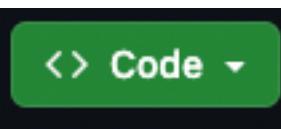


Welcome! We'll get started shortly!

- Before we start, visit the github repository for today's training here:
<https://github.com/ktmbiome-niaid/16S-data-processing>
 - You can also search for the username *ktmbiome-niaid* and find the 16S-data-processing repository
- Click on the green <Code> button 
- Click Download ZIP
- Also, make sure you have **R** and **Rstudio** installed. If you don't, visit NIH Self Service and install both.
- If you brought your own data to work on today, let us know in the chat!

Any questions? Also share in the chat!

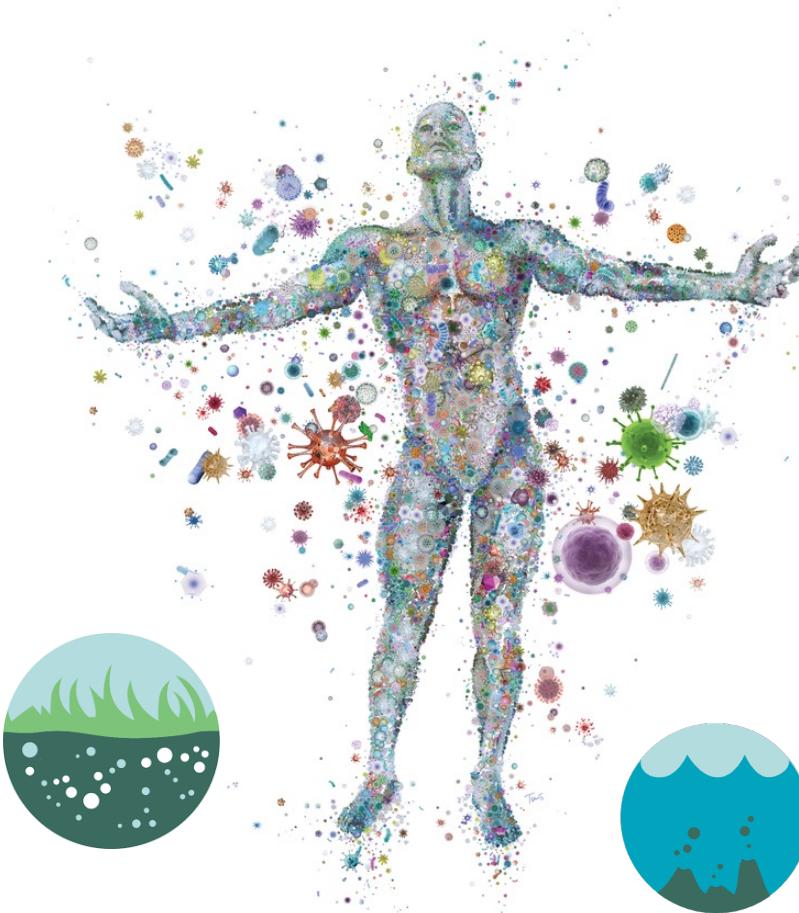
National Institute of Allergy and Infectious Diseases

16S rRNA amplicon sequencing

Kathryn McCauley, MPH
and
Lauren Krausfeldt, PhD

Bioinformatics and Computational Biosciences Branch (BCBB)
OCICB/OSMO/OD/NIAID/NIH

October 2023



NIAID



National Institute of
Allergy and
Infectious Diseases

Today's instructors

Kathryn (Katie) McCauley, MPH

Metagenomics Specialist

kathryn.mccauley@nih.gov

Lauren Krausfeldt, PhD

(Viral) Metagenomics Specialist

lauren.krausfeldt@nih.gov

Science Support: Metagenomics/Microbiome

Bioinformatics and Computational Biosciences Branch (BCBB)

National Institute of Allergies and Infectious Diseases (NIAID)

National Institute of Health (NIH)

Bethesda, MD, USA

Find out more about what is offered at the BCBB:

Website: bioinformatics.niaid.nih.gov

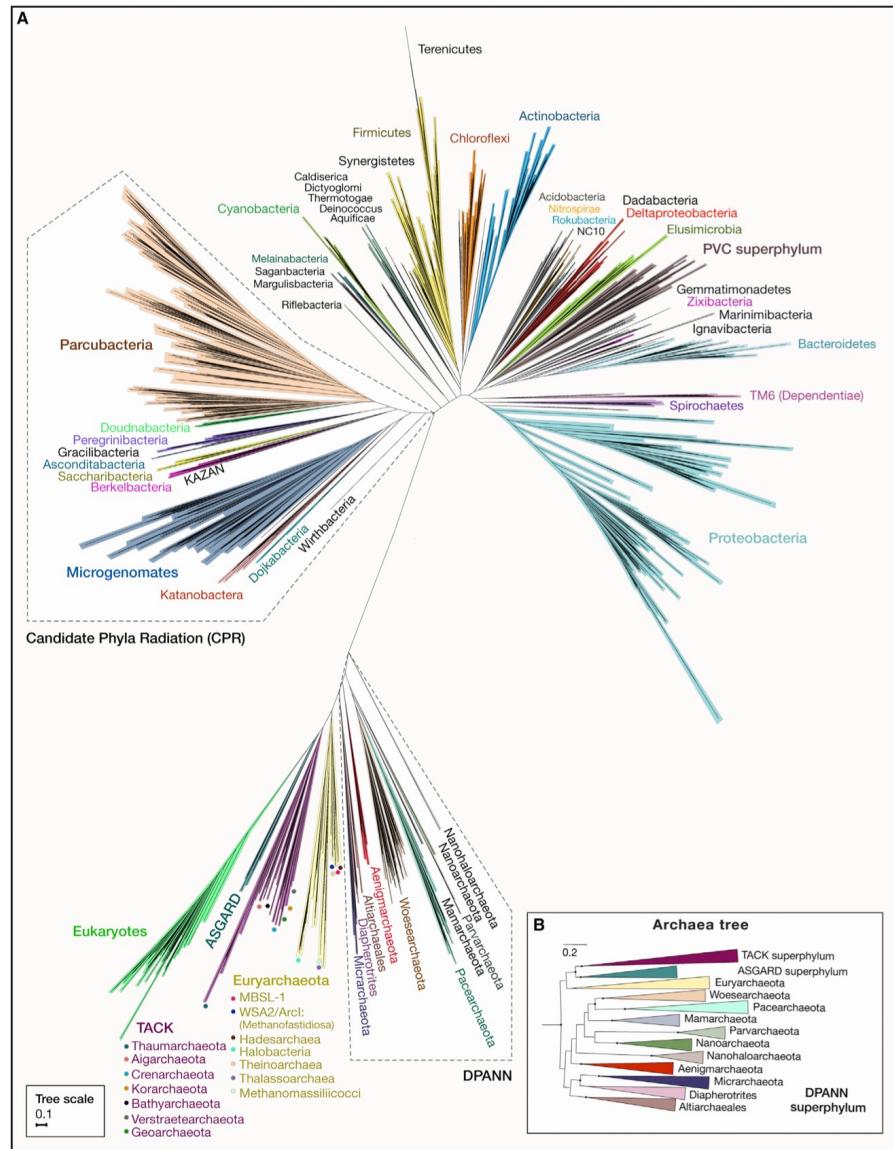
Contact: bioinformatics@niaid.nih.gov

Services are FREE!



