No Template Control Characterization

Nate Olson

2017-04-11

Objective

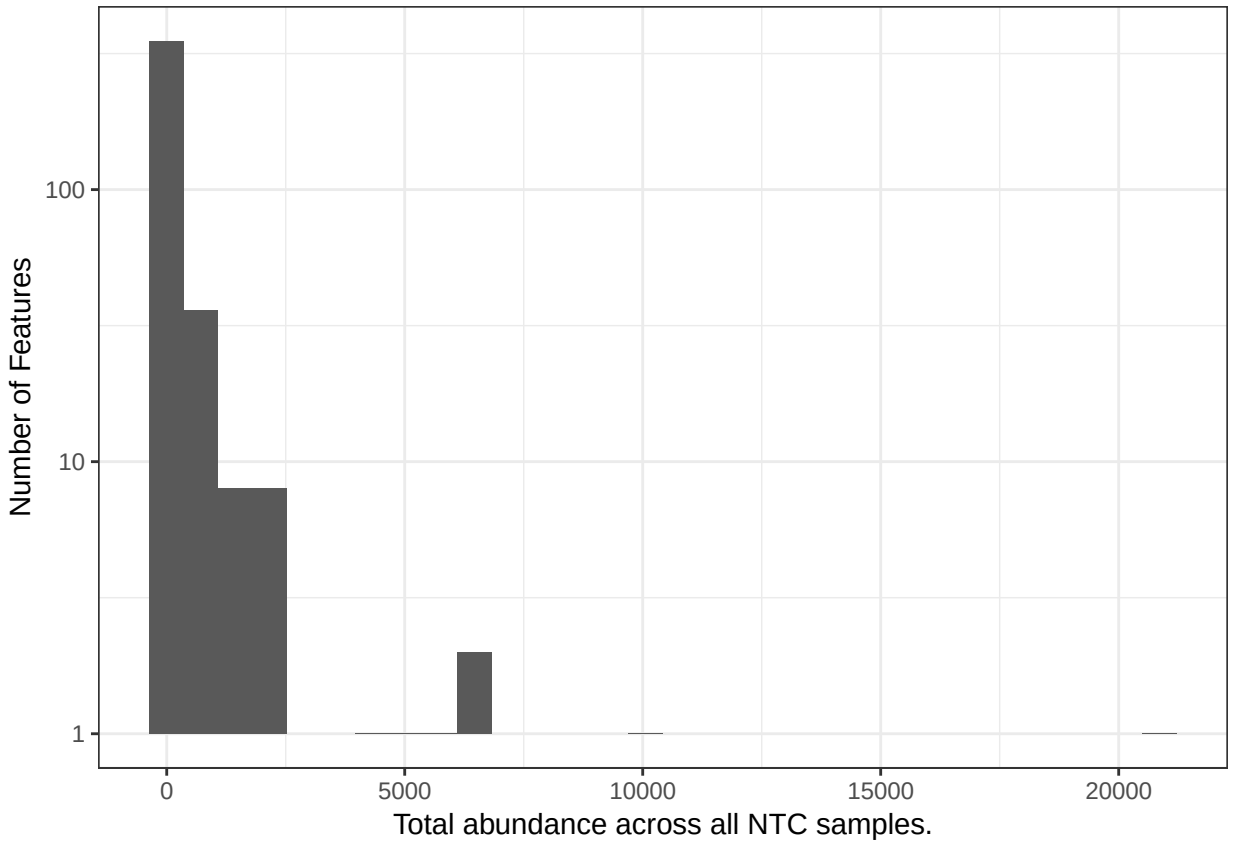
Identify and characterize features present in No Template Controls.

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

      mrobj <- cumNorm(mrobj, p = 0.75)
      mrobj %>%
            # not sure whether or not to normalize counts prior to analysis
            MRcounts(norm = TRUE, log = FALSE, sl = 1000) %>%
            as.data.frame() %>%
            rownames_to_column(var = "feature_id") %>%
            gather("id","count", -feature_id)
}

pipeline_dir <- "../../mgtst_pipelines"
mrexp <- get_mrexp(pipeline_dir)

