MOCHIS: Senescence Data Analysis

Alan Aw

3/29/2022

1 Introduction

MOCHIS is a software that allows the user to perform flexible non-parametric tests of differential gene expression. Such tests include the popular Mann-Whitney (Wilcoxon rank sum) test, which was recently promoted by Li et al. (2022) as an approach to perform differential analysis on RNA-seq data without incurring an inflated false positive rate.

In this markdown document, we explore how MOCHIS can detect multiple kinds of differential gene expression signatures, including mean shifts or dispersion shifts. Dispersion shifts have recently been shown to characterize age-related changes in gene expression (see Schaum et al., 2020 and Yamamoto and Chung et al., 2022+). In particular, we:

- perform multiple kinds of two-sample tests on all single-cell tissue data provided in *Tabula muris senis*
- report and compare findings across the different kinds of tests

For **Section 3 (Analysis)**, all our analyses are followed by a summary of key findings, to help the reader quickly grasp the main points.

```
## Setup
library(tidyverse) # for data analysis
tab_mur_dir <- "~/Documents/research/spacing_stats/030222/"
source("~/Documents/research/spacing_stats/012522/main_draft0.R")
```

2 Data

Publicly available *mus musculus* (house mice) single-cell RNA-seq data from the Chan-Zuckerberg Initiative (also known as *Tabula Muris Senis*) is used. We download senescence datasets from here. These datasets are made up of single cell gene transcript levels measured using Smart-Seq2, across 22 distinct mice tissues. For each tissue, the cells originate from mice that are either 3 months, 18 months or 24 months old (with the exception of the mammary gland tissue, which has 3 months, 18 months and 21 months). There are also other cell labels like tissue location (identified with guidance from biologists) and mice sex.

Below, we perform the Mann-Whitney test to identify genes that are differentially expressed, also known as differentially expressed genes (DEGs), across age groups. We compare each pair of age group, so that for each gene $\binom{3}{2} = 3$ tests are performed.

We restrict our analysis to those regions where the zero counts are the fewest, using an 80% cut-off. This avoids running tests on genes that have pronounced zero inflation, which hinders the detection of differential expression.

We additionally compute a "ratio of variances" index, which heuristic measures of the difference in dispersion across the pair of age groups. The larger the ratio of variances, the more differentially dispersed the gene expression between the pair of age groups.

```
## Perform analysis for each tissue
## There are 22 tissues
all_tissues <- read.table(paste0(tab_mur_dir,"tissues/tissues.txt"))$V1
for (tissue in all_tissues) {
  message(date(), paste0(": Reading in and analyzing data from ", tissue))

  if (file.exists(paste0(tab_mur_dir, "tissues/", tissue, "/p_val_table.csv"))) {
    message(date(), ": p-value table already generated, moving onto next tissue...")
  } else {
    tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))
    class(tissue_smartseq2_data)

    smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
    smartseq2_sparse_mat %>% dim()

    ## Get cutoff and restrict to only those genes
    message(date(),
            ": Using cutoff = 0.8 to select genes with non-zero reads at least cutoff...")
    cutoff <- round(0.8*dim(smartseq2_sparse_mat@counts)[2])
    rowSums(smartseq2_sparse_mat@counts[1:dim(smartseq2_sparse_mat@counts)[1],
                                        1:dim(smartseq2_sparse_mat@counts)[2]] != 0) %>%
      summary()
    row_ids <- which(
      rowSums(
        smartseq2_sparse_mat@counts[1:dim(smartseq2_sparse_mat@counts)[1],
                                    1:dim(smartseq2_sparse_mat@counts)[2]] != 0) > cutoff)
    smartseq2_high_exp_sparse_mat <-
      smartseq2_sparse_mat@counts[row_ids, 1:dim(smartseq2_sparse_mat@counts)[2]]
    message(date(), paste0(": Found ",
                           length(row_ids),
                           " genes out of ",
                           dim(smartseq2_sparse_mat@counts)[2],
                           " genes meeting the cutoff threshold..."))

    ## Grab age labels
    identical(names(tissue_smartseq2_data$age), colnames(smartseq2_high_exp_sparse_mat))
    smartseq2_df <- t(as.matrix(smartseq2_high_exp_sparse_mat)) %>% as.data.frame()
    smartseq2_df$age <- tissue_smartseq2_data$age

    ## Run Mann-Whitney test for genes
    gene_names <- colnames(smartseq2_df)[1:length(row_ids)]
    results_df <- data.frame(TRANSCRIPT = character(),
                             MANN_WHITNEY_3_18 = numeric(),
                             MANN_WHITNEY_18_24 = numeric(),
                             MANN_WHITNEY_24_3 = numeric(),
                             VAR_3_18 = numeric(),
                             VAR_18_24 = numeric(),
                             VAR_24_3 = numeric())

    message("How many cells of each age group?")
    print(table(smartseq2_df$age))

    ## Run test for each gene
```

```
    for (i in 1:length(gene_names)) {
      #message(date(), paste0(": Performing two-sample test for ", gene_names[i]))

      # Subset to that transcript
      to_run_test <- smartseq2_df %>% select(c(gene_names[i], "age"))

      # Separate out the 3m, 18m, and 24m reads (counts/million)
      if (tissue == "mammary-gland") {
        message(date(),
                ": Reminder that mammary-gland has 3m,
                18m and 21m age groups, so interpret 24m as 21m...")
        age_3m <- (to_run_test %>% subset(age == "3m"))[,1]
        age_18m <- (to_run_test %>% subset(age == "18m"))[,1]
        age_24m <- (to_run_test %>% subset(age == "21m"))[,1]
      } else {
        age_3m <- (to_run_test %>% subset(age == "3m"))[,1]
        age_18m <- (to_run_test %>% subset(age == "18m"))[,1]
        age_24m <- (to_run_test %>% subset(age == "24m"))[,1]
      }

      #message(date(), ": Running Mann-Whitney tests...")
      wrs_test_3_18 <- wilcox.test(x = age_3m, y = age_18m, alternative = "two.sided")
      wrs_test_18_24 <- wilcox.test(x = age_18m, y = age_24m, alternative = "two.sided")
      wrs_test_24_3 <- wilcox.test(x = age_3m, y = age_24m, alternative = "two.sided")

      #message(date(), ": Computing ratio of variances...")
      var_3_18 <- max(var(age_3m)/var(age_18m), var(age_18m)/var(age_3m))
      var_18_24 <- max(var(age_18m)/var(age_24m),var(age_24m)/var(age_18m))
      var_24_3 <- max(var(age_24m)/var(age_3m),var(age_3m)/var(age_24m))

      results_df <- rbind(results_df,
                          data.frame(TRANSCRIPT = gene_names[i],
                                     MANN_WHITNEY_3_18 = wrs_test_3_18$p.value,
                                     MANN_WHITNEY_18_24 = wrs_test_18_24$p.value,
                                     MANN_WHITNEY_24_3 = wrs_test_24_3$p.value,
                                     VAR_3_18 = var_3_18,
                                     VAR_18_24 = var_18_24,
                                     VAR_24_3 = var_24_3))

    }
    message(date(), paste0(": Saving results for ", tissue))
    write.csv(results_df, file = paste0(tab_mur_dir,"tissues/",tissue,"/p_val_table.csv"))
  }
}
```

2.1 Mann-Whitney DEGs

Given we have the tables of p-values and ratios of variances from the previous step, we now select genes whose p-values, after a Benjamini-Hochberg adjustment procedure, lie below or equal to a 0.05 significance level. These are Mann-Whitney significant genes that would be flagged as potentially carrying biological signal in a typical differential expression analysis procedure.

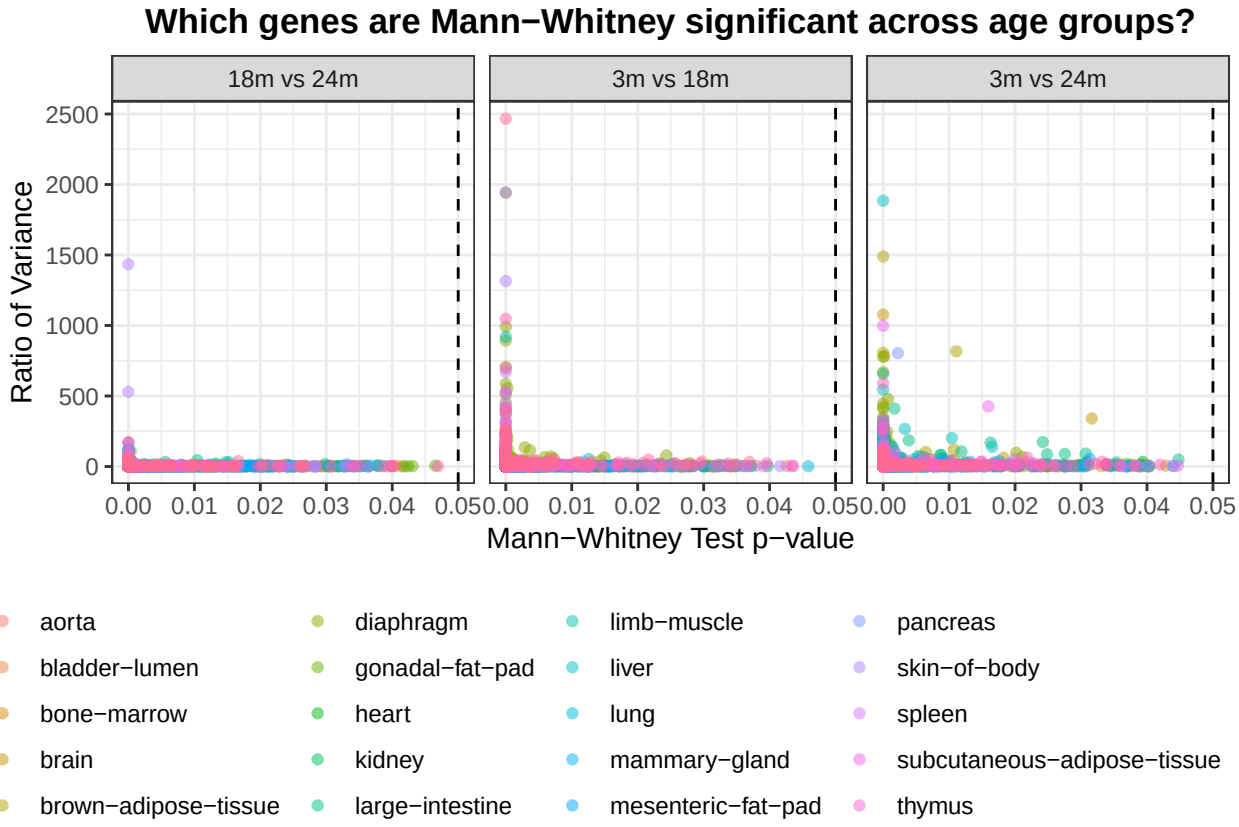
```
tissue_transcript_3_18 <- data.frame(TRANSCRIPT = character(),
                                     MANN_WHITNEY = numeric(),
```

```
                                    VARIANCE_RATIO = numeric(),
                                    TISSUE = character())

tissue_transcript_18_24 <- data.frame(TRANSCRIPT = character(),
                                     MANN_WHITNEY = numeric(),
                                     VARIANCE_RATIO = numeric(),
                                     TISSUE = character())

tissue_transcript_24_3 <- data.frame(TRANSCRIPT = character(),
                                    MANN_WHITNEY = numeric(),
                                    VARIANCE_RATIO = numeric(),
                                    TISSUE = character())

for (tissue in all_tissues) {
  message(date(), paste0(": Reading in summary of p-values and ratios of variances for ",
                         tissue))
  tissue_mann_whitney_df <- read.csv(paste0(tab_mur_dir,
                                            "tissues/",
                                            tissue,
                                            "/p_val_table.csv"))

  ## Pick genes where one of the three pairs (3m, 18m, 24m)
  ## has significant p-value at FDR 0.05 control
  selected_genes_3_18 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MANN_WHITNEY_3_18,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MANN_WHITNEY_3_18", "VAR_3_18")) # [!] Benjamini-Hochberg
  colnames(selected_genes_3_18) <- c("TRANSCRIPT", "MANN_WHITNEY", "VARIANCE_RATIO")
  selected_genes_3_18$TISSUE <- rep(tissue, nrow(selected_genes_3_18))
  tissue_transcript_3_18 <- rbind(tissue_transcript_3_18, selected_genes_3_18)

  selected_genes_18_24 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MANN_WHITNEY_18_24,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MANN_WHITNEY_18_24", "VAR_18_24"))
  colnames(selected_genes_18_24) <- c("TRANSCRIPT", "MANN_WHITNEY", "VARIANCE_RATIO")
  selected_genes_18_24$TISSUE <- rep(tissue, nrow(selected_genes_18_24))
  tissue_transcript_18_24 <- rbind(tissue_transcript_18_24, selected_genes_18_24)

  selected_genes_24_3 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MANN_WHITNEY_24_3,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MANN_WHITNEY_24_3", "VAR_24_3"))
  colnames(selected_genes_24_3) <- c("TRANSCRIPT", "MANN_WHITNEY", "VARIANCE_RATIO")
  selected_genes_24_3$TISSUE <- rep(tissue, nrow(selected_genes_24_3))
  tissue_transcript_24_3 <- rbind(tissue_transcript_24_3, selected_genes_24_3)
}

write.csv(tissue_transcript_3_18,
          file = paste0(tab_mur_dir,"tissues/mw_sig_3m_18m.csv"),
```

```
          row.names = FALSE)
write.csv(tissue_transcript_18_24,
          file = paste0(tab_mur_dir,"tissues/mw_sig_18m_24m.csv"),
          row.names = FALSE)
write.csv(tissue_transcript_24_3,
          file = paste0(tab_mur_dir,"tissues/mw_sig_24m_3m.csv"),
          row.names = FALSE)
```

Let us visualize the raw p-values and variance ratios of the Mann-Whitney DEGs fished out from the above procedure.

```
library(tidyverse)
df_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_3m_18m.csv"))
df_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_18m_24m.csv"))
df_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_24m_3m.csv"))

df_3_18$PAIR = rep("3m vs 18m", nrow(df_3_18))
df_18_24$PAIR = rep("18m vs 24m", nrow(df_18_24))
df_24_3$PAIR = rep("3m vs 24m", nrow(df_24_3))

grand_df <- rbind(rbind(df_3_18,df_18_24), df_24_3)
ggplot(grand_df, aes(x = MANN_WHITNEY, y = VARIANCE_RATIO)) +
  geom_point(aes(color = TISSUE), alpha = 0.5) +
  geom_vline(xintercept = 0.05, lty = "dashed") +
  facet_wrap(.~PAIR) +
  xlab("Mann-Whitney Test p-value") +
  ylab("Ratio of Variance") +
  ggtitle("Which genes are Mann-Whitney significant across age groups?") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5,
                                  face = "bold"),
        legend.position = "bottom",
        legend.title = element_blank())
```



Next we look at the distribution, across tissues, of Mann-Whitney significant genes.

```
## Create a theme for plotting pie charts
blank_theme <- theme_minimal()+
  theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank(),
  plot.title=element_text(hjust = 0.5, size=14, face="bold")
  )

## Report the number of MW significant genes
print(paste0(
  "No. MW significant genes for 3m vs 18m: ", nrow(df_3_18)))
```

```
## [1] "No. MW significant genes for 3m vs 18m: 5571"
```

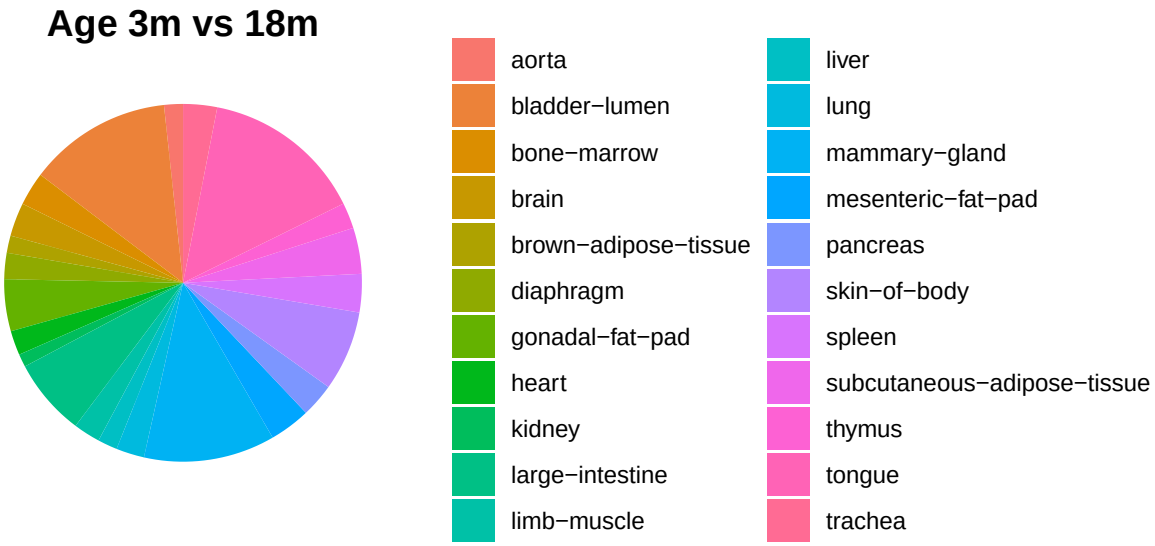
```
print(paste0(
  "No. MW significant genes for 18m vs 24m: ", nrow(df_18_24)))
```

```
## [1] "No. MW significant genes for 18m vs 24m: 5305"
```

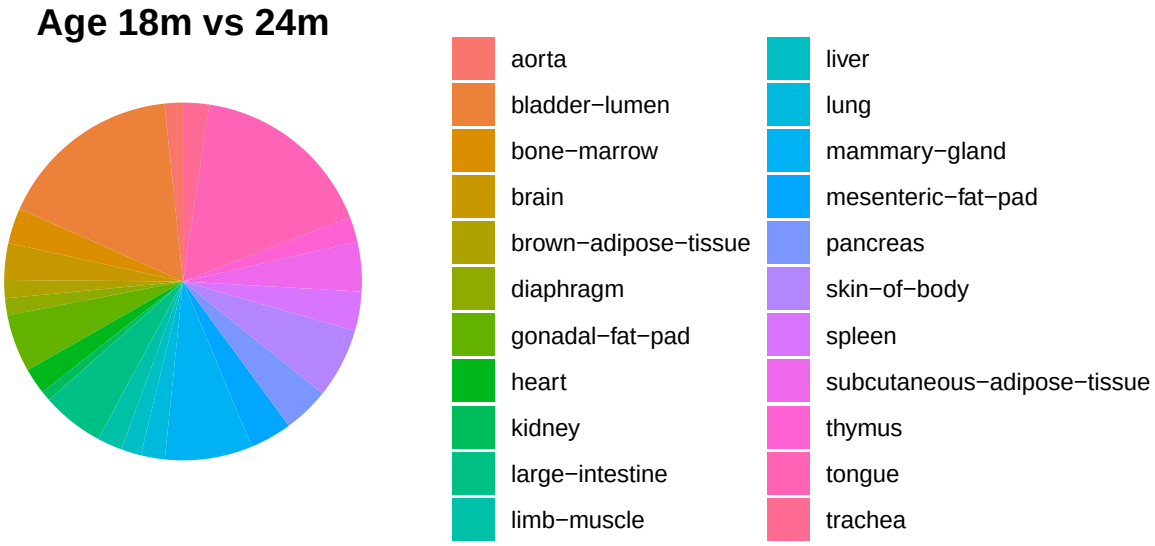
```
print(paste0(
  "No. MW significant genes for 24m vs 3m: ", nrow(df_24_3)))
```

```
## [1] "No. MW significant genes for 24m vs 3m: 5634"
```