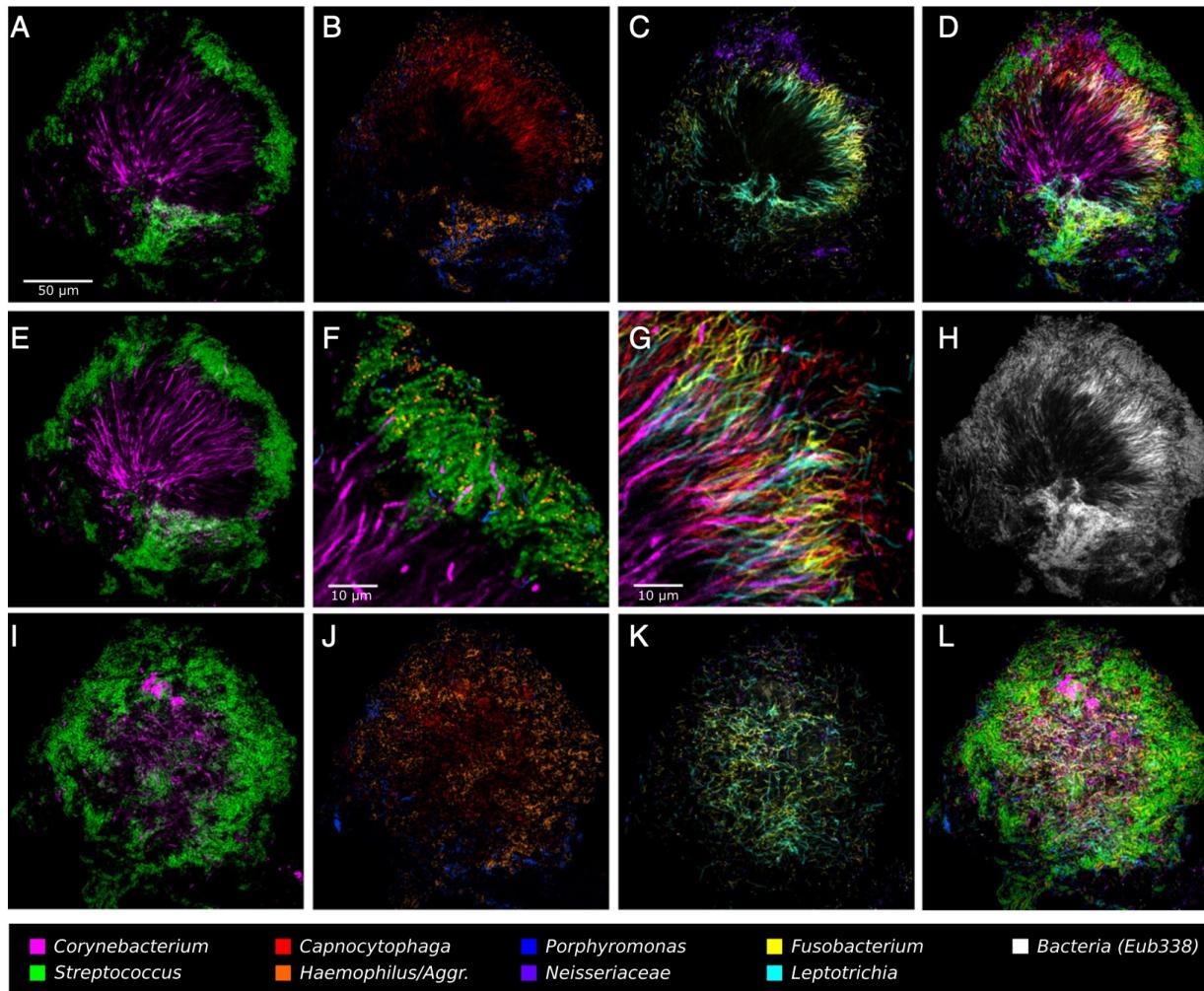
bioinformatics
@NIAID

Studying microbial communities



- The microbial world is extremely diverse (biomass construct ~1/2 of Earth's biomass)
- Less than 1% of microbes are culturable
- Cannot re-create natural conditions in the lab
 - Media (food, energy sources), temperature, light, salinity, etc.)

Studying microbial communities



- Community complexity – microbial interactions disrupted
- Molecular tools allow us to learn a lot more about microbes!

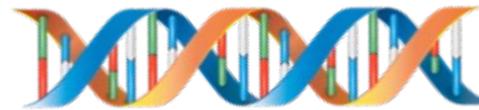
Molecular techniques to study microbial communities



16S AMPLICON SEQUENCING

DNA

- Who is there?
- Relative abundances
- Microbial diversity
- plus more!



(META)GENOMICS

DNA

- Who is there?
- Relative abundances
- Microbial diversity
- What can they do?
- Functional diversity



(META)TRANSCRIPTOMICS

RNA

- Who is active?
- How are they responding?
- What genes and functional pathways are activated?
- Who is doing* what?

Molecular techniques to study microbial communities



(META)PROTEOMICS

PROTEINS



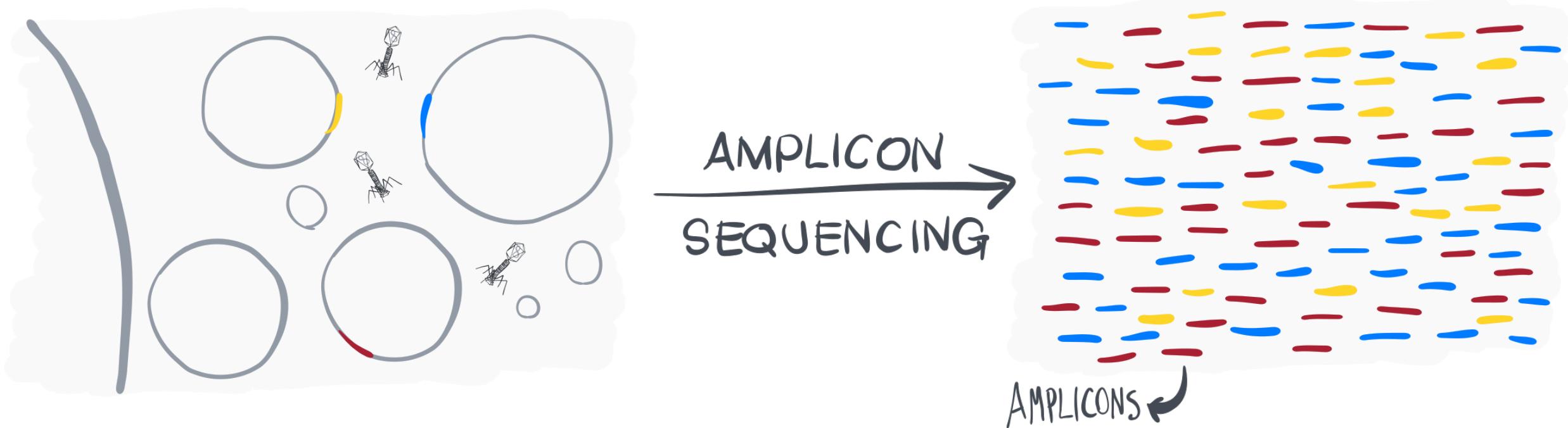
METABOLOMICS

METABOLITES

- Who is doing what?
- What proteins were actually produced?

- What are they doing?
- What processes have actually occurred?

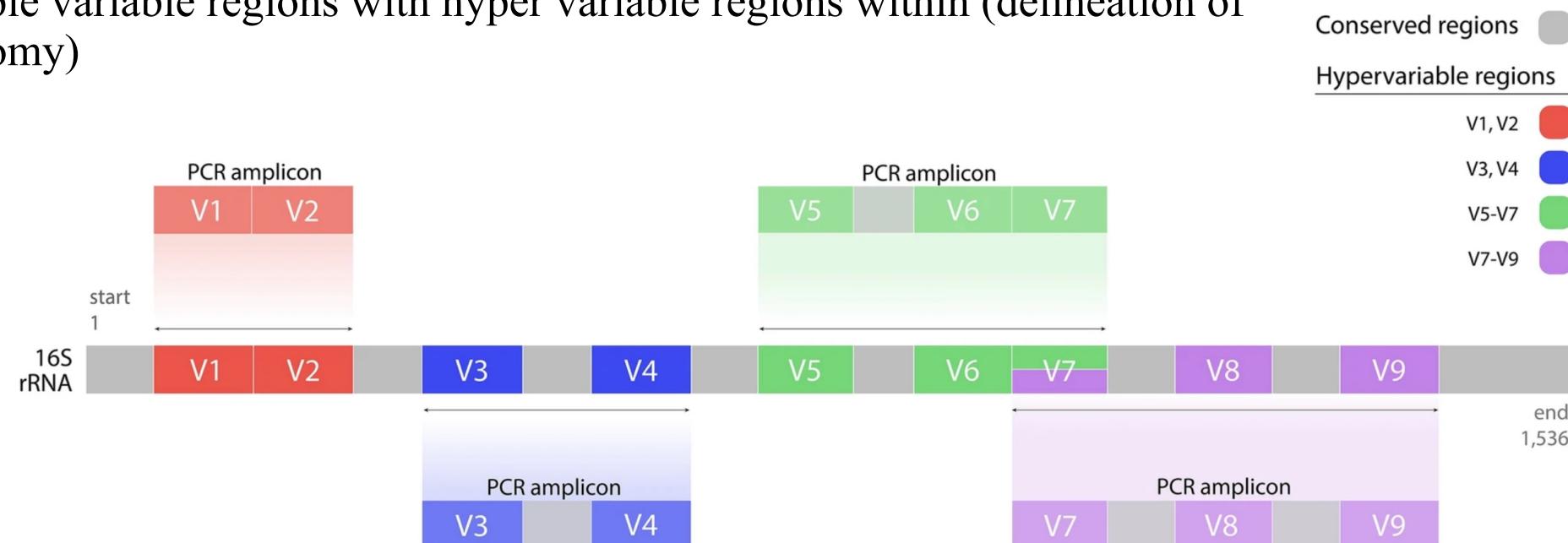
16S rRNA amplicon sequencing



Why 16S rRNA for microbial communities?

