Microbial genomics

Charlotte Soneson
University of Zurich
Brixen 2016

What is the “microbiome”?

Logos: https://commons.wikimedia.org/wiki/File:Human_Microbiome_Project_logo.jpg

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

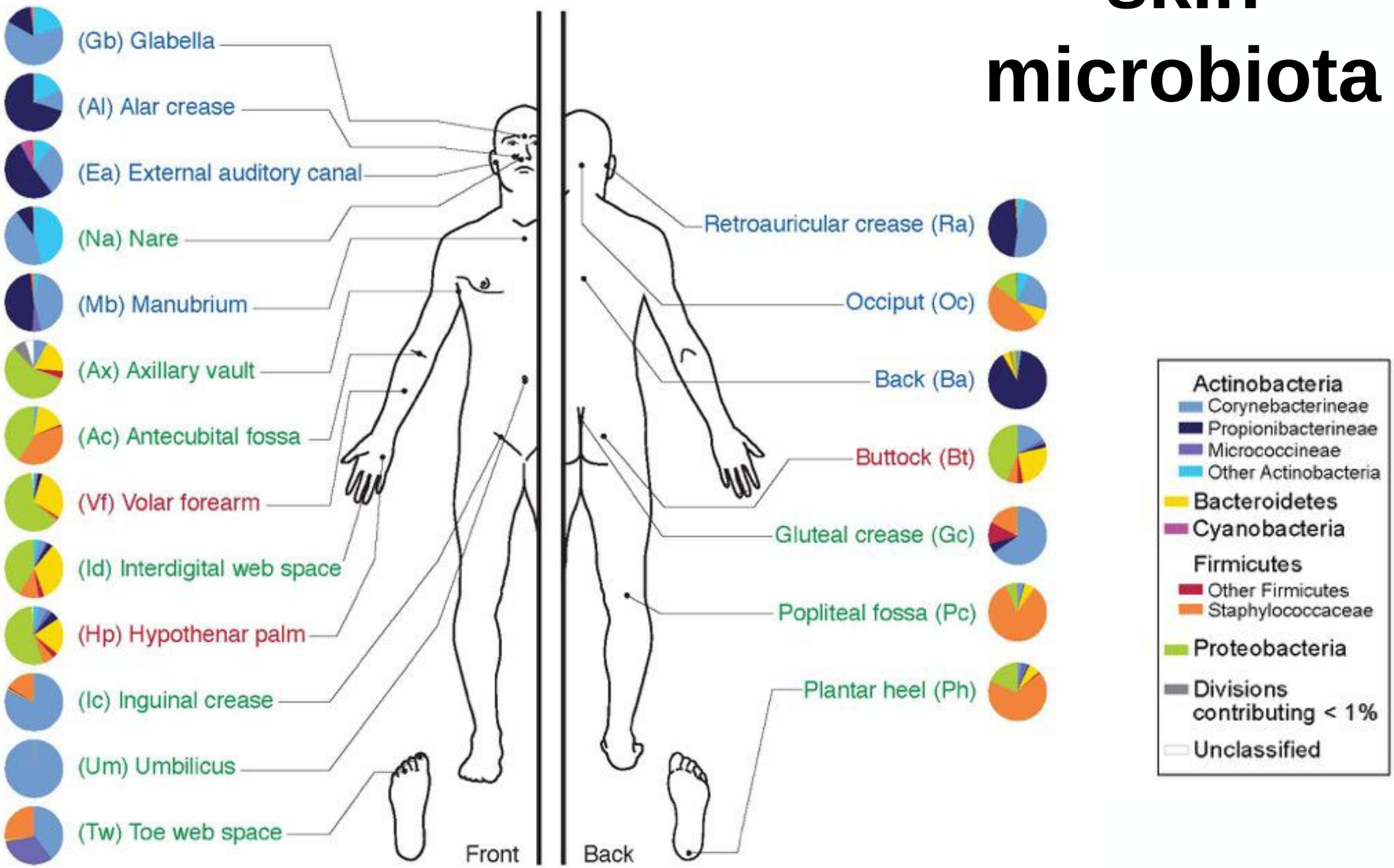
Cox et al., Human Molecular Genetics 2013

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

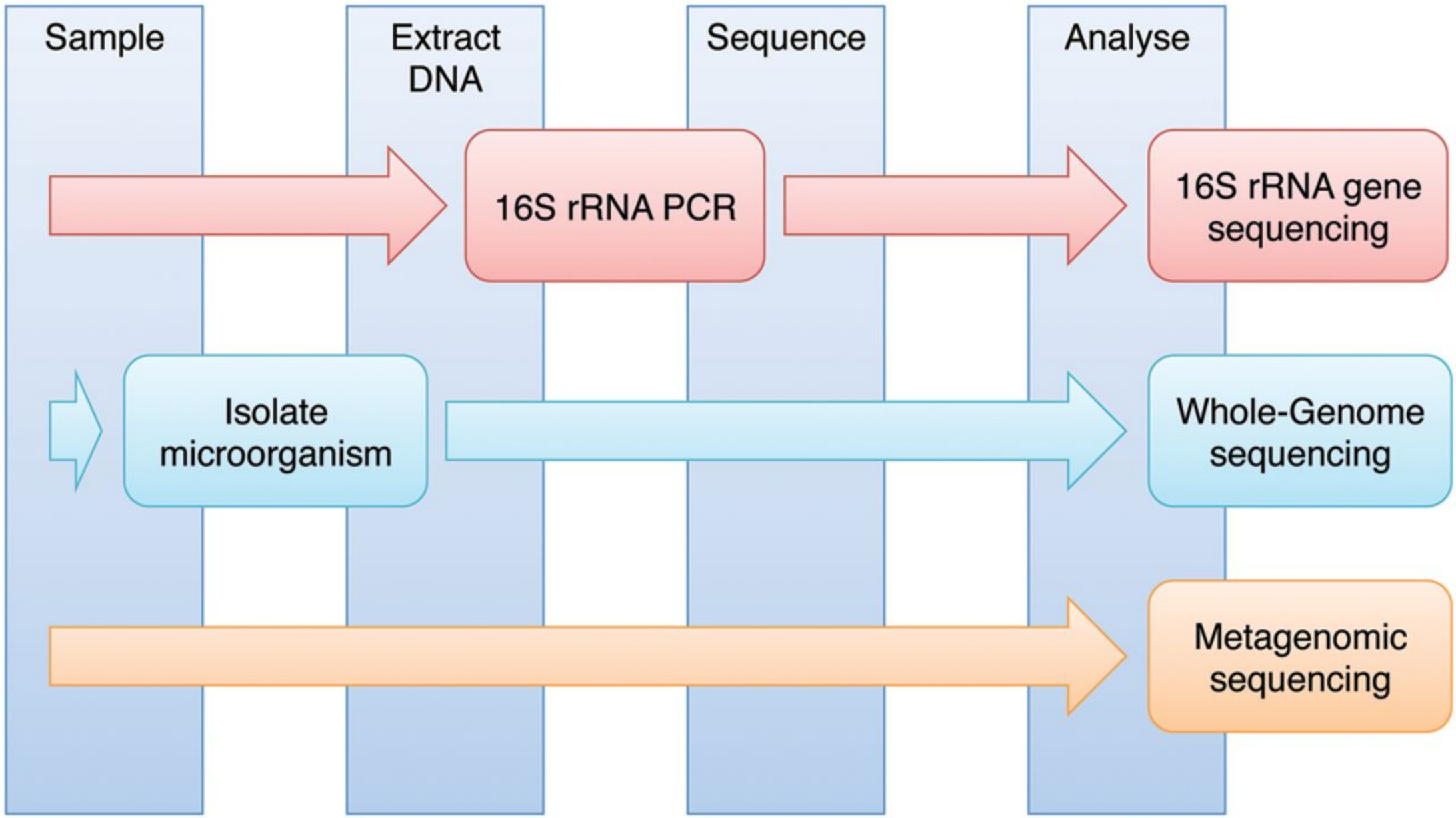
Cox et al., Human Molecular Genetics 2013

Human skin microbiota



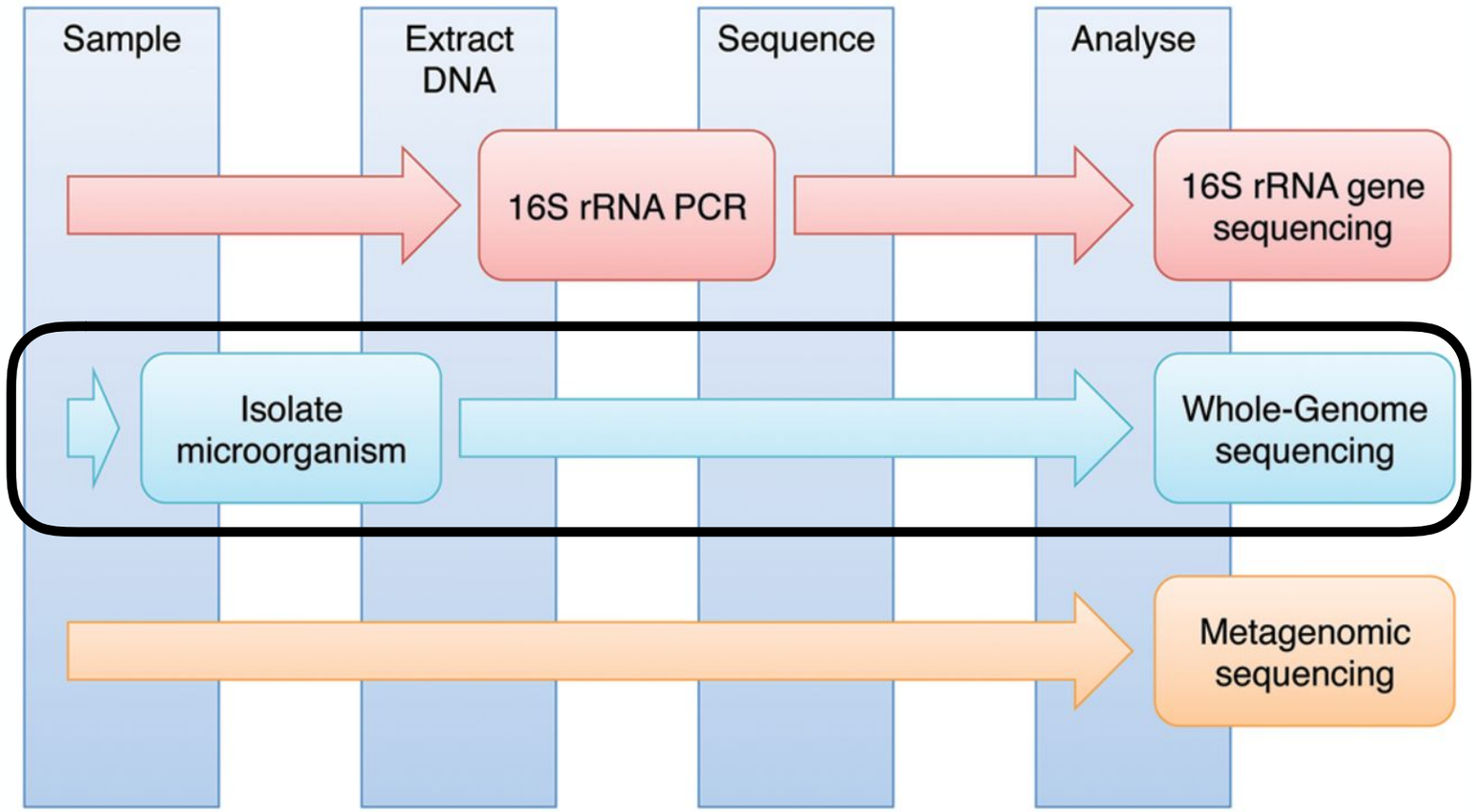
https://commons.wikimedia.org/wiki/File:Skin_Microbiome20169-300.jpg

How can we analyze the microbiome?



