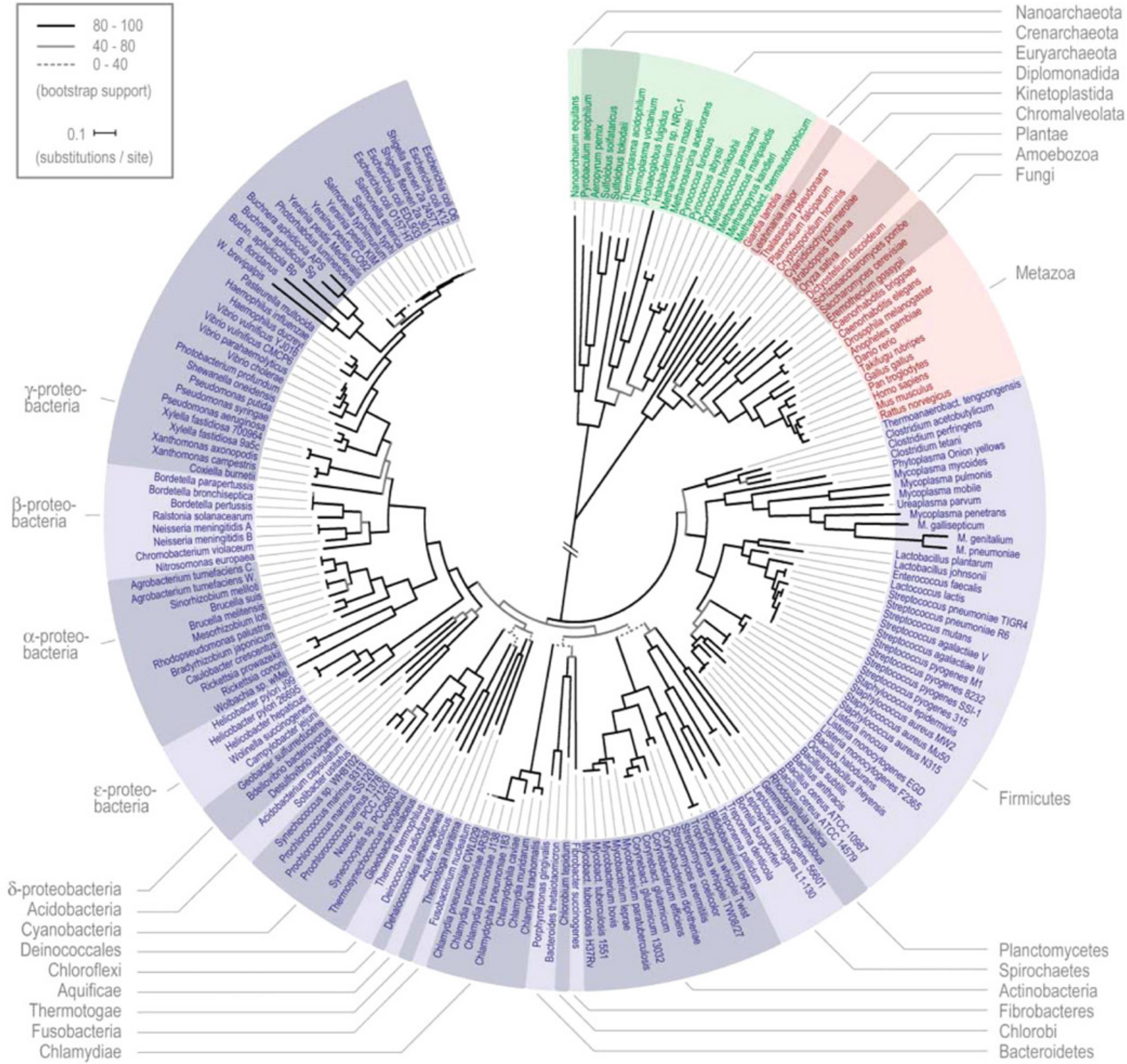
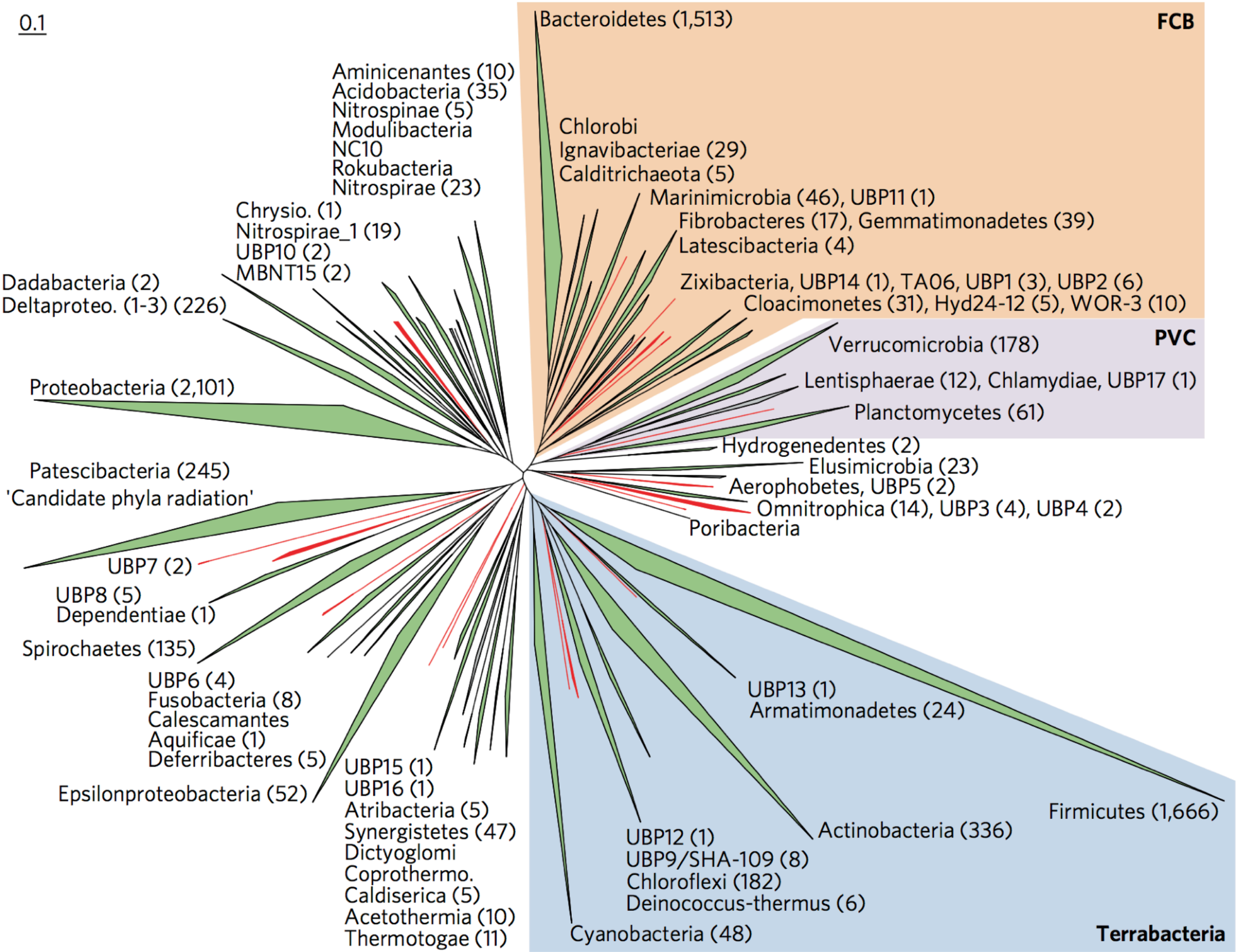
Microbiome Module

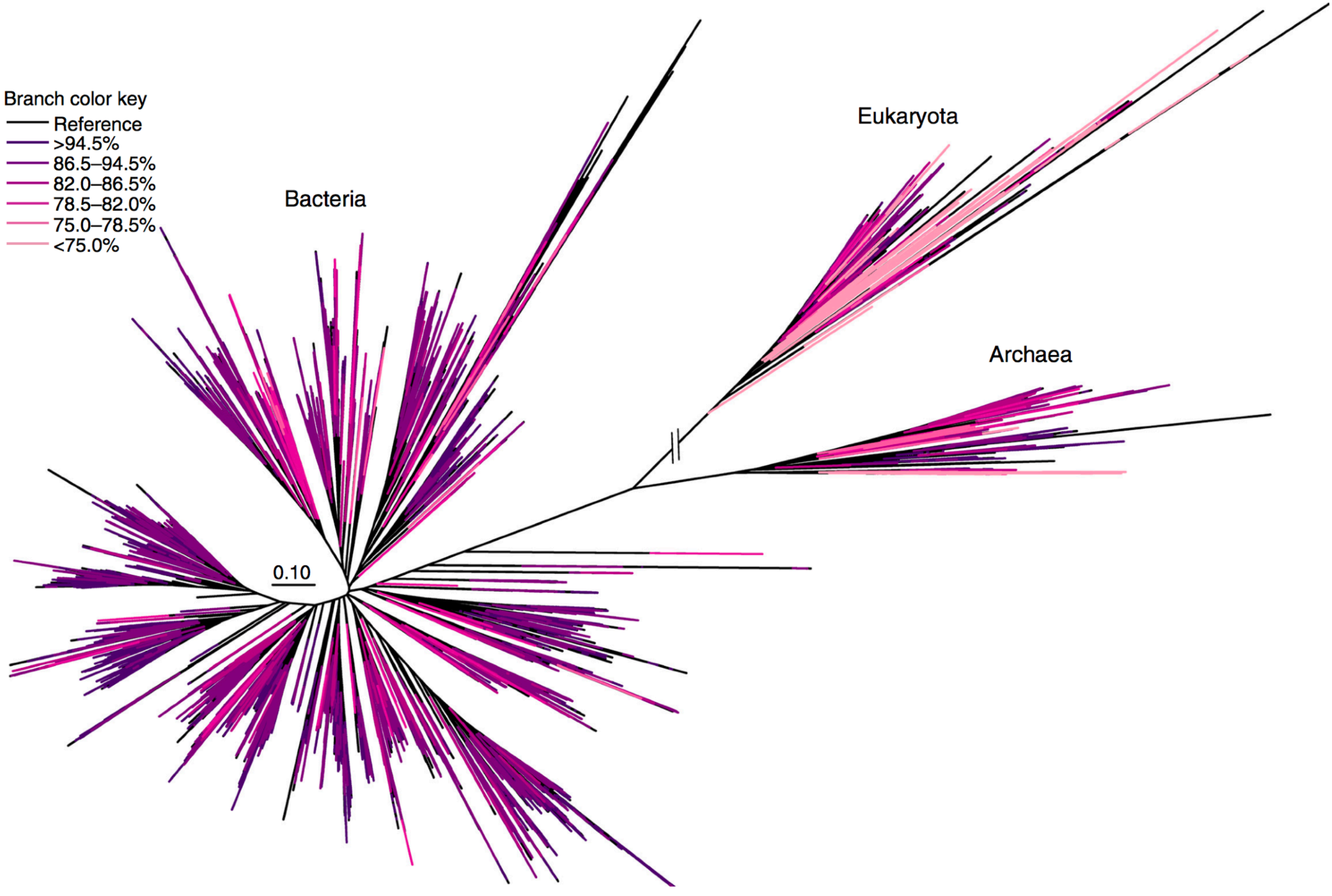
Instructor: Erik Wright

The microbiome

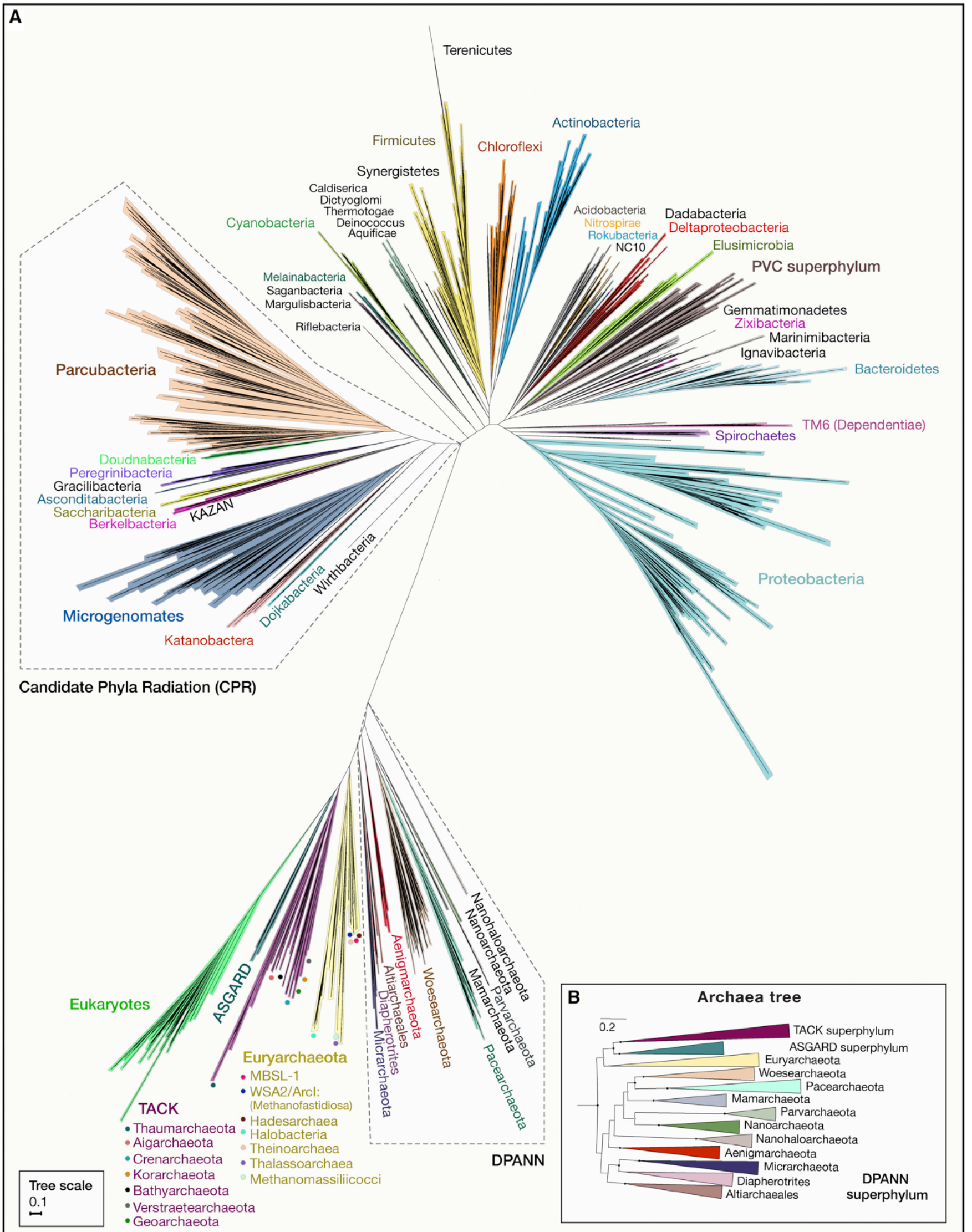

80 - 100
40 - 80
0 - 40
(bootstrap support)
0.1
(substitutions / site)
Nanoarchaeota
Crenarchaeota
Euryarchaeota
Diplomonadida
Kinetoplastida
Chromalveolata
Plantae
Amoebozoa
Fungi
Metazoa
Firmicutes
Planctomycetes
Spirochaetes
Actinobacteria
Fibrobacteres
Chlorobi
Bacteroidetes
γ-proteo-
bacteria
β-proteo-
bacteria
α-proteo-
bacteria
ε-proteo-
bacteria
δ-proteobacteria
Acidobacteria
Cyanobacteria
Deinococcales
Chloroflexi
Aquificae
Thermotogae
Fusobacteria
Chlamydiae



Parks, D. H., *et al*. (2017). *Nature Microbiology*.

