

Microbiome quantification using amplicon sequencing approaches

BIOS² Workshop – Day 1

Steven Kembel – Sciences biologiques, UQAM

Outline



The basics of amplicon sequencing



Study design considerations



From sequences to ecological data matrices

The basics of amplicon sequencing

Microbial diversity

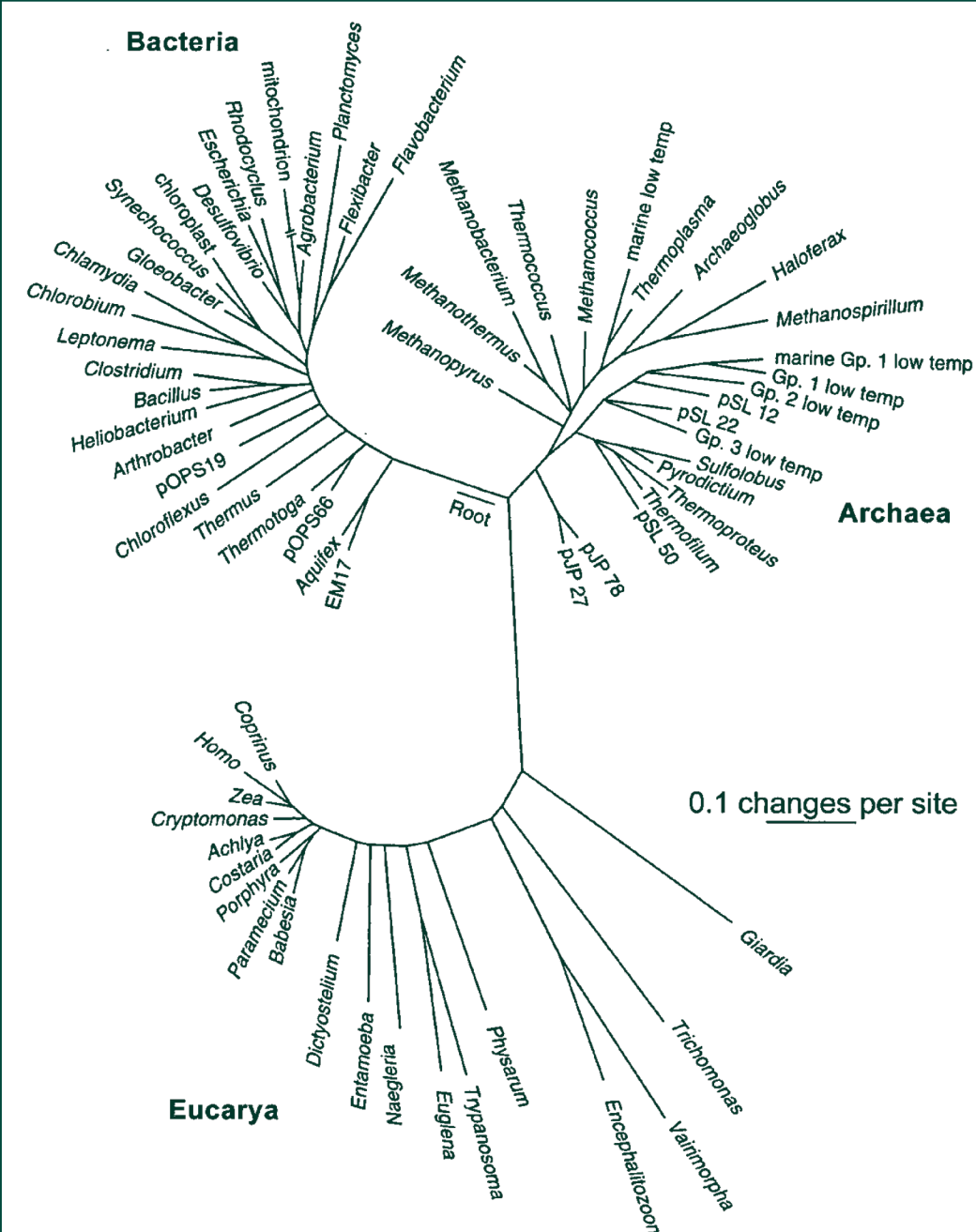
- from culture to sequencing

Metagenomic sequencing

- Sequencing DNA from the environment
- Metagenomic shotgun sequencing
- Amplicon sequencing

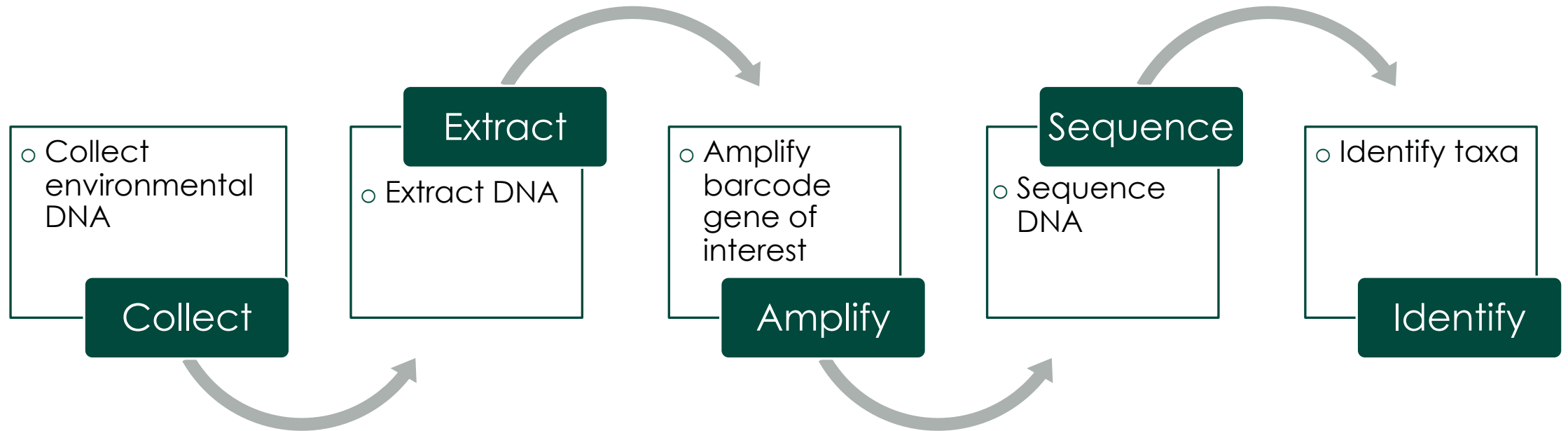
Amplicon sequencing

- a.k.a. metabarcoding, barcode gene sequencing, marker gene sequencing
- Amplify and sequence a gene of interest from an environmental sample



Pace, N.1997. A molecular view of microbial diversity and the biosphere. Science276:734-740.

