Sample Selection

Nate Olson

September 19, 2016

Sample Selection

```
load("../data/etecExp.rda")
## needed for mypar
library(metagenomeSeq)
```

```
## Loading required package: Biobase
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: limma
```

```
##
## Attaching package: 'limma'
```

```
## The following object is masked from 'package:BiocGenerics':
##
##     plotMA
```

```
## Loading required package: glmnet
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach

## Loaded glmnet 2.0-5

## Loading required package: RColorBrewer
```

```
library(Biobase)
library(BiocGenerics)
library(parallel)
library(rafalib)

## experiment count data
pd <- pData(etecExp)

## log 2 qPCR data
qpcr <- log2(as.numeric(as.character(pd$QPCRLT)) + 1)
o <- order(pd$Day,pd$Subj)
days <- unique(pd$Day)
sampIds <- unique(pd$Subj)

tmp <- matrix(NA, nr=length(sampIds), nc=length(days))
rownames(tmp) <- sampIds
colnames(tmp) <- days
for (d in seq(along=days)) {
    ii <- pd$Day == days[d]
    m <- match(pd$Subj[ii], sampIds)
    tmp[m,d] <- qpcr[ii]
}
qpcr <- tmp

maxday <- apply(qpcr,1,which.max)
qpcr2 <- cbind(qpcr[,1], qpcr[cbind(seq(len=nrow(qpcr)),maxday)])

## object otus not found updated name to etecExp
#ecoliFeatures <- which(pData(otus)$species == "Escherichia coli")
ecoliFeatures <- which(fData(etecExp)$species == "Escherichia coli")
tmp <- log2(MRcounts(etecExp[ecoliFeatures,], norm=TRUE) + 1)

# changes day to -1 to match ngs day
ind <- sapply(rownames(qpcr), function(x) which(pd$Subj == x & pd$Day == -1))
drop <- sapply(ind, function(x) length(x)==0)
ind <- unlist(ind[!drop])

tmpPre <- tmp[,ind]

ind <- sapply(names(ind), function(x) which(pd$Subj == x & pd$Day == colnames(qpcr)[maxday[x]]))
tmpPost <- tmp[,ind]

cnts <- tmp
aggTmp <- colSums(cnts,na.rm=TRUE)

tmp <- matrix(NA, nr=length(ind), nc=length(days))
rownames(tmp) <- names(ind)
colnames(tmp) <- days
for (d in seq(along=days)) {
```

```
    ii <- which(pd$Day == days[d])
    m <- match(pd$Subj[ii], names(ind))
    tmp[m[!is.na(m)],d] <- aggTmp[ii[!is.na(m)]]
}
ngs <- tmp

maxday <- apply(ngs,1,which.max)
ngs2 <- cbind(ngs[,2], ngs[cbind(seq(len=nrow(ngs)),maxday)])


sampsToUse <- c(2,3,6,11)
sampIdsToUse <- sampIds[sampsToUse]

qpcr2 <- cbind(qpcr[,1], qpcr[cbind(seq(len=nrow(qpcr)),maxday)])

indPre <- sapply(sampIdsToUse, function(x) which(pd$Subj == x & pd$Day == -1))
cntPre <- cnts[,indPre]

indPost <- sapply(sampIdsToUse, function(x) which(pd$Subj == x & pd$Day == colnames(qpcr)[maxday[as.cha
cntPost <- cnts[,indPost]


## defining plotmat
plotmat <- cnts[,c(indPre,indPost)]
```

```
library(dplyr)
```

