

Microbial genomics

Charlotte Soneson
University of Zurich
Brixen 2017

What is the “microbiome”?



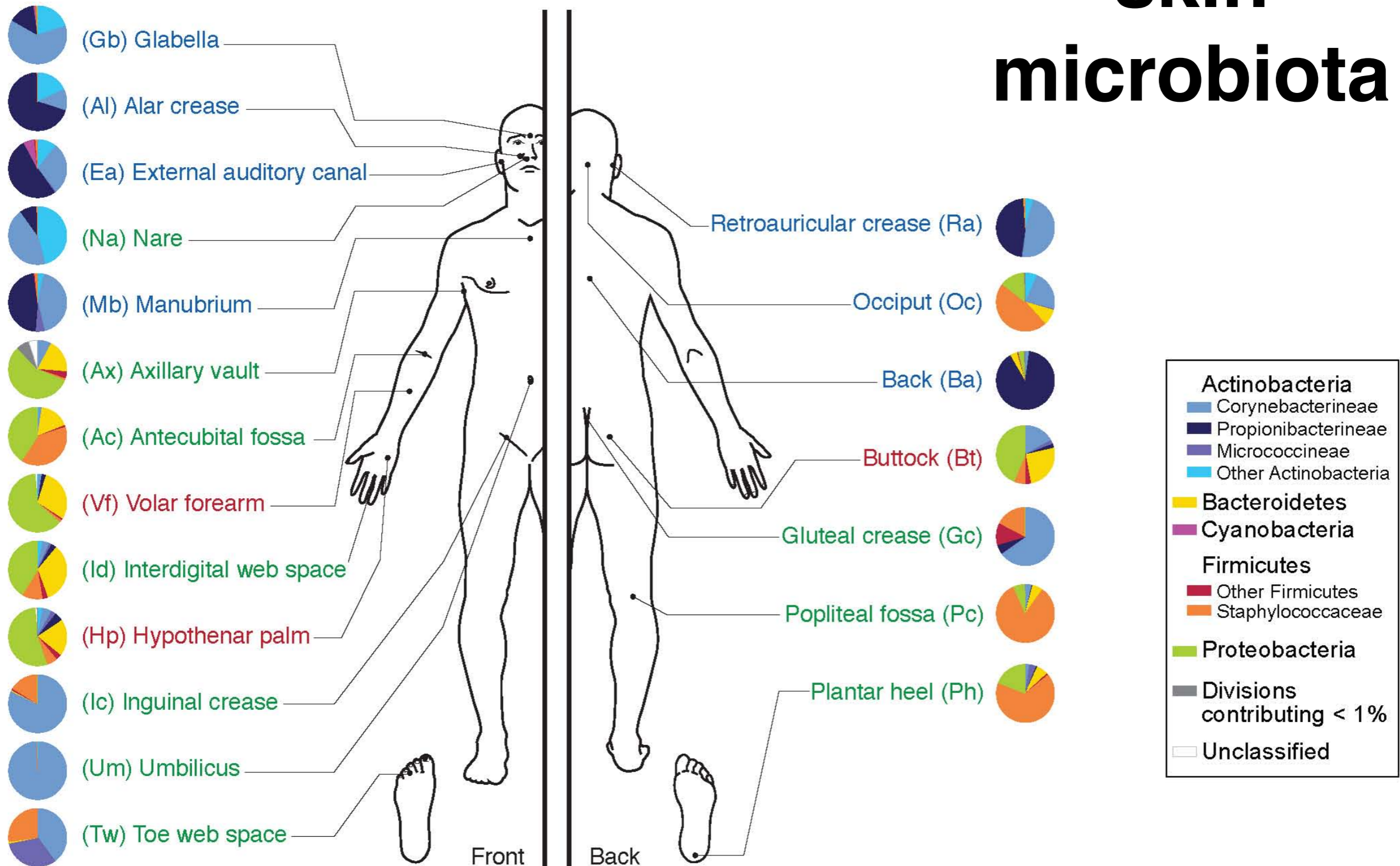
What is the “microbiome”?

- micro**biota** = the assemblage of microorganisms (e.g., bacteria, archaea, viruses, fungi)
- micro**biome** = the ecosystem comprising all microorganisms in an environment, as well as their genes and environmental interactions

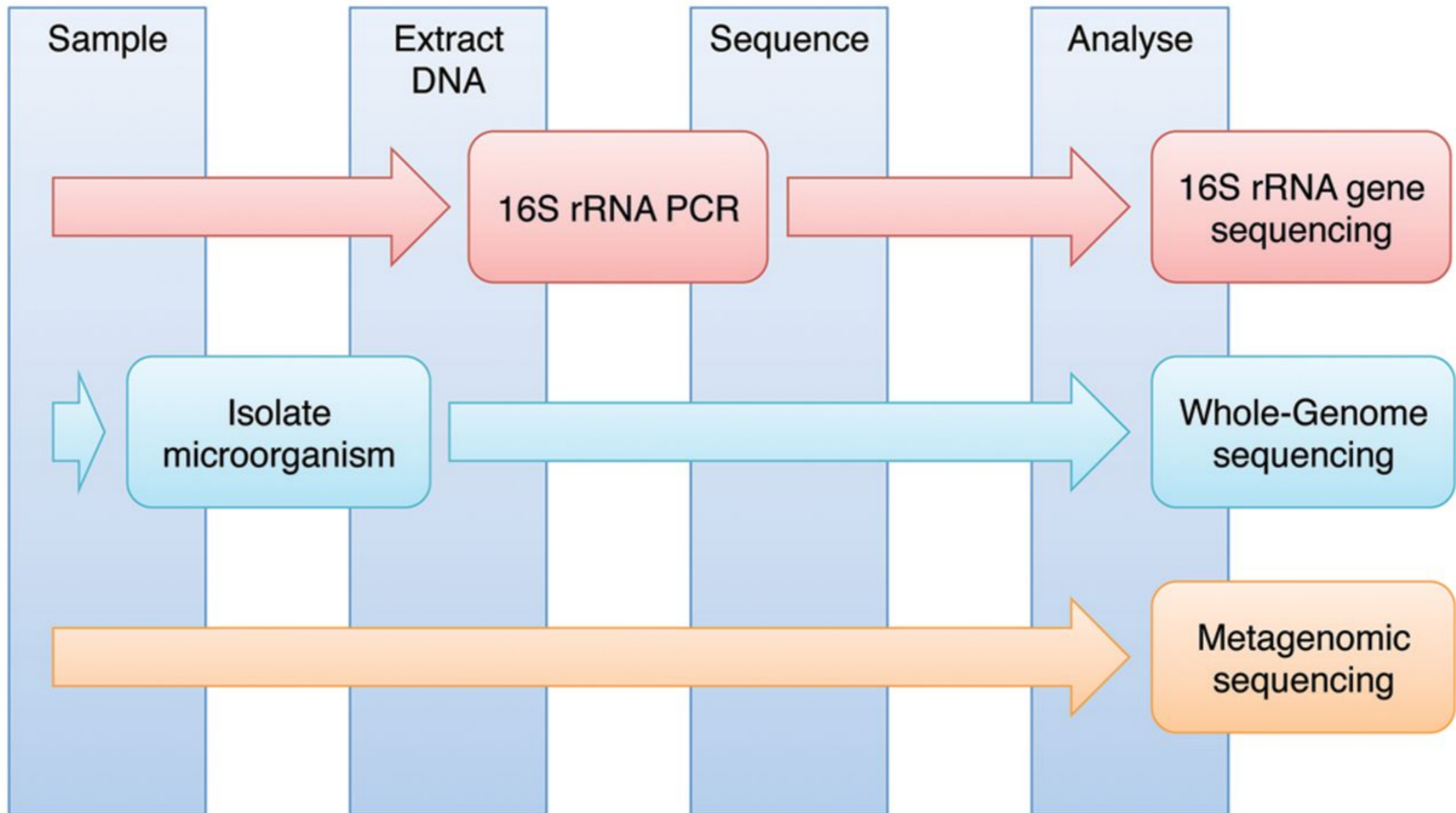
The human microbiota

- “microorganisms that exist upon, within or in close proximity to the human body”
- widely varying composition between body sites and individuals
- important for health: building vitamins, breaking down food etc.
- ratio of microbial to human genes in the body is estimated between 1:1 and 100:1

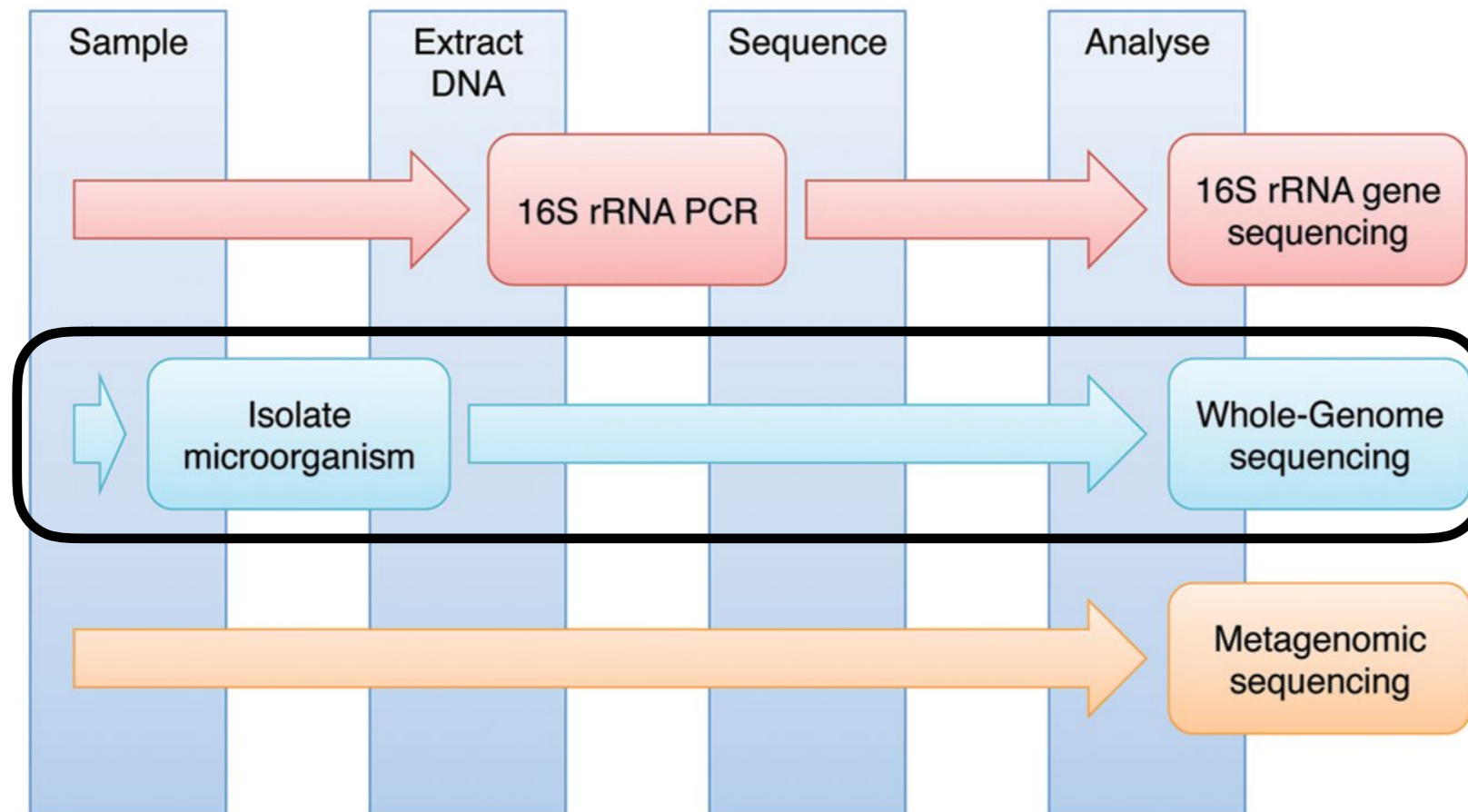
Human skin microbiota



How can we analyze the microbiome?