count_df <- mrexp %>% map_df(get_count_df, .id = "pipe") %>%
      left_join(pData(mrexp$dada2))
```

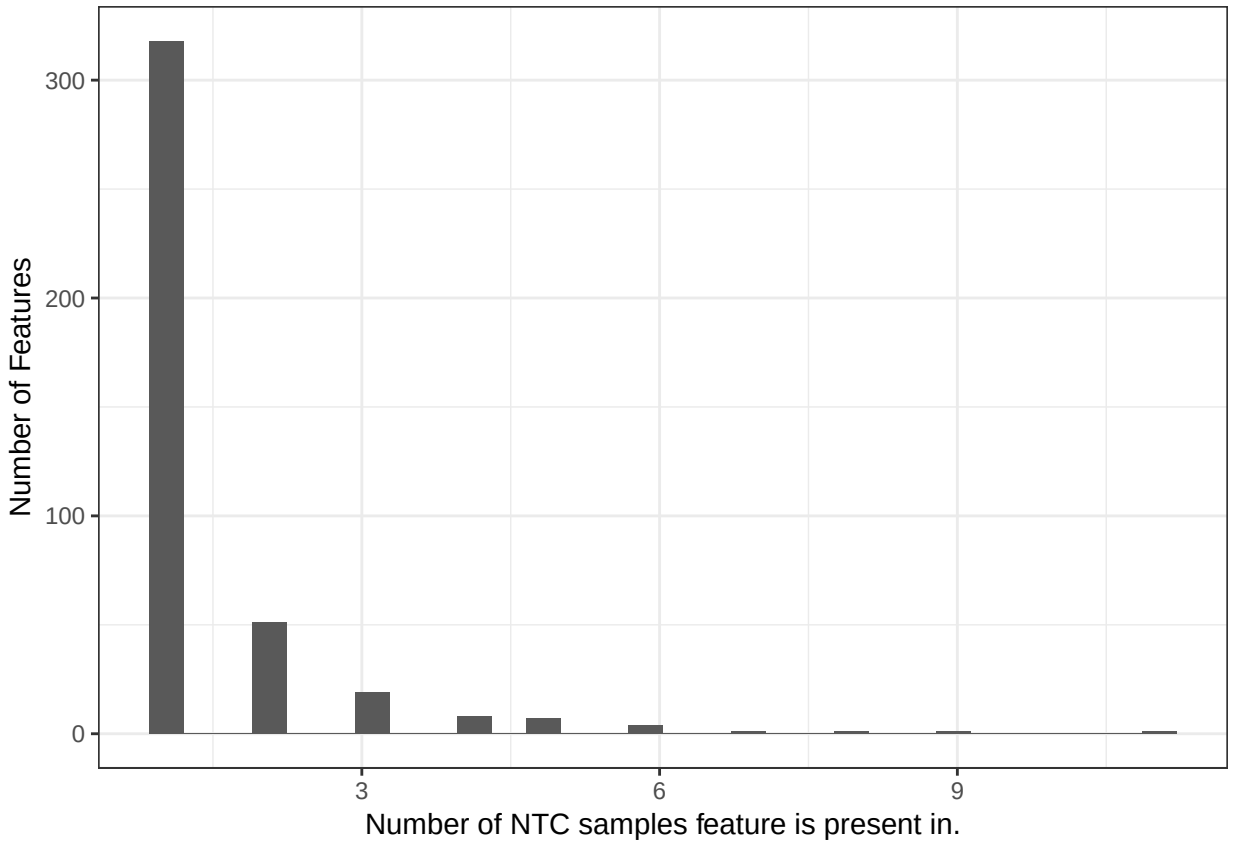
```
ntc_summary <- count_df %>% filter(biosample_id == "NTC") %>%
      filter(count != 0) %>%
      group_by(pipe, feature_id) %>%
      summarise(total_count = sum(count),
                med_count = median(count),
