Generate PA Data Frame

Nate Olson

2017-08-25

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

```
## Warning in .load.config(override.config): Your configuration file
## is missing the following entries: data_loading_header, data_ignore,
## logging_level, cache_loaded_data, sticky_variables. Defaults will be used.
```

```
## Warning in .check.version(config): Your configuration is compatible with version 0.7 of the ProjectT
##   Please run ProjectTemplate::migrate.project() to migrate to the installed version 0.8.
```

```
setwd(cwd)
pipeline_dir <- "../../mgtst_pipelines"
mrexp <- get_mrexp(pipeline_dir)
```

Objective

Generate data frame with feature-level presence absence information.

- generate count data frame
- convert counts to 0-1
- summarise by total, Titration, and mix

```
get_count_df <- function(mrobj, agg_genus = FALSE, css = TRUE){
      if(agg_genus){
            mrobj <- aggregateByTaxonomy(mrobj, lvl = "Rank6",
                                         norm = FALSE, log = FALSE, sl = 1)
      }

      if(css == TRUE){
            mrobj <- cumNorm(mrobj, p = 0.75)
            count_mat <- MRcounts(mrobj, norm = TRUE, log = FALSE, sl = 1000)
      }else{
            count_mat <- MRcounts(mrobj, norm = FALSE, log = FALSE, sl = 1)
      }
       count_mat %>%
            as.data.frame() %>%
            rownames_to_column(var = "feature_id") %>%
            gather("id","count", -feature_id)
}

## Converting count matrix to data frame
counts_df <- mrexp %>%
      map_df(get_count_df, css = FALSE, .id = "pipe") %>%
      ## Adding sample metadata
      left_join(pData(mrexp$dada2)) %>%
      filter(biosample_id != "NTC") %>%
```

```
      ## Making t_fctr a charater string with two numeric positions - maintains ordering
      mutate(t_fctr = paste0("T", str_pad(t_fctr, 2, side = "left", pad = "0")))
```

```
## Joining, by = "id"
```

```
## Converting counts to presence/ absence
pa_counts_df <- counts_df %>%
      mutate(pa_count = if_else(count > 0, 1, 0)) %>%
      select(pipe, biosample_id, id, feature_id, t_fctr, pa_count)

## Total, Titration, and Unmixed total observed values
pa_df <- pa_counts_df %>%
      ## Number of PCR reps by titration
      group_by(biosample_id, pipe, t_fctr, feature_id) %>%
      summarise(pa_titration = sum(pa_count)) %>%
      ## Number of of PCR reps total
      group_by(biosample_id, pipe, feature_id) %>%
      mutate(pa_total = sum(pa_titration)) %>%
      ## Unmixed and Mixed PA counts
      spread(t_fctr, pa_titration) %>%
      mutate(pa_unmixed = T00 + T20,  ## Number of PCR reps unmixed pre/post
             pa_mixed = pa_total - pa_unmixed) ## Number of PCR reps mixed

## Summary PA Values
pa_summary <- pa_df %>%
      filter(pa_total != 0) %>%
      select(biosample_id, pipe, feature_id, T00, T20,
             pa_unmixed, pa_mixed, pa_total)


## PA Annotation
pa_summary_anno <- pa_summary %>%
      mutate(full_pre = if_else(T00 == 4, 1, 0),
             full_post = if_else(T20 == 4, 1, 0),
             full_unmixed = if_else(T00 + T20 == 8, 1, 0),
             pre_specific = if_else(T00 == 4 & T20 == 0, 1, 0),
             post_specific = if_else(T00 == 0 & T20 == 4, 1, 0),
             pa_specific = case_when(
                    pa_unmixed != 0 & pa_mixed != 0 ~ "unspecific",
                    pa_unmixed != 0 & pa_mixed == 0 ~ "unmix",
                    pa_unmixed == 0 & pa_mixed != 0 ~ "mixed"))

saveRDS(pa_summary_anno, "~/Desktop/pa_summary_anno_df.RDS")
```

