Nanopore\_16S\_analysis

## Analysing nanopore 16S rRNA sequences

Data source is from article “RESCUE: a validated Nanopore pipeline to classify bacteria through long-read, 16S-ITS-23S rRNA sequencing.” by Petrone et. al (2023). Sequences were downloaded from SRA and grouped into three.

Link to article: https://doi.org/10.3389/fmicb.2023.1201064

Zymo research mock community sample and gut mock community each contain 4 replicate samples. Oral microbiome set contains 20 saliva samples.

### Preprosessing

All reads were trimmed with cutadapt using the 1492R primer sequence

* cutadapt -g CGGTTACCTTGTTACGACTT –rc -o file input\_file

*Note, that sequences are reverse complement and 16S rRNA sequence starts from 3’ end*

*About 85-90 % contained correct primer sequence on default error rate 0.1.*

### Qiime 2 vsearch otu clustering workflow

Workflow: Importing sequences, dereplication, clustering, chimera check, chimera and singleton removal and taxa classification.

*Manifest file contains all sequence files with absolute path*

1. qiime tools import –type ‘SampleData[SequencesWithQuality]’  –input-path seq/manifest  –output-path output.qza –input-format SingleEndFastqManifestPhred33V2
2. qiime vsearch dereplicate-sequences –i-sequences seqs.qza –o-dereplicated-table dereplicated\_table.qza –o-dereplicated-sequences dereplicated\_sequences.qza
3. qiime vsearch cluster-features-de-novo –i-sequences dereplicated\_sequences.qza –i-table dereplicated\_table.qza –p-perc-identity 0.97 –o-clustered-sequences clustered\_otu\_sequences.qza –o-clustered-table clustered\_otu\_table.qza
4. qiime vsearch uchime-denovo –i-sequences clustered\_otu\_sequences.qza –output-dir chimeras –i-table clustered\_otu\_table.qza
5. qiime feature-table filter-features –i-table clustered\_otu\_table.qza –m-metadata-file chimeras/nonchimeras.qza —p-min-frequency 2  –o-filtered-table  final\_otu\_table.qza
6. qiime feature-table filter-seqs   –i-data clustered\_otu\_sequences.qza –i-table final\_otu\_table.qza  –o-filtered-data representative\_sequences.qza
7. qiime feature-classifier classify-sklearn –i-classifier silva-138-99-nb-classifier.qza –i-reads representative\_sequences.qza –p-read-orientation reverse –o-classification taxonomy.qza

### Qiime 2 dada denoising workflow

Workflow: Importing sequences, denoising and feature classification.

Note. Using value 1 error per 100 bp sequence produces best compromise.

1. qiime tools import –type ‘SampleData[SequencesWithQuality]’  –input-path seq/manifest  –output-path output.qza –input-format SingleEndFastqManifestPhred33V2
2. qiime dada2 denoise-single –p-max-ee 14 —p-trunc-len 0 —i-demultiplexed-sequences —output-dir dada
3. qiime feature-classifier classify-sklearn —i-classifier silva-138-99-classifier.qza —i-reads dada/representative\_sequences.qza —p-read-orientation reverse —o-classification dada/taxonomy.qza

### Loading R libraries

#required libraries  
library(mia)  
library(miaViz)  
library(tidyverse)  
library(ggplot2)  
library(ggthemes)  
library(kableExtra)  
library(reshape2)

### Zymo Research mock community samples

Importing data from qiime 2 files

#dada2 results  
zymo\_mock\_dada <- loadFromQIIME2(featureTableFile="Qiime2\_files/mock\_community/mock\_dada\_asv\_table.qza", taxonomyTableFile="Qiime2\_files/mock\_community/mock\_dada\_taxa.qza", sampleMetaFile="Qiime2\_files/mock\_community/meta\_dada.tsv")  
#clustering\_results  
zymo\_mock\_vsearch <- loadFromQIIME2(featureTableFile="Qiime2\_files/mock\_community/mock\_vsearch\_otu\_table.qza", taxonomyTableFile="Qiime2\_files/mock\_community/mock\_vsearch\_taxa.qza", sampleMetaFile="Qiime2\_files/mock\_community/meta\_clust.tsv")

Creating dataframe from zymo kit data

#create vectors and alphabetically sorted dataframe containing real mock composition  
genus <- c("Bacillus", "Enterococcus", "Escherichia-Shigella", "Lactobacillus", "Listeria", "Pseudomonas", "Salmonella", "Staphyloccus")  
abundance <- c(0.174,0.10,0.101,0.184,0.141,0.042,0.104,0.155)  
comparison <- data.frame(abundance)  
rownames(comparison) <- genus  
comparison <- tibble::rownames\_to\_column(comparison)