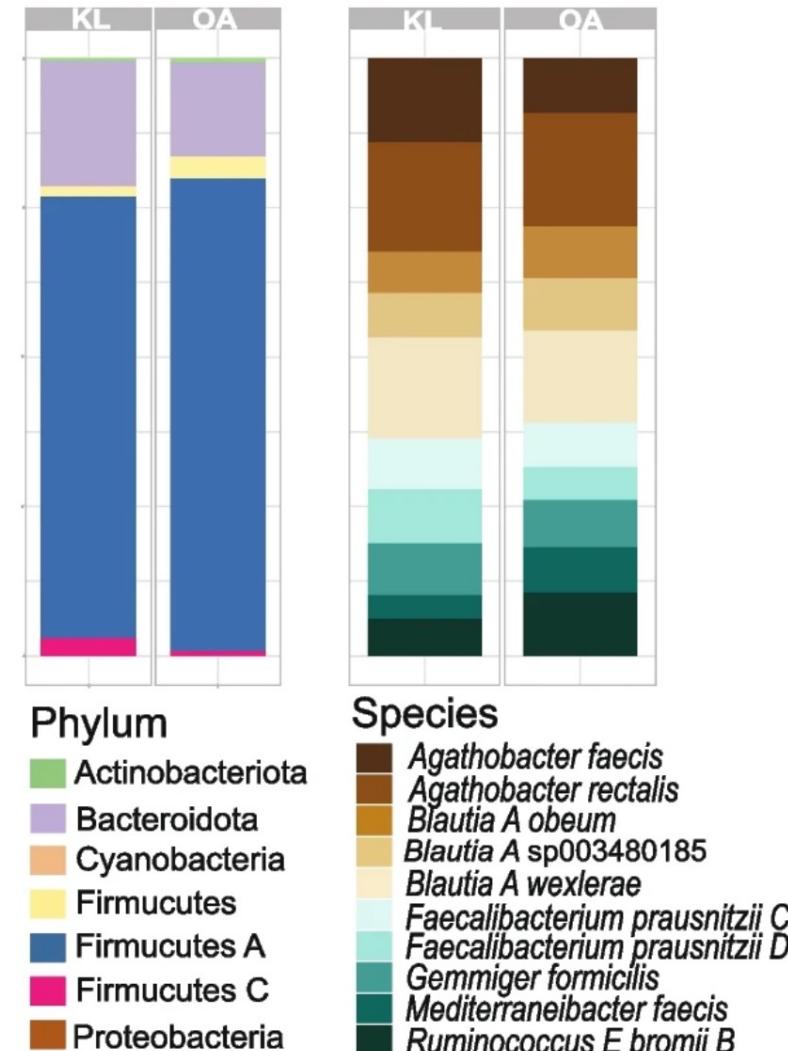
- Highly conserved gene in bacteria and archaea
 - Universal marker gene for prokaryotes
- Function has not changed over time – good for studying time (evolution)
- Length (1500bp) is enough for informatics
- Has multiple highly conserved regions (allows for primer binding across multiple species)
- Multiple variable regions with hyper variable regions within (delineation of taxonomy)

- ** Common marker genes for eukaryotes:
- 18S rRNA/23S rRNA
For Fungi/Algae
 - ITS region



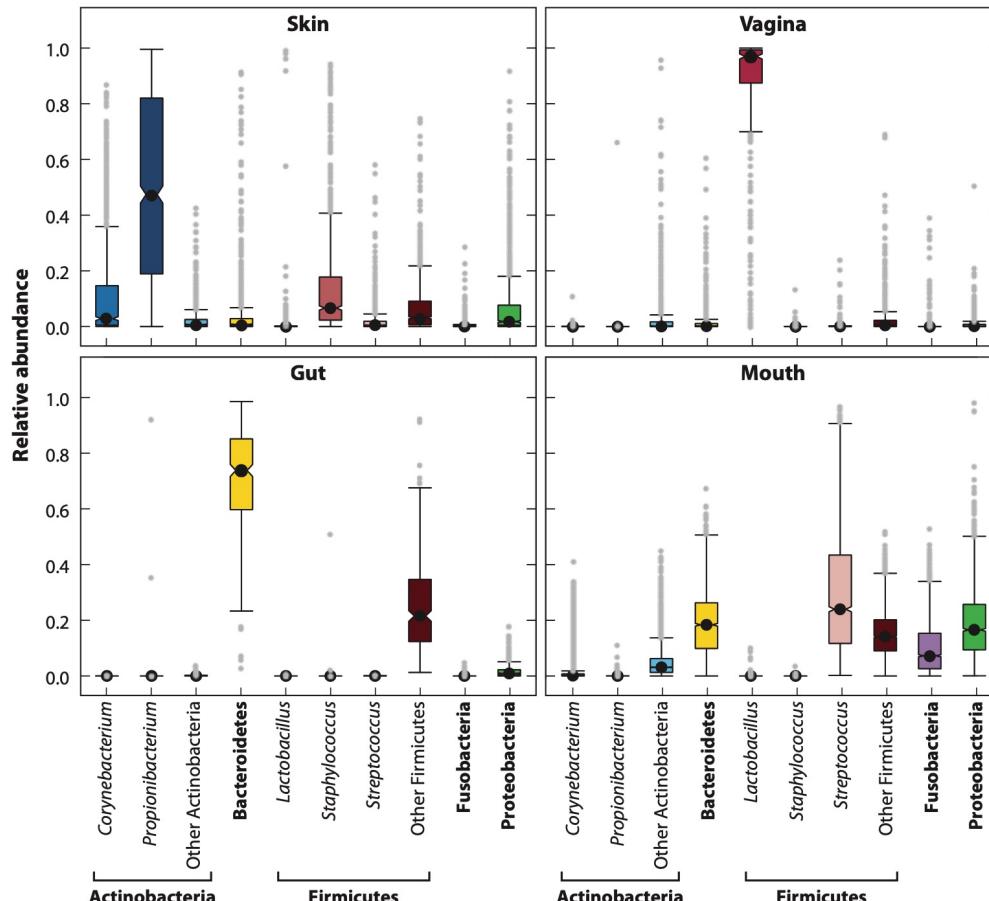
What can we learn/what questions can we ask?

- **Community Composition:** What organisms (prokaryotes) are present?
- **Community Structure:** What is the proportion of prokaryotes in each microbiome?



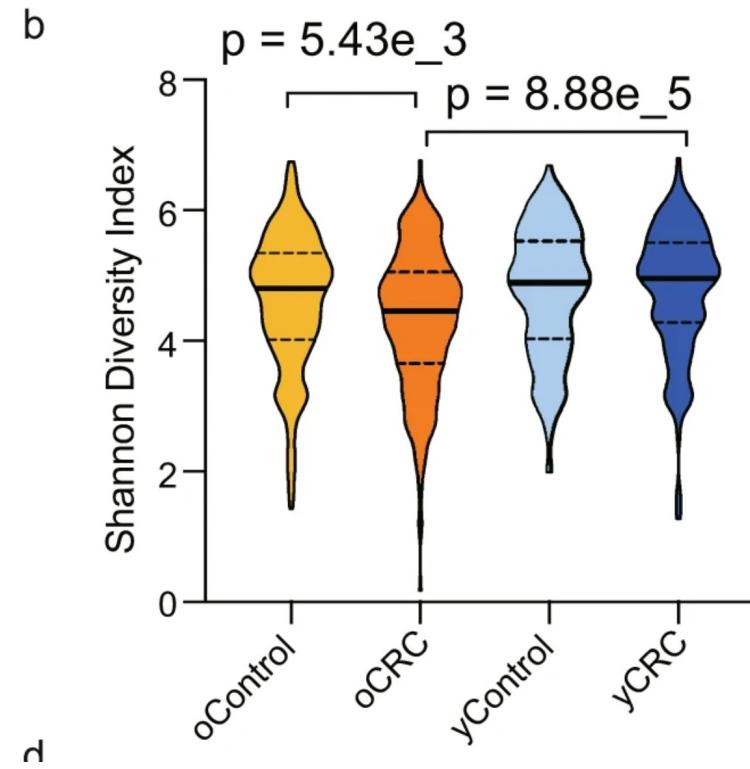
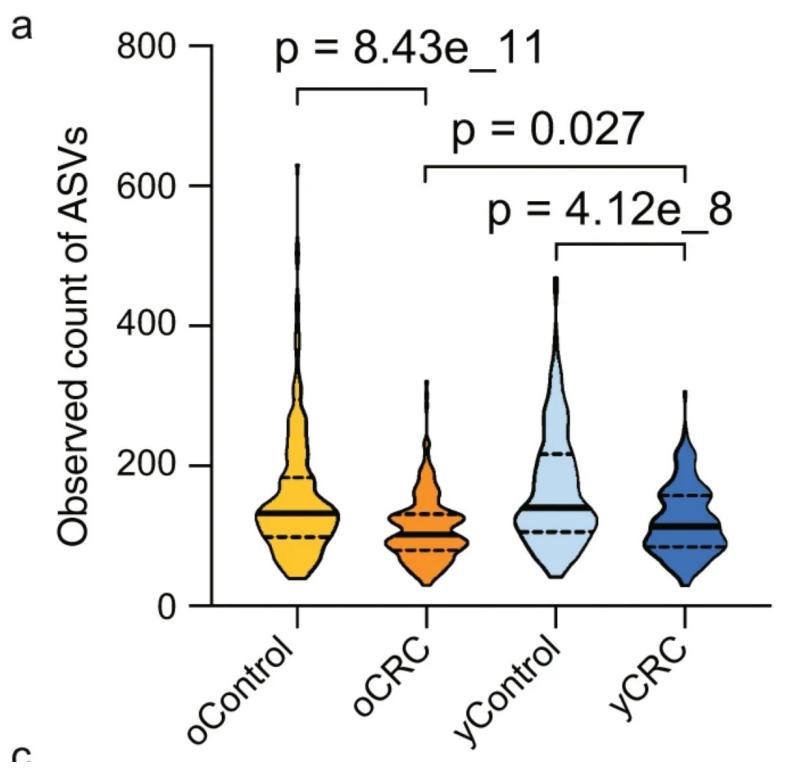
What can we learn/what questions can we ask?

- **Community Differences:** What are the differences in composition and structure between communities?