Cox et al., Human Molecular Genetics 2013

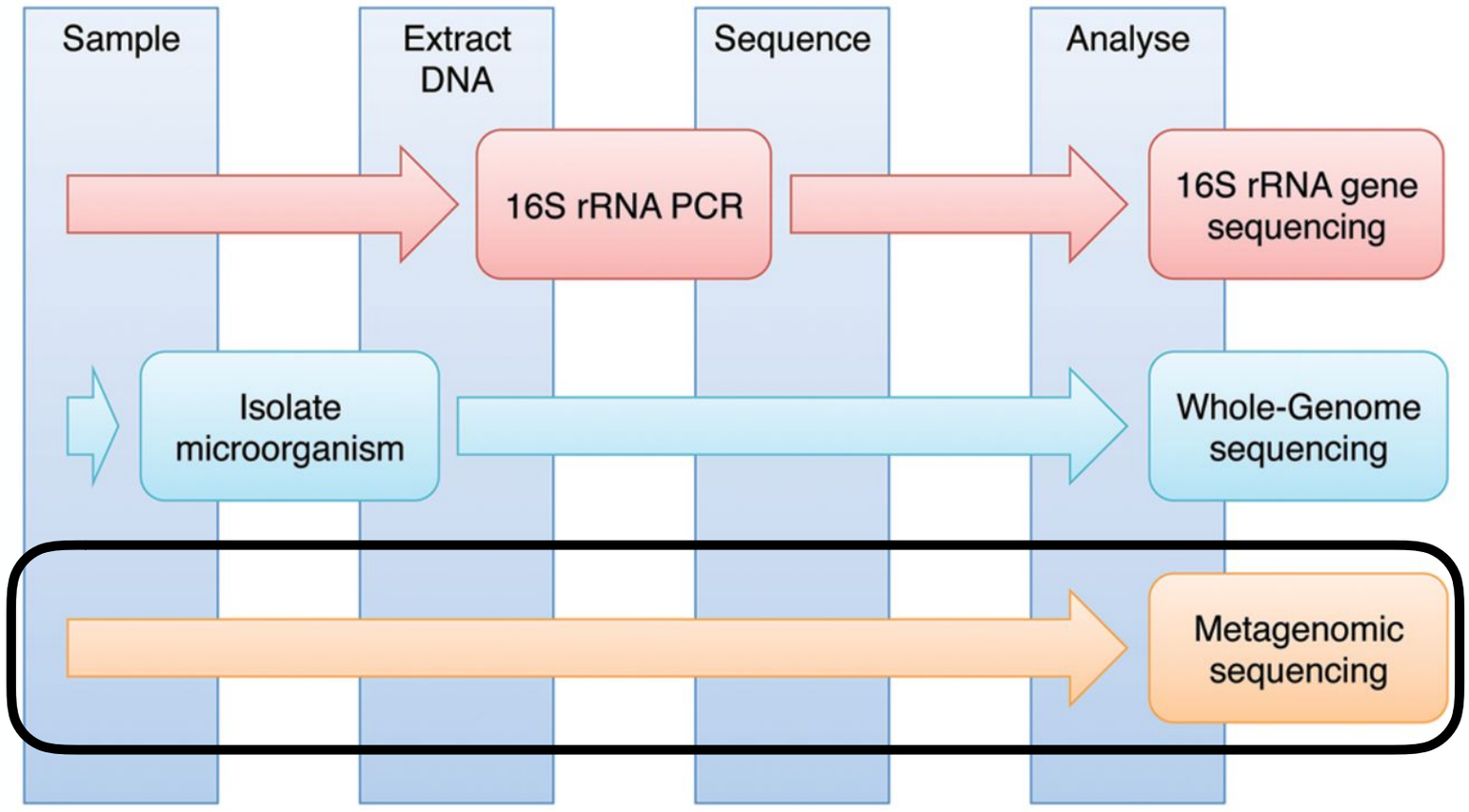
How can we analyze the microbiome?



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

Cox et al., Human Molecular Genetics 2013

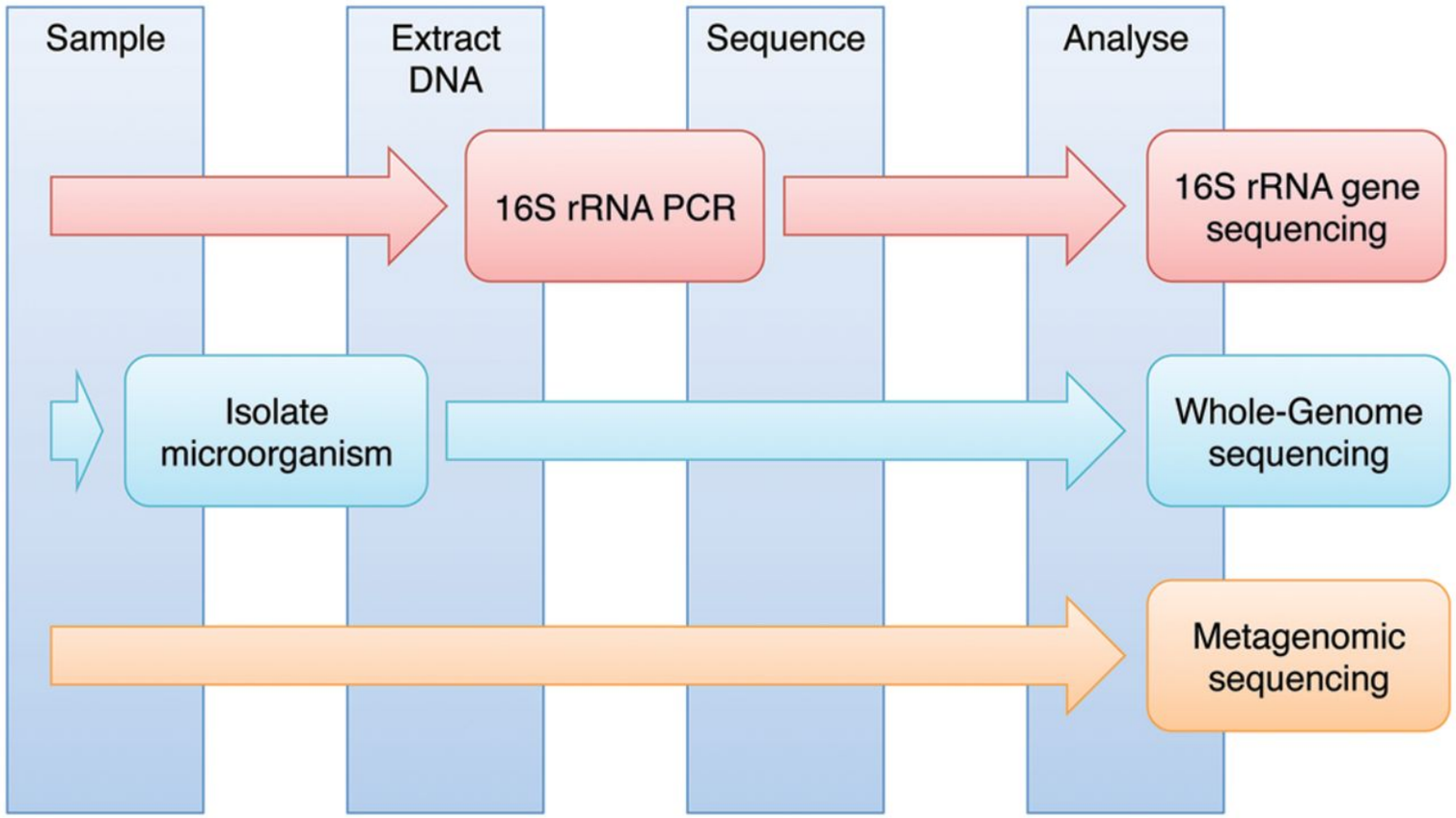
How can we analyze the microbiome?



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

Cox et al., Human Molecular Genetics 2013

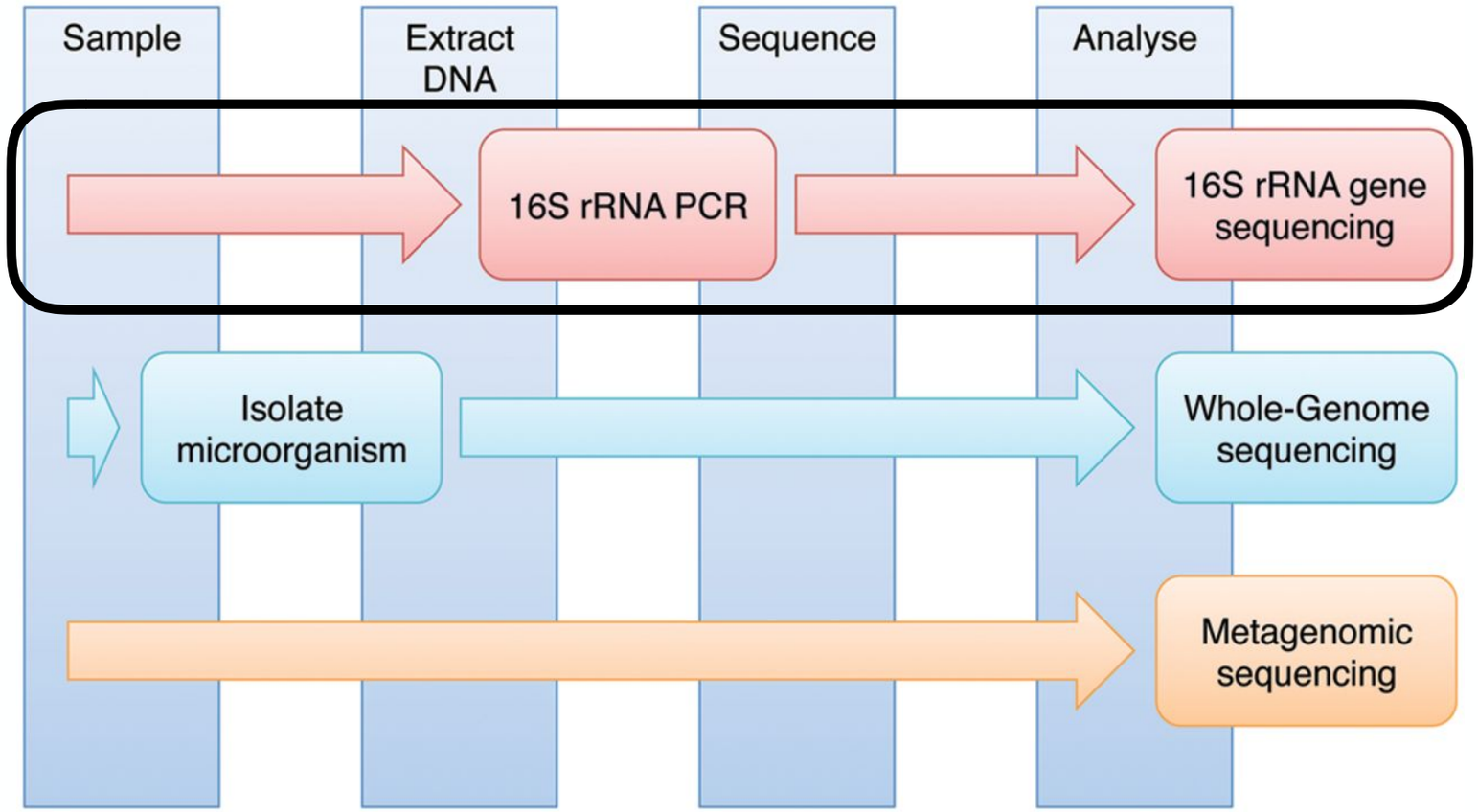
How can we analyze the microbiome?



- meta*transcriptomics*
- meta*proteomics*
- meta*bolomics*
- …

Cox et al., Human Molecular Genetics 2013

How can we analyze the microbiome?



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

Cox et al., Human Molecular Genetics 2013

Amplicon sequencing - basic idea

- Amplify (part of) the 16S rRNA gene from all microbes - sequence amplified part
- Cluster sequences together in so called OTUs (= 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

How to cluster?

