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

```r
#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_a
#clustering_results
zymo_mock_vsearch <- loadFromQIIME2(featureTableFile="Qiime2_files/mock_community/mock_vse
```

Creating dataframe from zymo kit data

```
#create vectors and alphabetically sorted dataframe containing real mock composition
genus <- c("Bacillus", "Enterococcus", "Escherichia-Shigella", "Lactobacillus", "Listeria"
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

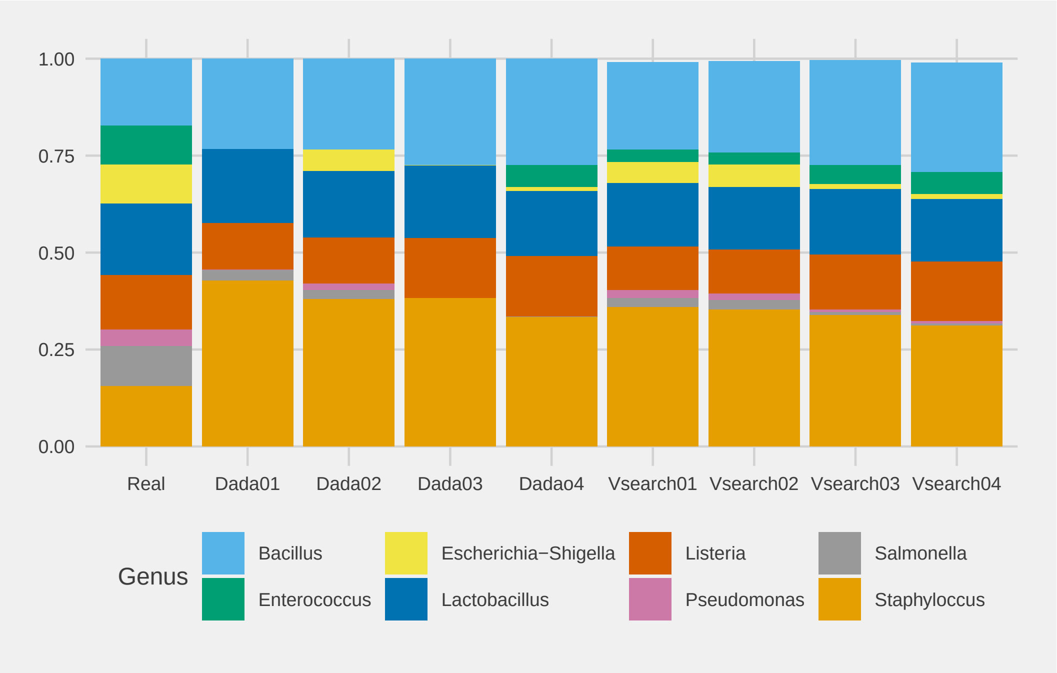
```
#agglomeration to genus and calculating relabundance
zymo_mock_vsearch <- agglomerateByRank(zymo_mock_vsearch, rank="Genus")
zymo_mock_vsearch <- transformAssay(zymo_mock_vsearch, assay.type="counts",
                                    method="relabundance", onRankOnly=TRUE)
#create dataframe and 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", "Vsearc
```

```
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
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.