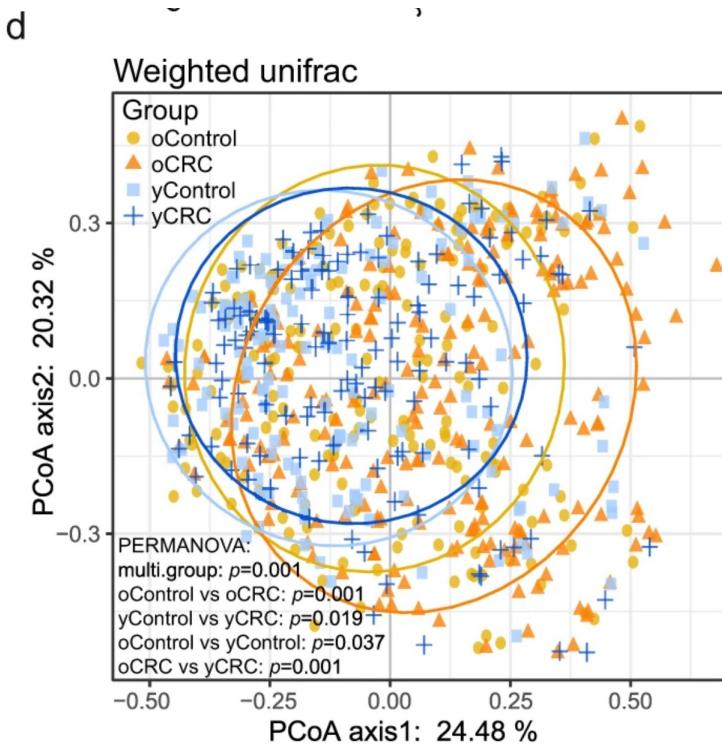
What can we learn/what questions can we ask?

- **Community Diversity:** What is the diversity of prokaryotes within each microbiome?



What can we learn/what questions can we ask?

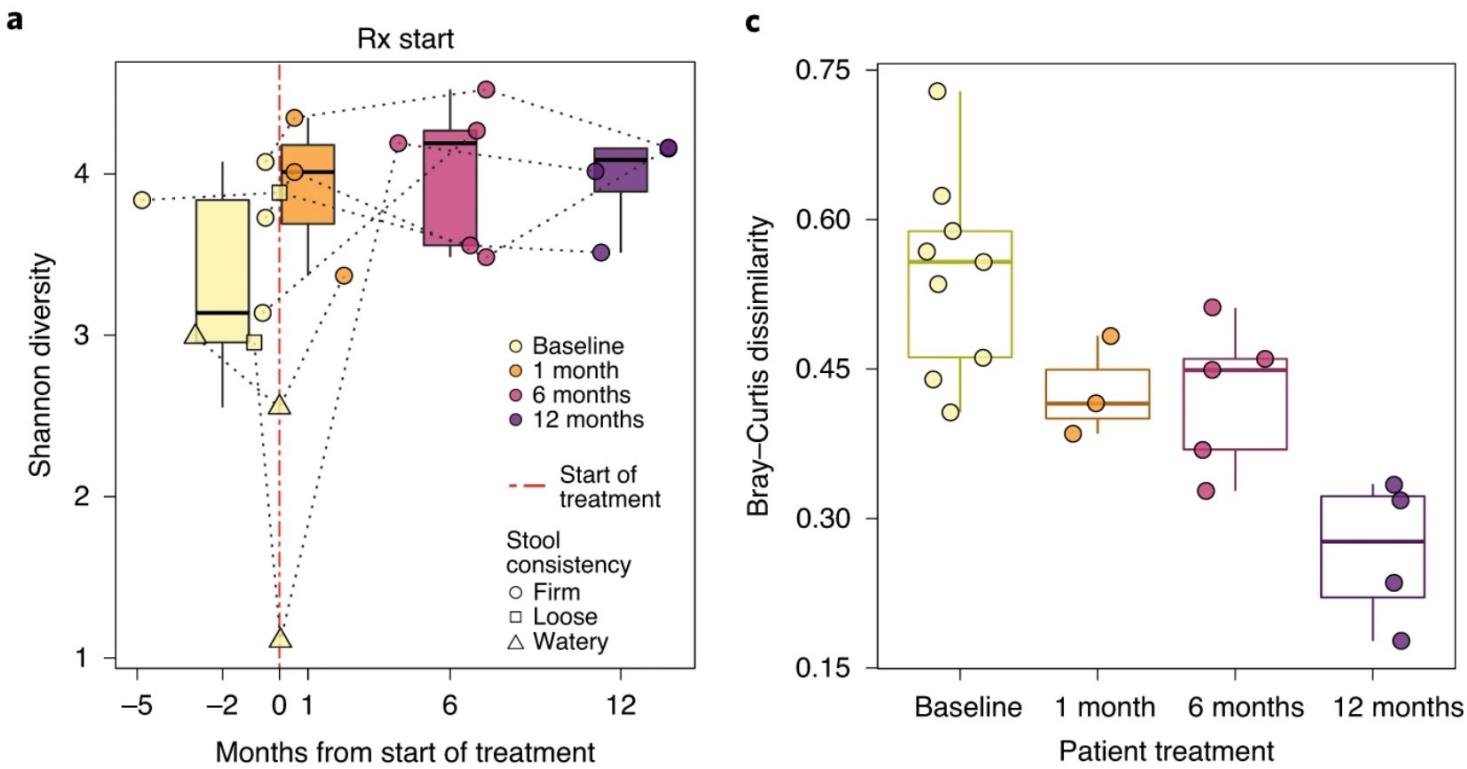
- **Community Response:** What are the changes in entire community in response to a factor (e.g. treatment, changing environmental condition)?



Yang, Y., Du, L., Shi, D., Kong, C., Liu, J., Liu, G., ... & Ma, Y. (2021). Dysbiosis of human gut microbiome in young-onset colorectal cancer. *Nature communications*, 12(1), 6757.

What can we learn/what questions can we ask?

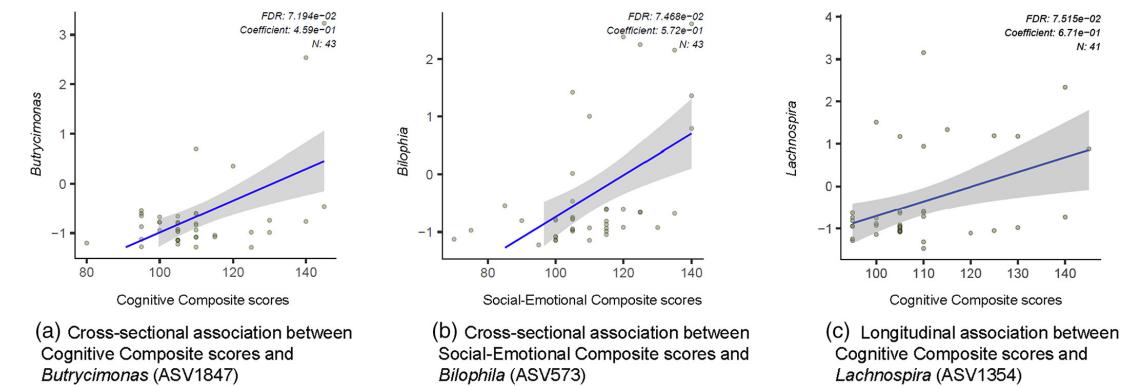
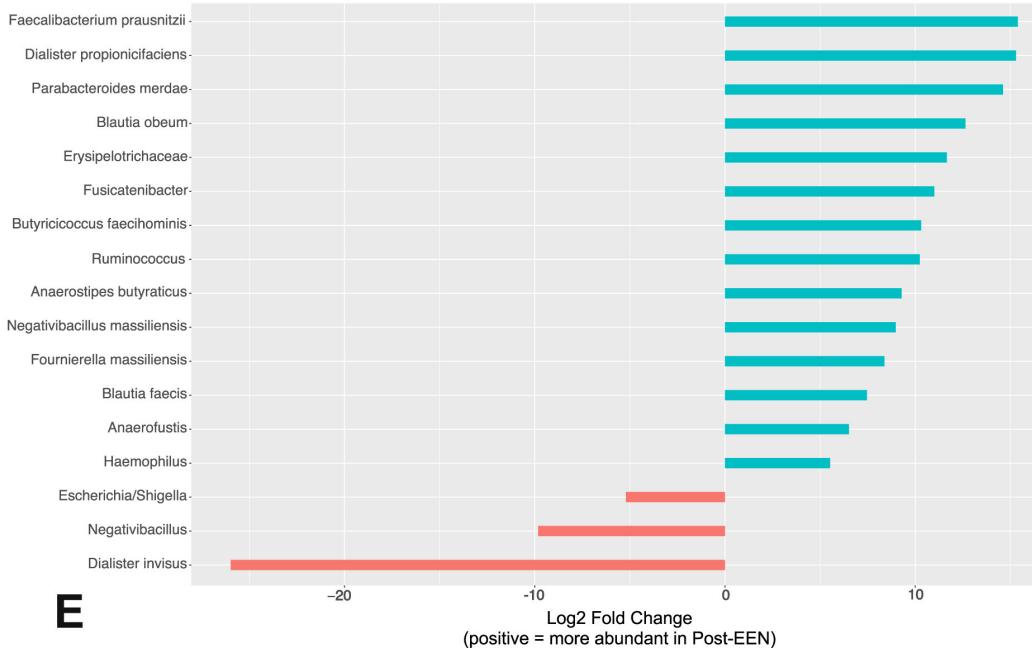
- **Community Temporal Dynamics:** What are the changes in microbial composition and diversity over time?