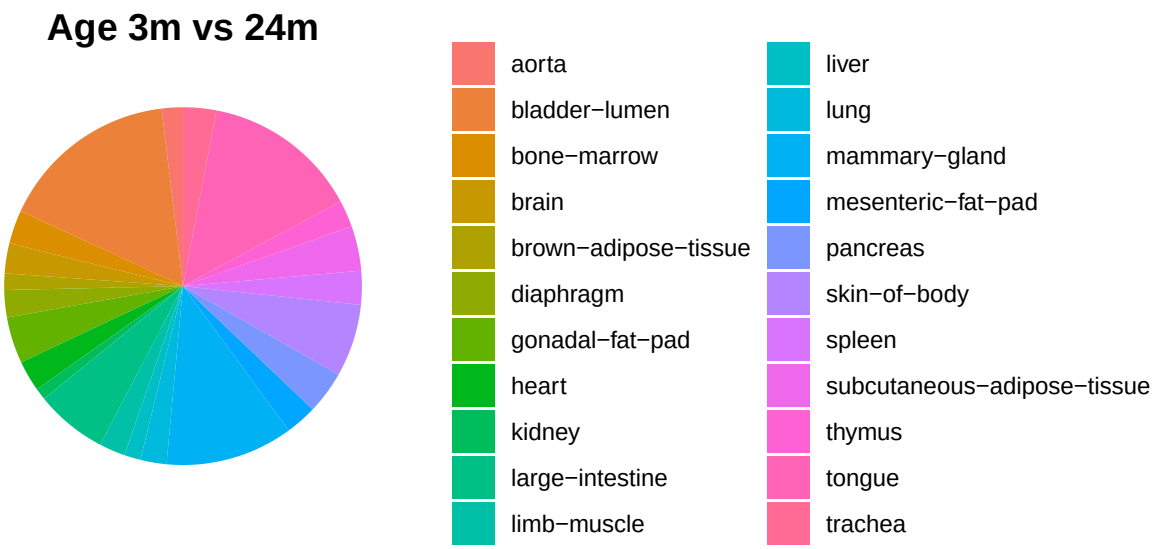
```
piechart_3_18 <- data.frame(group = df_3_18 %>% select(TISSUE) %>%
                              table() %>% names(),
                            value = df_3_18 %>% select(TISSUE) %>%
                              table() %>% as.vector())
ggplot(piechart_3_18, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  ggtitle("Age 3m vs 18m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```




```
piechart_18_24 <- data.frame(group = df_18_24 %>% select(TISSUE) %>%
                              table() %>% names(),
                            value = df_18_24 %>% select(TISSUE) %>%
                              table() %>% as.vector())
ggplot(piechart_18_24, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  ggtitle("Age 18m vs 24m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```




```
piechart_24_3 <- data.frame(group = df_24_3 %>% select(TISSUE) %>%
                              table() %>% names(),
                            value = df_24_3 %>% select(TISSUE) %>%
                              table() %>% as.vector())
ggplot(piechart_24_3, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  ggtitle("Age 3m vs 24m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```



2.2 MOCHIS

We now repeat the DEG identification procedure above, now using our flexible non-parametric testing software MOCHIS. We run MOCHIS with test statistic $||S_{n,k}||^p_{p,\boldsymbol{w}}$. We choose the following parametrization:

- $p = 1$
- $\boldsymbol{w} = \left((\frac{j}{k} - \frac{1}{2})^2 : j = 1, \ldots, k\right)$

This parametrization optimizes detection of dispersion shifts between two samples.

Step 1. Compute p-values.

When computing the p-values, we apply a tie-breaking routine (adding noise ranging from -0.25 to 0.25, which is less than the minimum spacing width of integer counts). To ensure that this routine does not overly contaminate the data, we also compute Mann-Whitney p-values and check that the Mann-Whitney DEGs identified after applying the tie-breaking routine are not markedly different from the original DEGs identified in Section 2.1. We report this latter comparison between post-contamination and original DEGs in Section 2.3. (Heads up: We find little difference.)

```
## Perform analysis for each tissue
## There are 22 tissues
all_tissues <- read.table(paste0(tab_mur_dir,"tissues/tissues.txt"))$V1
for (tissue in all_tissues) {
  message(date(), paste0(": Reading in and analyzing data from ", tissue))

  if (CRAP) {
  #if (file.exists(paste0(tab_mur_dir, "tissues/", tissue, "/mochis_p_val_table.csv"))) {
    message(date(), ": p-value table already generated, moving onto next tissue...")
  } else {
    tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))
    class(tissue_smartseq2_data)

    smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
    smartseq2_sparse_mat %>% dim()

    ## Get cutoff and restrict to only those genes
    message(date(),
            ": Using cutoff = 0.8 to select genes with non-zero reads at least cutoff...")
    cutoff <- round(0.8*dim(smartseq2_sparse_mat@counts)[2])
    rowSums(smartseq2_sparse_mat@counts[1:dim(smartseq2_sparse_mat@counts)[1],
                                        1:dim(smartseq2_sparse_mat@counts)[2]] != 0) %>%
      summary()
    row_ids <- which(
      rowSums(
        smartseq2_sparse_mat@counts[1:dim(smartseq2_sparse_mat@counts)[1],
                                    1:dim(smartseq2_sparse_mat@counts)[2]] != 0) > cutoff)
    smartseq2_high_exp_sparse_mat <-
      smartseq2_sparse_mat@counts[row_ids, 1:dim(smartseq2_sparse_mat@counts)[2]]
    message(date(), paste0(": Found ",
                           length(row_ids),
                           " genes out of ",
                           dim(smartseq2_sparse_mat@counts)[2],
                           " genes meeting the cutoff threshold..."))

    ## Grab age labels
    identical(names(tissue_smartseq2_data$age), colnames(smartseq2_high_exp_sparse_mat))
    smartseq2_df <- t(as.matrix(smartseq2_high_exp_sparse_mat)) %>% as.data.frame()
    smartseq2_df$age <- tissue_smartseq2_data$age

    ## Run Mann-Whitney test for genes
    gene_names <- colnames(smartseq2_df)[1:length(row_ids)]
    results_df <- data.frame(TRANSCRIPT = character(),
                             MOCHIS_3_18 = numeric(),
```

```
                            MW_3_18 = numeric(),
                            MOCHIS_18_24 = numeric(),
                            MW_18_24 = numeric(),
                            MOCHIS_24_3 = numeric(),
                            MW_24_3 = numeric(),
                            VAR_3_18 = numeric(),
                            INV_3_18 = logical(),
                            VAR_18_24 = numeric(),
                            INV_18_24 = logical(),
                            VAR_24_3 = numeric(),
                            INV_24_3 = logical())

  message("How many cells of each age group?")
  print(table(smartseq2_df$age))

  ## Run test for each gene
  for (i in 1:length(gene_names)) {
    #message(date(), paste0(": Performing two-sample test for ", gene_names[i]))

    # Subset to that transcript
    to_run_test <- smartseq2_df %>% select(c(gene_names[i], "age"))

    # Separate out the 3m, 18m, and 24m reads (counts/million)
    if (tissue == "mammary-gland") {
      message(date(),
              ": Reminder that mammary-gland has 3m,
              18m and 21m age groups, so interpret 24m as 21m...")
      age_3m <- (to_run_test %>% subset(age == "3m"))[,1]
      age_18m <- (to_run_test %>% subset(age == "18m"))[,1]
      age_24m <- (to_run_test %>% subset(age == "21m"))[,1]
    } else {
      age_3m <- (to_run_test %>% subset(age == "3m"))[,1]
      age_18m <- (to_run_test %>% subset(age == "18m"))[,1]
      age_24m <- (to_run_test %>% subset(age == "24m"))[,1]
    }

    #  Add noise to break ties
    set.seed(2022)
    noisy_age_3m <- sort(age_3m + runif(n = length(age_3m), min = -1/4, max = 1/4))
    noisy_age_18m <- sort(age_18m + runif(n = length(age_18m), min = -1/4, max = 1/4))
    noisy_age_24m <- sort(age_24m + runif(n = length(age_24m), min = -1/4, max = 1/4))

    #message(date(), ": Running Mann-Whitney tests...")
    wrs_test_3_18 <- wilcox.test(x = noisy_age_3m, y = noisy_age_18m, correct = FALSE)
    wrs_test_18_24 <- wilcox.test(x = noisy_age_18m, y = noisy_age_24m, correct = FALSE)
    wrs_test_24_3 <- wilcox.test(x = noisy_age_3m, y = noisy_age_24m, correct = FALSE)

    if (length(noisy_age_3m) > length(noisy_age_18m)) {
      k <- length(age_18m)+1
      mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
      mochis_test_3_18 <- mochis.test(x = noisy_age_18m,
                                      y = noisy_age_3m,
                                      p = 1,
```

```
                                    wList = mochis_weights,
                                    alternative = "two.sided",
                                    approx = "chebyshev",
                                    n_mom = 100,
                                    python_backend = FALSE)
  } else {
    k <- length(age_3m)+1
    mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
    mochis_test_3_18 <- mochis.test(x = noisy_age_3m,
                                    y = noisy_age_18m,
                                    p = 1,
                                    wList = mochis_weights,
                                    alternative = "two.sided",
                                    approx = "chebyshev",
                                    n_mom = 100,
                                    python_backend = FALSE)
  }

  if (length(noisy_age_18m) > length(noisy_age_24m)) {
    k <- length(age_24m)+1
    mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
    mochis_test_18_24 <- mochis.test(x = noisy_age_24m,
                                    y = noisy_age_18m,
                                    p = 1,
                                    wList = mochis_weights,
                                    alternative = "two.sided",
                                    approx = "chebyshev",
                                    n_mom = 100,
                                    python_backend = FALSE)
  } else {
    k <- length(age_18m)+1
    mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
    mochis_test_18_24 <- mochis.test(x = noisy_age_18m,
                                    y = noisy_age_24m,
                                    p = 1,
                                    wList = mochis_weights,
                                    alternative = "two.sided",
                                    approx = "chebyshev",
                                    n_mom = 100,
                                    python_backend = FALSE)
  }

  if (length(noisy_age_3m) > length(noisy_age_24m)) {
    k <- length(age_24m)+1
    mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
    mochis_test_24_3 <- mochis.test(x = noisy_age_24m,
                                    y = noisy_age_3m,
                                    p = 1,
                                    wList = mochis_weights,
                                    alternative = "two.sided",
                                    approx = "chebyshev",
                                    n_mom = 100,
                                    python_backend = FALSE)
```

```
      } else {
        k <- length(age_3m)+1
        mochis_weights <- sapply(1:k, function(x) {(x/k-0.5)^2})
        mochis_test_24_3 <- mochis.test(x = noisy_age_3m,
                                        y = noisy_age_24m,
                                        p = 1,
                                        wList = mochis_weights,
                                        alternative = "two.sided",
                                        approx = "chebyshev",
                                        n_mom = 100,
                                        python_backend = FALSE)
      }

      #message(date(), ": Computing ratio of variances...")
      var_3_18 <- max(var(age_3m)/var(age_18m), var(age_18m)/var(age_3m))
      var_18_24 <- max(var(age_18m)/var(age_24m),var(age_24m)/var(age_18m))
      var_24_3 <- max(var(age_24m)/var(age_3m),var(age_3m)/var(age_24m))

      invert_3_18 <- ifelse(max(var(age_3m)/var(age_18m),
                                var(age_18m)/var(age_3m)) == var(age_3m)/var(age_18m),
                            TRUE, FALSE)
      invert_18_24 <- ifelse(max(var(age_18m)/var(age_24m),
                                 var(age_24m)/var(age_18m)) == var(age_18m)/var(age_24m),
                             TRUE, FALSE)
      invert_24_3 <- ifelse(max(var(age_3m)/var(age_24m),
                                var(age_24m)/var(age_3m)) == var(age_3m)/var(age_24m),
                            TRUE, FALSE)

      results_df <- rbind(results_df,
                          data.frame(TRANSCRIPT = gene_names[i],
                                     MOCHIS_3_18 = mochis_test_3_18,
                                     MW_3_18 = wrs_test_3_18$p.value,
                                     MOCHIS_18_24 = mochis_test_18_24,
                                     MW_18_24 = wrs_test_18_24$p.value,
                                     MOCHIS_24_3 = mochis_test_24_3,
                                     MW_24_3 = wrs_test_24_3$p.value,
                                     VAR_3_18 = var_3_18,
                                     INV_3_18 = invert_3_18,
                                     VAR_18_24 = var_18_24,
                                     INV_18_24 = invert_18_24,
                                     VAR_24_3 = var_24_3,
                                     INV_24_3 = invert_24_3))

    }
    message(date(), paste0(": Saving results for ", tissue))
    write.csv(results_df, file = paste0(tab_mur_dir,"tissues/",tissue,"/mochis_p_val_table.csv"))
  }
}
```

Step 2. Identify MOCHIS significant genes (with FDR control at 0.05).

```
tissue_transcript_3_18 <- data.frame(TRANSCRIPT = character(),
                                     MOCHIS = numeric(),
                                     VARIANCE_RATIO = numeric(),
```

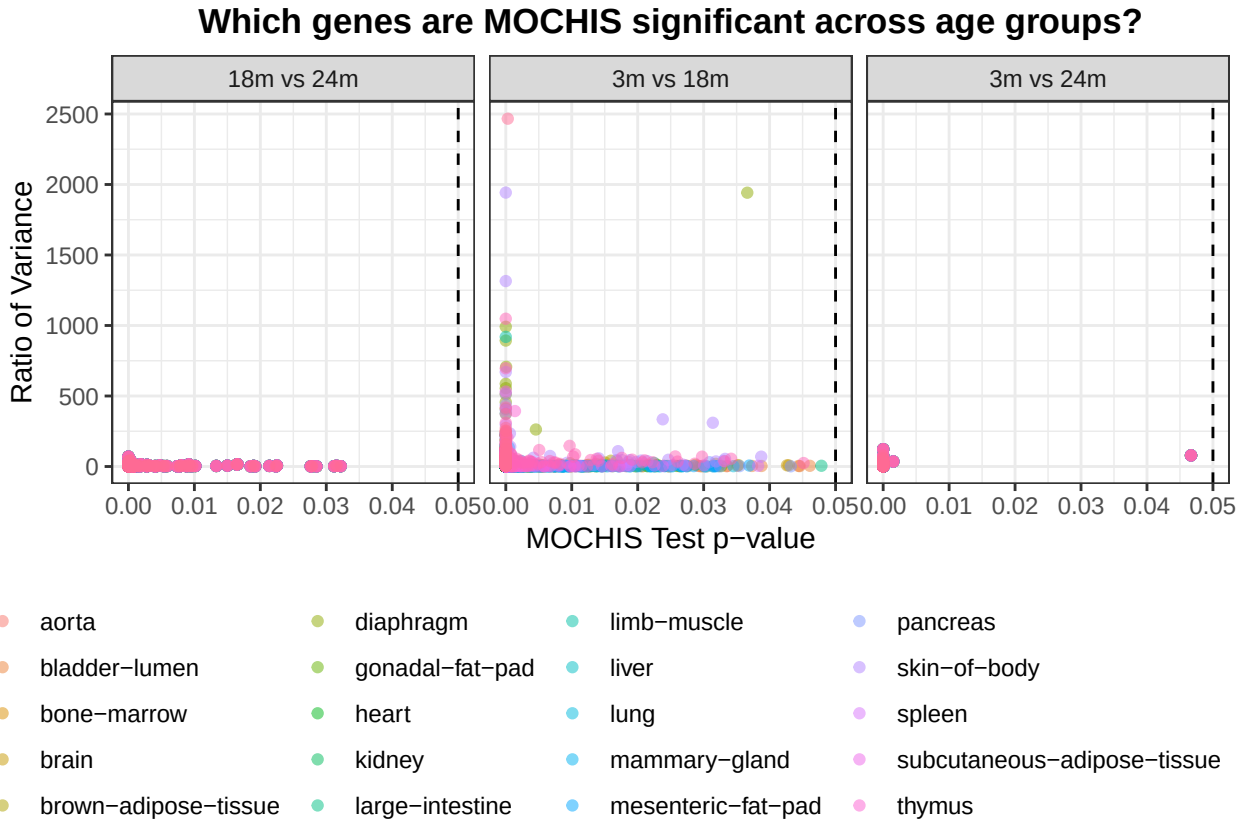
```
                                    TISSUE = character())

tissue_transcript_18_24 <- data.frame(TRANSCRIPT = character(),
                                      MOCHIS = numeric(),
                                      VARIANCE_RATIO = numeric(),
                                      TISSUE = character())

tissue_transcript_24_3 <- data.frame(TRANSCRIPT = character(),
                                     MOCHIS = numeric(),
                                     VARIANCE_RATIO = numeric(),
                                     TISSUE = character())

for (tissue in all_tissues) {
  message(date(), paste0(": Reading in summary of p-values and ratios of variances for ",
                         tissue))
  tissue_mochis_df <- read.csv(paste0(tab_mur_dir,
                                      "tissues/",
                                      tissue,
                                      "/mochis_p_val_table.csv"))

  ## Pick genes where one of the three pairs (3m, 18m, 24m)
  ## has significant p-values
  selected_genes_3_18 <- tissue_mochis_df[
    which(p.adjust(
      tissue_mochis_df$MOCHIS_3_18,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MOCHIS_3_18", "VAR_3_18")) # [!] Benjamini-Hochberg
  colnames(selected_genes_3_18) <- c("TRANSCRIPT", "MOCHIS", "VARIANCE_RATIO")
  selected_genes_3_18$TISSUE <- rep(tissue, nrow(selected_genes_3_18))
  tissue_transcript_3_18 <- rbind(tissue_transcript_3_18, selected_genes_3_18)

  selected_genes_18_24 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MOCHIS_18_24,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MOCHIS_18_24", "VAR_18_24"))
  colnames(selected_genes_18_24) <- c("TRANSCRIPT", "MOCHIS", "VARIANCE_RATIO")
  selected_genes_18_24$TISSUE <- rep(tissue, nrow(selected_genes_18_24))
  tissue_transcript_18_24 <- rbind(tissue_transcript_18_24, selected_genes_18_24)

  selected_genes_24_3 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MOCHIS_24_3,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MOCHIS_24_3", "VAR_24_3"))
  colnames(selected_genes_24_3) <- c("TRANSCRIPT", "MOCHIS", "VARIANCE_RATIO")
  selected_genes_24_3$TISSUE <- rep(tissue, nrow(selected_genes_24_3))
  tissue_transcript_24_3 <- rbind(tissue_transcript_24_3, selected_genes_24_3)
}

write.csv(tissue_transcript_3_18,
          file = paste0(tab_mur_dir,"tissues/mochis_sig_3m_18m.csv"),
          row.names = FALSE)
```

```
write.csv(tissue_transcript_18_24,
          file = paste0(tab_mur_dir,"tissues/mochis_sig_18m_24m.csv"),
          row.names = FALSE)
write.csv(tissue_transcript_24_3,
          file = paste0(tab_mur_dir,"tissues/mochis_sig_24m_3m.csv"),
          row.names = FALSE)
```

Step 3. Visualization.

First, let us visualize the raw p-values and variance ratios of the MOCHIS DEGs fished out from the above procedure.

