logFC Error Metrics

Nate Olson

2017-11-14

0.1 Objective

Generate data frames for log fold-change assessment.

Data frames

- `logFC_pre` - log fold-change estimates and expected values for pre-specific and pre-dominant features.
- `logFC_feature_summary` - feature-level linear model fit summary

```
### Loading Data --------------------------------------------------------------
pa_summary_df <- readRDS("~/Desktop/pa_summary_anno_df.RDS") %>%
    select(pipe, biosample_id, feature_id,
           T00, T20, pa_mixed) #%>%
    # filter(biosample_id %in% paste0("E01JH00", c("04","11","16")))

logFC_df <- readRDS("~/Desktop/logFC_edgeR_df.rds") %>%
    # filter(biosample_id %in% paste0("E01JH00", c("04","11","16"))) %>%
    rename(feature_id = OTUname)

## Combining logFC estimates
logFC_df <- readRDS("~/Desktop/logFC_nb_df.rds") %>%
    rename(nb_logFC = logFC) %>%
    right_join(logFC_df)
```

```
## Joining, by = c("biosample_id", "T1", "T2", "pipe", "feature_id")
```

```
## Fixing taxonomy for consistency across pipelines
logFC_df <- logFC_df %>%
    mutate(Rank2 = str_replace(Rank2, "p__",""),
           Rank3 = str_replace(Rank3, "c__",""),
           Rank4 = str_replace(Rank4, "o__",""),
           Rank5 = str_replace(Rank5, "f__",""),
           Rank6 = str_replace(Rank6, "g__",""),
           Rank7 = str_replace(Rank7, "s__","")) %>%
    left_join(pa_summary_df)
```

```
## Joining, by = c("biosample_id", "pipe", "feature_id")
```

0.2 log fold-change error metrics

The slope estimate and R^2 for feature level linear models of the expected and estimated log fold-change. The bias metric is 1-slope estimate and the model R^2 value is the variance metric.

```
## Fitting a linear model to individual features
logFC_feature_fit <- logFC_pre %>%
    filter(T1 != 0, T2 != 20) %>%
    select(pipe, biosample_id, feature_id, exp_logFC, logFC, nb_logFC) %>%
    rename(edgeR = logFC) %>%
    gather("logFC_est","logFC",
```

```r
            -pipe, -biosample_id, -feature_id, -exp_logFC) %>%
    group_by(pipe, biosample_id, feature_id, logFC_est) %>%
    nest() %>%
    mutate(data = map(data, filter,
                      ## Excluding comparisons where logFC estimates are undefined for Inf
                      ## this is for negative binomial logFC estimates
                      !is.na(logFC), logFC != Inf, logFC != -Inf)) %>%
    mutate(fit = map(data, ~lm(logFC ~ exp_logFC, data = .)))

logFC_feature_glance <- logFC_feature_fit %>%
    mutate(fit_glance = map(fit, glance)) %>%
    select(-data, -fit) %>%
    unnest()

logFC_feature_tidy <- logFC_feature_fit %>%
    mutate(fit_glance = map(fit, tidy)) %>%
    select(-data, -fit) %>%
    unnest()

logFC_feature_slope <- logFC_feature_tidy %>%
    filter(term == "exp_logFC") %>%
    select(-term) %>%
    rename(slope = estimate)

logFC_feature_summary <- logFC_feature_slope %>%
    select(-statistic, -p.value) %>%
    left_join(logFC_feature_glance)
```

```
## Joining, by = c("pipe", "biosample_id", "feature_id", "logFC_est")
```

0.3 Saving data frames

```r
saveRDS(logFC_pre, "~/Desktop/logFC_pre.RDS")
saveRDS(logFC_feature_summary, "~/Desktop/logFC_feature_summary.RDS")
```