                n_present = n())
```

Note count values are CSS normalized and scaled.

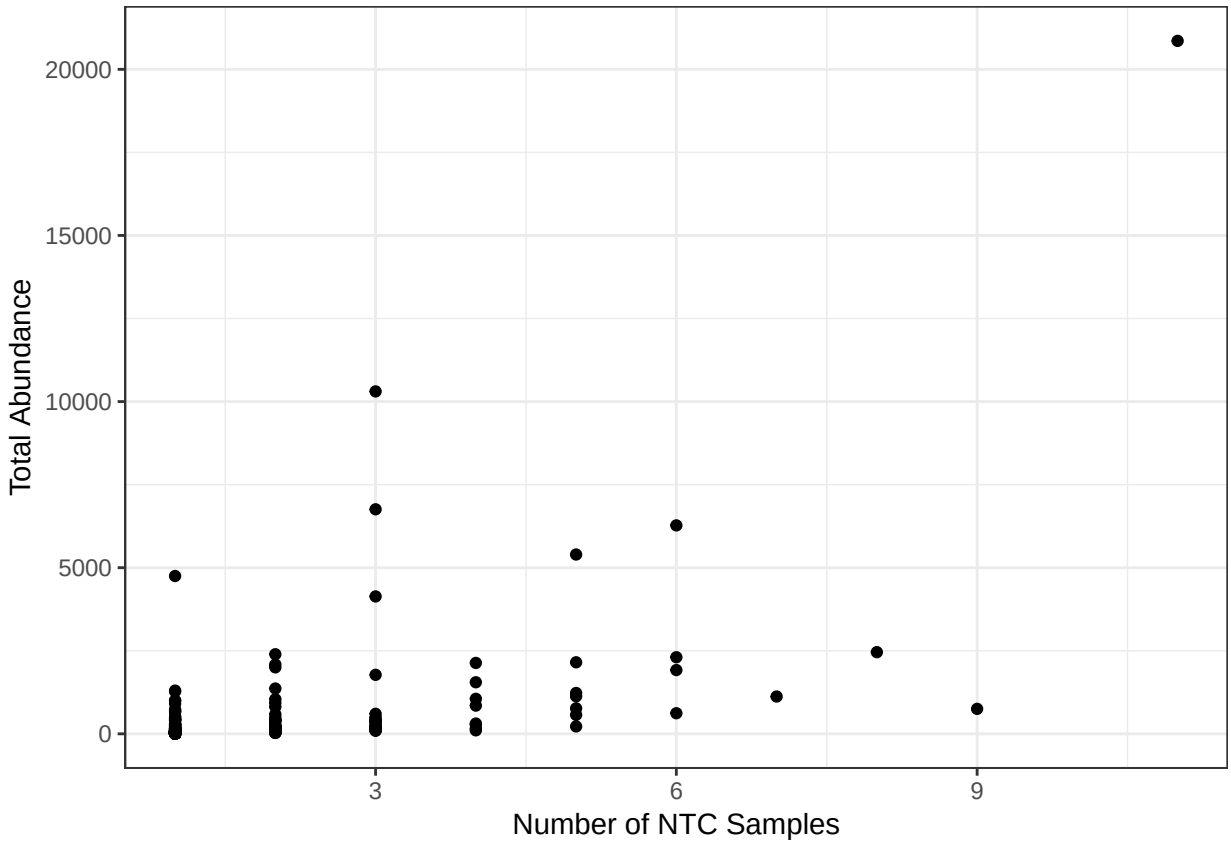
```
ntc_summary %>% ggplot() +
      geom_histogram(aes(x = total_count)) + scale_y_log10() + theme_bw() +
      labs(x = "Total abundance across all NTC samples.", y = "Number of Features")
```



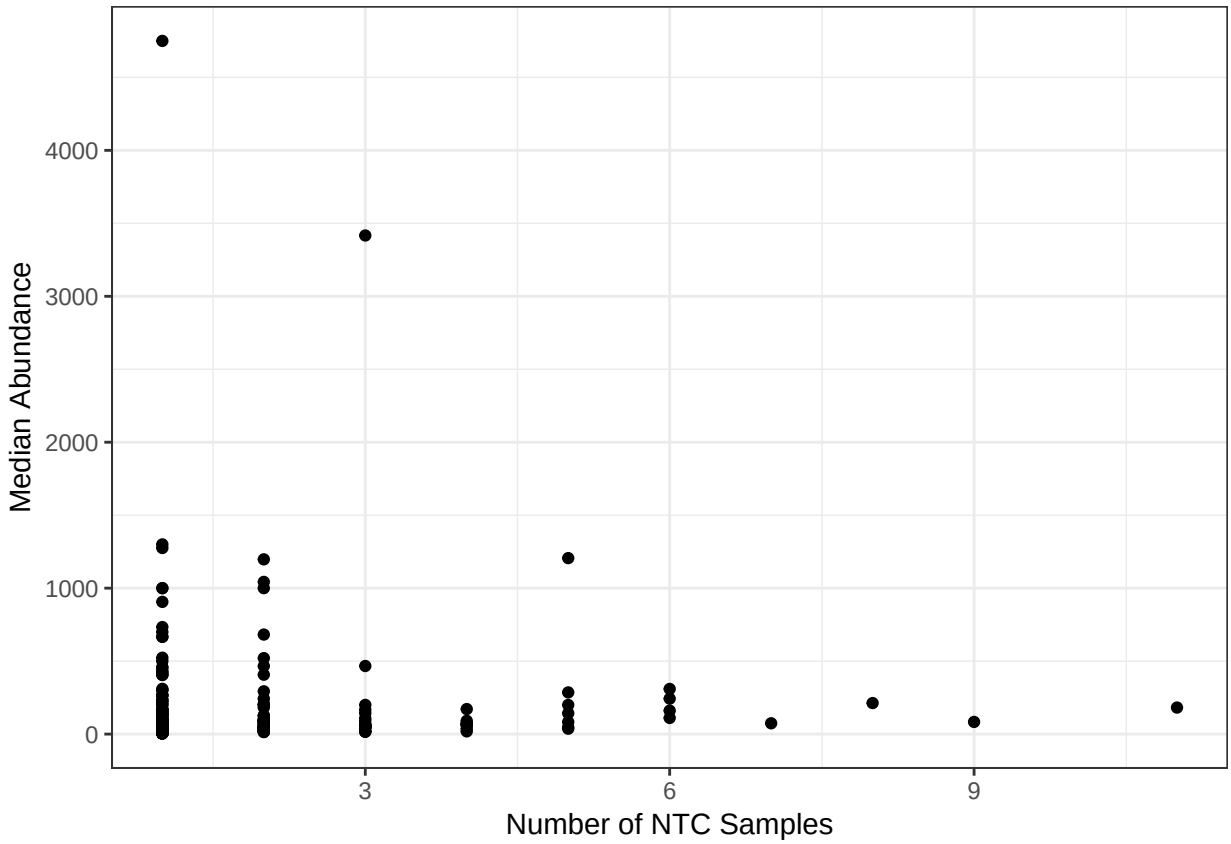