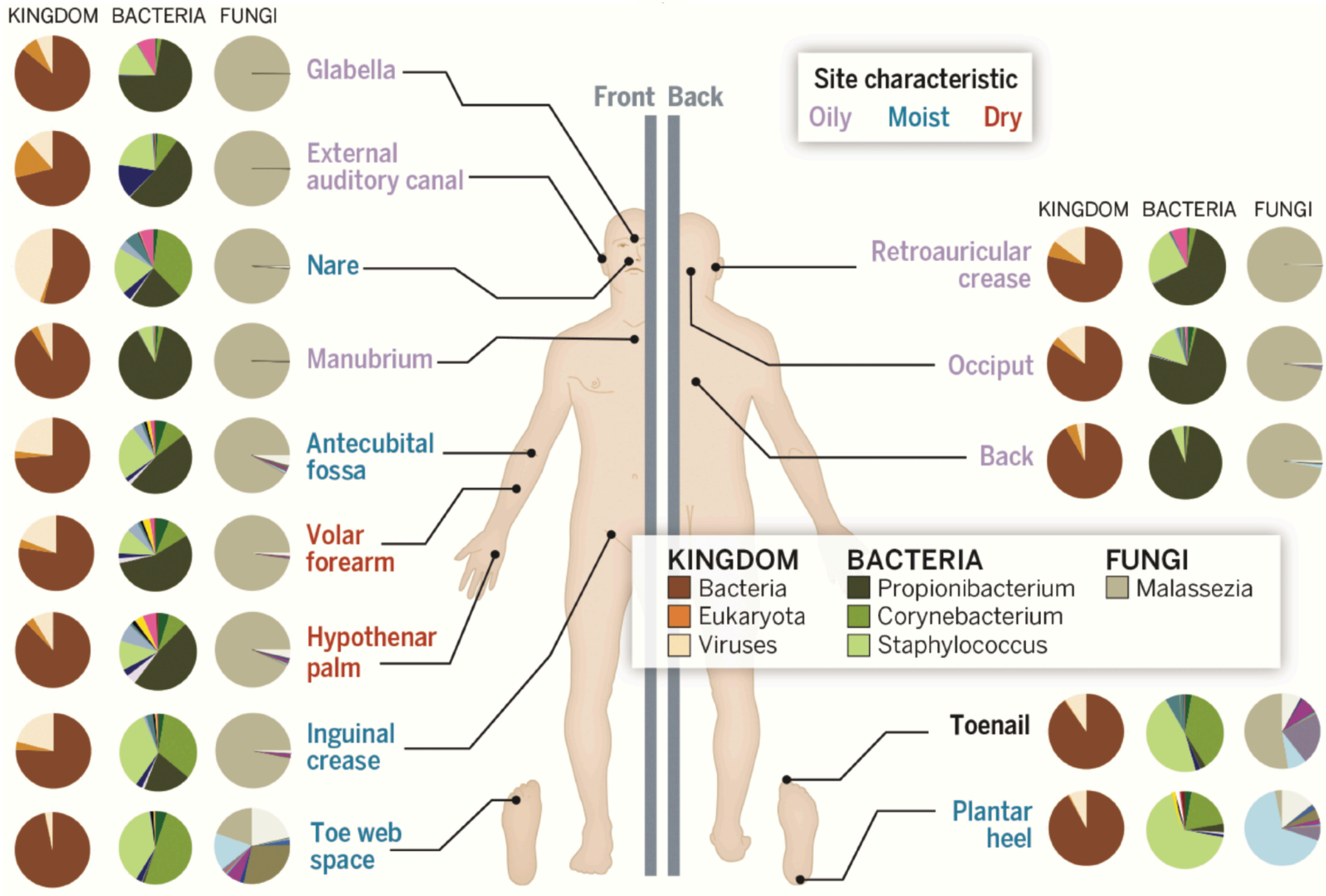


Karst, S. M., *et al*. (2018). *Nature Biotechnology*.



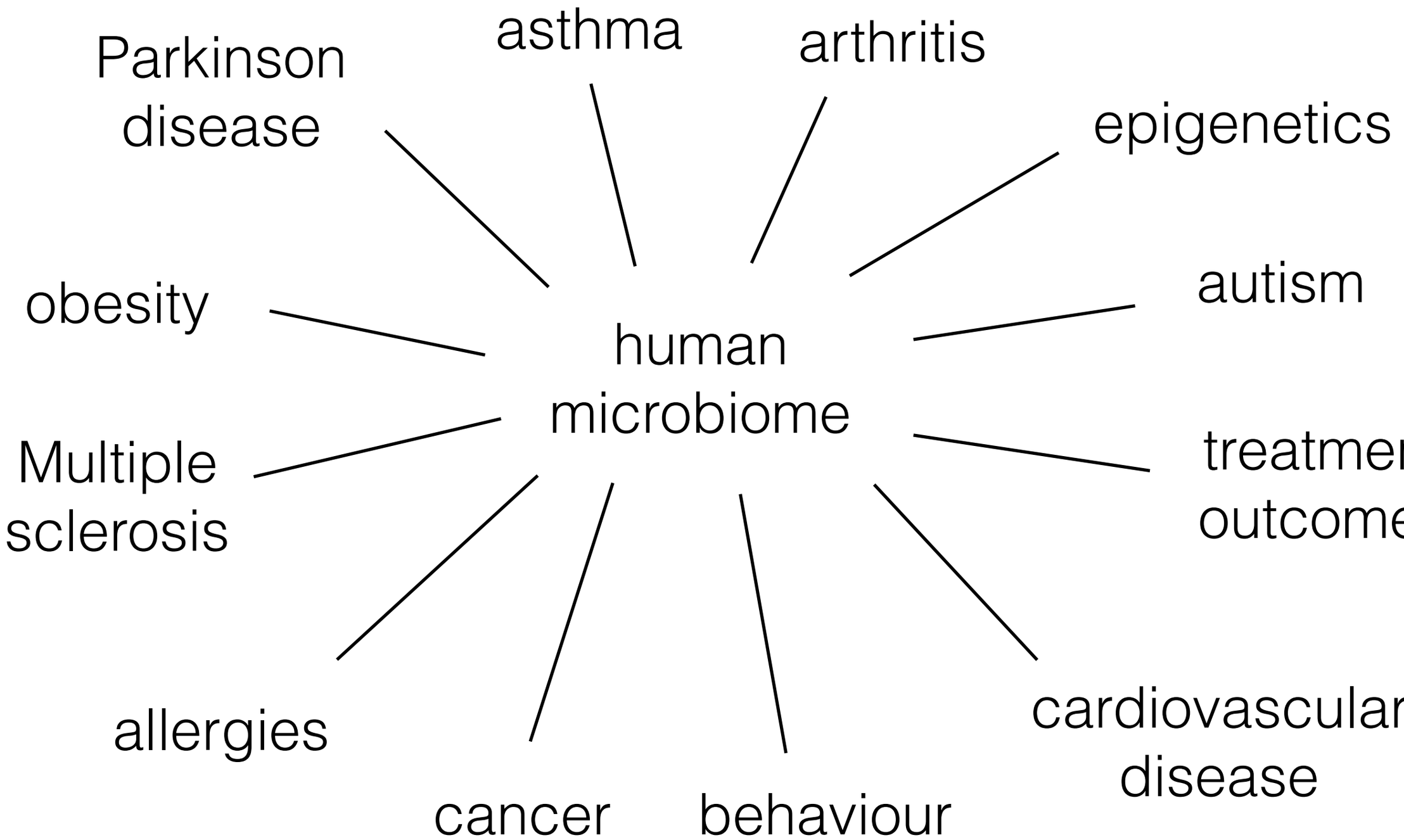
Castelle, C. & Banfield, J. (2018). *Cell*.

Microbiome composition is important to health



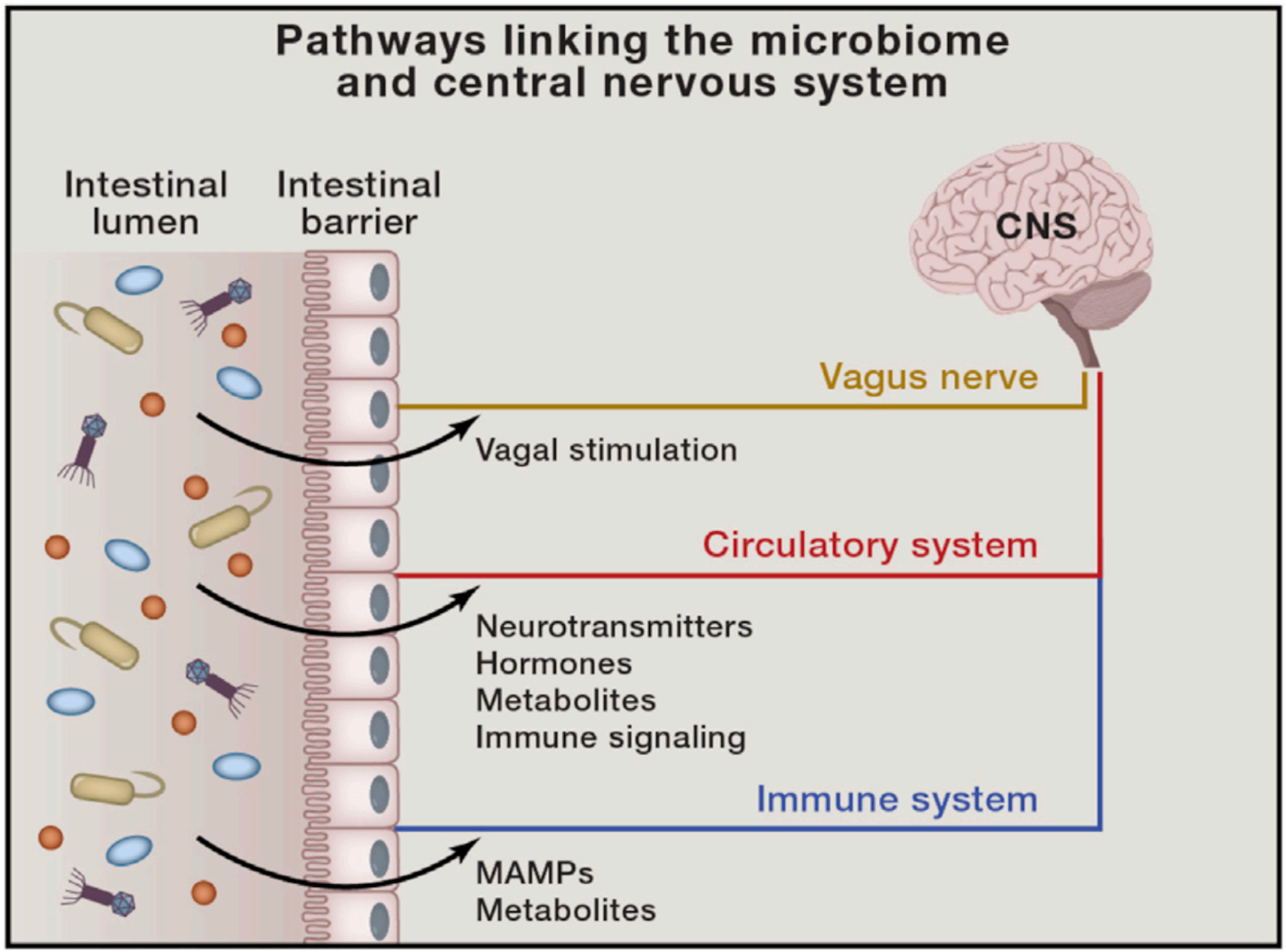
Belkaid, Y. & Segre, J. (2014). *Science.*

Microbiome composition is important to health



Nussinov, R. & Papin, J. (2017). *PLoS Computational Biology.*

The gut microbiota has many connections



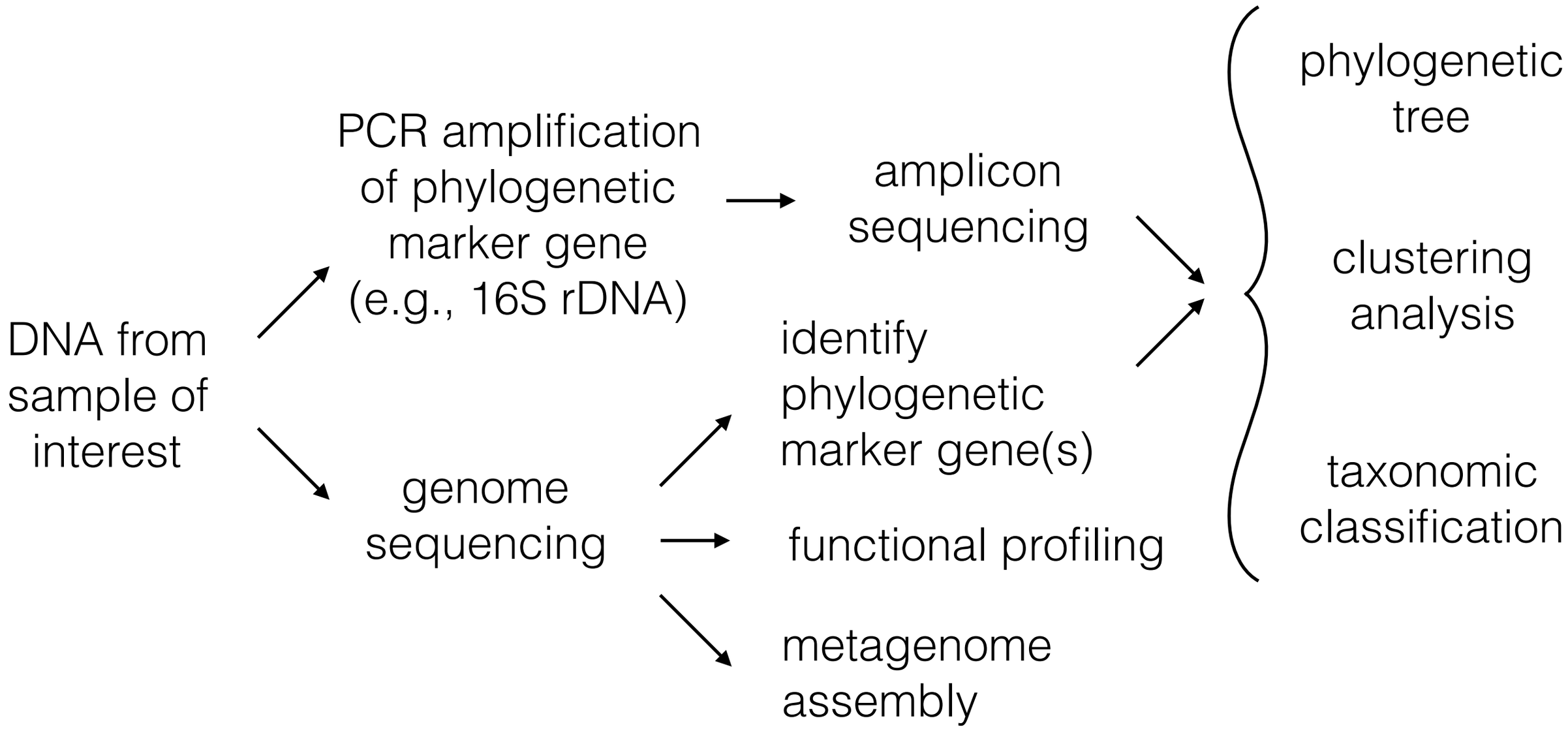
Sampson, T. & Mazmanian, S. (2015). *Cell Host & Microbe.*

