Post-review analysis

```
library(MASS)

library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(bio3d)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v tibble  3.1.8     v purrr   0.3.4
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
library(colorspace)
library(cowplot)
library(ggpubr)
```

```
##
## Attaching package: 'ggpubr'
##
## The following object is masked from 'package:cowplot':
##
##     get_legend
```

```
library(patchwork)
```

```
##
## Attaching package: 'patchwork'
##
## The following object is masked from 'package:cowplot':
##
```

```
##     align_plots
##
## The following object is masked from 'package:MASS':
##
##     area
```

```
source("dms_analysis_utilities.R")

#source("oct1_dms_read_enrich.R")
```

Read score files

```
# Enrich2 score files
oct1_combined_scores_file ="../data/oct1_combined_scores.csv"
oct1_combined_scores <- read_csv(oct1_combined_scores_file)
```

```
## Rows: 11573 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): hgvs, mutation_type, variants, wt_pos
## dbl (23): SM73_0_SE, SM73_0_epsilon, SM73_1_SE, SM73_1_epsilon, GFP_SE, GFP_...
## lgl  (1): is.wt
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
oct1_combined_scores <- oct1_combined_scores %>% mutate(pos = as.integer(pos),
                                                        len = as.integer(len))
```

```
oct1_wt = "MPTVDDILEQVGESGWFQKQAFLILCLLSAAFAPICVGIVFLGFTPDHHCQSPGVAELSQRCGWSPAEELNYTVPGLGPAGEAFLGQCRRYE
```

Plot the correlation between the two scores and quantify it.

Plots all (in black) and synonymous (in red) variants, regression line, as well as pearson correlation.

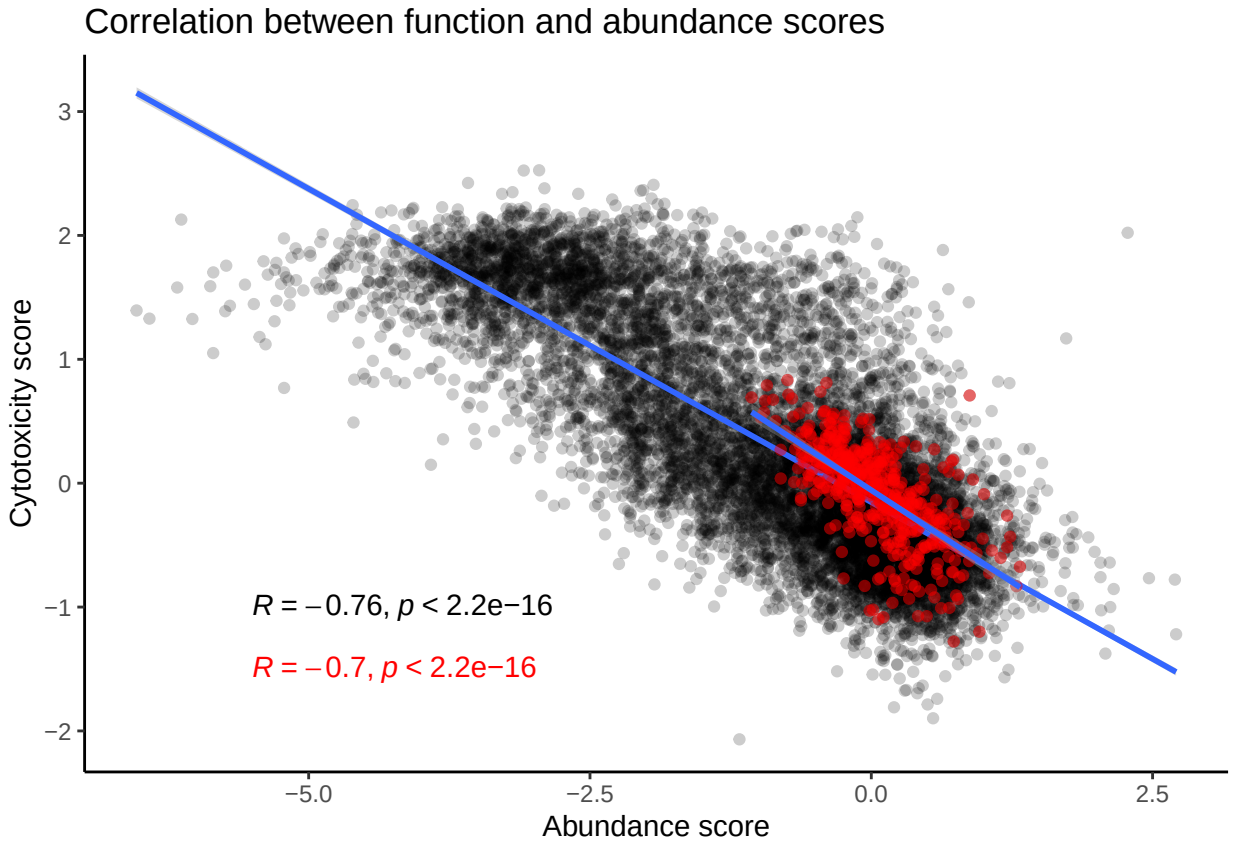
```
score_plot <- ggplot(oct1_combined_scores %>% filter(mutation_type != "X"),
                     aes(y = SM73_1_score, x = GFP_score)) +
  geom_point(alpha = 0.2) +
  ggtitle('Correlation between function and abundance scores') +
  stat_cor(method = "spearman", label.x = -5.5, label.y = -1, color = 'black') +
  geom_smooth(method='lm', se = TRUE) +
  geom_point(data = oct1_combined_scores %>% filter(mutation_type == "S"),
             color = 'red', alpha = 0.5) +
  stat_cor(data = oct1_combined_scores %>% filter(mutation_type == "S"),
           method = "spearman", label.x = -5.5, label.y = -1.5, color = 'red') +
  geom_smooth(data = oct1_combined_scores %>% filter(mutation_type == "S"),
              method='lm', se = TRUE) +
  ylab("Cytotoxicity score") +
  xlab("Abundance score") +
  theme_classic()

score_plot
```

```
## Warning: Removed 907 rows containing non-finite values (stat_cor).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 907 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9 rows containing non-finite values (stat_cor).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 9 rows containing non-finite values (stat_smooth).
## Warning: Removed 907 rows containing missing values (geom_point).
## Warning: Removed 9 rows containing missing values (geom_point).
```