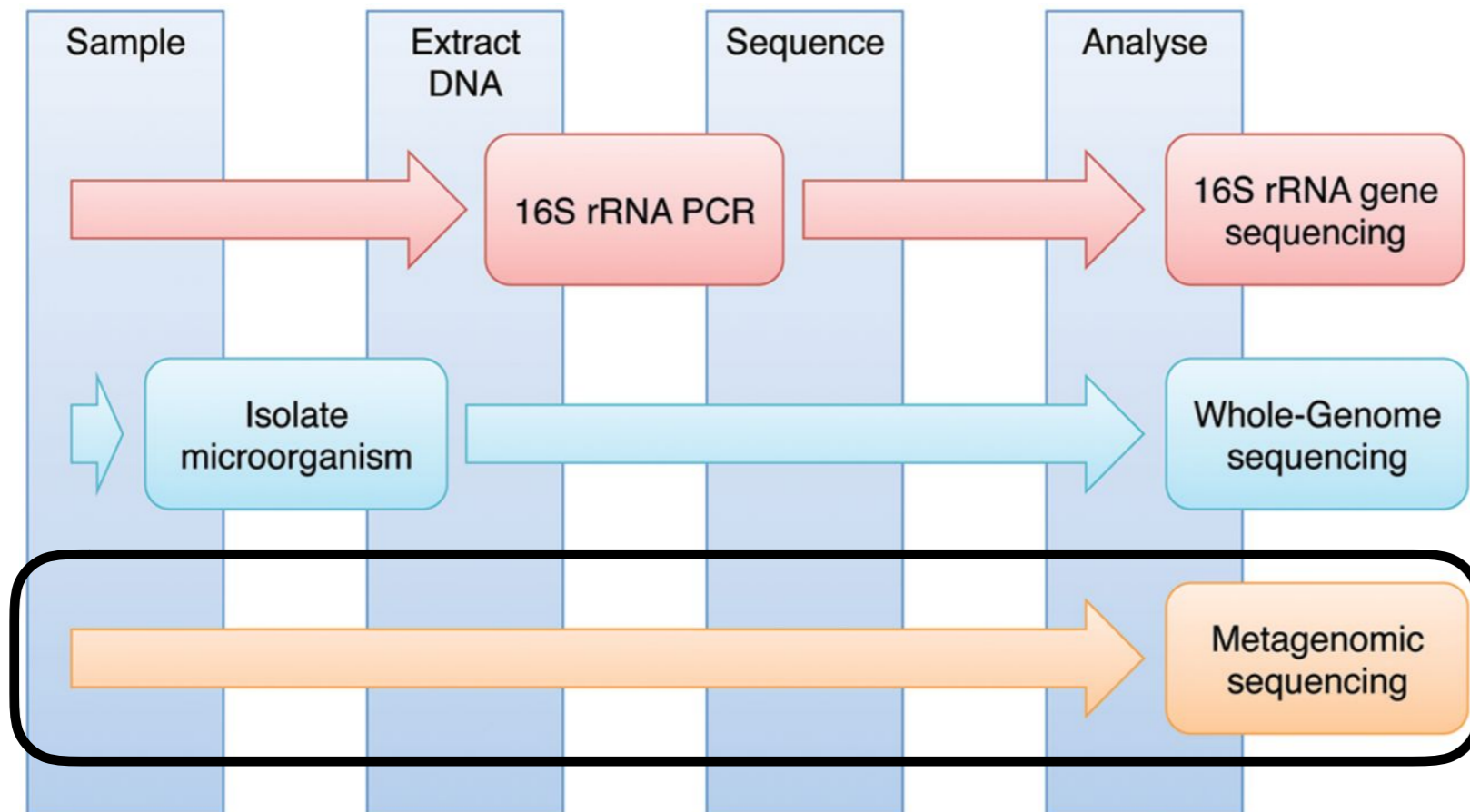


How can we analyze the microbiome?



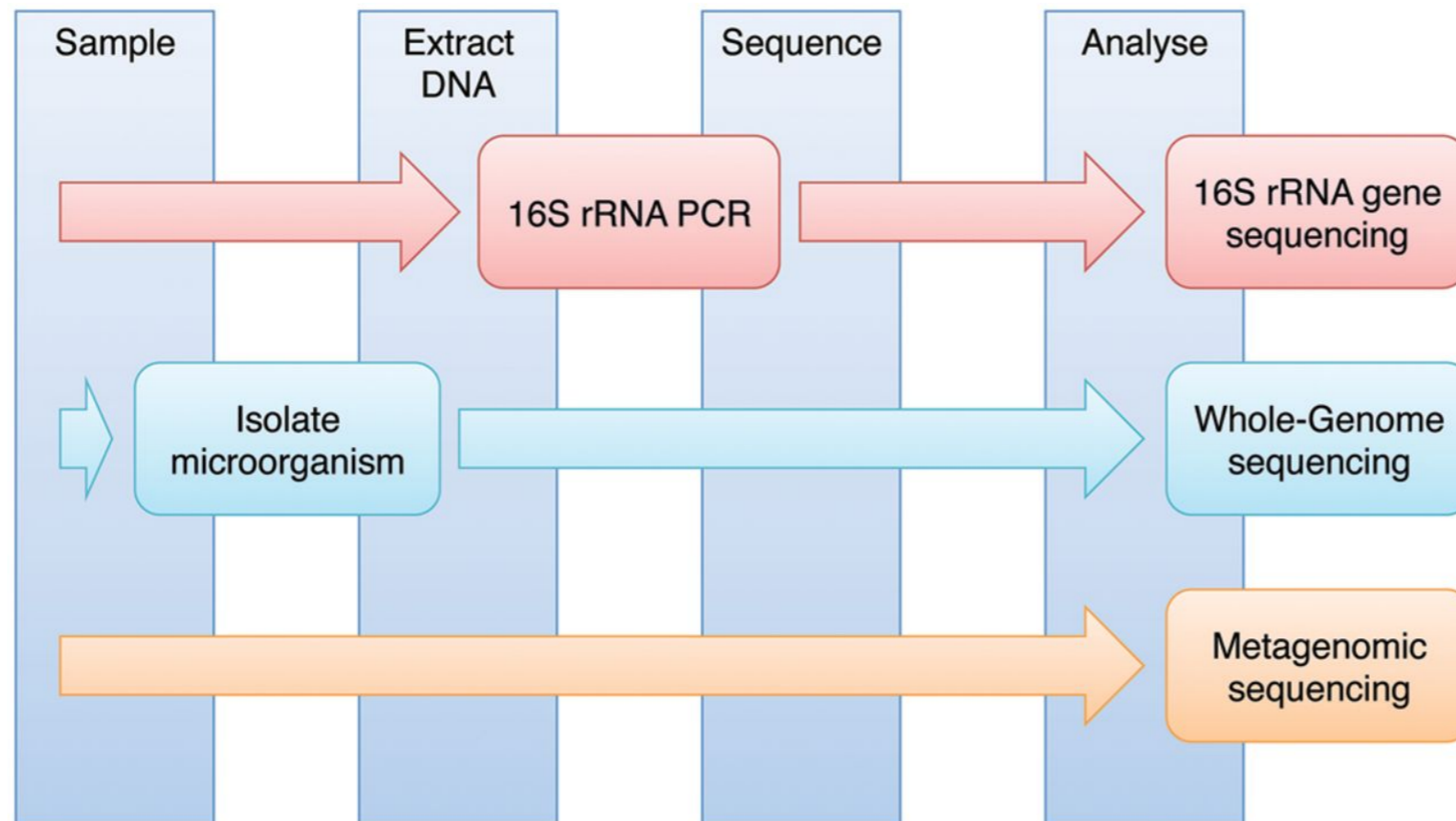
- Whole-genome sequencing: **characterize specific isolate**

How can we analyze the microbiome?