Pre-Post Specific Sanity Check - should not be any features with values > 1

```
filter(pa_summary_anno, pre_specific + post_specific + full_unmixed > 1)
```

```
## # A tibble: 0 x 14
## # Groups:   biosample_id, pipe, feature_id [0]
## # ... with 14 variables: biosample_id <chr>, pipe <chr>, feature_id <chr>,
## #   T00 <dbl>, T20 <dbl>, pa_unmixed <dbl>, pa_mixed <dbl>,
## #   pa_total <dbl>, full_pre <dbl>, full_post <dbl>, full_unmixed <dbl>,
## #   pre_specific <dbl>, post_specific <dbl>, pa_specific <chr>
```

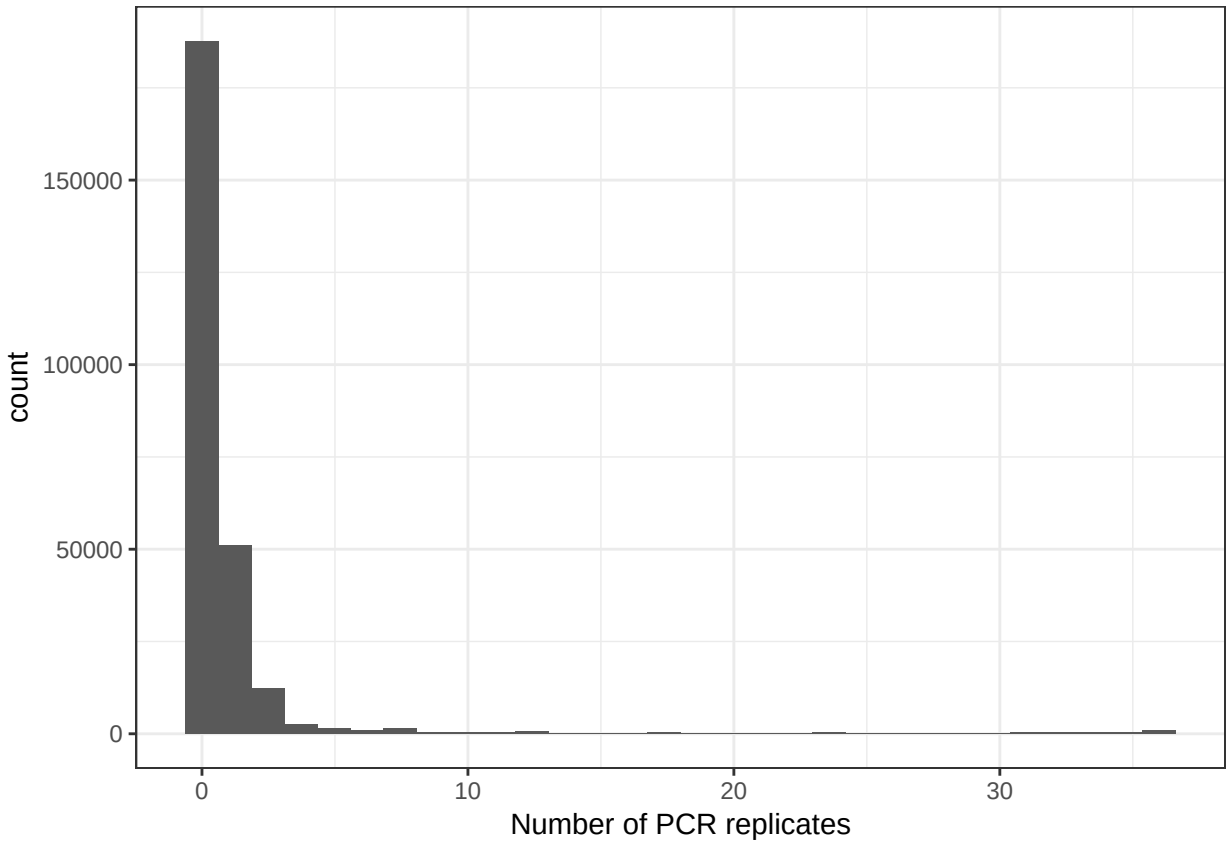


Figure 1: Distribution in the number of PCR replicates a feature is observed in for a pipeline and biological replicate. The maximum number of PCR replicates is 36, 4 PCR replicates for the seven titrations and the two endpoints (unmixed samples).