Investigate the correlation between the baseline counts and abundance scores

```
getwd()
```

```
## [1] "/Users/bartleby/Desktop/Projects/OCT1/OCT1_DMS/Figures"
```

```
oct1_counts_1SM73_T0_R1_file ="../data/counts/OCT1_full/Cy1a.csv"
oct1_counts_1SM73_T0_R2_file ="../data/counts/OCT1_full/Cy1b.csv"
oct1_counts_1SM73_T0_R3_file ="../data/counts/OCT1_full/C.csv"

oct1_counts_1SM73_T0_R1 <- read_delim(oct1_counts_1SM73_T0_R1_file, col_select = !1)
```

```
## New names:
## Rows: 11572 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): mutation_type, name, codon, mutation, hgvs dbl (5): count, pos, chunk_pos,
## chunk, length
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
oct1_counts_1SM73_T0_R2 <- read_delim(oct1_counts_1SM73_T0_R2_file, col_select = !1)
```

```
## New names:
## Rows: 11572 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): mutation_type, name, codon, mutation, hgvs dbl (5): count, pos, chunk_pos,
## chunk, length
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
oct1_counts_1SM73_T0_R3 <- read_delim(oct1_counts_1SM73_T0_R3_file, col_select = !1)
```

```
## New names:
## Rows: 11572 Columns: 10
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): mutation_type, name, codon, mutation, hgvs dbl (5): count, pos, chunk_pos,
## chunk, length
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
oct1_scores_counts <- full_join(oct1_counts_1SM73_T0_R1, oct1_combined_scores)
```

```
## Joining, by = c("pos", "mutation_type", "hgvs")
```