Manipulations of the gut microbiome

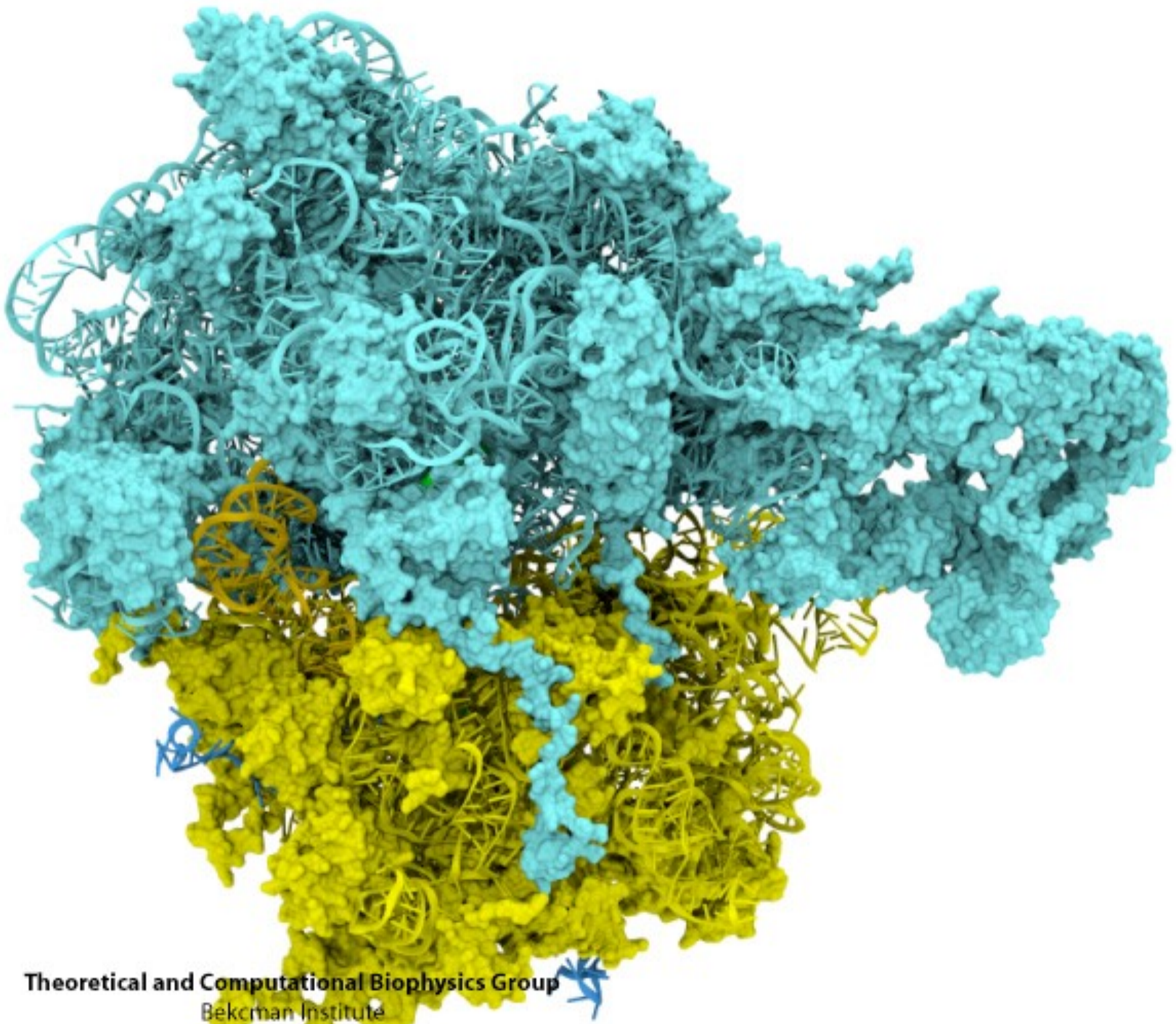
fecal transplants

Popular approaches to studying the microbiome

Typical microbiome bioinformatic workflows

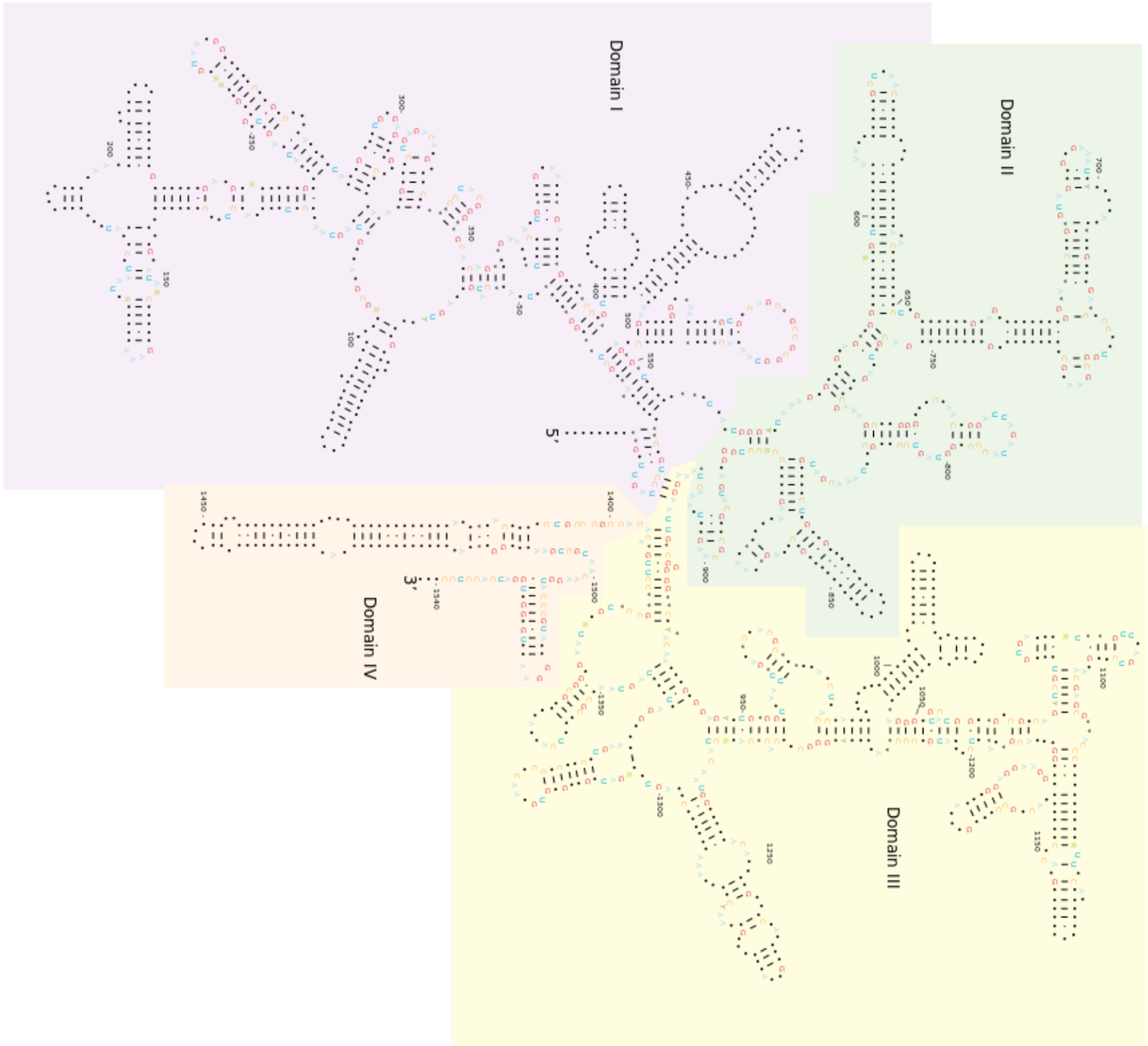


Quaternary structure of the ribosome

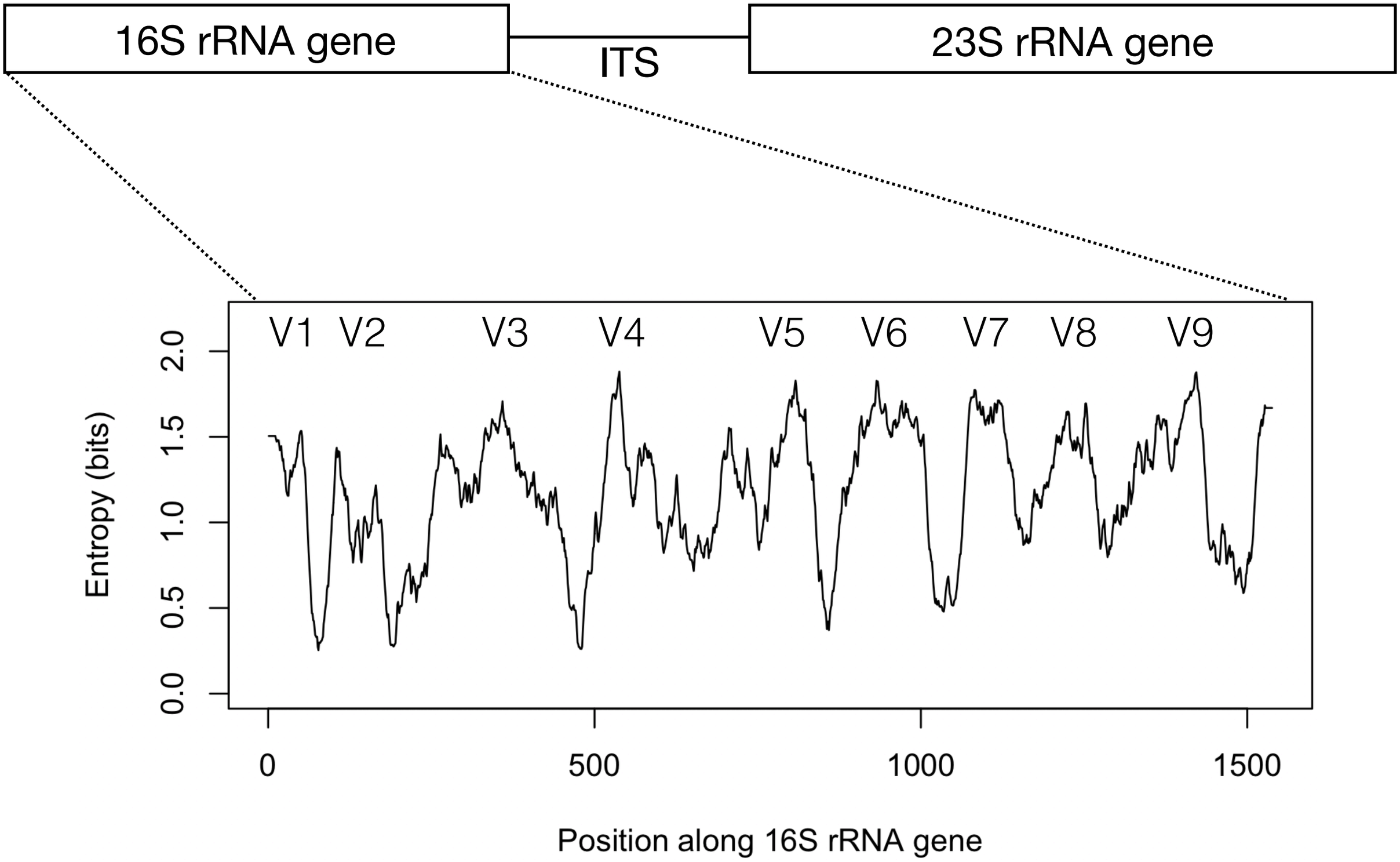


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

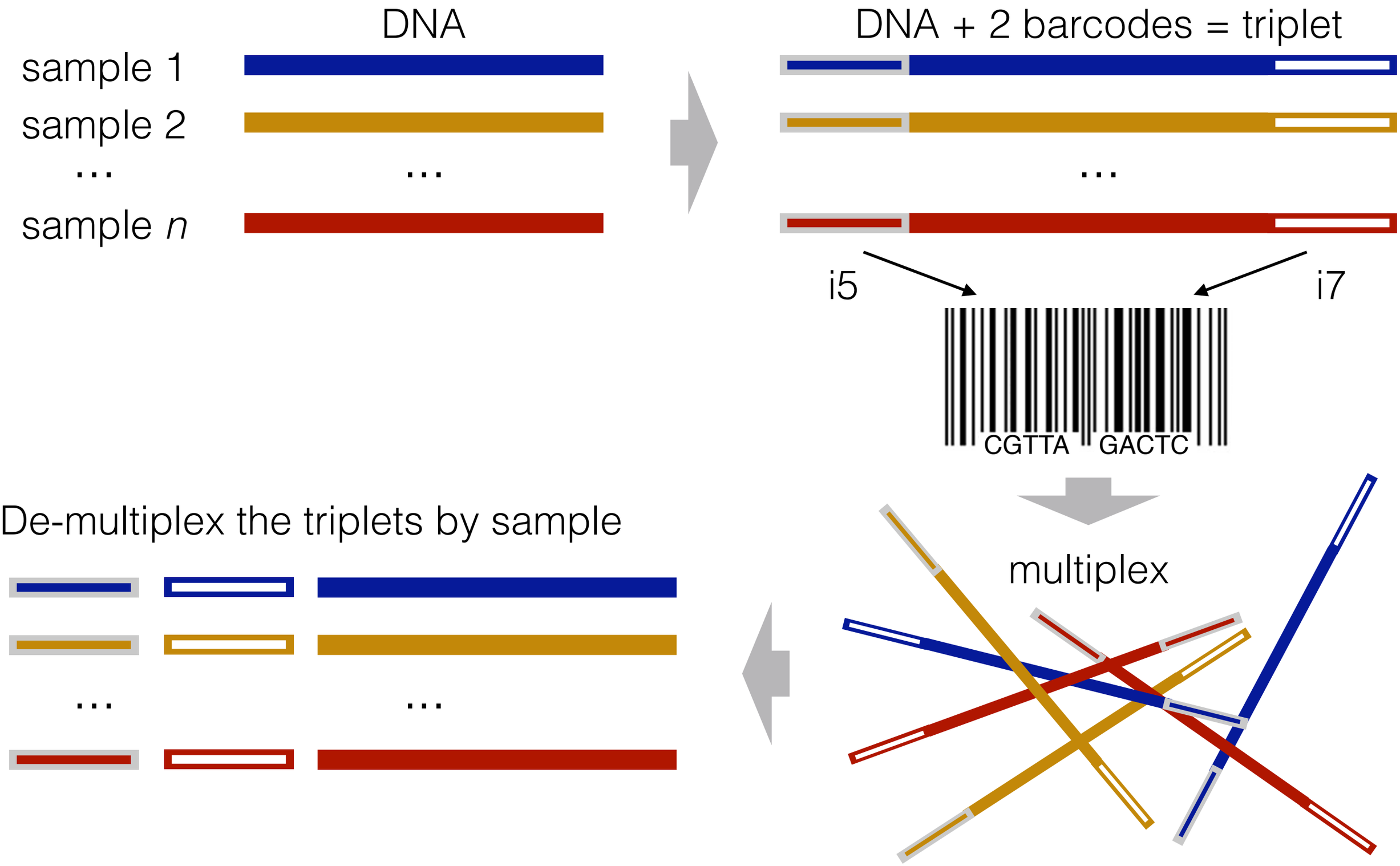