Summary Plots

Overall PCR Replicate Count Distributions

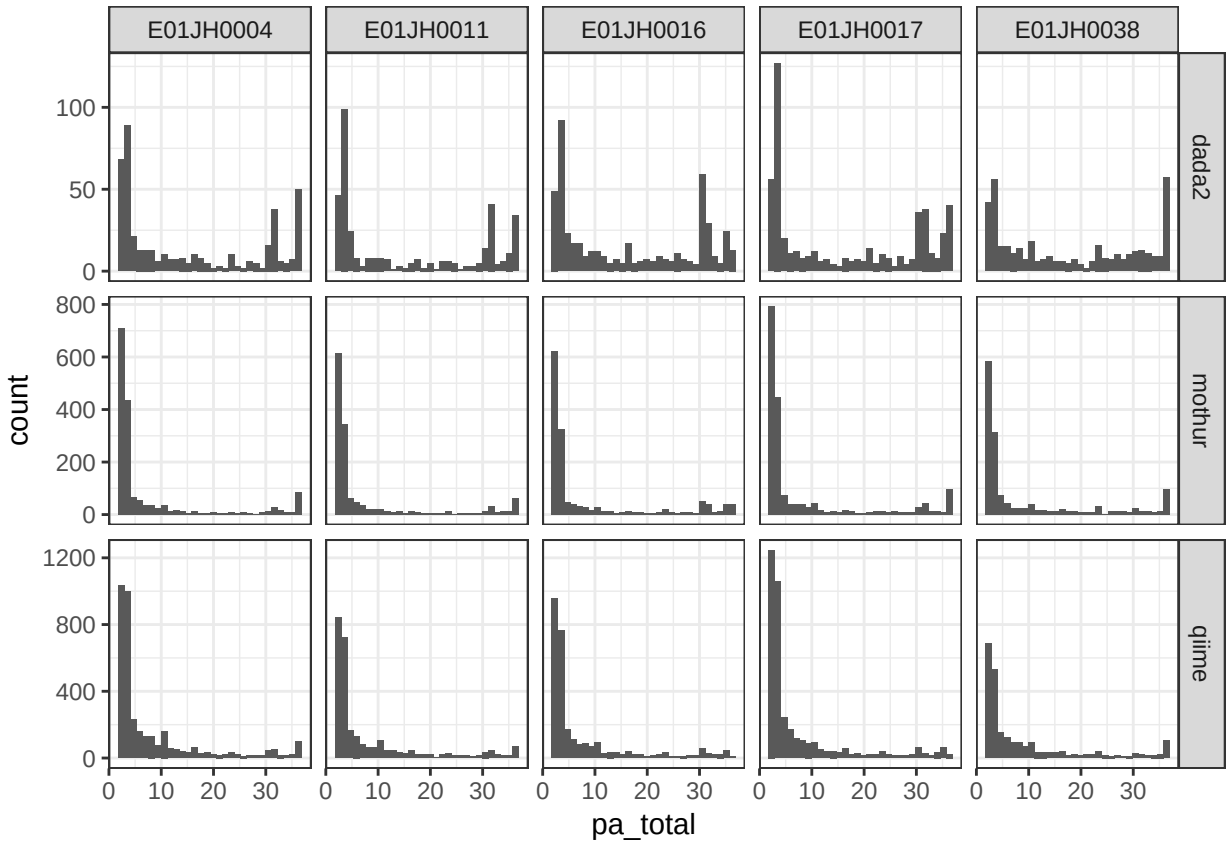
Most of the features are only observed in 1 of the 36 PCR replicates. Of the features observed in more than 1 PCR replicates, DADA2 has the highest proportion of features observed in more than 20 PCR replicates.

```
pa_df %>% ggplot() +
      geom_histogram(aes(x = pa_total)) +