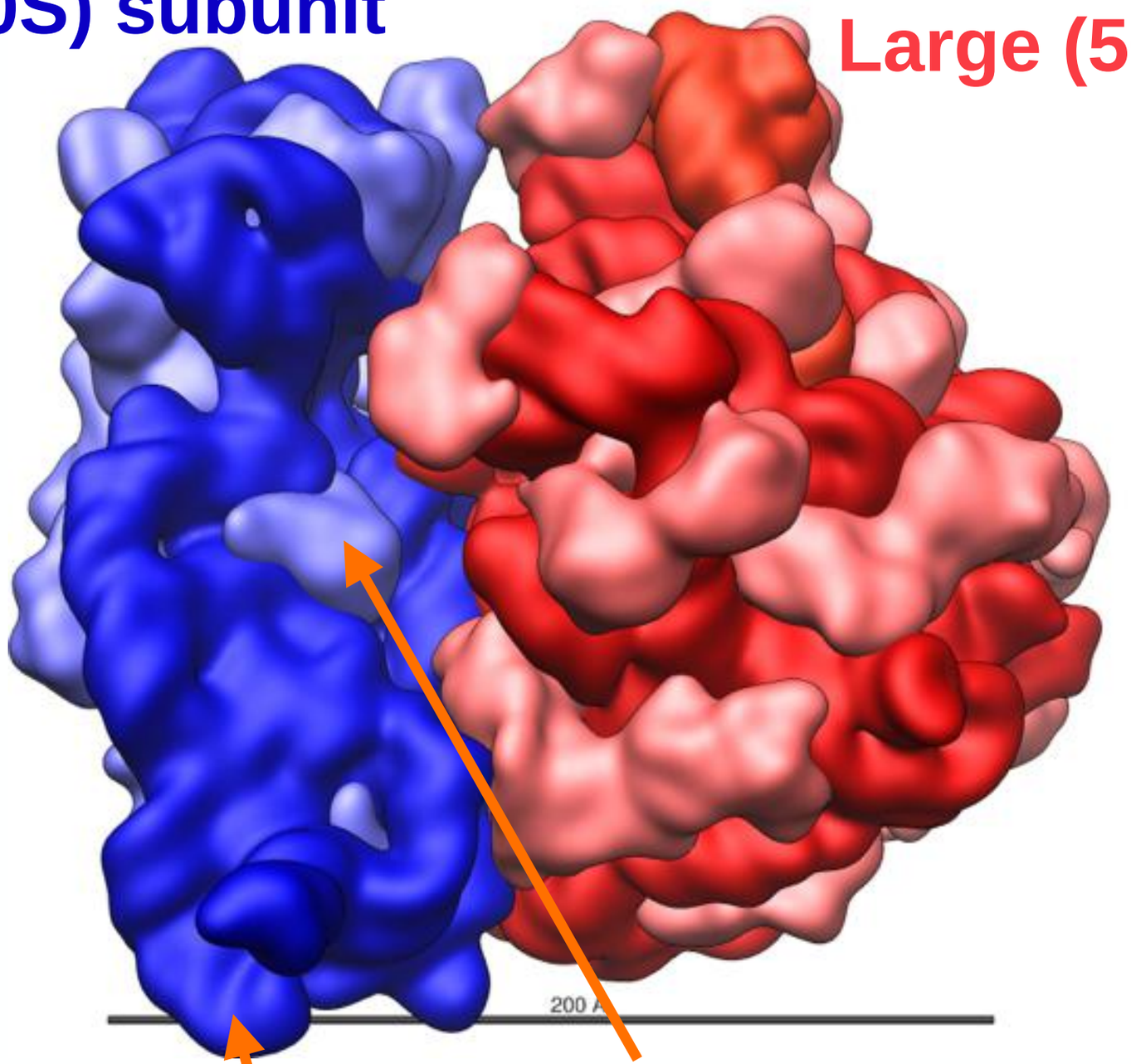
- Cluster sequences together in so called OTUs (= clusters of similar sequences ~ “species”)

- Get the number of sequences in each cluster/OTU for each sample

How to analyze?

- Generate an abundance table (OTUs x samples)

An E.coli ribosome

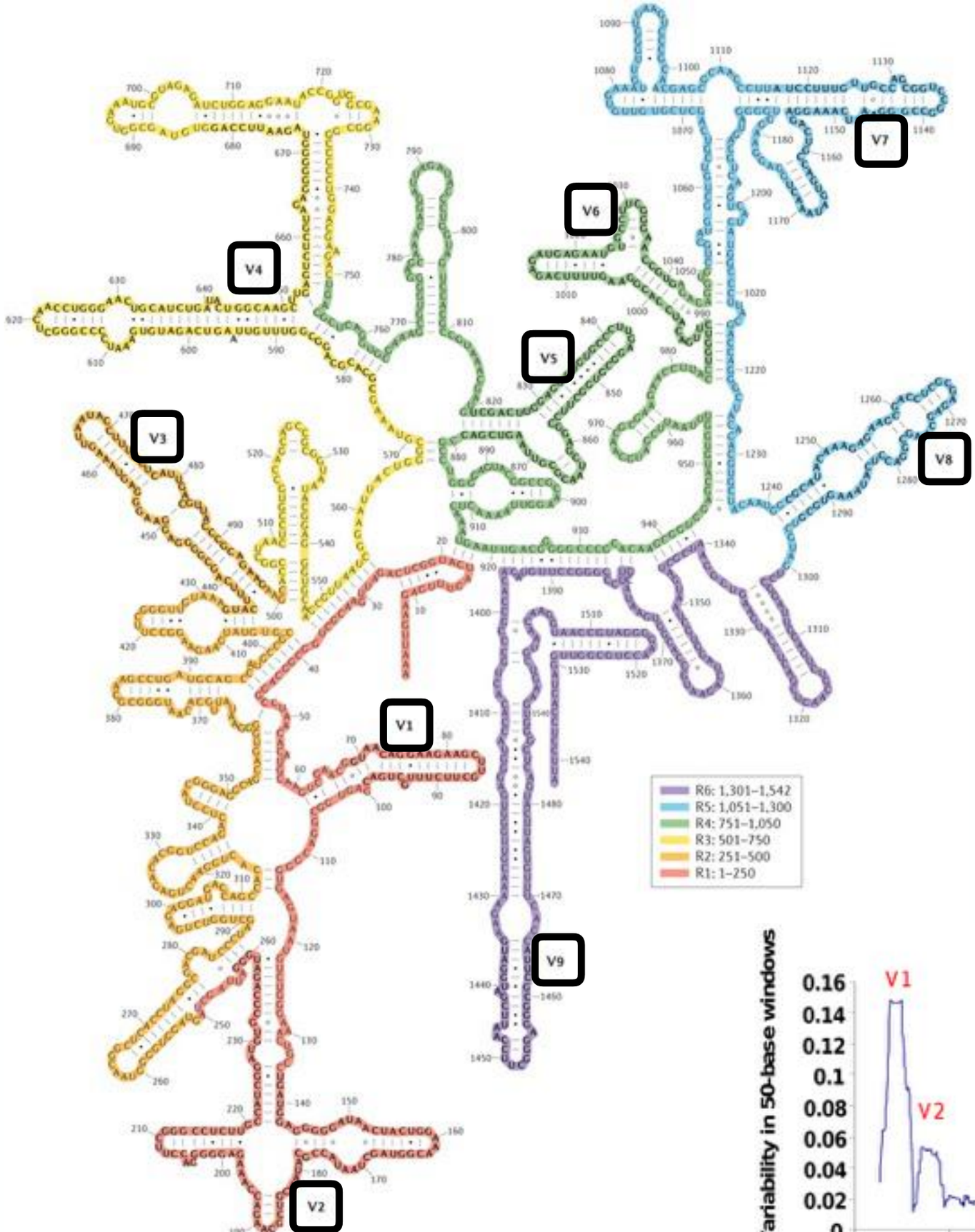


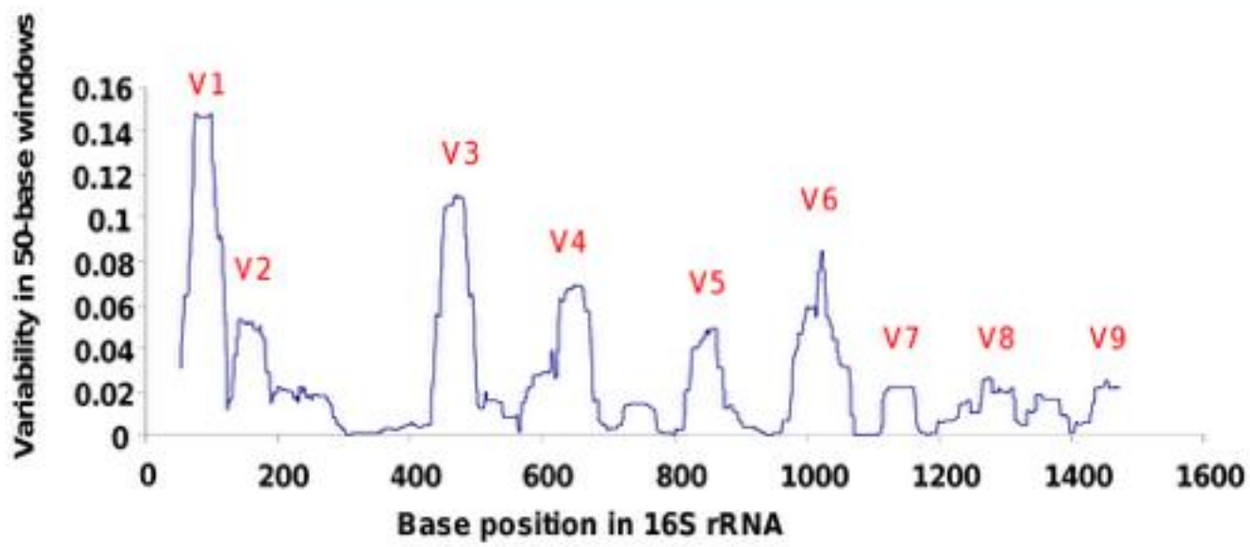
https://commons.wikimedia.org/wiki/File:Ribosome_shape.png

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)



