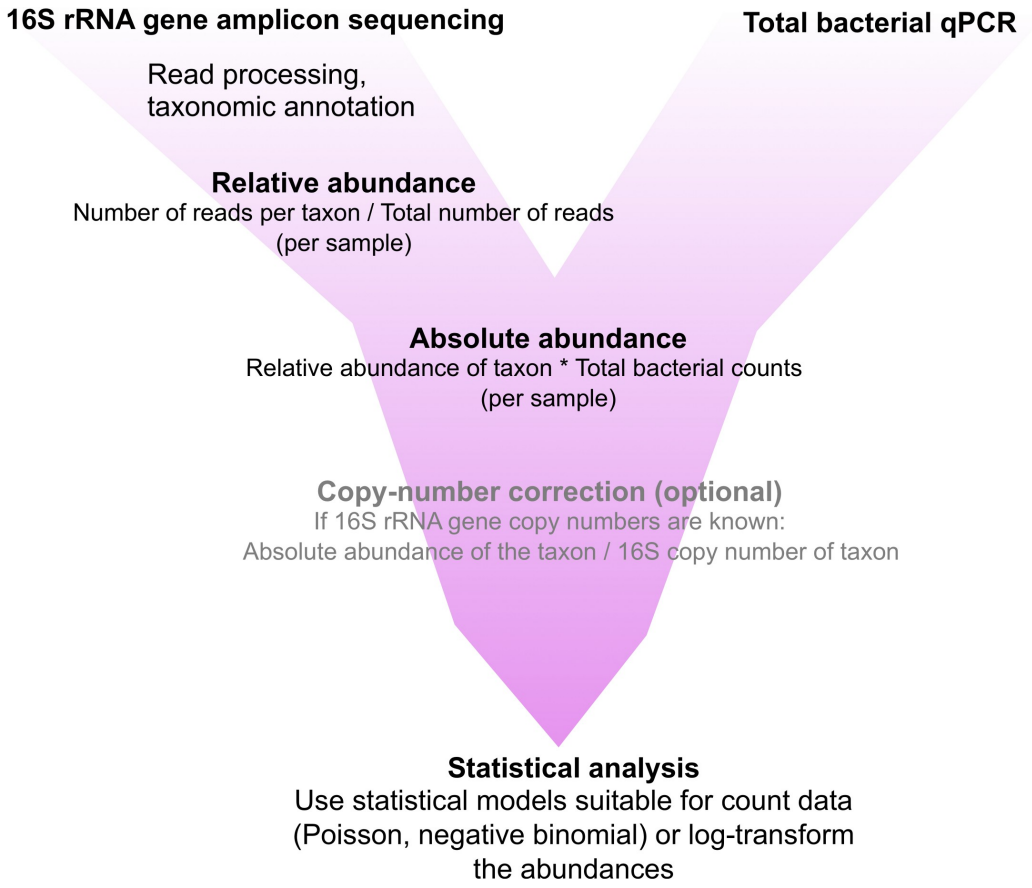
Jian et al 2020, PLOS ONE, https://doi.org/10.1371/journal.pone.0227285

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

Dealing with compositionality

Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM



Gloor et al 2017, Frontier Microbiol, https://doi.org/10.3389/fmicb.2017.02224

Jian et al 2020, PLOS ONE, https://doi.org/10.1371/journal.pone.0227285

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

4b. Diversity analysis (β-diversity)

Consider detection limits and marker rlevance

E.g. if I need to analyse archaeal ammonia oxidizers

- knowing that they paricipate in ~1/10,000 16S rRNA gene ASVs
- it is preferable to choose the *amoA* marker instead of the 16S rRNA gene

Also if I need to identify correlations between different markers

- I need to have basic understanding of thei relative frequencies in the microbial communities with respect to detection limits (e.g. comparison of ARG counts with 16S rRNA gene diversity)

Summary – Key points

- Consider carefully the setup (incl. replicates)!
- Select the appropriate marker gene and primers
- Develop the wet lab protocols to facilitate specificity/sensitivity (annealing temps, cycles etc.)
- Enhance target sequence analysis in masking/complex environments (e.g. host systems)
- Choose sequencing method/depth/read-size (low budget is not necssarily restrictive)

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

Summary – Key points

- Choose the type of your microbial reference unit (ASVs, 97% id OTUs?)
- Opt for the best taxonomy annotation method
- Set the questions before the analysis
- Consider if the analysis should reflect compositionality or abundance
- Consider the methodological limits of detection

DEPARTMENT OF
Biochemistry & Biotechnology
UNIVERSITY OF THESSALY

Some useful links

- Free statistical software:
  - Programming knowledge and/or good will: R/RStudio (https://www.r-project.org, https://rstudio.com) and proposed tutorial (https://www.guru99.com/r-tutorial.html)
  - *No programming knowledge*: PAST (https://folk.uio.no/ohammer/past/)
- Mothur tutorial for Illumina data (MiSeq-SOP): www.mothur.org and https://www.mothur.org/wiki/MiSeq_SOP
- Dada2 tutorial for 16S rRNA gene and ITS Illumina data in R: https://benjjneb.github.io/dada2/tutorial_1_8.html, https://benjjneb.github.io/dada2/ITS_workflow.html
- Amplicon sequencing analysis Dada2 workshop for long-read data data in R: http://web.stanford.edu/class/bios221/Workshops/
- Multivariate statistics with GUSTA-ME of Alban Ramette (https://mb3is.megx.net/gustame)

Schloss, P.D., et al. (2009). Appl Environ Microbiol 75 doi: 10.1128/aem.01541-09
Callahan et al (2016). F1000Research 5 doi: 10.12688/f1000research.8986.2
Buttigieg, P.L. and Ramette, A. (2014). FEMS Microbiol Ecol 90, 543-550, doi: 10.1111/1574-6941.12437

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ & ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

DEPARTMENT BIOCHEMISTRY & BIOTECHNOLOGY
UNIVERSITY OF THESSALY

25 YEARS

Plant & Environmental Biotechnology Laboratory

Key & frequently overlooked points in amplicon sequencing analysis of environmental microbiomes

Thank you for your attention!!!

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΕΤΟΣ ΙΔΡΥΣΗΣ 1984

From: Bioinformatics for Microbiomes
Microbial diversity analysis with amplicon sequencing

HosMic MSc programme!!!

MSc contact person: Prof K. Kormas (kkormas@uth.gr)

ACTION