```
library(tidyverse)
df_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_3m_18m.csv"))
df_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_18m_24m.csv"))
df_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_24m_3m.csv"))

df_3_18$PAIR = rep("3m vs 18m", nrow(df_3_18))
df_18_24$PAIR = rep("18m vs 24m", nrow(df_18_24))
df_24_3$PAIR = rep("3m vs 24m", nrow(df_24_3))

grand_df <- rbind(rbind(df_3_18,df_18_24), df_24_3)
ggplot(grand_df, aes(x = MOCHIS, y = VARIANCE_RATIO)) +
  geom_point(aes(color = TISSUE), alpha = 0.5) +
  geom_vline(xintercept = 0.05, lty = "dashed") +
  facet_wrap(.~PAIR) +
  xlab("MOCHIS Test p-value") +
  ylab("Ratio of Variance") +
  ggtitle("Which genes are MOCHIS significant across age groups?") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5,
                                  face = "bold"),
        legend.position = "bottom",
        legend.title = element_blank())
```



Next we look at the distribution, across tissues, of MOCHIS significant genes.

```
## Report the number of MW significant genes
print(paste0(
  "No. MOCHIS significant genes for 3m vs 18m: ", nrow(df_3_18)))
```

```
## [1] "No. MOCHIS significant genes for 3m vs 18m: 5723"
```

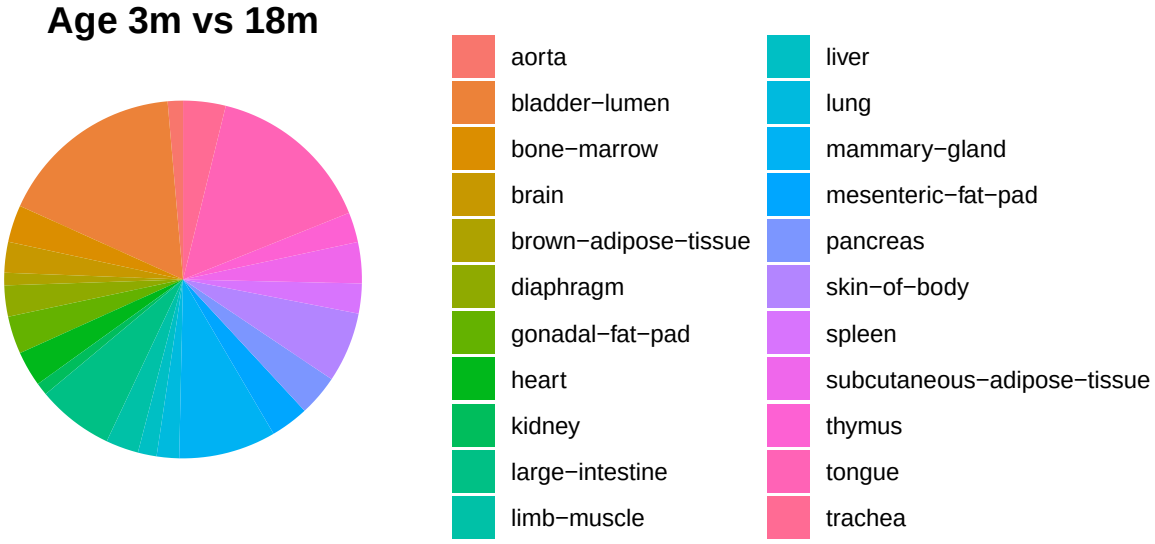
```
print(paste0(
  "No. MOCHIS significant genes for 18m vs 24m: ", nrow(df_18_24)))
```

```
## [1] "No. MOCHIS significant genes for 18m vs 24m: 3410"
```

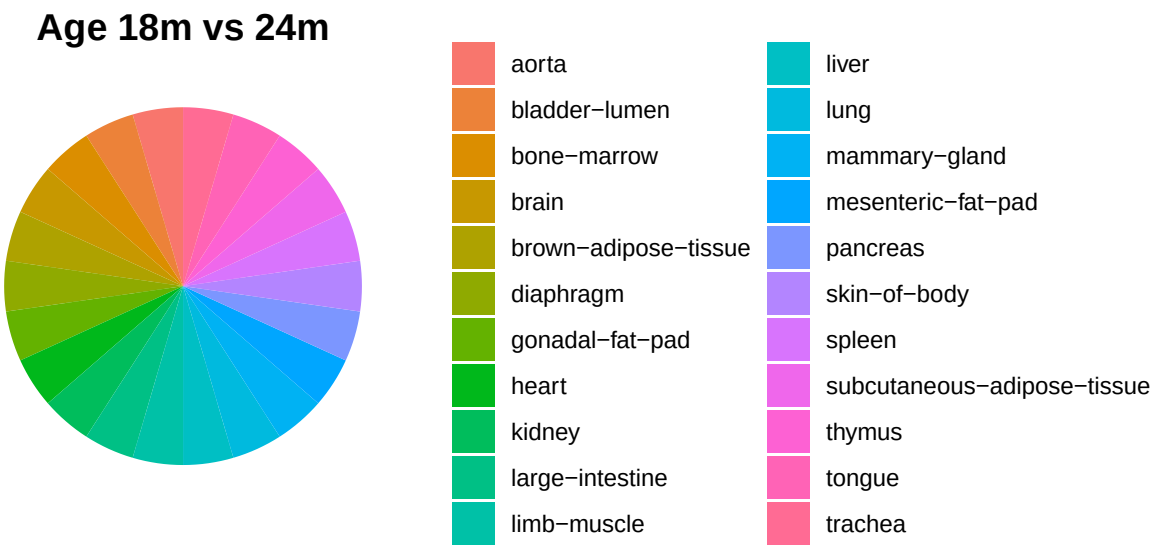
```
print(paste0(
  "No. MOCHIS significant genes for 24m vs 3m: ", nrow(df_24_3)))
```

```
## [1] "No. MOCHIS significant genes for 24m vs 3m: 4730"
```

```
piechart_3_18 <- data.frame(group = df_3_18 %>% select(TISSUE) %>%
                              table() %>% names(),
                            value = df_3_18 %>% select(TISSUE) %>%
                              table() %>% as.vector())
ggplot(piechart_3_18, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  ggtitle("Age 3m vs 18m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```

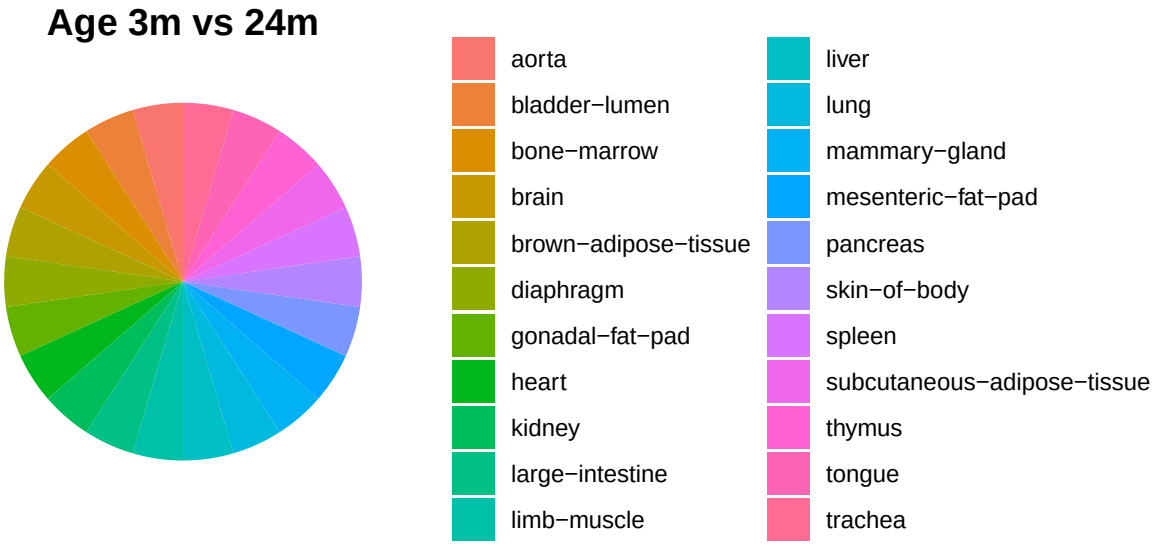



```
piechart_18_24 <- data.frame(group = df_18_24 %>% select(TISSUE) %>%
                               table() %>% names(),
                             value = df_18_24 %>% select(TISSUE) %>%
                               table() %>% as.vector())
ggplot(piechart_18_24, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
  ggtitle("Age 18m vs 24m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```




```
piechart_24_3 <- data.frame(group = df_24_3 %>% select(TISSUE) %>%
                              table() %>% names(),
                            value = df_24_3 %>% select(TISSUE) %>%
                              table() %>% as.vector())
ggplot(piechart_24_3, aes(x="", y=value, fill=group)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y") +
```

```
  ggtitle("Age 3m vs 24m") +
  blank_theme +
  theme(axis.text.x=element_blank(),
        legend.title=element_blank())
```



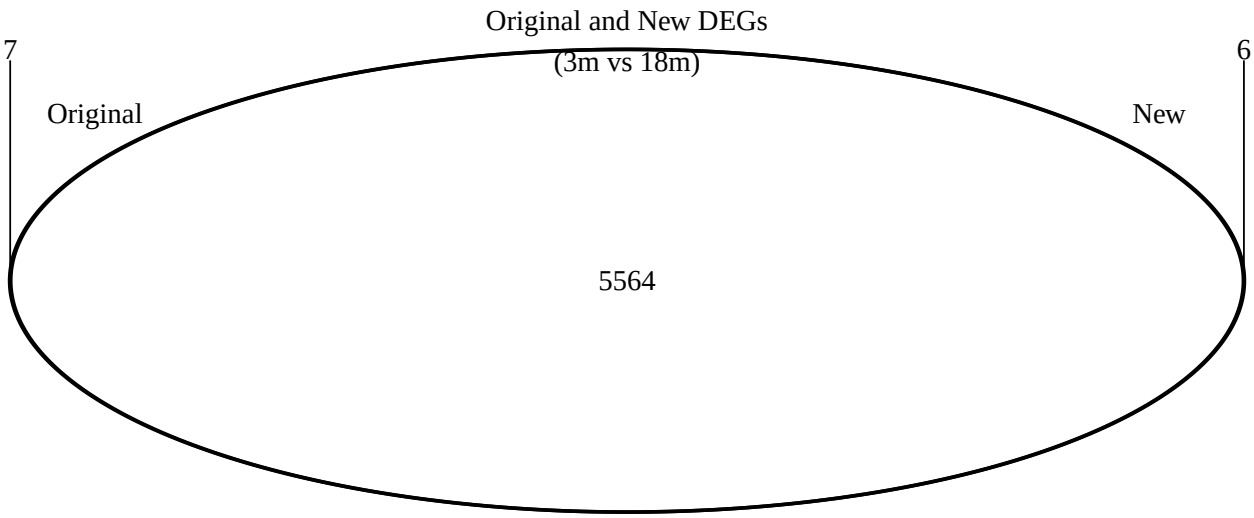
2.3 Impact of Tie Breaking on Mann-Whitney DEGs

In Section 2.2, we raised the issue of tie breaking potentially affecting the significance of Mann-Whitney DEGs. Here, we check for difference between post-contaminated Mann-Whitney DEGs (this Section) and original DEGs (Section 2.1).

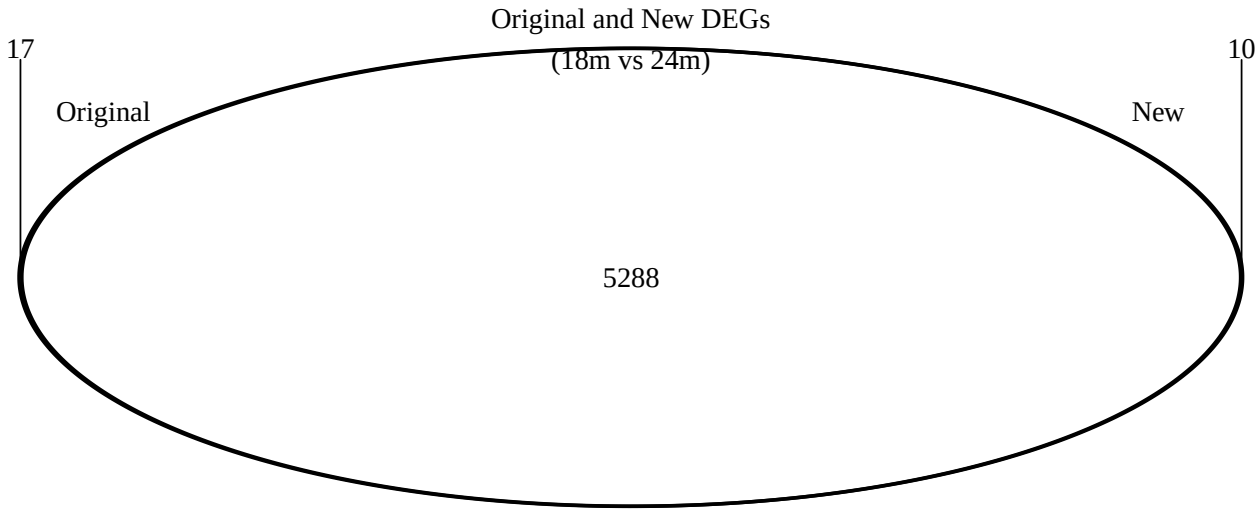
```
## Compare between all pairs (3m vs 18m, 18m vs 24m, 24m vs 3m)
# Define all_tissues again
all_tissues <- read.table(paste0(tab_mur_dir,"tissues/tissues.txt"))$V1

# Comparison of results
tissue_transcript_3_18 <- data.frame(TRANSCRIPT = character(),
                                     NEW_MANN_WHITNEY = numeric(),
                                     TISSUE = character())

tissue_transcript_18_24 <- data.frame(TRANSCRIPT = character(),
                                      NEW_MANN_WHITNEY = numeric(),
                                      TISSUE = character())

tissue_transcript_24_3 <- data.frame(TRANSCRIPT = character(),
                                     NEW_MANN_WHITNEY = numeric(),
                                     TISSUE = character())

for (tissue in all_tissues) {
  message(date(), paste0(": Reading in summary of p-values and ratios of variances for ",
                         tissue))
  tissue_mann_whitney_df <- read.csv(paste0(tab_mur_dir,
                                            "tissues/",
                                            tissue,
                                            "/mochis_p_val_table.csv"))
```

```
  ## Pick genes where one of the three pairs (3m, 18m, 24m)
  ## has significant p-value at FDR 0.05 control
  selected_genes_3_18 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MW_3_18,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MW_3_18")) # [!] Benjamini-Hochberg
  colnames(selected_genes_3_18) <- c("TRANSCRIPT", "NEW_MANN_WHITNEY")
  selected_genes_3_18$TISSUE <- rep(tissue, nrow(selected_genes_3_18))
  tissue_transcript_3_18 <- rbind(tissue_transcript_3_18, selected_genes_3_18)

  selected_genes_18_24 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MW_18_24,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MW_18_24"))
  colnames(selected_genes_18_24) <- c("TRANSCRIPT", "NEW_MANN_WHITNEY")
  selected_genes_18_24$TISSUE <- rep(tissue, nrow(selected_genes_18_24))
  tissue_transcript_18_24 <- rbind(tissue_transcript_18_24, selected_genes_18_24)

  selected_genes_24_3 <- tissue_mann_whitney_df[
    which(p.adjust(
      tissue_mann_whitney_df$MW_24_3,
      method = "BH") <= 0.05),] %>%
    select(c("TRANSCRIPT", "MW_24_3"))
  colnames(selected_genes_24_3) <- c("TRANSCRIPT", "NEW_MANN_WHITNEY")
  selected_genes_24_3$TISSUE <- rep(tissue, nrow(selected_genes_24_3))
  tissue_transcript_24_3 <- rbind(tissue_transcript_24_3, selected_genes_24_3)
}

## Compare against original MW significant genes
# Load library
library(VennDiagram)
library(RColorBrewer)

# Load original DEGs from Section 2.1
og_transcript_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_3m_18m.csv"))
og_transcript_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_18m_24m.csv"))
og_transcript_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_24m_3m.csv"))

# Compare 3m vs 18m
#print("Comparing DEGs between 3m and 18m...")
#og_tissue_transcripts <- og_transcript_3_18 %>% select(c("TRANSCRIPT","TISSUE"))
#new_tissue_transcripts <- tissue_transcript_3_18 %>% select(c("TRANSCRIPT","TISSUE"))
#print(paste0("No. genes in common = ",
#             nrow(dplyr::intersect(og_tissue_transcripts,new_tissue_transcripts))))
#print(paste0("No. genes lost after contamination = ",
#             nrow(og_tissue_transcripts) - nrow(dplyr::intersect(og_tissue_transcripts,new_tissue_tran
#print(paste0("No. new DEGs after contamination = ",
#             nrow(new_tissue_transcripts) - nrow(dplyr::intersect(og_tissue_transcripts,new_tissue_tra
set1 <- paste0(og_transcript_3_18$TRANSCRIPT,"_",og_transcript_3_18$TISSUE)
set2 <- paste0(tissue_transcript_3_18$TRANSCRIPT,"_",tissue_transcript_3_18$TISSUE)
```

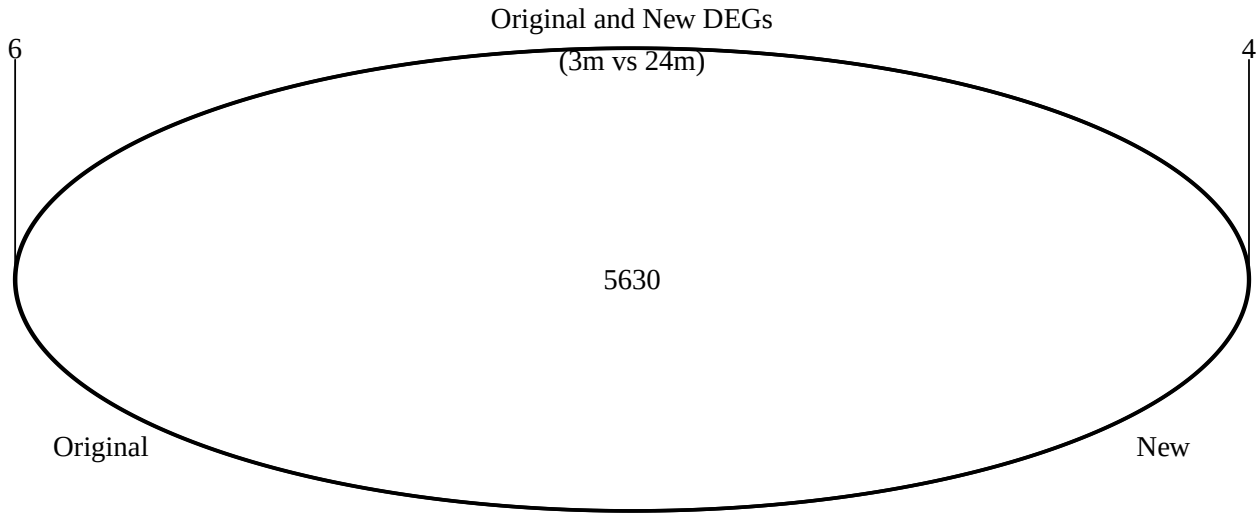
```
# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("Original" , "New"),
        main = "Original and New DEGs\n(3m vs 18m)",
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```




```
# Compare 18m vs 24m
set1 <- paste0(og_transcript_18_24$TRANSCRIPT,"_",og_transcript_18_24$TISSUE)
set2 <- paste0(tissue_transcript_18_24$TRANSCRIPT,"_",tissue_transcript_18_24$TISSUE)

# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("Original" , "New"),
        main = "Original and New DEGs\n(18m vs 24m)",
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```




```
# Compare 24m vs 3m
set1 <- paste0(og_transcript_24_3$TRANSCRIPT,"_",og_transcript_24_3$TISSUE)
set2 <- paste0(tissue_transcript_24_3$TRANSCRIPT,"_",tissue_transcript_24_3$TISSUE)

# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("Original" , "New"),
        main = "Original and New DEGs\n(3m vs 24m)",
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```



We see that there are very few original Mann-Whitney DEGs that are no longer significant after tie breaking, and conversely there are also very few new Mann-Whitney DEGs that were originally non-significant. This suggests that the tie-breaking procedure hardly affected the gene expression distributions between age groups.