Secondary structure of the 16S rRNA



16S marker gene sequencing

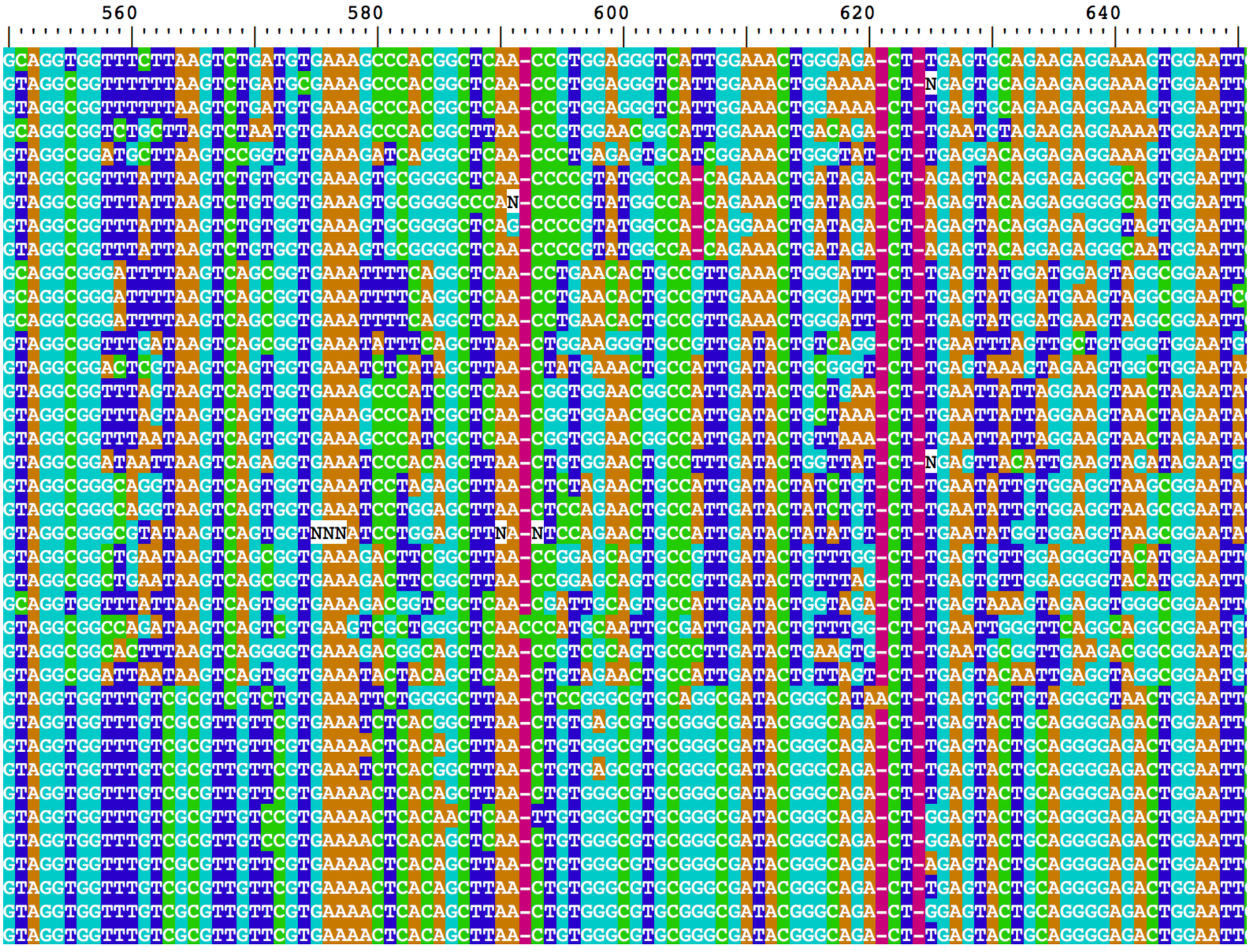


Multiplexing on the Illumina platform

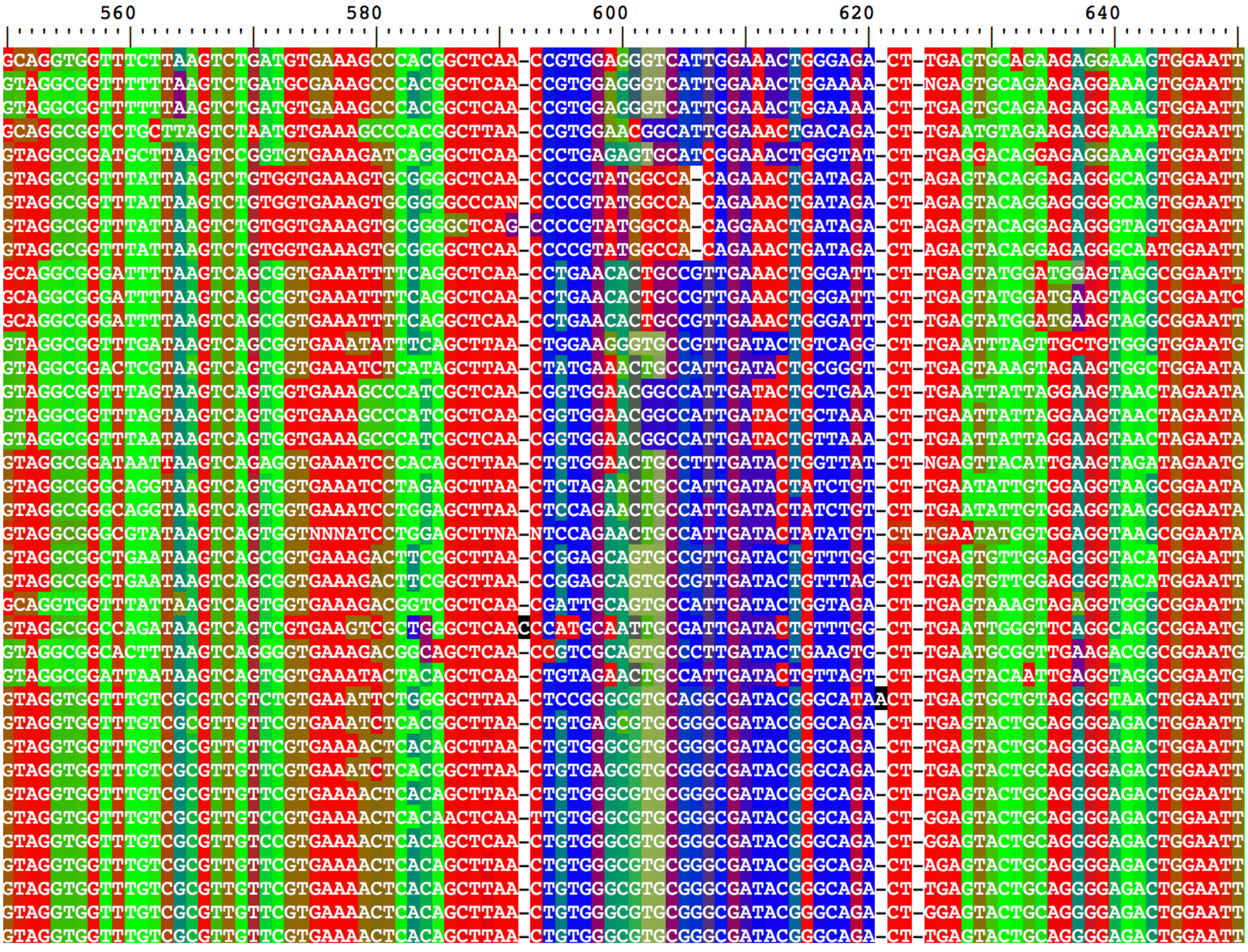


Reference-free analysis approaches

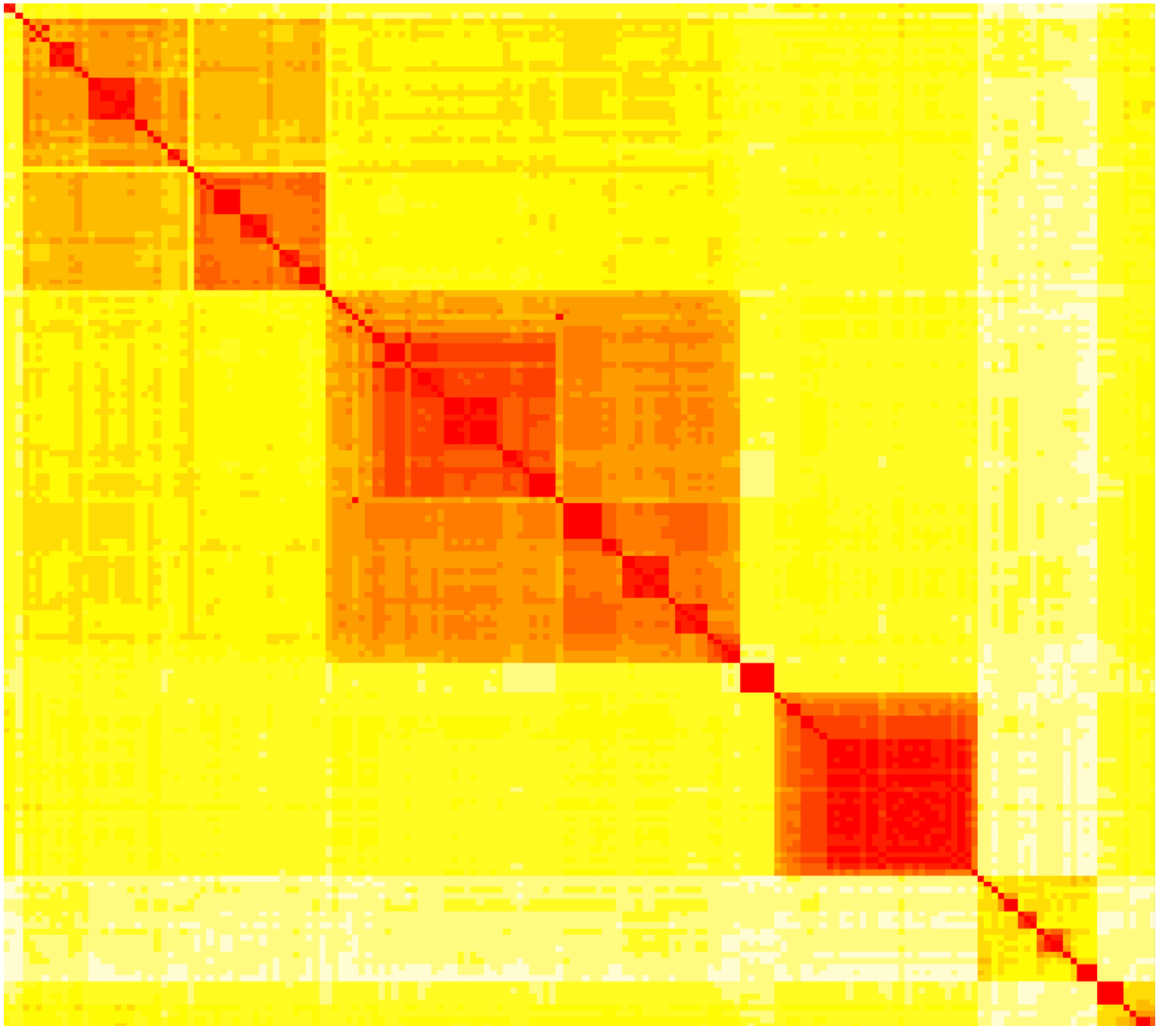
Constructing/interpreting phylogenetic trees



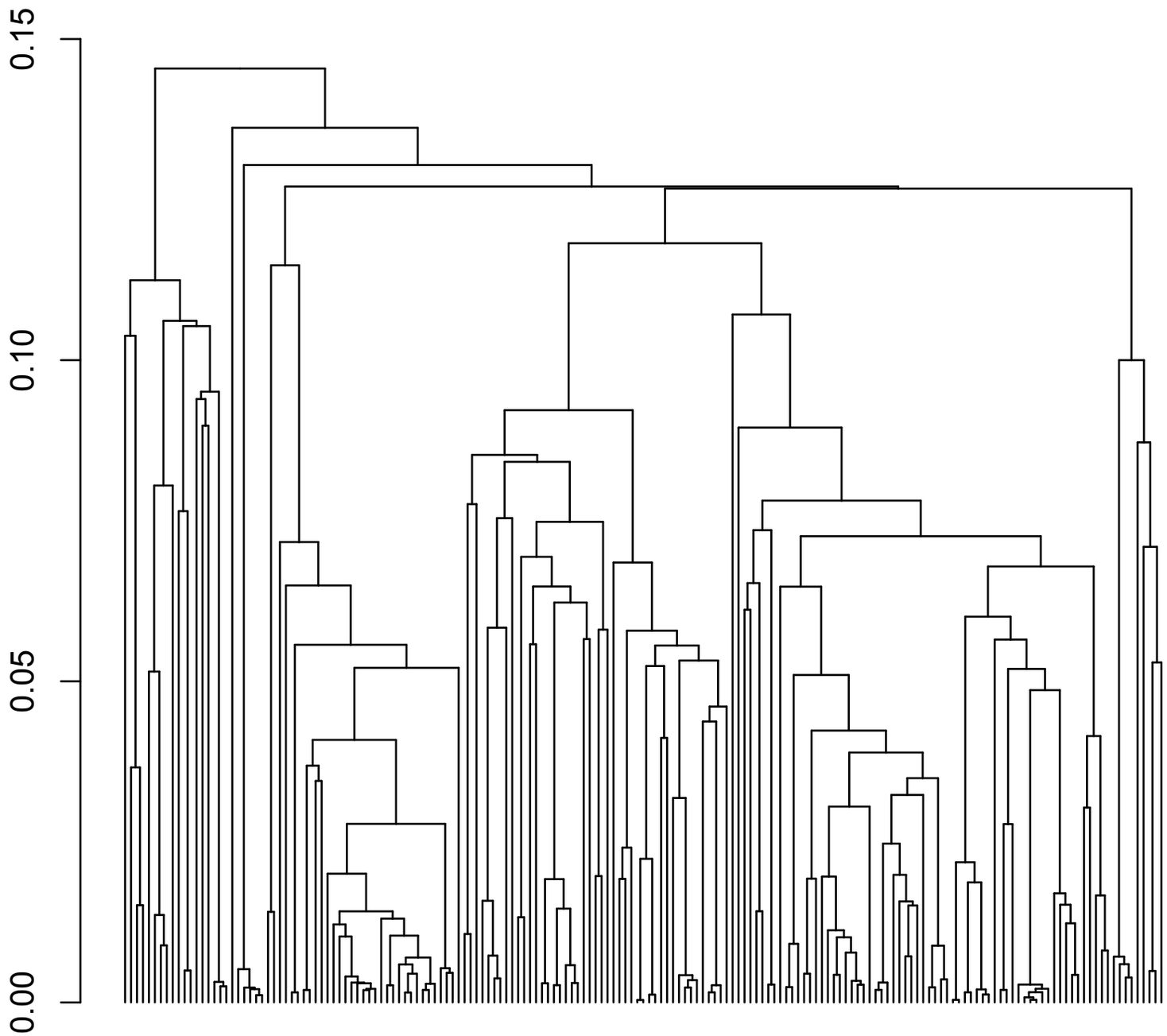
Constructing/interpreting phylogenetic trees



