Microbiome Data Analysis

Kaiden Liu

30/06/2021

Motivation

What is DNA sequencing

- DNA sequencing is the process of determining the nucleic acid sequence.
- Comparing healthy and mutated DNA sequences can diagnose different diseases

High-throughput sequencing

- also known as Next Generation Sequencing
- allows the entire DNA strand to be sequenced at once by breaking it into small pieces, and sequenced them all at once
- lower computational cost per sample/ quicker

Applications

- Reduced Lactobacillus species in the vagina is a risk factor for premature birth [1][2]
- detect pathogen in human respiratory system and brain biopsies
- Low biomass biological specimens have less reproducible sequences [3][4]

---

[1] Callahan et al., 2017; DiGiulio et al., 2015

[2] Jeganathan et al., 2021

[3] Gu et al., 2019; Langelier et al., 2018; Schlaberg et al., 2017b; Wilson et al., 2014; Brown et al., 2018

[4] Jeganathan et al., 2021

Low biomass specimens

- eg) saliva, blood
- Low biomass specimens produce low abundance of DNA

Challenges

- sensitive to contamination (eg: reagent to extract DNA, lab environment)

Challenges

- library depth
  - microbiome samples are sequenced at the same time, but they often result in total different numbers of sequences
    - proportional abundance
    - rarefy abundances
  - Improvements provided by with hierarchical mixture model
- batch effect
  - non-biological factors in an experiment causes changes in the data.

[5] Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible(McMurdie and Holmes, 2014)

Phyloseq

1) OTU table

- row: taxa/OTU
  - groups of closely related bacteria based on sequence similarity
- column: sample
- value: the number of reads

```
##       Sample1 Sample2 Sample3 Sample4 Sample5 Sample6 Sample7 Sample8 Sample9
## OTU1        0       0      54      18       0       0       0       0      91
## OTU2        0      69      71      65      33       0       0      98      17
## OTU3        0      90      34      80      47       0      34       0      73
## OTU4        4      25       3       0      85      79       0      26       0
## OTU5        1       4      92       0      35       0      79       0      37
## OTU6       51       0       0       0      29      73      11       0      46
## OTU7        0      63       0      64       0       0      25       0       8
## OTU8        0       9       0      97      37       0      65       0      59
## OTU9       75       0       0       0      81      76      95       0      56
## OTU10      36      25       0       2       0       0      76      98      80
##       Sample10
## OTU1        71
## OTU2        57
## OTU3        82
## OTU4        36
## OTU5        67
## OTU6         9
## OTU7         0
## OTU8        74
## OTU9        85
## OTU10       85
```

2) taxonomy table

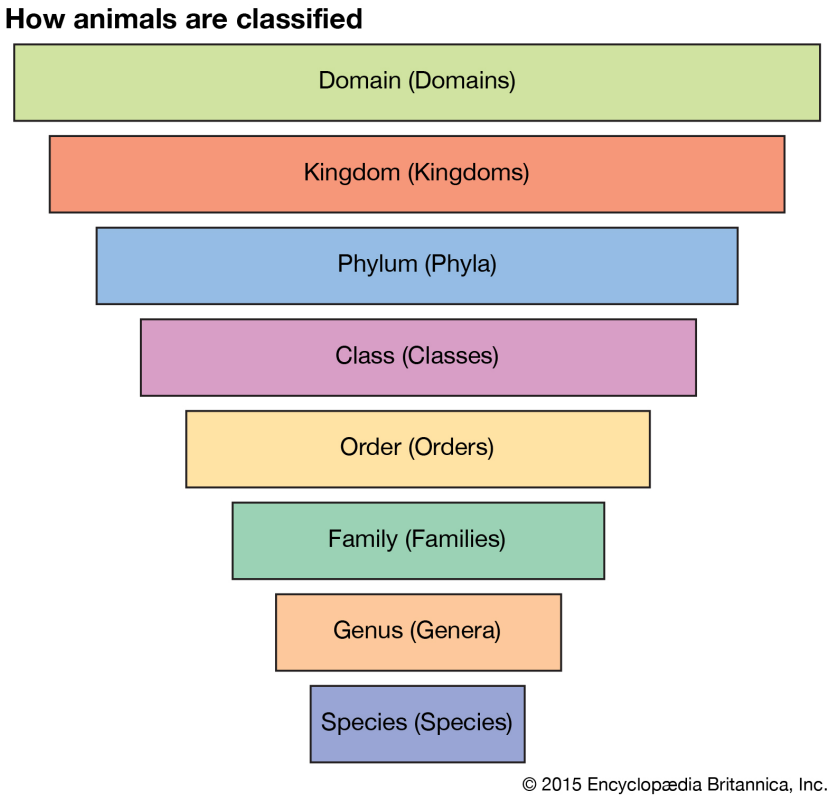


Figure 1: taxonomy rank

Taxonomy table

Table of names of the taxonomic rank of the data

- row: OTU
- column: taxonomy rank(levels)
- value: names of the taxonomy

```
##       Domain Phylum Class Order Family Genus Species
## OTU1  "f"    "z"    "t"   "b"   "j"    "l"   "u"
## OTU2  "s"    "r"    "v"   "w"   "h"    "e"   "q"
## OTU3  "u"    "j"    "n"   "h"   "z"    "r"   "f"
## OTU4  "e"    "r"    "p"   "s"   "s"    "n"   "k"
## OTU5  "s"    "d"    "t"   "o"   "o"    "y"   "w"
## OTU6  "r"    "g"    "m"   "z"   "f"    "b"   "z"
## OTU7  "i"    "o"    "r"   "r"   "x"    "s"   "e"
## OTU8  "u"    "b"    "r"   "h"   "w"    "m"   "y"
## OTU9  "c"    "x"    "l"   "h"   "o"    "s"   "v"
## OTU10 "r"    "w"    "w"   "l"   "h"    "e"   "l"
```