3 Analysis

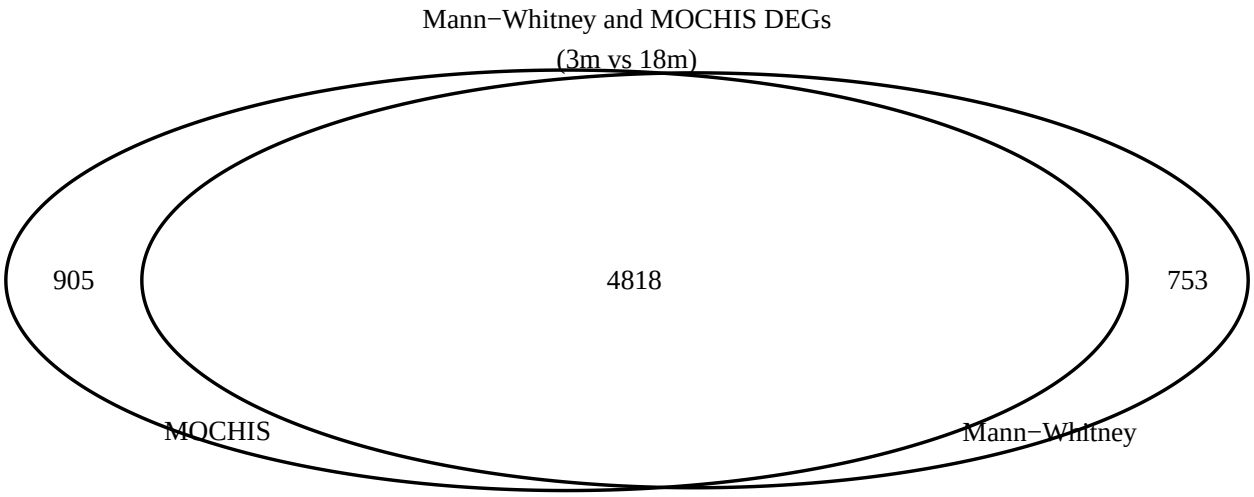
We examine more closely the differences between Mann-Whitney DEGs and MOCHIS DEGs. Recall that Mann-Whitney DEGs are genes that are typically picked up by standard differential analysis routines, whereas MOCHIS DEGs are genes that are differentially expressed owing to shifts in dispersion. Below, we perform some analyses to answer the following questions.

- How many MOCHIS DEGs were previously not detected by Mann-Whitney?
- Does MOCHIS really pick up shifts in dispersion?
- Are there other interesting questions we may answer with our newly detected MOCHIS DEGs?

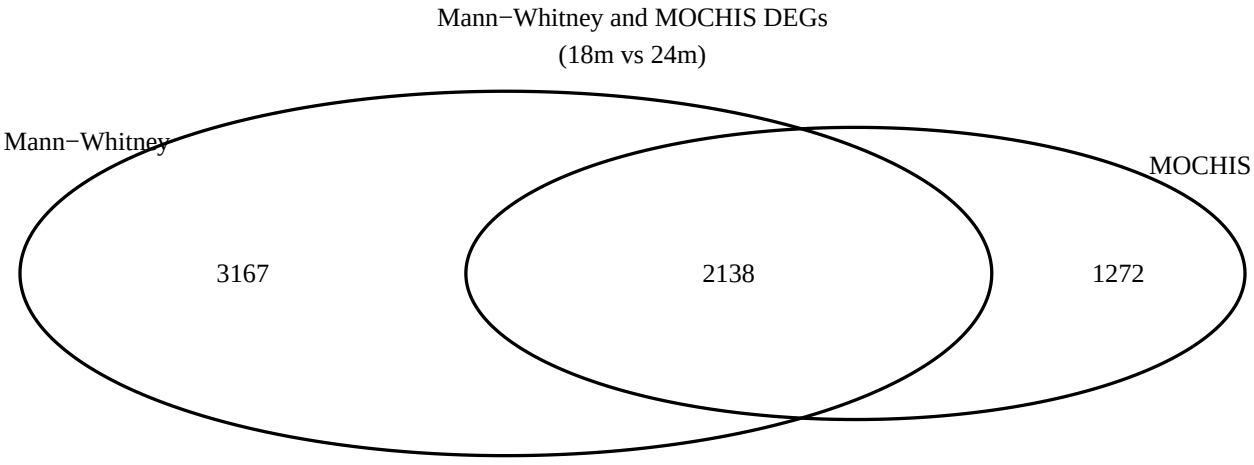
3.1 Counts Comparison

We draw Venn diagrams to illustrate the agreement between MOCHIS and Mann-Whitney significant genes.

```
## Compare counts
# Load original DEGs from Section 2.1
og_transcript_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_3m_18m.csv"))
og_transcript_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_18m_24m.csv"))
og_transcript_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mw_sig_24m_3m.csv"))

# Load MOCHIS DEGs from Section 2.2
tissue_transcript_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_3m_18m.csv"))
tissue_transcript_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_18m_24m.csv"))
tissue_transcript_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_24m_3m.csv"))

# Compare 3m vs 18m
set1 <- paste0(og_transcript_3_18$TRANSCRIPT,"_",og_transcript_3_18$TISSUE)
set2 <- paste0(tissue_transcript_3_18$TRANSCRIPT,"_",tissue_transcript_3_18$TISSUE)

# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("MOCHIS" , "Mann-Whitney"),
        main = "Mann-Whitney and MOCHIS DEGs\n(3m vs 18m)",
        #filename = paste0(tab_mur_dir,"3m_18m_mw_vs_mochis_venn_diagram.png"),
        #output=TRUE
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```

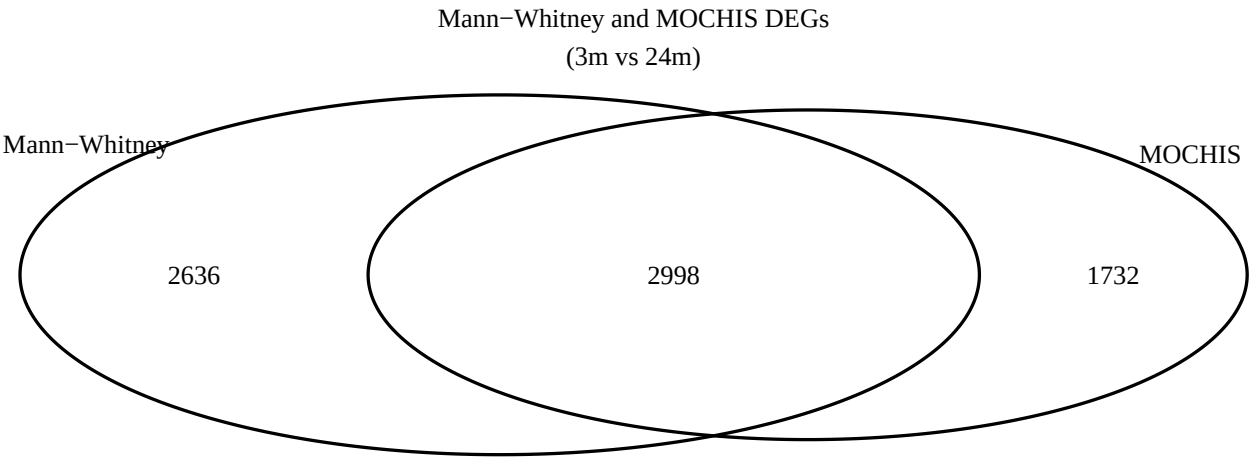



```
# Compare 18m vs 24m
set1 <- paste0(og_transcript_18_24$TRANSCRIPT,"_",og_transcript_18_24$TISSUE)
set2 <- paste0(tissue_transcript_18_24$TRANSCRIPT,"_",tissue_transcript_18_24$TISSUE)

# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("Mann-Whitney" , "MOCHIS"),
        main = "Mann-Whitney and MOCHIS DEGs\n(18m vs 24m)",
        #filename = paste0(tab_mur_dir,"18m_24m_mw_vs_mochis_venn_diagram.png"),
        #output=TRUE
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```




```
# Compare 24m vs 3m
set1 <- paste0(og_transcript_24_3$TRANSCRIPT,"_",og_transcript_24_3$TISSUE)
set2 <- paste0(tissue_transcript_24_3$TRANSCRIPT,"_",tissue_transcript_24_3$TISSUE)

# Chart
my.vd <- venn.diagram(
        x = list(set1, set2),
        category.names = c("Mann-Whitney" , "MOCHIS"),
```

```
        main = "Mann-Whitney and MOCHIS DEGs\n(3m vs 24m)",
        #filename = paste0(tab_mur_dir,"3m_24m_mw_vs_mochis_venn_diagram.png"),
        #output=TRUE
        filename = NULL
)
grid.newpage()
grid.draw(my.vd)
```



Summary of Findings. We can see from the Venn diagrams that

1. In general, there are considerable differences in the genes picked up by Mann-Whitney and MOCHIS. For any pair of age groups, MOCHIS picks up at least 900 DEGs that were not picked up by Mann-Whitney.
2. The number of new genes picked up by MOCHIS is the largest for the pair "3m vs 24m" (= 1732), and smallest for the pair "3m vs 18m" (= 905).
3. The number of Mann-Whitney significant genes that are not MOCHIS significant is greatest for the pair "18m vs 24m" (= 3167) and smallest for the pair "3m vs 18m" (= 753).

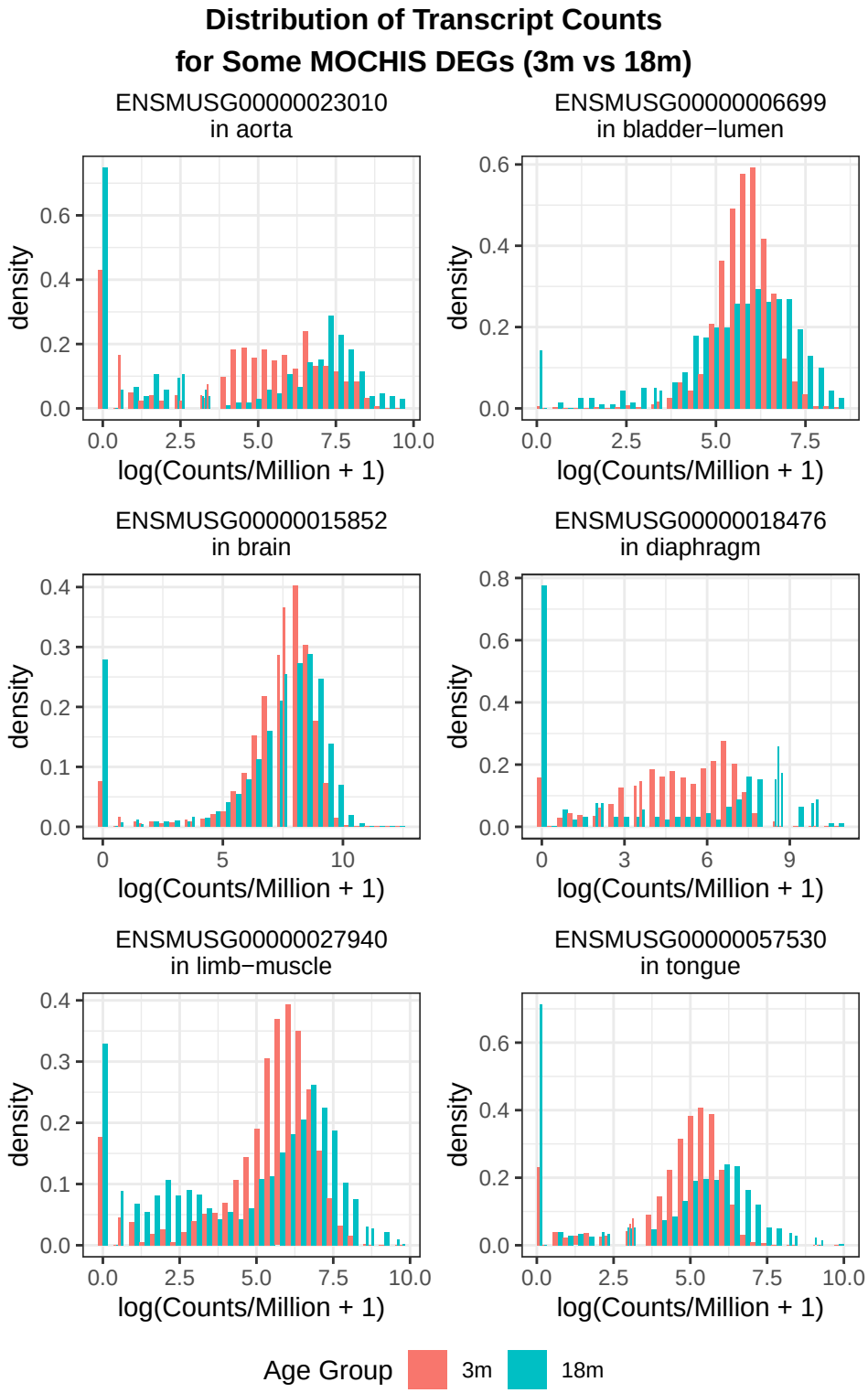
3.2 Visualizing Changes in Dispersion

The skeptical reader may wonder if MOCHIS is really picking up a shift in dispersion between the two age groups. Since we realistically cannot compare gene expression distributions between age groups for each MOCHIS significant gene, here we show some gene expression visualizations of MOCHIS significant genes. We focus on MOCHIS DEGs that were *not detected by Mann-Whitney*. We show visualizations for each pair of age groups ("3m vs 18m", "18m vs 24m" and "3m vs 24m").