Ozen, A., Kasap, N., Vujkovic-Cvijin, I., Apps, R., Cheung, F., Karakoc-Aydiner, E., ... & Lenardo, M. J. (2021). Broadly effective metabolic and immune recovery with C5 inhibition in CHAPLE disease. *Nature immunology*, 22(2), 128-139.

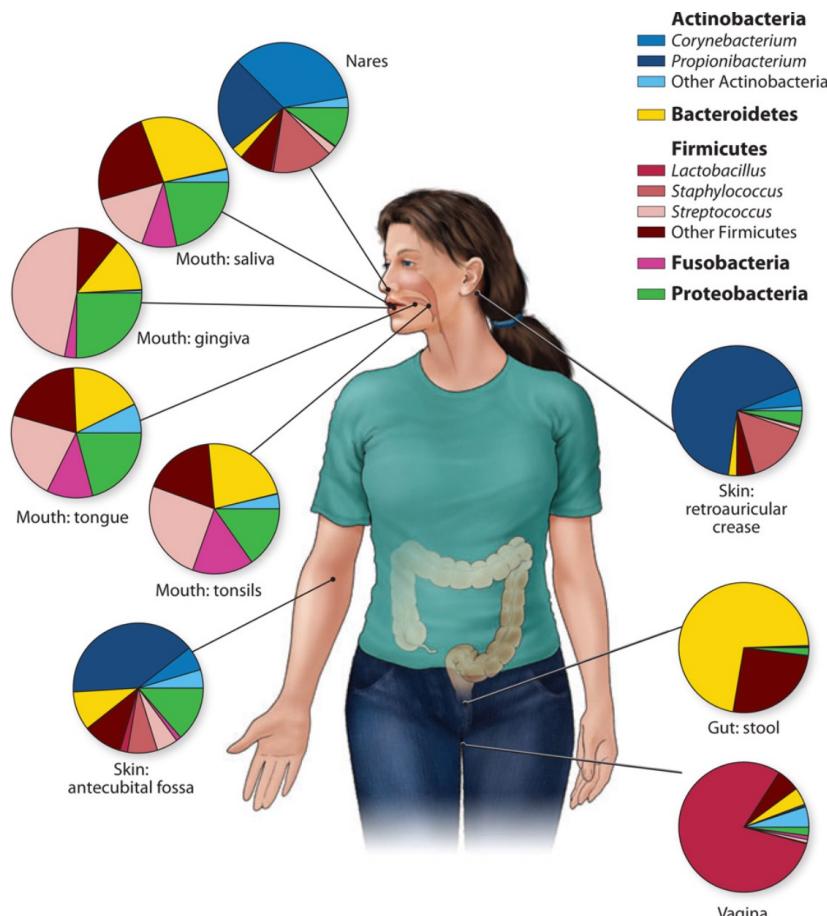
What can we learn/what questions can we ask?

- **Signature Organisms:** Are there specific organisms that are characteristic for, or respond to, specific conditions?



What can we learn/what questions can we ask?

- **Core Community / Niche Specificity:** Is there a stable sub-community characterizing a habitat / microbiome?



Grice, E. A., & Segre, J. A. (2012). The human microbiome: our second genome. *Annual review of genomics and human genetics*, 13, 151-170.

What can we learn/what questions can we ask?

- **Clinical diagnosis:** Identifying the microorganism(s) causing a disease in order to determine appropriate treatments.

TABLE 3

TABLE 3 Performance of heart valve 16S rRNA gene PCR/sequencing versus blood culture

Case classification	No. of samples	Identical findings	No. (%)	
			Organisms not detected by 16S rRNA gene PCR/sequencing	New organisms detected by 16S rRNA gene PCR/sequencing
Noninfective valvular disease	11	10 (91)	1 (9) ^a	0
Infective endocarditis	40	25 (63)	4 (10)	11 (28)
Cured infective endocarditis	3	1 (33)	2 (67)	0

Hong, H. L., Flurin, L., Greenwood-Quaintance, K. E., Wolf, M. J., Pritt, B. S., Norgan, A. P., & Patel, R. (2023). 16S rRNA Gene PCR/Sequencing of Heart Valves for Diagnosis of Infective Endocarditis in Routine Clinical Practice. *Journal of Clinical Microbiology*, 61(8), e00341-23

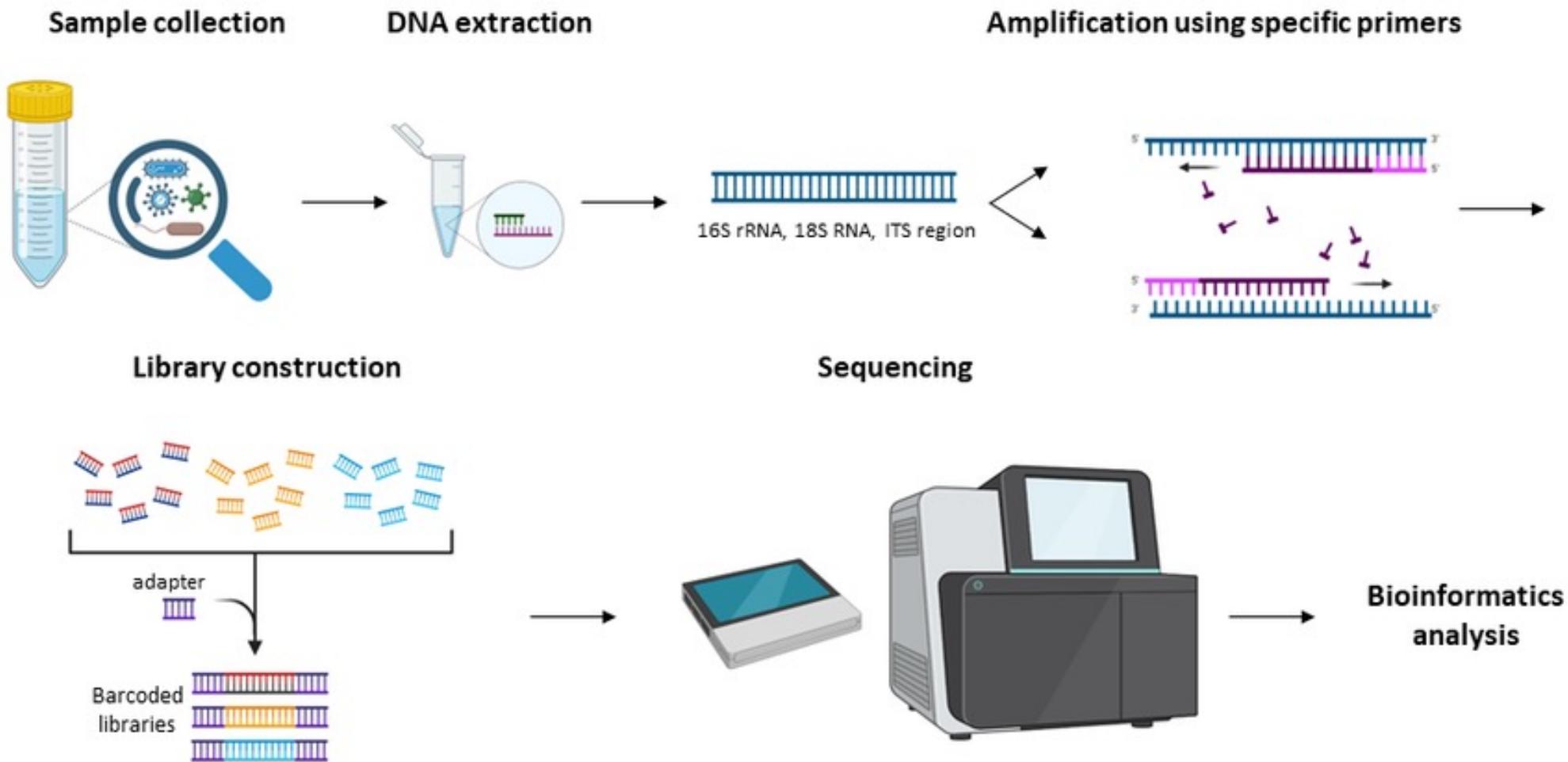
Study design

- Considerations
 - Appropriate biological controls (i.e. healthy control, healthy control with similar diet)
 - Collection of appropriate metadata to rule out confounding factors
 - Diet
 - Medicines
 - Sex
 - Gender
 - Age
 - Weight
 - Date of sample collection
 - Disease history
 - Blood work

