Microbiome 101

An Introduction to Microbiome Studies and Bioinformatics Tools

MEDINI ANNAVAJHALA, PHD

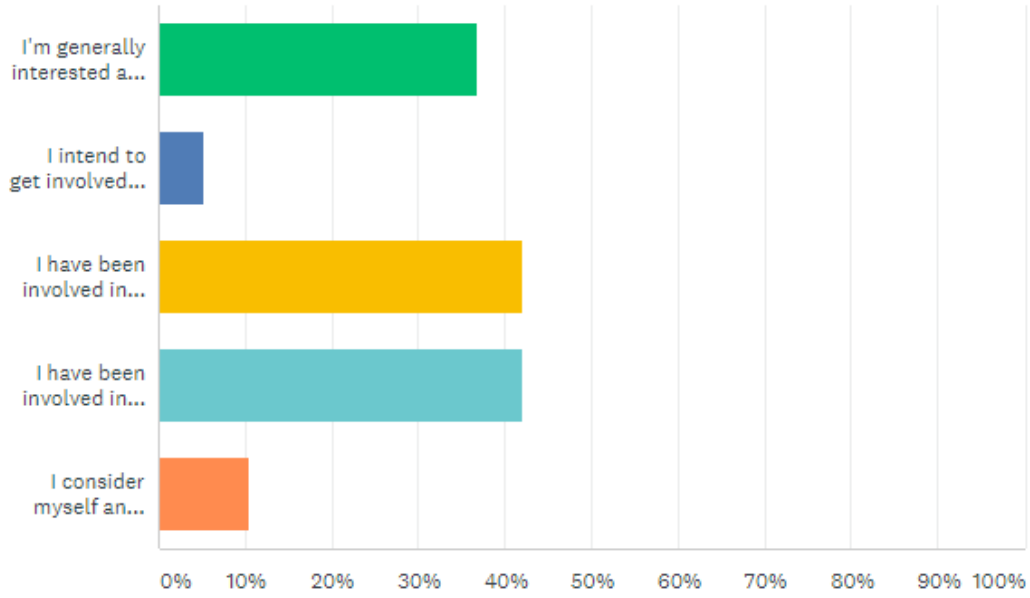
UHLEMANN LABORATORY, DEPT. OF MED. MICROBIOME CORE LABORATORY

CUMC, DIVISION OF INFECTIOUS DISEASES

MICROBIOME WORKING GROUP SEMINAR SERIES, FALL 2017

What is your experience level with microbiome research?

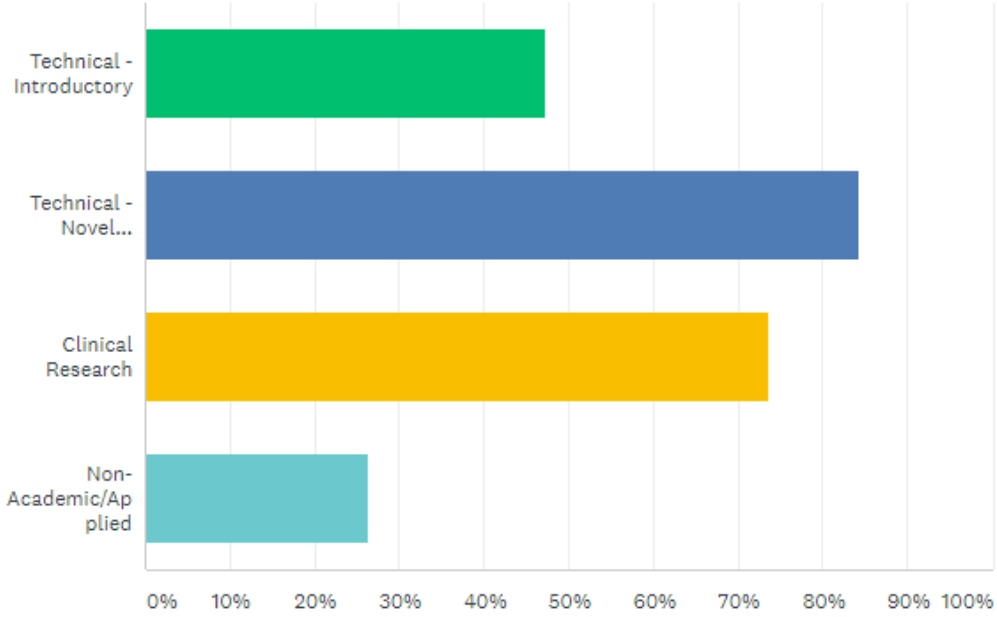
Answered: 19 Skipped: 0



ANSWER CHOICES	RESPONSES	
I'm generally interested and want to learn more about microbiome research	36.84%	7
I intend to get involved in microbiome research in the future but have no experience yet	5.26%	1
I have been involved in some microbiome research projects but want to learn more about techniques/analysis	42.11%	8
I have been involved in multiple microbiome research projects and have basic understanding of microbiome data analysis	42.11%	8
I consider myself an expert on some aspect of the microbiome	10.53%	2
Total Respondents: 19		

What types of seminars are you most interested in attending?

Answered: 19 Skipped: 0



ANSWER CHOICES	RESPONSES	
Technical - Introductory	47.37%	9
Technical - Novel Techniques/Analyses	84.21%	16
Clinical Research	73.68%	14
Non-Academic/Applied	26.32%	5
Total Respondents: 19		

Comments (1)

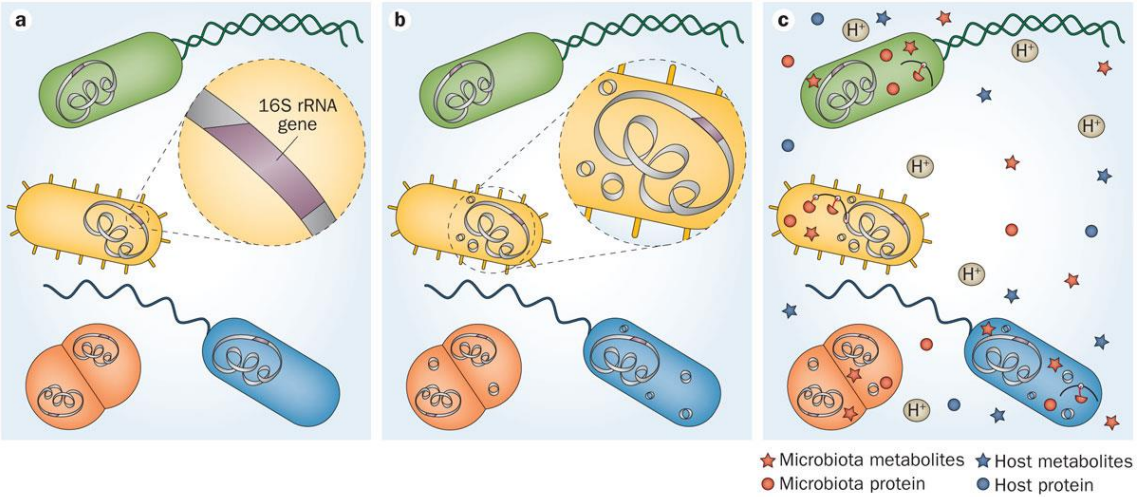
Overview

Microbiome – components, host interactions, and dysbiosis

NGS and Bioinformatics Tools

Microbiome Study Design

Terminology



Microbiota
16S rRNA
Taxonomic identification

Metagenome
Genes and genomes of microbiota

Microbiome
Genes, genomes, products, host proteins

Whiteside, S. A., et al. (2015). "The microbiome of the urinary tract– a role beyond infection."
Nat Rev Urol **12**(2): 81-90.

NIH Human Microbiome Project

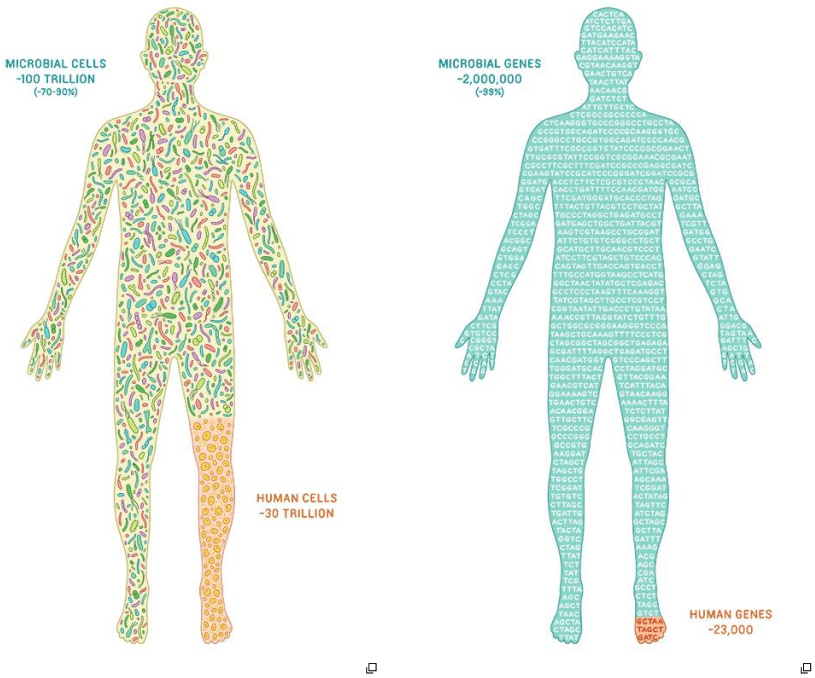
Phase I (2008 – 2012)

- cataloging microbes in / on human body
- 242 healthy American adults (18 – 44 years old)
- ~ 10,000 bacterial species