      theme_bw() +
      labs(x = "Number of PCR replicates")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

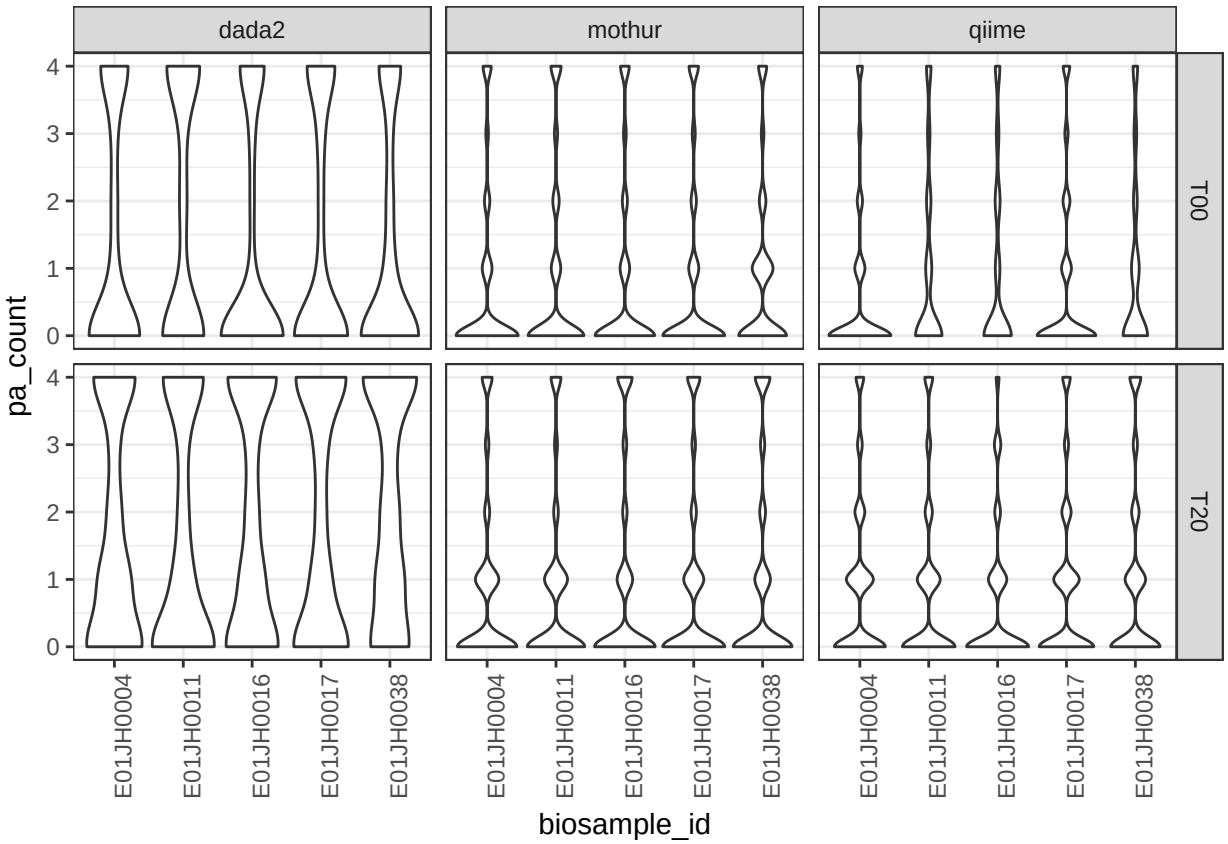
```
pa_df %>% filter(pa_total > 1) %>% ggplot() +
      geom_histogram(aes(x = pa_total)) +
      facet_grid(pipe~biosample_id, scales = "free_y") +
      theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