Microbiome Data Analysis

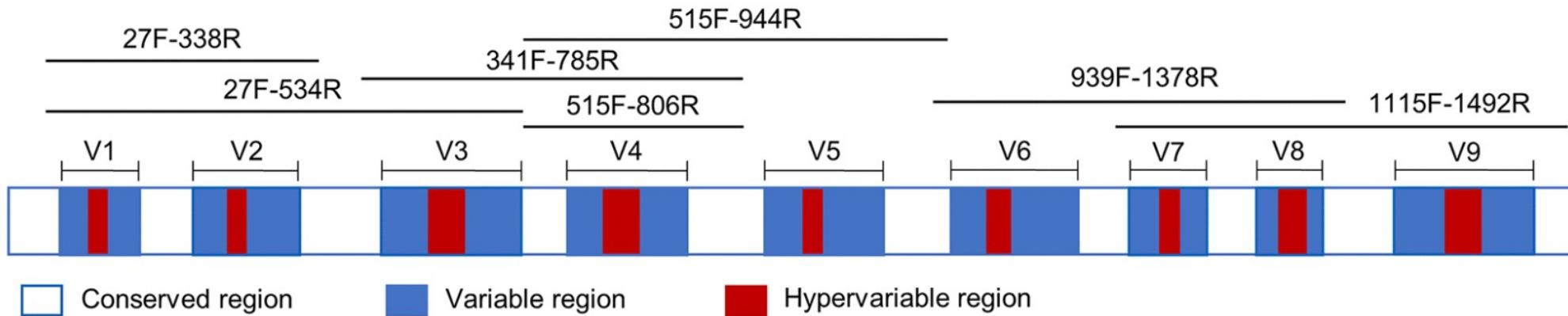
Monday, November 6th, 2:00PM – 5:00PM [EST]

Basic workflow



16S rRNA primers

- Fragment size appropriate for 16S amplicon sequencing will depend on sequencer (Illumina, Nanopore, PacBio)



- V1 – V2/V3, V3-V4/V5, V4-V5 are the most commonly used
- Taxonomy calls can vary depending on which region you are analyzing
- Examples:

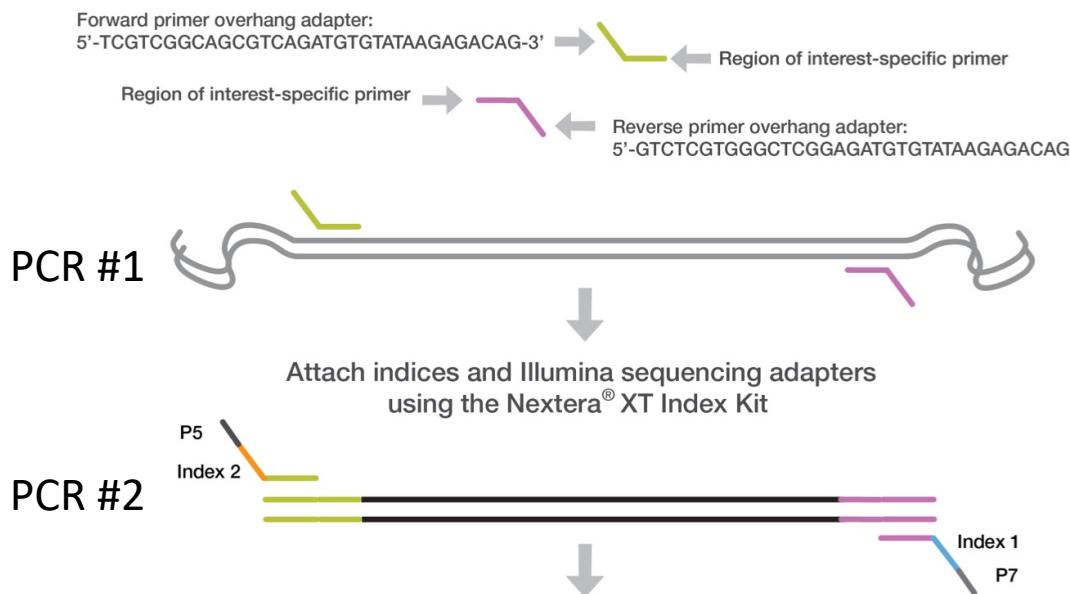
- 515F: GTGYCAGCMGCCGCGGTAA
- 806R: GGACTACNVGGGTWTCTAAT

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., ... & Neuhaus, K. (2021). Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *MspHERE*, 6(1), 10-1128.

