Pipeline QA

Nate Olson

September 21, 2016

Sequence Processing

Loading sequencing data

```
mrexp_filenames <- list(mothur = "../data/mrexp_mothur.RDS",
                        qiime_denovo_chimerafilt = "../data/mrexp_qiime_denovo_chimera_filt.RDS",
                        qiime_denovo_nochimerafilt = "../data/mrexp_qiime_denovo_nochimera.RDS",
                        qiime_openref_chimerafilt = "../data/mrexp_qiime_refclus_chimera_filt.RDS",
                        qiime_openref_nochimerafilt = "../data/mrexp_qiime_refclus_nochimera.RDS")

mrexp_obj <- mrexp_filenames %>% map(readRDS)
fvarLabels(mrexp_obj$qiime_openref_chimerafilt) <- paste0("taxonomy",1:7)
```

```
## Loading required package: metagenomeSeq
## Loading required package: limma
##
## Attaching package: 'limma'
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
## Loading required package: glmnet
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following object is masked from 'package:S4Vectors':
##
##     expand
## The following object is masked from 'package:tidyr':
##
##     expand
## Loading required package: foreach
##
## Attaching package: 'foreach'
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
## Loaded glmnet 2.0-5
## Loading required package: RColorBrewer
```

Rename qiime samples for consistent set of ids

```
get_new_ids <- function(mr_qiime, sample_sheet){

      qiime_id_set <- pData(mr_qiime) %>% rownames()

      id_fix_df <- sample_sheet %>%
            filter(seq_lab == "JHU",barcode_lab == "JHU") %>%
            select(id, pcr_16S_plate, pos) %>%
            mutate(pos = str_replace(pos, "_",""),
                   qiime_id = str_c(pcr_16S_plate, pos, sep = "-")) %>%
            filter(qiime_id %in% qiime_id_set) %>%
            group_by(id) %>%
            mutate(id2 = if_else(grepl(x = id, pattern = "B0_M0"),
                                 paste(id,1:n(),sep = "_"),id))
      id_fix_df$id2[match(id_fix_df$qiime_id, qiime_id_set)]
}


id_set <- get_new_ids(mrexp_obj$qiime_denovo_chimerafilt, sample_sheet)
rownames(pData(mrexp_obj$qiime_denovo_chimerafilt)) <- id_set
colnames(assayData(mrexp_obj$qiime_denovo_chimerafilt)$counts) <- id_set

id_set <- get_new_ids(mrexp_obj$qiime_denovo_nochimerafilt, sample_sheet)
rownames(pData(mrexp_obj$qiime_denovo_nochimerafilt)) <- id_set
colnames(assayData(mrexp_obj$qiime_denovo_nochimerafilt)$counts) <- id_set

id_set <- get_new_ids(mrexp_obj$qiime_openref_chimerafilt, sample_sheet)
rownames(pData(mrexp_obj$qiime_openref_chimerafilt)) <- id_set
colnames(assayData(mrexp_obj$qiime_openref_chimerafilt)$counts) <- id_set

id_set <- get_new_ids(mrexp_obj$qiime_openref_nochimerafilt, sample_sheet)
rownames(pData(mrexp_obj$qiime_openref_nochimerafilt)) <- id_set
colnames(assayData(mrexp_obj$qiime_openref_nochimerafilt)$counts) <- id_set
```

Pipeline characteristics

- Section objectives
  - make non-quantitative statements
  - capturing differences in quality across samples
- Characterization of different pipelines
  - number of clusters
  - different taxonomic assignments
- Statements/ Figures showing how datasets behave
- number of assigned vs. non-assigned
- **TODO** difference in richness
  - need to figure out how I want to normalize/ transform the data prior to calculating diversity values
- number of features found across samples and replicates
- **TODO** Table - pipeline sequence budget
  - number of reads filtered due to low quality
  - number of reads merged
  - number of chimeras

Developing Code for characterizing pipeline results

Number of OTUs

```
mrexp_obj %>% map(nrow)
```

```
## $mothur
## Features
##    25739
##
## $qiime_denovo_chimerafilt
## Features
##    14326
##
## $qiime_denovo_nochimerafilt
## Features
##    24617
##
## $qiime_openref_chimerafilt
## Features
##     2832
##
## $qiime_openref_nochimerafilt
## Features
##    11381
```