```
ntc_summary %>% ggplot() +
      geom_histogram(aes(x = n_present)) + theme_bw() +
      labs(x = "Number of NTC samples feature is present in.",
           y = "Number of Features")
```




```
ntc_summary %>% ggplot() +
      geom_point(aes(x = n_present, y = total_count)) + theme_bw() +
      labs(x = "Number of NTC Samples", y = "Total Abundance")
```




```
ntc_summary %>% ggplot() +
      geom_point(aes(x = n_present, y = med_count)) + theme_bw() +
      labs(x = "Number of NTC Samples", y = "Median Abundance")
```



High Abundance Features * Features with a high total abundance or count and low median count are features with a high abundance count in few samples and low abundance in other samples.

```
ntc_summary %>% filter(total_count > 4000) %>%
      arrange(desc(total_count)) %>% knitr::kable()
```

pipe	feature_id	total_count	med_count	n_present
mothur	Otu00002	20855.382	181.81818	11
dada2	SV727	10305.556	3416.66667	3
mothur	Otu00010	6757.330	45.45455	3
qiime	513445	6273.634	242.85714	6
dada2	SV1	5397.178	1205.88235	5
dada2	SV90	4750.000	4750.00000	1
qiime	348304	4134.931	466.66667	3

High Frequency Features

```
ntc_summary %>% filter(n_present > 6) %>%
      arrange(desc(n_present)) %>% knitr::kable()
```

pipe	feature_id	total_count	med_count	n_present
mothur	Otu00002	20855.3824	181.81818	11
mothur	Otu00003	751.5403	83.33333	9
mothur	Otu00001	2457.8397	212.53071	8
mothur	Otu00004	1122.8116	74.07407	7

Taxonomy of high abundance and fequency features

- Taxonomy of high abundance no template control features is not consistent between pipelines.
- May not want to drop all no template controls features as *Escherichia/Shigella* features that are relevant to the study were present in NTCs.