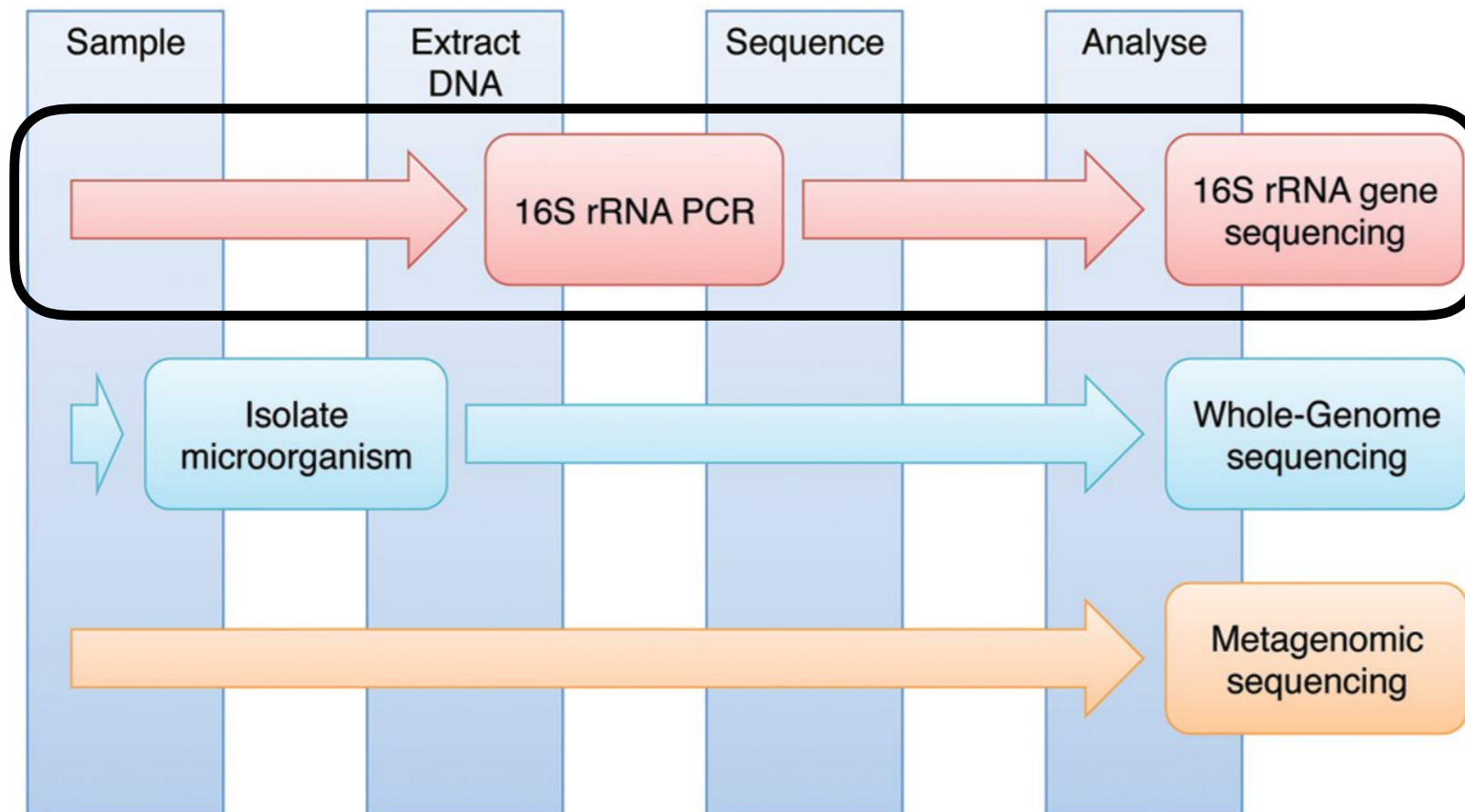
- Metagenomic (shotgun) sequencing: **sequence complete set of DNA in a sample**

How can we analyze the microbiome?



- *metatranscriptomics*
- *metaproteomics*
-

How can we analyze the microbiome?



- [16S] rRNA amplicon/marker gene sequencing:
infer microbial composition

Amplicon sequencing - basic idea

- Amplify (part of) the 16S rRNA gene from all microbes - sequence amplified part
- Cluster sequences together in so called OTUs (operational taxonomic units = clusters of similar sequences ~ “species”)
- Get the number of sequences in each cluster/OTU for each sample
- Generate an abundance table (OTUs x samples)

Amplicon sequencing - basic idea

Which part?

We need primers!

- Amplify (part of) the 16S rRNA gene from all microbes - sequence amplified part
- Cluster sequences together in so called OTUs (operational taxonomic units = clusters of similar sequences ~ “species”)
- Get the number of sequences in each cluster/OTU for each sample
- Generate an abundance table (OTUs x samples)

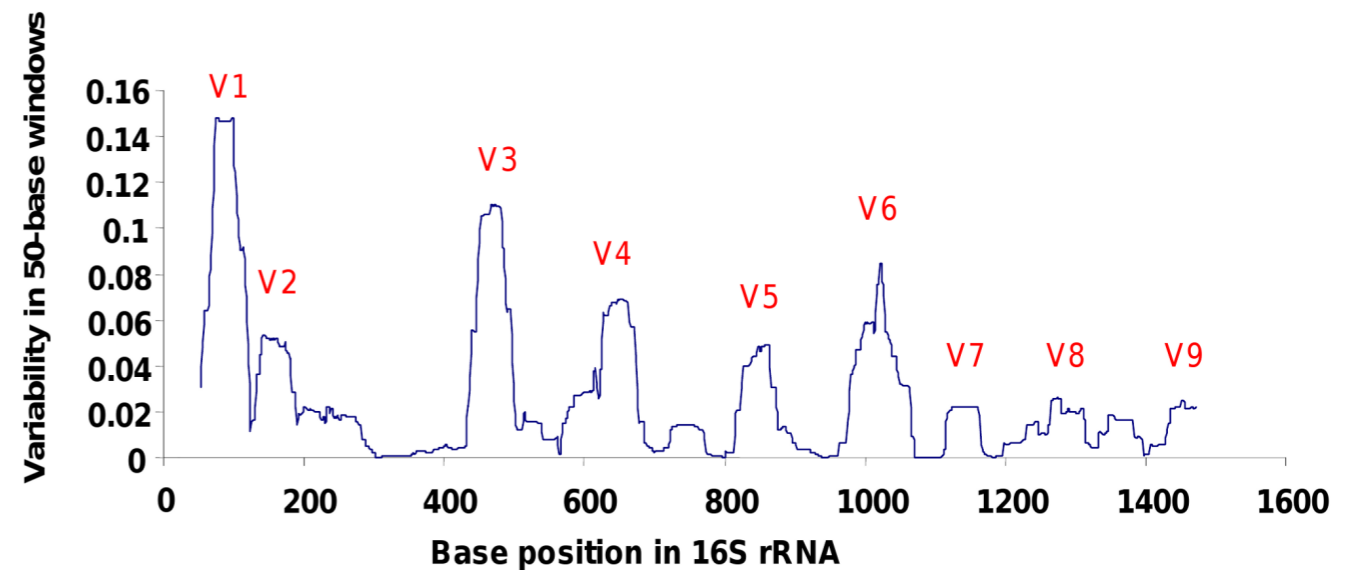
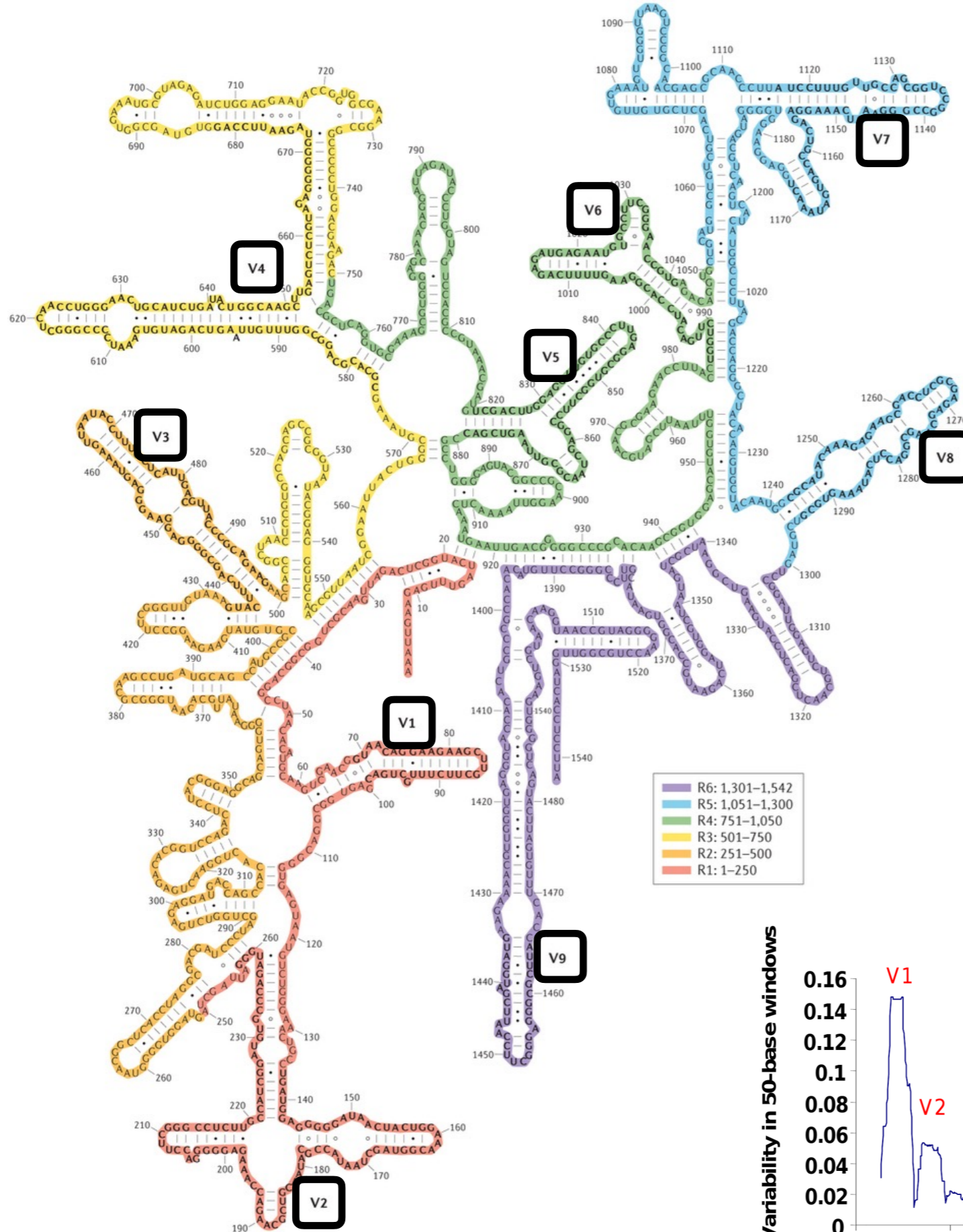
How to cluster?

How to analyze?

Why [16S] rRNA?