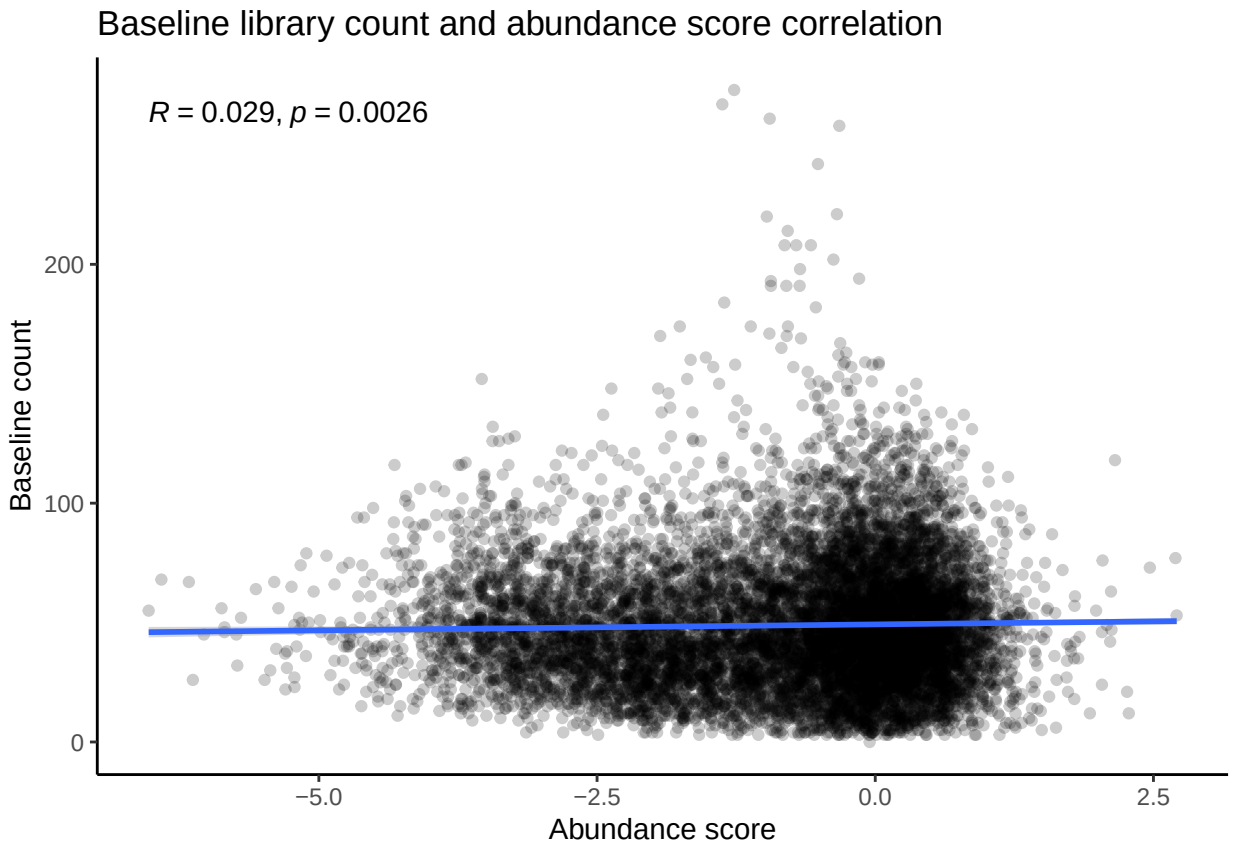
```
score_baseline_plot_abundance <- ggplot(oct1_scores_counts %>% filter(mutation_type != "X"),
                     aes(y = count, x = GFP_score)) +
  ggtitle('Baseline library count and abundance score correlation') +
  geom_point(alpha = 0.2) +
  stat_cor(method = "spearman", color = 'black') +
  geom_smooth(method='lm', se = TRUE) +
  ylab("Baseline count") +
  xlab("Abundance score") +
  theme_classic()

score_baseline_plot_abundance
```

```
## Warning: Removed 816 rows containing non-finite values (stat_cor).
```

```
## `geom_smooth()` using formula 'y ~ x'
```