Creating data frame from dada2 results

#agglomeration to genus and calculating relabundance  
zymo\_mock\_dada <- agglomerateByRank(zymo\_mock\_dada, rank="Genus")  
zymo\_mock\_dada <- transformAssay(zymo\_mock\_dada, assay.type="counts",  
 method="relabundance", onRankOnly=TRUE)  
#create dataframe and table alphabetically sorted  
zm\_dada <- data.frame(assays(zymo\_mock\_dada)$relabundance)  
zm\_dada <- zm\_dada[order(row.names(zm\_dada)),]  
rownames(zm\_dada) <- genus   
zm\_dada <- tibble::rownames\_to\_column(zm\_dada)  
kable(zm\_dada, digits=2)

| rowname | Mock01 | Mock02 | Mock03 | Mock04 |
| --- | --- | --- | --- | --- |
| Bacillus | 0.23 | 0.23 | 0.28 | 0.27 |
| Enterococcus | 0.00 | 0.00 | 0.00 | 0.06 |
| Escherichia-Shigella | 0.00 | 0.06 | 0.00 | 0.01 |
| Lactobacillus | 0.19 | 0.17 | 0.19 | 0.17 |
| Listeria | 0.12 | 0.12 | 0.15 | 0.16 |
| Pseudomonas | 0.00 | 0.02 | 0.00 | 0.00 |
| Salmonella | 0.03 | 0.02 | 0.00 | 0.00 |
| Staphyloccus | 0.43 | 0.38 | 0.38 | 0.33 |

Creating data.frame from vsearch results

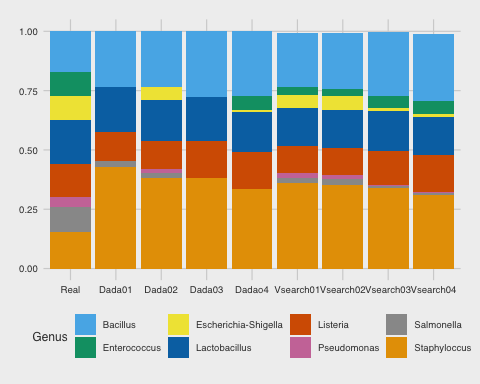
#agglomeration to genus and calculating relabundance  
zymo\_mock\_vsearch <- agglomerateByRank(zymo\_mock\_vsearch, rank="Genus")  
zymo\_mock\_vsearch <- transformAssay(zymo\_mock\_vsearch, assay.type="counts",  
 method="relabundance", onRankOnly=TRUE)  
#create dataframe and table alphabetically sorted  
zm\_vsearch <- data.frame(assays(zymo\_mock\_vsearch)$relabundance)  
zm\_vsearch <- zm\_vsearch[1:8,]  
zm\_vsearch <- zm\_vsearch[order(row.names(zm\_vsearch)),]  
rownames(zm\_vsearch) <- genus  
zm\_vsearch <- tibble::rownames\_to\_column(zm\_vsearch)  
kable(zm\_vsearch, digits=2)

| rowname | Mock01 | Mock02 | Mock03 | Mock04 |
| --- | --- | --- | --- | --- |
| Bacillus | 0.23 | 0.24 | 0.27 | 0.28 |
| Enterococcus | 0.03 | 0.03 | 0.05 | 0.06 |
| Escherichia-Shigella | 0.05 | 0.06 | 0.01 | 0.01 |
| Lactobacillus | 0.16 | 0.16 | 0.17 | 0.16 |
| Listeria | 0.11 | 0.11 | 0.14 | 0.15 |
| Pseudomonas | 0.02 | 0.02 | 0.01 | 0.01 |
| Salmonella | 0.02 | 0.02 | 0.01 | 0.01 |
| Staphyloccus | 0.36 | 0.35 | 0.34 | 0.31 |

Merging data frames

#merging dataframes and renaming columns  
summary\_df <- left\_join(comparison, zm\_dada, by="rowname")  
summary\_df <- left\_join(summary\_df, zm\_vsearch, by="rowname")  
colnames(summary\_df) <- c("Genus", "Real", "Dada01", "Dada02", "Dada03", "Dadao4", "Vsearch01", "Vsearch02", "Vsearch03", "Vsearch04")

molten\_summary <- melt(summary\_df, id.vars = 1)  
plot <- ggplot(molten\_summary, aes(x=variable,y=value, fill=Genus)) +   
geom\_bar(position="stack", stat = "identity") + theme\_fivethirtyeight(base\_size=9) + scale\_fill\_pander()  
plot



In this sample set, differences between dada and vsearch look minor. However, dada looks more sensitive as two out of four samples have less variants than supposed to.

When compared to real composition, some genus are underpresented (Salmonella, E. coli) and some overpresented (Bacillus, Streptococcus.