- rRNA is one of the few gene products present in all cells
- 16S rRNA has 9 **hypervariable regions** allowing species identification, as well as **conserved regions** allowing primer construction
- conserved function
- sequence has been characterized for many species

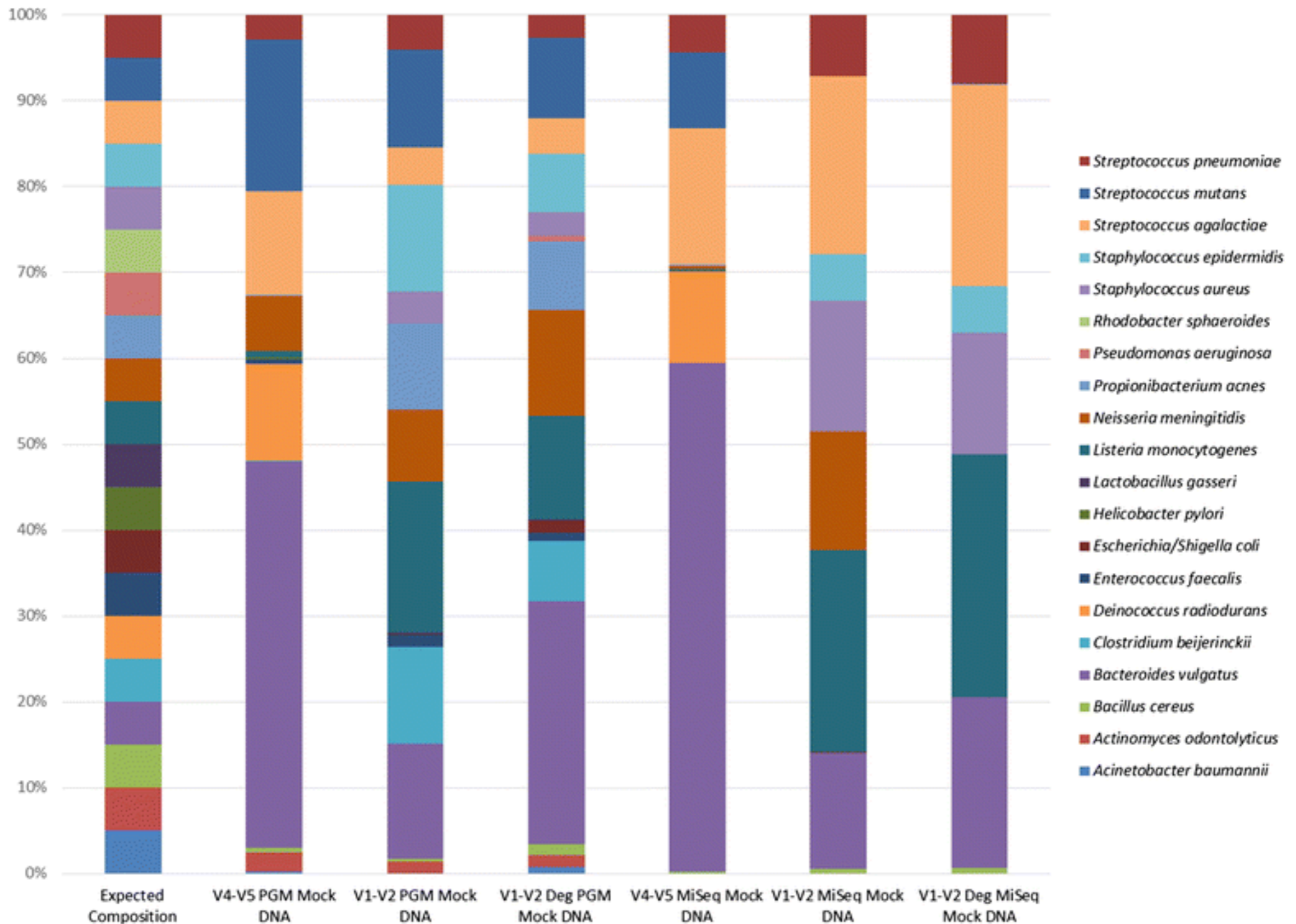
16S rRNA (*E. coli*)



16S is not perfect

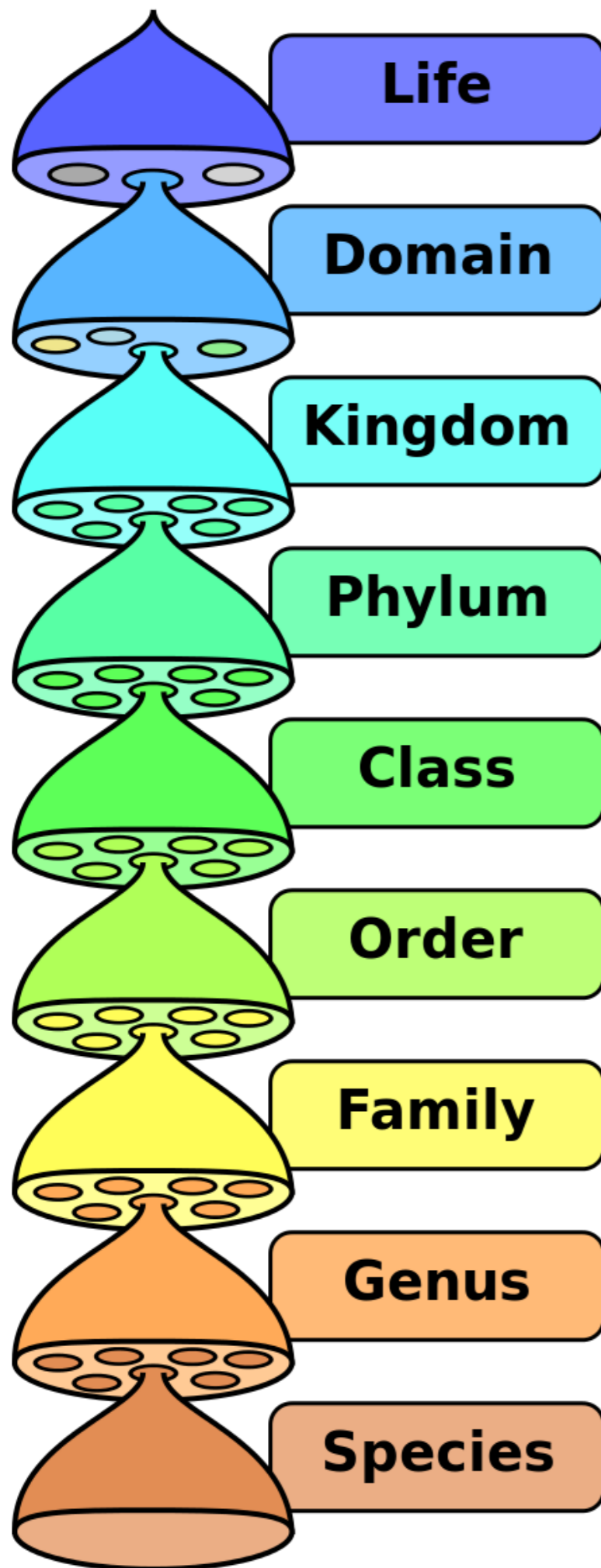
- 16S doesn't capture all differences between the full DNA sequences
- Different species can have similar 16S sequences
- A single species can have paralogs that are not identical
- Results can depend on which variable region is considered, and which sequencer is used

There are still challenges to overcome



OTU generation

- “**closed-reference** clustering”: compare sequences to a reference catalog, group together sequences that are similar to the same references.
- “**distance-based/*de novo*** clustering”: cluster based on pairwise distances among sequences.
- “**open-reference** clustering”: closed-reference clustering followed by *de novo* clustering of unclassified sequences



Which similarity threshold?

- Typical (but arbitrary) similarity threshold: 97% (for species level)
- This means different things depending on the clustering method that was used!

Representation in R - phyloseq object

```
> head(sample_data(GlobalPatterns))
```

```
Sample Data:      [6 samples by 7 sample variables]:
```

	X.SampleID	Primer	Final_Barcode	Barcode_truncated_plus_T	Barcode_full_length	SampleType
CL3	CL3	ILBC_01	AACGCA	TGCGTT	CTAGCGTGCGT	Soil
CC1	CC1	ILBC_02	AACTCG	CGAGTT	CATCGACGAGT	Soil
SV1	SV1	ILBC_03	AACTGT	ACAGTT	GTACGCACAGT	Soil
M31Fcsw	M31Fcsw	ILBC_04	AAGAGA	TCTCTT	TCGACATCTCT	Feces
M11Fcsw	M11Fcsw	ILBC_05	AAGCTG	CAGCTT	CGACTGCAGCT	Feces
M31Plmr	M31Plmr	ILBC_07	AATCGT	ACGATT	CGAGTCACGAT	Skin

Description

CL3	Calhoun South Carolina Pine soil, pH 4.9
CC1	Cedar Creek Minnesota, grassland, pH 6.1
SV1	Sevilleta new Mexico, desert scrub, pH 8.3
M31Fcsw	M3, Day 1, fecal swab, whole body study
M11Fcsw	M1, Day 1, fecal swab, whole body study
M31Plmr	M3, Day 1, right palm, whole body study