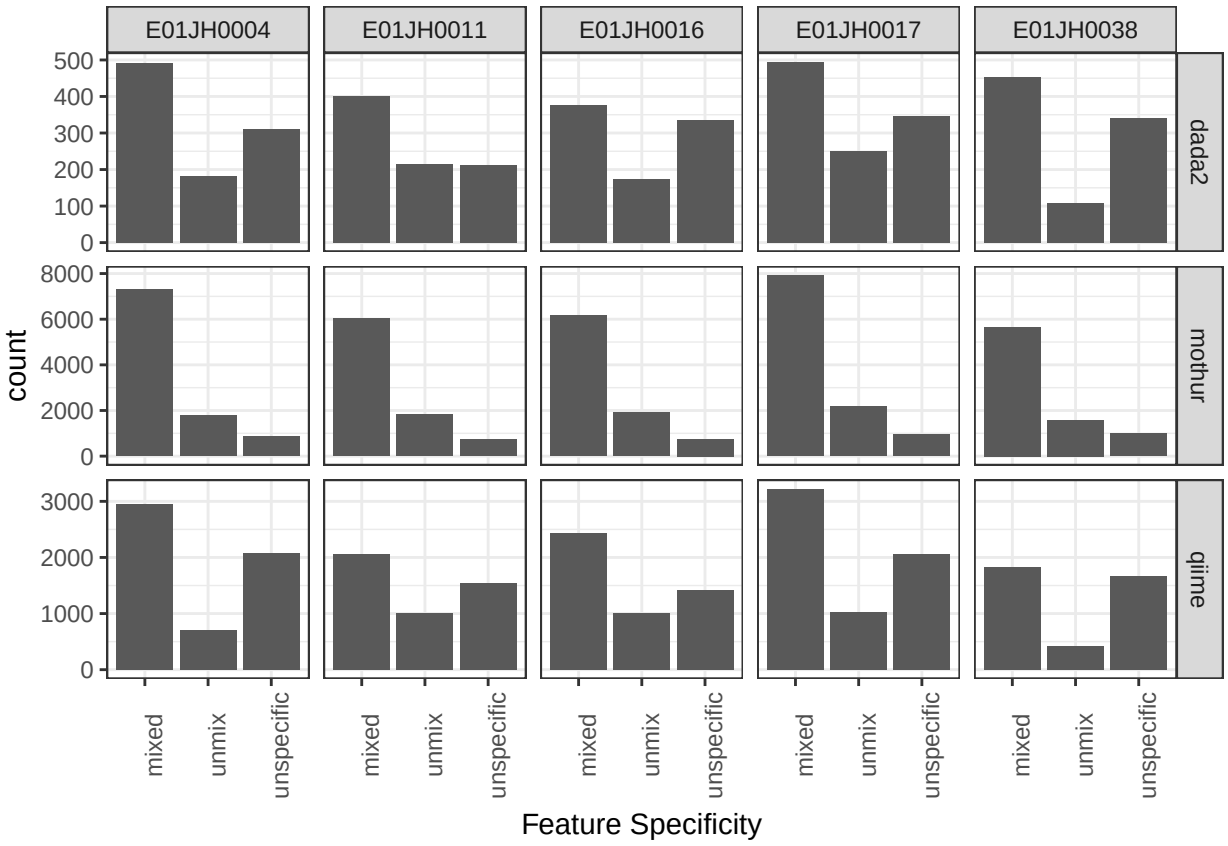
Pre and Post Specific Features

```
pa_summary_long <- pa_summary %>%
      filter(pa_total > 1)  %>%
      select(-pa_total, -pa_unmixed, -pa_mixed) %>%
      gather("count_class", "pa_count", -biosample_id, -pipe, -feature_id)
pa_summary_long %>% ggplot() +
      geom_violin(aes(y = pa_count, x = biosample_id)) +
      facet_grid(count_class ~ pipe, scales = "free_y") +
      theme_bw() +
      theme(axis.text.x = element_text(angle = 90))
```