Taxonomy Table

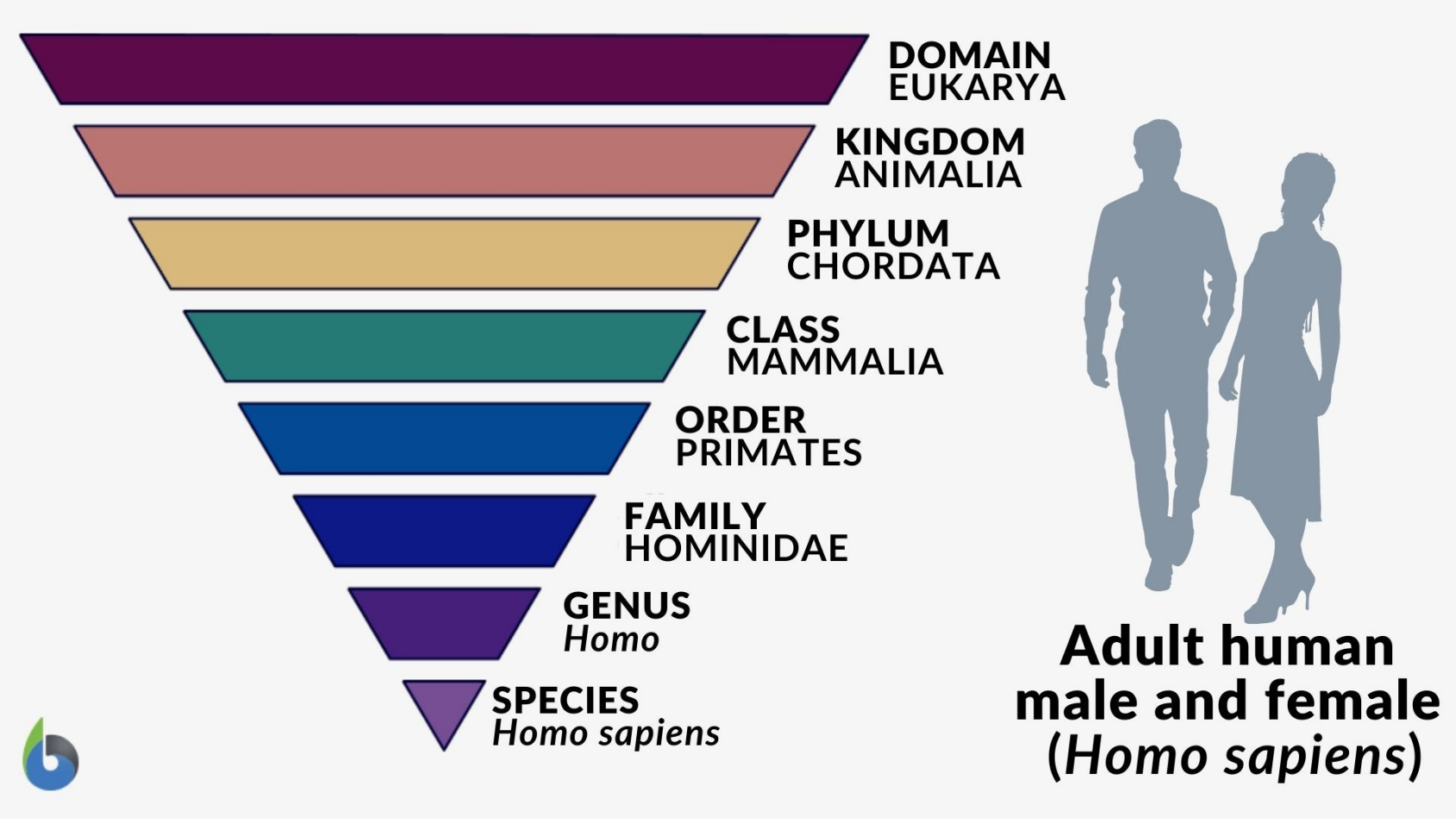


Figure 2: human

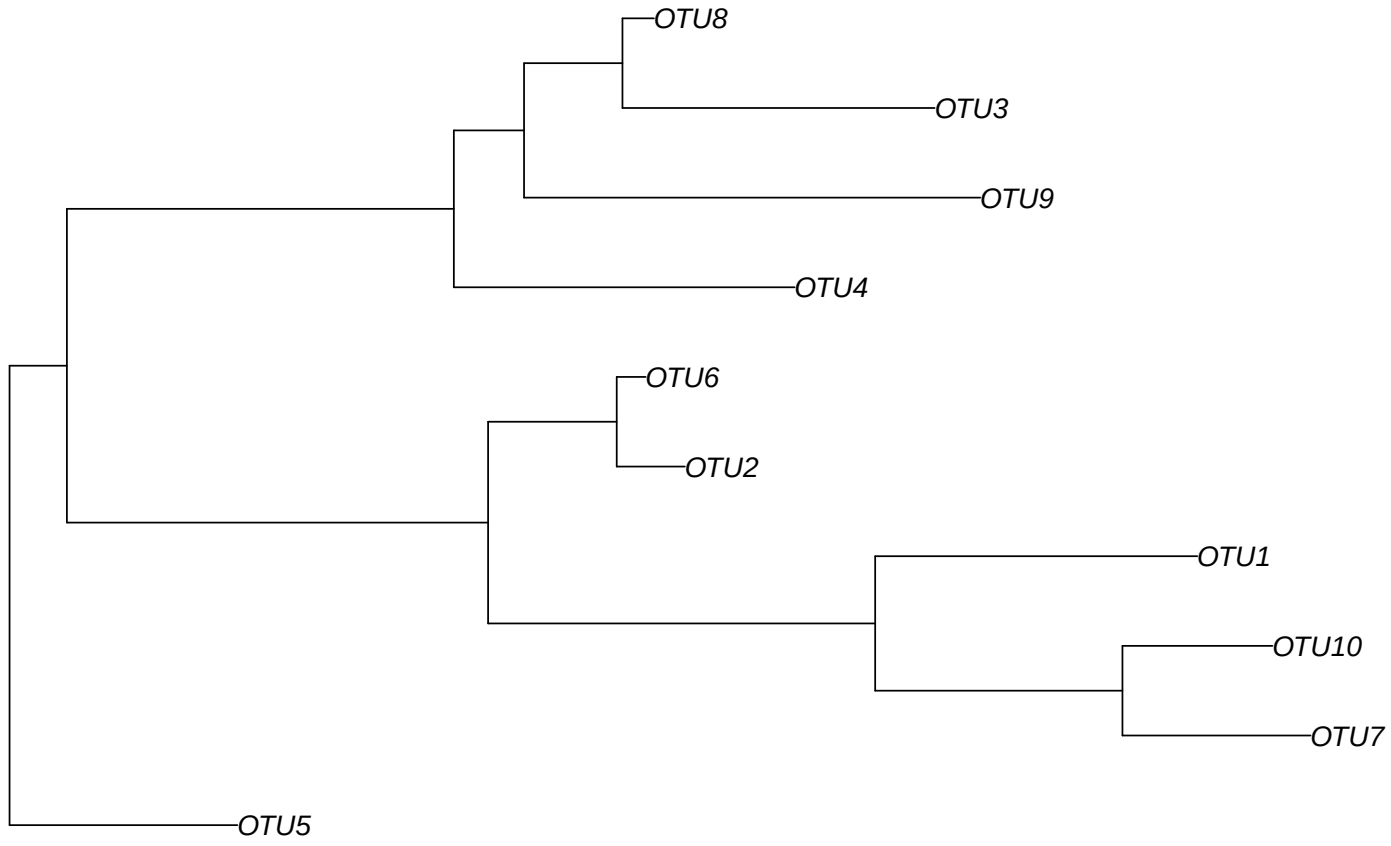
3) sample variables

- Location: Location of where the sample is collected(eg: Feces, Blood, Skin)
- Depth: Number of times sample has been read

```
##          Location Depth
## Sample1         C   340
## Sample2         D   328
## Sample3         A   256
## Sample4         C   630
## Sample5         A   440
## Sample6         B   727
## Sample7         A   623
## Sample8         B   452
## Sample9         B   540
## Sample10        B   131
```