3m vs 18m.

```
## Compare 3m vs 18m
# Identify those in MOCHIS but not Mann-Whitney
set1 <- paste0(og_transcript_3_18$TRANSCRIPT,"_",og_transcript_3_18$TISSUE)
set2 <- paste0(tissue_transcript_3_18$TRANSCRIPT,"_",tissue_transcript_3_18$TISSUE)
mochis_unique <- tissue_transcript_3_18[which(!(set2 %in% set1)),]
rownames(mochis_unique) <- 1:nrow(mochis_unique)

# View(mochis_unique %>%
#        group_by(TISSUE) %>%
#        mutate(
#          MinPValue = min(MOCHIS)) %>%
#        arrange(TISSUE) %>%
#        subset(MOCHIS == MinPValue))
```

```r
# Pick genes by hand (I choose the ones with smallest p-values in each tissue)
curated_degs_df <- mochis_unique[c(10,92,356,372,523,842),]
rownames(curated_degs_df) <- 1:nrow(curated_degs_df)

# Generate plots
plot_list <- list()
for (i in 1:nrow(curated_degs_df)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_degs_df[i,]$TISSUE
  transcript <- curated_degs_df[i,]$TRANSCRIPT

  # Open tissue-specific data and select gene
  tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))

  smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
  this_gene_exp_level <- data.frame(
    TRANSCRIPT = smartseq2_sparse_mat@counts[transcript,
                                             1:dim(smartseq2_sparse_mat@counts)[2]],
    AGE = tissue_smartseq2_data$age)

  # Separate out the 3m and 18m reads (counts/million)
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "18m")

  plot_list[[i]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue)) +
    xlab("log(Counts/Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
}

library(ggpubr)
combined_plot <- ggarrange(plot_list[[1]], plot_list[[2]],
          plot_list[[3]], plot_list[[4]],
          plot_list[[5]], plot_list[[6]],
          nrow = 3, ncol = 2,
          common.legend = TRUE, legend = "bottom")
annotate_figure(combined_plot,
                top = text_grob("Distribution of Transcript Counts\nfor Some MOCHIS DEGs (3m vs 18m)",
                                face = "bold", size = 12))
```

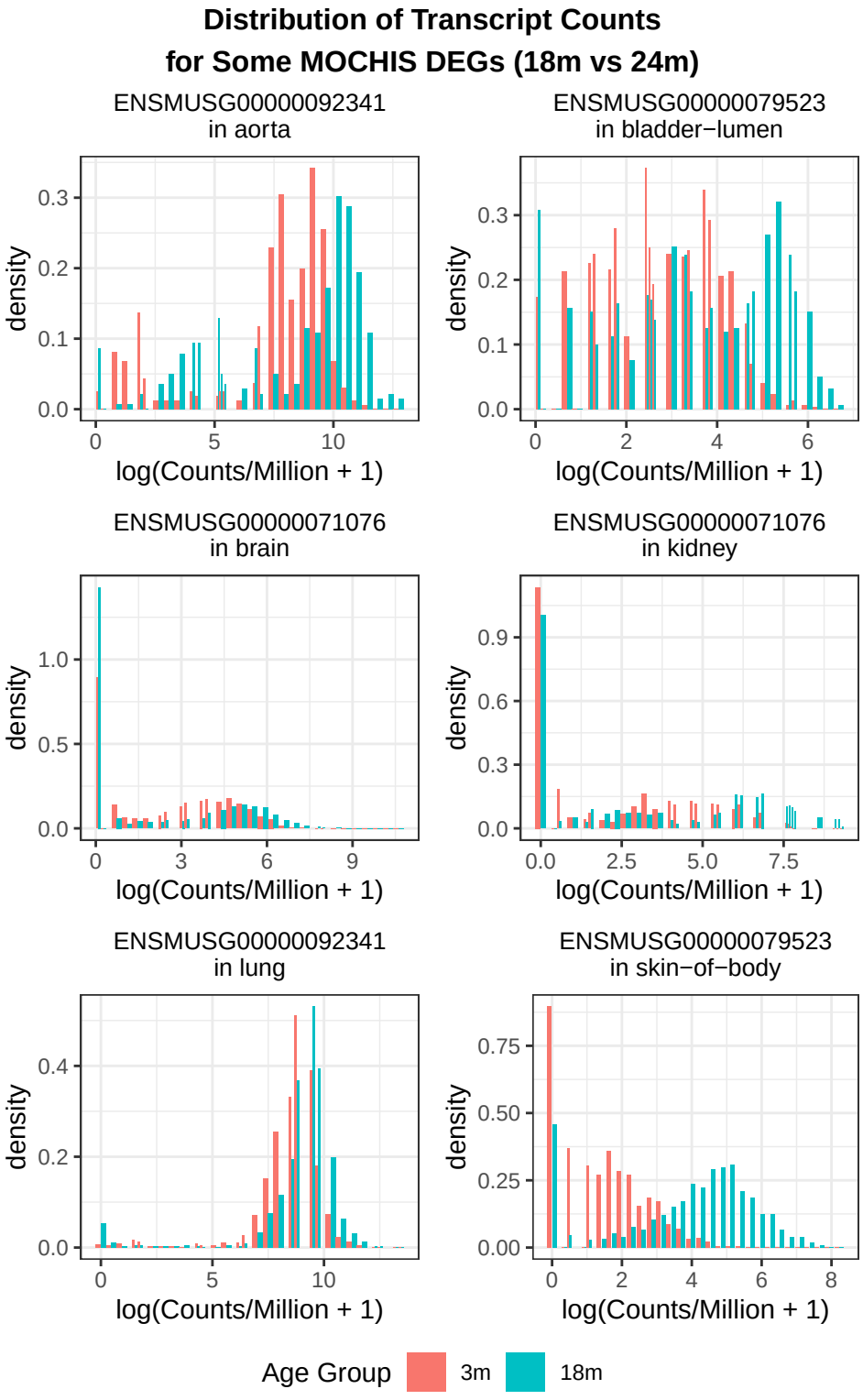


18m vs 24m.

```
## Compare 18m vs 24m
# Identify those in MOCHIS but not Mann-Whitney
set1 <- paste0(og_transcript_18_24$TRANSCRIPT,"_",og_transcript_18_24$TISSUE)
set2 <- paste0(tissue_transcript_18_24$TRANSCRIPT,"_",tissue_transcript_18_24$TISSUE)
```

```r
mochis_unique <- tissue_transcript_18_24[which(!(set2 %in% set1)),]
rownames(mochis_unique) <- 1:nrow(mochis_unique)

# View(mochis_unique %>%
#        group_by(TISSUE) %>%
#        mutate(
#          MinPValue = min(MOCHIS)) %>%
#        arrange(TISSUE) %>%
#        subset(MOCHIS == MinPValue))

# Pick genes by hand (I choose the ones with smallest p-values in each tissue)
curated_degs_df <- mochis_unique[c(88,101,179,541,896,1060),]
rownames(curated_degs_df) <- 1:nrow(curated_degs_df)

# Generate plots
plot_list <- list()
for (i in 1:nrow(curated_degs_df)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_degs_df[i,]$TISSUE
  transcript <- curated_degs_df[i,]$TRANSCRIPT

  # Open tissue-specific data and select gene
  tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))

  smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
  this_gene_exp_level <- data.frame(
    TRANSCRIPT = smartseq2_sparse_mat@counts[transcript,
                                             1:dim(smartseq2_sparse_mat@counts)[2]],
    AGE = tissue_smartseq2_data$age)

  # Separate out the 3m and 18m reads (counts/million)
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "18m")

  plot_list[[i]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue)) +
    xlab("log(Counts/Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
}

library(ggpubr)
combined_plot <- ggarrange(plot_list[[1]], plot_list[[2]],
          plot_list[[3]], plot_list[[4]],
          plot_list[[5]], plot_list[[6]],
          nrow = 3, ncol = 2,
          common.legend = TRUE, legend = "bottom")
```

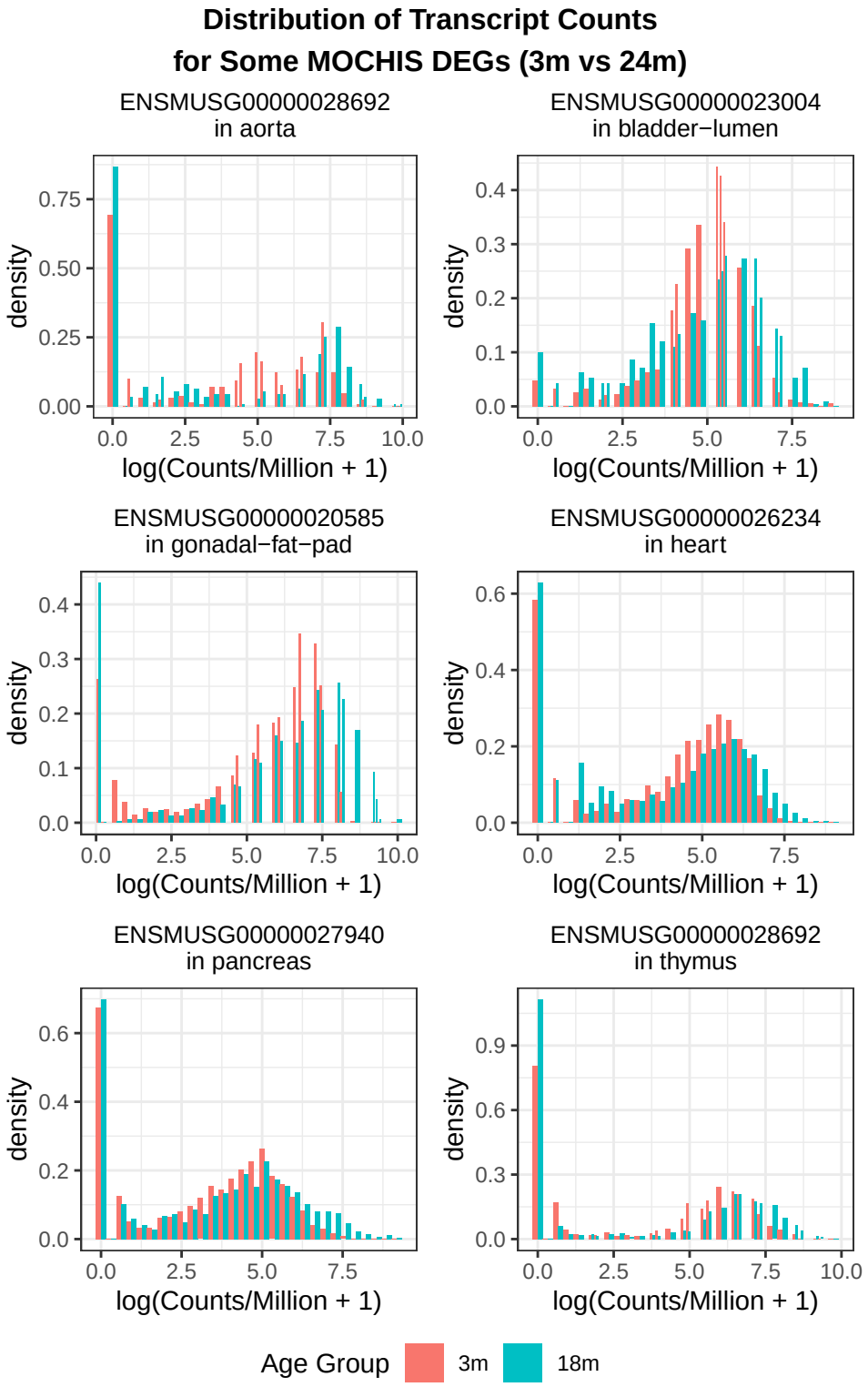
```
annotate_figure(combined_plot,
                top = text_grob("Distribution of Transcript Counts\nfor Some MOCHIS DEGs (18m vs 24m)",
                                face = "bold", size = 12))
```



3m vs 24m.

```r
## Compare 3m vs 24m
# Identify those in MOCHIS but not Mann-Whitney
set1 <- paste0(og_transcript_24_3$TRANSCRIPT,"_",og_transcript_24_3$TISSUE)
set2 <- paste0(tissue_transcript_24_3$TRANSCRIPT,"_",tissue_transcript_24_3$TISSUE)
mochis_unique <- tissue_transcript_24_3[which(!(set2 %in% set1)),]
rownames(mochis_unique) <- 1:nrow(mochis_unique)

# View(mochis_unique %>%
#        group_by(TISSUE) %>%
#        mutate(
#          MinPValue = min(MOCHIS)) %>%
#        arrange(TISSUE) %>%
#        subset(MOCHIS == MinPValue))

# Pick genes by hand (I choose the ones with smallest p-values in each tissue)
curated_degs_df <- mochis_unique[c(38,117,593,639,1412,1583),]
rownames(curated_degs_df) <- 1:nrow(curated_degs_df)

# Generate plots
plot_list <- list()
for (i in 1:nrow(curated_degs_df)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_degs_df[i,]$TISSUE
  transcript <- curated_degs_df[i,]$TRANSCRIPT

  # Open tissue-specific data and select gene
  tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))

  smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
  this_gene_exp_level <- data.frame(
    TRANSCRIPT = smartseq2_sparse_mat@counts[transcript,
                                             1:dim(smartseq2_sparse_mat@counts)[2]],
    AGE = tissue_smartseq2_data$age)

  # Separate out the 3m and 18m reads (counts/million)
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "18m")

  plot_list[[i]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue)) +
    xlab("log(Counts/Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
}

library(ggpubr)
combined_plot <- ggarrange(plot_list[[1]], plot_list[[2]],
```

```
          plot_list[[3]], plot_list[[4]],
          plot_list[[5]], plot_list[[6]],
          nrow = 3, ncol = 2,
          common.legend = TRUE, legend = "bottom")
annotate_figure(combined_plot,
                top = text_grob("Distribution of Transcript Counts\nfor Some MOCHIS DEGs (3m vs 24m)",
                                face = "bold", size = 12))
```



Summary of Findings. We find that

1. MOCHIS detects shifts in dispersions. These shifts can be in either direction (positive or negative).
2. Some of the shifts can be attributed to more pronounced zero inflation in one age group than another. This raises an important caveat in our analysis, namely, that our first step of filtering out genes that have more than 20% zero-inflation rate effectively removes all contribution by technical noise to the

data. If we are skeptical, then we must find other ways to effectively remove contribution by technical noise.

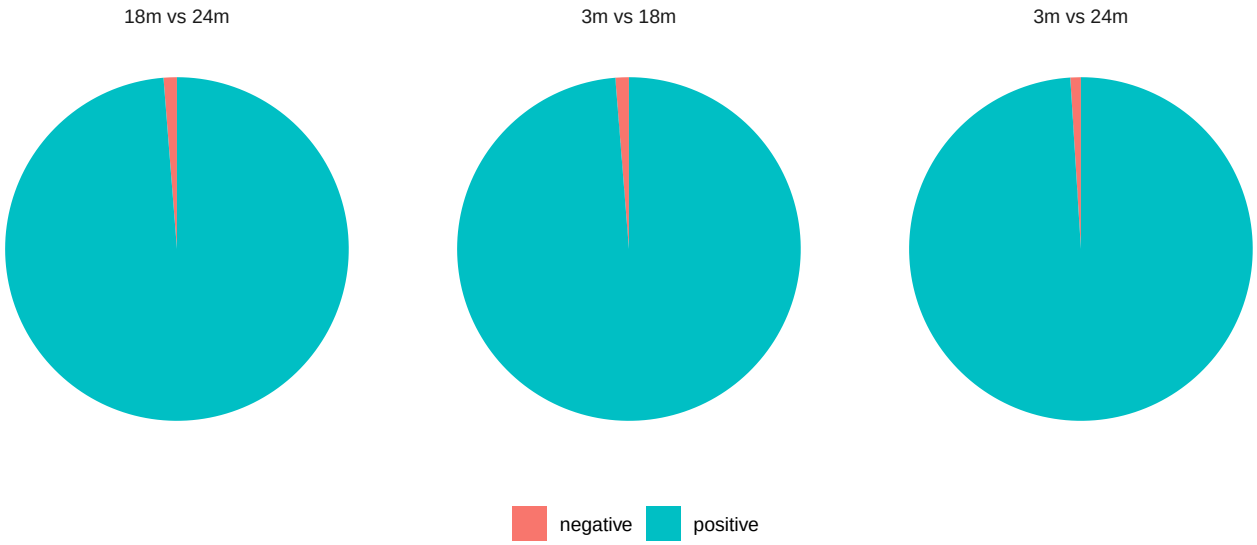
3.3 Other Interesting Results

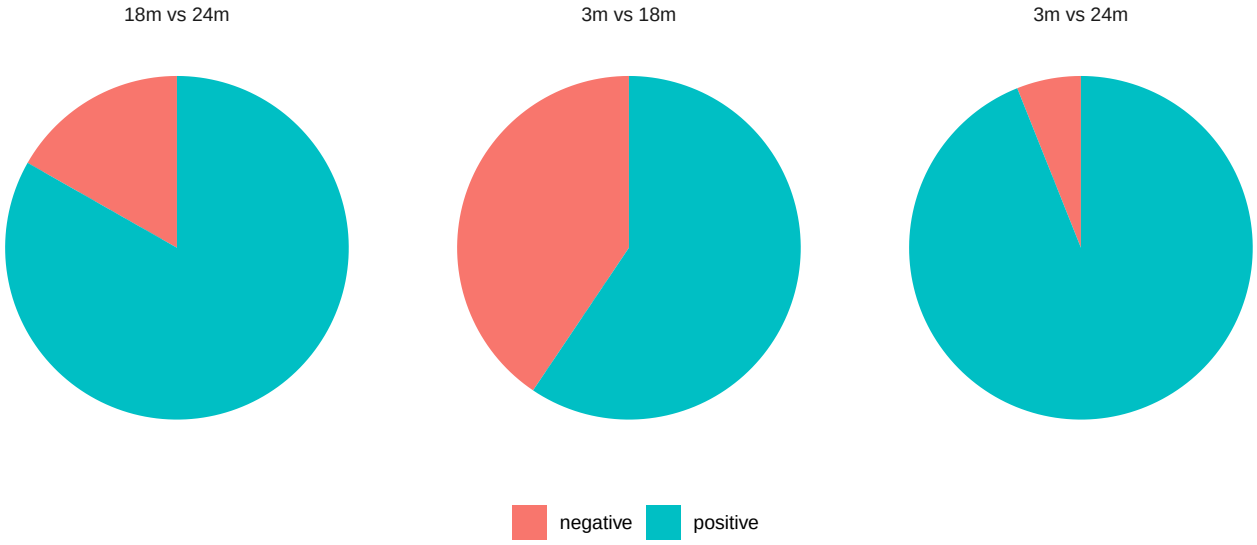
We show how we can further interpret our results to answer biologically meaningful questions.

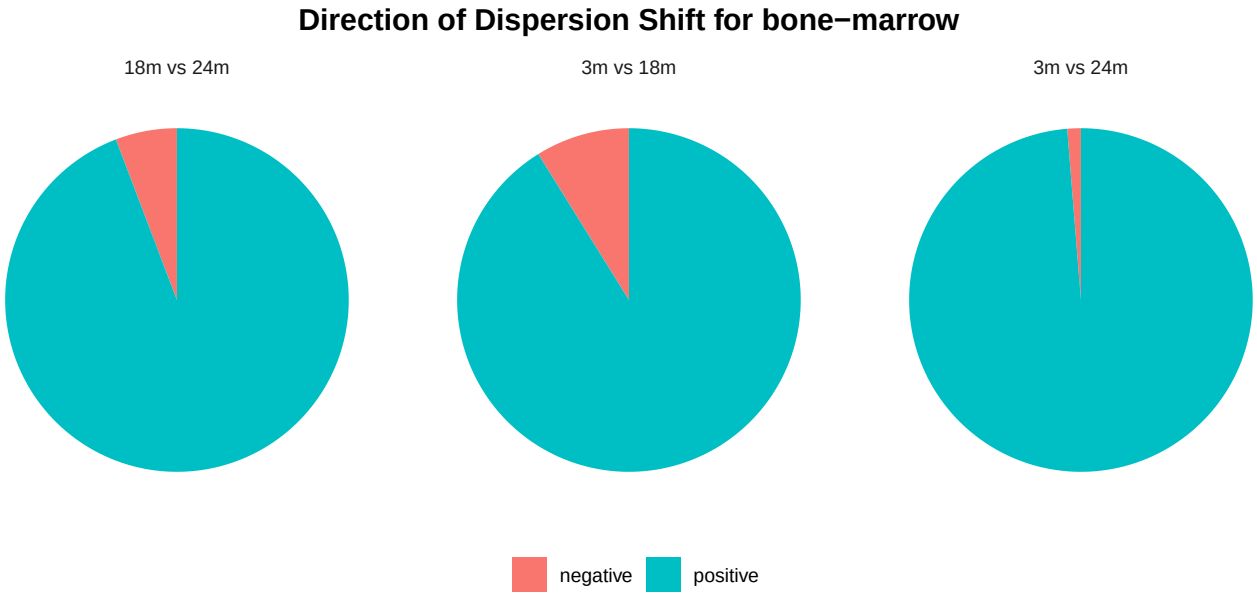
Gene Up-regulation vs Down-regulation. We have only looked at changes in dispersion, without explicitly tracking the directionality of change. We report, for each tissue and the corresponding pair of age groups, the fraction of positive and negative changes in dispersion, as measured by the ratio of variances.

