data_summary_statistics

Stefan Meinke

2024-12-19

Script for some (summary) statistics of the DGE results

load required libraries

```
required_libraries <- c("tidyverse",
                        "magrittr",
                        "ggplot2",
                        "readxl",
                        "knitr")
```

Check and install/load packages

Load the DGE result file

```
DESeq_results <- read_xlsx("results/DESeq_results.xlsx")

DESeq_results_sig <- DESeq_results %>%
  filter(padj < 0.05 & abs(log2FoldChange) > log2(1.5))
```

calculate the total number of significantly regulated genes

```
DESeq_results_sig %>%
  group_by(group) %>%
  summarize(n = n())
```

```
## # A tibble: 2 x 2
##   group                        n
##   <chr>                    <int>
## 1 3.weeks.MM vs baseline    1904
## 2 3.weeks.RMPI vs baseline   399
```

number of up- and downregulated genes per group

```
DESeq_results_sig %>%
  mutate(regulation = ifelse(log2FoldChange > 0, "upregulated", "downregulated")) %>%
  group_by(group, regulation) %>%
  summarize(n = n())
```

```
## `summarise()` has grouped output by 'group'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 4 x 3
## # Groups:   group [2]
##   group                  regulation        n
##   <chr>                  <chr>         <int>
## 1 3.weeks.MM vs baseline   downregulated   891
## 2 3.weeks.MM vs baseline   upregulated    1013
## 3 3.weeks.RMPI vs baseline downregulated   115
## 4 3.weeks.RMPI vs baseline upregulated     284
```

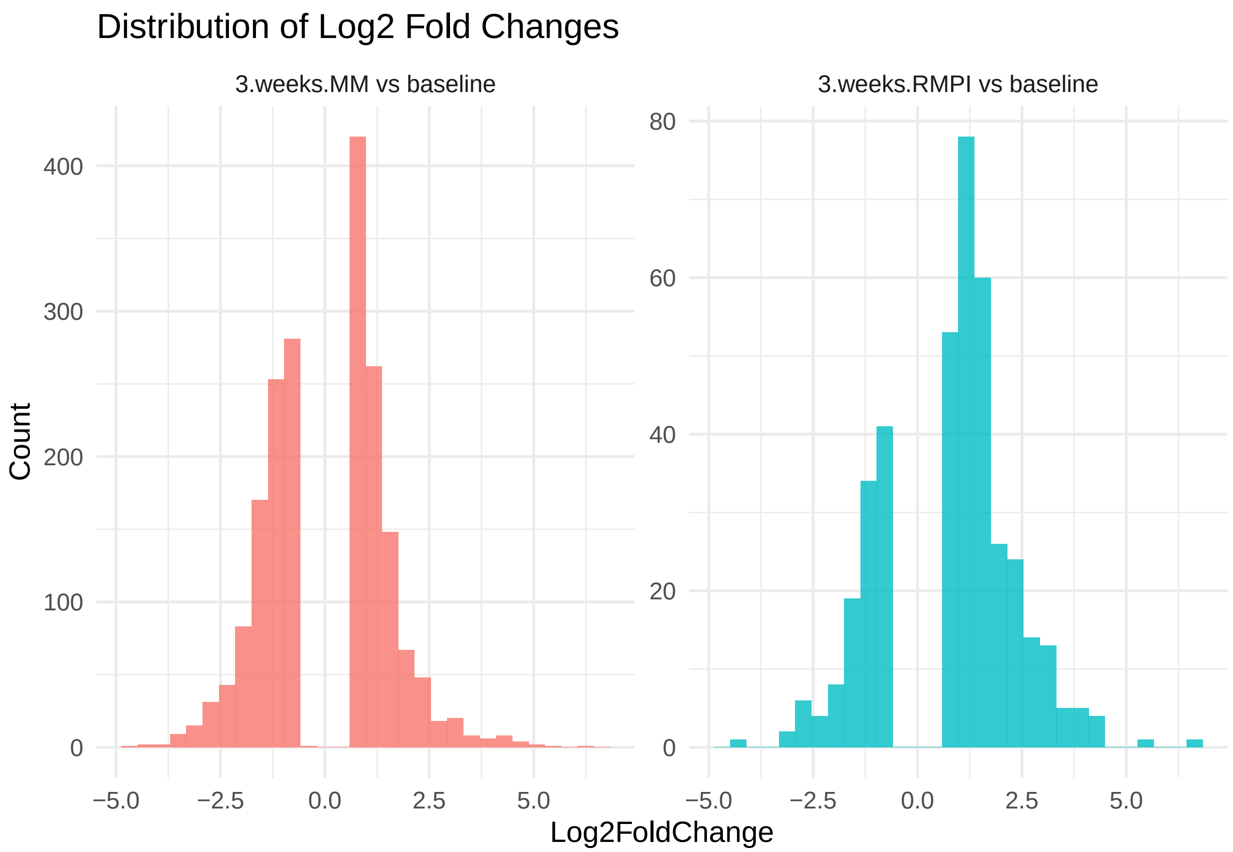
summary statistics for fold changes per group

```
DESeq_results_sig %>%
  group_by(group) %>%
  summarize(
    mean_log2fc = mean(log2FoldChange, na.rm = TRUE),
    median_log2fc = median(log2FoldChange, na.rm = TRUE),
    sd_log2fc = sd(log2FoldChange, na.rm = TRUE),
    min_log2fc = min(log2FoldChange, na.rm = TRUE),
    max_log2fc = max(log2FoldChange, na.rm = TRUE)
  )
```

```
## # A tibble: 2 x 6
##   group                mean_log2fc median_log2fc sd_log2fc min_log2fc max_log2fc
##   <chr>                      <dbl>         <dbl>     <dbl>      <dbl>      <dbl>
## 1 3.weeks.MM vs basel~      0.0569         0.645      1.52      -4.83       6.39
## 2 3.weeks.RMPI vs bas~      0.808          1.10       1.60      -4.22       6.50
```