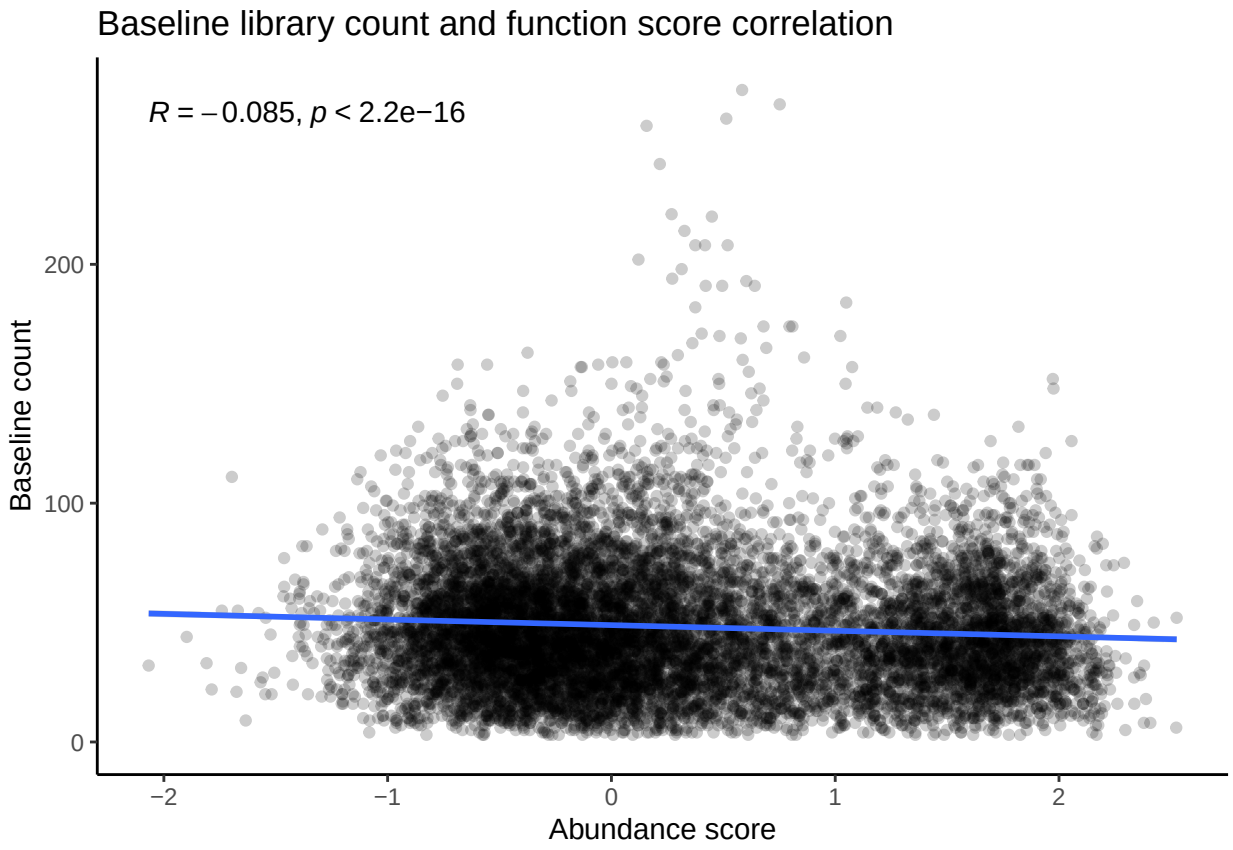
```
## Warning: Removed 816 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 816 rows containing missing values (geom_point).
```




```
score_baseline_plot_sm73 <- ggplot(oct1_scores_counts %>% filter(mutation_type != "X"),
                      aes(y = count, x = SM73_1_score)) +
  ggtitle('Baseline library count and function score correlation') +
  geom_point(alpha = 0.2) +
  stat_cor(method = "spearman", color = 'black') +
  geom_smooth(method='lm', se = TRUE) +
  ylab("Baseline count") +
  xlab("Abundance score") +
  theme_classic()

score_baseline_plot_sm73
```

```
## Warning: Removed 460 rows containing non-finite values (stat_cor).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 460 rows containing non-finite values (stat_smooth).
## Warning: Removed 460 rows containing missing values (geom_point).
```


Baseline library count and function score correlation
R = −0.085, p < 2.2e−16
Baseline count
200
100
0
−2
−1
0
1
2
Abundance score