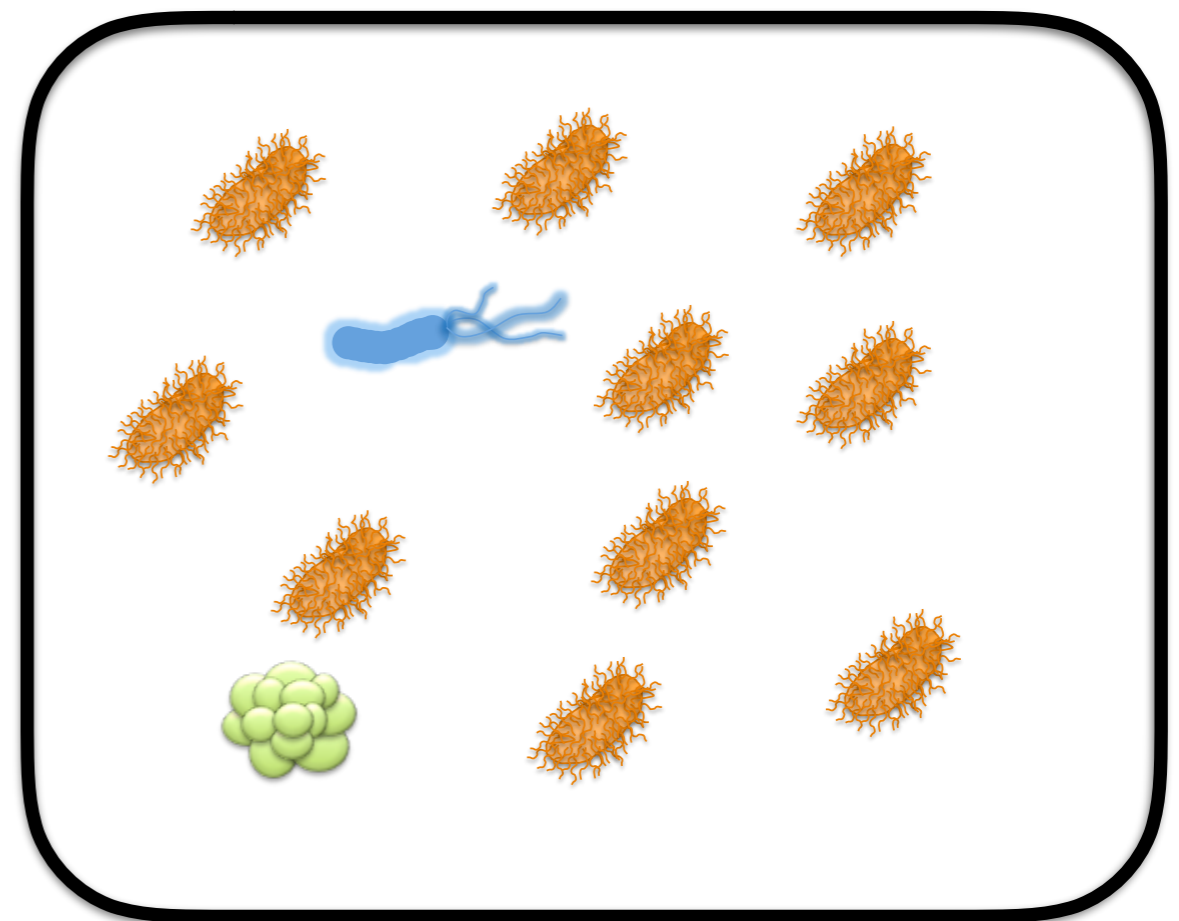
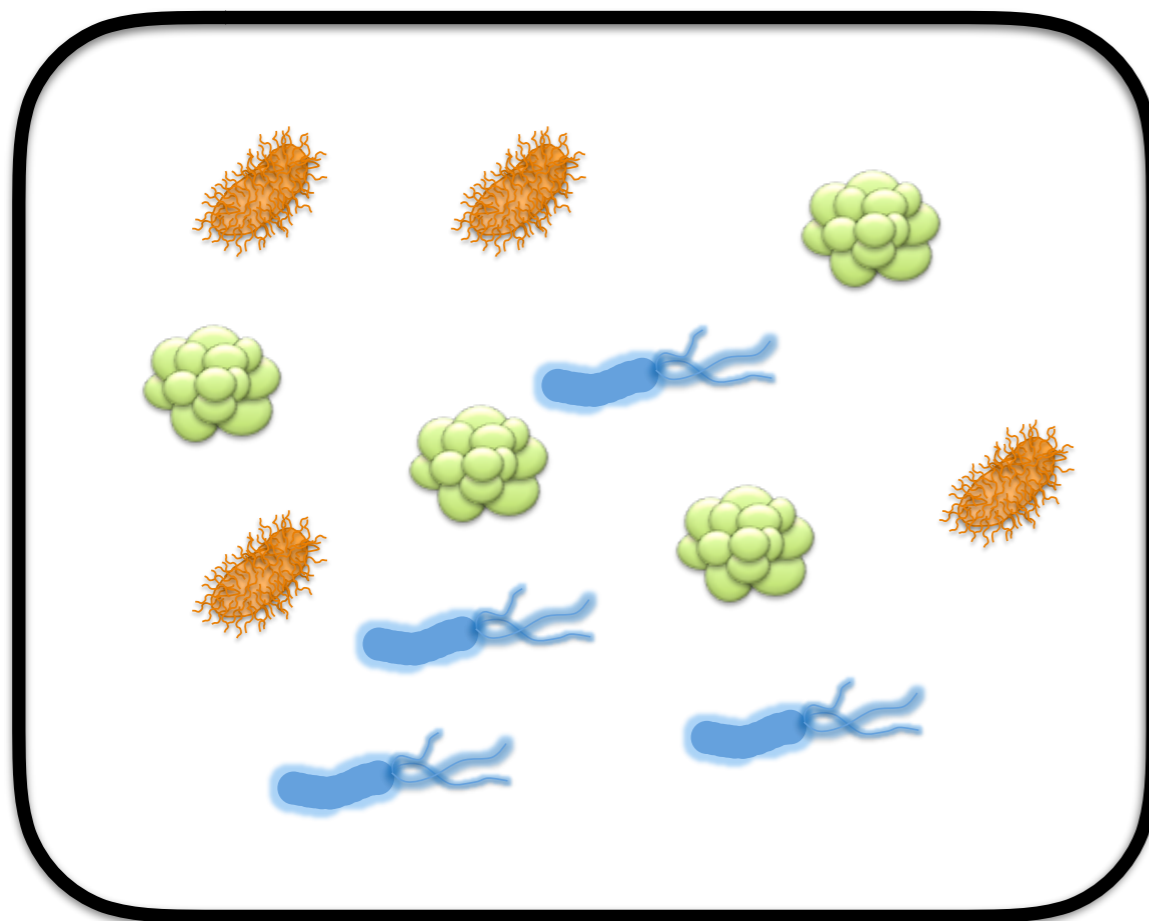
```
> head(tax_table(GlobalPatterns))
```

```
Taxonomy Table:      [6 taxa by 7 taxonomic ranks]:
```

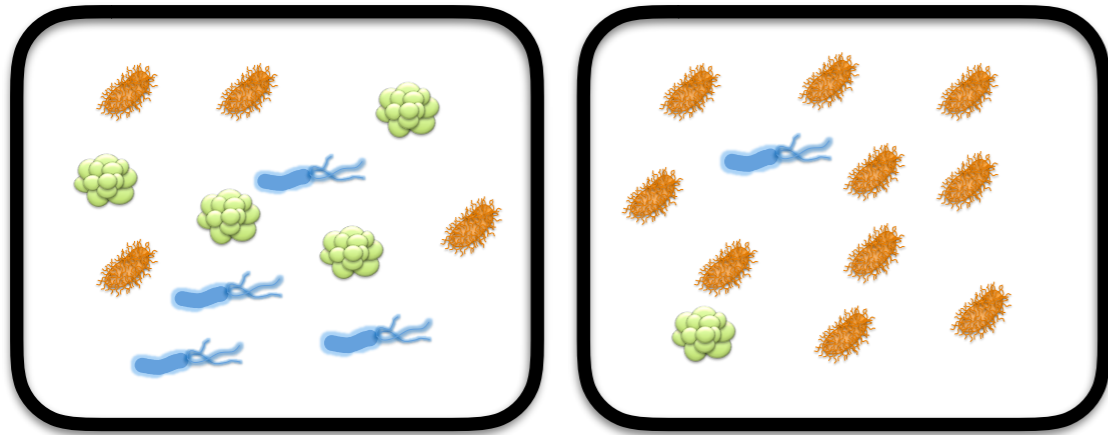
	Kingdom	Phylum	Class	Order	Family	Genus	Species
549322	"Archaea"	"Crenarchaeota"	"Thermoprotei"	NA	NA	NA	NA
522457	"Archaea"	"Crenarchaeota"	"Thermoprotei"	NA	NA	NA	NA
951	"Archaea"	"Crenarchaeota"	"Thermoprotei"	"Sulfolobales"	"Sulfolobaceae"	"Sulfolobus"	"Sulfolobusacidocaldarius"
244423	"Archaea"	"Crenarchaeota"	"Sd-NA"	NA	NA	NA	NA
586076	"Archaea"	"Crenarchaeota"	"Sd-NA"	NA	NA	NA	NA
246140	"Archaea"	"Crenarchaeota"	"Sd-NA"	NA	NA	NA	NA

Richness and alpha diversity

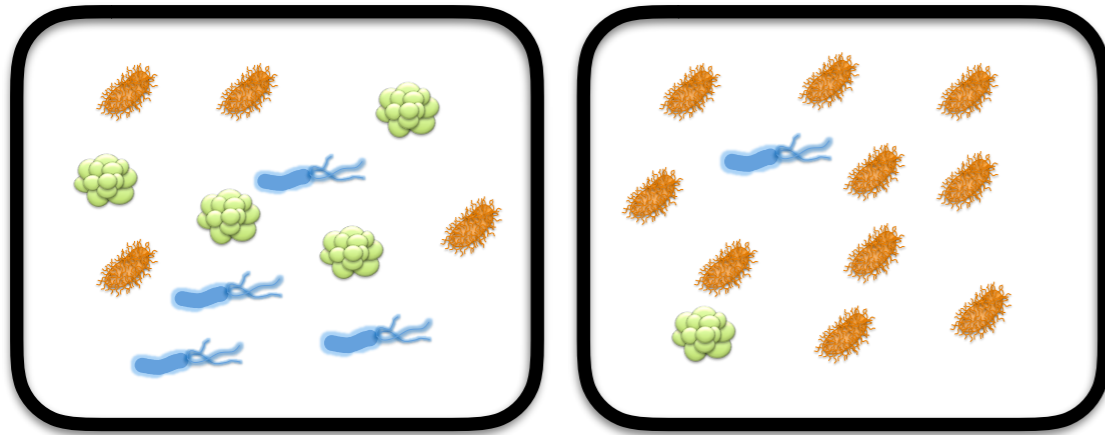
- richness = number of species observed in a sample
- alpha diversity ~ diversity (“unevenness”) of species abundances within a sample



Richness and alpha diversity



Richness and alpha diversity



```
> phx
```

```
OTU Table: [3 taxa and 2 samples]  
taxa are rows
```

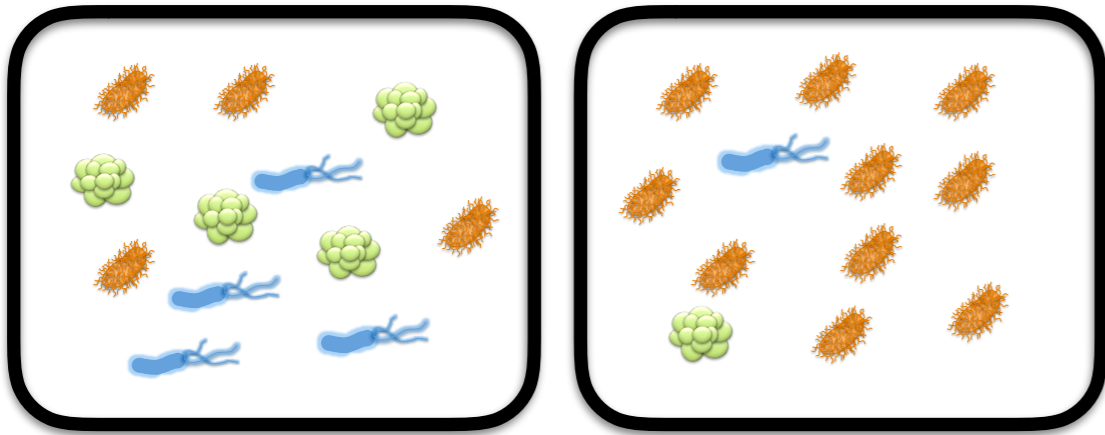
```
      sample1 sample2  
OTU1      4      10  
OTU2      4       1  
OTU3      4       1
```

```
>
```

```
> estimate_richness(phx, measures = "Observed")
```

```
      Observed  
sample1      3  
sample2      3
```