Basics of the amplicon sequencing approach



Collect environmental DNA

DNA collection

There are many different methods

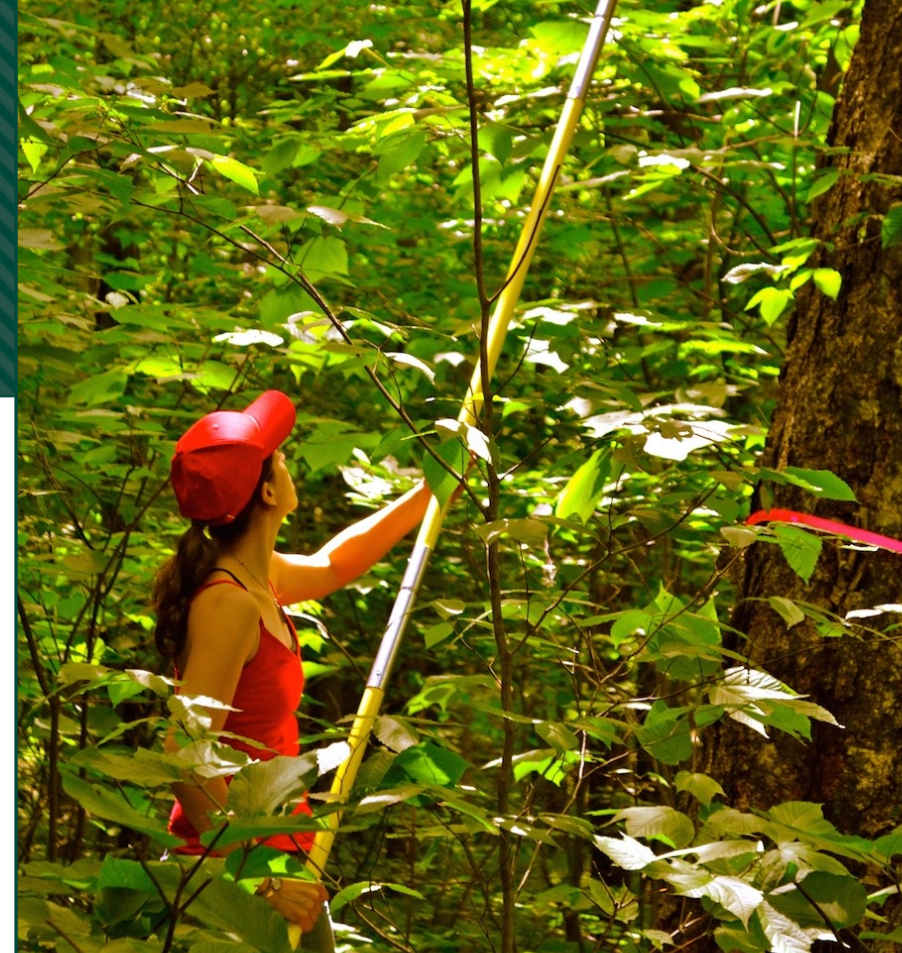
- Filters (air, water)
- Swabs and washes (surfaces)
- Bulk samples (tissues, substrate)

Issues to consider

Contamination

- Microbial biomass – need sufficient biomass to avoid issues with contaminants
- Host DNA

Don't forget to collect **negative control** samples (« blanks ») at the same time as 'real' samples.



Extract DNA

DNA extraction kits and protocols

There are many options available:

- Check the literature on **your study system** to see what is most widely **used in the community**
- i.e., for soil and plant samples: QIAGEN PowerSoil (Pro) kit
- Compare **results quality** and **cost**
- Consider the **collection method** used

DNeasy PowerSoil Kit Procedure

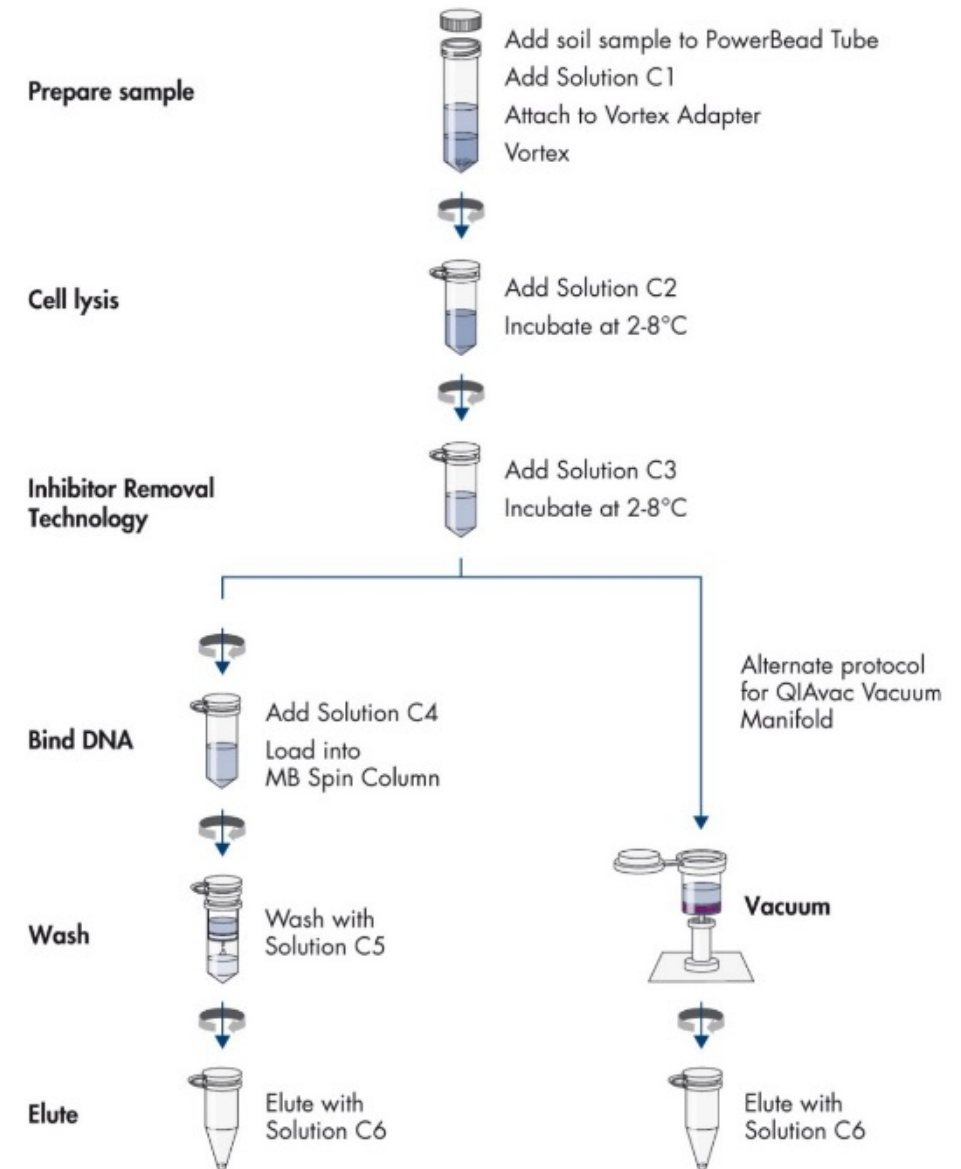


Figure 2. DNeasy PowerSoil Kit procedure.

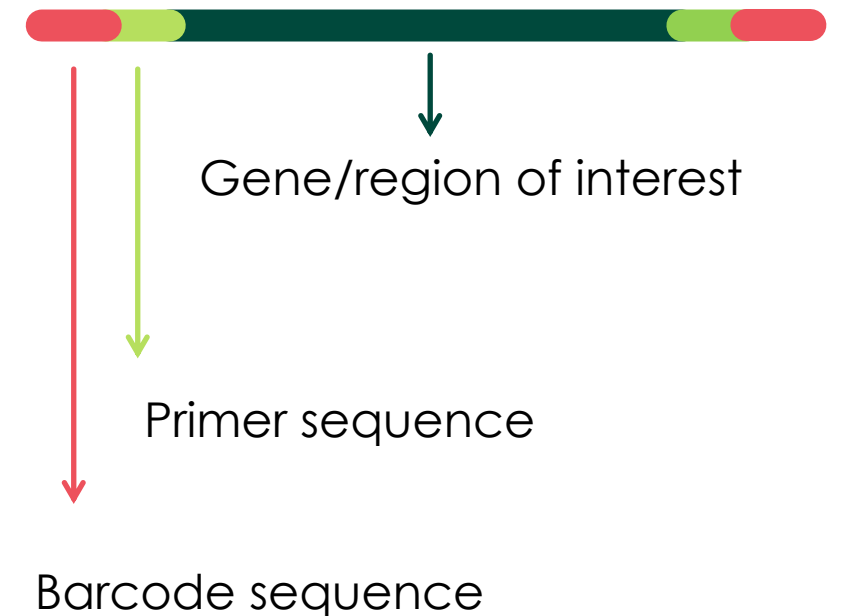
Amplify barcode gene of interest

PCR to amplify the gene of interest

- Bacteria: 16S
- Fungi: ITS, 28S, 18S
- Eukaryotes: 18S, COI
- Plants: rbcL, matK