Phase II (2013 – 2015)

- Biological properties of both microbiome and host
- microbial composition, gene expression, proteins and metabolities
- longitudinal cohort studies

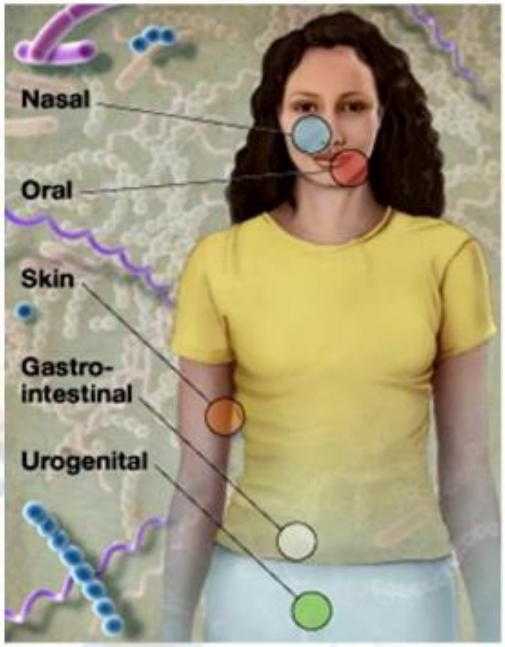
Microbial genomes may code for 100x as many genes as human genome

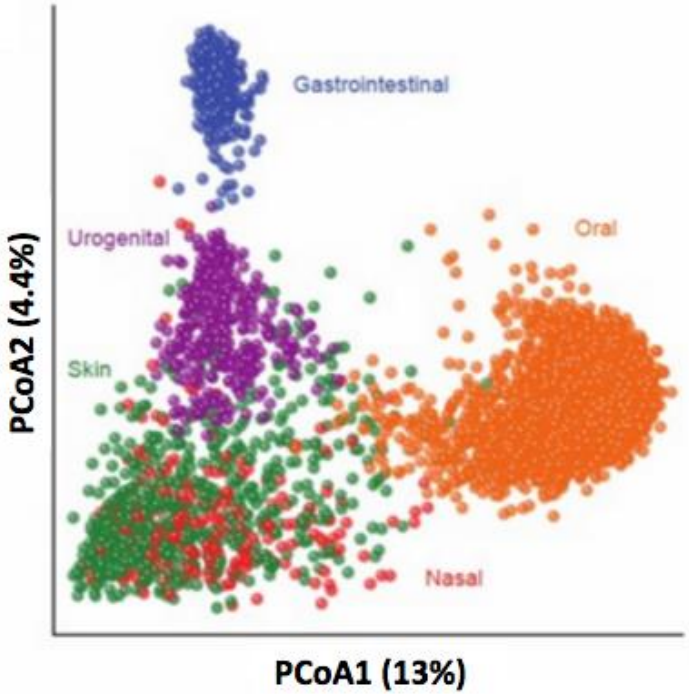


~2,000,000 bacterial genes
~23,000 human genes

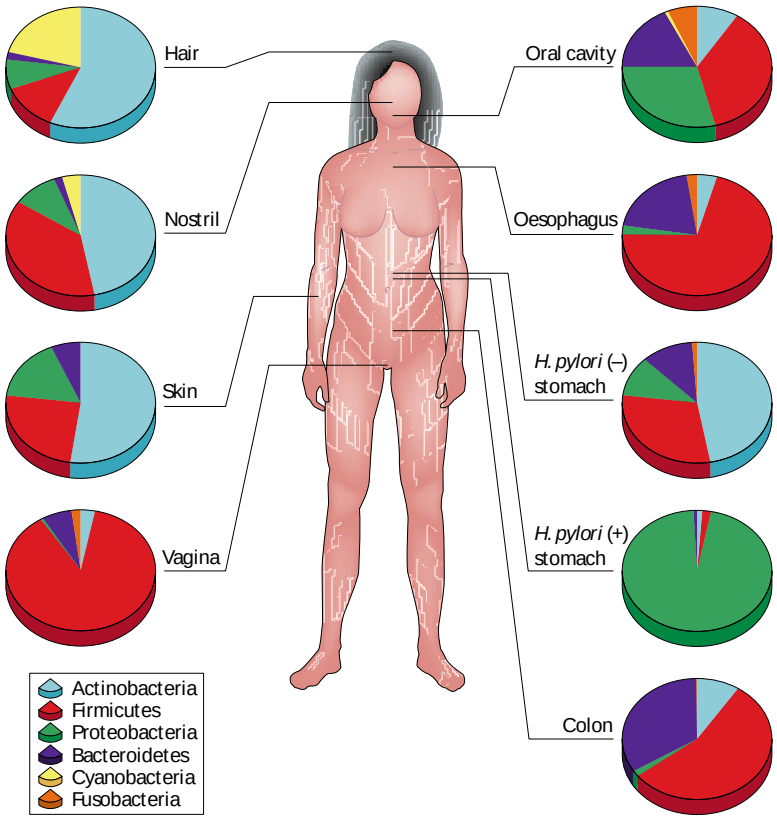
Many bacterial species previously not recognized because unculturable with current methods

Healthy adults: Unique microbial community composition in each body part





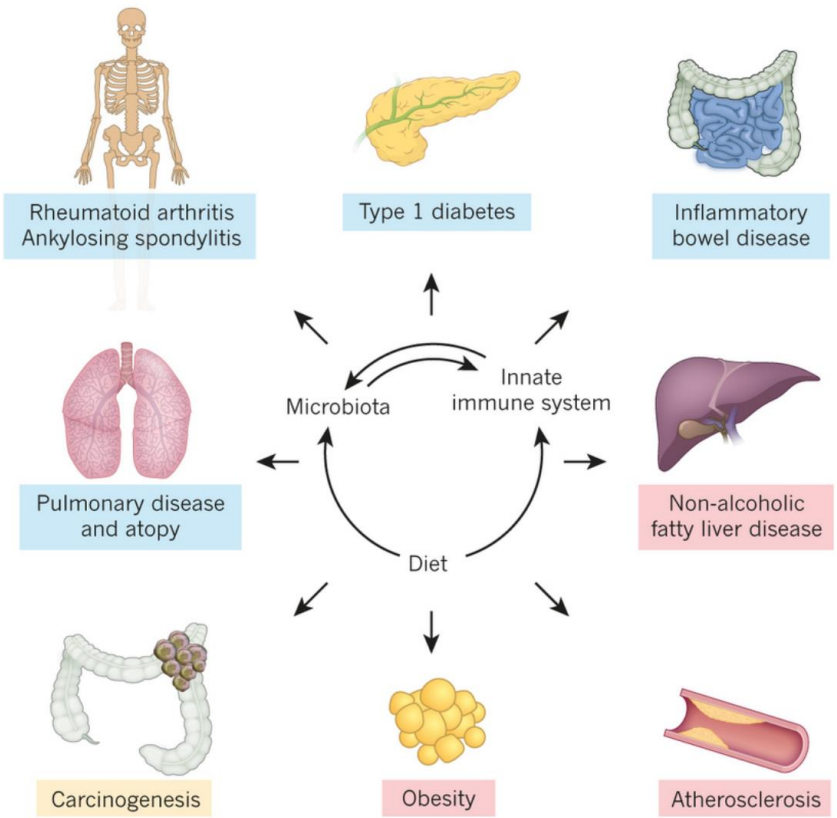
Relative abundance of taxa at the phylum level by anatomical site



- Substantial within individual variation across anatomical site
- Presence of pathogens associated with ecology
- Variations between individuals

Cho, I. and M. J. Blaser (2012). "The human microbiome: at the interface of health and disease." Nat Rev Genet **13**(4): 260-270.