4) phylogenetic Tree

Shows how the different taxa are related.



5) Phyloseq object

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 10 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 2 sample variables ]
## tax_table()   Taxonomy Table:    [ 10 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 10 tips and 9 internal nodes ]
```

Application

Zymo

```
path <- file.path("RawSeq")
list.files(path)
```

```
##  [1] "filtered"           "in1299_1_R1.fastq"  "in1299_1_R2.fastq"
##  [4] "in1299_10_R1.fastq" "in1299_10_R2.fastq" "in1299_11_R1.fastq"
##  [7] "in1299_11_R2.fastq" "in1299_12_R1.fastq" "in1299_12_R2.fastq"
## [10] "in1299_13_R1.fastq" "in1299_13_R2.fastq" "in1299_14_R1.fastq"
## [13] "in1299_14_R2.fastq" "in1299_15_R1.fastq" "in1299_15_R2.fastq"
## [16] "in1299_16_R1.fastq" "in1299_16_R2.fastq" "in1299_17_R1.fastq"
## [19] "in1299_17_R2.fastq" "in1299_18_R1.fastq" "in1299_18_R2.fastq"
## [22] "in1299_2_R1.fastq"  "in1299_2_R2.fastq"  "in1299_3_R1.fastq"
## [25] "in1299_3_R2.fastq"  "in1299_4_R1.fastq"  "in1299_4_R2.fastq"
## [28] "in1299_5_R1.fastq"  "in1299_5_R2.fastq"  "in1299_6_R1.fastq"
## [31] "in1299_6_R2.fastq"  "in1299_7_R1.fastq"  "in1299_7_R2.fastq"
## [34] "in1299_8_R1.fastq"  "in1299_8_R2.fastq"  "in1299_9_R1.fastq"
## [37] "in1299_9_R2.fastq"
```

fastq file

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

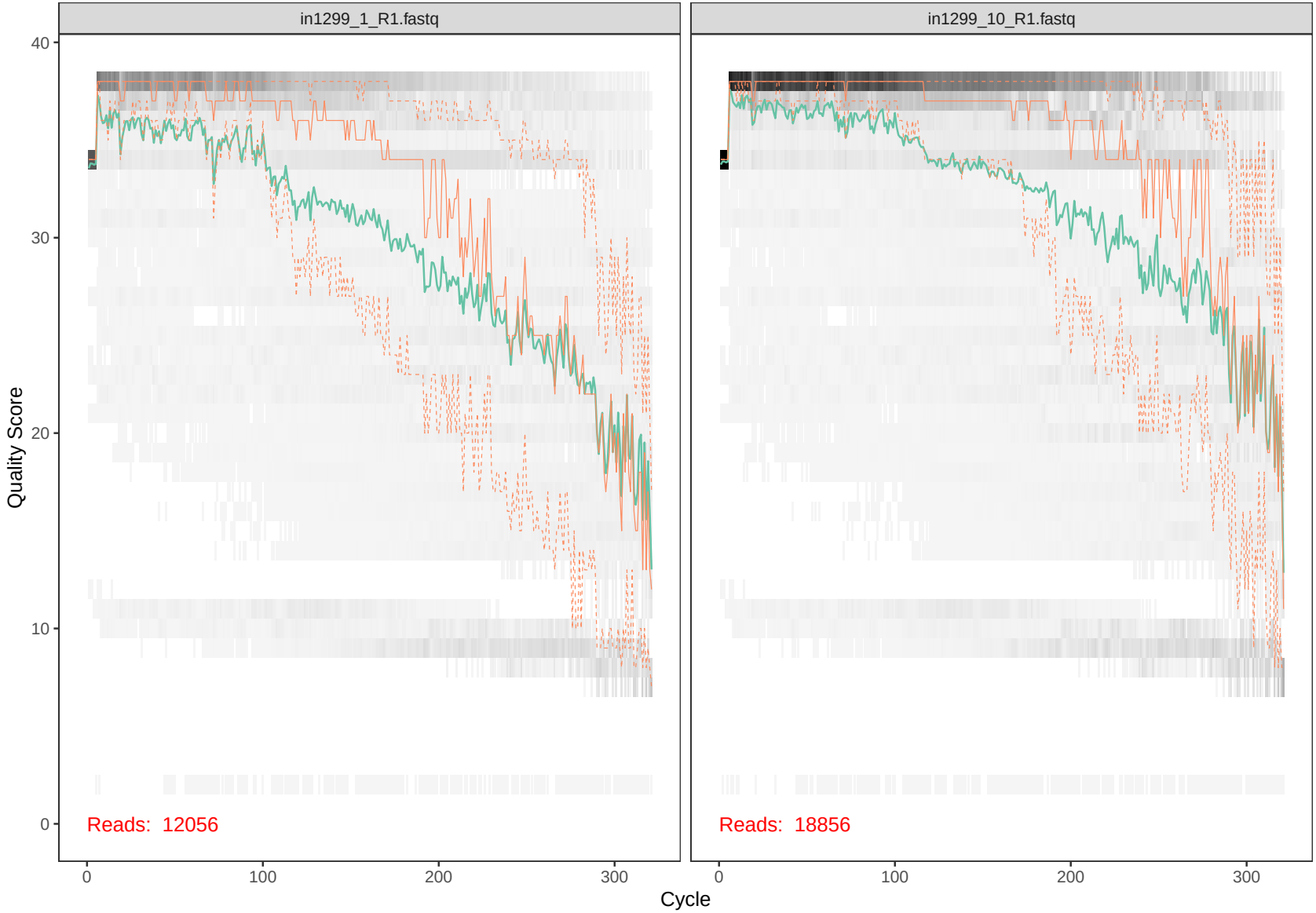
Figure 3: fastq file

1) Name of the read
2) Sequence
3) place holder line
4) Quality score with respect to each base in the sequence.

Let's learn about the quality score

Quality Score