Constructing/interpreting phylogenetic trees



Constructing/interpreting phylogenetic trees

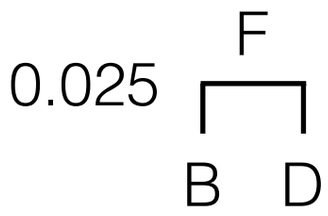


Constructing/interpreting phylogenetic trees

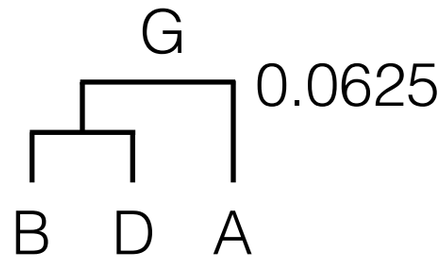
	A	B	C	D
A	0			
B	0.1	0		
C	0.3	0.2	0	
D	0.15	0.05	0.25	0

Constructing/interpreting phylogenetic trees

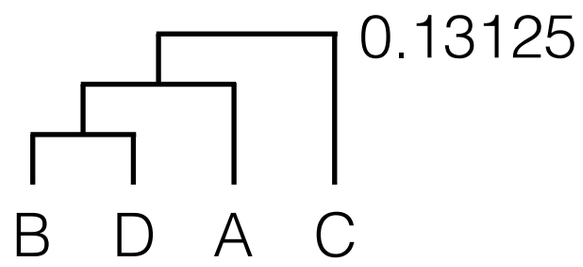
	A	B	C	D
A	0			
B	0.1	0		
C	0.3	0.2	0	
D	0.15	**0.05**	0.25	0



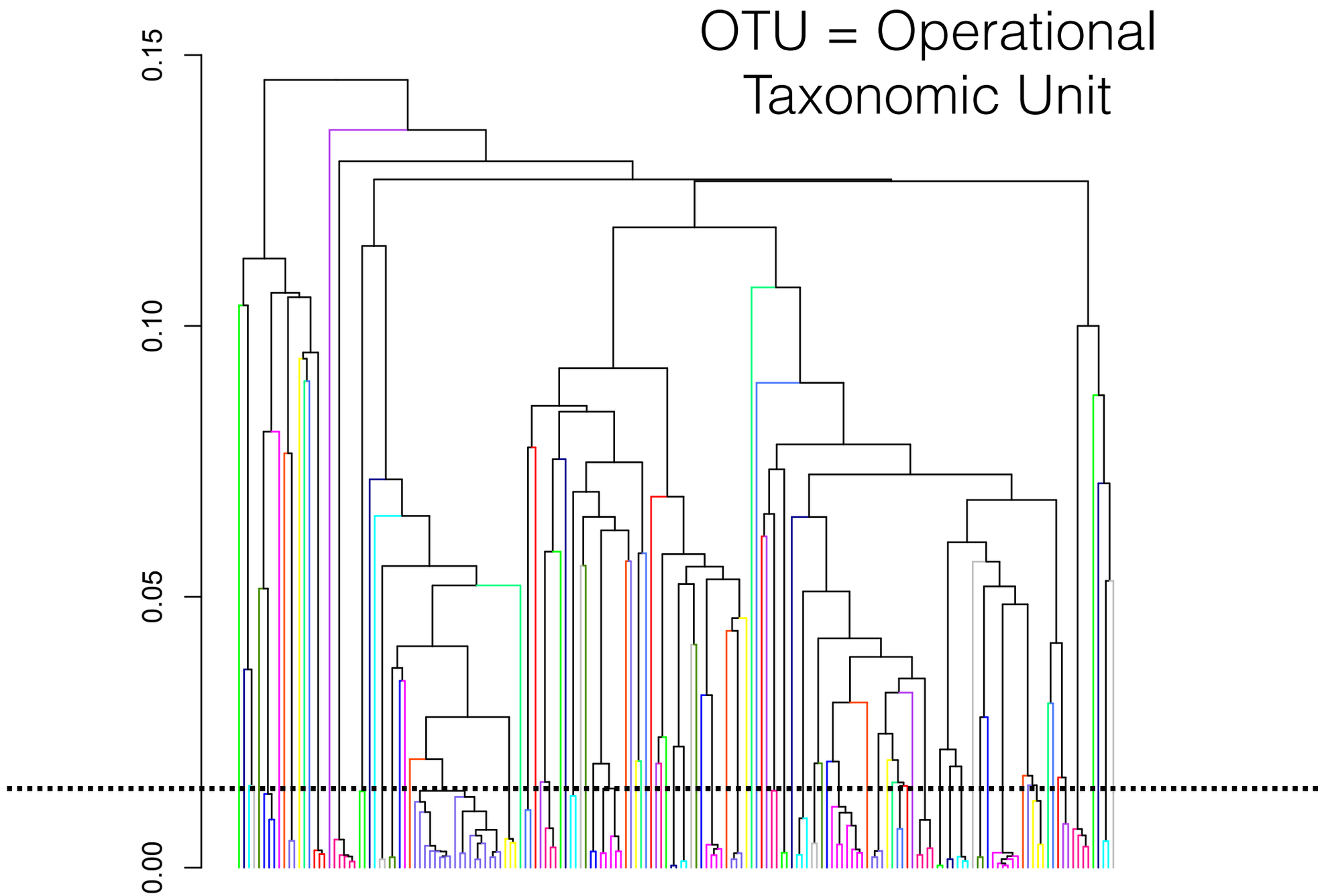
	A	F	C
A	0		
F	**0.12**	0	
C	0.3	0.23	0



	G	C
G	0	
C	**0.26**	0



Clustering into OTUs at ≥ 97% similarity

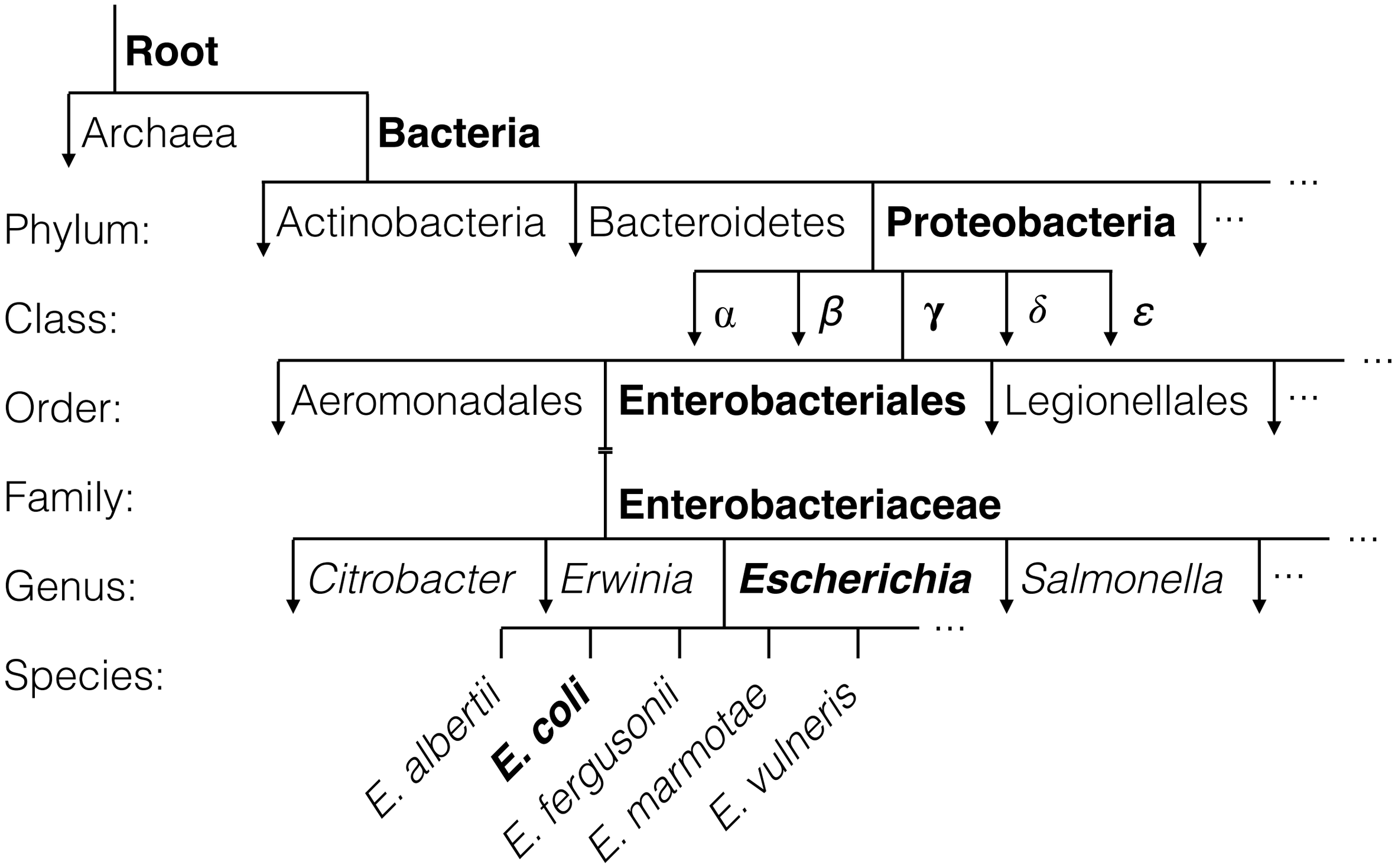


Generating an OTU table

#OTU ID	700099606	700106946	700103482	700103479	Consensus Lineage
OTU_97.12271	0	0	0	0	Root;p__Proteobacteria;c__Alphaproteobacteria;o__;f__;g__
OTU_97.12272	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces
OTU_97.12273	0	0	0	0	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella
OTU_97.12274	0	0	0	0	Root;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Neisseria
OTU_97.12275	2	0	0	0	Root;p__Fusobacteria;c__Fusobacteria;o__Fusobacteriales;f__Fusobacteriaceae;g__Leptotrichia
OTU_97.12276	0	0	0	0	Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus
OTU_97.12277	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium
OTU_97.12278	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium
OTU_97.12279	0	5	0	11	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides
OTU_97.1228	0	0	0	2	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides
OTU_97.12280	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium
OTU_97.12281	0	0	0	0	Root;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga
OTU_97.12282	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales
OTU_97.12283	0	0	0	0	Root;p__GN02
OTU_97.12284	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Citricoccus
OTU_97.12285	0	0	0	0	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas
OTU_97.12286	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium
OTU_97.12287	0	0	0	0	Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus

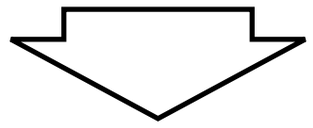
Reference-based analysis

Classifying sequences into a taxonomy



Classifying sequences into a taxonomy

AGCGGCAGCACAGAGGAACTTGTTCCTTGGGTGGCGAGCG...

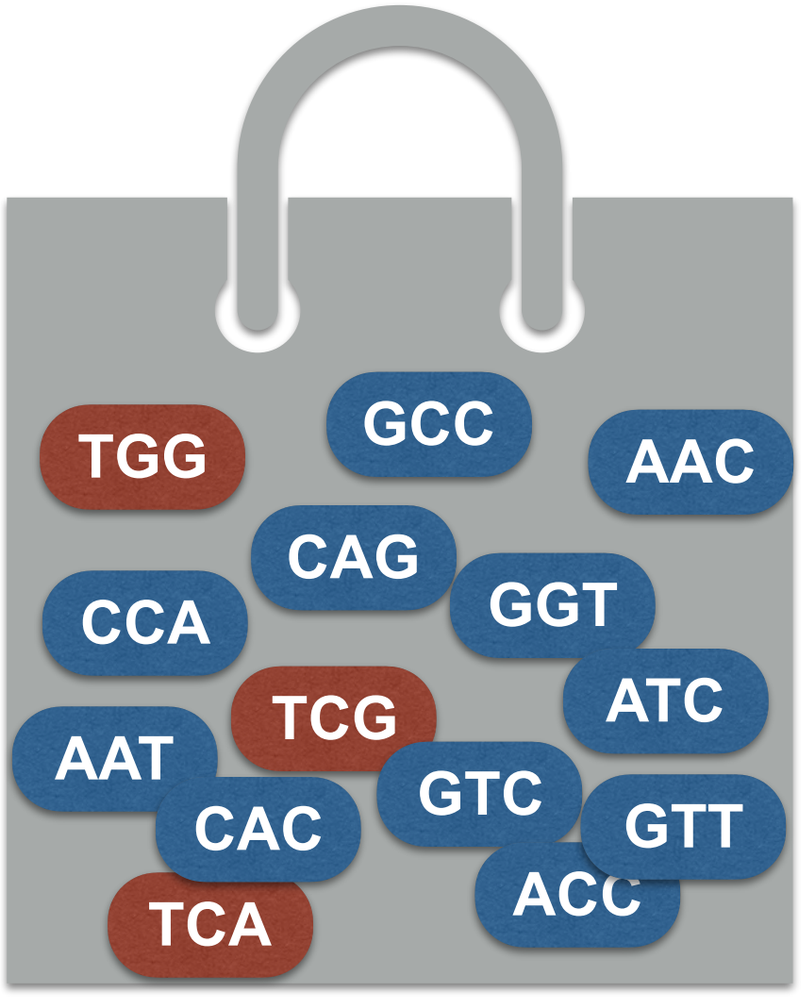


Root (97.8%);
  Bacteria (97.8%);
    Proteobacteria (97.8%);
      Gammaproteobacteria (97.8%);
        Enterobacteriales (97.8%);
          Enterobacteriaceae (97.8%);
            Escherichia (95%)

Classifying via machine learning

CACCGGTTCAGTCG...