Microbiome – innate immune system interactions



Diseases:
Inflammatory
Metabolic
Neoplastic

Thaiss et al. Nature 2016

Dysbiosis

Microbial imbalance

Changes in quantity and quality

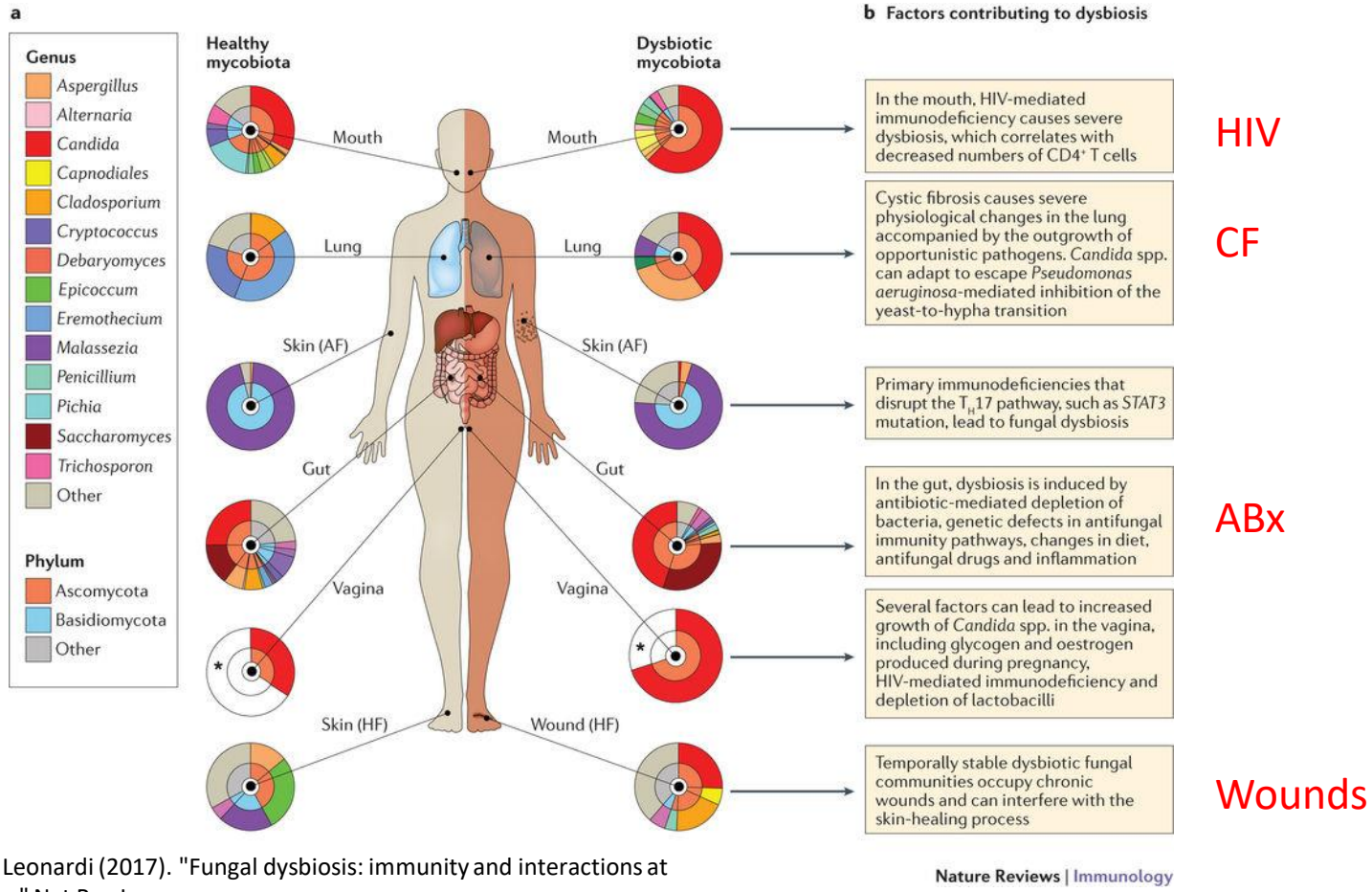
Causes:

- antibiotics
- lifestyle, stress
- age
- genetic predisposition

Consequences:

- immune stimulation

Fungal Dysbiosis



HIV

CF

ABx

Wounds

Iliev, I. D. and I. Leonardi (2017). "Fungal dysbiosis: immunity and interactions at mucosal barriers." Nat Rev Immuno

What do we want to ask?

Microbiome/Dysbiosis

- Presence/Absence
- Differential abundance
- Diversity
- Functional capabilities
- Activity

Disease/Outcomes

Overview

Microbiome – components, host interactions, and dysbiosis

NGS and Bioinformatics Tools

Microbiome Study Design

What is my target?

A lot of data describing many organisms in multiple samples

16S rRNA (Bacteria)

ITS/ 18S rRNA (Fungi)

Shotgun metagenome (Mixed community)

RNA-Seq

Other Ampli-Seq

16s rRNA Sequencing

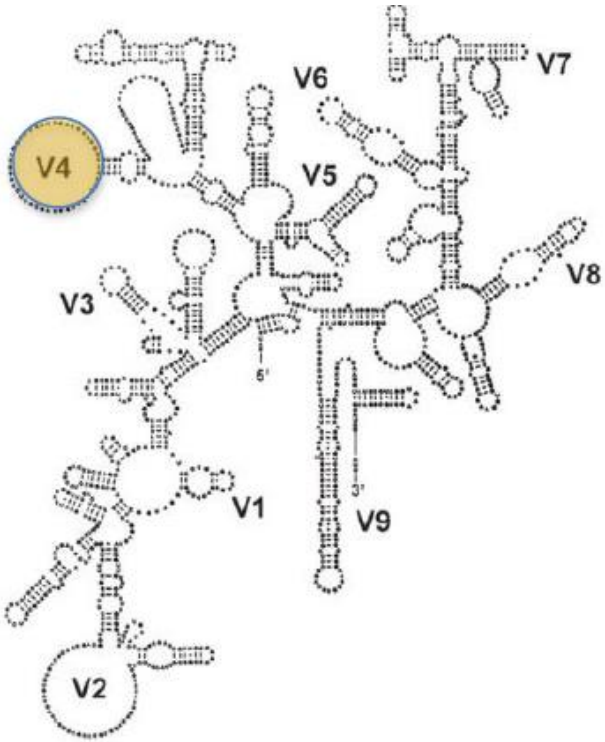
16S rRNA gene present in all bacterial species

Highly conserved and variable sequences

Variable = “molecular fingerprint”

Amplification with degenerate primers targeting conserved regions

Large public database for comparisons



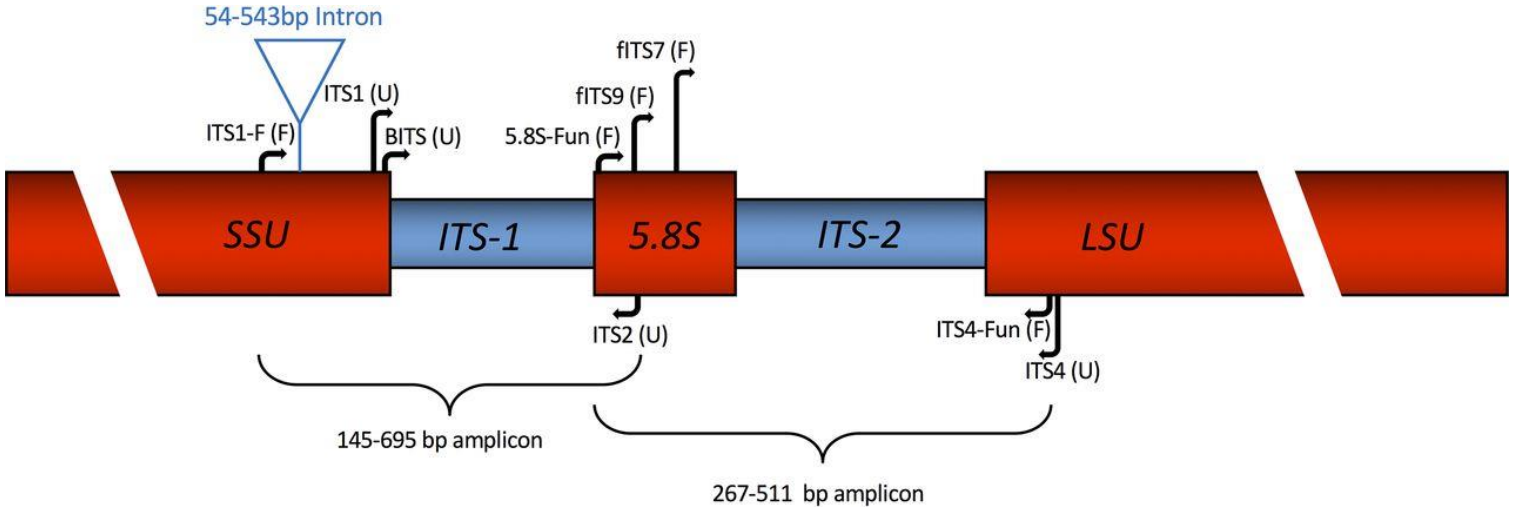
18S rRNA/ITS Sequencing

18S rRNA eukaryotic gene homologous to 16S rRNA in bacteria

For fungi, variability in 18S rRNA may not be sufficient to classify species

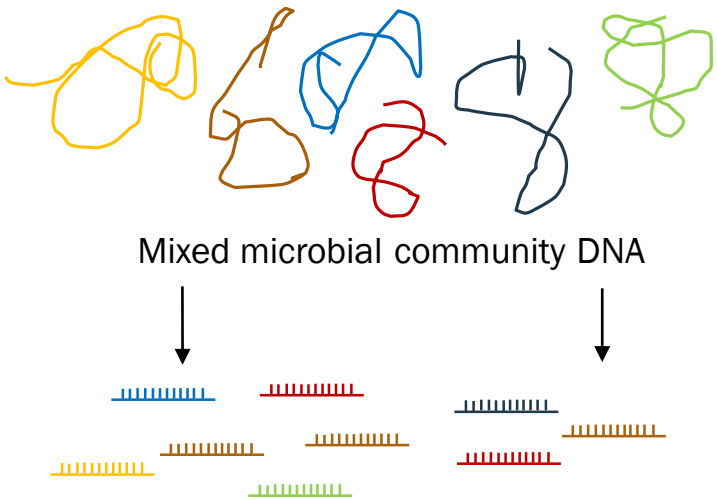
ITS = Intergenic Transcribed Spacer → 2 ITS regions in each eukaryotic cistron

- Higher resolution for fungi, but longer reads generated



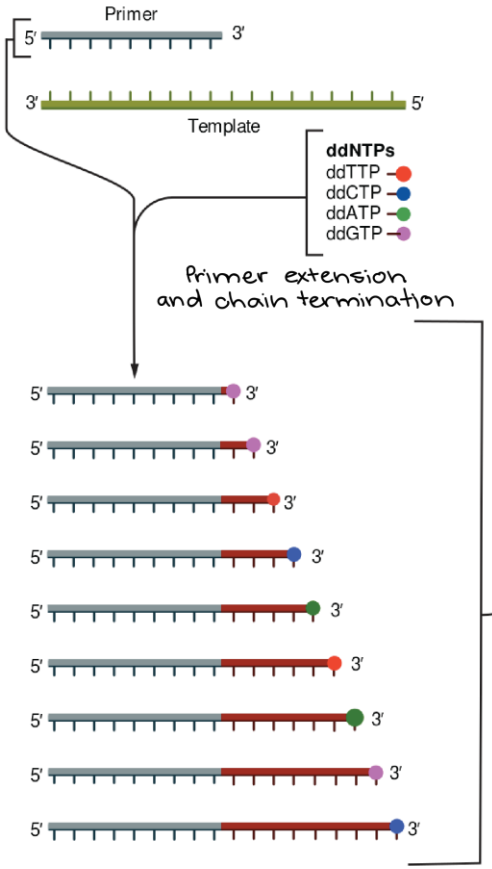
Taylor, D. L., et al. (2016). "Accurate Estimation of Fungal Diversity and Abundance through Improved Lineage-Specific Primers Optimized for Illumina Amplicon Sequencing." Applied and Environmental Microbiology **82**(24): 7217-7226.