Amplification and library preparation

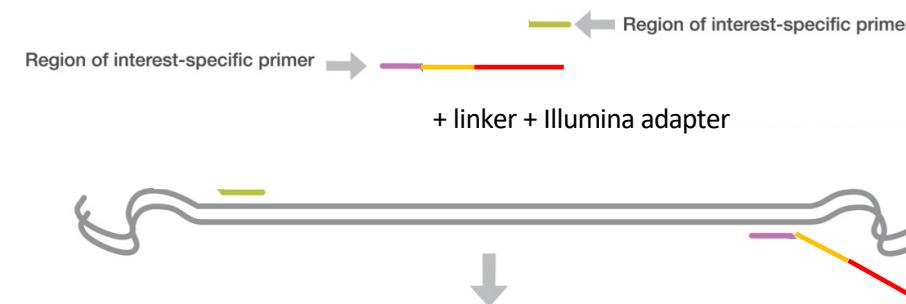
Illumina library prep

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters



Earth Microbiome library prep

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters

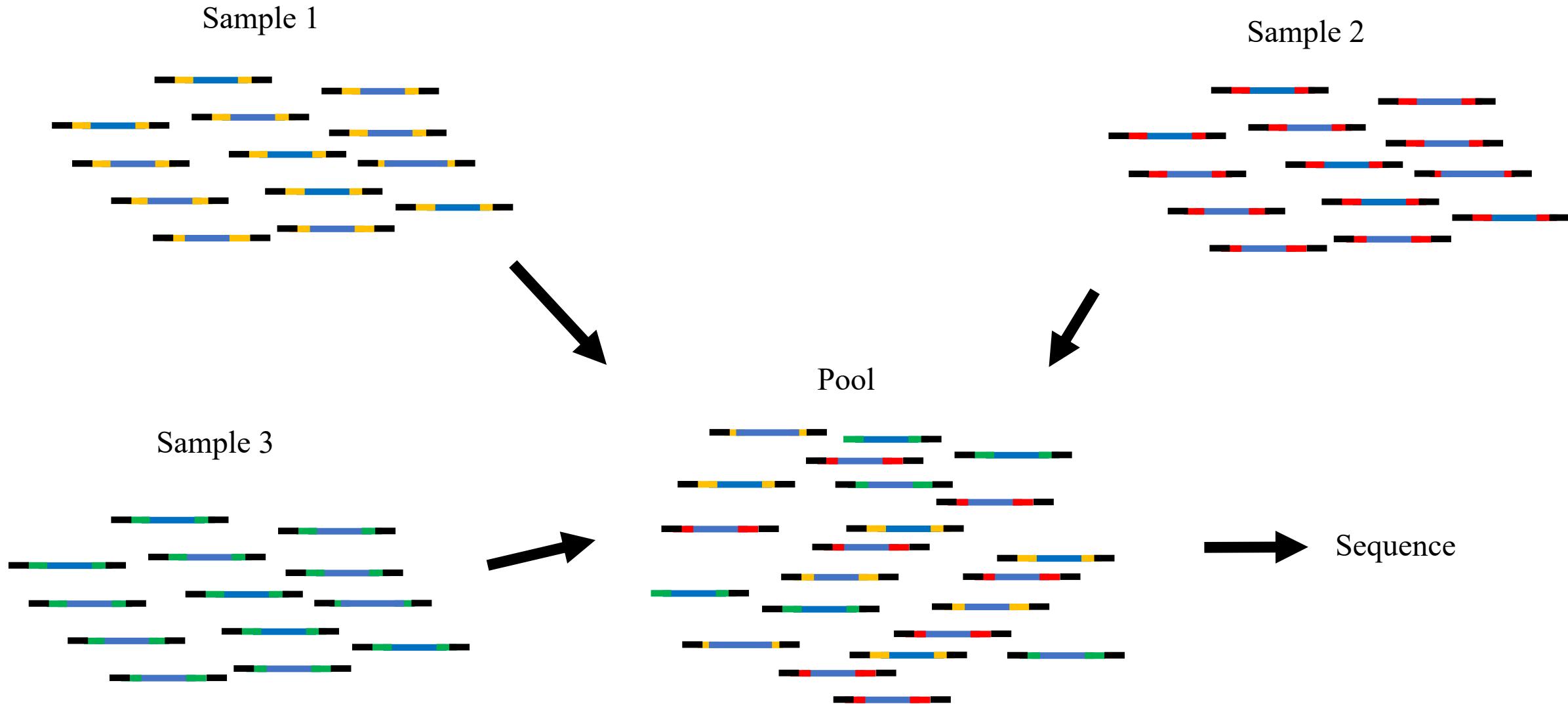


**Option for multiplexing many more samples

PCR #1

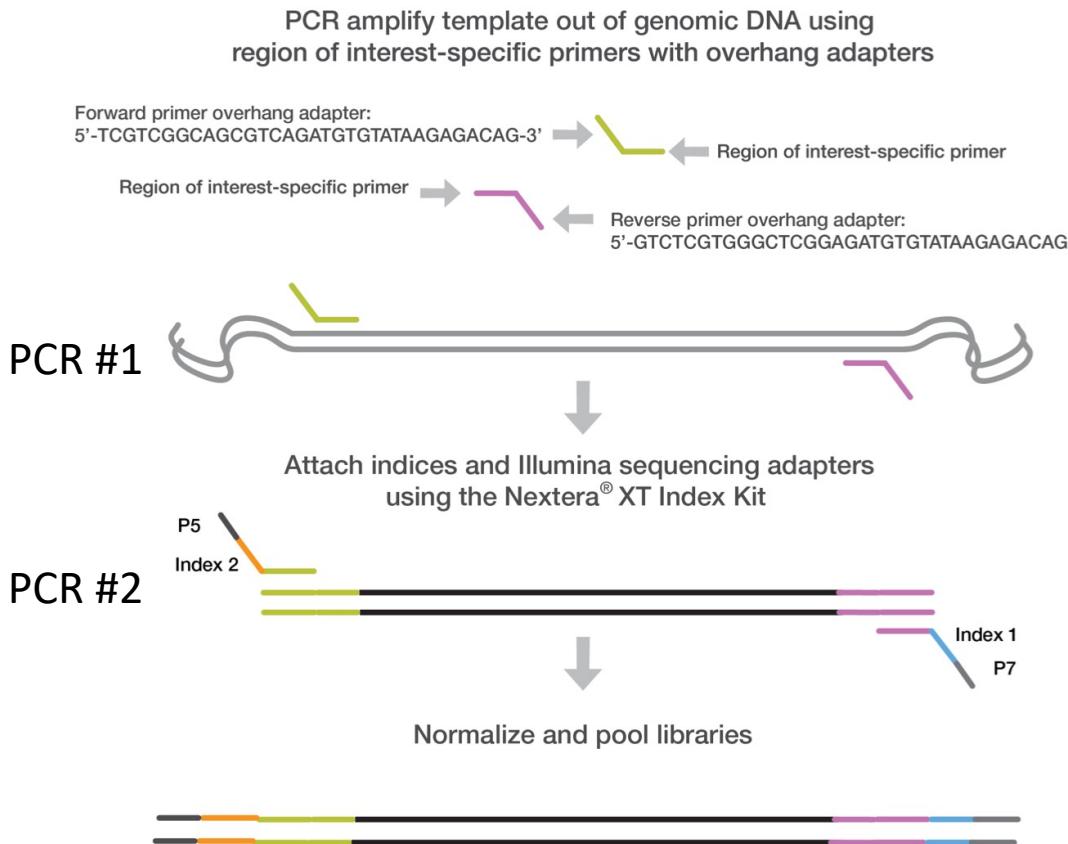
**need custom sequencing index and sequencing primers

Adapter addition and barcoding (adding indices)



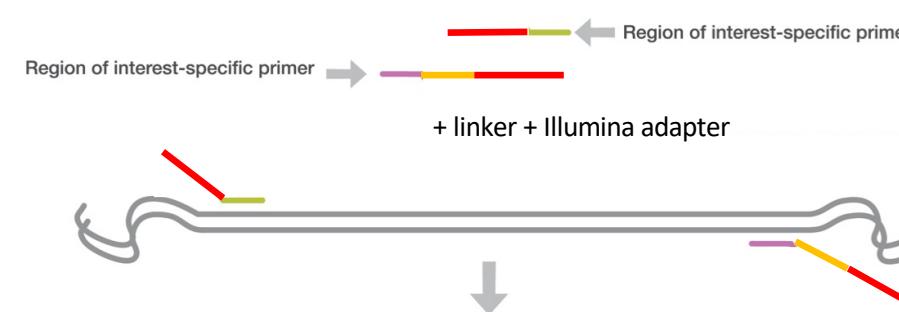
Amplification and library preparation

Illumina library prep



Earth Microbiome Project library prep

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters



**need custom sequencing index and sequencing primers

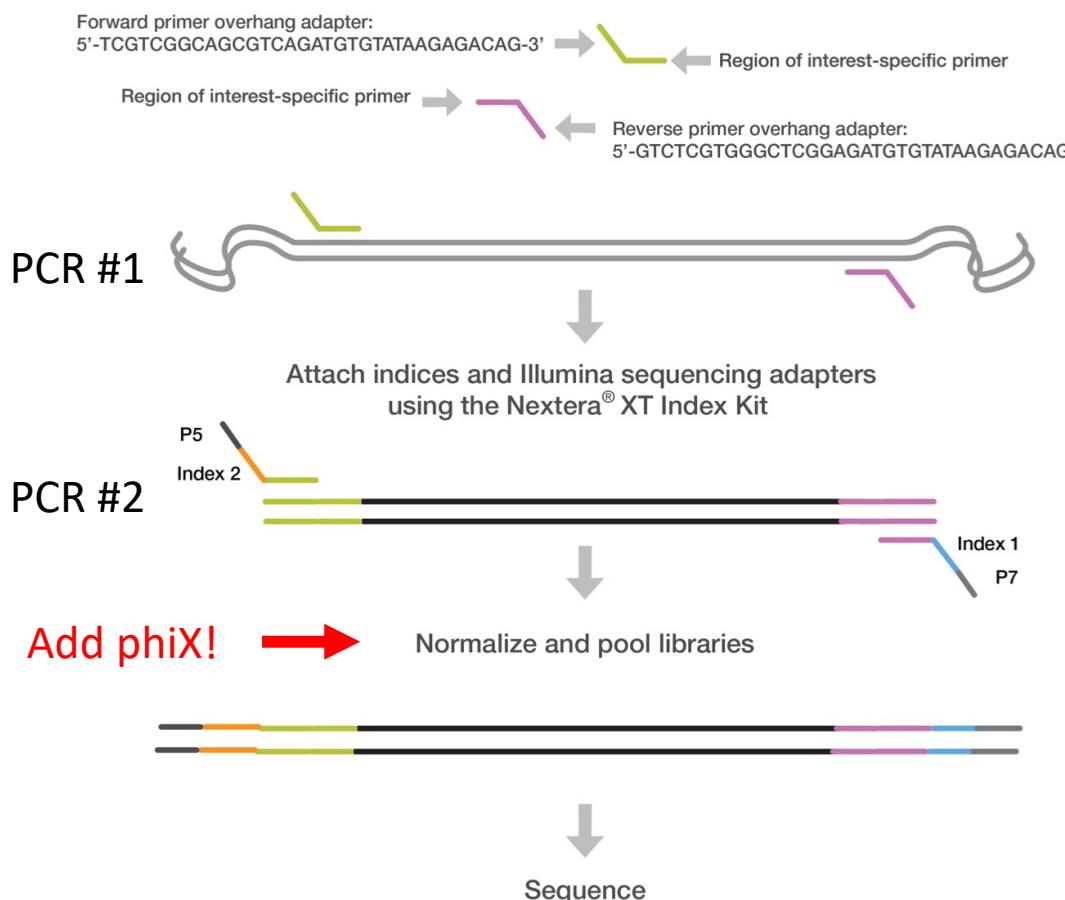
Normalize and pool libraries

**Option for multiplexing many more samples

Amplification and library preparation

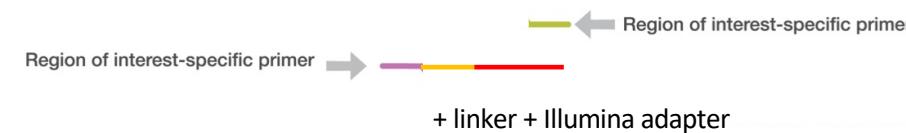
Illumina library prep

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters



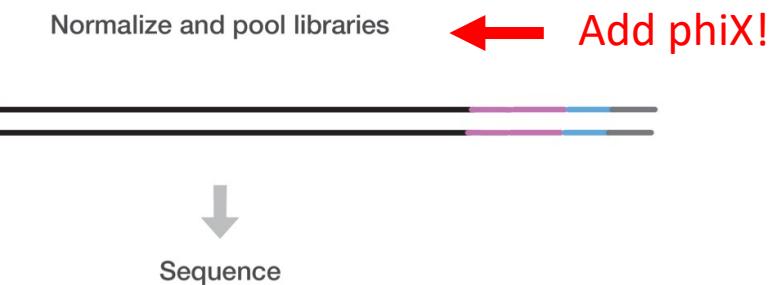
Earth Microbiome library prep

PCR amplify template out of genomic DNA using region of interest-specific primers with overhang adapters



**Option for multiplexing many more samples

**need custom sequencing index and sequencing primers



Sequencing platforms

- Illumina (short read, max 600bp)
 - MiSeq
 - NextSeq
 - NovaSeq
- ThermoFisher (short read, max 600bp)
 - IonTorrent S5
 - Ion Torrent PGM
- Oxford Nanopore Technologies (long read)
 - MinION
 - PromethION
 - GridION
- PacBio (long read and now short read)
 - Sequel II
 - Revio
 - ONSO

Illumina sequencing technology

- Sequencing by synthesis
 1. Clonal amplification
 1. Bridge amplification
 2. Cluster generation
 2. Sequencing
 1. Fluorescently labeled nucleotides added
 2. Base calls made based on emission of fluorescence

Take away:

