Log Fold-Change Analysis

Nate Olson

2017-01-29

```
library(ProjectTemplate)
cwd <- getwd()
setwd("../")
load.project()
setwd(cwd)
```

Overview

Code for Analysis

Loading Pipeline Data

```
mrexp_files <- list(
      dada2  = "../data/mrexp_dada2.RDS",
      mothur = "../data/mrexp_mothur.RDS",
      qiime  = "../data/mrexp_qiime_refclus_nochimera.RDS"
)
mrexp <- mrexp_files %>% map(readRDS)

#Extracting metadata

meta_dat <- mrexp$mothur %>% pData()

##labeling PCR replicates
half1 <- paste(rep(c("A","B","C","D","E","F","G","H"), each = 6), 1:6, sep = "_")
sam_dat <- meta_dat %>%
      mutate(pcr_half = if_else(pos %in% half1, "1","2"),
             pcr_rep = paste0(pcr_16S_plate,":",pcr_half)) %>%
      select(sampleID, dilution,sam_names, pcr_rep) %>%
      dplyr::rename(samID = sam_names)
```

Subsetting data to focus on one biological replicate

Only looking at biological replicate E01JH0004, to avoid overfitting the data.

```
E01JH004_sams <- meta_dat %>%
      filter(sampleID == "E01JH0004") %>% .$sam_names
mrexp_004 <- mrexp %>%
      map(~.[,which(colnames(.) %in% E01JH004_sams)]) %>%
      map(~.[which(rowSums(MRcounts(.)) > 0), ])
```

Start of Log-Fold Change (Differential Abundance) Analysis

Log-Fold Change Variance

Looking at pre and post samples first

- Calculate variance for the observed log-fold change differences for pairwise combinations of PCR replicates
  - Does it make sense to look at the distribution of the logFC values - to estiamte 95% CI
  - Can also get quantiles (95% CI) from pairwise log-fold change values

```
E01JH004_pre_post_sams <- meta_dat %>%
      filter(sampleID == "E01JH0004", dilution %in% c(0,-1)) %>% .$sam_names
mrexp_004_pre_post <- mrexp %>%
      map(~.[,which(colnames(.) %in% E01JH004_pre_post_sams)]) %>%
      map(~.[which(rowSums(MRcounts(.)) > 0), ])
```

```
pre_post_meta <- meta_dat %>% filter(sampleID == "E01JH0004", dilution %in% c(0,-1))
pre_sams <- pre_post_meta %>% filter(dilution == 0) %>% .$sam_names
post_sams <- pre_post_meta %>% filter(dilution == -1) %>% .$sam_names
```

```
pre_post_mat <- mrexp_004_pre_post$dada2 %>% metagenomeSeq::cumNormMat()
```

```
## Default value being used.
```

```
get_logFC <- function(pre, post){
      pre_post_mat[,pre]/pre_post_mat[,post]
}

perm_logFC <- map2(rep(pre_sams,4), rep(post_sams, each = 4), get_logFC) %>%
      set_names(paste0("X", 1:16)) %>% as_data_frame() %>%
      add_column(feature_id = rownames(pre_post_mat)) %>%
      gather("perm","FC", -feature_id)
```

```
perm_logFC_summary <- perm_logFC %>% mutate(logFC = log2(FC+1)) %>%
      group_by(feature_id) %>%
      summarise(logFC_mean = mean(logFC), logFC_median = median(logFC),
                logFC_lq = quantile(logFC, 0.025, na.rm = TRUE),
                logFC_uq = quantile(logFC, 0.975, na.rm = TRUE),
                logFC_var = var(logFC))
```

Exploring FC Values

FC 0 - Post specific feature FC NaN - Pre and Post 0 FC Inf - Pre specific feature

Example Features for Exploration

Mean 0

Mean Inf

```
perm_logFC_summary %>% filter(feature_id == "Seq10")
```

```
## # A tibble: 1 × 6
##   feature_id logFC_mean logFC_median logFC_lq logFC_uq logFC_var
##        <chr>      <dbl>        <dbl>    <dbl>    <dbl>     <dbl>
## 1      Seq10        Inf          Inf 5.277137      Inf       Inf
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq10"),]
```

```
##       1-A1       1-A7      1-F12       1-F6       2-A1       2-A7
## 1135.10204 1015.19757   18.07664    0.00000 1137.83692  984.77673
##      2-F12       2-F6
##   26.36719    0.00000
```

```
perm_logFC %>% filter(feature_id == "Seq10")
```

```
## # A tibble: 16 × 3
##    feature_id  perm       FC
##         <chr> <chr>    <dbl>
## 1       Seq10    X1 62.79384
## 2       Seq10    X2 56.16073
## 3       Seq10    X3 62.94514
## 4       Seq10    X4 54.47785
## 5       Seq10    X5      Inf
## 6       Seq10    X6      Inf
## 7       Seq10    X7      Inf
## 8       Seq10    X8      Inf
## 9       Seq10    X9 43.04980
## 10      Seq10   X10 38.50231
## 11      Seq10   X11 43.15352
## 12      Seq10   X12 37.34857
## 13      Seq10   X13      Inf
## 14      Seq10   X14      Inf
## 15      Seq10   X15      Inf
## 16      Seq10   X16      Inf
```