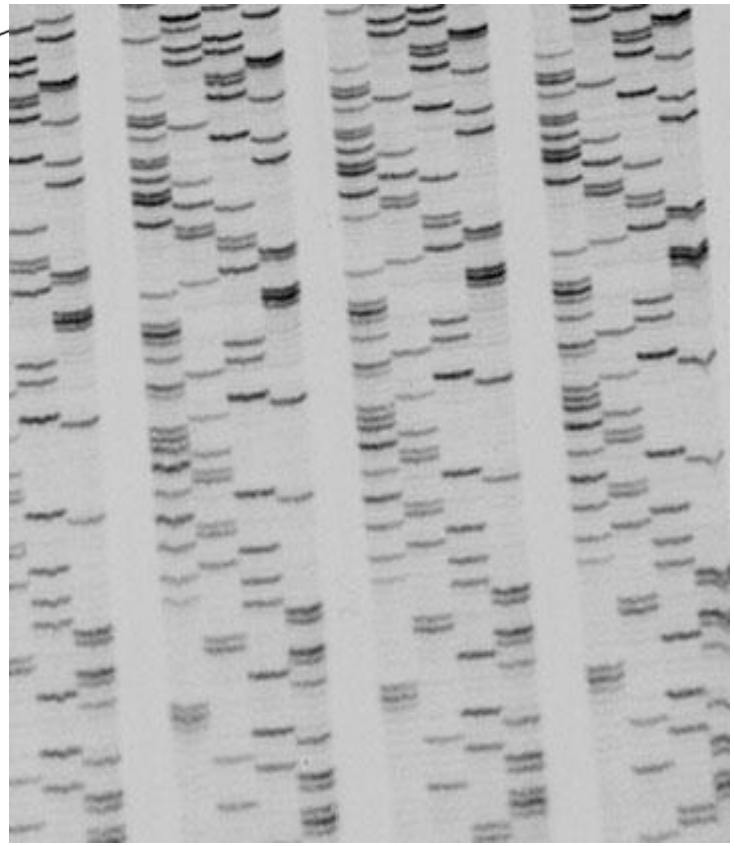
“Shotgun” Metagenomics



- What genes are present (from which organisms– maybe) → Functional potential of the system

Traditional Sanger Sequencing





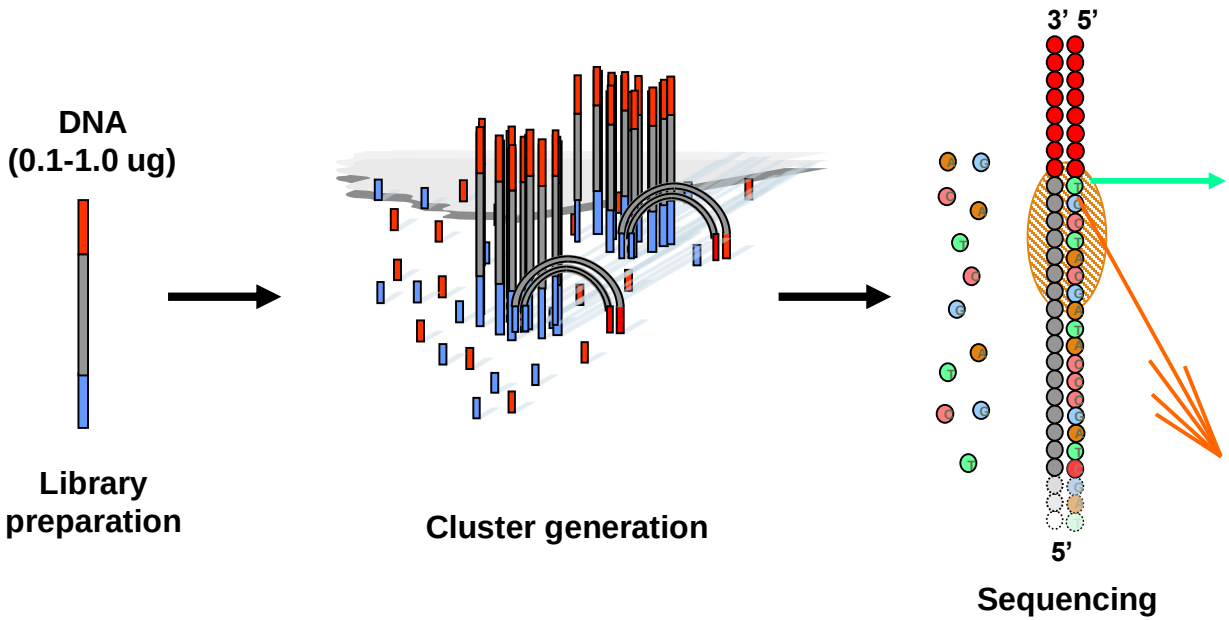
https://www.khanacademy.org/science/biology/biotech-dna-technology/dna-sequencing-pcr-electrophoresis/a/dna-sequencing
https://unlockinglifescode.org/timeline/11