Richness and **alpha diversity**

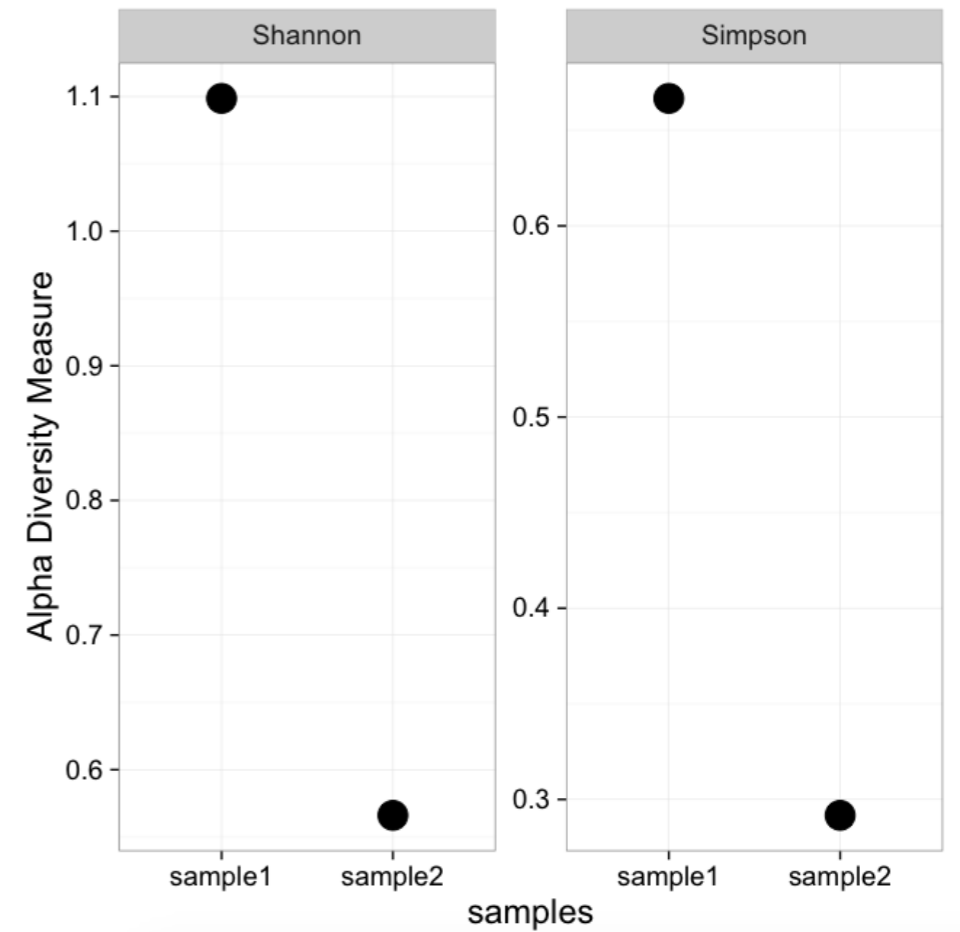


```
> estimate_richness(phx, measures = c("Shannon", "Simpson"))
      Shannon Simpson
sample1 1.0986123 0.6666667
sample2 0.5660857 0.2916667
>
> plot_richness(phx, measures = c("Shannon", "Simpson")) +
+   theme_bw(base_size = 18) + geom_point(size = 7)
```

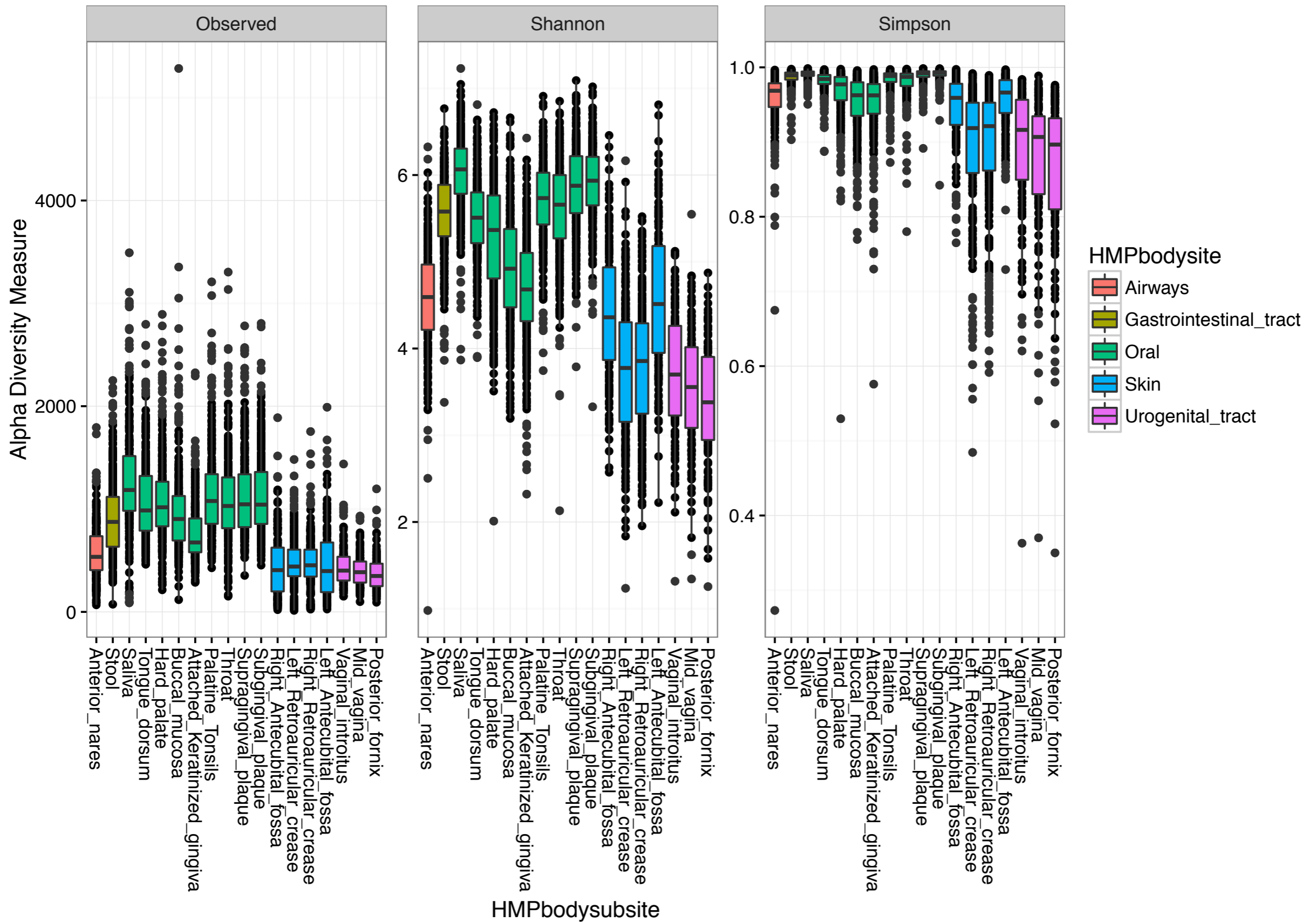
$$\text{Shannon: } H = - \sum_i p_i \log p_i$$

$$\text{Simpson: } D = 1 - \sum_i p_i^2$$

relative abundance of species i

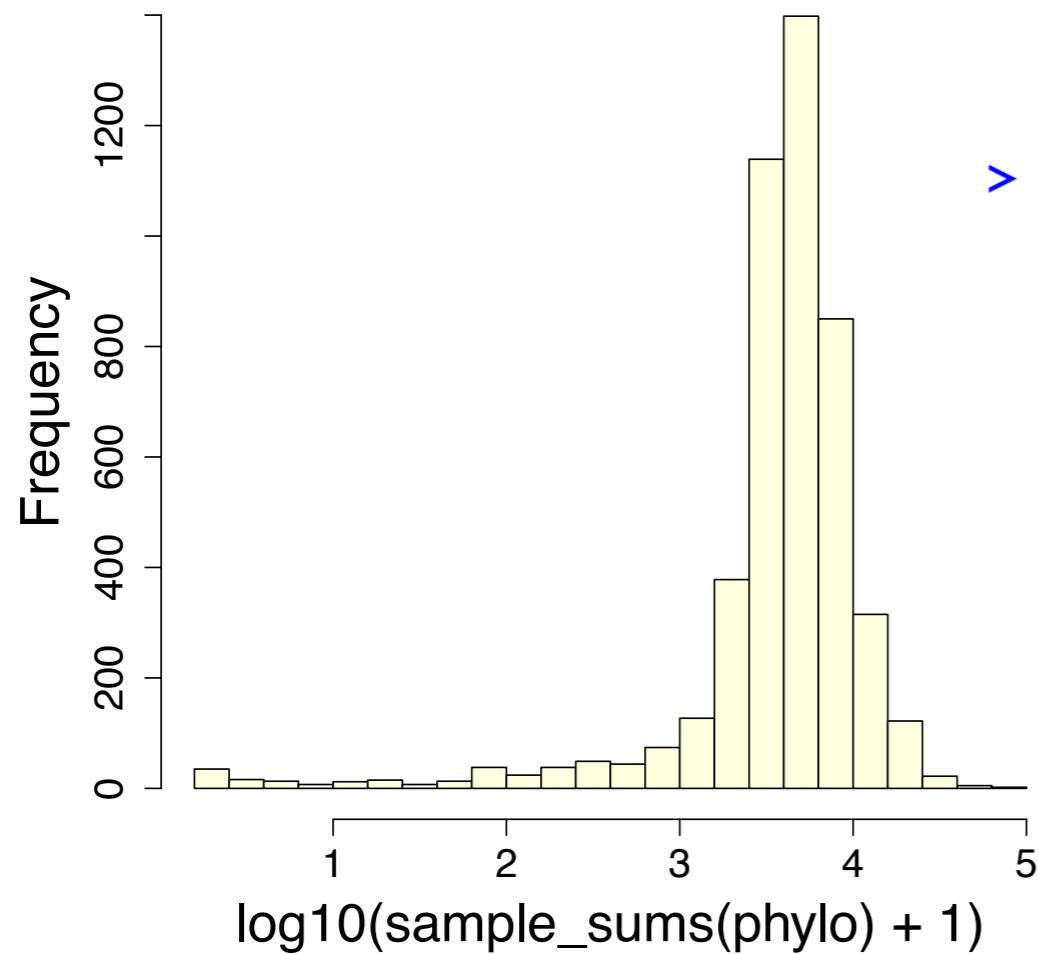


Richness and alpha diversity - HMP data



Normalization

- Library sizes vary greatly between samples



```
> summary(sample_sums(phylo))
```

