Dataset with Well Behaved Features

Nate Olson

2017-04-06

Objective

Generate a dataset with phenotype data and count data for only informative features, thoes assigned to full, pre, and post categories.

Extracting count data from MRexperiments

```
## Extracting a tidy dataframe with count values from MRexpiment objects
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 = FALSE, log = FALSE, sl = 1) %>%
            as.data.frame() %>%
            rownames_to_column(var = "feature_id") %>%
            gather("id","count", -feature_id)
}

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

Feature categories from `2017-03-29-Feature-Categorization-Take2.Rmd`.

```
feature_cat <- readRDS("../data/feature_categories_df.rds")
```

```
annotated_counts <- left_join(count_df, feature_cat) %>%
      filter(cat %in% c("cat_full","cat_post","cat_pre")) %>%
      select(-titration) %>%
      dplyr::rename(pcr_id = id)
annotated_counts %>% saveRDS("../data/raw_counts_good_feature_categories.rds")
```

Column description

pipe - bioinformatic pipeline used
feature_id - feature id assigned by the bioinformatic pipeline
pcr_id - unique id for a sample PCR replicate, used as column names in `annotated_count_matrix`
count - raw counts
biosample_id - unique id for individual biological replicates
t_fctr - titration factor
pcr_16S_plate - id for the replicate PCR plate
pos - well position in the PCR plate
pcr_half - PCR plate blocking indicator, first or second half of the PCR plate

pcr_rep - blocking indicator for the four sets of PCR replicates
cat - feature category assignment

```
glimpse(annotated_counts)
```

```
## Observations: 102,996
## Variables: 11
## $ pipe          <chr> "dada2", "dada2", "dada2", "dada2", "dada2", "da...
## $ feature_id    <chr> "SV1", "SV2", "SV4", "SV7", "SV8", "SV9", "SV10"...
## $ pcr_id        <chr> "1-A1", "1-A1", "1-A1", "1-A1", "1-A1", "1-A1", ...
## $ count         <dbl> 1072, 10205, 2865, 2259, 2324, 36, 885, 2864, 49...
## $ biosample_id  <chr> "E01JH0004", "E01JH0004", "E01JH0004", "E01JH000...
## $ t_fctr        <fctr> 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20,...
## $ pcr_16S_plate <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ pos           <chr> "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", ...
## $ pcr_half      <chr> "1", "1", "1", "1", "1", "1", "1", "1", "1", "1"...
## $ pcr_rep       <chr> "1:1", "1:1", "1:1", "1:1", "1:1", "1:1", "1:1",...
## $ cat           <chr> "cat_full", "cat_full", "cat_full", "cat_full", ...
```

Feature Category assignment definitions

- Full - features present in at least one PCR replicates for all samples and absent in less than 12 of the 36 total PCR replicates of a biological replicate, and pipeline.
- Pre - present in three or more PCR replicates for unmixed pre-treatment samples, not observed in any PCR replicates of the unmixed post treatment samples, and present in at least 24 total PCR replicates.
- Post - present in three or more PCR replicates for the unmixed post-treatment samples, not observed in any PCR replicates of the unmixed pre-treatment samples, and present in at least 12 total PCR replicates.

Generating count matrix

```
annotated_count_matrix <- annotated_counts %>% select(feature_id, pcr_id, count) %>%
      spread(pcr_id,count,fill = 0)
rownames(annotated_count_matrix) <- annotated_count_matrix$feature_id
annotated_count_matrix <- annotated_count_matrix %>%
      select(-feature_id) %>% as.matrix()

annotated_count_matrix %>% saveRDS("../data/raw_counts_matrix_good_feature_categories.rds")
```

pData for `annotated_count_matrix`

```
anno_counts_pdata <- annotated_counts %>% select(-count, -feature_id, -cat, -pipe) %>% unique()
rownames(anno_counts_pdata) <- anno_counts_pdata$pcr_id
```

```
anno_counts_pdata %>% saveRDS("../raw_counts_pdata_good_feature_categories.rds")
```

```
annotated_counts %>% select(feature_id, pipe) %>% unique() %>%
       saveRDS("../feature_id_pipeline_good_feature_categories.rds")
```