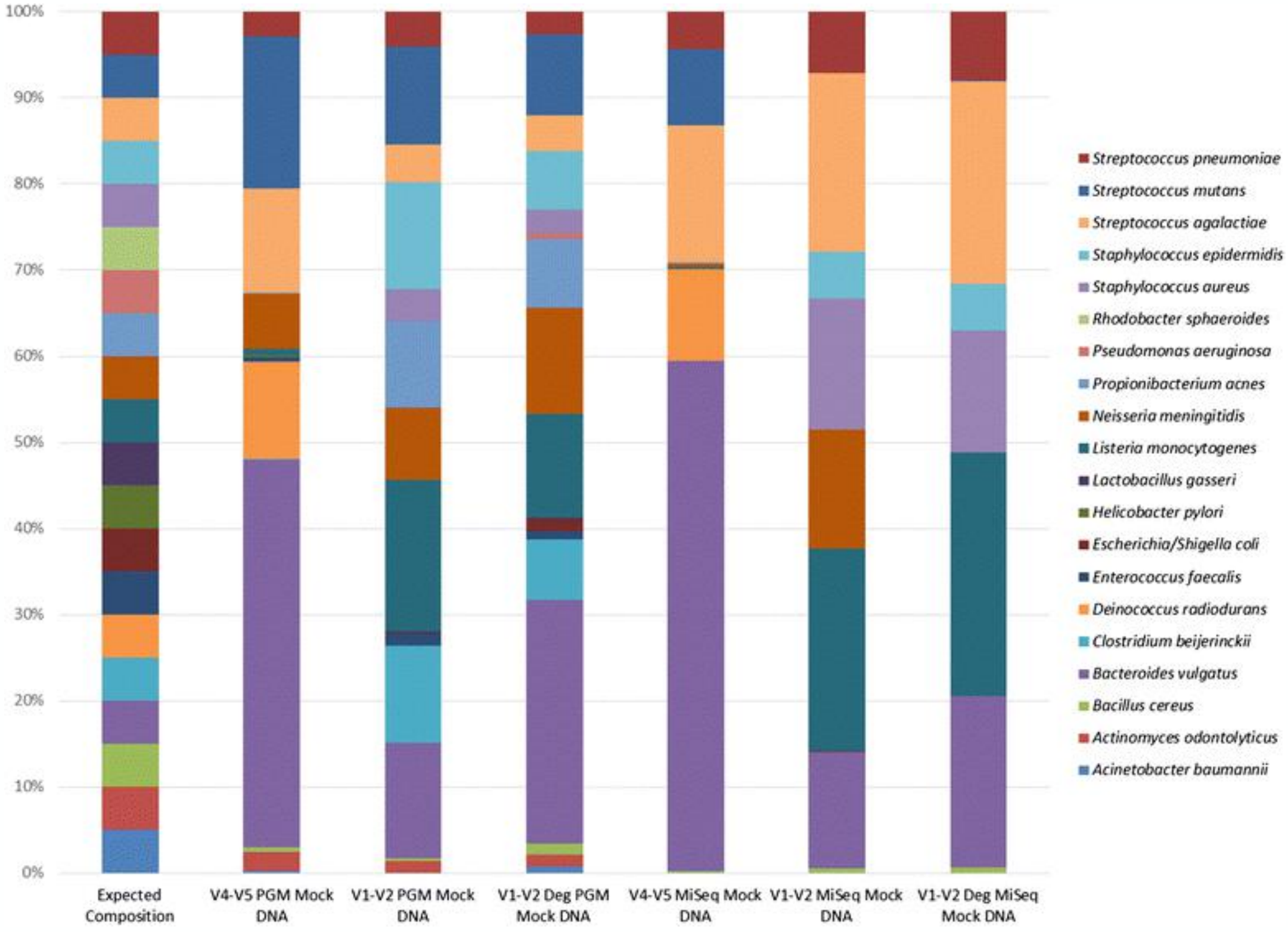


Yarsa et al., Nature Rev Microbiology 2014; Bodilis et al., PLoS One 2012

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



Fouhy et al., BMC Microbiology 2016

Preprocessing of reads

- Merge read pairs and remove low-quality reads
- Align reads to reference 16S sequence

16S sequence databases

- SILVA (http://www.arb-silva.de/)
- RDP (https://rdp.cme.msu.edu/)
- GreenGenes (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)

```
>AF515816.1
.....................................A------T-G--A-C-G-C--T-G-G-C--G-G--C-A-T-G----------C----T-T--TA-C--AC-A----T-G--C----A-
>AY688433.1
.....................................A---C--G----A-C-G-C--T-G-G-C--G-G--C-A-T-G----------C----T-T--A--C--AC-A----T-G--C----A-
>Z22781.1
.....................................A---C--G----A-C-G-T--T---G-C--G-A--T-G-C-G----------T----C-T--TA-A--GC-A----T-G--C----A-
>AJ582031.1
.....................................A---C--G-A----C-G-C--T-G-G-C--G-G--C-A-G-G----------C----C-T--AA-T--AC-A----T-G--C----A-
>AB070566.1
.....................................A---C--G-A----C-G-C--T-G-G-C--G-G--C-A-G-G----------C----T-T--AA-C--AC-A----T-G--C----A-
>AB070570.1
.....................................A---C--G-A----CAG-C--T-G-G-C--G-G--C-A-G-G----------C----T-T--AA-C--AC-A----T-G--C----A-
>AY033301.1
.....................................A---C--G-A--A-C-G-C--T-G-G-C--G-G--C-A-G-G----------C----C-T--AA-C--AC-A----T-G--C----A-
>AY035307.1
```

16S sequence databases

- SILVA (http://www.arb-silva.de/)
- RDP (https://rdp.cme.msu.edu/)
- GreenGenes (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi)

```
AB000389.1      Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Pseudoalteromonadaceae;Pseudoalteromonas;
AB000699.1      Bacteria;Proteobacteria;Betaproteobacteria;Nitrosomonadales;Nitrosomonadaceae;Nitrosomonas;
AB000700.1      Bacteria;Proteobacteria;Betaproteobacteria;Nitrosomonadales;Nitrosomonadaceae;Nitrosomonas;
AB000701.1      Bacteria;Proteobacteria;Betaproteobacteria;Nitrosomonadales;Nitrosomonadaceae;Nitrosomonas;
AB000702.1      Bacteria;Proteobacteria;Betaproteobacteria;Nitrosomonadales;Nitrosomonadaceae;Nitrosomonas;
AB001518.1      Bacteria;Proteobacteria;Alphaproteobacteria;Rickettsiales;Rickettsiaceae;
AB001724.1      Bacteria;Cyanobacteria;Chroococcales;Microcystis;
AB001774.1      Bacteria;Chlamydiae;Chlamydiales;Chlamydiaceae;Chlamydophila;
AB001775.1      Bacteria;Chlamydiae;Chlamydiales;Chlamydiaceae;Chlamydophila;
```

Preprocessing of reads

mothur

- Merge read pairs and remove low-quality reads
- Align reads to reference 16S sequence
- Denoise (cluster very similar sequences)
- Identify and remove chimeras and contaminants
- Cluster reads into **O**perational **T**axonomic **U**nits (OTUs)

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

OTU generation

- "**closed-reference** clustering": compare sequences to a reference catalog, group together sequences that are similar to the same references.

- Fast, parallelizable
- OTU assignment independent of other sequences
- Comparable OTUs across studies

- Relies on accuracy and completeness of reference catalog
- Sequence can be similar to multiple reference sequences
- Similarity among clustered sequences may be lower than the similarity between each of them and the reference

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.

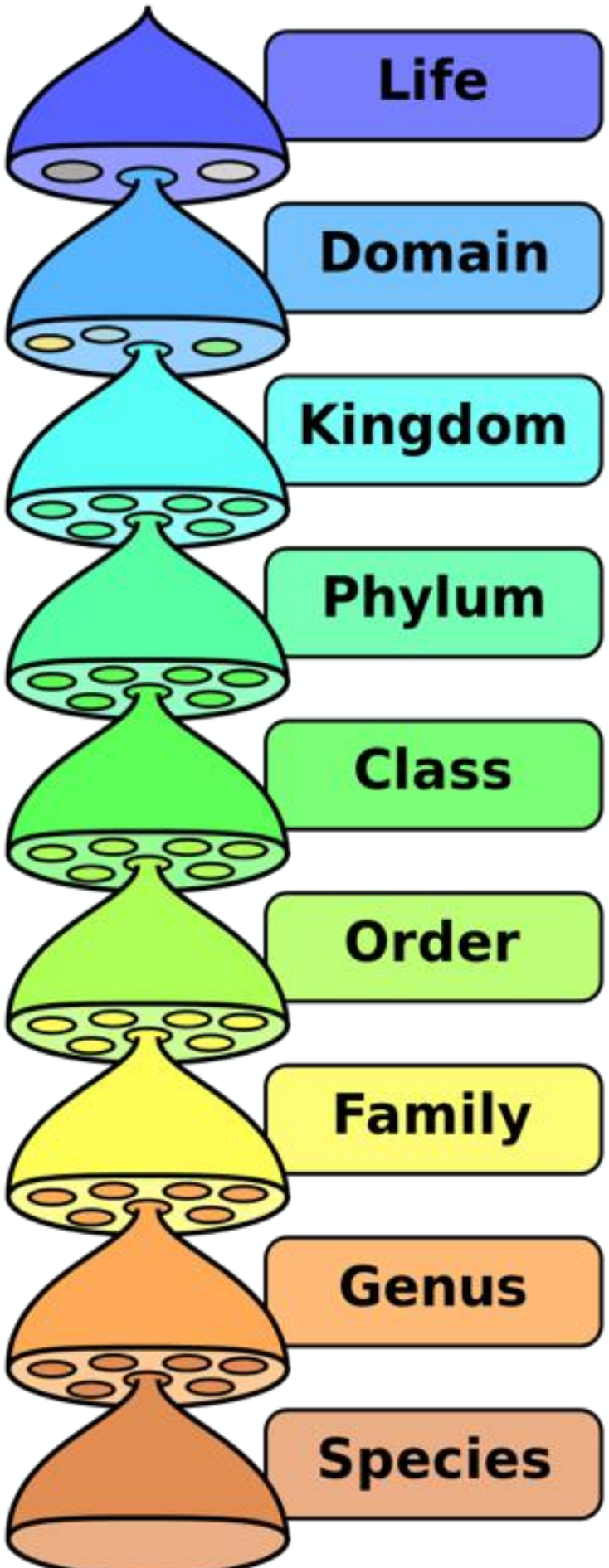
  - Independent of reference catalog

  - Scales quadratically with number of sequences
  - OTU assignment depends on which other sequences are present

de novo OTU clustering

- results depend on clustering method
- single-linkage tends to be (too?) “inclusive” while average-linkage/complete-linkage are more “exclusive”

	700114607	700114380	700114716	700114798	Consensus Lineage
OTU_97.4499	0	0	0	0	Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__
OTU_97.44990	0	0	0	0	Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae;g__Odoribacter
OTU_97.44991	0	1	4	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium
OTU_97.44992	0	0	0	0	Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Mycobacteriaceae;g__Mycobacterium
OTU_97.44993	0	1	1	0	Root;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;f__Campylobacteraceae;g__Campylobacter



Which similarity threshold?

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

https://en.wikipedia.org/wiki/File:Biological_classification_L_Pengo_vflip.svg

Representation in R - phyloseq object

```
> library(phyloseq)
> data(GlobalPatterns)
> GlobalPatterns
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 19216 taxa and 26 samples ]
sample_data() Sample Data:       [ 26 samples by 7 sample variables ]
tax_table()   Taxonomy Table:    [ 19216 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 19216 tips and 19215 internal nodes ]
> head(otu_table(GlobalPatterns))
OTU Table:          [6 taxa and 26 samples]
                     taxa are rows
       CL3 CC1 SV1 M31Fcsw M11Fcsw M31Plmr M11Plmr F21Plmr M31Tong M11Tong LMEpi24M SLEpi20M AQC1cm AQC4cm
549322   0   0   0       0       0       0       0       0       0       0        0        1     27    100
522457   0   0   0       0       0       0       0       0       0       0        0        0      0      2
951      0   0   0       0       0       0       1       0       0       0        0        0      0      0
244423   0   0   0       0       0       0       0       0       0       0        0        0      0     22
586076   0   0   0       0       0       0       0       0       0       0        0        0      0      2
246140   0   0   0       0       0       0       0       0       0       0        0        0      0      1
       AQC7cm NP2 NP3 NP5 TRRsed1 TRRsed2 TRRsed3 TS28 TS29 Even1 Even2 Even3
549322    130   1   0   0        0        0        0    0    0     0     0     0
522457      6   0   0   0        0        0        0    0    0     0     0     0
951         0   0   0   0        0        0        0    0    0     0     0     0
244423     29   0   0   0        0        0        0    0    0     0     0     0
586076      1   0   0   0        0        0        0    0    0     0     0     0
246140      3   0   0   0        0        0        0    0    0     0     0     0
```