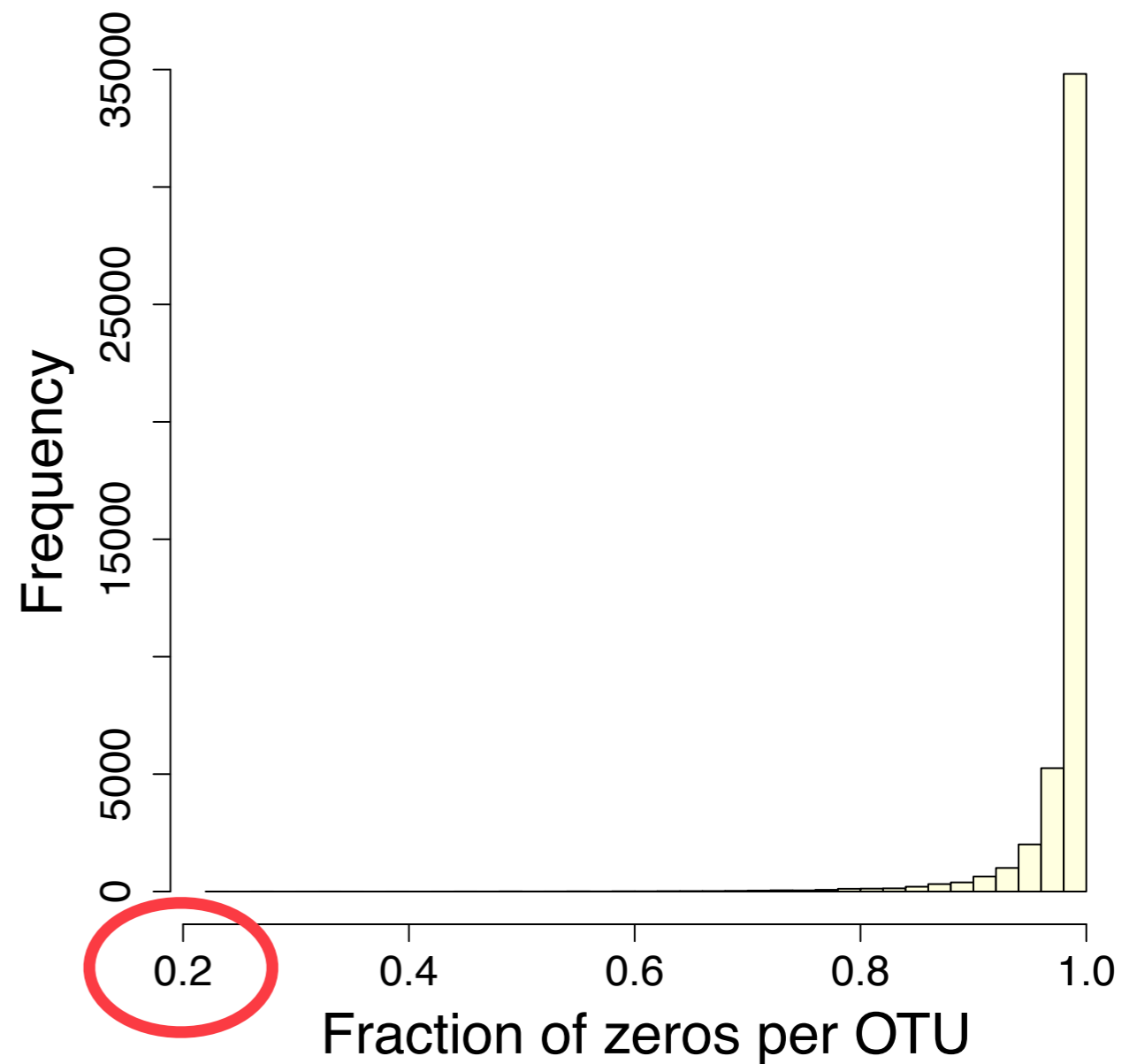
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
105	3054	4562	5604	6818	87840

Normalization

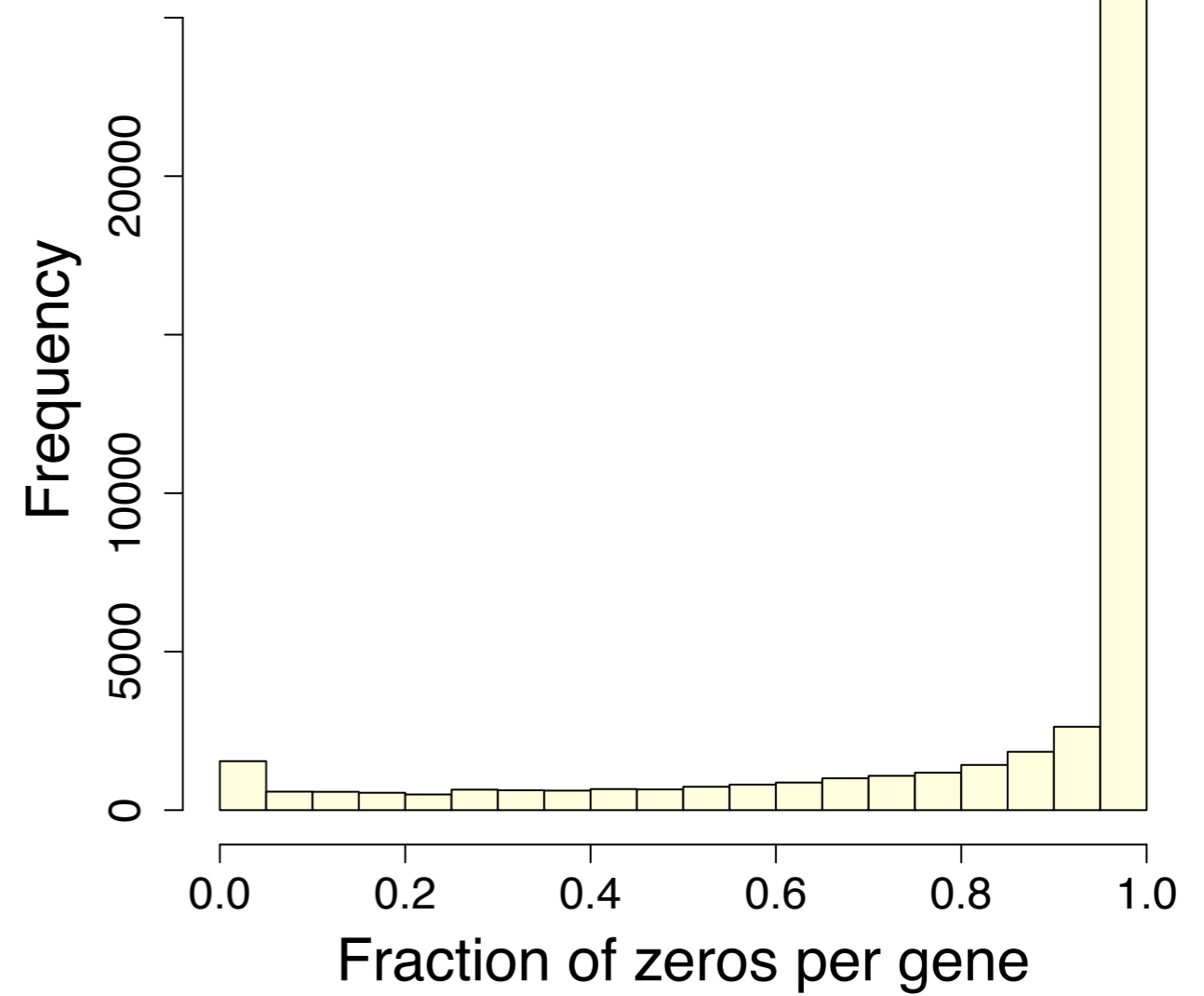
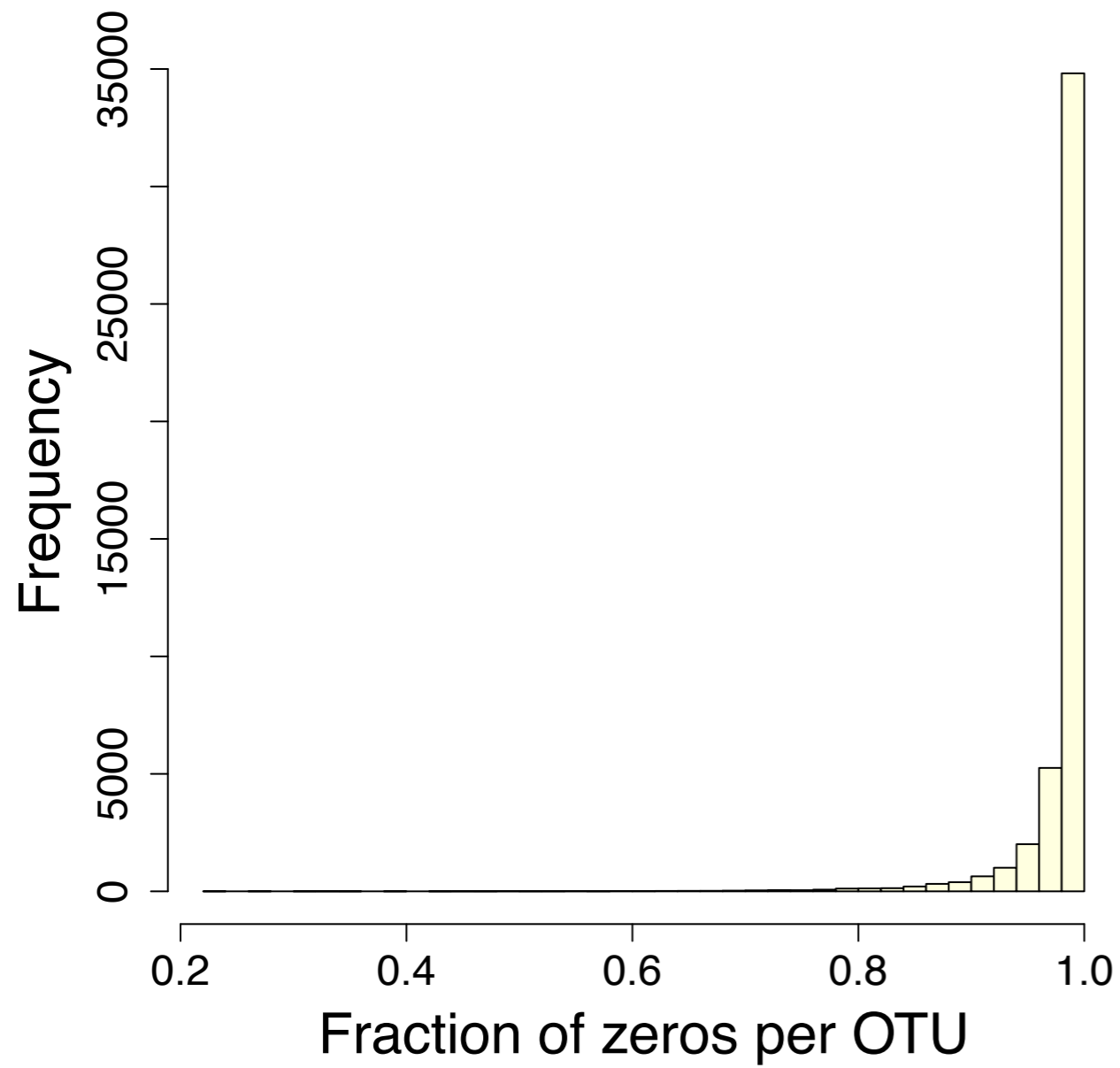
- Library sizes vary greatly between samples
- OTU abundances are often normalized by **rarefying** (subsampling to equal sequencing depth across samples) or by representing them as **relative abundances**.
- Recent studies have suggested using scaling normalization (similar to RNA-seq).