```
plotQualityProfile(fnFs[1:2])
```



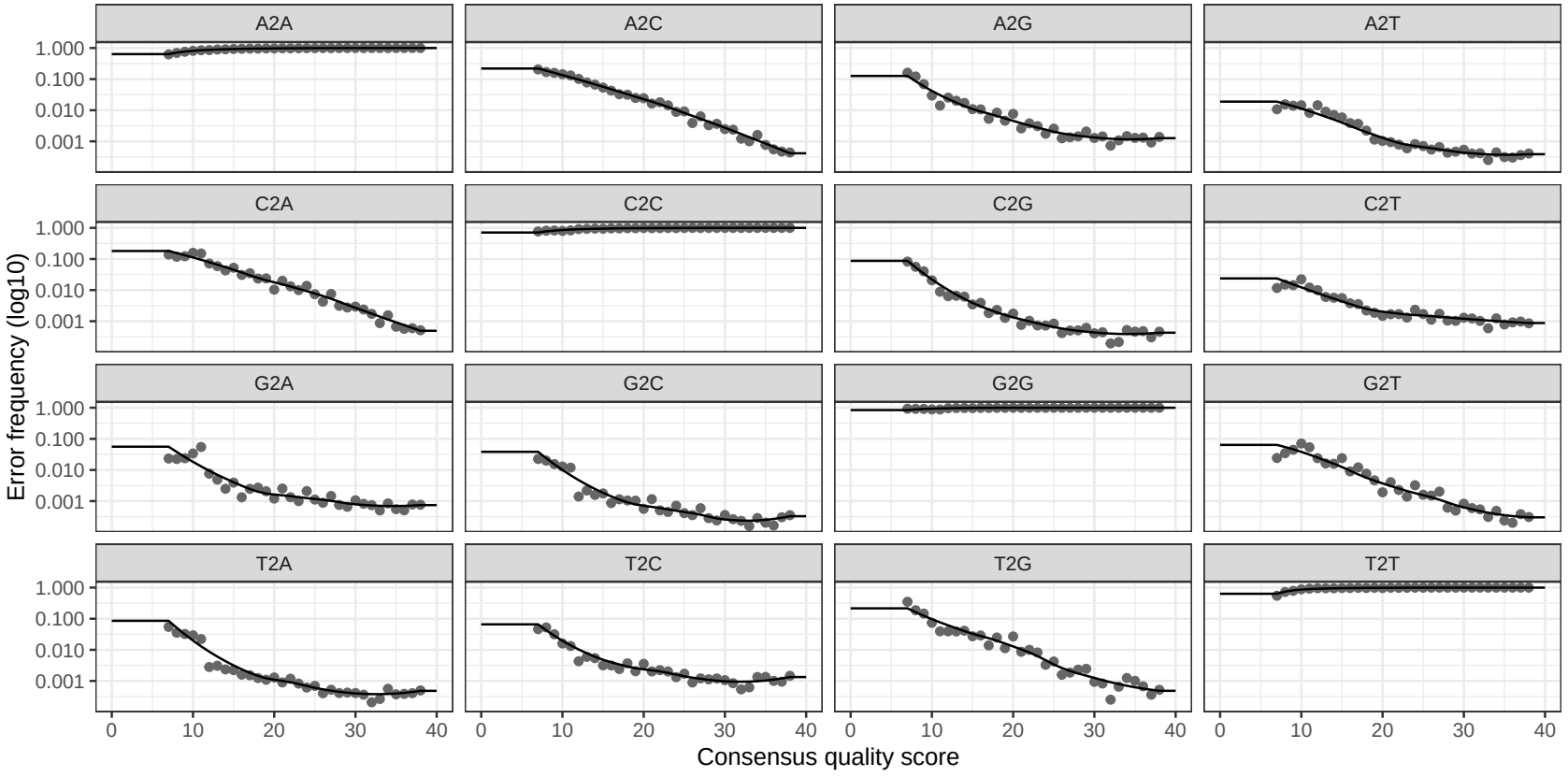
Trim

```
head(out)
```

```
##                          reads.in reads.out
## in1299_1_R1.fastq           12056      10502
## in1299_10_R1.fastq          18856      17435
## in1299_11_R1.fastq          15372      14515
## in1299_12_R1.fastq          24292      23059
## in1299_13_R1.fastq          16309      14967
## in1299_14_R1.fastq          22322      21134
```

Generating model of our data

```
plotErrors(errF)
```



Black line: estimated error rates

Black dots: observed error rates for each consensus quality score.

Sample Inference

```
dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
dadaRs <- dada(filtRs, err=errR, multithread=TRUE)
```

Removes all sequencing errors to reveal the true biological sequences.

6

[6] DADA2: High-resolution sample inference from Illumina amplicon data, https://www.nature.com/articles/nmeth.3869#methods

Merger

```
mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE, minOverlap = 12)
```

Reconstruct the full target sequence by merging each denoised pair of forward and reverse reads, rejecting any pairs which do not sufficiently overlap

7

[7] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016).

Track reads through the pipeline

```
##              input filtered merged
## in1299_1   12056    10502  10134
## in1299_10 18856    17435  16957
## in1299_11 15372    14515  13283
## in1299_12 24292    23059  22125
## in1299_13 16309    14967  14658
## in1299_14 22322    21134  20655
```

Most reads drops in the filter step but not the merge steps, which is a good sign.

Assign taxonomy