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
```

Construct phyloseq object

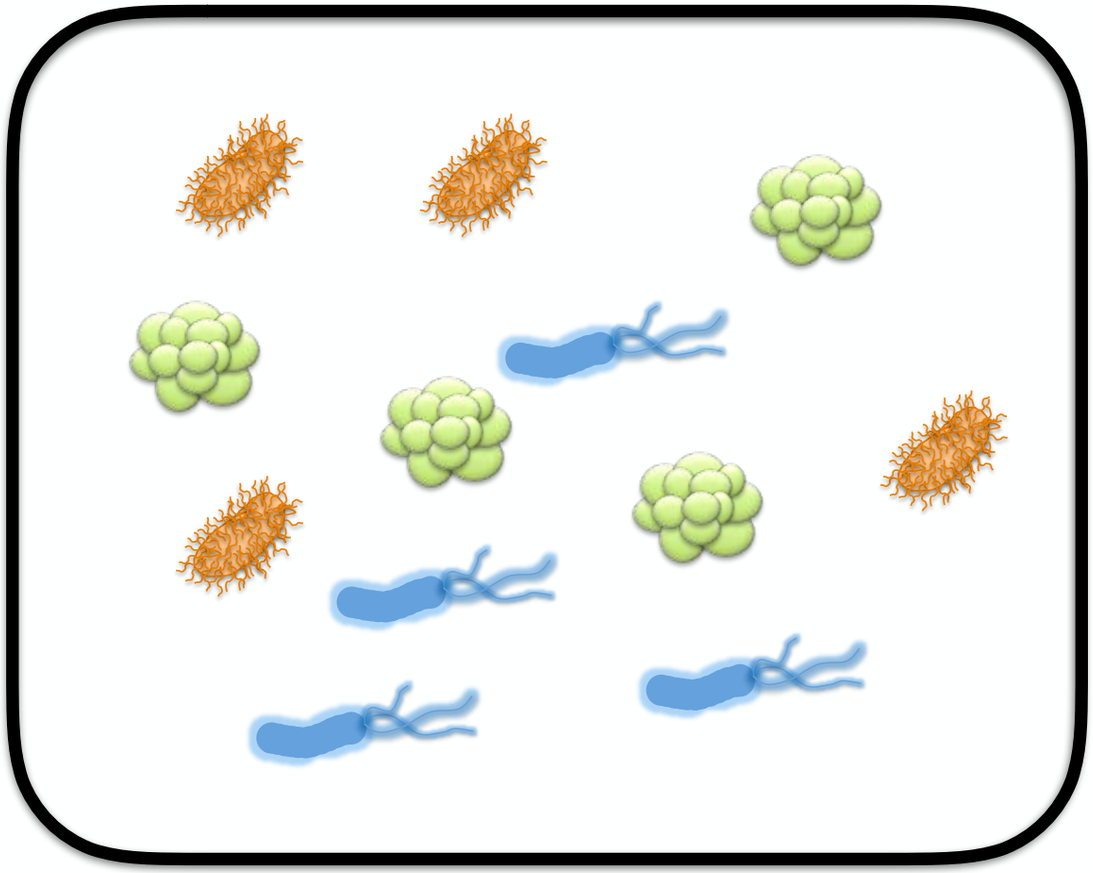
```
> otu_table[1:3, 1:5]
           700013549 700014386 700014403 700014409 700014412
OTU_97.1           0         0         0         0         0
OTU_97.10          0         0         0         0         0
OTU_97.100         0         0         0         0         0
>
> sample_info[1:3, ]
               RSID visitno    sex RUNCENTER HMPbodysubsite Mislabeled Contaminated
700013549 158013734       1 female       BCM          Stool         NA           NA
700014386 158398106       1   male    BCM,BI          Stool         NA           NA
700014403 158398106       1   male    BCM,BI         Saliva         NA           NA
                                                                                                  Description
700013549 HMP_Human_metagenome_sample_700013549_from_subject_158013734__sex_female_
700014386   HMP_Human_metagenome_sample_700014386_from_subject_158398106__sex_male_
700014403   HMP_Human_metagenome_sample_700014403_from_subject_158398106__sex_male_
                    HMPbodysite
700013549 Gastrointestinal_tract
700014386 Gastrointestinal_tract
700014403                  Oral
>
> lineage2[1:3, ]
           Phylum           Class                 Order              Family             Genus
OTU_97.1   "Firmicutes"     "Bacilli"             "Lactobacillales"  "Streptococcaceae" "Streptococcus"
OTU_97.10  "Proteobacteria" "Betaproteobacteria" "Neisseriales"     "Neisseriaceae"    "Neisseria"
OTU_97.100 "Bacteroidetes"  "Bacteroidia"         "Bacteroidales"    "Bacteroidaceae"   "Bacteroides"
>
> phylo <- phyloseq(otu_table = otu_table(otu_table, taxa_are_rows = TRUE),
+                   sample_data = sample_data(sample_info),
+                   tax_table = tax_table(lineage2))
>
> phylo
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 45383 taxa and 4743 samples ]
sample_data() Sample Data:       [ 4743 samples by 9 sample variables ]
tax_table()   Taxonomy Table:    [ 45383 taxa by 5 taxonomic ranks ]
```

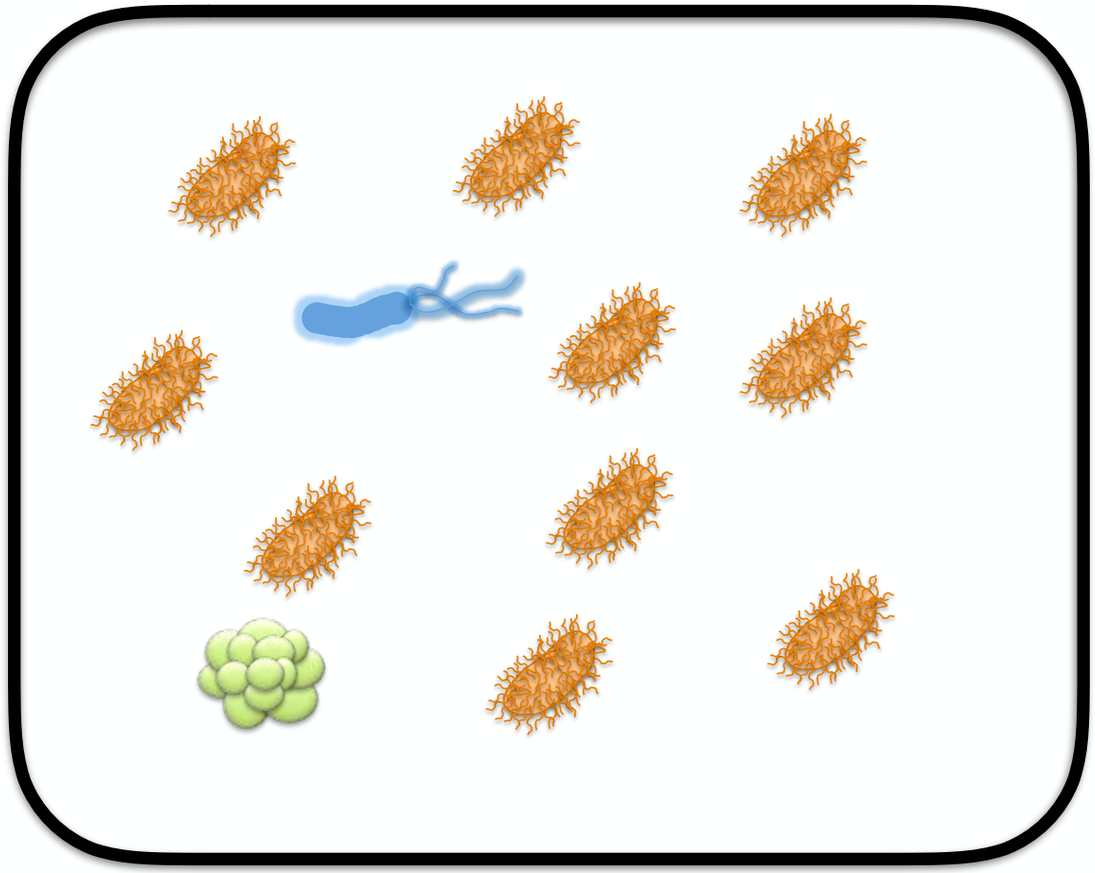
Filtering and subsetting