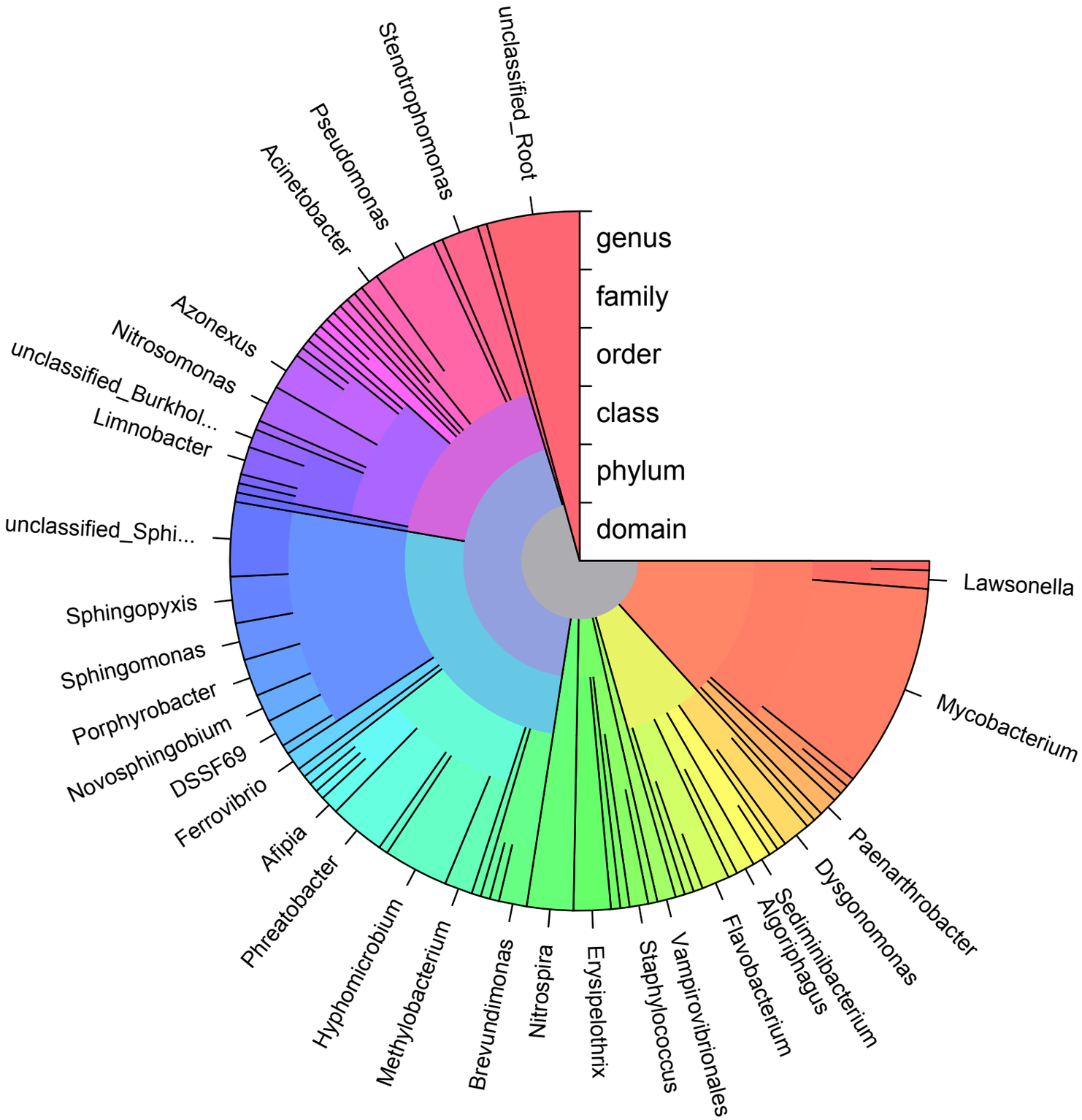
k-mers (K=3)



Bootstrap replicates (B=100)

	Subsamples (S=3)			
1	GCC	CAG	GGT	Escherichia
2	CAG	TCG	AAT	Escherichia
3	GCC	GCC	TCA	Escherichia
4	TTT	TTT	TGG	Salmonella
5	CCA	GCC	GTC	Escherichia
6	AAT	TCG	TTT	Salmonella
7	GCC	CAG	AAC	Escherichia
8	AAT	GTC	GGT	Escherichia
...	...	...		
100	CAC	TTT	GGT	Escherichia

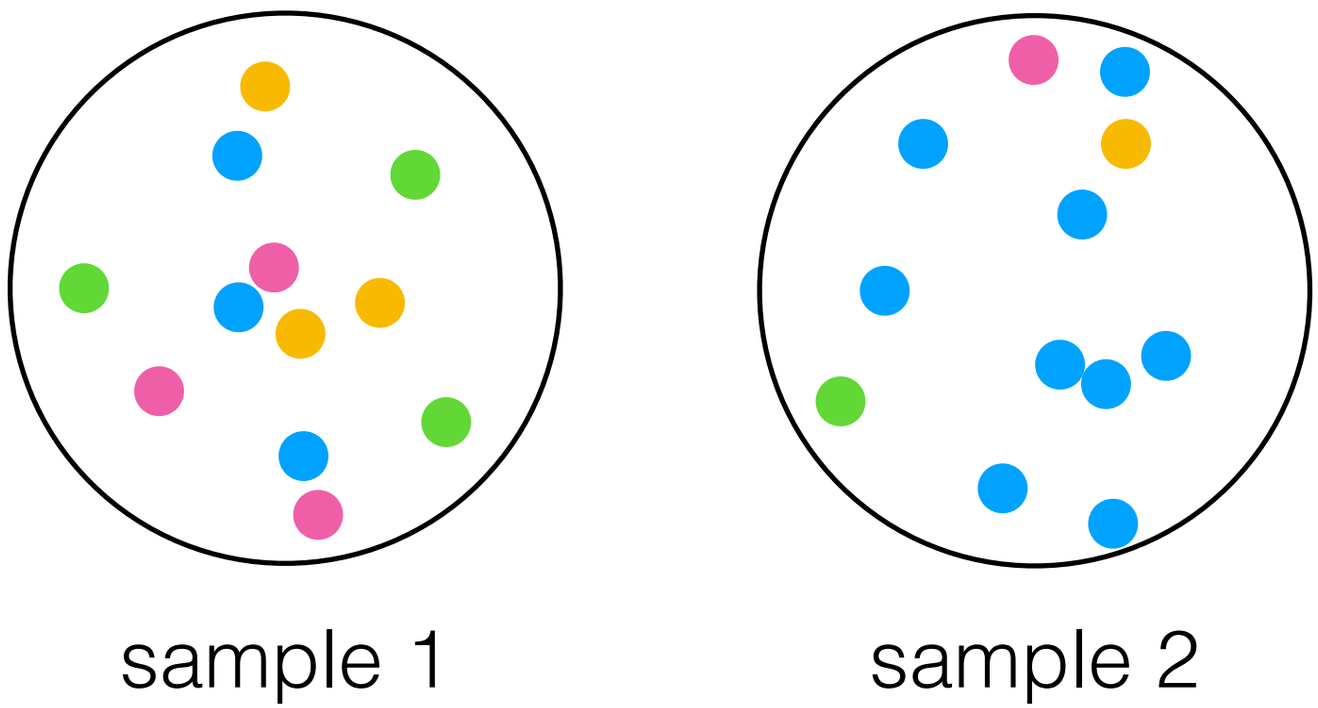
Visualizing the results of classification



Analyzing multiple samples

Measures of diversity within samples

- α-Diversity: within-sample taxonomic diversity
  - Simply count the number of OTUs
    - only measures richness
  - Shannon Index = $-sum_i(p_i * ln(p_i))$, p_i = proportion of each OTU
    - measures richness and evenness
  - Simpson Index = $sum_i(p_i^2)$, p_i = proportion of each OTU
    - measures richness and evenness



sample 1

sample 2

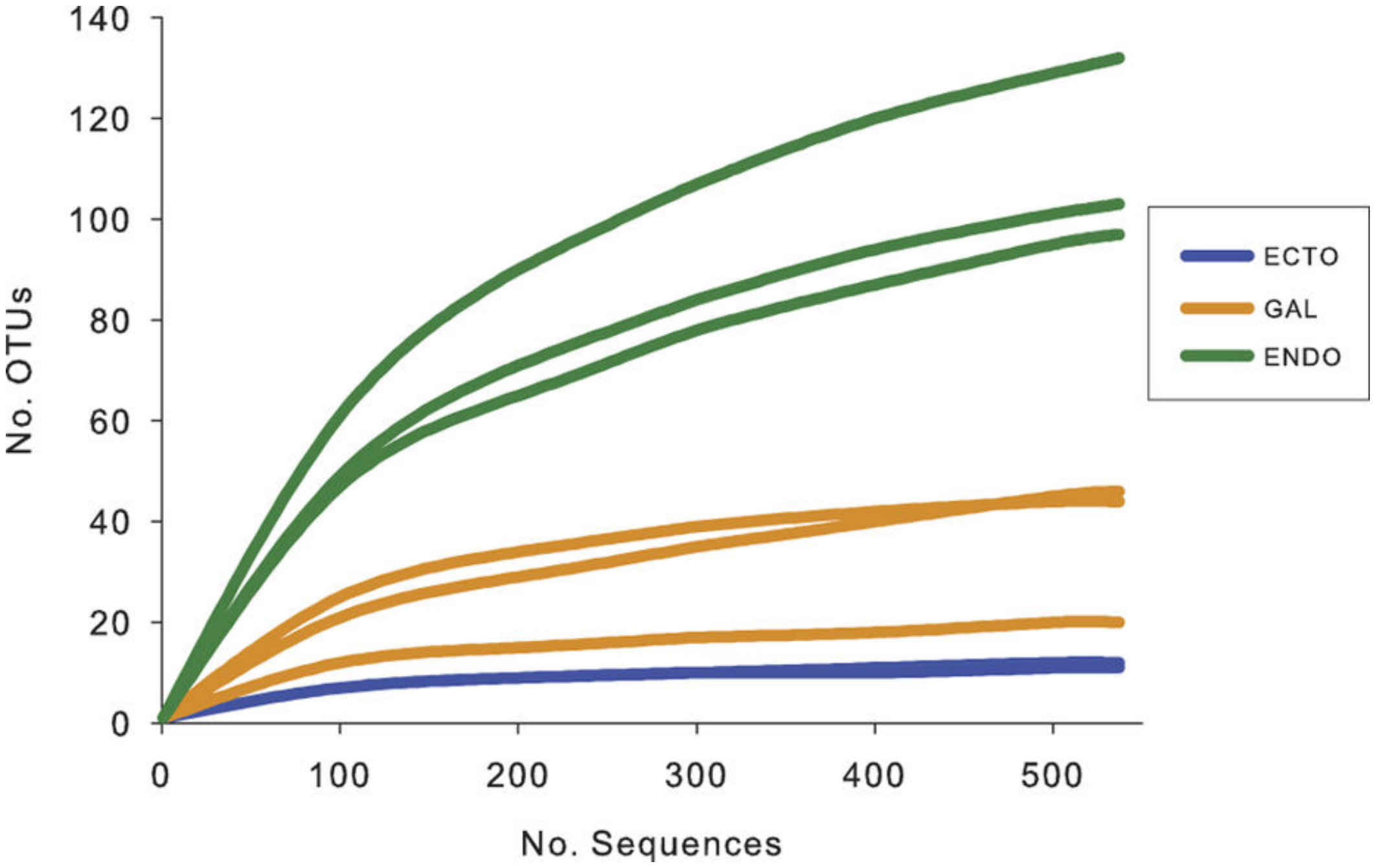
Measures of diversity within samples

	Sample 1	Sample 2	Sample 3
OTU1	0.0	**0.1**	**0.1**
OTU2	**0.4**	**0.2**	0.0
OTU3	**0.2**	**0.1**	**0.8**
OTU4	0.0	**0.4**	0.0
OTU5	**0.2**	**0.2**	**0.1**
α-Diversity			
Shannon Index			
Simpson Index			

Measures of diversity within samples

	Sample 1	Sample 2	Sample 3
OTU1	0.0	**0.1**	**0.1**
OTU2	**0.4**	**0.2**	0.0
OTU3	**0.2**	**0.1**	**0.8**
OTU4	0.0	**0.4**	0.0
OTU5	**0.2**	**0.2**	**0.1**
α-Diversity	3	5	3
Shannon Index	1.01	1.47	0.64
Simpson Index	0.24	0.26	0.66

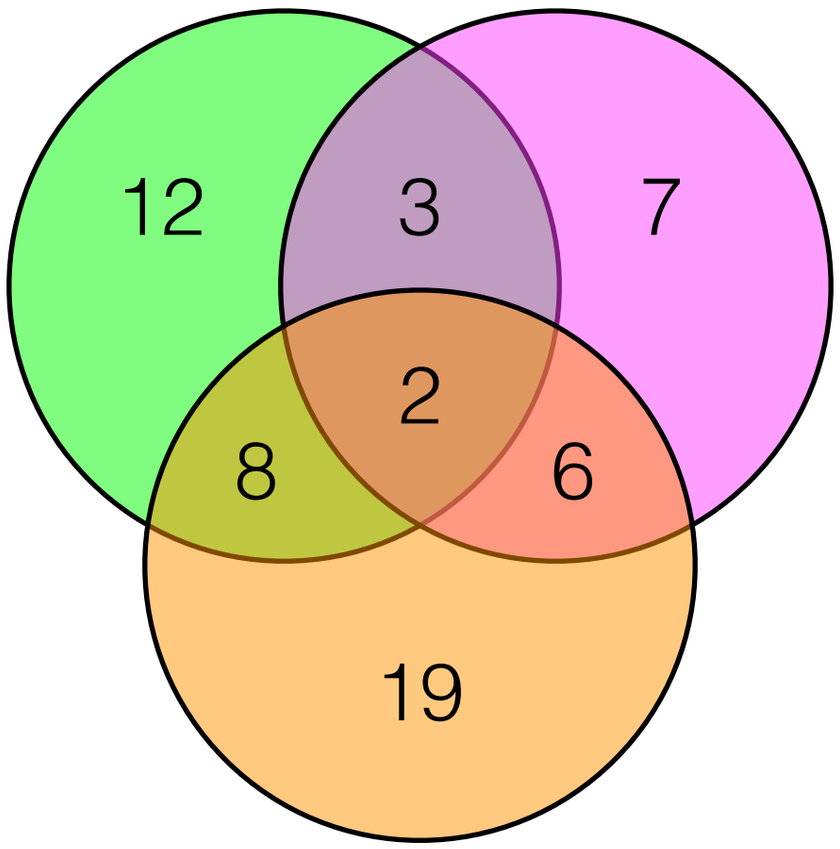
Rarefaction curves to estimate total diversity



Durand, A., *et al*. (2015). *Scientific Reports.*

Measures of diversity among samples

- β-Diversity: between-sample taxonomic diversity
  - Simply look at OTU overlap between samples
  - Bray-Curtis Dissimilarity = $1 - 2*C/(S_i + S_j)$
    - S_i & S_j are the sum of OTUs at sites *i* and *j*
    - *C* is the number of OTUs in common between S_i & S_j
  - UniFrac = similar idea but take into account phylogenetic breadth



Measures of diversity among samples

	Sample 1	Sample 2	Sample 3
OTU1	0.0	**0.1**	**0.1**
OTU2	**0.4**	**0.2**	0.0
OTU3	**0.2**	**0.1**	**0.8**
OTU4	0.0	**0.4**	0.0
OTU5	**0.2**	**0.2**	**0.1**