Scaling normalization - challenges

- Lots of zero counts!
- Assumption that “most things don’t change” across samples may not be valid.
- RNA-seq normalization methods require (e.g.) at least one OTU which is observed in all samples.

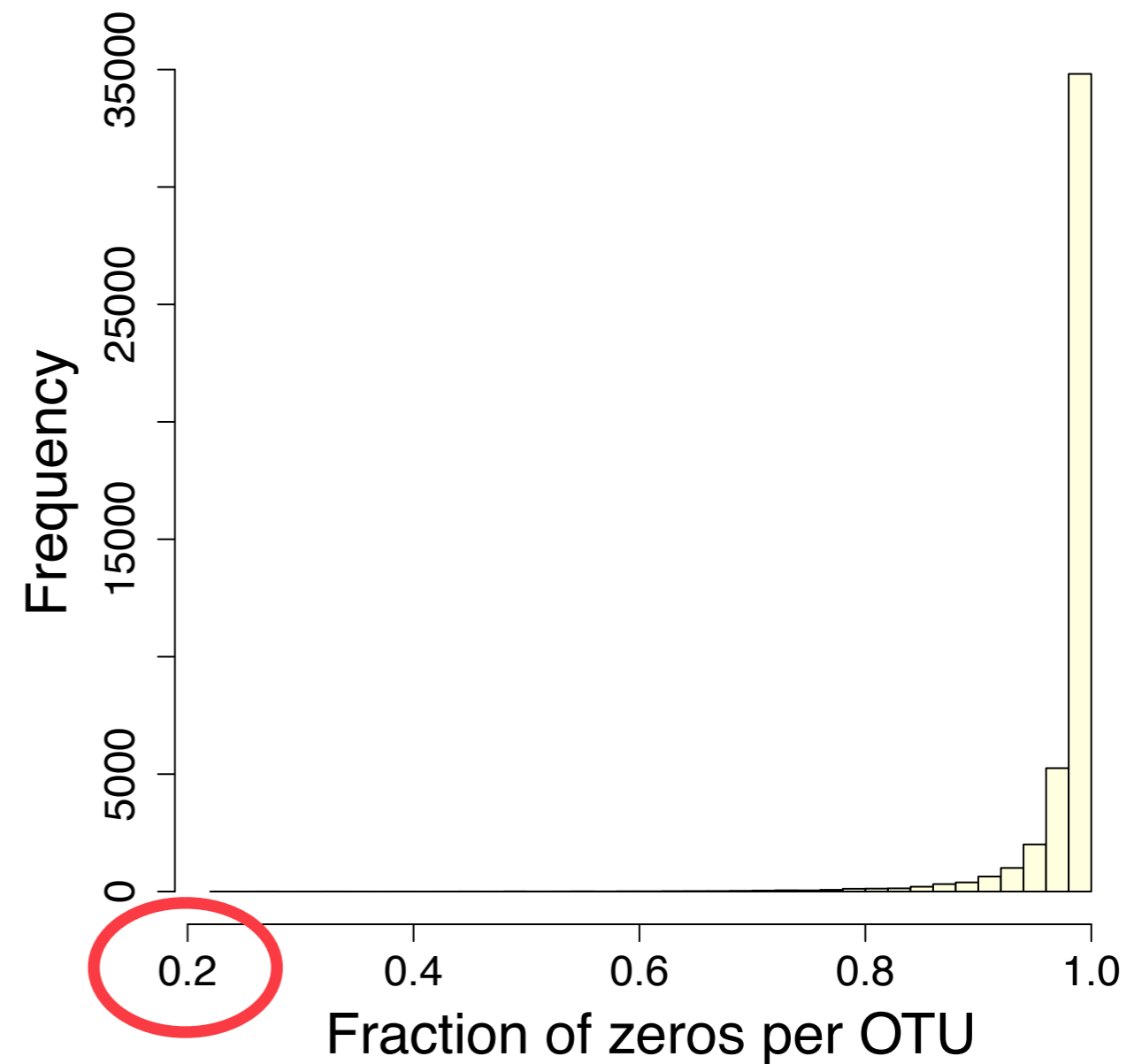


Similarity to single-cell RNA-seq



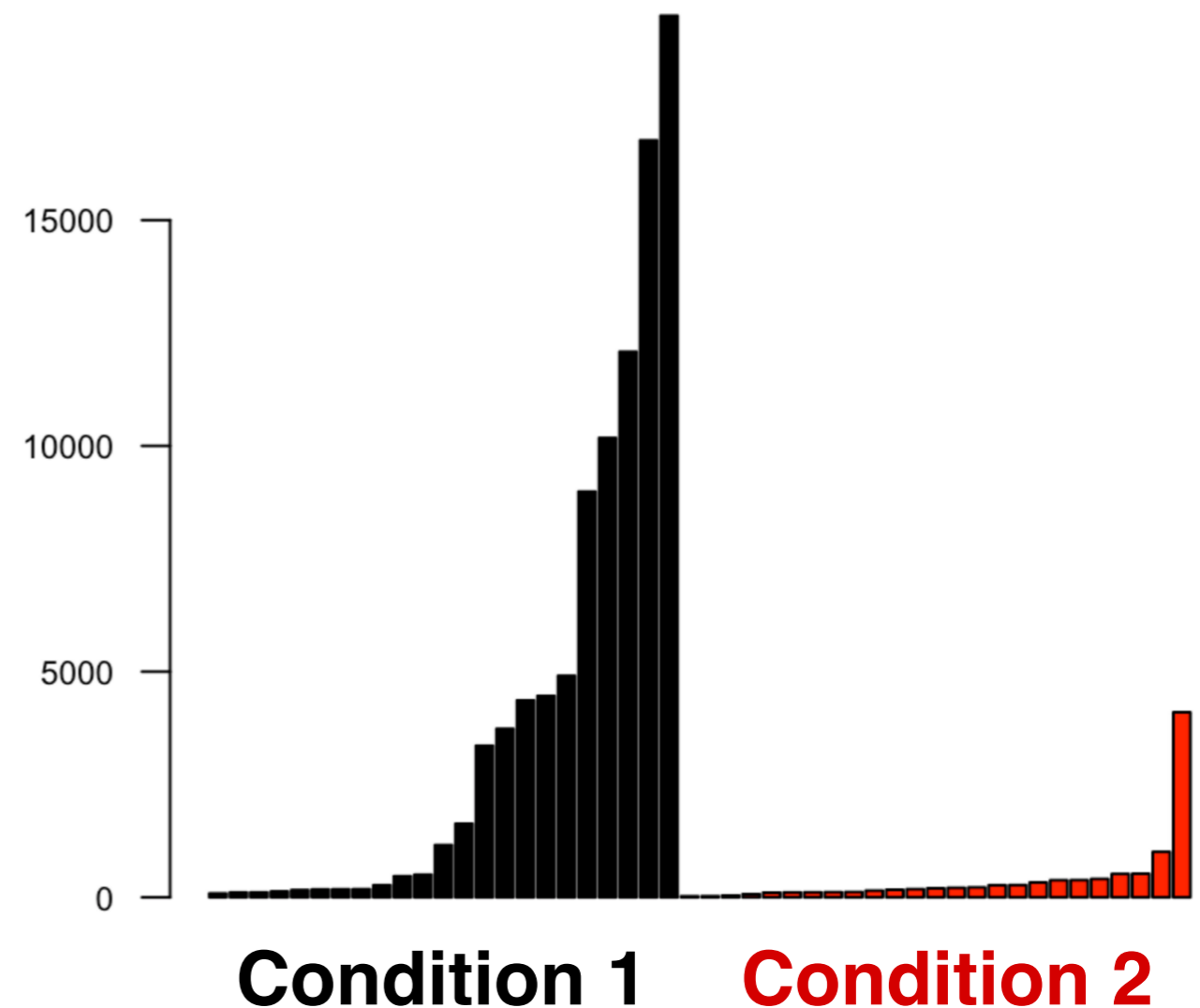
Normalization of sparse count data

- metagenomeSeq (CSS)
- scran
- SCnorm
- Wrench



Differential abundance testing

- What do we want to test?
 - Difference between mean abundance
 - Difference in fraction of zeros
 - Difference between mean abundance conditioning on being present
 - Overall difference in OTU composition



References

- Paulson et al.: Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* 10(12):1200-1203 (2013) - **metagenomeSeq**
- Mandal et al.: Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* 26:27663 (2015) - **ANCOM**
- McMurdie & Holmes: Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology* 10(4):e1003531 (2014)
- Lun et al.: Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* 17:75 (2016) - **scran**
- Bacher et al.: SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods* 14:584-586 (2017) - **SCnorm**
- Senthil Kumar et al.: Analysis and correction of compositional bias in sparse sequencing count data. *bioRxiv* doi: 10.1101/142851 (2017) - **Wrench**
- Weiss et al.: Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ Preprints* (2015)
- The Human Microbiome Project Consortium: Structure, function and diversity of the healthy human microbiome. *Nature* 486:207-214 (2012) - **Human Microbiome Project**
- Lovell et al.: Proportionality: a valid alternative to correlation for relative data. *PLoS Computational Biology* 11(3):e1004075 (2015) - **log-ratio analysis**
- Gloor et al.: It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology* (2016) - **log-ratio analysis**
- Westcott & Schloss: De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 3:e1487 (2015) - **OTU clustering**
- Schmidt et al.: Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environmental Microbiology* 17(5):1689-1706 (2015) - **OTU clustering**
- Cox et al.: Sequencing the human microbiome in healthy and disease. *Human Molecular Genetics* 22:R88-R94 (2013)
- Yarza et al.: Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* 12:635-645 (2014)
- Bodilis et al.: Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in *Pseudomonas*. *PLoS One* 7(4):e35647 (2012)
- Janda & Abbott: 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils and pitfalls. *Journal of Clinical Microbiology* 45(9):2761-2764 (2007)
- Fouhy et al.: 16S rRNA gene sequencing of mock microbial populations - impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology* 16:123 (2016)