Mean NaN

```
perm_logFC_summary %>% filter(feature_id == "Seq1135")
```

```
## # A tibble: 1 × 6
##   feature_id logFC_mean logFC_median logFC_lq logFC_uq logFC_var
##        <chr>      <dbl>        <dbl>    <dbl>    <dbl>     <dbl>
## 1    Seq1135        NaN           NA        0        0       NaN
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq1135"),]
```

```
##     1-A1     1-A7    1-F12     1-F6     2-A1     2-A7    2-F12     2-F6
##  0.00000  0.00000 10.12292  0.00000  0.00000  0.00000  0.00000  0.00000
```

```
perm_logFC %>% filter(feature_id == "Seq1135")
```

```
## # A tibble: 16 × 3
##    feature_id  perm    FC
##         <chr> <chr> <dbl>
## 1     Seq1135    X1     0
## 2     Seq1135    X2     0
## 3     Seq1135    X3     0
## 4     Seq1135    X4     0
```

```
## 5     Seq1135    X5   NaN
## 6     Seq1135    X6   NaN
## 7     Seq1135    X7   NaN
## 8     Seq1135    X8   NaN
## 9     Seq1135    X9   NaN
## 10    Seq1135   X10   NaN
## 11    Seq1135   X11   NaN
## 12    Seq1135   X12   NaN
## 13    Seq1135   X13   NaN
## 14    Seq1135   X14   NaN
## 15    Seq1135   X15   NaN
## 16    Seq1135   X16   NaN
```

Mean 0

```
perm_logFC_summary %>% filter(feature_id == "Seq108")
```

```
## # A tibble: 1 × 6
##   feature_id logFC_mean logFC_median logFC_lq logFC_uq logFC_var
##        <chr>      <dbl>        <dbl>    <dbl>    <dbl>     <dbl>
## 1     Seq108          0            0        0        0         0
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq108"),]
```

```
##     1-A1     1-A7    1-F12     1-F6     2-A1     2-A7    2-F12     2-F6
##  0.00000  0.00000 48.80694 33.00790  0.00000  0.00000 50.29297 41.31175
```

```
perm_logFC %>% filter(feature_id == "Seq108")
```

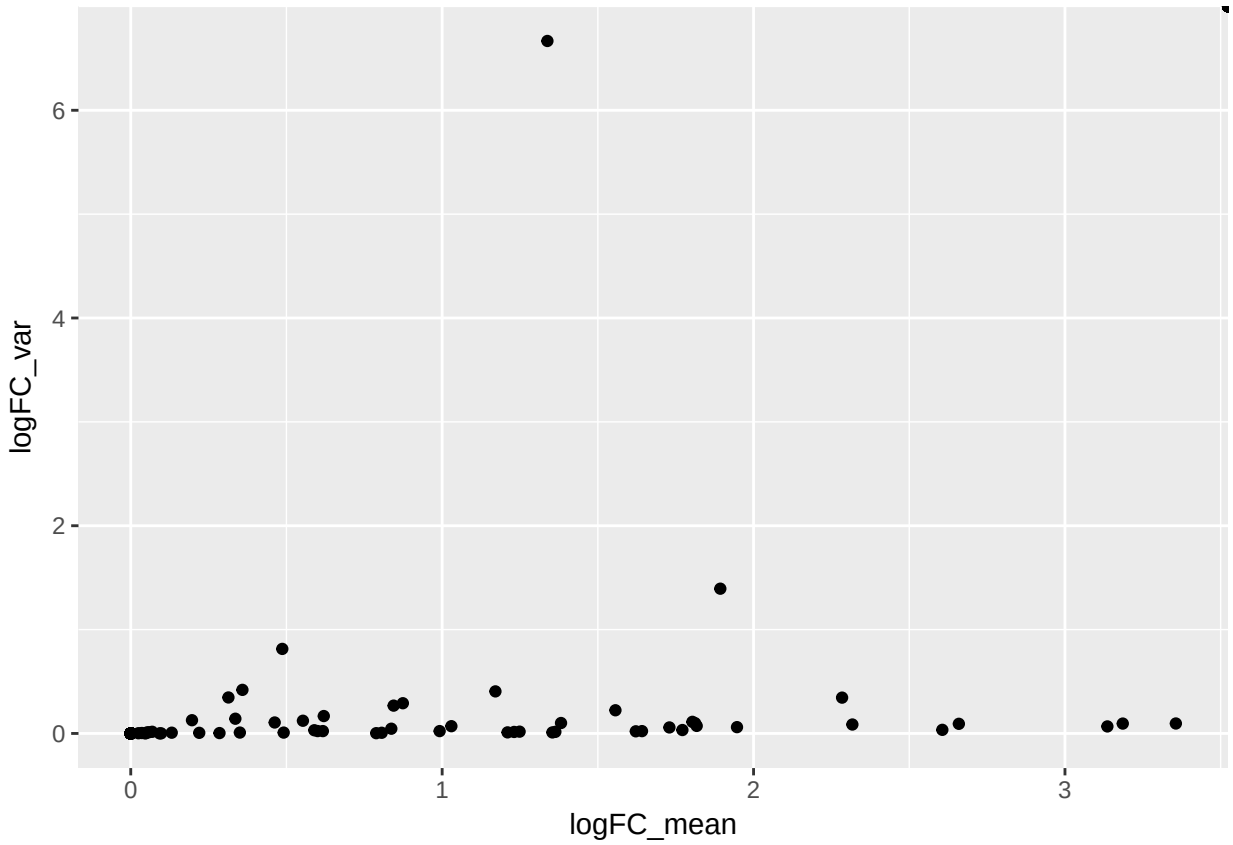
```
## # A tibble: 16 × 3
##    feature_id  perm    FC
##         <chr> <chr> <dbl>
## 1      Seq108    X1     0
## 2      Seq108    X2     0
## 3      Seq108    X3     0
## 4      Seq108    X4     0
## 5      Seq108    X5     0
## 6      Seq108    X6     0
## 7      Seq108    X7     0
## 8      Seq108    X8     0
## 9      Seq108    X9     0
## 10     Seq108   X10     0
## 11     Seq108   X11     0
## 12     Seq108   X12     0
## 13     Seq108   X13     0
## 14     Seq108   X14     0
## 15     Seq108   X15     0
## 16     Seq108   X16     0
```