```
taxa <- assignTaxonomy(seqtab.nochim, "./tax/silva_nr99_v138.1_train_set.fa.gz", multithread = TRUE)

taxa.print <- taxa
```

Assign taxonomy

```
head(taxa.print)
```

```
##      Kingdom    Phylum           Class                  Order
## [1,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Burkholderiales"
## [2,] "Bacteria" "Firmicutes"     "Bacilli"             "Bacillales"
## [3,] "Bacteria" "Firmicutes"     "Bacilli"             "Lactobacillales"
## [4,] "Bacteria" "Firmicutes"     "Bacilli"             "Staphylococcales"
## [5,] "Bacteria" "Firmicutes"     "Bacilli"             "Lactobacillales"
## [6,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Enterobacterales"
##      Family               Genus
## [1,] "Burkholderiaceae"   "Ralstonia"
## [2,] "Bacillaceae"        "Bacillus"
## [3,] "Listeriaceae"       "Listeria"
## [4,] "Staphylococcaceae"  "Staphylococcus"
## [5,] "Lactobacillaceae"   "Limosilactobacillus"
## [6,] "Enterobacteriaceae" "Escherichia-Shigella"
##      Species
## [1,] "detusculanense/insidiosa/mannitolilytica/pickettii/solanacearum/syzygii"
## [2,] "altitudinis/amyloliquefaciens/firmus/halotolerans/intestinalis/licheniformis/mojavensis/siamensis
## [3,] "innocua/ivanovii/marthii/monocytogenes/phage/seeligeri/welshimeri"
## [4,] "argenteus/aureus/equorum/phage/schweitzeri/simiae"
## [5,] NA
## [6,] NA
```

Assign taxonomy

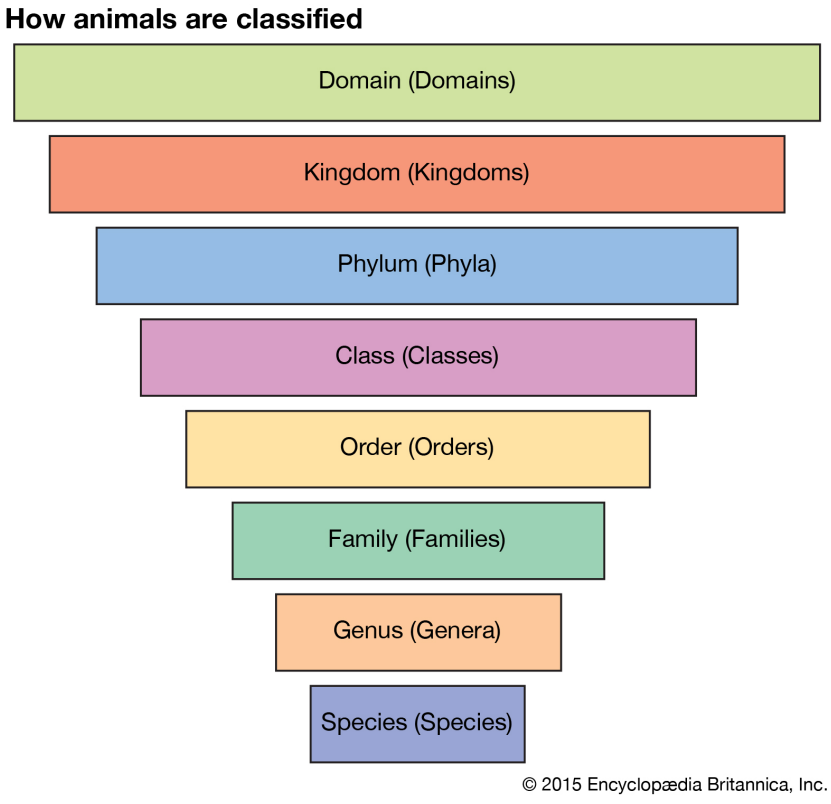


Figure 4: taxonomy rank

```
head(rownames_ex)
```

```
## [1] "AGTGGGGAATTTTGGACAATGGGCGAAAGCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTGTCCGGAAAGAA
## [2] "AGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAGAA
## [3] "AGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGTATGAAGAAGGTTTTCGGATCGTAAAGTACTGTTGTTAGAGAAGAA
## [4] "AGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGAAGGTCTTCGGATCGTAAAACTCTGTTATTAGGGAAGAA
## [5] "AGTAGGGAATCTTCCACAATGGGCGCAAGCCTGATGGAGCAACACCGCGTGAGTGAAGAAGGGTTTCGGCTCGTAAAGCTCTGTTGTTAAAGAAGAA
## [6] "AGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAA
```

Phyloseq