Illumina MiSeq - SBS

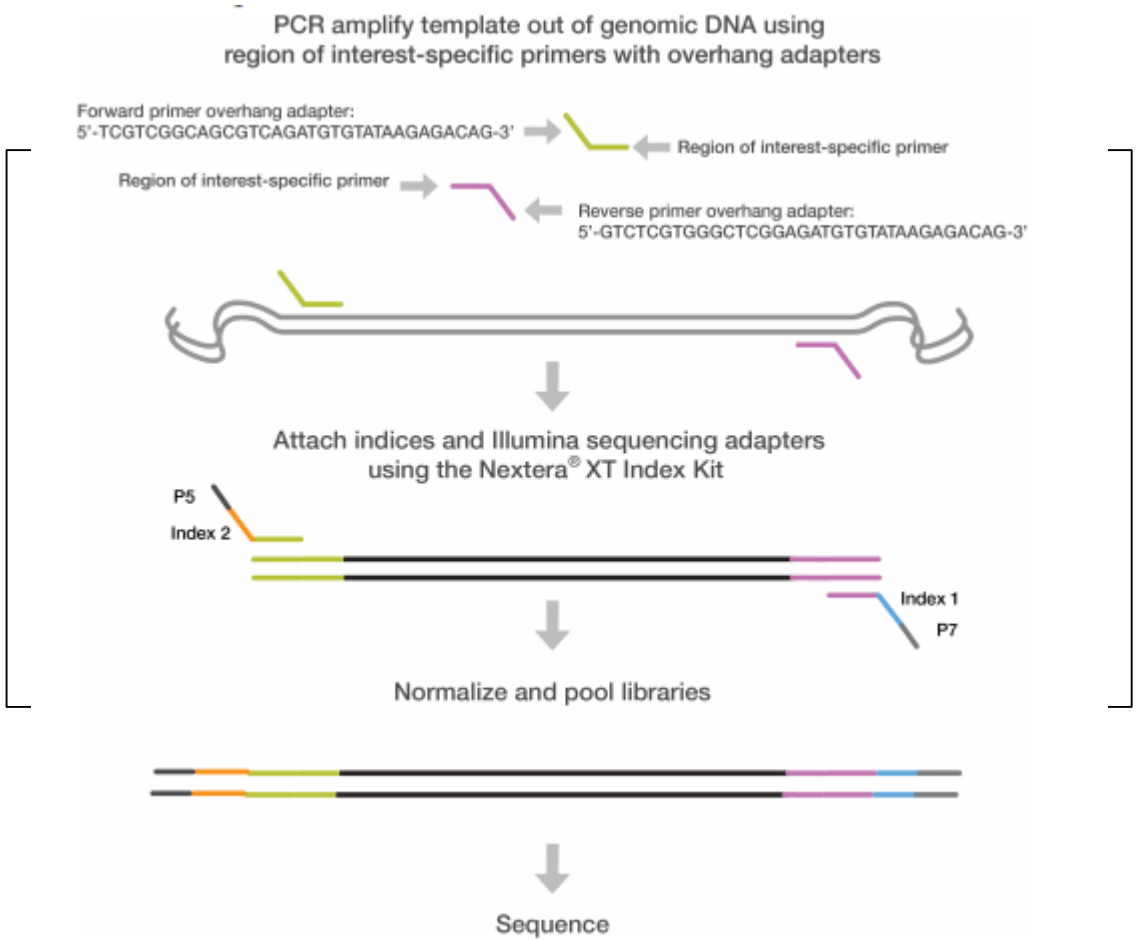
SBS = Sequencing By Synthesis

Multiplexing samples into one run using indexes

Multiple reads per base pair

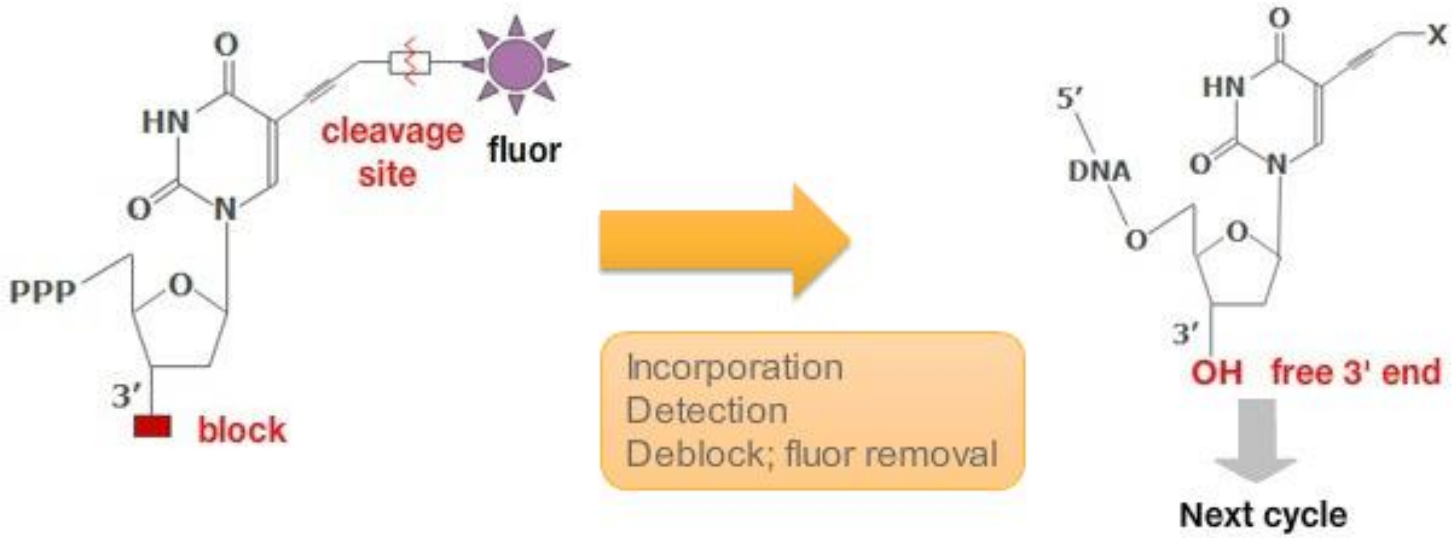


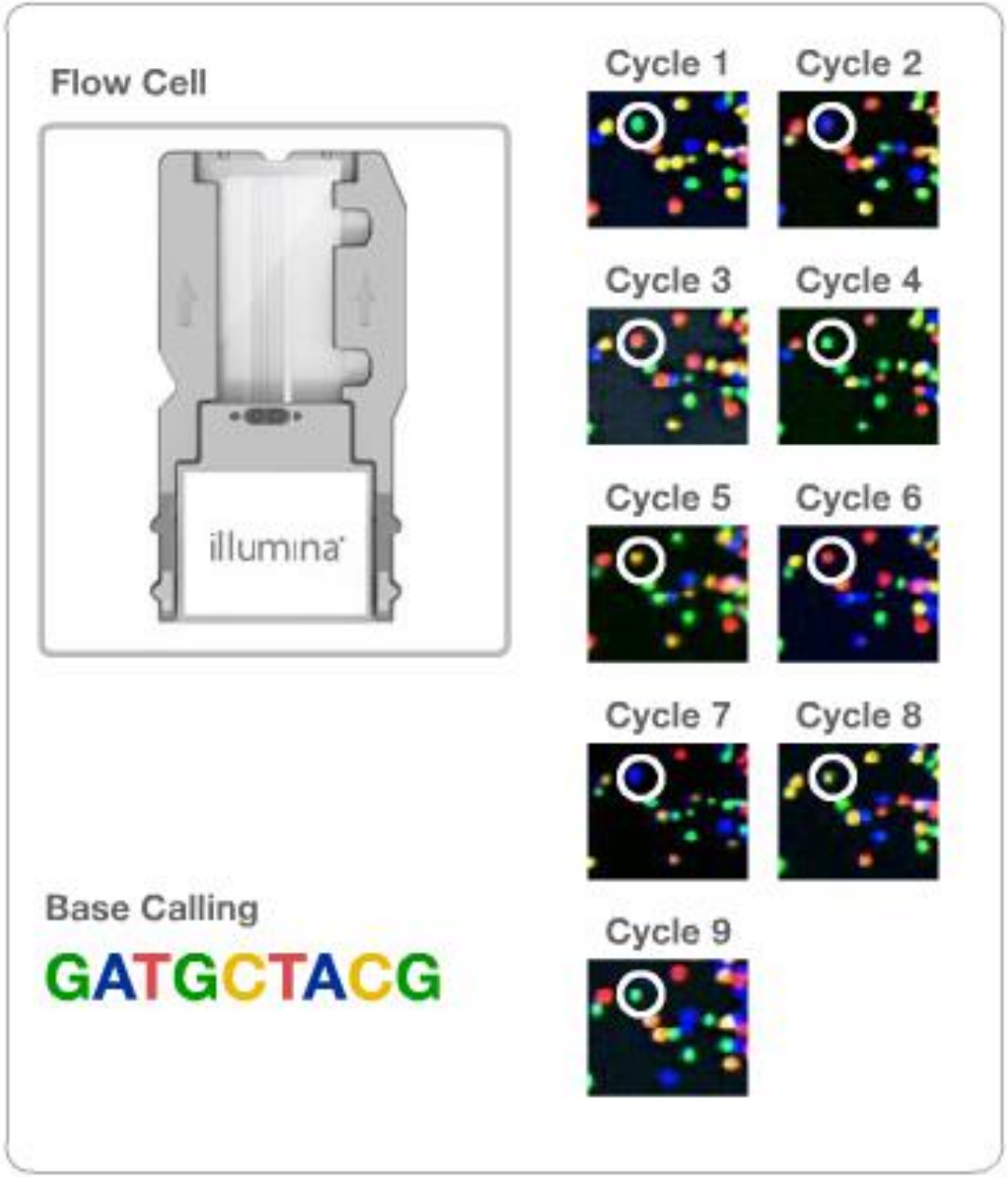
Adapter Ligation



Cluster Generation on Flow Cell

Sequencing


O
HN
O
N
PPP
O
3′
block
cleavage site
fluor
Incorporation
Detection
Deblock; fluor removal
O
X
HN
5′
DNA
O
N
O
O
3′
OH
free 3' end
Next cycle


Flow Cell
illumina
Cycle 1
Cycle 2
Cycle 3
Cycle 4
Cycle 5
Cycle 6
Cycle 7
Cycle 8
Cycle 9
Base Calling
GATGCTACG

Sequencing Output

```
AAACTGGAAAAAATGATCGGCGATCTGCTGTAATCATTCTTAGCGTGACCGGGAAGTCGGTCACGCTACCTCT
TCTGAAGCCTGTCTGTCACTCCCTTCGCAGTGTATCATTCTGTTTAACGAGACTGTTTAAACGGAAAAATCTT
GATGAATACTTTACGTATTGGCTTAGTTTCCATCTCTGATCGCGCATCCAGCGGCGTTTATCAGGATAAAGGC
ATCCCTGCGCTGGAAGAATGGCTGACATCGGCGCTAACCACGCCGTTTGAACTGGAAACCCGCTTAATCCCCG
ATGAGCAGGCGATCATCGAGCAAACGTTGTGTGAGCTGGTGGATGAAATGAGTTGCCATCTGGTGCTCACCAC
GGGCGGAACTGGCCCGGCGCGTCGTGACGTAACGCCCGATGCGACGCTGGCAGTAGCGGACCGCGAGATGCCT
GGCTTTGGTGAACAGATGCGCCAGATCAGCCTGCATTTTGTACCAACTGCGATCCTTTCGCGTCAGGTGGGCG
TGATTCGCAAACAGGCGCTGATCCTTAACTTACCCGGTCAGCCGAAGTCGATTAAAGAGACGCTGGAAGGCGT
GAAGGACGCTGCGGGTAACGTTGTGGTGCATGGTATTTTTGCCAGCGTACCGTACTGCATTCAGTTGCTGGAA
GGGCCATACATTGAAACGGCACCGGAAGTGGTTGCAGCATTCAGACCGAAGAGTGCAAGACGCGACGTTAGCG
AATAAAAAAATCCCCCCGAGCGGGGGGATCTCAAAACAATTAGTGGGATTCACCAATCGGCAGAACGGTGCGA
CCAAACTGCTCGTTCAGTACTTCACCCATCGCCAGATAGATTGCGCTGGCACCGCAGATCAGACCAATCCAGC
CGGCAAAGTGGATGATTGCGGCGTTACCGGCAATGTTACCGATCGCCAGCAGGGCAAACAGCACGGTCAGGCT
AAAGAAAACGAATTGCAGAACGCGTGCGCCTTTCAGCGTGCCGAAGAACATAAACAGCGTAAATACGCCCCAC
AGACCCAGGTAGACACCAAGGAACTGTGCATTTGGCGCATCGGTCAGACCCAGTTTCGGCATCAGCAGAATCG
CAACCAGCGTCAGCCAGAAAGAACCGTAAGAGGTGAATGCGGTTAAACCGAAAGTGTTGCCTTTTTTGTACTC
CAGCAGACCAGCAAAAATTTGCGCGATGCCGCCGTAGAAAATGCCCATGGCAAGAATAATACCGTCCAGAGCA
AAATAACCCACGTTGTGCAGGTTAAGCAGAATGGTGGTCATGCCGAAGCCCATCAGGCCCAGCGGTGCCGGAT
TAGCCAACTTAGTGTTGCCCATAATTCCTCAAAAATCATCATCGAATGAATGGTGAAATAATTTCCCTGAATA
ACTGTAGTGTTTTCAGGGCGCGGCATAATAATCAGCCAGTGGGGCAGTGTCTACGATCTTTTGAGGGGAAAAT
GAAAATTTTCCCCGGTTTCCGGTATCAGACCTGAGTGGCACTAACCATCCGGCGCAGGCAGGCGATTTGCAGT
ACGGCTGGAATTGTCACGCGATAGGCAATGCCGCTGACCGCTTTAACCCCATTTAGTGCCGCGCCTACAGGGC
CTCCCAGACCCGCGCCGCGCAGCAAACCATGCCCAAGTACGCTCATTGCTGCGTGGGTGCGTAAAATGCGGGT
CAATTGGCTGGAAAGCAAATGCGACACGCCTTTTGCCAATAATTTGTCTTTCATCAGCAGCGGCAGCAGCTCT
TCCAGCTCATTCACCCTGGCATCGACCGCGTGCAGAAACTCCTGCTTATGTTCCTCGTCCATTTTCTTCCAGG
TGTTACGCAGAAATTGTTCCAGTAACTGTTGCTCAATCTCAAACGTAGACATCTCTTTGTCGGCTTTCAGCTT
CAATCGCTTTGAAACATCGAGCAAAATGGCCCGATACAATTTACCGTGTCCACGCAGTTTGTTGGCGATACTA
TCGCCACCAAAATGCTGTAATTCTCCGGCAATCAGCTGCCAGTTGCGGCGATGTTGCTCGGGATGTCCCTCCA
TCGATTTAAACAGTTCGTTGCGCATCAGTACGCTGGAGAGGCGAGTTTTGCCTTTTTCATTATGGGTGAGCAA
TCGGGCGAAATTTGCCAATTGTTCCTCACTACAATGCTGGAGAAAATCCAGATCTGAATCATTCAGGTAATTA
ACATTCATTTTTTGTGGCTTCTATATTCTGGCGTTAGTCGTCGCCGATAATTTTCAGCGTGGCCATATCCGAT
```