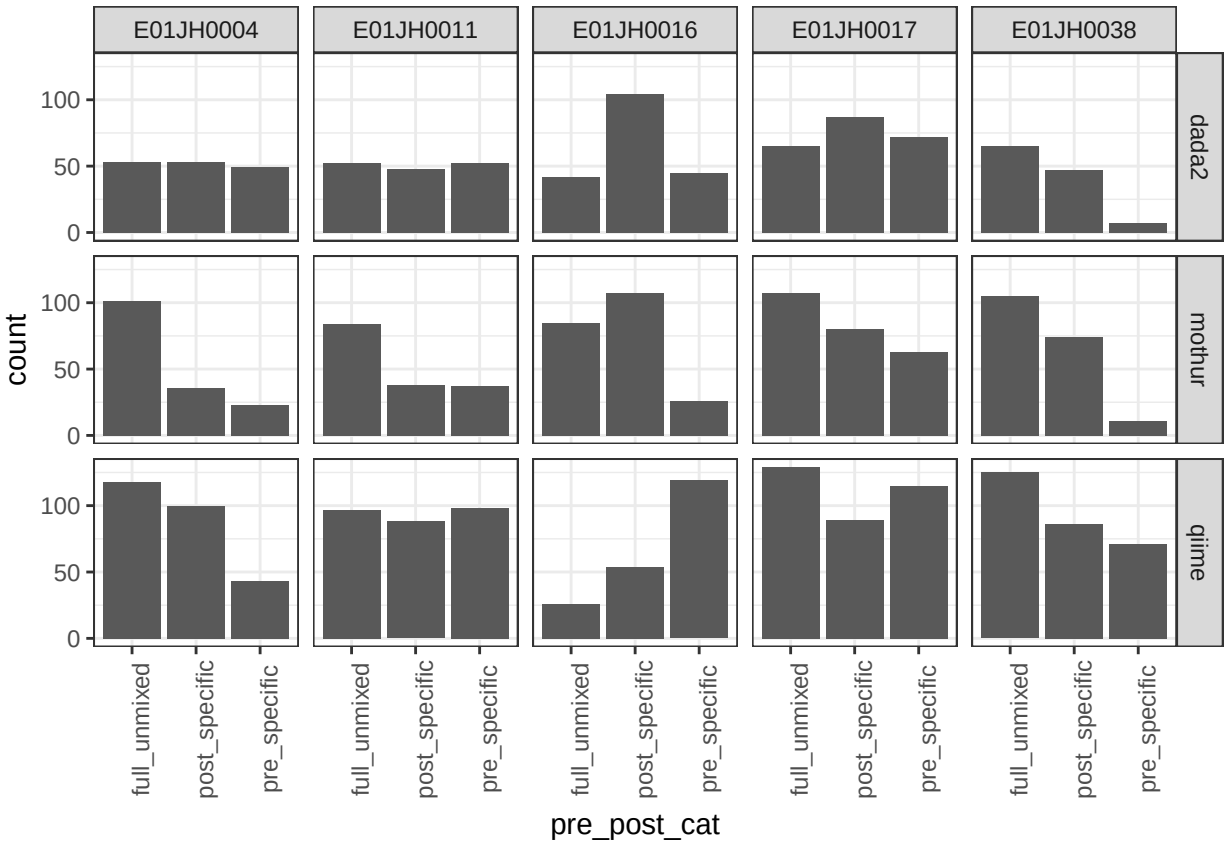
Breakdown of Feature Types

```
pa_summary_anno %>% ggplot() +
      geom_bar(aes(x = pa_specific)) +
      facet_grid(pipe ~ biosample_id, scales = "free_y") +
      theme_bw() +
      labs(x = "Feature Specificity") +
      theme(axis.text.x = element_text(angle = 90))
```