```
ps <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows=FALSE),
               sample_data(samdf),
               tax_table(taxa))
ps
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 195 taxa and 18 samples ]
## sample_data() Sample Data:       [ 18 samples by 4 sample variables ]
## tax_table()   Taxonomy Table:    [ 195 taxa by 7 taxonomic ranks ]
```

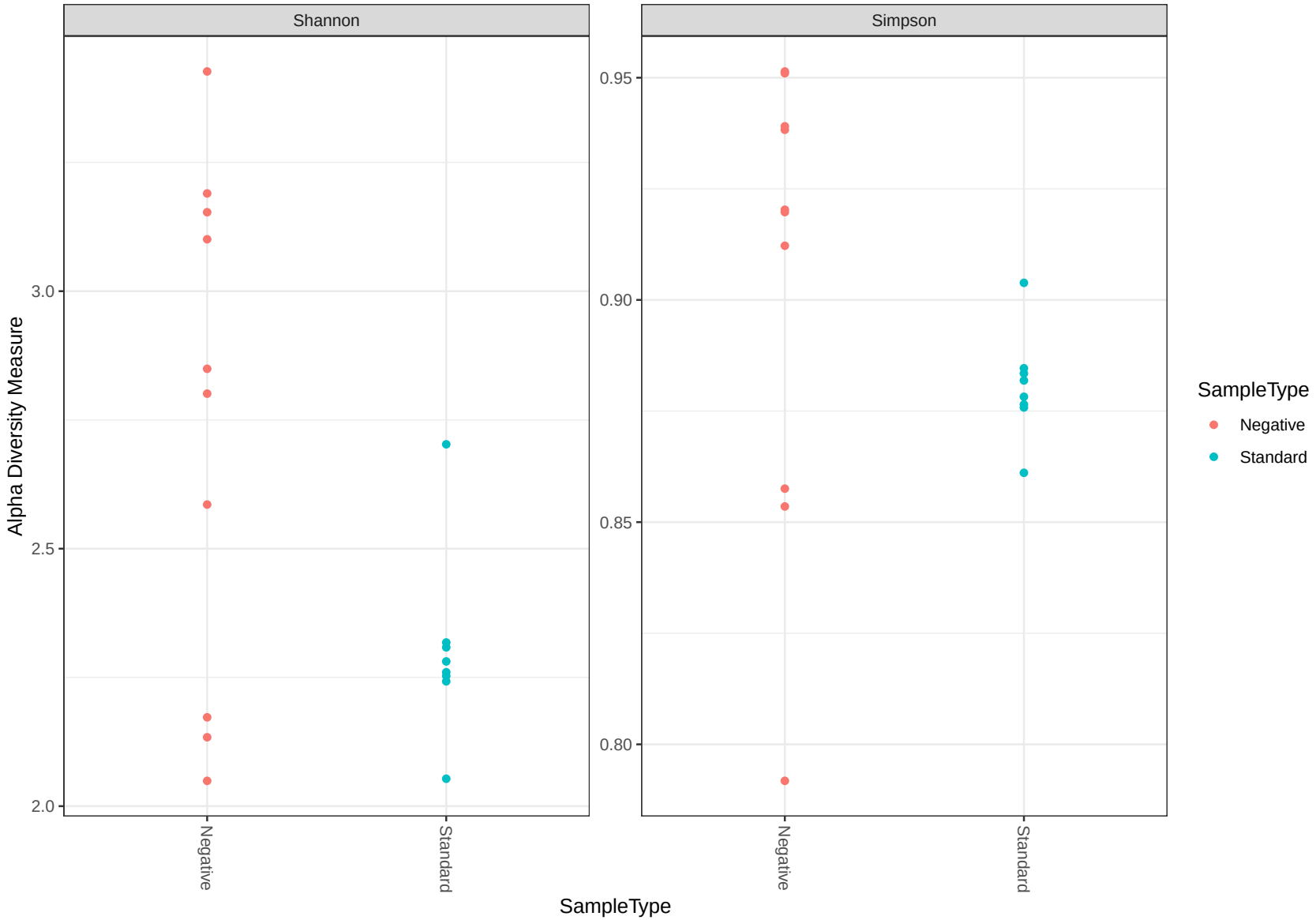
Exploratory analysis

Sample data

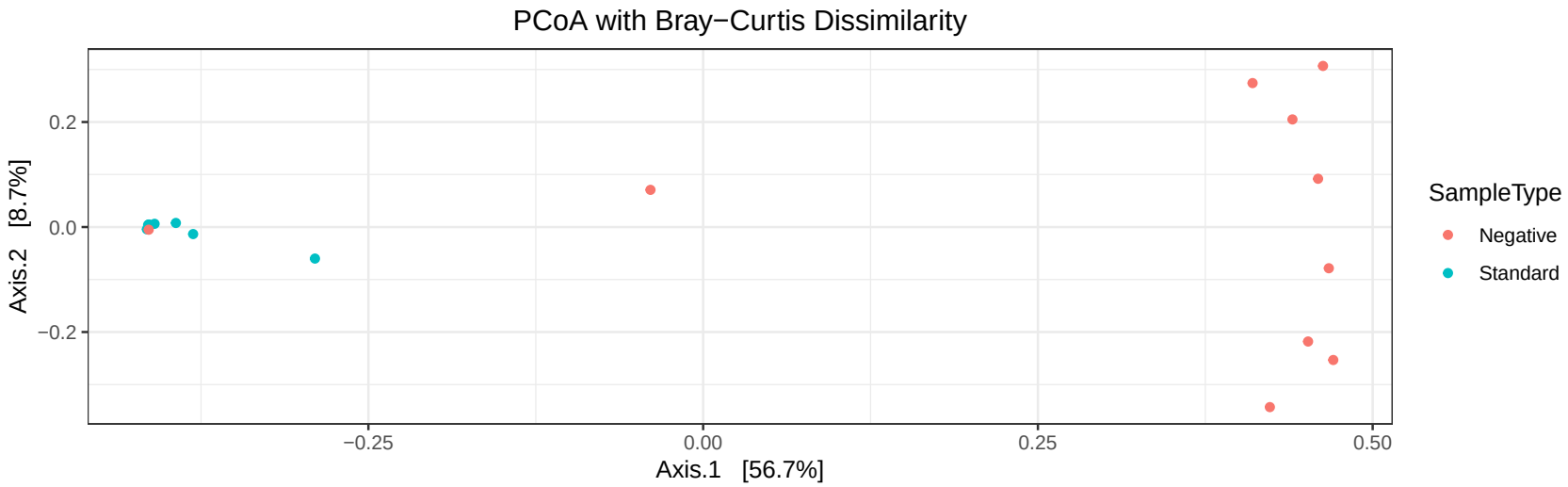
	SubjectID	sampleID	SampleType	Name
in1299_1	in1299.1	in1299_1	Negative	NegativeControl.1
in1299_10	in1299.10	in1299_10	Negative	NegativeControl.10
in1299_11	in1299.11	in1299_11	Standard	Standard.Dilution.1.1
in1299_12	in1299.12	in1299_12	Standard	Standard.Dilution.1.6
in1299_13	in1299.13	in1299_13	Standard	Standard.Dilution.1.36
in1299_14	in1299.14	in1299_14	Standard	Standard.Dilution.1.216
in1299_15	in1299.15	in1299_15	Standard	Standard.Dilution.1.1296
in1299_16	in1299.16	in1299_16	Standard	Standard.Dilution.1.7776
in1299_17	in1299.17	in1299_17	Standard	Standard.Dilution.1.46656
in1299_18	in1299.18	in1299_18	Standard	Standard.Dilution.1.279936
in1299_2	in1299.2	in1299_2	Negative	NegativeControl.2
in1299_3	in1299.3	in1299_3	Negative	NegativeControl.3
in1299_4	in1299.4	in1299_4	Negative	NegativeControl.4
in1299_5	in1299.5	in1299_5	Negative	NegativeControl.5
in1299_6	in1299.6	in1299_6	Negative	NegativeControl.6
in1299_7	in1299.7	in1299_7	Negative	NegativeControl.7
in1299_8	in1299.8	in1299_8	Negative	NegativeControl.8
in1299_9	in1299.9	in1299_9	Negative	NegativeControl.9

Alpha-Diversity