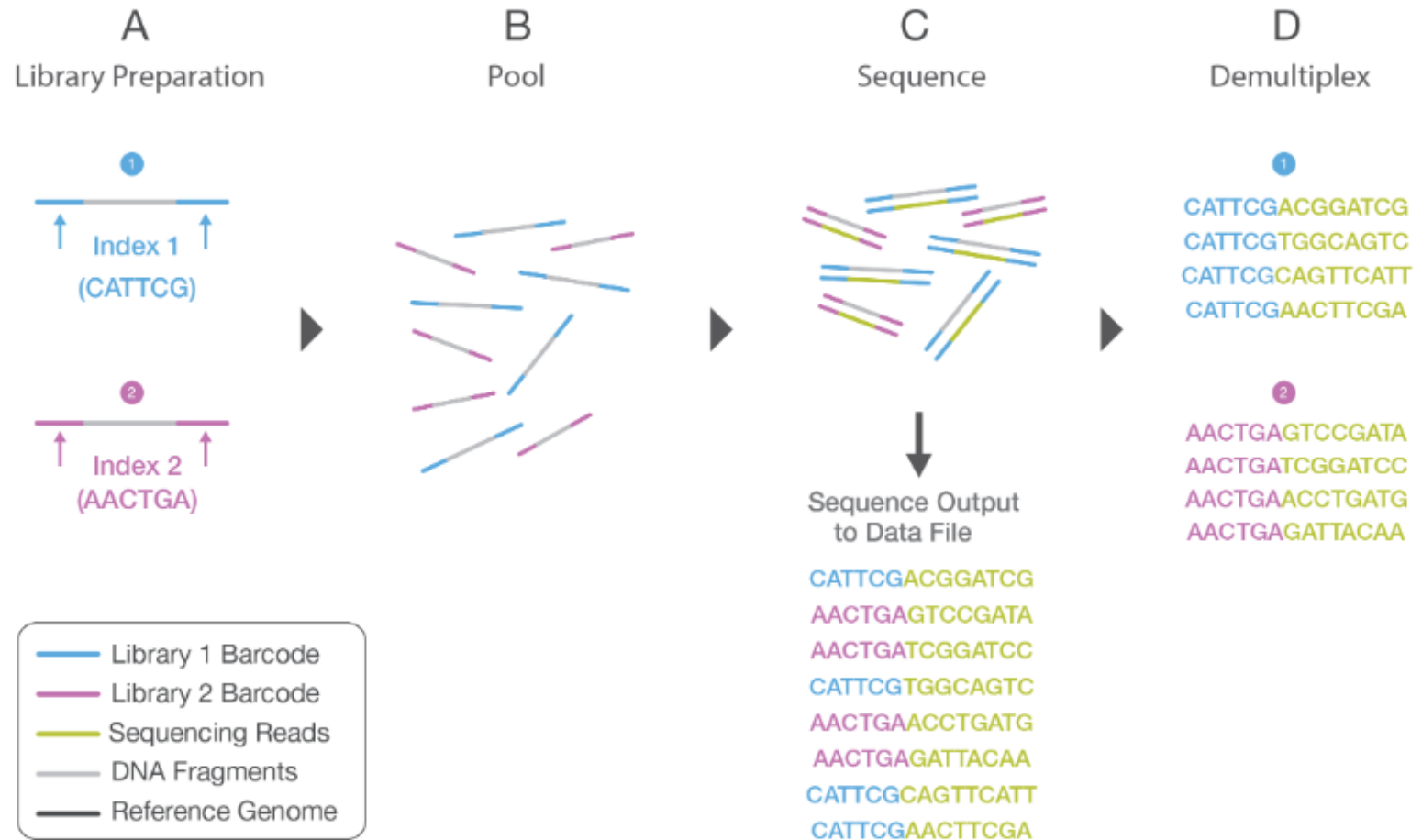
Multiplexing

- We add **unique barcode sequences** to DNA from each sample
- Pool samples together for sequencing
- Demultiplex bioinformatically after sequencing



Library preparation and sequencing

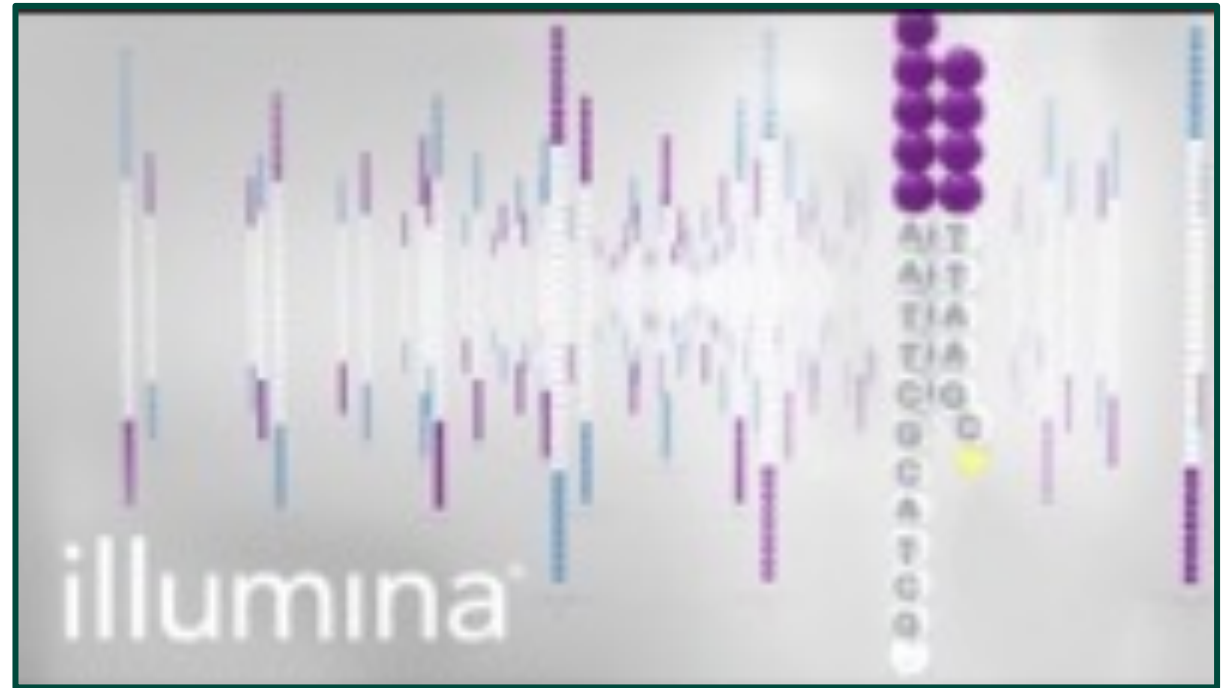
- Library preparation
- Multiplexing
- PCR
- Quality control
- Sequencing