Quality Control

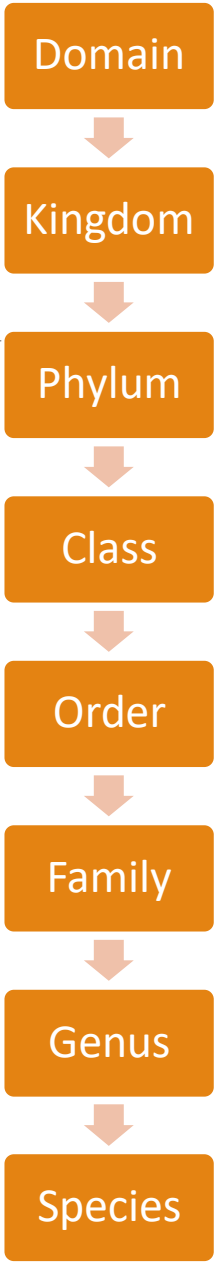
- Pair-ended read generation reduces sequence error rates
- But, low quality reads can still be present
- Quality (Phred Score, Q Score) = Function of confidence in base calling

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy	Q-score
10	1 in 10	90%	Q10
20	1 in 100	99%	Q20
30	1 in 1000	99.9%	Q30
40	1 in 10000	99.99%	Q40

Alignment – the key step

Sequencing projects are limited by the available reference databases

- Sequences clustered by similarity into Operational Taxonomic Units (OTUs) (97%,99%)
- Representative sequence from each cluster is aligned to a reference database (e.g. Greengenes, Silva, UNITE)
- Challenges: multiple matches, no matches (new OTU)
- Some species may share >97% similarity, no resolution at species level
- Taxonomic databases (16S/18S/ITS studies) vs. Functional databases (NCBI RefSeq, nr/nt)



Overview

Microbiome – components, host interactions, and dysbiosis

NGS and Bioinformatics Tools

Microbiome Study Design

Microbiome Study Design

Microbiome/Dysbiosis

- Presence/Absence
- Differential abundance
- Diversity
- Functional capabilities
- Activity

Disease/Outcomes

- What microorganisms are present (who is there)?

- What functionality does this community represent (what can they do)?

- Alpha-diversity

- Beta-diversity

Microbiome Study Design

Considerations:

- Outcomes vs. predictors
  - A variable may be both, but need to collect enough information on all potential predictors and outcomes

- Power calculation (1-β), effect size → often **not really known**

β = Type II Error (False Positives)

↓ β = ↑ Power

↑ Sample Size leads to ↑ Power

Microbiome Study Design

Sample collection:

- Site of sampling

- Samples unlikely to be contaminated by other sites or after collection
  - → ANY BACTERIA/FUNGI/ETC WILL BE AMPLIFIED

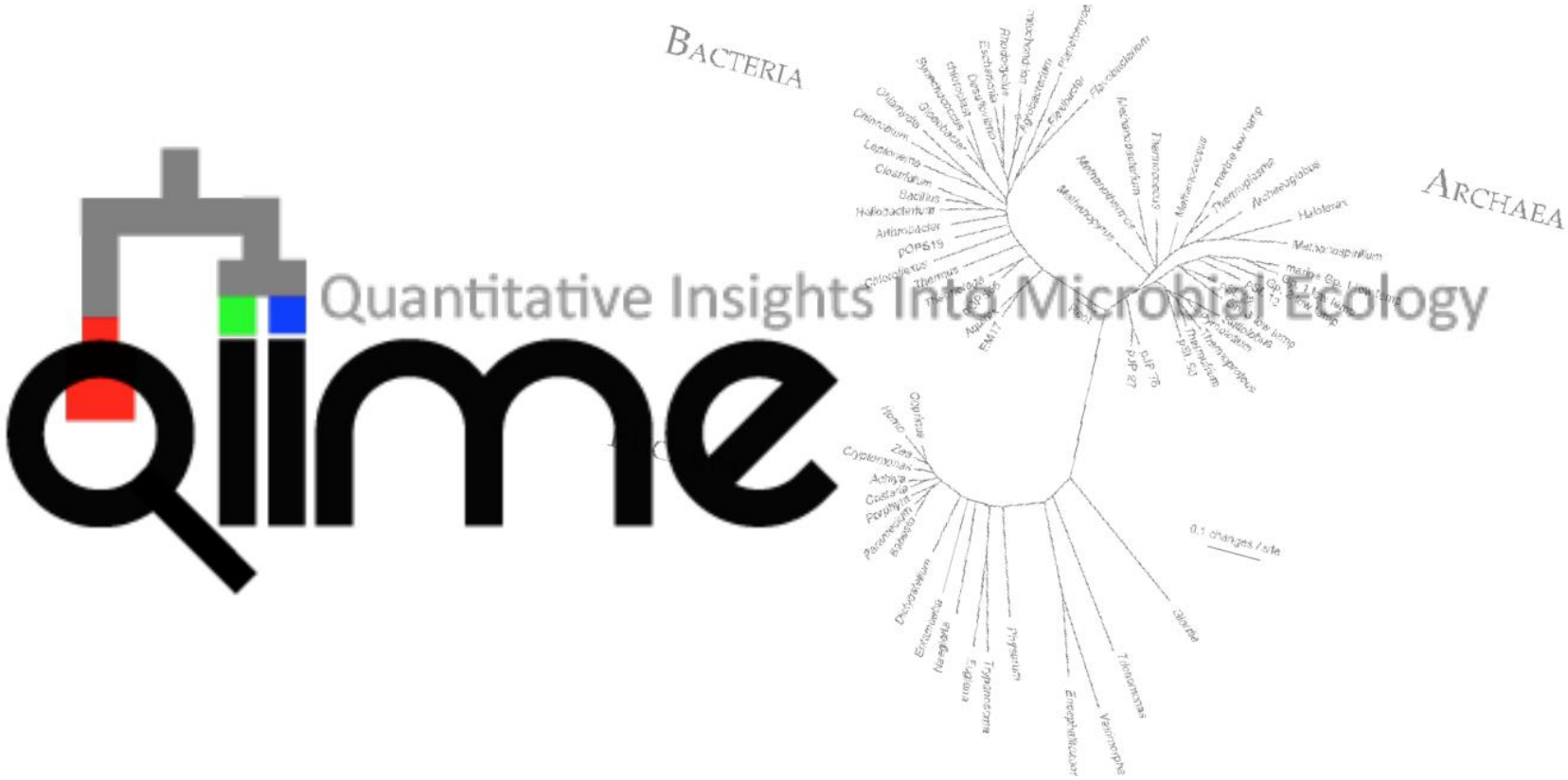
- Expected major bacterial/fungal organisms → may affect DNA extraction needed

- How much DNA? Depends on the type of sample (is it a mix of human and bacterial DNA? Fungi? etc.)

Metadata, Metadata, Metadata!