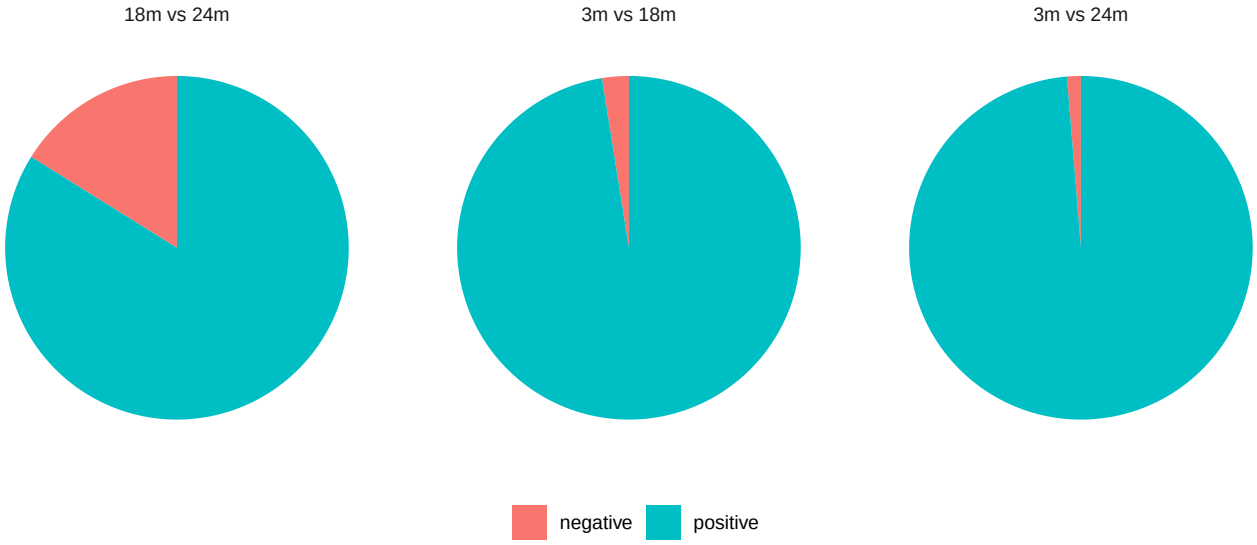
```
## Report tissue-specific distribution of up-regulated
## and down-regulated genes between age groups
# Define all_tissues again
all_tissues <- read.table(paste0(tab_mur_dir,"tissues/tissues.txt"))$V1

# Pie chart plotting hack (https://github.com/tidyverse/ggplot2/issues/2815)
cp <- coord_polar(theta = "y")
cp$is_free <- function() TRUE

mochis_degs_3_18 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_3m_18m.csv"))
mochis_degs_18_24 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_18m_24m.csv"))
mochis_degs_24_3 <- read.csv(paste0(tab_mur_dir,"tissues/mochis_sig_24m_3m.csv"))

up_down_reg_df <- data.frame(TISSUE = character(),
                             DOWN_3_18 = numeric(),
                             UP_3_18 = numeric(),
                             DOWN_18_24 = numeric(),
                             UP_18_24 = numeric(),
                             DOWN_3_24 = numeric(),
                             UP_3_24 = numeric())
for (tissue in all_tissues) {
  results_df <- read.csv(paste0(tab_mur_dir,"tissues/",tissue,"/mochis_p_val_table.csv"))

  # Analysis for 3m vs 18m
  n_3_18_down <- results_df %>%
    subset(TRANSCRIPT %in% (mochis_degs_3_18 %>% subset(TISSUE == tissue))$TRANSCRIPT) %>%
    select(INV_3_18) %>% sum()
  n_3_18_up <- nrow(results_df %>%
                      subset(TRANSCRIPT %in% (mochis_degs_3_18 %>%
                                                subset(TISSUE == tissue))$TRANSCRIPT)) -
    n_3_18_down

  # Analysis for 18m vs 24m
  n_18_24_down <- results_df %>%
    subset(TRANSCRIPT %in% (mochis_degs_18_24 %>% subset(TISSUE == tissue))$TRANSCRIPT) %>%
    select(INV_18_24) %>% sum()
  n_18_24_up <- nrow(results_df %>%
                       subset(TRANSCRIPT %in% (mochis_degs_18_24 %>%
                                                 subset(TISSUE == tissue))$TRANSCRIPT)) -
    n_18_24_down

  # Analysis for 3m vs 24m
  n_3_24_down <- results_df %>%
    subset(TRANSCRIPT %in% (mochis_degs_24_3 %>% subset(TISSUE == tissue))$TRANSCRIPT) %>%
```

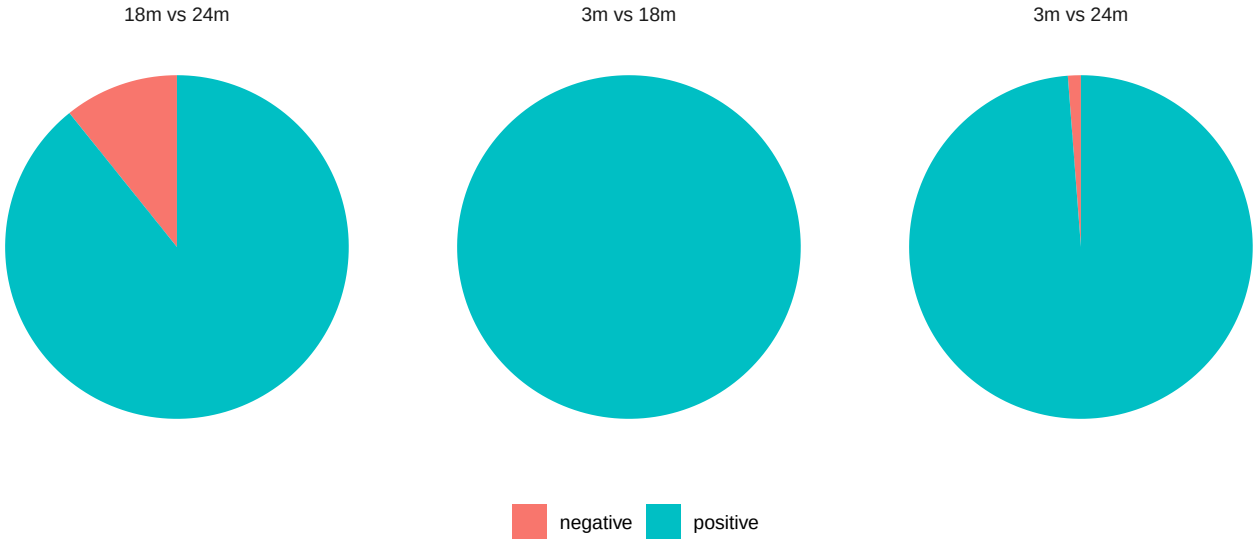
```
    select(INV_24_3) %>% sum()
  n_3_24_up <- nrow(results_df %>%
                      subset(TRANSCRIPT %in% (mochis_degs_24_3 %>%
                                                subset(TISSUE == tissue))$TRANSCRIPT)) -
    n_3_24_down

  up_down_reg_df <- rbind(up_down_reg_df,
                          data.frame(TISSUE = tissue,
                                     DOWN_3_18 = n_3_18_down,
                                     UP_3_18 = n_3_18_up,
                                     DOWN_18_24 = n_18_24_down,
                                     UP_18_24 = n_18_24_up,
                                     DOWN_3_24 = n_3_24_down,
                                     UP_3_24 = n_3_24_up))

  # Show pie charts for visualization
  plotting_df <- data.frame(PAIR = c(rep("3m vs 18m",2),
                                     rep("18m vs 24m",2),
                                     rep("3m vs 24m",2)),
                            GROUP = c(rep(c("negative", "positive"),3)),
                            FREQ = c(n_3_18_down,n_3_18_up,
                                     n_18_24_down,n_18_24_up,
                                     n_3_24_down, n_3_24_up))
  plot(ggplot(plotting_df, aes(x="", y=FREQ, fill=GROUP)) +
         geom_bar(width = 1, stat = "identity") +
         coord_polar("y") +
         cp +
         facet_wrap(.~PAIR, scales = "free") +
         theme_minimal()+
         theme(
           axis.title.x = element_blank(),
           axis.title.y = element_blank(),
           panel.border = element_blank(),
           panel.grid=element_blank(),
           axis.ticks = element_blank(),
           plot.title=element_text(hjust = 0.5, size=14, face="bold"),
           aspect.ratio = 1
         ) +
         ggtitle(paste0("Direction of Dispersion Shift for ", tissue)) +
         theme(axis.text.x=element_blank(),
               legend.title=element_blank(),
               legend.position = "bottom"))
}
```

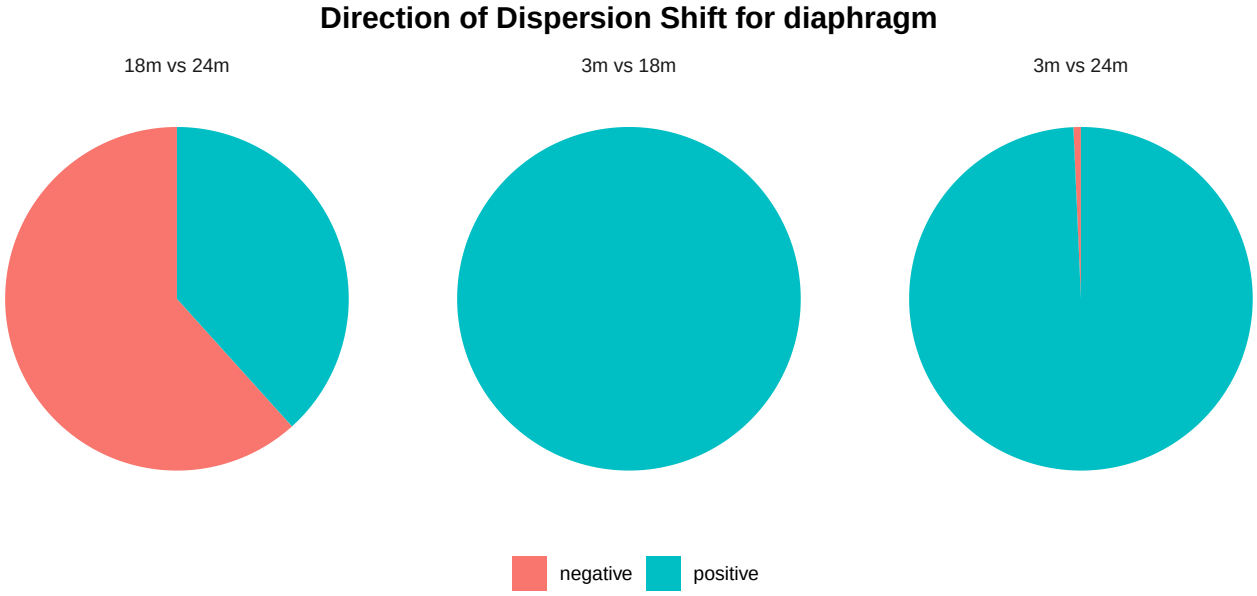

Direction of Dispersion Shift for aorta
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

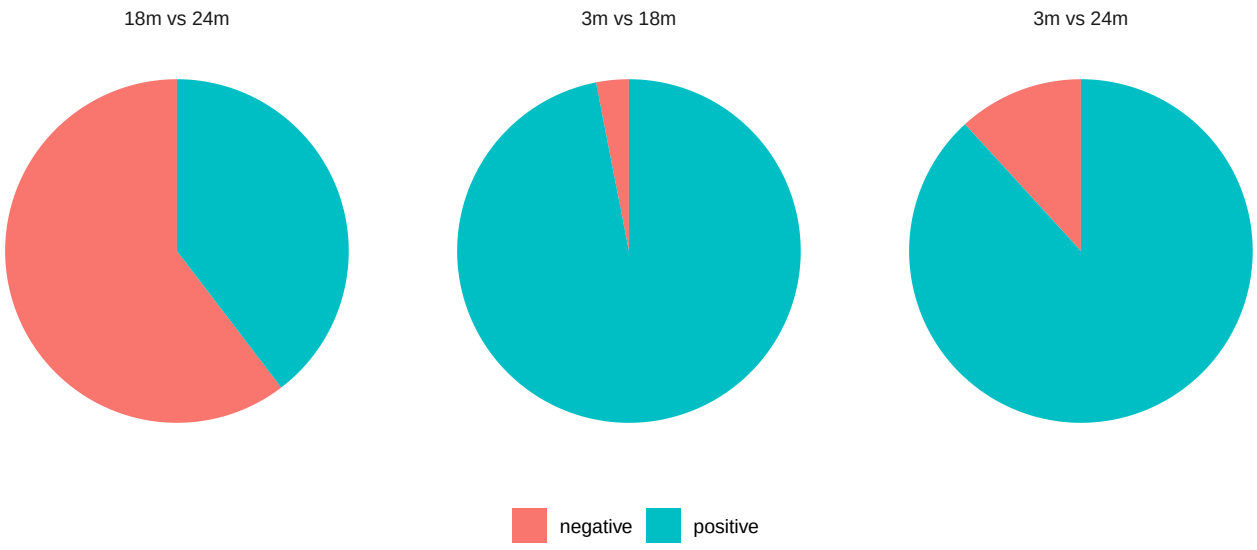

Direction of Dispersion Shift for bladder−lumen
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

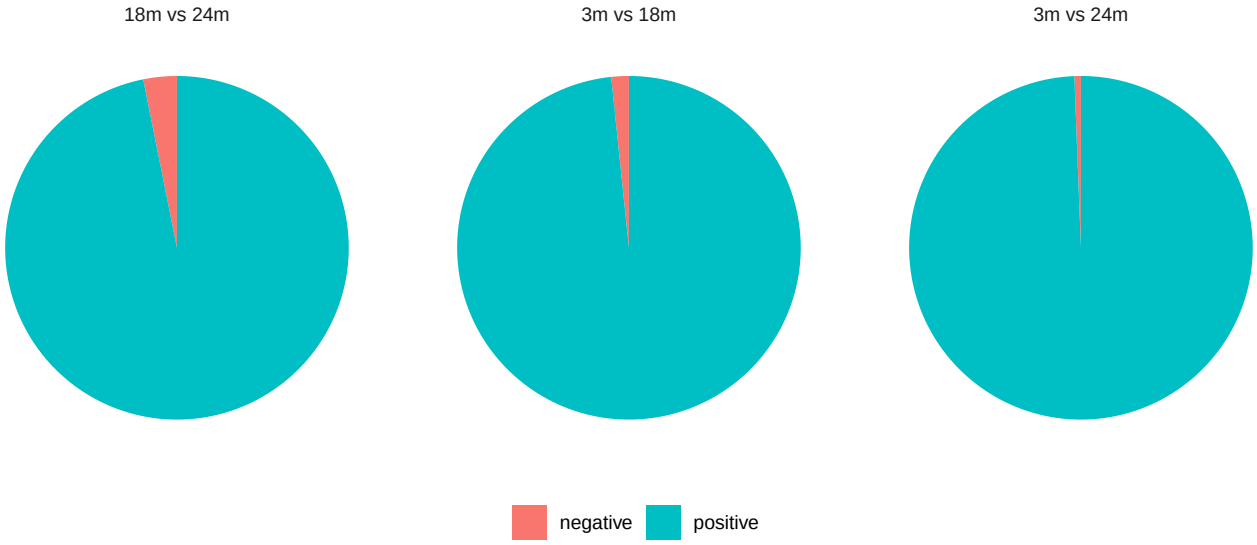

Direction of Dispersion Shift for bone−marrow
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

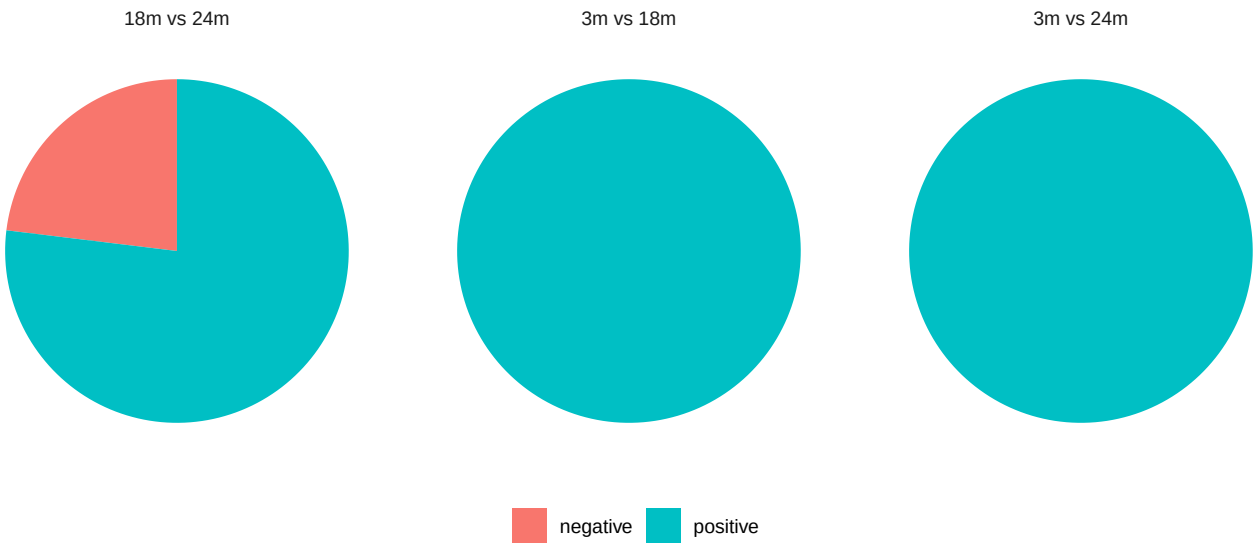

Direction of Dispersion Shift for brain
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

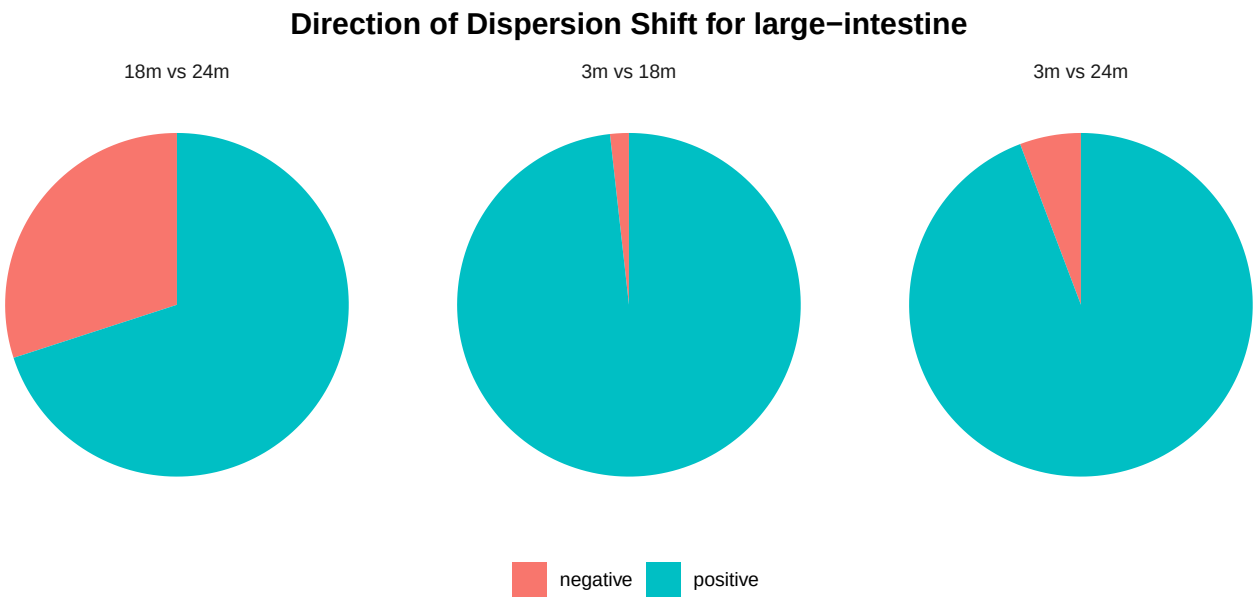

Direction of Dispersion Shift for brown−adipose−tissue
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