Illumina sequencing

Illumina MiSeq

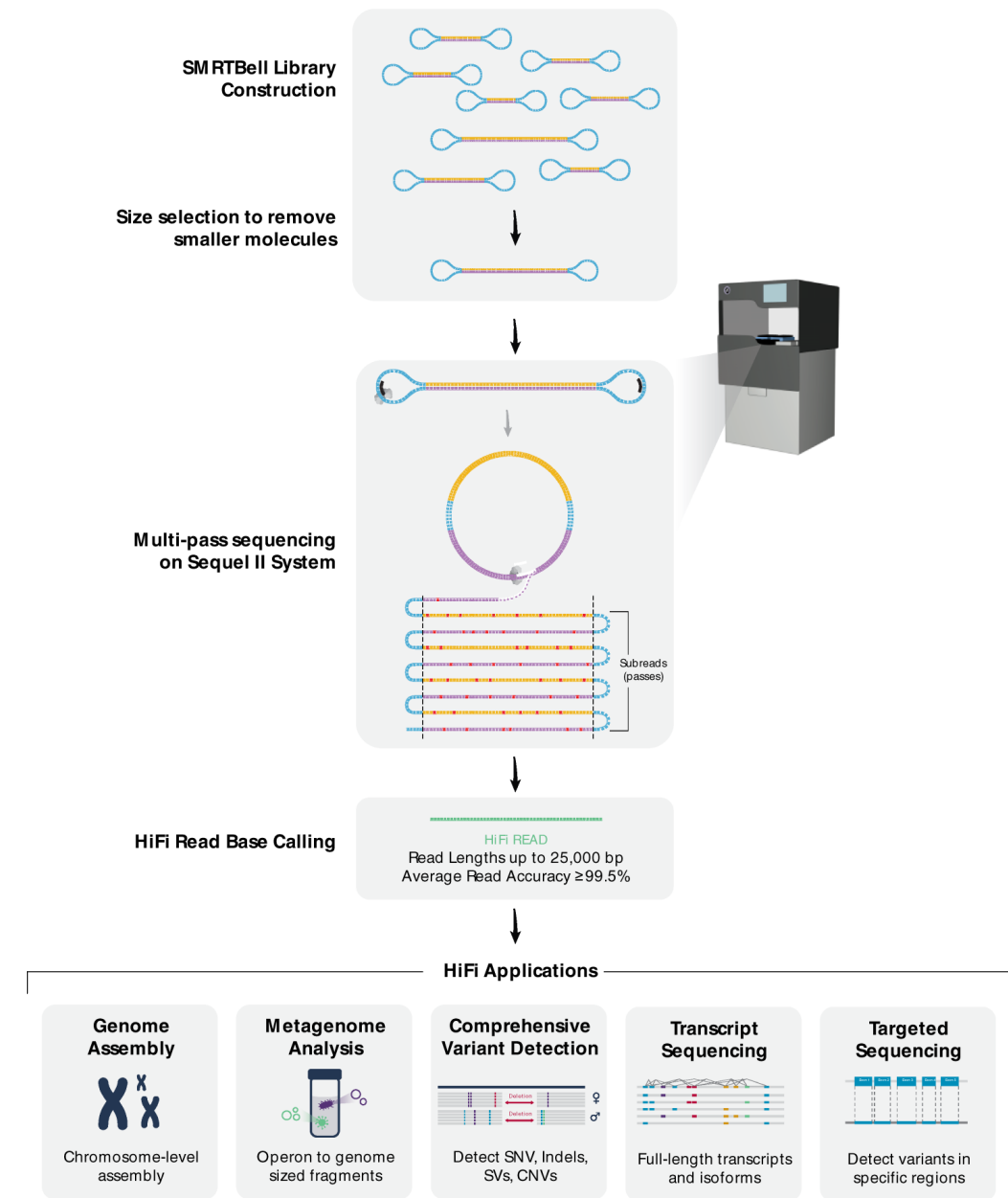
- Most widely used sequencing technology for amplicon sequencing
- Provides large amounts of short reads
 - Read length may not span entire gene/region of interest
- 15-20 million reads/run
- 2 x 250bp or 2x300bp per sequence



<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

Long read sequencing

- Several long-read technologies are becoming available
 - Oxford Nanopore
 - Pacific Biosciences
- Gives very long reads than span entire genes/regions of genome
- Just starting to be applied to amplicon sequencing



Hon, T., Mars, K., Young, G. et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* 7, 399 (2020). doi: [10.1038/s41597-020-00743-4](https://doi.org/10.1038/s41597-020-00743-4)

Identifying taxa from sequence data: ASVs and OTUs approaches

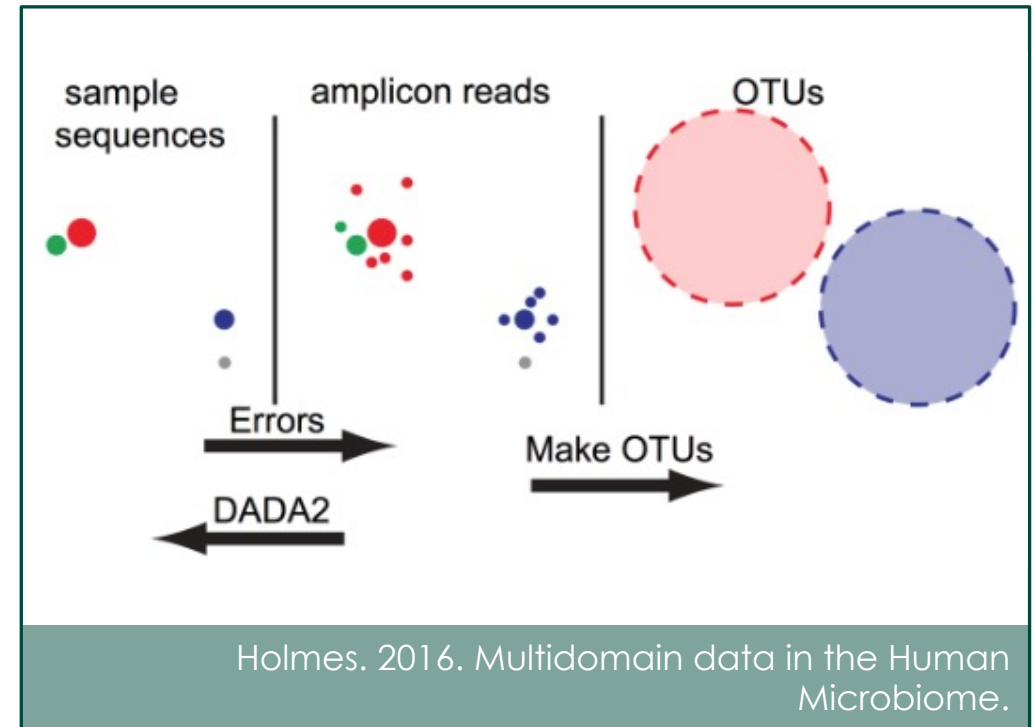
To carry out ecological analysis, we want to identify taxa in our sequencing data

Operational taxonomic unit (OTU)

- Clusters reads together based on sequence similarity into OTUs
 - Bacteria 16S – 97% OTU similarity cutoff

Amplicon sequence variant (ASV)

- Identify error-corrected identical sequences (ASVs)
- Advantages compared with OTU approach
 - Can compare ASVs among studies
 - Can easily classify future sequences into ASVs
 - Higher resolution (100% versus 97% similarity)



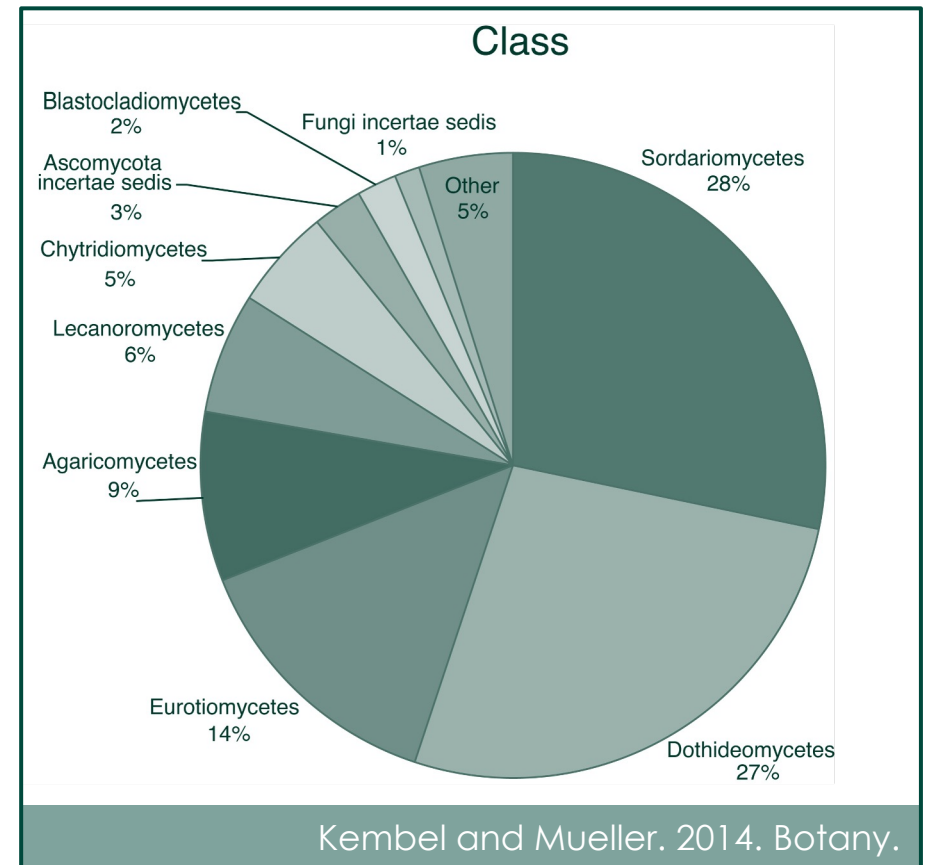
Identifying taxa from sequence data: Higher-level taxonomic annotation