```
> phylo
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 45383 taxa and 4743 samples ]
sample_data() Sample Data:        [ 4743 samples by 9 sample variables ]
tax_table()   Taxonomy Table:     [ 45383 taxa by 5 taxonomic ranks ]
>
> phylo <- prune_taxa(taxa_sums(phylo) > 0, phylo)
> phylo <- prune_samples(sample_sums(phylo) > 100, phylo)
>
> phylo
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 45369 taxa and 4586 samples ]
sample_data() Sample Data:        [ 4586 samples by 9 sample variables ]
tax_table()   Taxonomy Table:     [ 45369 taxa by 5 taxonomic ranks ]
>
> phylosub <- subset_taxa(phylo, Phylum == "Acidobacteria")
>
> phylosub
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 26 taxa and 4586 samples ]
sample_data() Sample Data:        [ 4586 samples by 9 sample variables ]
tax_table()   Taxonomy Table:     [ 26 taxa by 5 taxonomic ranks ]
```

Richness and alpha diversity

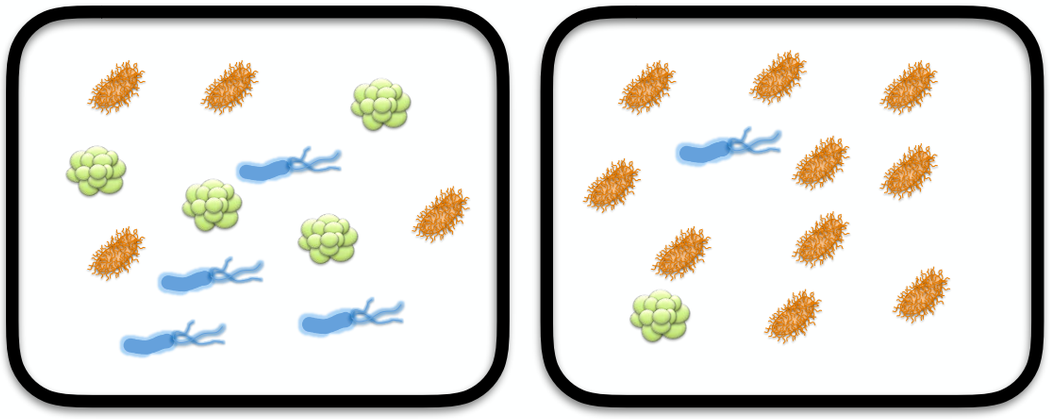
- richness = number of species observed in a sample
- alpha diversity ~ diversity ("unevenness") of species abundances within a sample



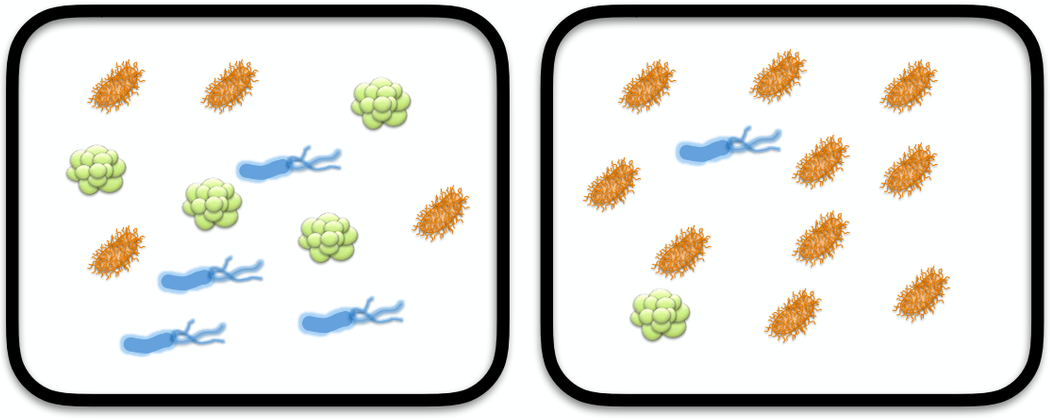


http://www.knightslab.org/#!bugs/c1sjx

Richness and alpha diversity

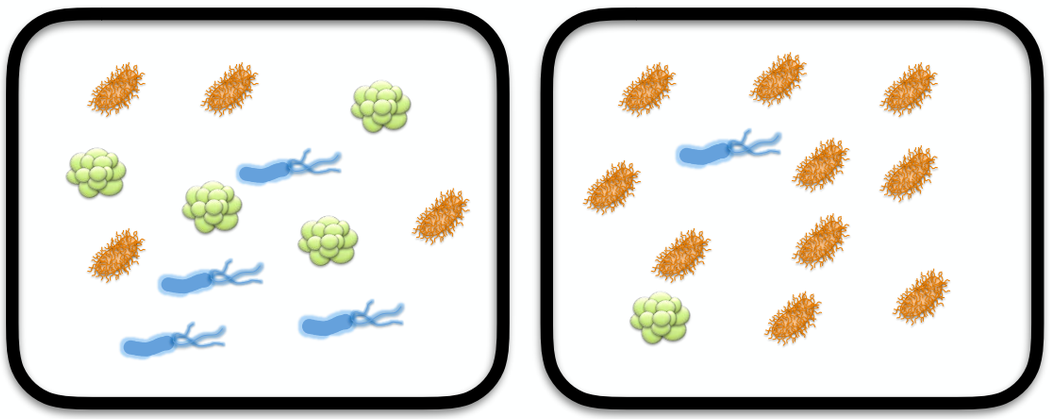


Richness and alpha diversity

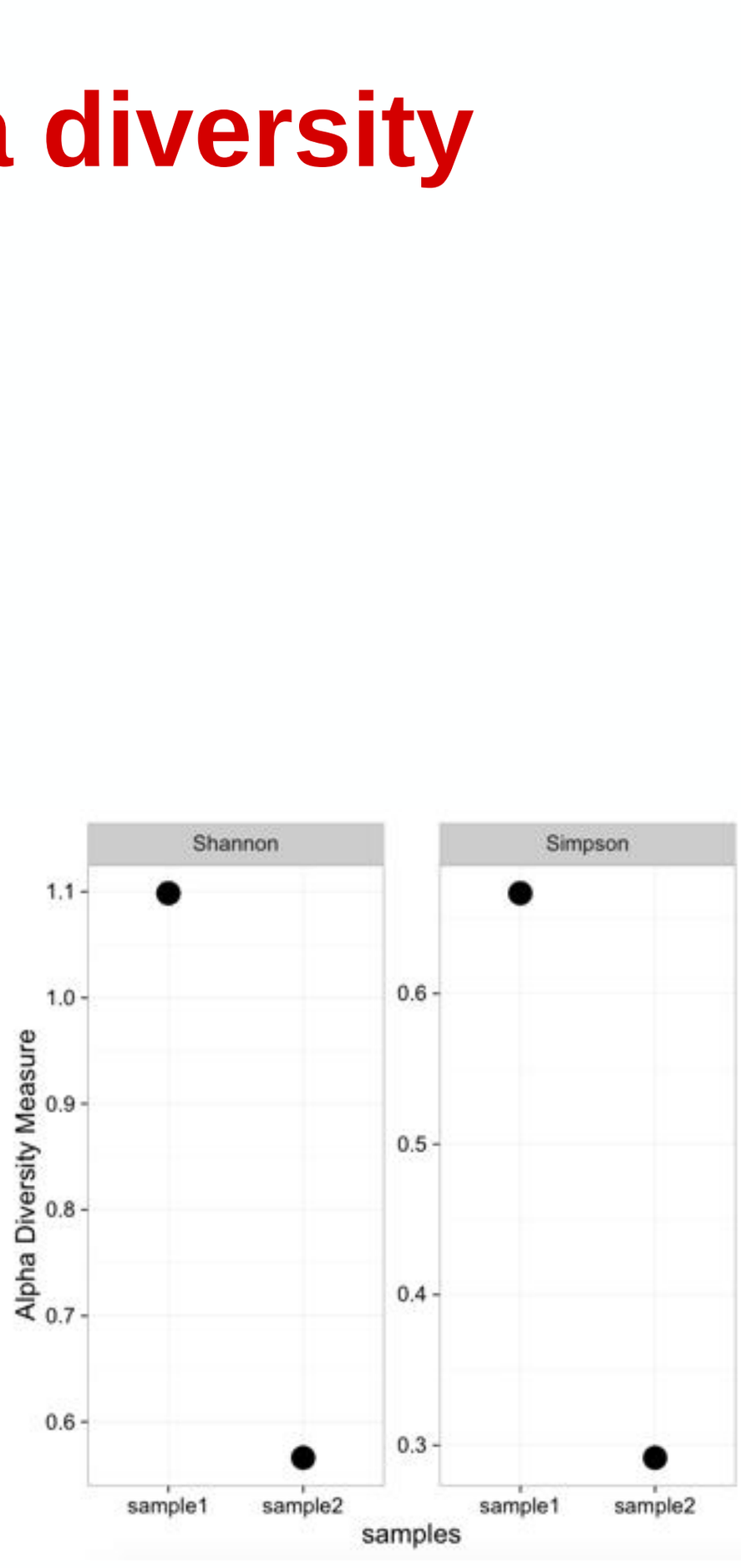


```
> phx
OTU Table:          [3 taxa and 2 samples]
                     taxa are rows
     sample1 sample2
OTU1       4      10
OTU2       4       1
OTU3       4       1
>
> estimate_richness(phx, measures = "Observed")
        Observed
sample1        3
sample2        3
```

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)
```