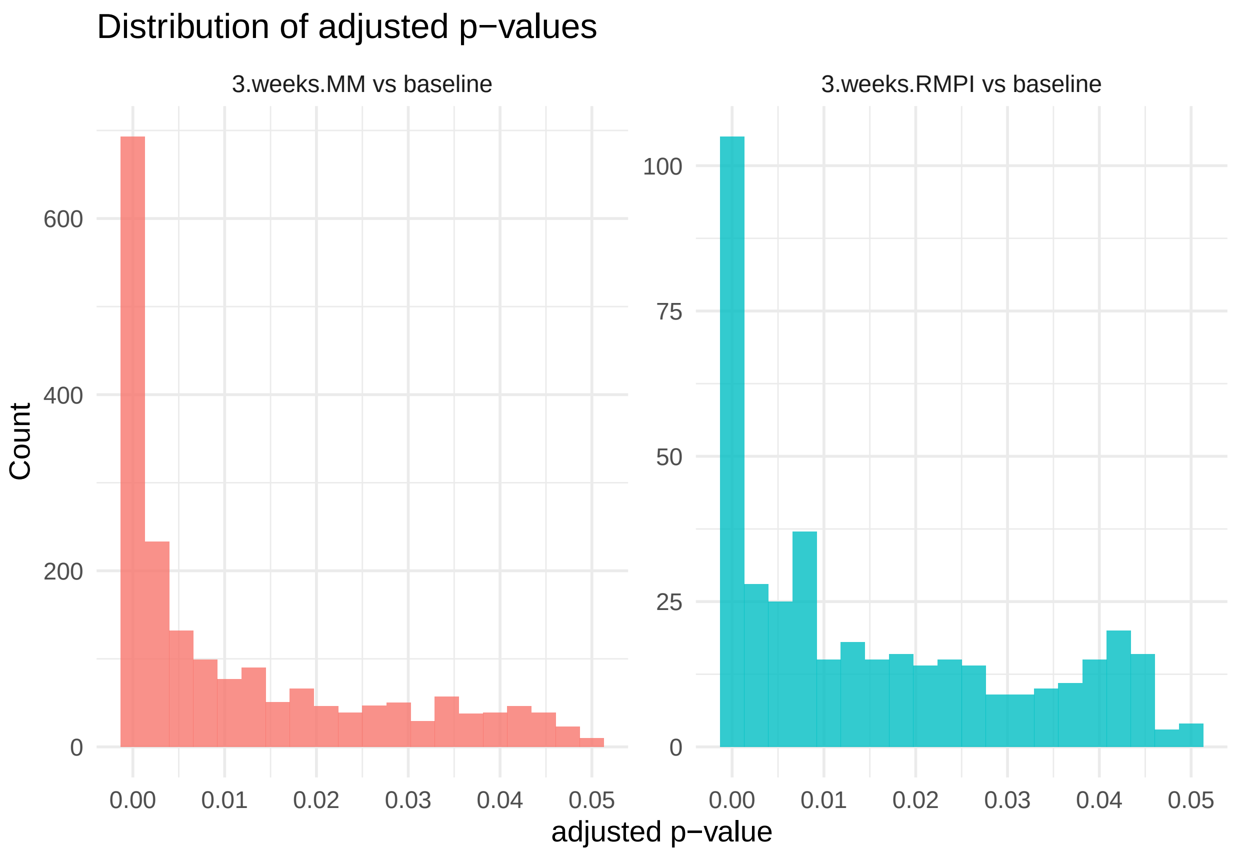
Histogram of fold changes

```
ggplot(DESeq_results_sig, aes(x = log2FoldChange, fill = group)) +
  geom_histogram(bins = 30, alpha = 0.8, position = "identity") +
  facet_wrap(~ group, scales = "free_y") +
  theme_minimal() +
  labs(x = "Log2FoldChange", y = "Count", title = "Distribution of Log2 Fold Changes") +
  theme(legend.position = "none")
```



p-value distribution

```r
ggplot(DESeq_results_sig, aes(x = padj, fill = group)) +
  geom_histogram(bins = 20, alpha = 0.8, position = "identity") +
  facet_wrap(~ group, scales = "free_y") +
  theme_minimal() +
  labs(x = "adjusted p-value", y = "Count", title = "Distribution of adjusted p-values") +
  theme(legend.position = "none")
```



Top genes

```
top_genes <- DESeq_results_sig %>%
  drop_na %>%
  group_by(group) %>%
  arrange(desc(log2FoldChange)) %>%
  slice_head(n = 10) %>%
  mutate(regulation = "upregulated") %>%
  bind_rows(
    DESeq_results_sig %>%
      group_by(group) %>%
      arrange(log2FoldChange) %>%
      slice_head(n = 10) %>%
      mutate(regulation = "downregulated")
  )
```