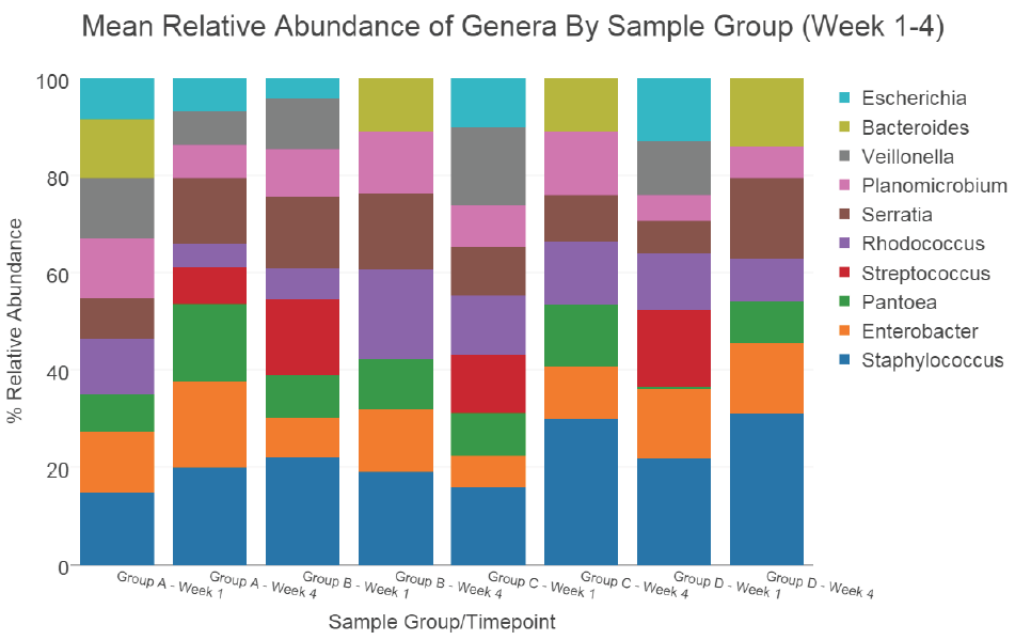
	A	B	C	D	E	F	G
1	#SampleID	BarcodeSequence	LinkerPrimerSequence	ForwardFastqFile	ReverseFastqFile	TreatmentGroup	Days_Post_Tx
2	30005			30005_1.fastq.gz	30005_2.fastq.gz	post	12
3	30017			30017_1.fastq.gz	30017_2.fastq.gz	post	33
4	30019			30019_1.fastq.gz	30019_2.fastq.gz	post	40
5	30029			30029_1.fastq.gz	30029_2.fastq.gz	post	55
6	30036			30036_1.fastq.gz	30036_2.fastq.gz	post	70
7	30107			30107_1.fastq.gz	30107_2.fastq.gz	post	161
8	30014			30014_S136_L001_R1_00	30014_S136_L001_R2	post	20
9	30008			30008_S139_L001_R1_00	30008_S139_L001_R2	post	8
10	30041			30041_S85_L001_R1_001	30041_S85_L001_R2_	post	66
11	30123			30123_S85_L001_R1_001	30123_S85_L001_R2_	post	178
12	30010			30010_S151_L001_R1_00	30010_S151_L001_R2	post	9
13	30012			30012_S49_L001_R1_001	30012_S49_L001_R2_	post	20
14	30016			30016_S163_L001_R1_00	30016_S163_L001_R2	post	6
15	30204			30204_S14_L001_R1_001	30204_S14_L001_R2_	post	211
16	30596			30596_S53_L001_R1_001	30596_S53_L001_R2_	post	385
17	30020			30020_S175_L001_R1_00	30020_S175_L001_R2	post	10
18	30035			30035_S21_L001_R1_001	30035_S21_L001_R2_	post	35
19	30215			30215_S50_L001_R1_001	30215_S50_L001_R2_	post	208
20	30544			30544_S88_L001_R1_001	30544_S88_L001_R2_	post	356
21	30026			30026_S80_L001_R1_001	30026_S80_L001_R2_	post	5
22	30168			30168_S2_L001_R1_001.f	30168_S2_L001_R2_0	post	178
23	30117			30117_S73_L001_R1_001	30117_S73_L001_R2_	post	151
24	30606			30606_S132_L001_R1_00	30606_S132_L001_R2	pre	-32
25	30661			30661_S61_L001_R1_001	30661_S61_L001_R2_	post	4

Variables should be consistently collected, coded, and easy to interpret (BY A COMPUTER)

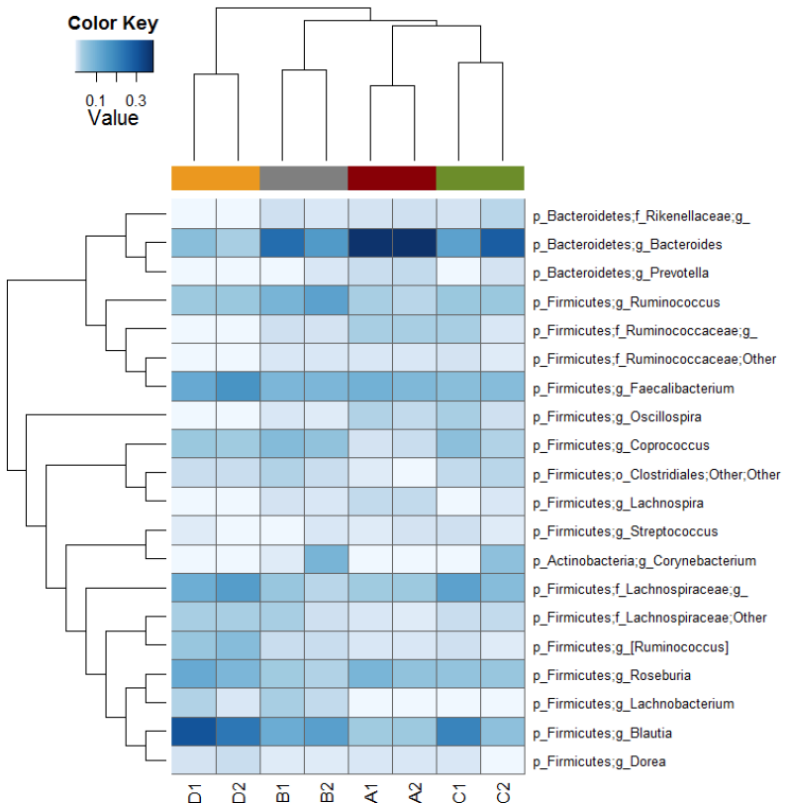

BACTERIA
ARCHAEA
Quantitative Insights Into Microbial Ecology
qiime
0.1 changes / site

Taxonomic Distribution

Bar chart



Heat map



Alpha diversity

Diversity within a sample – based on OTU assignments

Richness – number of species present (Chao index)

Evenness – abundance of different species (Shannon index)