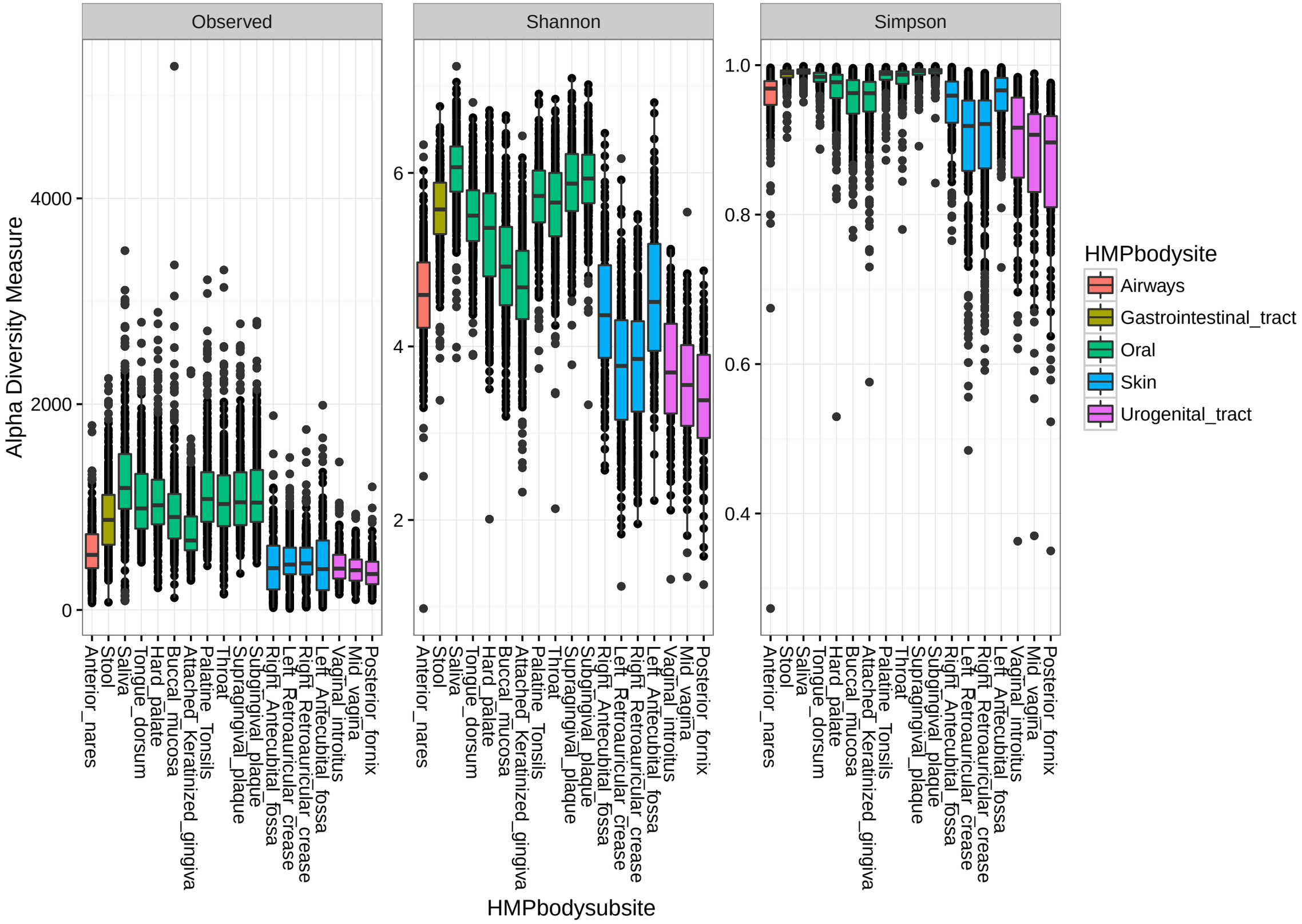


$$\text{Shannon: } H = -\sum_i p_i \log p_i$$

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

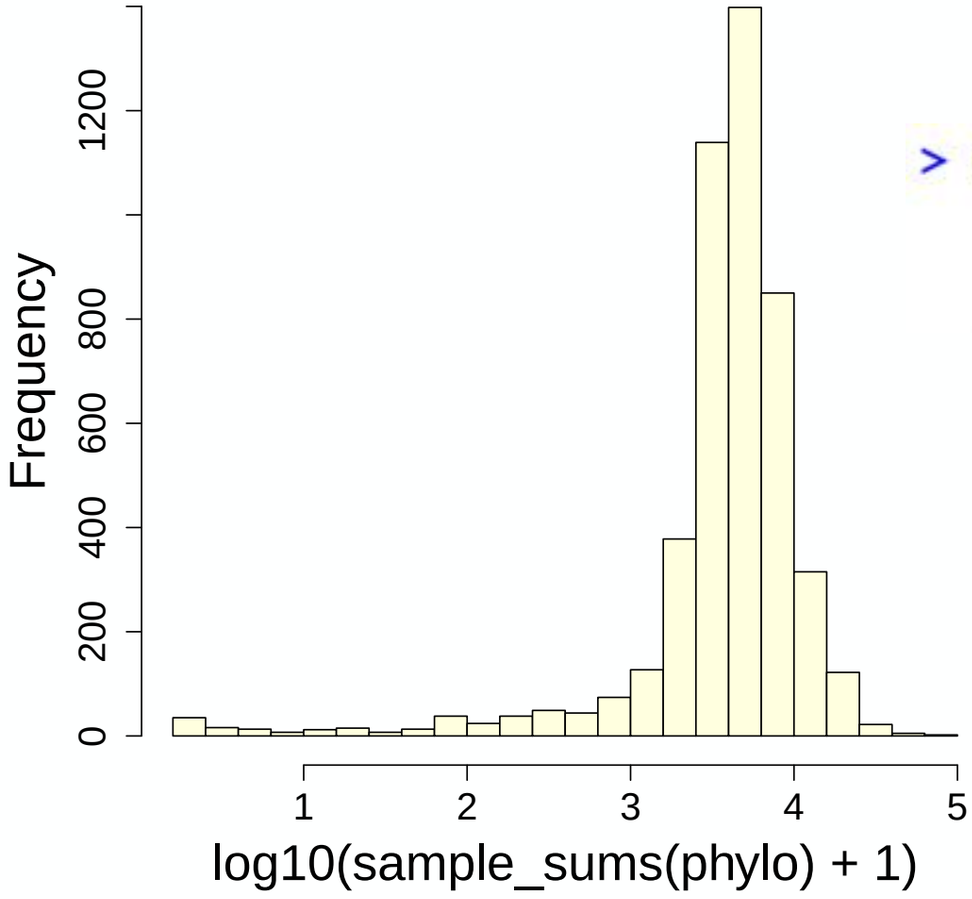
relative abundance of species *i*

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
```

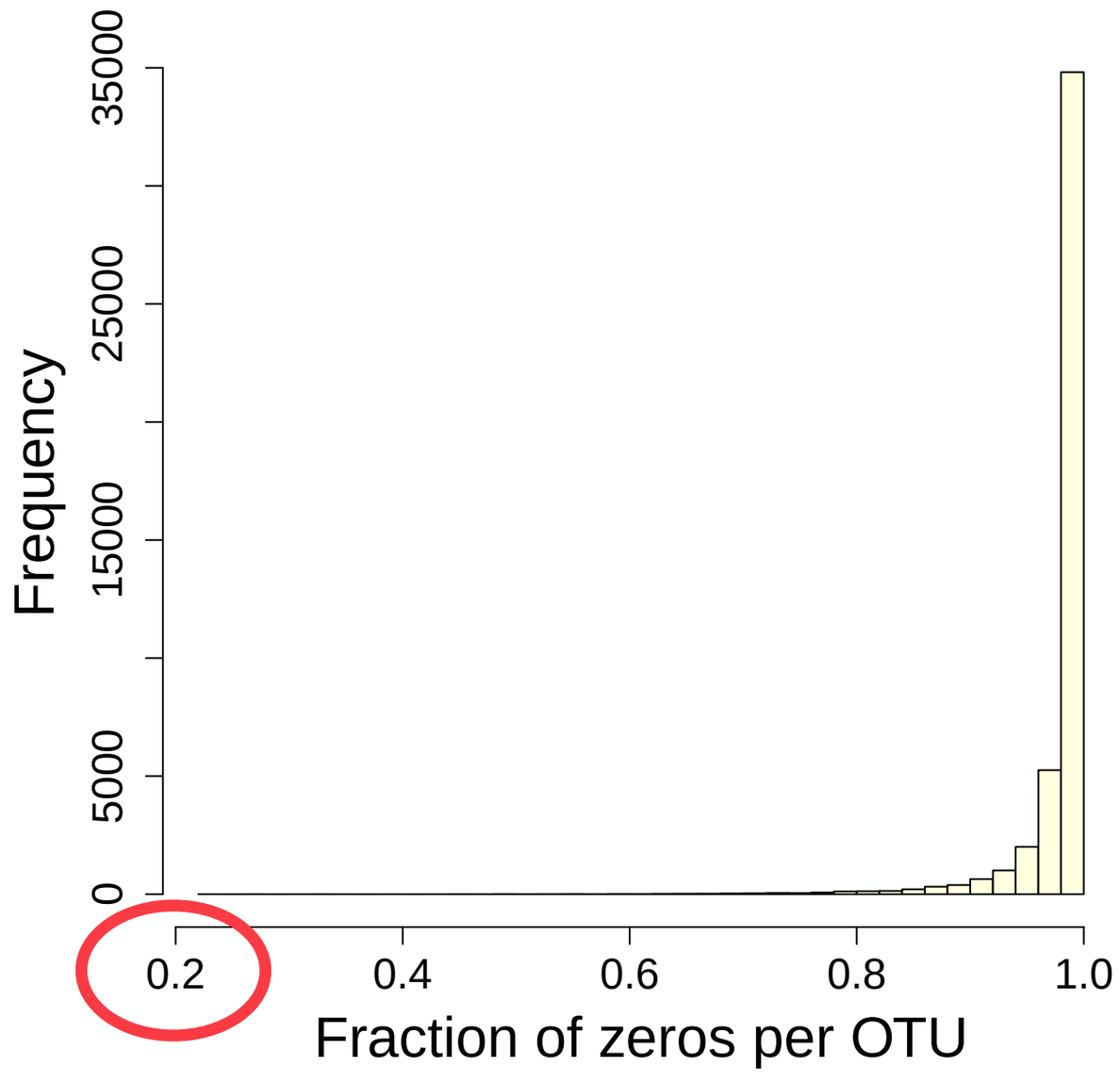
Normalization

- 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).

McMurdie & Holmes (2014); Weiss et al. (2015)

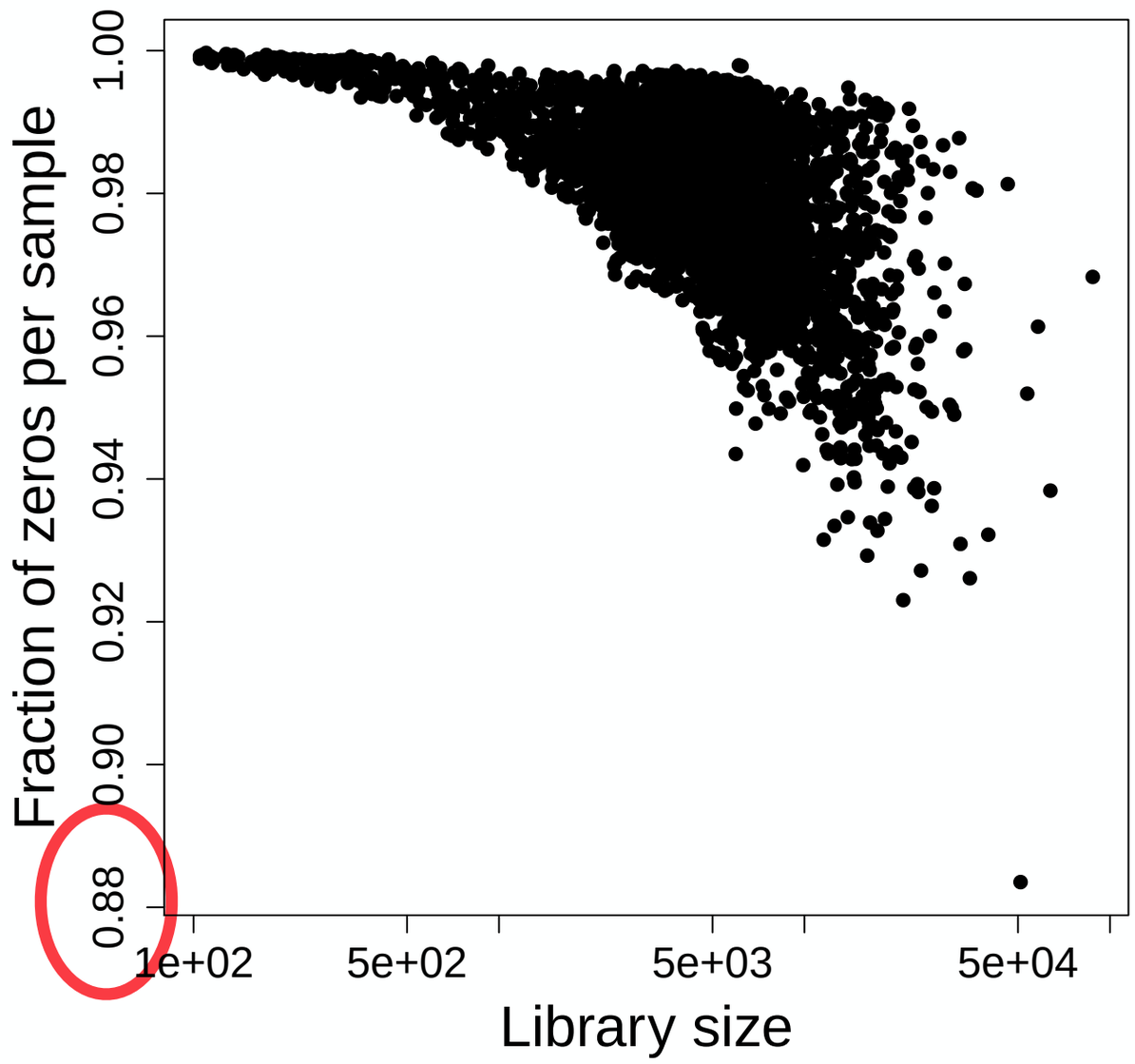
Scaling normalization - challenges

- 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.



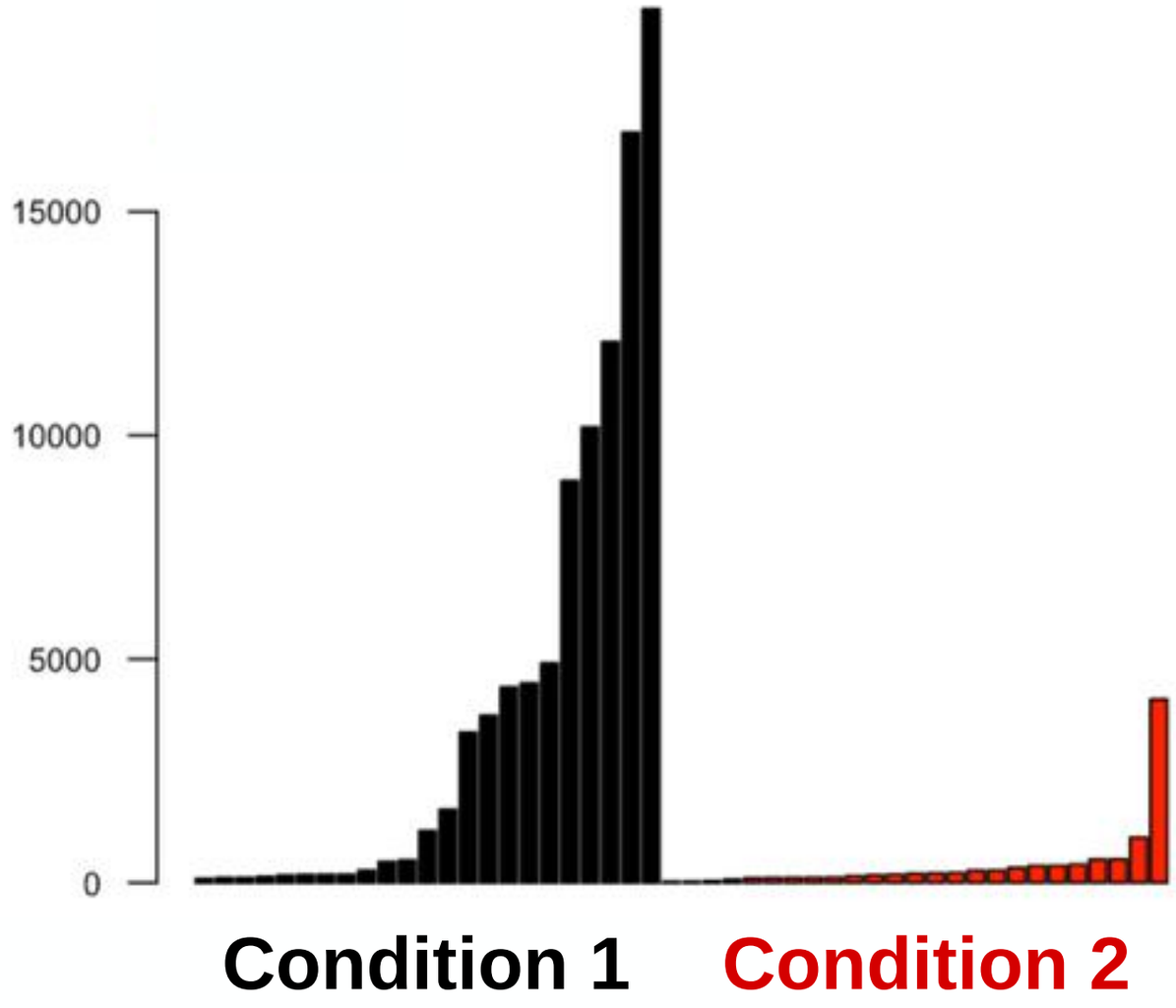
Scaling normalization - challenges

- Lots of zero counts!
- Many of them are likely due to undersampling - would be nonzero if the library size increased



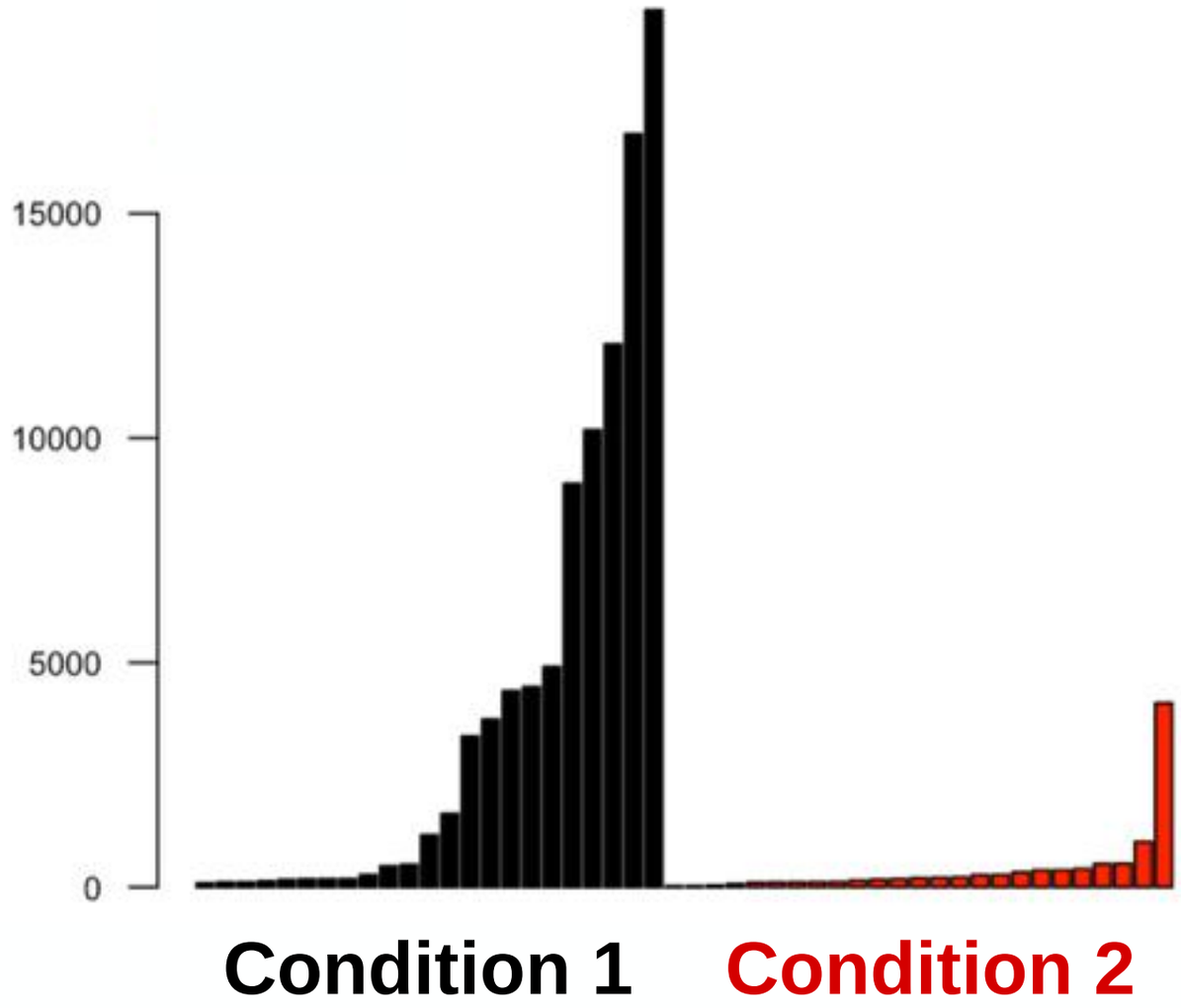
Differential abundance testing

- 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



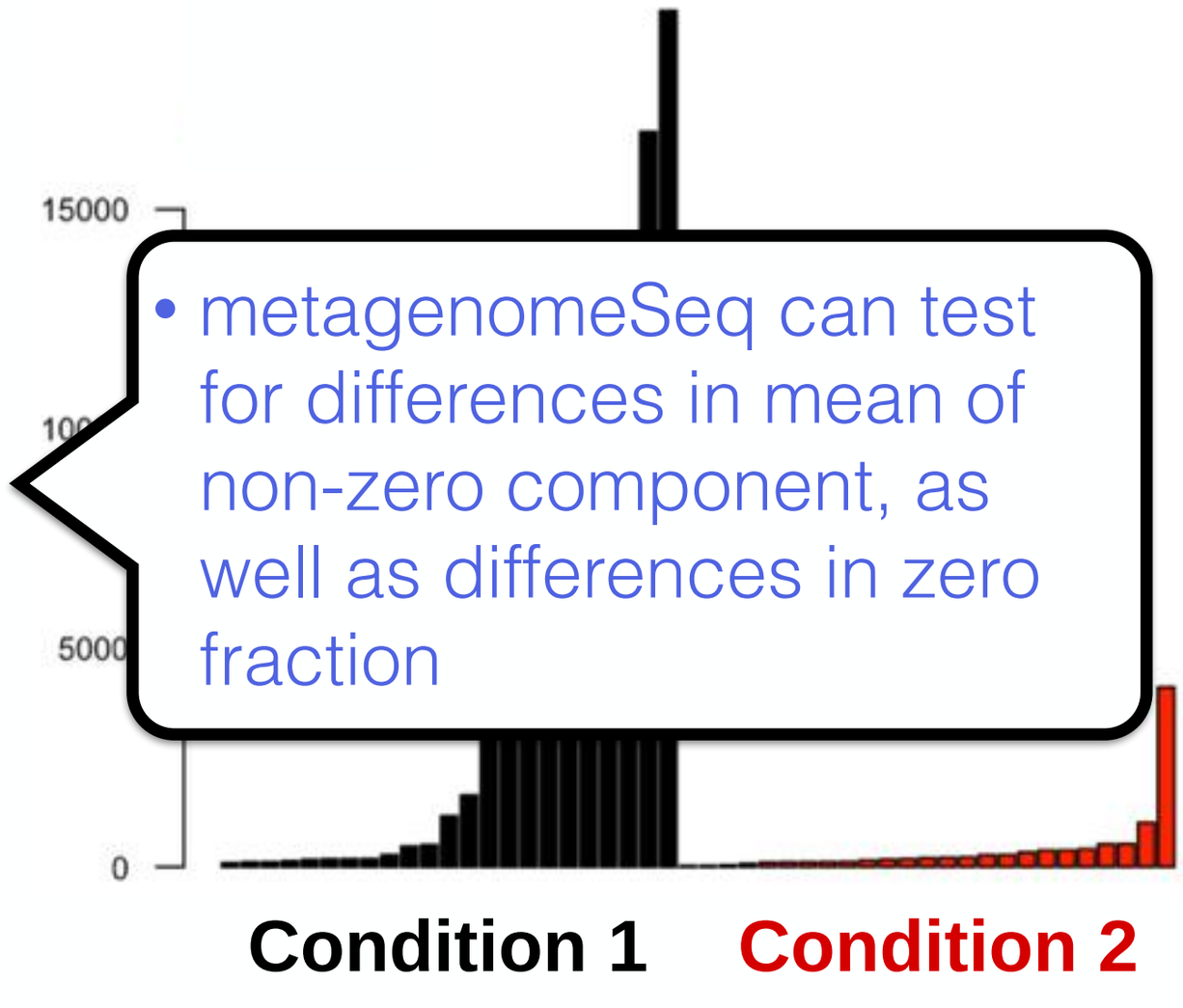
Differential abundance testing

- What do we want to test?
  - Difference between mean abundance
    - RNA-seq methods like edgeR and DESeq2 can be used, but
      - be careful when comparing very different habitats