Reference databases used to identify our sequences (ASVs)

- SILVA – rRNA - 16S, 18S, 28S
- GTDB – genome taxonomy database
- UNITE – eukaryotic ITS (fungi)
- UniEuk/EukRef – eukaryotic 18S

Limitations

- Reference databases are incomplete
- Focus on human-associated microbes and environments
- Many sequences are unidentified or can only be identified to higher taxonomic ranks



Study design considerations

- Amplicon sequencing technology and bioinformatics approaches are constantly evolving
- Decisions made at each step in the process can affect the outcome and ability to use and publish results
- Involve a microbiome collaborator in the project early

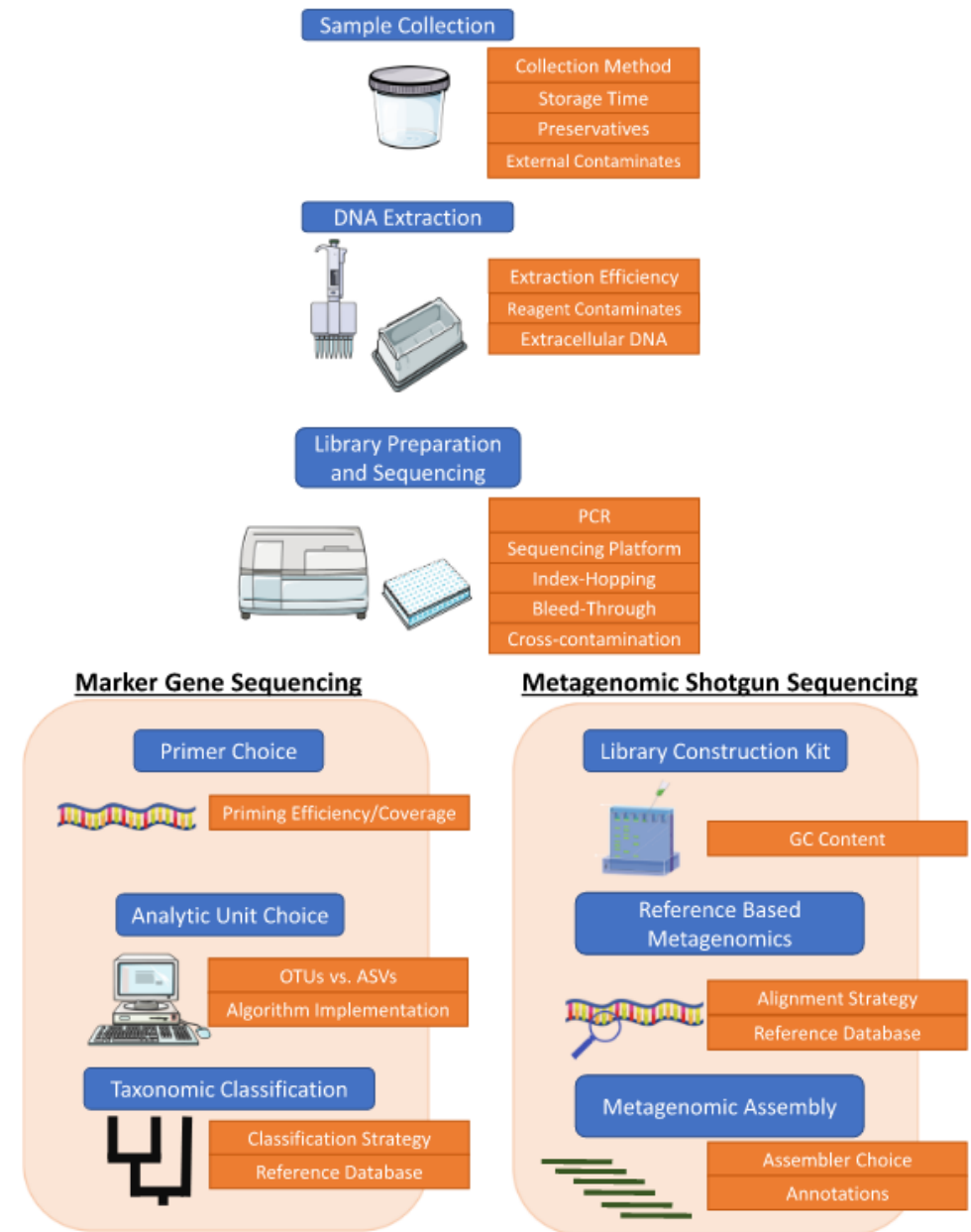


Fig. 1 The various stages that can introduce bias in sequenced-based human microbiome studies. Each blue box represents a stage in either DNA marker gene sequencing or DNA shotgun sequencing experiments. Orange boxes represent the various areas within a stage that can result in the introduction of systemic bias. Figure created using images from Servier Medical Art (<http://smart.servier.com>)

DNA extraction kit choice

- Many different DNA extraction kits and protocols can be used
- Check the literature on your study system to see what extraction kits are used most widely in the community
- Careful – it is hard to compare results among different kits and different lab protocols!
- Considerations
 - Sample storage and processing
 - DNA extraction kit / batch effects
 - Contamination

Goffau et al. *Microbiome* (2021) 9:6
<https://doi.org/10.1186/s40168-020-00949-z>

Microbiome

LETTER TO THE EDITOR **Open Access**


Batch effects account for the main findings of an in utero human intestinal bacterial colonization study

Marcus C. de Goffau¹, D. Stephen Charnock-Jones^{2,3}, Gordon C. S. Smith^{2,3} and Julian Parkhill^{1*}

Abstract

A recent study by Rackaityte et al. reported evidence for a low level of bacterial colonization, specifically of *Micrococcus luteus*, in the intestine of second trimester human fetuses. We have re-analyzed their sequence data and identified a batch effect which violates the underlying assumptions of the bioinformatic method used for contamination removal. This batch effect resulted in *Micrococcus* not being identified as a contaminant in the original work and being falsely assigned to the fetal samples. We further provide evidence that the micrographs presented by Rackaityte et al. are unlikely to show Micrococci or other bacteria as the size of the particles shown exceeds that of related bacterial cells. Finally, phylogenetic analysis showed that the microbes cultured from the fetal samples differed significantly from those detected by sequencing. Overall, our findings show that the presence of *Micrococcus* in the fetal gut is not supported by the primary sequence data. Our findings underline important aspects of the nature of contamination for both sequencing and culture approaches in microbiome studies and the appropriate use of automated contamination identification tools.