```
plot_richness(ps, x="SampleType", measures = c("Shannon", "Simpson"), color = "SampleType")
```



MDS plots


PCoA with Bray−Curtis Dissimilarity
0.2
0.0
−0.2
Axis.2 [8.7%]
−0.25
0.00
0.25
0.50
Axis.1 [56.7%]
SampleType
Negative
Standard

MDS plots

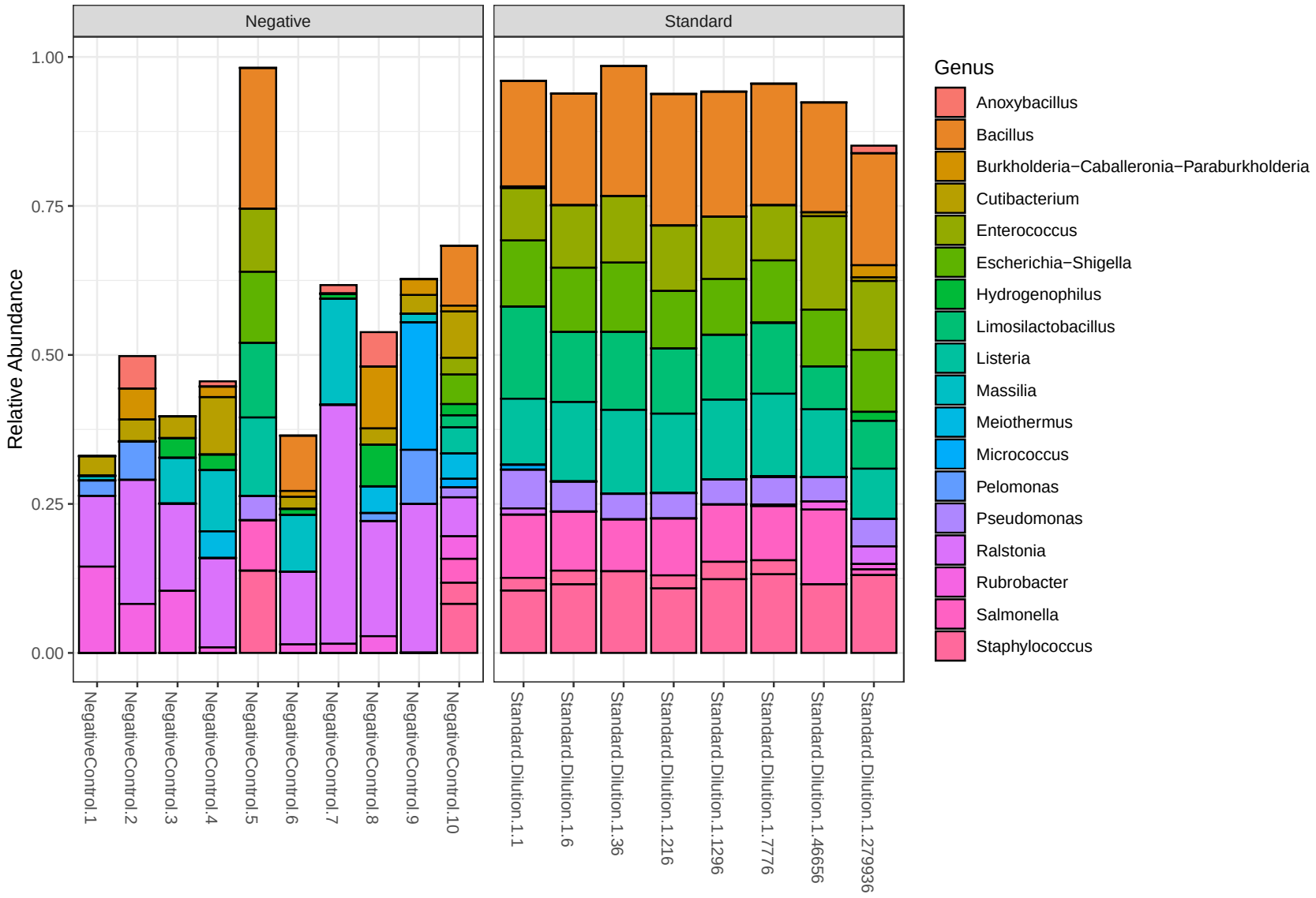
We want to map our data from a high dimension to a low dimension, so we can visualize the similarities of data(by how close they are).

The biggest difference of a PCoA is the construction of a distance matrix.

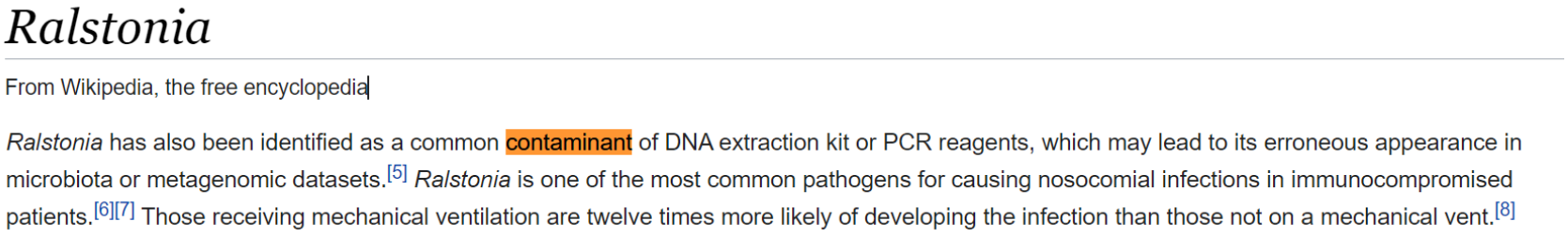
- For abundance data, Bray-Curtis distance is often recommended.

Top 20 ASVs in both control and dilution series samples