      - “outliers” may influence the results



Differential abundance testing

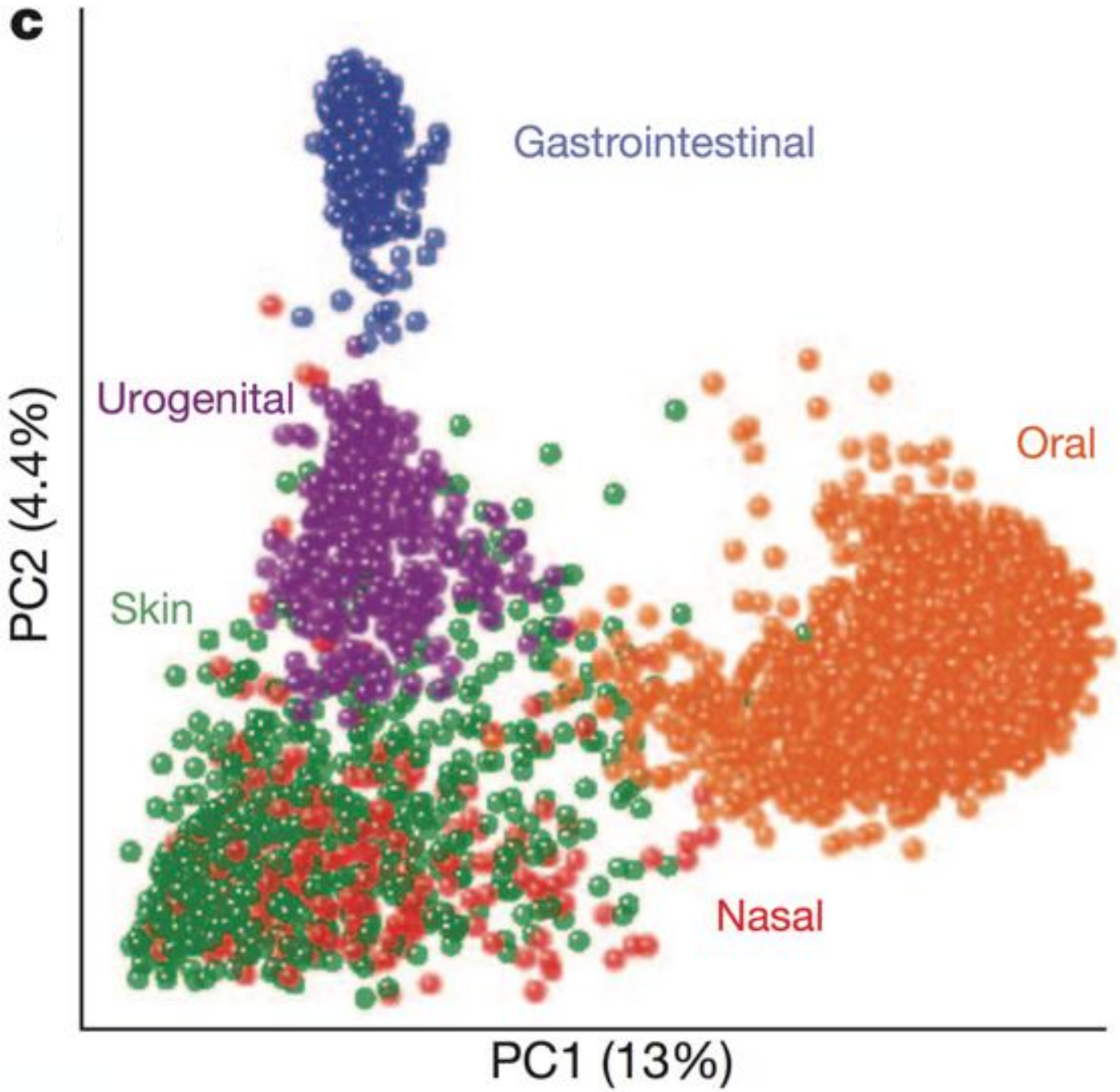
- What do we want to test?
  - Difference between mean abundance
  - Difference in fraction of zeros
  - Difference between mean abundance conditioning on being present
  - Overall difference in OTU composition



Visualization/ordination

- Calculate pairwise dissimilarities between samples (a.k.a. **beta diversity**)
  - Based on presence/absence of OTUs ("unweighted")
  - Incorporating abundances of OTUs ("weighted")
- Common dissimilarity measures: UniFrac, Bray-Curtis
- Generate low-dimensional representation that preserves these distances

Visualization/ordination



Human Microbiome Project (2012)

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)
- 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. Enviromental Microbiology 17(5):1689-1706 (2015) - **OTU clustering**
- Caporaso et al.: QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335-336 (2010) - **QIIME**
- Schloss et al.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75(23):7537-7541 (2009) - **mothur**
- 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)