```
feature_taxa <- mrexp %>% map_df(fData, .id = "pipe") %>%
      dplyr::rename(feature_id = OTUname)
ntc_summary %>% filter(n_present > 6 | total_count > 4000) %>%
      left_join(feature_taxa) %>% select(pipe, feature_id, n_present, total_count, Rank5, Rank6) %>%
      arrange(Rank5) %>% knitr::kable()
```

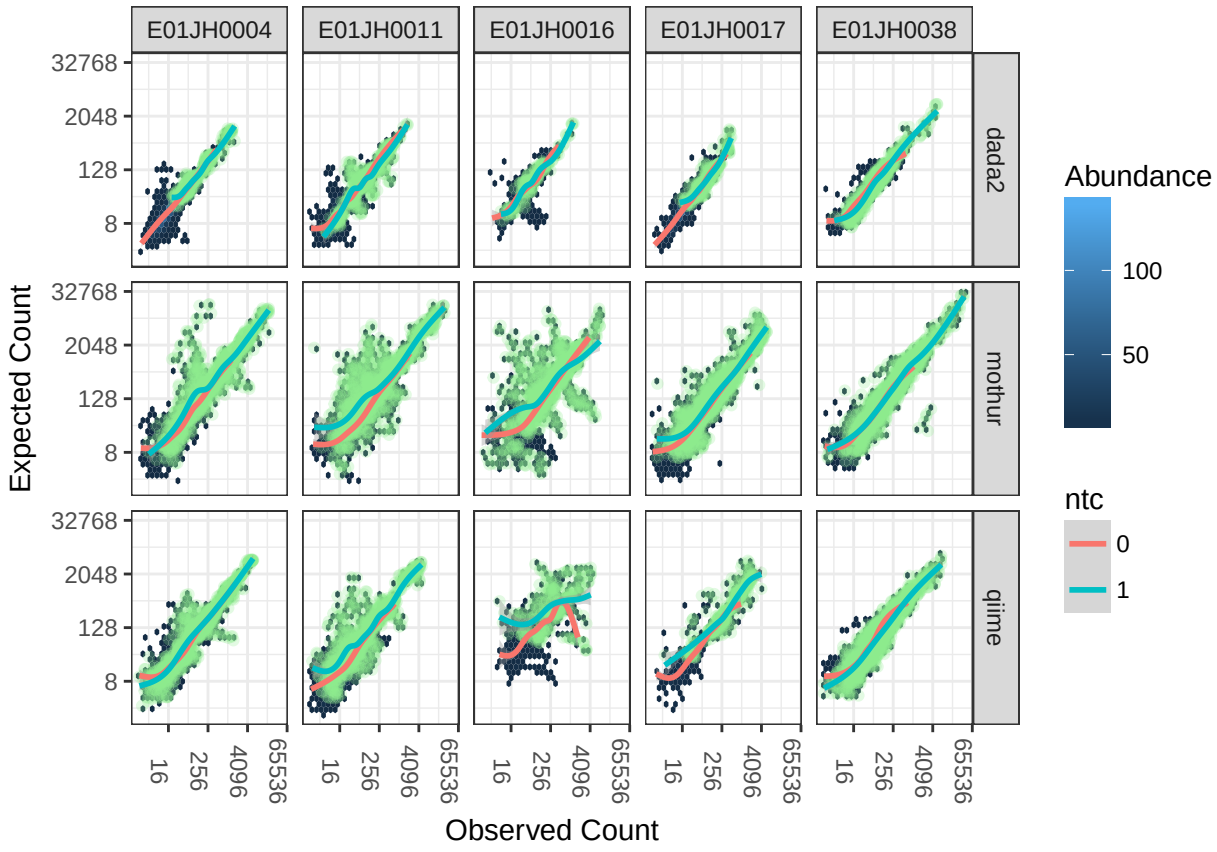
pipe	feature_id	n_present	total_count	Rank5	Rank6
mothur	Otu00002	11	20855.3824	Bacteroidaceae	Bacteroides
mothur	Otu00004	7	1122.8116	Bacteroidaceae	Bacteroides
mothur	Otu00010	3	6757.3301	Bacteroidaceae	Bacteroides
dada2	SV1	5	5397.1779	Enterobacteriaceae	Escherichia/Shigella
dada2	SV727	3	10305.5556	Enterobacteriaceae	Serratia
mothur	Otu00001	8	2457.8397	Enterobacteriaceae	Escherichia/Shigella
qiime	348304	3	4134.9313	f__Bacteroidaceae	g__Bacteroides
qiime	513445	6	6273.6337	f__Bacteroidaceae	g__Bacteroides
dada2	SV90	1	4750.0000	Prevotellaceae	Prevotella_9
mothur	Otu00003	9	751.5403	Prevotellaceae	Prevotella

NTC Feature Bias Analysis

```
count_exp_df <- readRDS("../data/expected_count_values_feature_df.rds")
feature_cat <- readRDS("../data/feature_categories_df.rds")
count_exp_df <- left_join(count_exp_df, feature_cat) %>%
      left_join(ntc_summary) %>%
      filter(cat %in% c("cat_full", "cat_pre", "cat_post")) %>%
      mutate(ntc = if_else(is.na(n_present), "0","1"),
             bad_sample = if_else(id %in% c("1-F9","1-F4"),id,"0"))
```

Features observed in at least one no template control sample overlaid on plot of all features. Overall the no template control features perform similarly in terms of bias relative to features not present in any no template control sample.