Keywords: Batch effects, Decontam, Colonization in utero, 16S rRNA



Primer choice

- Choice of a primer to use for amplicon sequencing depends on many factors
 - Targeted organisms
 - Host DNA?
 - Amplicon length
 - Community standards
- Keep biases in mind
 - Copy number variation
 - Universality / amplification bias
- It's difficult to compare sequences obtained from different primers and lab protocols

List of primers:

▼ 16S V5-V6 (Bacteria, excluding cyanobacteria)

▼ 16S V4-V5* (Universal Bacteria + Archaea)

▼ ITS1 (Fungi)

▼ Custom primers

<https://www.cermofc.uqam.ca/en/technological-platforms/genomics>

Replication and experimental design

Biological versus technical replicates

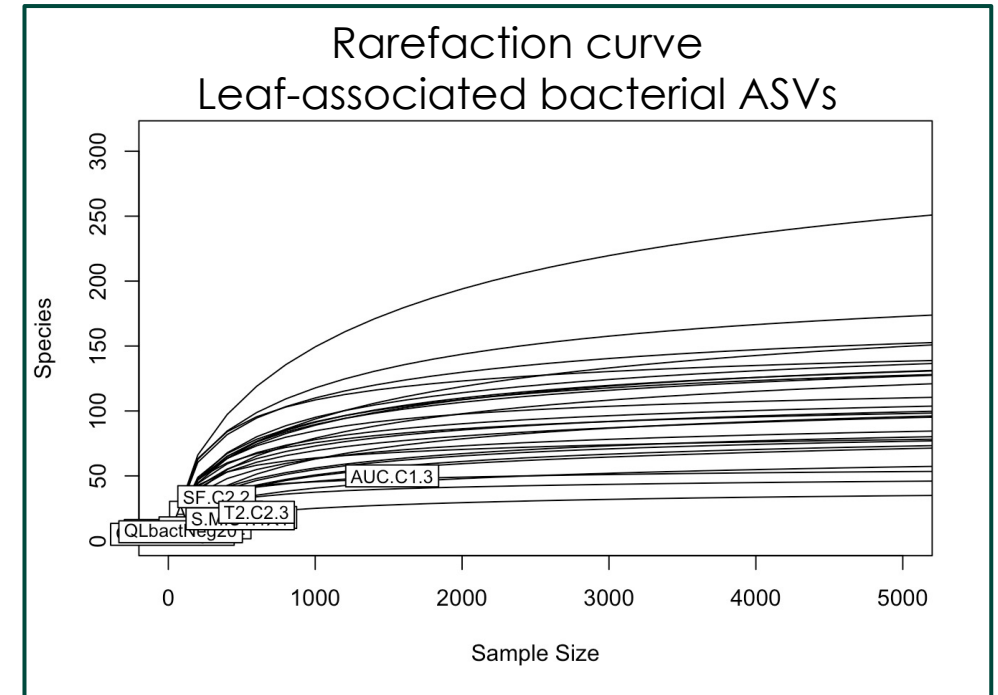
Kit and run effects and contaminants

Positive/negative controls

Compositional data

Biological vs technical replicates

- Need proper biological replicates to analyse data and respond to biological hypotheses/questions
 - Technical replicates are useful to quantify variability but won't help you answer your biological question
 - For a given amount of sequencing, there is a tradeoff between number of samples versus sequencing depth per sample (library size)
 - Think about the analyses you will do with your data, make sure you have enough replication (power analysis, etc.)
- Consider pilot studies or consult the literature to get an idea of variation among technical replicates and sequencing depth needed to capture the diversity of your samples



Kit and run effects and contaminants

- DNA extraction kit, PCR, and sequencing batch effects impact sequencing data
- Create and sequence negative extraction and PCR controls to identify contaminants
- Randomize samples among kits and sequencing runs
- It is difficult to fully correct for these batch effects

Salter *et al. BMC Biology* 2014, **12**:87
<http://www.biomedcentral.com/1741-7007/12/87>



RESEARCH ARTICLE

Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter^{1*}, Michael J Cox², Elena M Turek², Szymon T Calus³, William O Cookson², Miriam F Moffatt², Paul Turner^{4,5}, Julian Parkhill¹, Nicholas J Loman³ and Alan W Walker^{1,6*}


Olomu *et al. BMC Microbiology* (2020) 20:157
<https://doi.org/10.1186/s12866-020-01839-y>

BMC Microbiology

RESEARCH ARTICLE

Open Access

Elimination of “kitome” and “splashome” contamination results in lack of detection of a unique placental microbiome

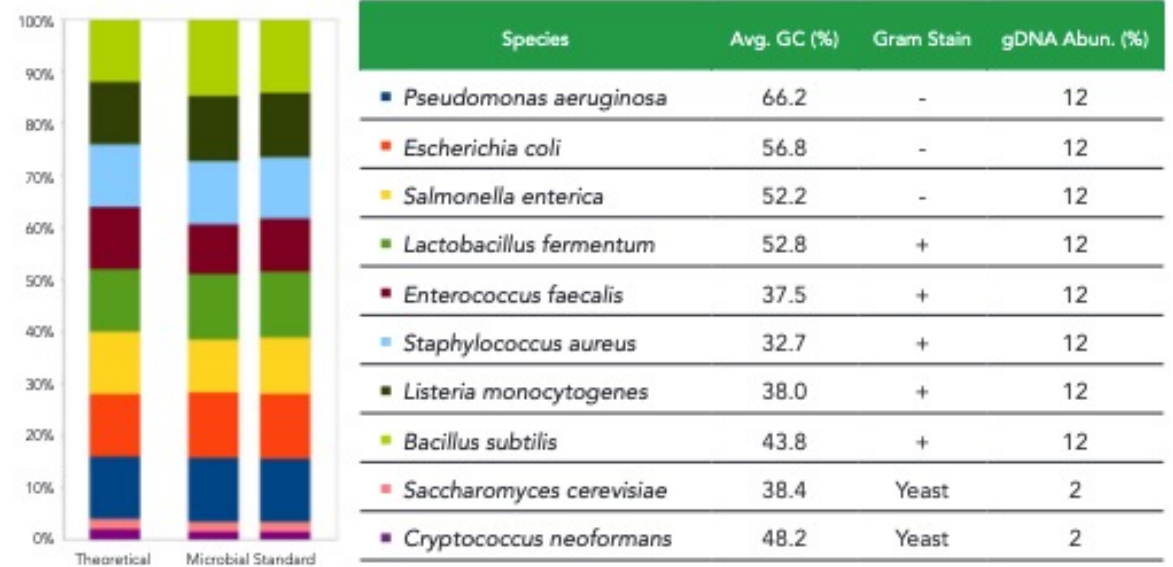
Isoken Nicholas Olomu^{1*} , Luis Carlos Pena-Cortes², Robert A. Long^{3,4}, Arpita Vyas⁵, Olha Krichevskiy³, Ryan Luellwitz⁶, Pallavi Singh⁷ and Martha H. Mulks⁸



Positive and negative controls

- Negative controls
 - You should create and sequence negative controls
 - Check for contaminants
 - Important when microbial biomass is low
 - Required by many microbial ecology journals
 - Blank/extraction controls
 - e.g. run blank sample through full sample collection and preparation protocol
 - PCR controls
 - no template DNA
- Positive controls
 - Sequence a mock community with known composition