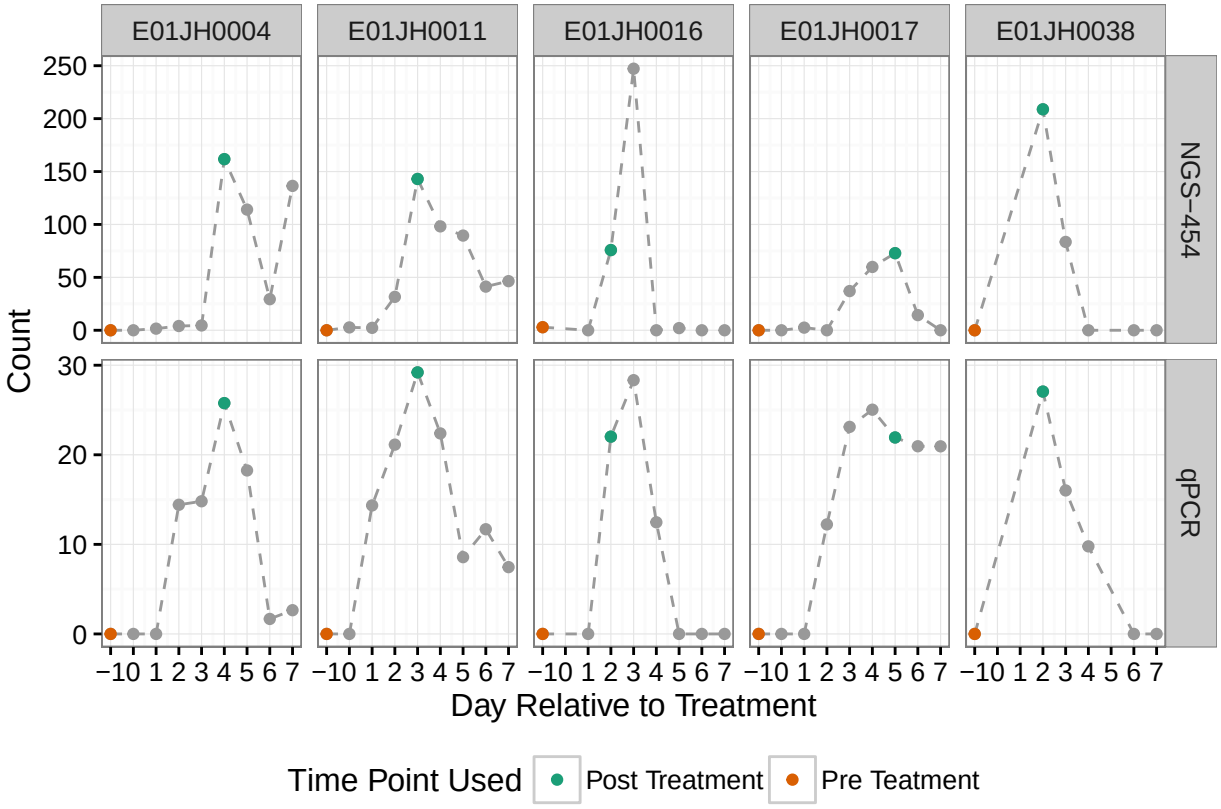
```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:Biobase':
##
##     combine
## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
## The following object is masked from 'package:Matrix':
##
##     expand
```

```
library(ggplot2)
```

```
qpcr_df  <- as.data.frame(qpcr) %>%
    mutate(ID = rownames(.), method = "qPCR")

ngs_df  <- as.data.frame(ngs) %>%
    mutate(ID = rownames(.), method = "NGS-454")

#samples_to_use <- c("E01JH0004", "E01JH0011", "E01JH0016",
#                    "E01JH0017", "E01JH0038")
post_sample <- c(E01JH0004 = 4, E01JH0011 = 3, E01JH0016 = 2,
                 E01JH0017 = 5, E01JH0038 = 2)

pre_post <- bind_rows(qpcr_df, ngs_df) %>%
    gather(key= "time",value = "count",-ID, -method) %>%
    group_by(ID, method) %>%
    filter(ID %in% names(post_sample)) %>%
    mutate(time = as.numeric(time),
           pre_post = ifelse(time == -1,"Pre Teatment",""),
           pre_post = ifelse(time == post_sample[ID],
                             "Post Treatment", pre_post))

## Subset of sample to use for mixtures
ts_df <- pre_post %>%
    filter(!is.na(count), time < 9)

## Samples abundance trends
ggplot(ts_df, aes(x = time, y = count)) + geom_point(color = "grey60") +
    geom_path(linetype = 2, color = "grey60") +
    geom_point(data = filter(ts_df, pre_post != ""),
               aes(x = time, y = count, color = pre_post)) +
    facet_grid(method~ID, scale = "free_y") +
    theme_bw() +
    scale_color_brewer(type = "qual", palette = 2) +
    scale_x_continuous(breaks = c(-1:7)) +
    theme(legend.position = "bottom") +
    labs(x = "Day Relative to Treatment", y = "Count",
         color = "Time Point Used")
```


E01JH0004
E01JH0011
E01JH0016
E01JH0017
E01JH0038
NGS−454
qPCR
Count
250
200
150
100
50
0
30
20
10
0
−10 1 2 3 4 5 6 7
Day Relative to Treatment
Time Point Used
Post Treatment
Pre Teatment