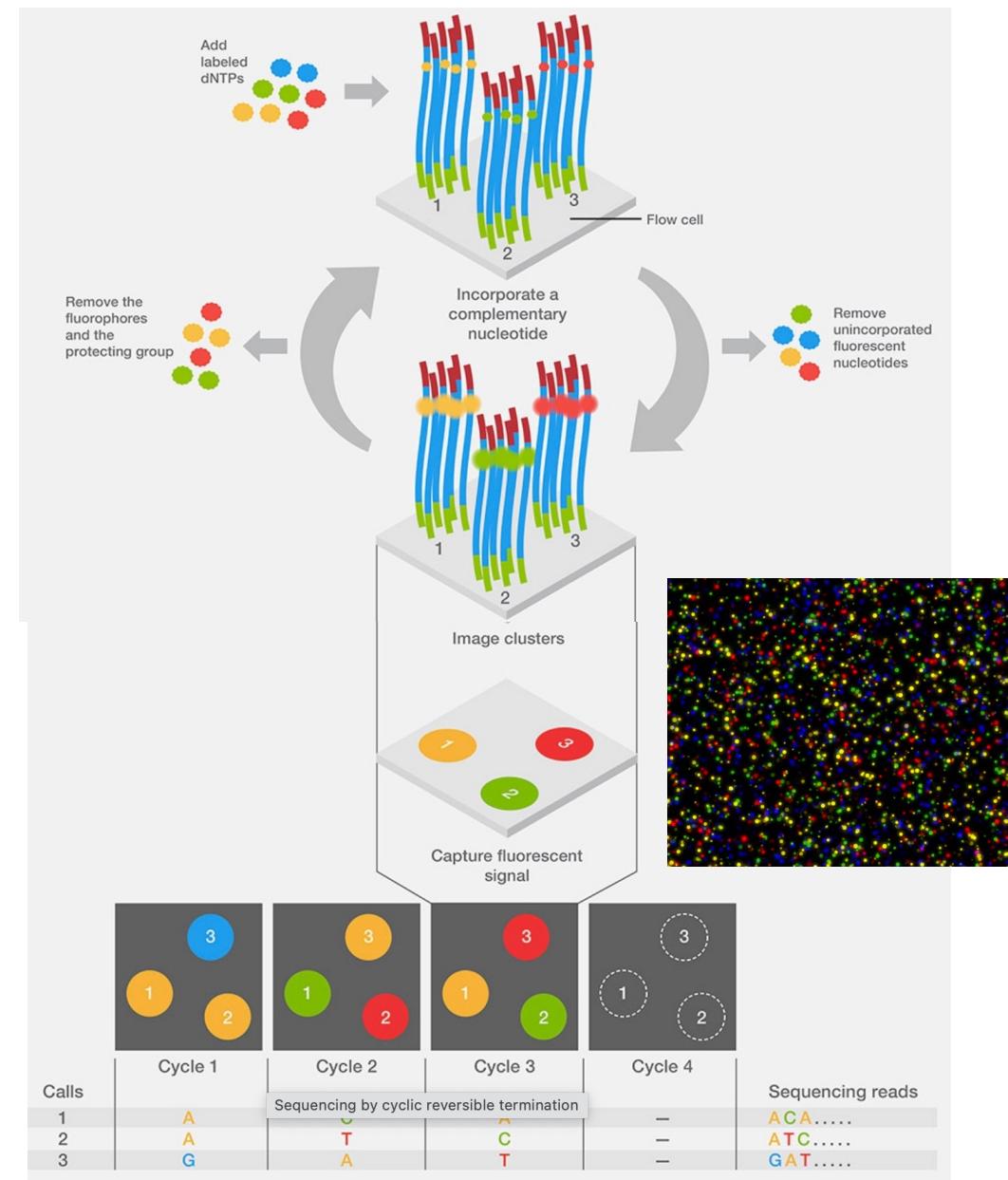
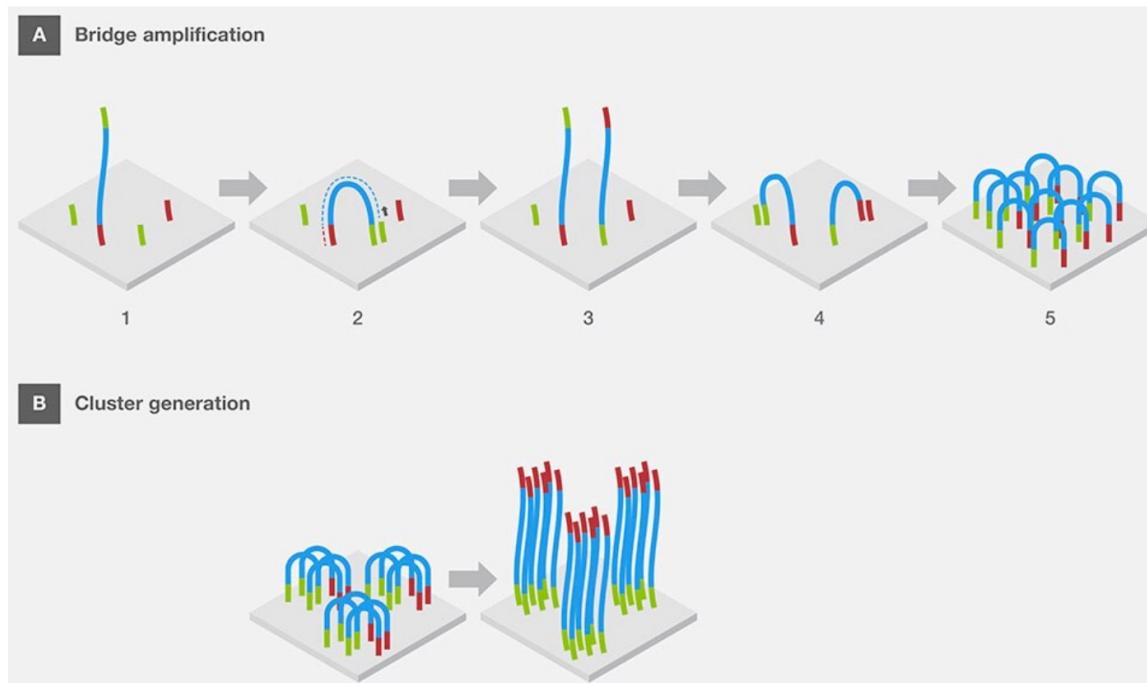
PhiX diversifies the clusters so signal can be read clearly

Barcode will get sequenced and sometimes the adapter can get sequenced in addition to your amplicon

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

https://www.ogc.ox.ac.uk/wp-content/uploads/2017/09/Illumina_Sequencing_Overview_15045845_D.pdf

Illumina sequencing technology



Caveats of amplicon-based explorations

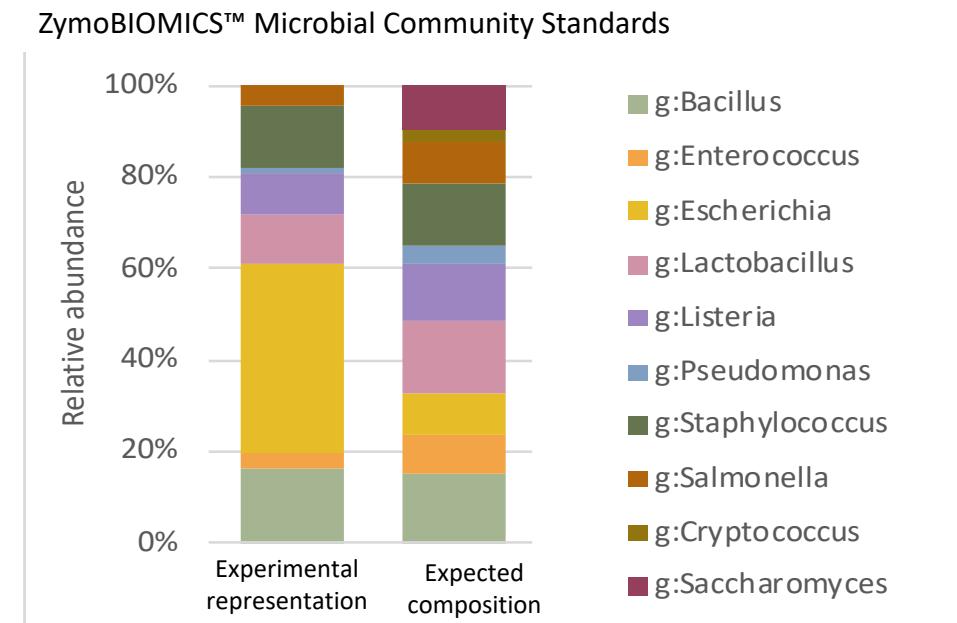
Technical considerations

- Primer bias: different primers have different annealing efficiency to different phylogenetic groups (sequence templates)
- Efficiency bias: the complexity of different sequence templates produce different sequencing rates, introducing inaccurate representation of phylogenetic group distributions
- PCR errors: Errors caused by Taq pol
- Chimera formation: Hybrid products between multiple parent sequences, especially common between closely related organisms

Caveats of amplicon-based explorations

Bias cannot always be avoided in amplification-based studies, however you can constrain with molecular biology techniques:

- Limit PCR cycles
- Degenerate base inclusion in primers
- Use replicate PCR products (debated)
- Avoid long extension times
- Correct for them bioinformatically
- Add controls!
 - negative controls
 - community standards
- Assess, acknowledge & report!



Assessment of PCR bias:
control samples holding microbial community standards, in order to assess the PCR bias of each experiment

Caveats of amplicon-based explorations

Biological considerations

- The 16S rRNA gene is *not* a single-copy gene
 - Gene number can vary from species to species
 - Gene sequence can vary between copies within the same organism
- The full capacity of the gene to distinguish lineages, cannot be captured by the sequence of any fraction of its variable regions (full gene seq is needed)
- Different variable regions have different capacity to differentiate between different lineages
- Inferring *true* phylogenetic relations from a single gene can be risky!
 - Even full length 16S gene cannot absolutely resolve the diversification of closely related organisms (species or strains)
- Only tells you who is there – will not provide functional information
 - There are tools to “infer” function based on taxonomy