Bray-Curtis Dissimilarity

	S1	S2	S3
S1	0		
S2		0	
S3			0

Measures of diversity among samples

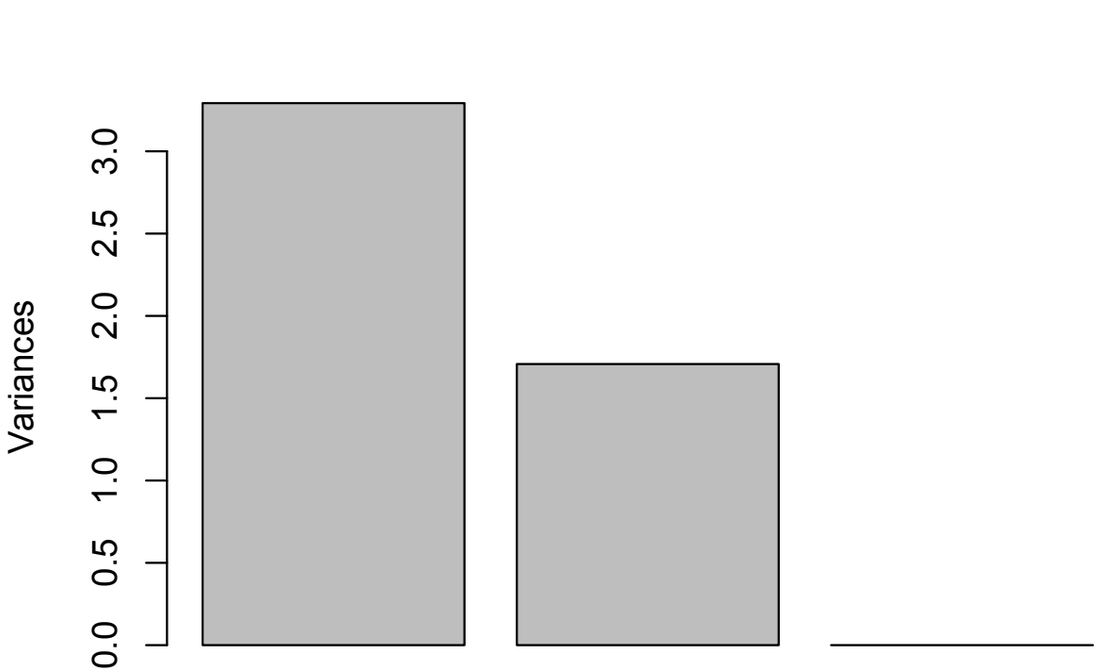
	Sample 1	Sample 2	Sample 3
OTU1	0.0	**0.1**	**0.1**
OTU2	**0.4**	**0.2**	0.0
OTU3	**0.2**	**0.1**	**0.8**
OTU4	0.0	**0.4**	0.0
OTU5	**0.2**	**0.2**	**0.1**

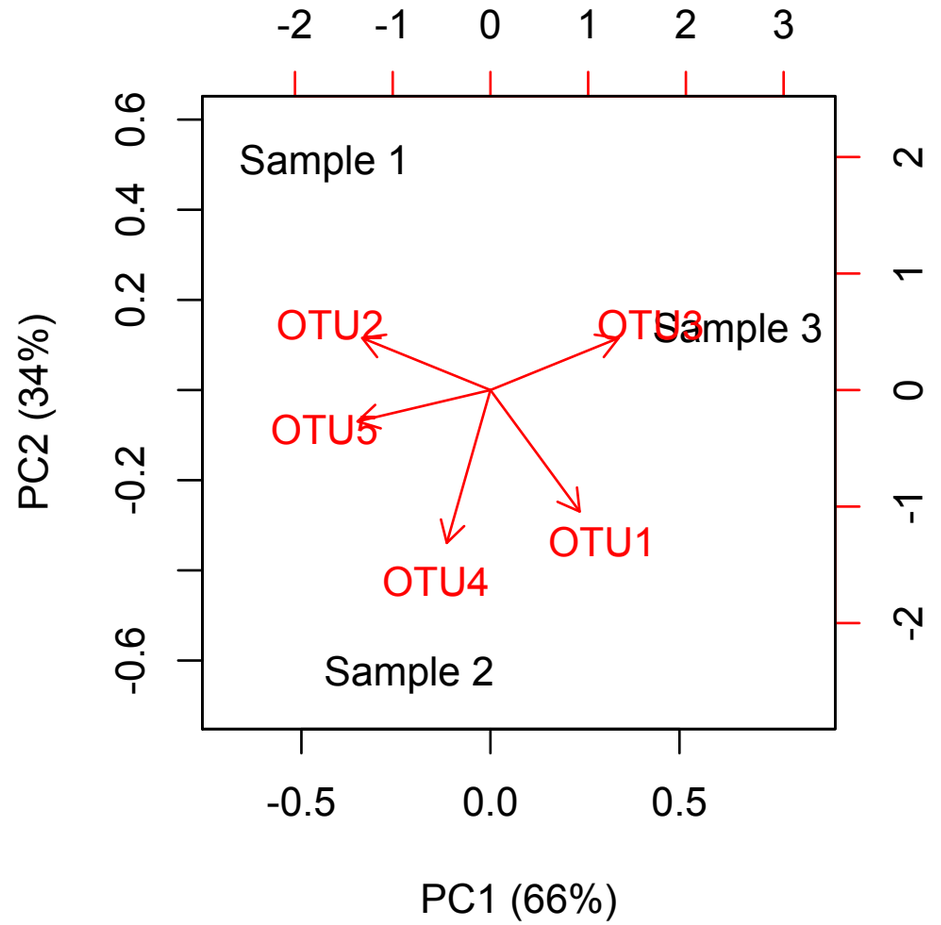
Bray-Curtis Dissimilarity

	S1	S2	S3
S1	0		
S2	0.25	0	
S3	0.33	0.25	0

Principle components analysis

	PC1	PC2	PC3
OTU1	0.3630540	-0.5757433	0.19782508
OTU2	-0.5217308	0.2465473	0.72185580
OTU3	0.5217308	0.2465473	0.01140291
OTU4	-0.1775562	-0.7244540	0.22067749
OTU5	-0.5406103	-0.1487108	-0.62526863





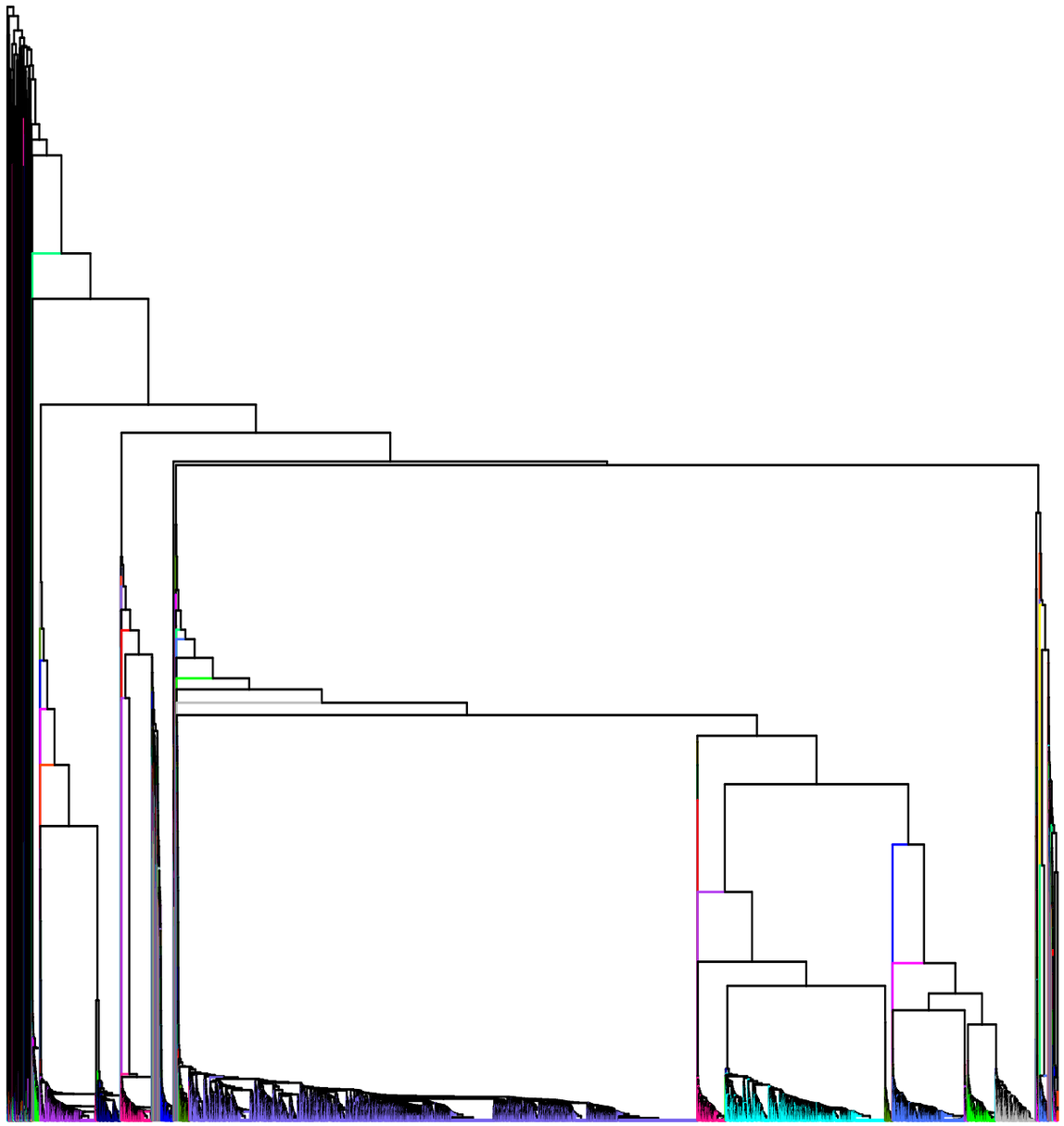
Pitfalls and caveats in microbiome analyses

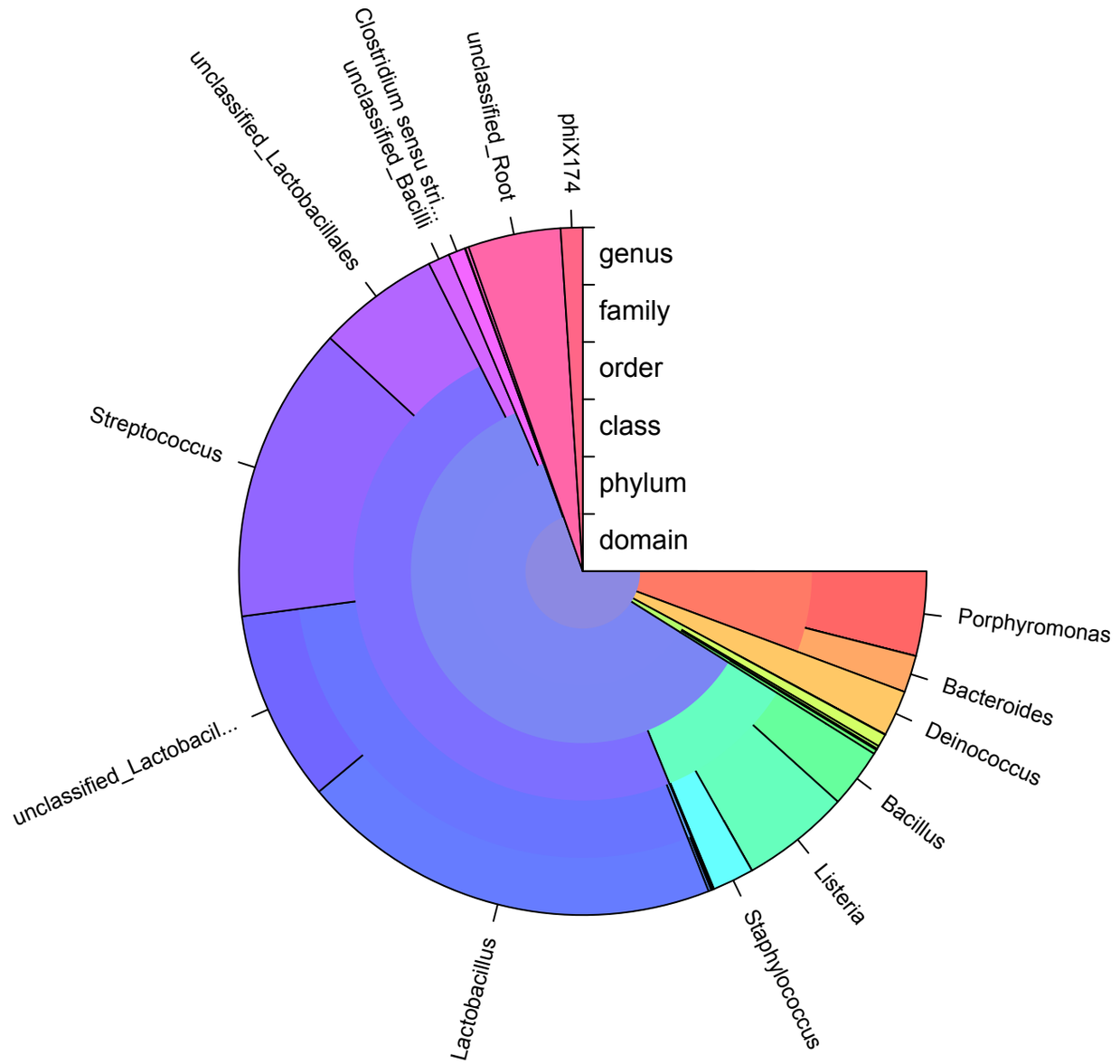
What do you think might be some limitations to 16S sequence surveys of the microbiome?

Potential experimental challenges in 16S studies

- "Kitome"
- PCR artifacts
  - Primer-dimer
  - Chimeras
- Biases:
  - DNA extraction
  - Amplification
  - Primers
  - Sequencing
- Multiple copies of 16S gene
- Sequencing error
- Demultiplexing error
- PhiX spike-in
- Limited length of the 16S gene
  - Short amplicon length
- Slow rate of evolution of the 16S gene
- Misses a lot of organisms
  - viruses
  - eukaryotes

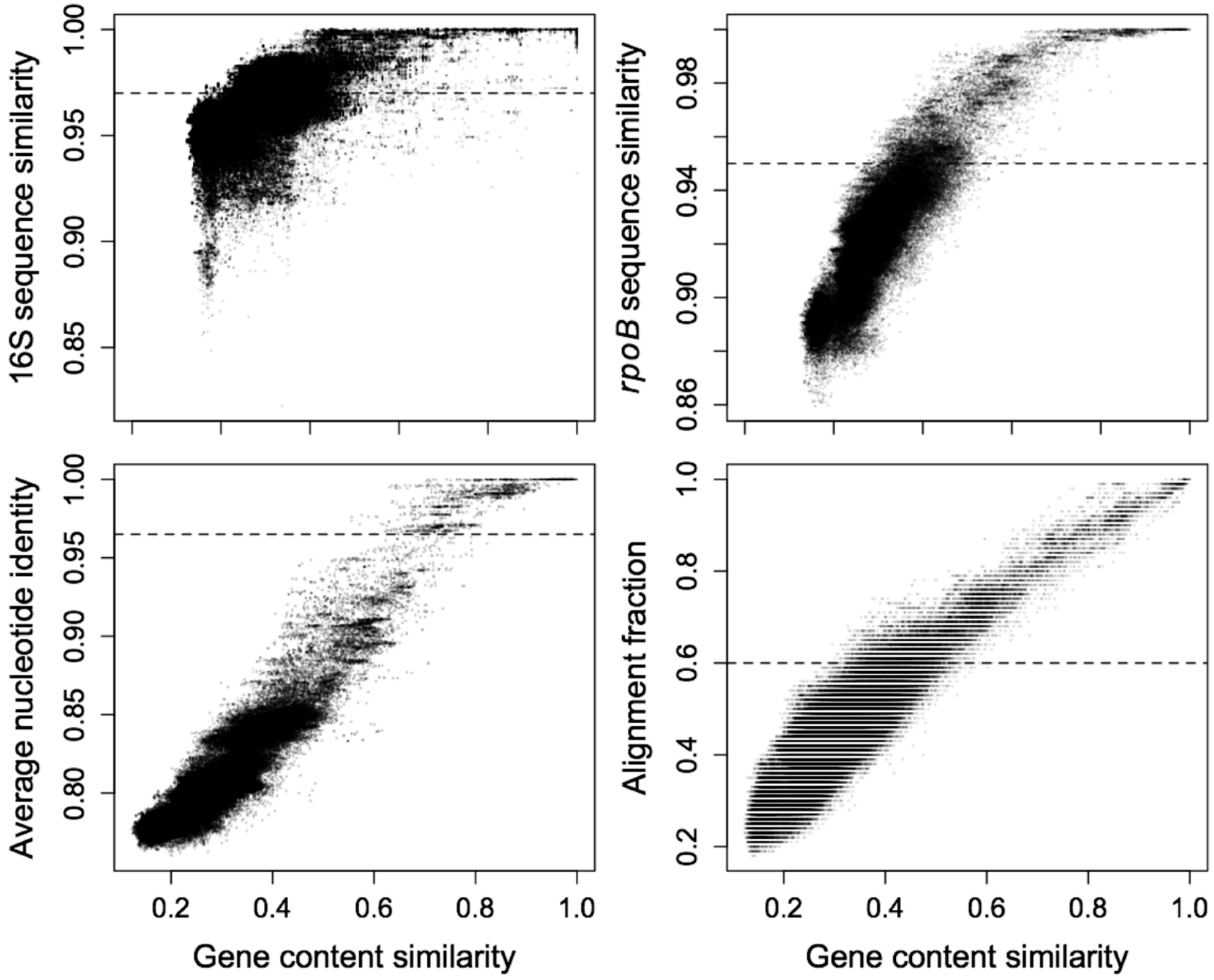
Sequencing a mock community of 20 strains





Fouhy, F., *et al*. (2016). *BMC Microbiology*. Accession: SRR3225706.