Defined Microbial Community



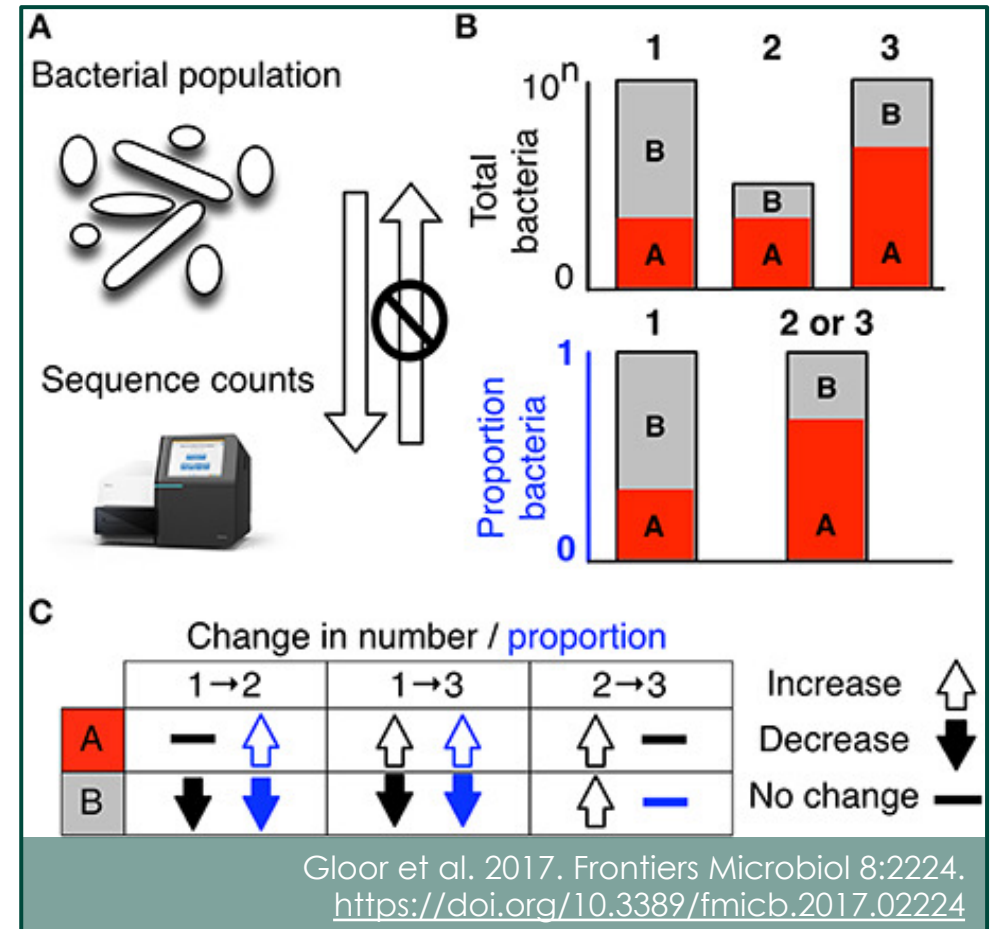
The ZymoBIOMICS® Microbial Community Standard contains three easy-to-lyse bacteria, five tough-to-lyse bacteria, and two tough-to-lyse yeasts.

Zymo Research.

<https://www.zymoresearch.com/collections/zymbiomics-microbial-community-standards/products/zymbiomics-microbial-community-dna-standard>

Microbiome data are compositional

- We normalize samples to approximately the same amount of DNA per sample for sequencing
 - The number of reads per sample (library size) does not tell us anything about absolute abundance
 - Amplicon sequencing gives us only relative abundance information (compositional data)
- Information on total/absolute abundance of organisms needs to be obtained using other approaches
 - qPCR / ddPCR
 - Spike-in approaches
 - Cell sorting / microscopy



From sequences to ecological data matrices

Sequencing files – FASTQ format

Demultiplexing

DADA2 algorithm for ASV identification

QC and visualization steps

ASV binning

Chimera removal

Taxonomic annotation

Sequencing files – FASTQ format

FASTQ files contain information on sequences and quality scores

header	@M02360:47:000000000-BC2LN:1:1101:21751:2134 1:N:0:1
sequence	GAGACTATATGCAGGGTTGCGCTCTTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACAGCCATGCA GCACCTGTGTGCGCGCCACCGAAGTGGACTGGGAATCTCTCCCATAACACGCCATGTCAAAGGATGGTAAGG TTC TGCGCGTTGCTTCGAATTAAACCACATGCTCCACCGCTTGTGCGGGCCCCCGTCAATTCCTTTGAGTTTTAATCTT GCGACCGTACTCCCAGGCGGAATGCTCAAAGCGTTAGCTGCGCTACTGAGGTGCAATCACCCACCCGCTGG
spacer	+
quality score	9CCCCGGGGGGF9FFGGFFGGGGGGGGFGGGG7FFGGGGGGDGGGGGGG9FCEGEGG7CCFFGG77C@FGFGGFF9 EF8FFGGGFF8@FG7CCEEEEE>FF, E9=BEF8, , EFEGFFGFCFE8, A, >FEG+@, CF, , 3CD8FGGGGFGCFGG GFC+@8EGDGFGGEGCDE9=; BCFF8F??ADGGGFEGGG4E6CC5) : CDCG3BE: DFFFFFFG?*7D9@FFFAFFF 9A6(49; ??? : AFFF(76<(4(-8:<:<2)<48;3:16:>2:6906>?<-41317))5)5:1;:(4(461(

Demultiplexing

- After sequencing, we need to demultiplex our samples to determine which sequences came from which samples
- You may have two files for each sample
 - Files labelled R1 and R2 contain forward/reverse reads, may contain barcodes and primers
 - Depends on the protocol used for library preparation and demultiplexing
 - Ask your sequencing centre if they can demultiplex your samples for you
- You may have two large files containing all samples
 - You will need to demultiplex your samples
 - idemp, mothur, QIIME2
 - Ask your sequencing centre for the list and format of barcode sequences for each sample

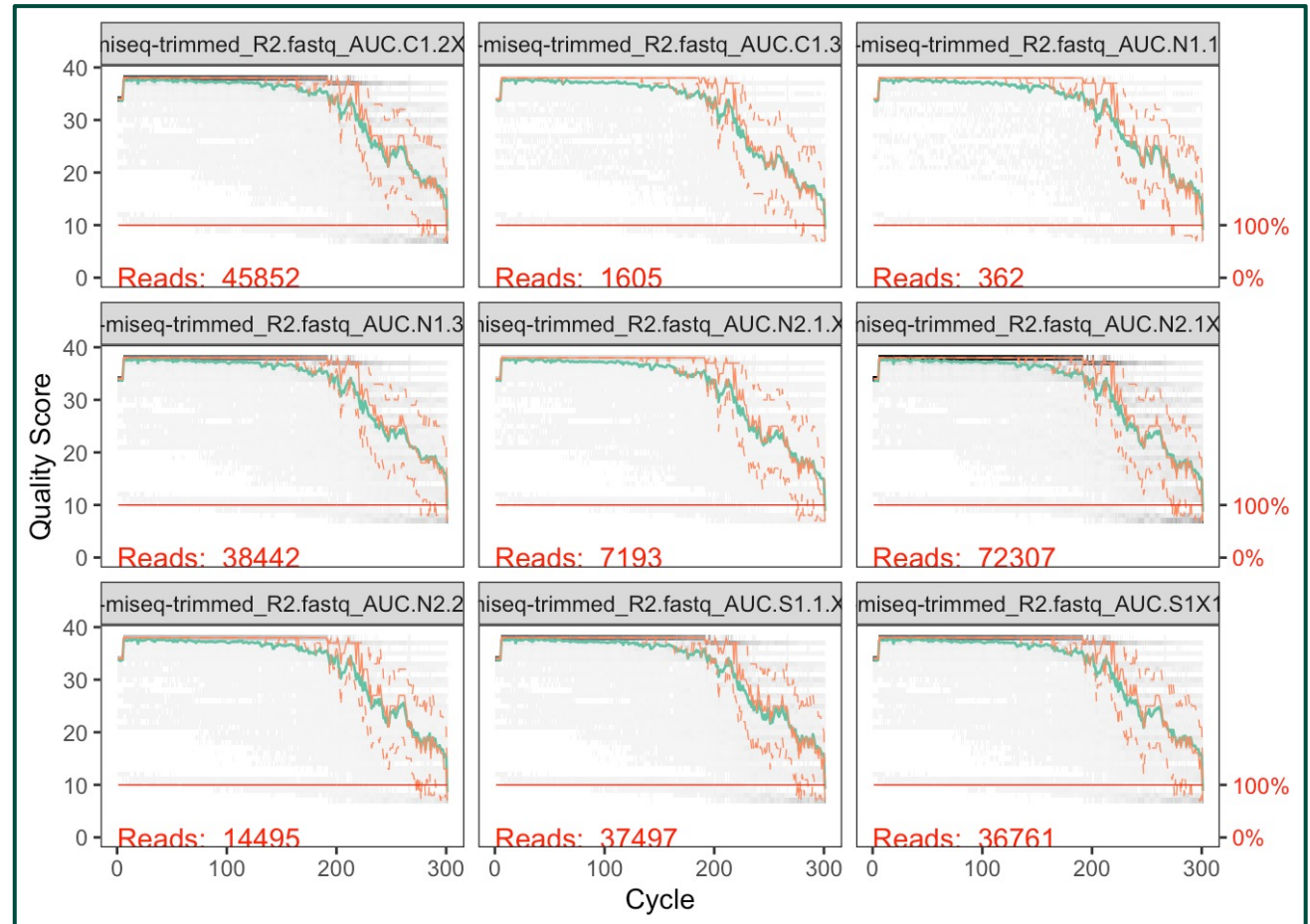
Barcode (6nt) Primer (16nt)