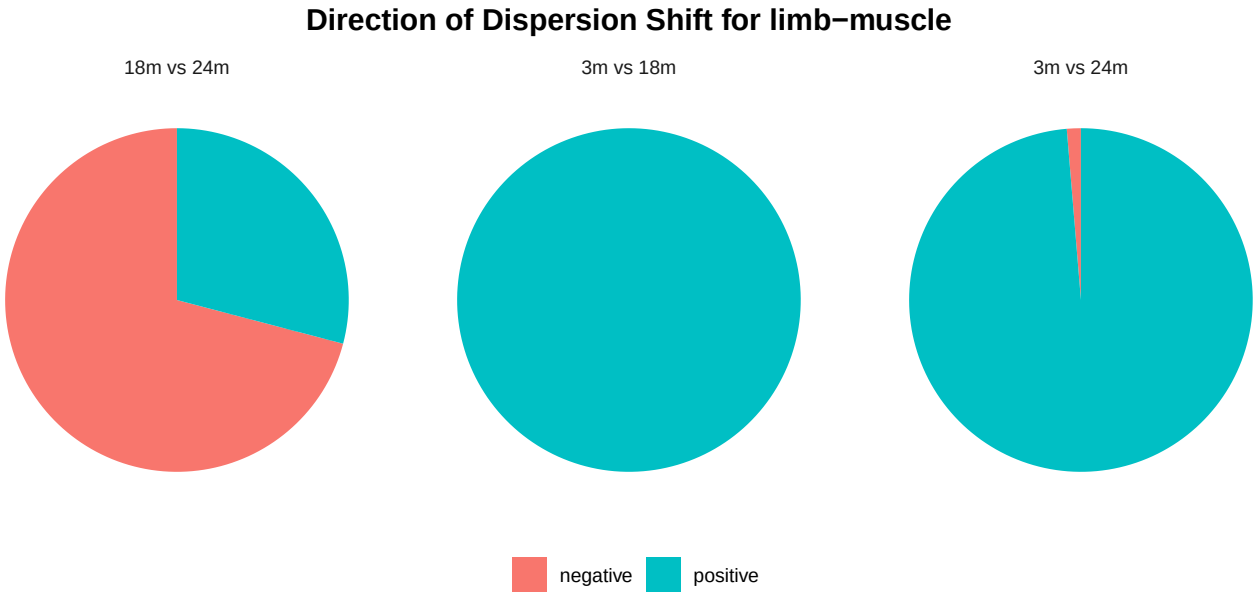

Direction of Dispersion Shift for diaphragm
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

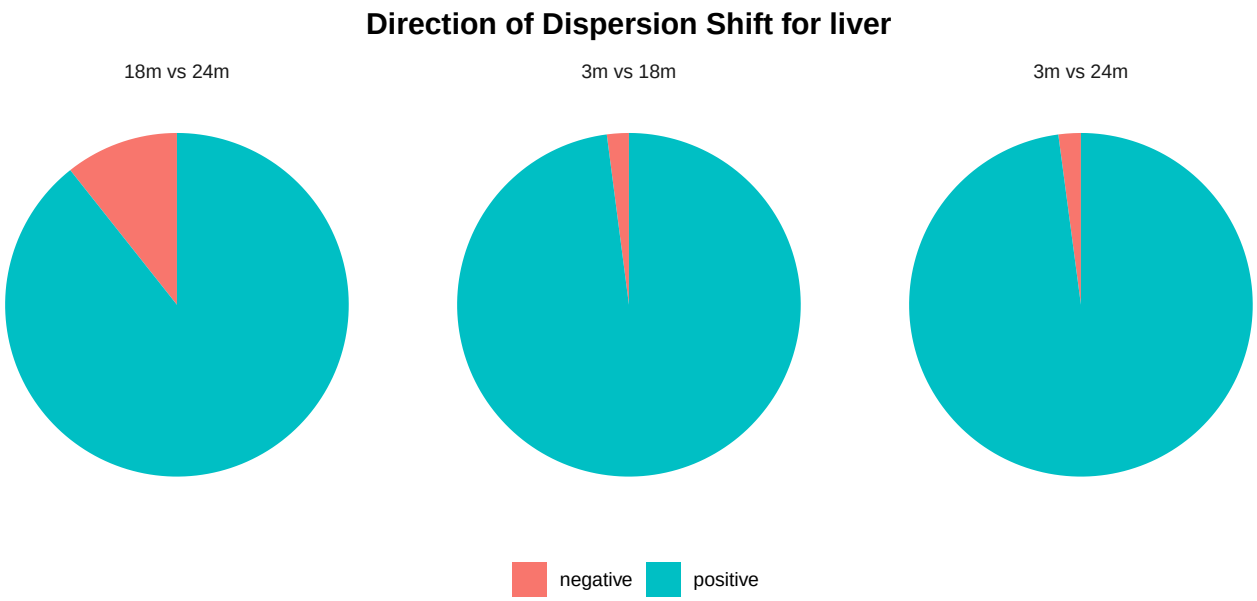

Direction of Dispersion Shift for gonadal−fat−pad
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

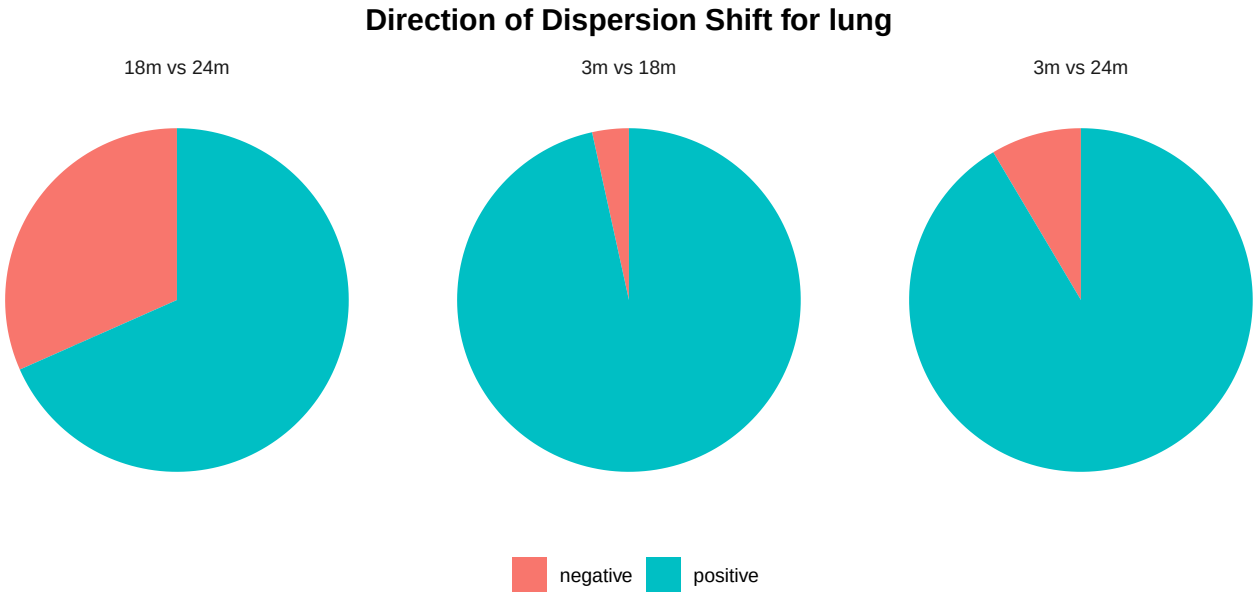

Direction of Dispersion Shift for heart
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

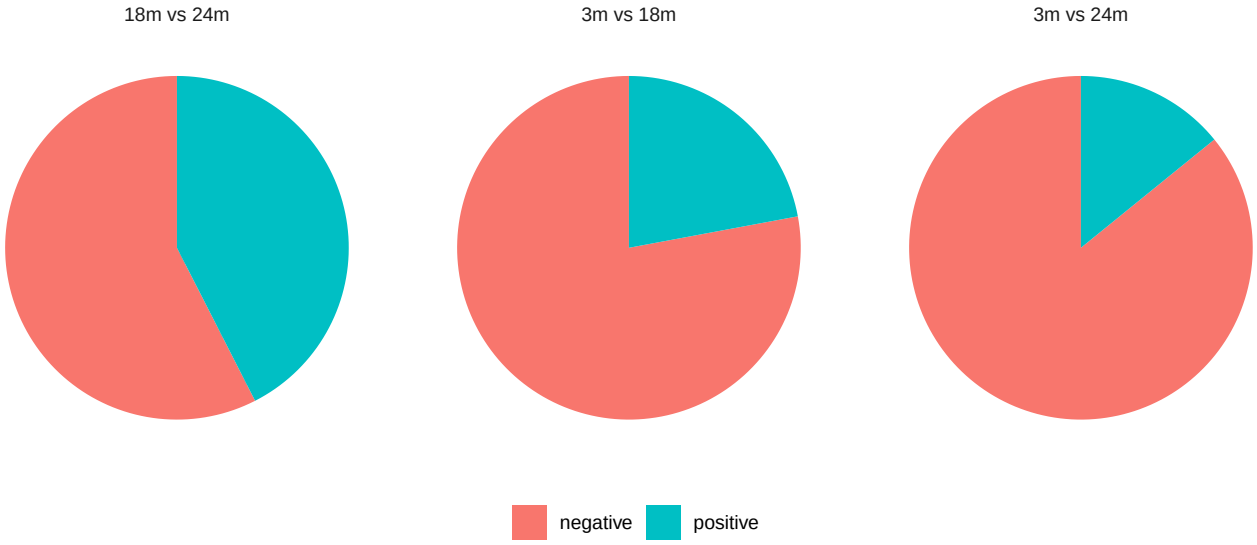

Direction of Dispersion Shift for kidney
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

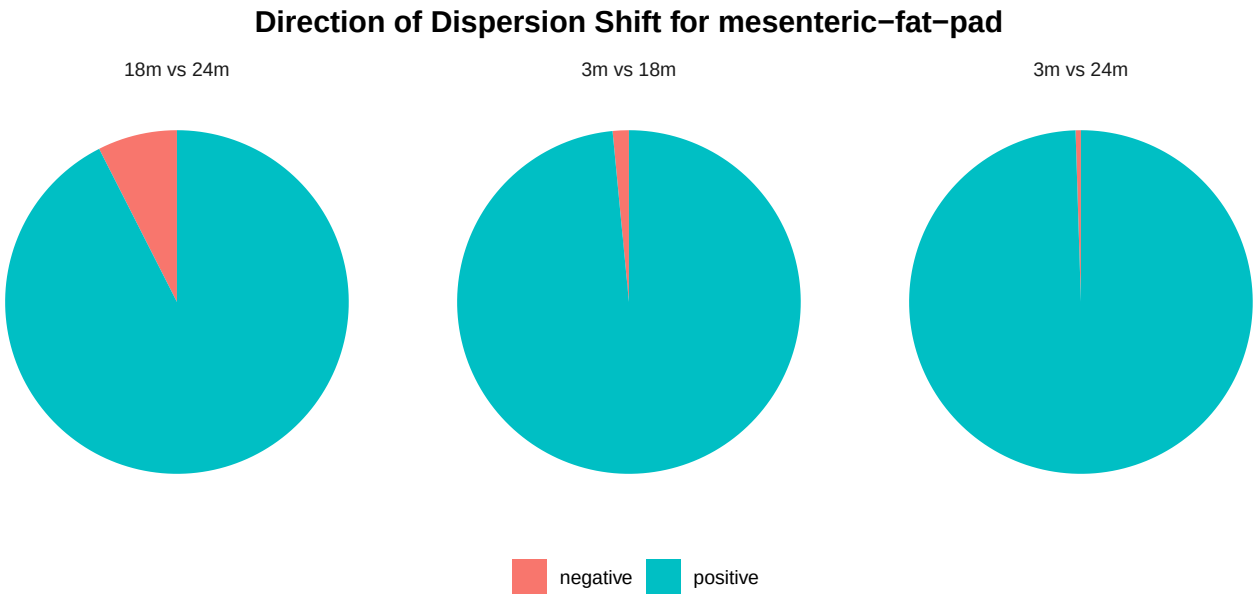

Direction of Dispersion Shift for large−intestine
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

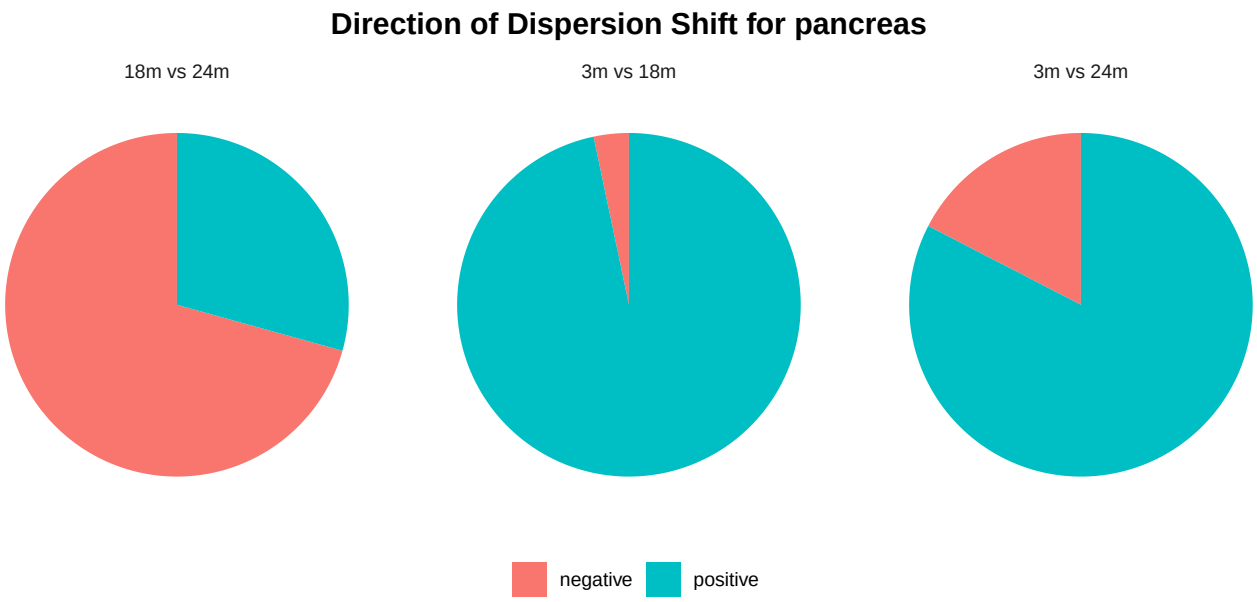

Direction of Dispersion Shift for limb−muscle
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

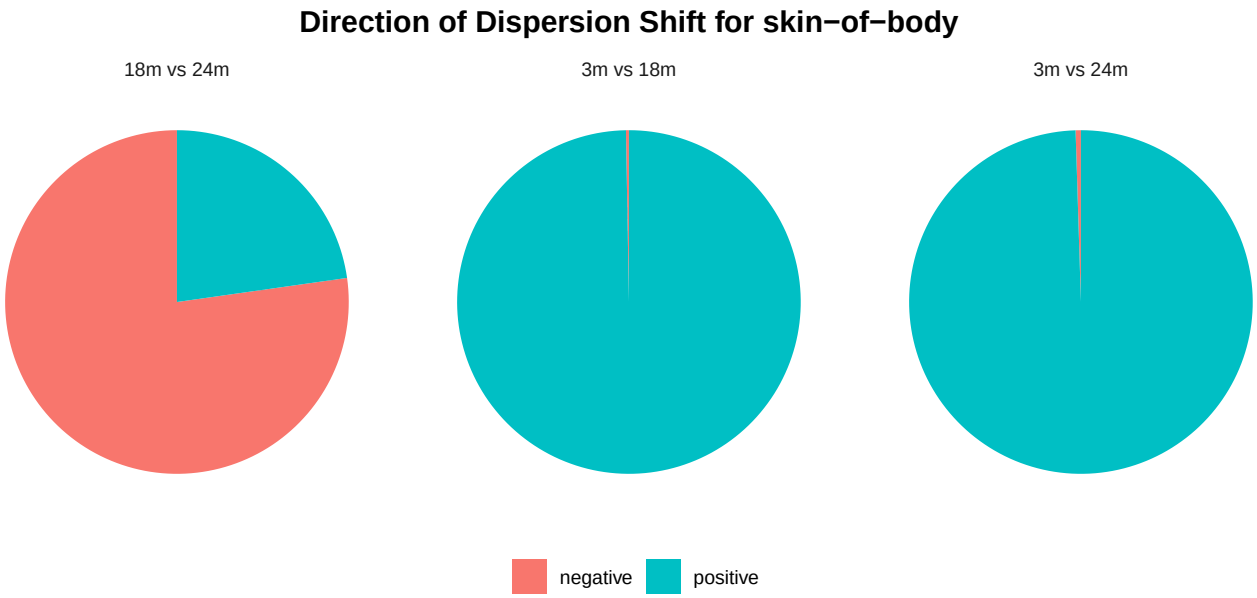

Direction of Dispersion Shift for liver
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