Filtering out features with mean premutation logFC values of 0, NaN, and Inf.

```
perm_logFC_summary <- perm_logFC_summary %>%
      filter(!is.nan(logFC_mean) | logFC_mean !=0 | !is.infinite(logFC_mean)) %>%
      arrange(-logFC_mean)
```

```
ggplot(perm_logFC_summary) + geom_point(aes(x = logFC_mean, y = logFC_var))
```

```
## Warning: Removed 362 rows containing missing values (geom_point).
```



Features with outlier variance values.

```
perm_logFC_summary %>% filter(logFC_var > 0.8)
```

```
## # A tibble: 69 × 6
##    feature_id logFC_mean logFC_median  logFC_lq logFC_uq logFC_var
##         <chr>      <dbl>        <dbl>     <dbl>    <dbl>     <dbl>
## 1       Seq10        Inf          Inf 5.2771374      Inf       Inf
## 2      Seq113        Inf          Inf       Inf      Inf       Inf
## 3      Seq114        Inf          Inf       Inf      Inf       Inf
## 4      Seq115        Inf          Inf       Inf      Inf       Inf
## 5      Seq117        Inf          Inf       Inf      Inf       Inf
## 6       Seq12        Inf          Inf       Inf      Inf       Inf
## 7       Seq13        Inf          Inf       Inf      Inf       Inf
## 8      Seq144        Inf          Inf 0.5795516      Inf       Inf
## 9      Seq149        Inf          Inf 2.3366579      Inf       Inf
## 10     Seq153        Inf          Inf       Inf      Inf       Inf
## # ... with 59 more rows
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq44"),]
```

```
##      1-A1      1-A7     1-F12      1-F6      2-A1      2-A7     2-F12
## 257.55102 167.17325  53.14534  35.79730 226.54384   0.00000  65.42969
##      2-F6
##  34.07155
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq39"),]
```

```
##       1-A1       1-A7      1-F12       1-F6       2-A1       2-A7
## 87.3469388  0.0000000  4.3383948 14.4119014  0.0000000  0.0000000
##      2-F12       2-F6
##  0.9765625  0.4258944
```

```
pre_post_mat[rownames(pre_post_mat) %in% c("Seq462"),]
```

```
##     1-A1     1-A7    1-F12     1-F6     2-A1     2-A7    2-F12     2-F6
## 5.714286 0.000000 1.084599 1.859600 0.000000 0.000000 3.417969 2.555366
```

NOTE Need to think about next steps * apply permutation to calculate logFC between all dilutions * compare permuted logFC to diff abu methods

Comparison of Permuted logFC value and variance with Diff Abu Method

- Compare variance values to values estimated by differential abundance methods, metagenomeSeq, DESeq2, EdgeR, others?

Log Fold-Change Bias

Treatment-Specific Features

Use the following function to get treatment specific features

```
#metagenomeSeq::uniqueFeatures()
```

- Expected logFC is the difference between titration factors for the titrations being compared
- Compare expected values to those made by differential abundance methods, maybe permutation based method as well

Non-Treatment Specific Features

Log Fold-Change Bias-Variance Relationship

- Use EDA - scatter plot?

Log Fold-Change Feature Exploration

- Correlating factors such as well position, primer matching, and GC content with observed variance and bias.
- Identify outliers for more detailed exploration