The diagram shows a sequencing read with a 6nt barcode and a 16nt primer highlighted. The barcode is GAGACTATATGC and the primer is AGGGTTGCGCTCTTTG. The read is shown in a black box with white text. The barcode is highlighted in green and the primer is highlighted in red. The read is shown in a black box with white text. The barcode is highlighted in green and the primer is highlighted in red.

```
@M02360:47:000000000-BC2LN:1:110
GAGACTATATGCAGGGTTGCGCTCTTTGCGGC
GCACCTGTGTGCGCGCCACCGAAGTGGACTGC
TGCGCGTTGCTTCGAATTAACACATGCTCC
GCGACCGTACTCCCAGGCGGAATGCTCAAAC
+
9CCCCGGGGGGF9FFGGFFGGGGGGGGFGGGC
EF8FFGGGFF8@FG7CCEEEEE>FF,E9=BEF
GFC+@8EGDGFGE9=;BCFF8F??AD
9A6(49;???:AFFF(76<(4(-8:<:<2)<4
```

QC and visualization steps

- We need to check the quality of our forward and reverse sequences
- We will trim our sequences and remove low quality sequences before proceeding



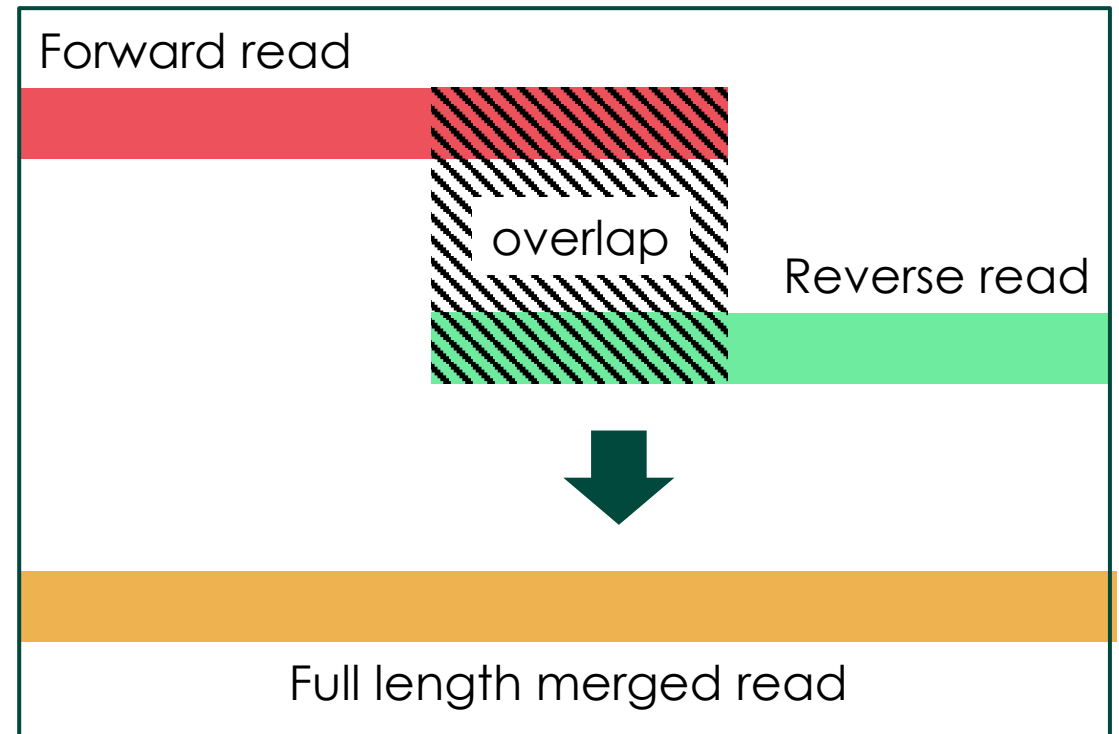
DADA2 algorithm for ASV identification

“This algorithm is built on a model of the errors in Illumina-sequenced amplicon reads

- The error model quantifies the rate λ_{ji} at which an amplicon read with sequence i is produced from sample *sequence* j as a function of sequence composition and quality.
- Then, a Poisson model for the number of repeated observations of sequence i , parameterized by the rate λ_{ji} , is used to calculate the p-value of the null hypothesis that the number of amplicon reads (the abundance) of sequence i is consistent with the error model.
- These p-values are used as the division criteria for an iterative partitioning algorithm, which continues dividing sequencing reads until all partitions are consistent with being produced from their central sequence.”

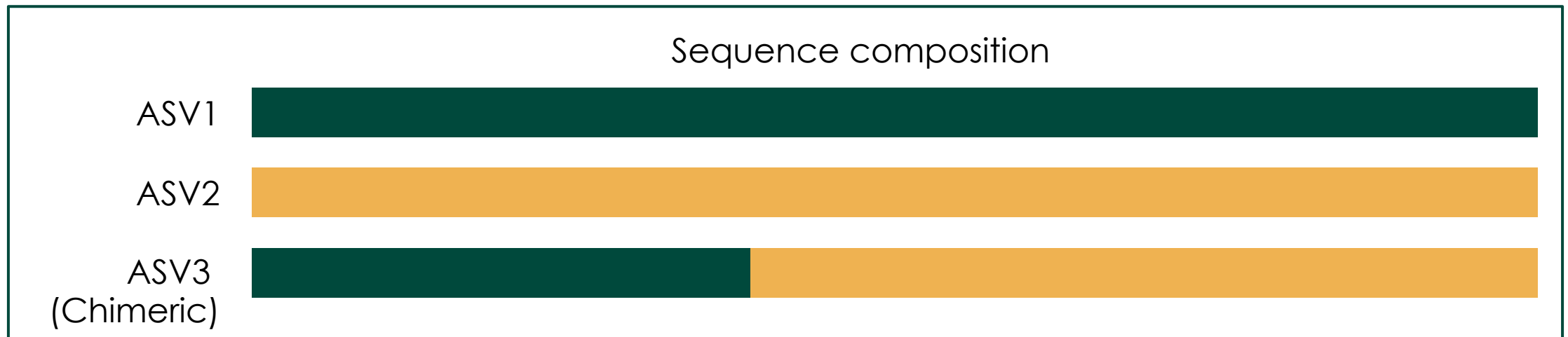
Merging denoised reads

- We can now merge together the denoised forward and reverse reads based on their overlapping nucleotides
- Remember that we need ensure the reads will overlap (as a function of amplicon length and read length)
- If the reads cannot be merged, we can still analyze the forward and reverse reads separately, but we cannot do exact species matching



Chimera removal

- Chimeric sequences contain DNA from different organisms
- Chimeras are common and arise during PCR and sequencing
- We need to remove chimeras from our data by identifying sequences with close matches to two different organisms/ASVs



Taxonomic annotation

1

- We annotate the taxonomic identity of ASVs by comparing them with a reference database using a **naive Bayesian classification approach**



2

- For species-level taxonomic annotation, we consider **only exact matches** to the reference database

From sequences to ecological data matrices

Summary of output files:

`seqtab.nochim.rds`

- Ecological data matrix – ASV abundances per sample

`taxa.sp.rds`

- Taxonomic annotation of ASVs

`Microbiome-sequence-analysis-workspace.Rdata`

- R workspace containing all data objects