16S is weakly correlated with gene content



Potential experimental challenges in 16S studies

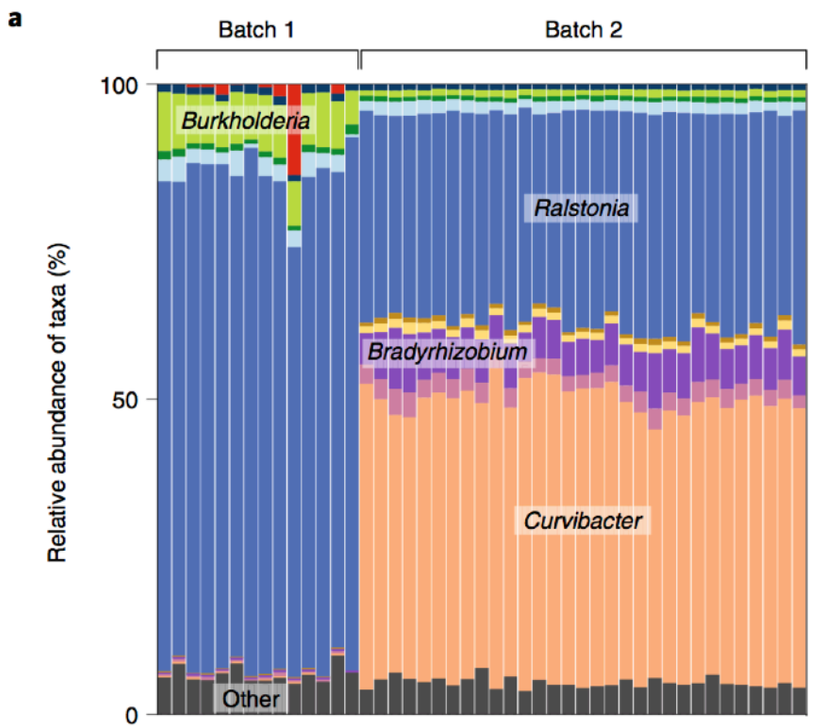
- "Kitome"
- PCR artifacts
  - Primer-dimer
  - Chimeras
- Biases:
  - DNA extraction
  - Amplification
  - Primers
  - Sequencing
- Multiple copies of 16S gene
- Sequencing error
- Demultiplexing error
- PhiX spike-in
- Limited length of the 16S gene
  - Short amplicon length
- Slow rate of evolution of the 16S gene
- Misses a lot of organisms
  - viruses
  - eukaryotes

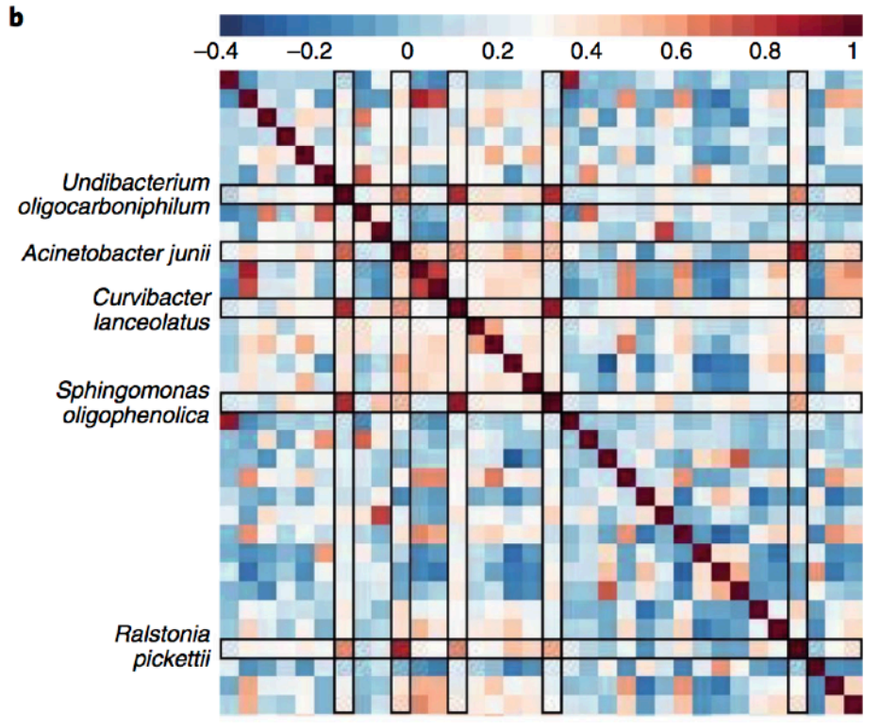
How would you control for each of these error modes?

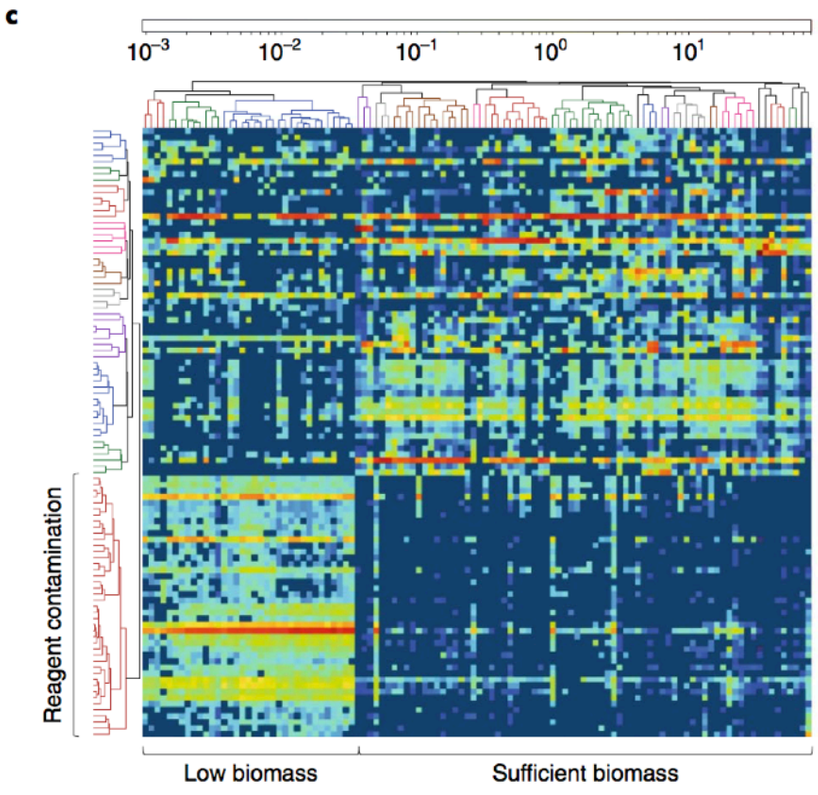
Addressing experimental challenges

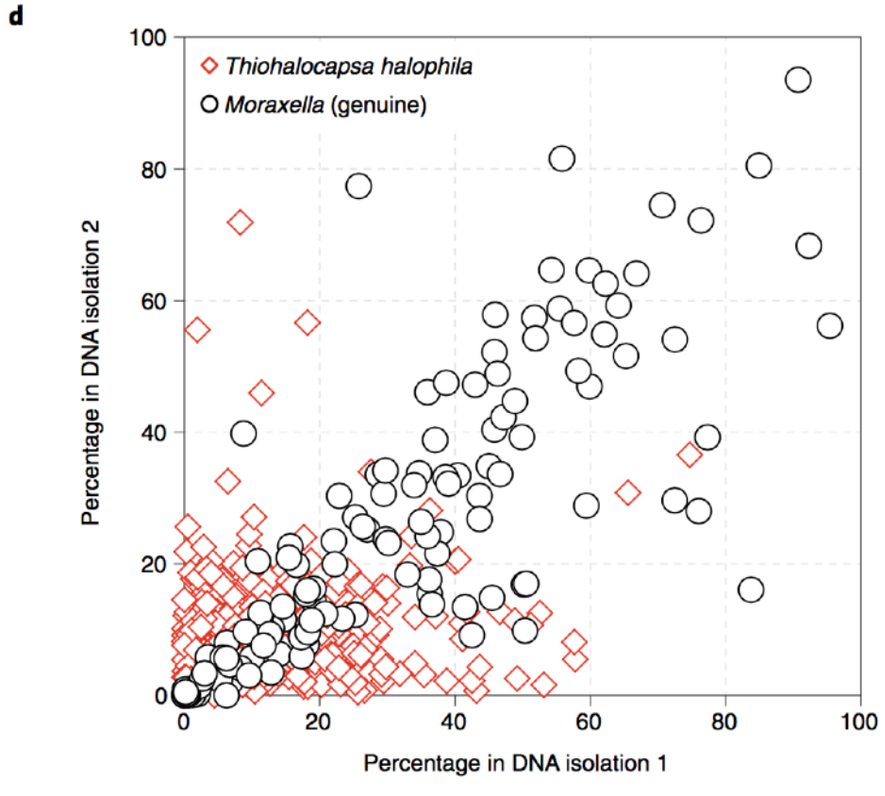
- Negative controls
- Positive controls
  - Mock microbial communities
  - Single strain controls
- Programs for correcting artifacts

Methods for detecting contamination





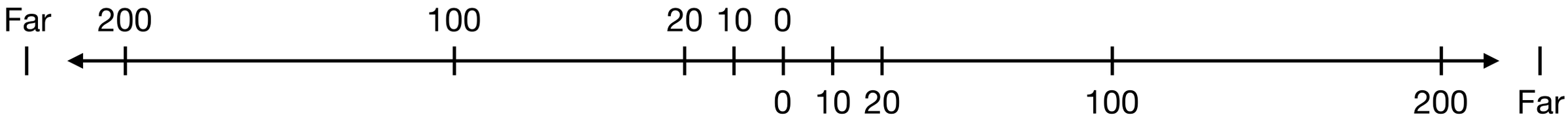


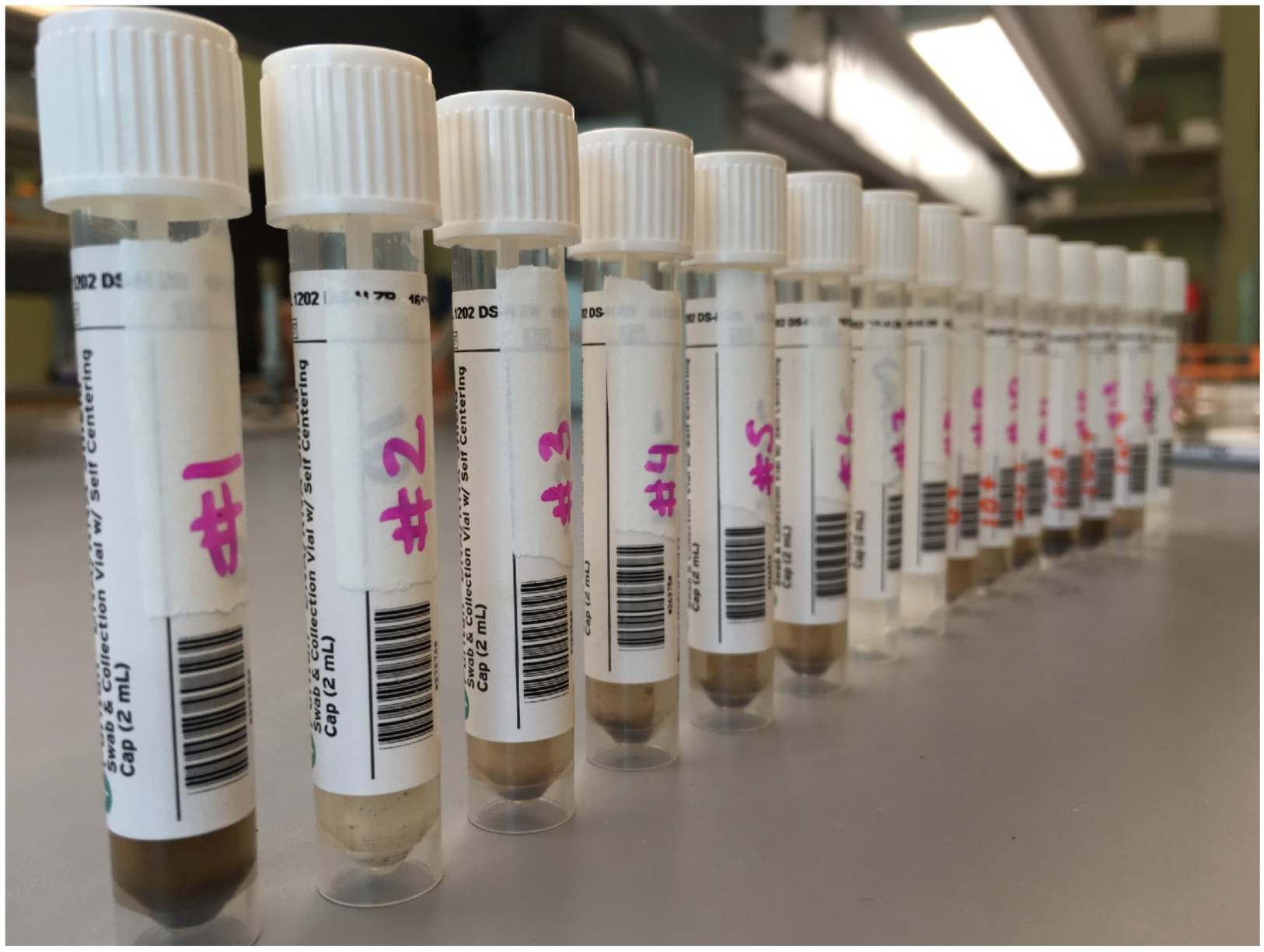


de Goffau, M., *et al.* (2018). *Nature Microbiology*.

Case study: sequencing the soil microbiome

Soil collected at varying distances





Also:

+ Control: *Klebsiella*

- Control: Nothing

20 spp. mock community