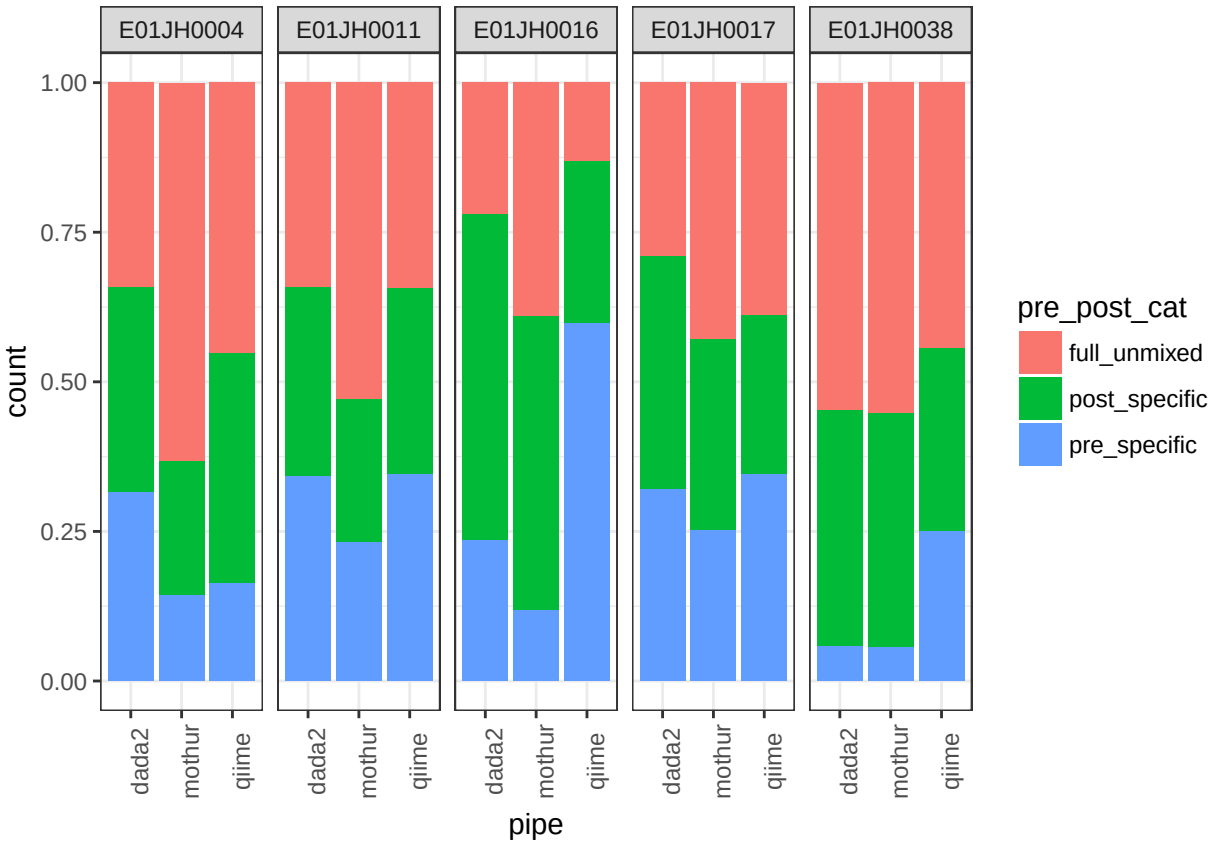
Useable features for analysis based on presence absence in unmixed samples. The

```
pre_post_cat_df <- pa_summary_anno %>%
      # excluding features only observed in endpoints
      filter(pa_specific != "unmixed") %>%
      select(biosample_id, pipe, feature_id,
             pre_specific, post_specific, full_unmixed) %>%
      gather("pre_post_cat", "value", -biosample_id, -pipe, -feature_id) %>%
      filter(value == 1)
ggplot(pre_post_cat_df) + geom_bar(aes(x = pre_post_cat)) +
      facet_grid(pipe~biosample_id) +
      theme_bw() + theme(axis.text.x = element_text(angle = 90))
```