```
count_exp_df %>%
    ggplot() +
    geom_hex(aes(x = obs_count + 1, y = exp_count + 1)) +

      geom_point(data =  filter(count_exp_df, ntc == "1"),
                 aes(x = obs_count + 1, y = exp_count + 1), color = "lightgreen", alpha = 0.25) +
        geom_smooth(aes(x = obs_count + 1, y = exp_count + 1, color = ntc)) +
    facet_grid(pipe~biosample_id)+ theme_bw() +
    labs(y = "Expected Count", x = "Observed Count", fill = "Abundance") +
    scale_y_continuous(trans = "log2") +
    scale_x_continuous(trans = "log2") +
    theme(axis.text.x = element_text(angle = 270))
```



Saving data

```
ntc_summary %>% saveRDS("../data/ntc_features.rds")
```

Session information

Git repo commit information

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

The current git commit of this file is 895febf566f36e330aec47d686d8dff2a1b35138, which is on the master branch and was made by nate-d-olson on 2017-04-11 16:16:44. The current commit message is bias analysis for outlier samples. 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 3.3.3 (2017-03-06)
##  system   x86_64, darwin15.6.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-04-11
```

Package Versions

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

package	version	date	source
bbmle	1.0.18	2016-02-11	CRAN (R 3.3.2)
Biobase	2.34.0	2016-11-07	Bioconductor
BiocGenerics	0.20.0	2016-11-07	Bioconductor
BiocParallel	1.8.1	2016-11-07	Bioconductor
Biostrings	2.42.1	2016-12-19	Bioconductor
DESeq	1.26.0	2016-11-28	Bioconductor
DESeq2	1.15.28	2017-02-02	bioc (readonly/DESeq2@125913)
dplyr	0.5.0	2016-06-24	CRAN (R 3.3.2)
edgeR	3.16.5	2017-02-02	Bioconductor
forcats	0.2.0	2017-01-23	CRAN (R 3.3.2)
foreach	1.4.3	2015-10-13	CRAN (R 3.3.1)
GenomeInfoDb	1.10.3	2017-03-28	Bioconductor
GenomicAlignments	1.10.1	2017-03-28	Bioconductor
GenomicRanges	1.26.4	2017-03-28	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.3.2)
git2r	0.18.0	2017-01-01	CRAN (R 3.3.2)
glmnet	2.0-5	2016-03-17	CRAN (R 3.3.1)
hexbin	1.27.1	2015-08-19	CRAN (R 3.3.1)
IRanges	2.8.2	2017-03-28	Bioconductor
knitr	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.3)
limma	3.30.13	2017-03-28	Bioconductor
locfit	1.5-9.1	2013-04-20	CRAN (R 3.3.1)
Matrix	1.2-8	2017-01-20	CRAN (R 3.3.3)
metagenomeSeq	1.16.0	2016-11-07	Bioconductor
modelr	0.1.0	2016-08-31	cran (@0.1.0)
permute	0.9-4	2016-09-09	CRAN (R 3.3.1)
phyloseq	1.19.1	2017-01-04	Bioconductor
ProjectTemplate	0.7	2016-08-11	CRAN (R 3.3.1)
purrr	0.2.2	2016-06-18	CRAN (R 3.3.1)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.1)
readr	1.1.0	2017-03-22	CRAN (R 3.3.2)
readxl	0.1.1	2016-03-28	cran (@0.1.1)
Rqc	1.8.0	2016-11-07	Bioconductor
Rsamtools	1.26.1	2016-11-07	Bioconductor
S4Vectors	0.12.2	2017-03-28	Bioconductor
sads	0.3.1	2016-05-13	CRAN (R 3.3.2)
savR	1.12.0	2016-11-07	Bioconductor

package	version	date	source
ShortRead	1.32.1	2017-03-28	Bioconductor
stringr	1.2.0	2017-02-18	CRAN (R 3.3.2)
SummarizedExperiment	1.4.0	2016-11-07	Bioconductor
tibble	1.3.0	2017-04-01	CRAN (R 3.3.3)
tidyr	0.6.1	2017-01-10	CRAN (R 3.3.2)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.3.2)
vegan	2.4-3	2017-04-07	CRAN (R 3.3.3)
XVector	0.14.1	2017-03-28	Bioconductor