Rich and even

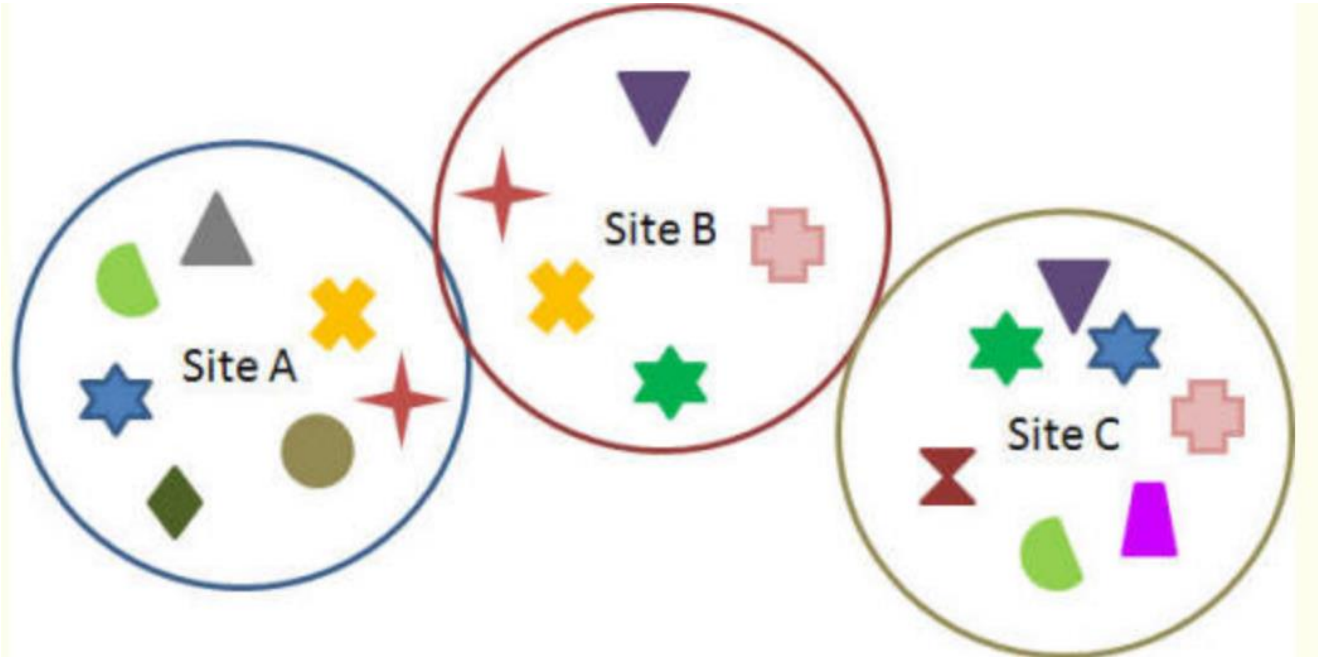
Not rich but even

Beta diversity

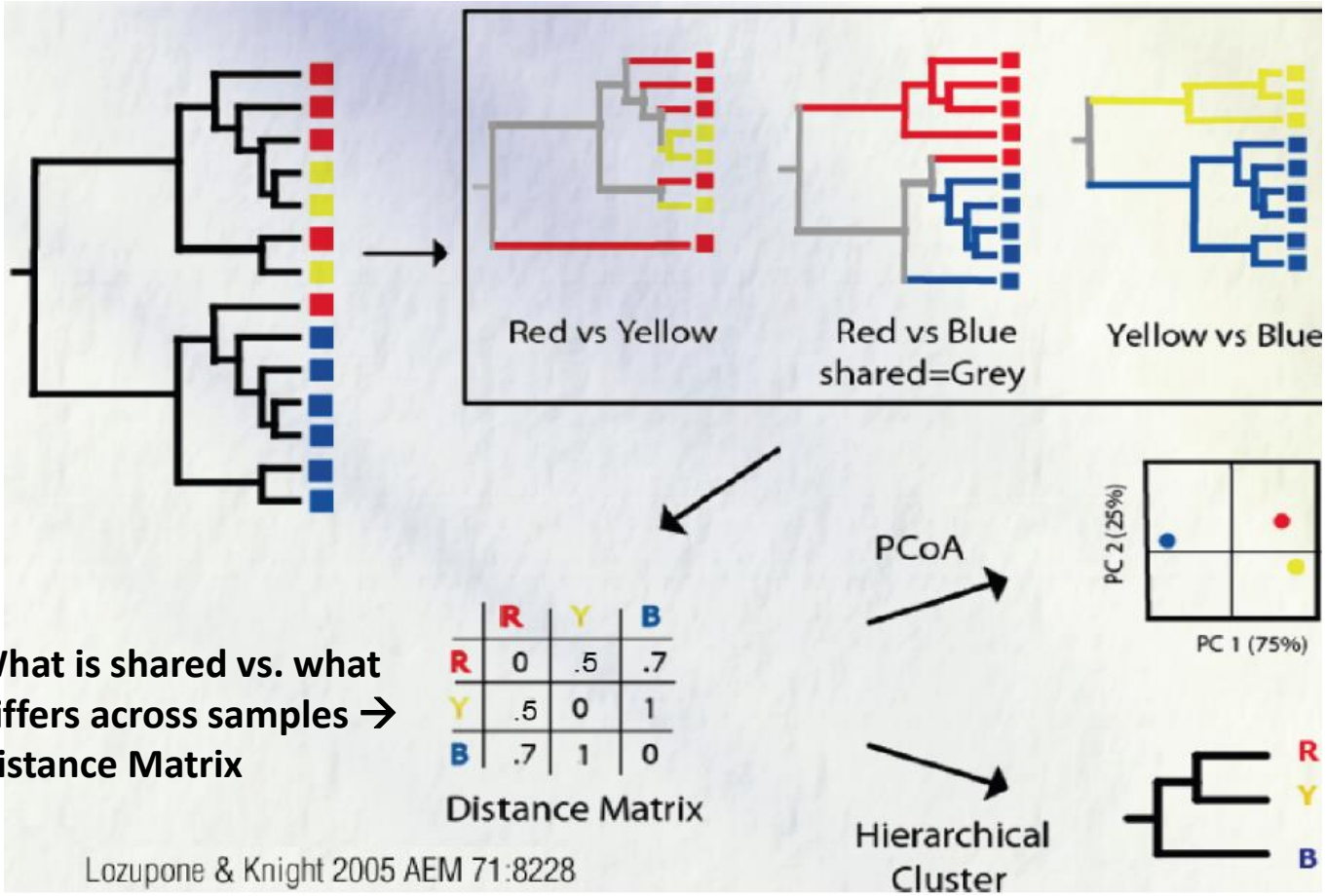
Cross-sample relatedness

How different are types present?

Measure of genetic distance / dissimilarity between sample pair



UniFraq to determine beta-diversity - sequence alignment based



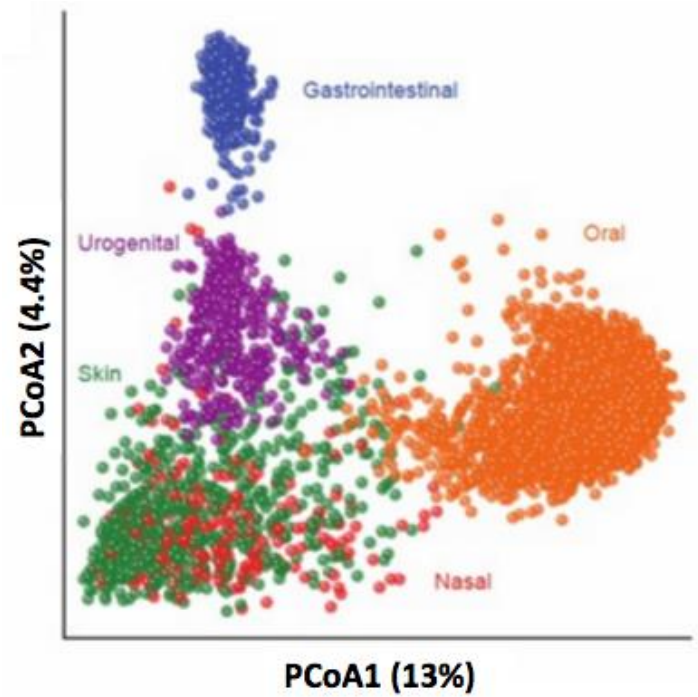
What is shared vs. what differs across samples → Distance Matrix

Principal Coordinate Analysis

Visualization of beta diversity matrix

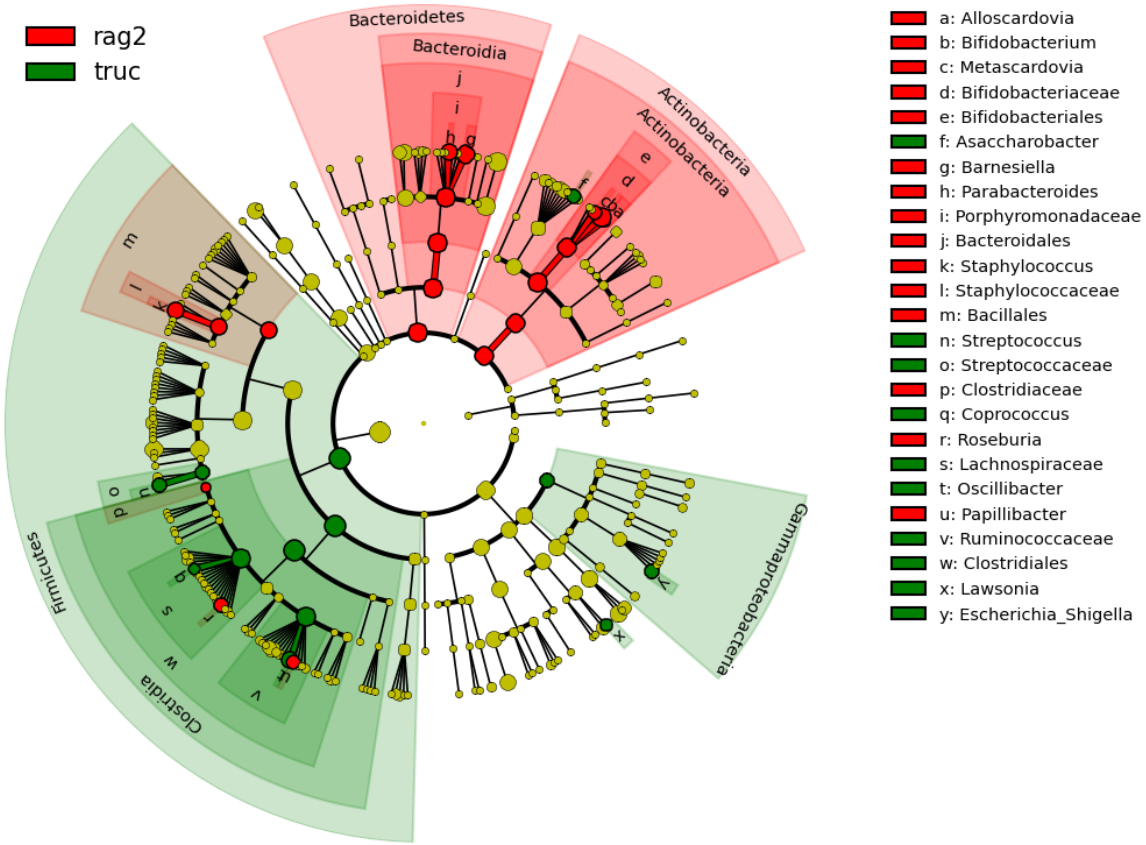
Transform distance matrix into new set of orthogonal axes

2D or 3D



LEfSe – Effect sizes across sample groups

Differential abundance of key taxa (biomarkers)



Segata, N., et al. (2011). "Metagenomic biomarker discovery and explanation." Genome Biol **12**(6): R60.

PICRUST – Inferring Functional Potential from Taxonomy

What do we want to ask?

Microbiome/Dysbiosis

- Presence/Absence
- Differential abundance
- Diversity
- Functional capabilities
- Activity

Disease/Outcomes

Summary

Microbiota can influence host in a number of ways → bacteria, fungi, viruses

16S rRNA sequencing remains the method of choice (cost, ease of use) for initial surveys of microbiome communities

ITS gaining ground as databases improve fungal-host connections are made

Metagenomics more informative but with logistical limitations

Taxonomic profiles, alpha-diversity, beta-diversity, differential abundance of taxa can all give different sides of the picture

Importance of honing study design and data collection prior to sequencing

Many opportunities for growth and collaborations: immunology, metabolomic studies, validation in animal models, mycobiome; analysis of existing datasets, epidemiology and biostatistics

Acknowledgements

Microbiome Core / Uhlemann Lab Group

Anne-Catrin Uhlemann

Stephania Stump

Marley Giddins

Sabrina Khan

Thomas McConville

Angela Gomez-Simmonds

Nenad Macesic

Thank You!