Of the useable features for the quantitative analysis section the proportion of features that are present in all four PCR replicates for the unmixed pre, post, or both samples varies by pipeline. One would expect these to be consistent across pipelines.

```
ggplot(pre_post_cat_df) + geom_bar(aes(x =pipe, fill = pre_post_cat),
                                   position = "fill") +
      facet_grid(.~biosample_id) +
      theme_bw() + theme(axis.text.x = element_text(angle = 90))
```



Session information

Git repo commit information

```
library(git2r)
```

```
##
## Attaching package: 'git2r'
## The following object is masked from 'package:foreach':
##
##     when
## The following objects are masked from 'package:Biobase':
##
##     content, notes
## The following object is masked from 'package:dplyr':
##
##     pull
## The following objects are masked from 'package:purrr':
##
##     is_empty, when
```

```
repo <- repository(path = "../")
last_commit <- commits(repo)[[1]]
```

The current git commit of this file is e2ad37393f6512d6d87ae85da4ee0963032decd4, which is on the master branch and was made by nate-d-olson on 2017-08-24 16:40:03. The current commit message is updated notes and outline. The repository is online at https://github.com/nate-d-olson/mgtst-pub

Platform Information

```
s_info <- devtools::session_info()
print(s_info$platform)
```

```
##  setting  value
##  version  R version 3.4.1 (2017-06-30)
##  system   x86_64, darwin16.6.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-08-25
```

Package Versions

```
s_info$packages %>% filter(`*` == "*") %>% select(-`*`) %>%
      knitr::kable()
```

package	version	date	source
base	3.4.1	2017-07-07	local
bindrcpp	0.2	2017-06-17	CRAN (R 3.4.0)
Biobase	2.36.2	2017-06-21	Bioconductor
BiocGenerics	0.22.0	2017-05-04	Bioconductor
datasets	3.4.1	2017-07-07	local
dplyr	0.7.2	2017-07-20	CRAN (R 3.4.1)
forcats	0.2.0	2017-01-23	CRAN (R 3.4.0)
foreach	1.4.3	2015-10-13	CRAN (R 3.4.0)
ggplot2	2.2.1	2016-12-30	CRAN (R 3.4.0)
git2r	0.19.0	2017-07-19	CRAN (R 3.4.1)
glmnet	2.0-10	2017-05-06	CRAN (R 3.4.0)
graphics	3.4.1	2017-07-07	local
grDevices	3.4.1	2017-07-07	local
knitr	1.17	2017-08-10	CRAN (R 3.4.1)
limma	3.32.3	2017-07-19	Bioconductor
Matrix	1.2-10	2017-05-03	CRAN (R 3.4.1)
metagenomeSeq	1.18.0	2017-05-04	Bioconductor
methods	3.4.1	2017-07-07	local
modelr	0.1.1	2017-07-24	CRAN (R 3.4.1)
parallel	3.4.1	2017-07-07	local
ProjectTemplate	0.8	2017-08-09	CRAN (R 3.4.1)
purrr	0.2.3	2017-08-02	CRAN (R 3.4.1)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.4.0)
readr	1.1.1	2017-05-16	CRAN (R 3.4.0)
readxl	1.0.0	2017-04-18	CRAN (R 3.4.0)
stats	3.4.1	2017-07-07	local
stringr	1.2.0	2017-02-18	CRAN (R 3.4.0)

package	version	date	source
tibble	1.3.3	2017-05-28	CRAN (R 3.4.0)
tidyr	0.6.3	2017-05-15	CRAN (R 3.4.0)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.4.0)
utils	3.4.1	2017-07-07	local