```
top20 <- names(sort(taxa_sums(ps), decreasing=TRUE))[1:20]
ps.top20 <- transform_sample_counts(ps, function(OTU) OTU/sum(OTU))
ps.top20 <- prune_taxa(top20, ps.top20)
plot_bar(ps.top20, x="Name", fill="Genus") + facet_wrap(~SampleType, scales="free_x") + labs(y="Relative A
```



Contaminant ASVs



Ralstonia

From Wikipedia, the free encyclopedia

Ralstonia has also been identified as a common contaminant of DNA extraction kit or PCR reagents, which may lead to its erroneous appearance in microbiota or metagenomic datasets.[5] *Ralstonia* is one of the most common pathogens for causing nosocomial infections in immunocompromised patients.[6][7] Those receiving mechanical ventilation are twelve times more likely of developing the infection than those not on a mechanical vent.[8]

Figure 5: ralstoniaWikipedia

Zymobiomic is a DNA extraction kit.

Plot heatmap

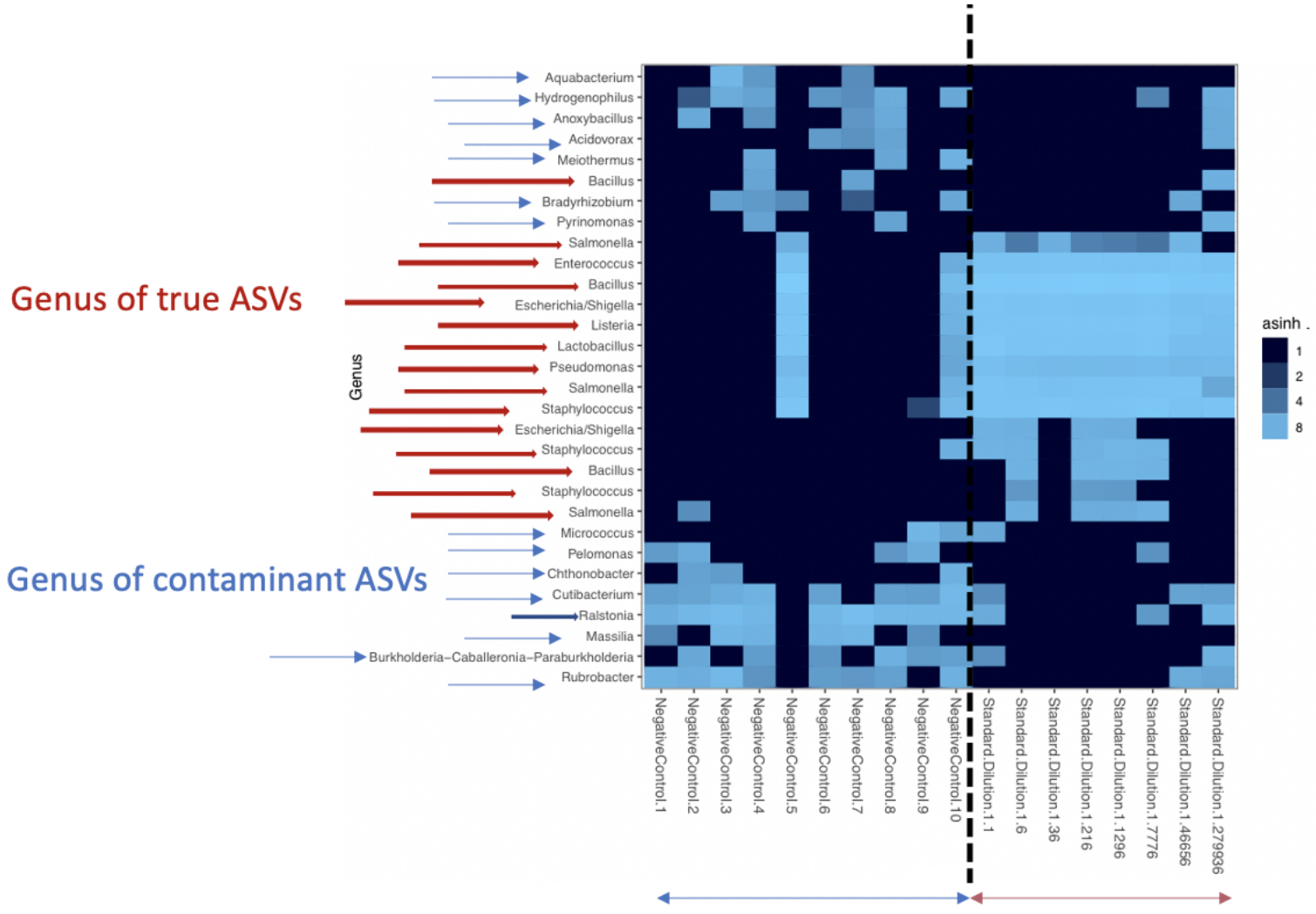
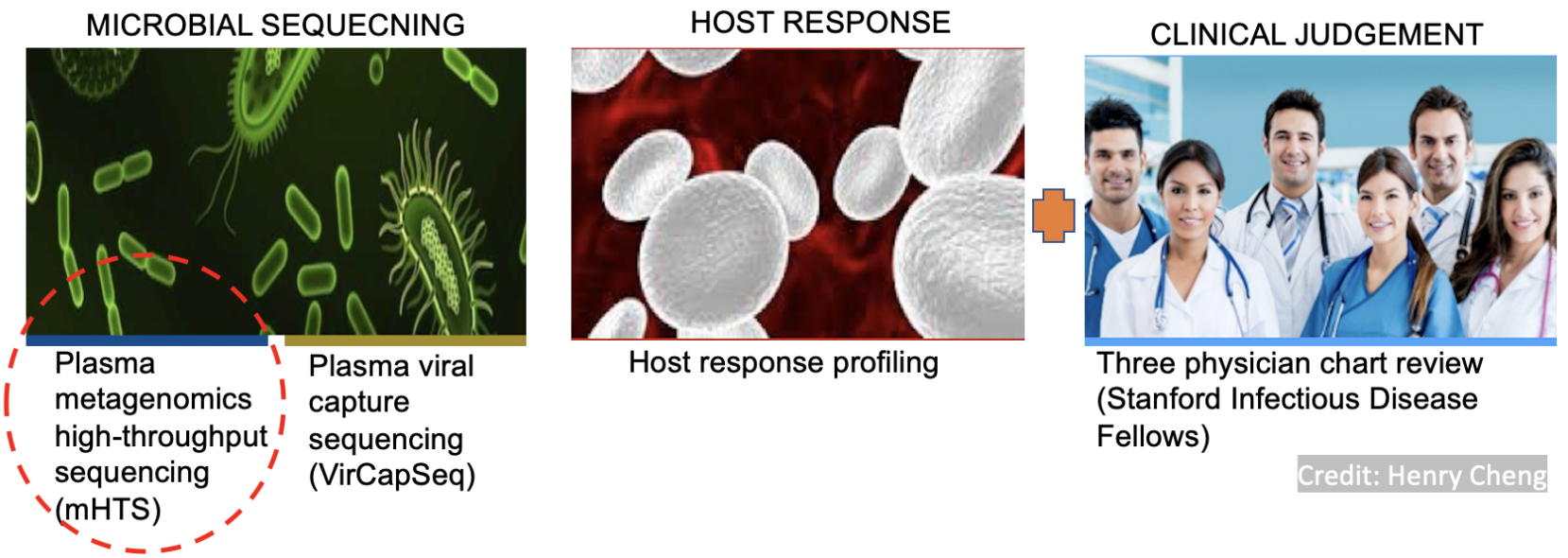


Figure 6: heatmap

What's next?

Challenges

Diagnosis of suspected sepsis



"Contaminant sequence identification and computational removal represents one of the greatest barriers to expanding the clinical application of metagenomic sequencing, especially in specimens with low microbial biomass such as blood."
[8]

[8] Combined use of metagenomic sequencing and host response profiling for the diagnosis of suspected sepsis (Cheng et al., 2019).

BARBI

Bayesian Reference analysis in the Background Interference

infers the true intensity of each taxon using of hierarchical gamma-Poisson mixture model

1) True reads intensity parameter
2) Contamination intensity parameter
3) Library depth effect
4) Metropolis-Hasting Markov chain Monte Carlo
5) 95% highest posterior density(HPD)

9

[9] A Bayesian Approach to Contamination Removal in Molecular Microbial Studies, (Jeganathan et al., 2021)