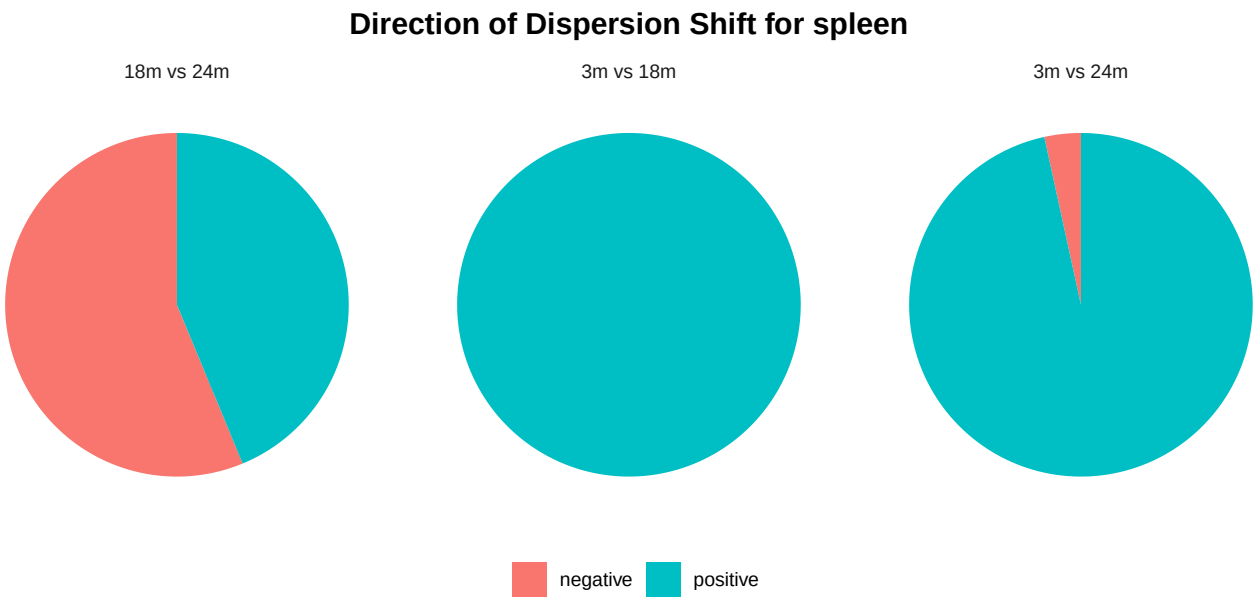

Direction of Dispersion Shift for lung
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

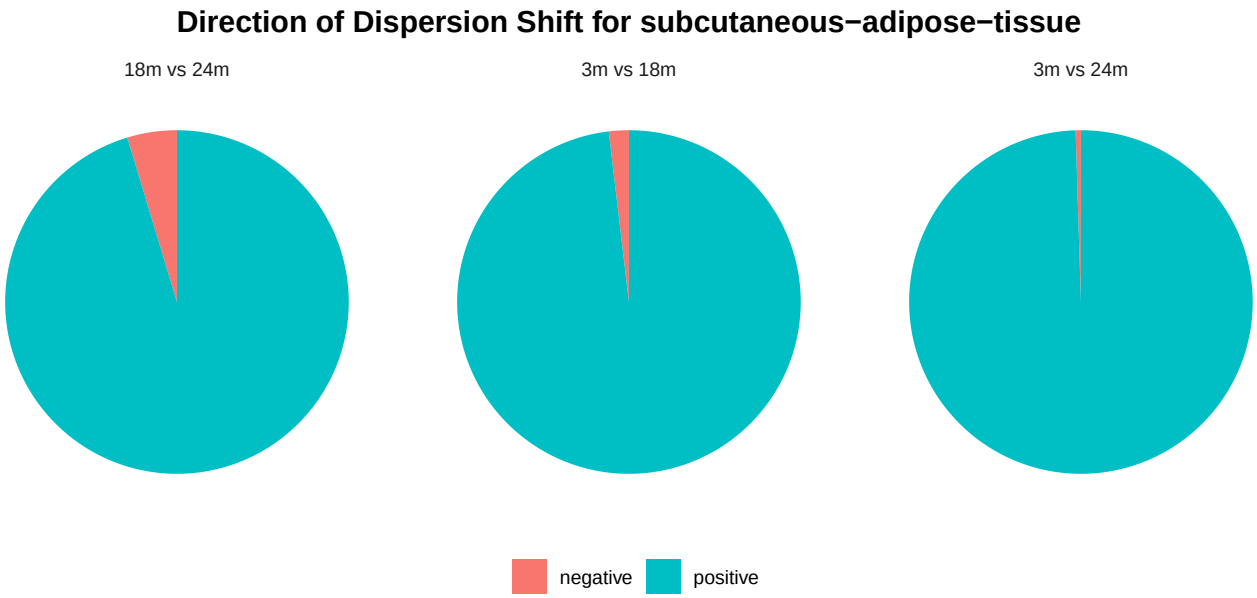

Direction of Dispersion Shift for mammary−gland
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

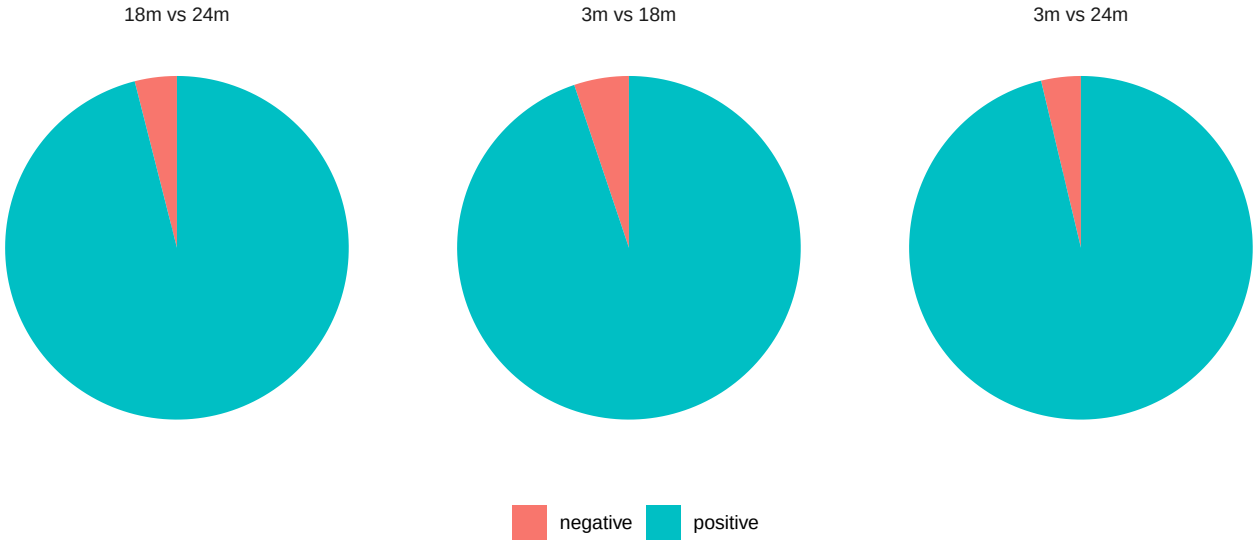

Direction of Dispersion Shift for mesenteric−fat−pad
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

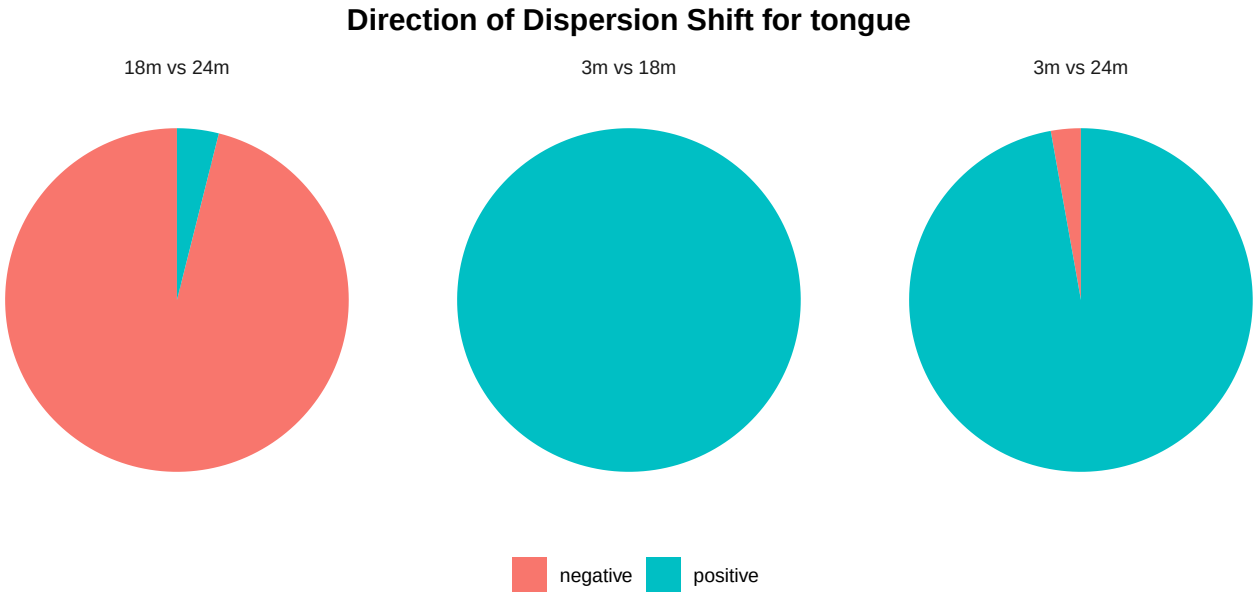

Direction of Dispersion Shift for pancreas
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive

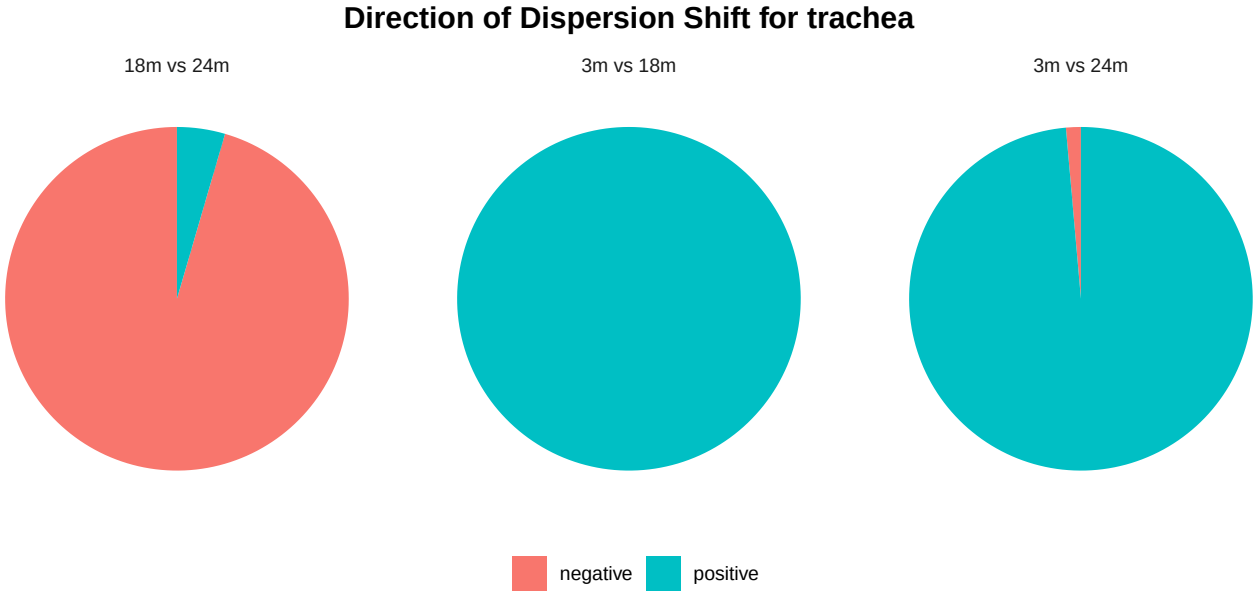

Direction of Dispersion Shift for skin−of−body
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive


Direction of Dispersion Shift for spleen
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive


Direction of Dispersion Shift for subcutaneous−adipose−tissue
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive


Direction of Dispersion Shift for thymus
18m vs 24m
3m vs 18m
3m vs 24m
negative
positive






```
# Print tissue-wide counts for dispersion shift directions
kableExtra::kable(up_down_reg_df) %>%
  kableExtra::kable_styling(position = "center",
                            latex_options="scale_down")
```

Given that a higher dispersion is indicative of increased gene regulation, we may further report which tissues tend to exhibit increased gene regulation between each age group. We do so below.

```
## Which tissues tend to exhibit increased gene regulation between each age group?
# We use 50% (majority rule) to decide whether a particular tissue 'tends' to exhibit
# increased gene regulation (i.e., positive change in dispersion)

# 3m vs 18m
print("The following tissues tend to exhibit increased gene regulation from 3m to 18m:")
```

```
## [1] "The following tissues tend to exhibit increased gene regulation from 3m to 18m:"
```

TISSUE	DOWN_3_18	UP_3_18	DOWN_18_24	UP_18_24	DOWN_3_24	UP_3_24
aorta	1	78	1	79	1	100
bladder-lumen	392	574	26	129	13	202
bone-marrow	17	175	7	114	2	155
brain	4	155	10	52	1	76
brown-adipose-tissue	0	65	7	58	1	81
diaphragm	0	161	66	41	1	142
gonadal-fat-pad	6	189	93	61	25	187
heart	3	179	4	123	1	162
kidney	0	65	12	40	0	61
large-intestine	7	389	42	98	11	180
limb-muscle	0	169	78	32	2	150
liver	2	95	8	67	2	93
lung	4	112	31	67	11	118
mammary-gland	392	111	88	65	182	30
mesenteric-fat-pad	3	194	11	136	1	199
pancreas	7	205	82	34	27	128
skin-of-body	1	362	112	33	1	200
spleen	0	155	63	49	5	141
subcutaneous-adipose-tissue	4	209	7	143	1	201
thymus	8	147	4	97	5	129
tongue	0	863	147	6	6	207
trachea	0	220	148	7	3	212

```
print(up_down_reg_df[
  which(up_down_reg_df$UP_3_18/(up_down_reg_df$UP_3_18 + up_down_reg_df$DOWN_3_18)>0.5),]$TISSUE)
```

```
##  [1] "aorta"                     "bladder-lumen"
##  [3] "bone-marrow"               "brain"
##  [5] "brown-adipose-tissue"      "diaphragm"
##  [7] "gonadal-fat-pad"           "heart"
##  [9] "kidney"                    "large-intestine"
## [11] "limb-muscle"               "liver"
## [13] "lung"                      "mesenteric-fat-pad"
## [15] "pancreas"                  "skin-of-body"
## [17] "spleen"                    "subcutaneous-adipose-tissue"
## [19] "thymus"                    "tongue"
## [21] "trachea"
```

```
# 18m vs 24m
print("The following tissues tend to exhibit increased gene regulation from 18m to 24m:")
```

```
## [1] "The following tissues tend to exhibit increased gene regulation from 18m to 24m:"
```

```
print(up_down_reg_df[
  which(up_down_reg_df$UP_18_24/(up_down_reg_df$UP_18_24 + up_down_reg_df$DOWN_18_24)>0.5),]$TISSUE)
```

```
##  [1] "aorta"                     "bladder-lumen"
##  [3] "bone-marrow"               "brain"
##  [5] "brown-adipose-tissue"      "heart"
##  [7] "kidney"                    "large-intestine"
##  [9] "liver"                     "lung"
## [11] "mesenteric-fat-pad"        "subcutaneous-adipose-tissue"
## [13] "thymus"
```

```
# 3m vs 24m
print("The following tissues tend to exhibit increased gene regulation from 3m to 24m:")
```

```
## [1] "The following tissues tend to exhibit increased gene regulation from 3m to 24m:"
```

```
print(up_down_reg_df[
  which(up_down_reg_df$UP_3_24/(up_down_reg_df$UP_3_24 + up_down_reg_df$DOWN_3_24)>0.5),]$TISSUE)
```

```
##  [1] "aorta"                      "bladder-lumen"
##  [3] "bone-marrow"                "brain"
##  [5] "brown-adipose-tissue"       "diaphragm"
##  [7] "gonadal-fat-pad"            "heart"
##  [9] "kidney"                     "large-intestine"
## [11] "limb-muscle"                "liver"
## [13] "lung"                       "mesenteric-fat-pad"
## [15] "pancreas"                   "skin-of-body"
## [17] "spleen"                     "subcutaneous-adipose-tissue"
## [19] "thymus"                     "tongue"
## [21] "trachea"
```

Persistently Differentially Expressed Genes. It is possible that for some tissues, gene regulation is so dynamic over the lifecourse, manifesting in detectable gene expression changes across time. To this end, we consider DEGs that are persistently differentiated; these are genes that are differentially expressed among *all* pairs of age groups. For brevity, we shall refer to them as *persistently DEGs*.

First, we ask how many such persistently DEGs there are.

```
## Identify persistently differentially expressed genes
# Find all unique tissue-transcript pairs in results
tissue_transcript_combined <- rbind(tissue_transcript_3_18 %>% select(c("TISSUE", "TRANSCRIPT")),
                                    tissue_transcript_18_24 %>% select(c("TISSUE", "TRANSCRIPT")),
                                    tissue_transcript_24_3 %>% select(c("TISSUE", "TRANSCRIPT"))) %>%
  distinct()


# Compute the persistent DEGs
age3m_vs_age18m <- c()
age18m_vs_age24m <- c()
age3m_vs_age24m <- c()

for (i in 1:nrow(tissue_transcript_combined)) {
  if (nrow(tissue_transcript_24_3 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age3m_vs_age24m <- c(age3m_vs_age24m, 1)
  } else {
    age3m_vs_age24m <- c(age3m_vs_age24m, 0)
  }
  if (nrow(tissue_transcript_3_18 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age3m_vs_age18m <- c(age3m_vs_age18m, 1)
  } else {
    age3m_vs_age18m <- c(age3m_vs_age18m, 0)
  }
  if (nrow(tissue_transcript_18_24 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age18m_vs_age24m <- c(age18m_vs_age24m, 1)
```

```
  } else {
    age18m_vs_age24m <- c(age18m_vs_age24m, 0)
  }
}

tissue_transcript_combined$`AGE3M_VS_AGE18M` <- age3m_vs_age18m
tissue_transcript_combined$`AGE18M_VS_AGE24M` <- age18m_vs_age24m
tissue_transcript_combined$`AGE3M_VS_AGE24M` <- age3m_vs_age24m
tissue_transcript_combined$PERSISTENCE <- (age3m_vs_age18m + age18m_vs_age24m + age3m_vs_age24m)

# How many persistently DEGs are there?
print(paste0("There are ",
             table(tissue_transcript_combined$PERSISTENCE)[3],
             " persistently DEGs."))
```

```
## [1] "There are 2174 persistently DEGs."
```

```
# Look closely at persistent DEGs
curated_persistent_degs <- tissue_transcript_combined %>% subset(PERSISTENCE == 3)
rownames(curated_persistent_degs) <- 1:nrow(curated_persistent_degs)
head(curated_persistent_degs)
```

```
##   TISSUE         TRANSCRIPT AGE3M_VS_AGE18M AGE18M_VS_AGE24M AGE3M_VS_AGE24M
## 1  aorta ENSMUSG00000030057               1                1               1
## 2  aorta ENSMUSG00000022982               1                1               1
## 3  aorta ENSMUSG00000060743               1                1               1
## 4  aorta ENSMUSG00000062825               1                1               1
## 5  aorta ENSMUSG00000044533               1                1               1
## 6  aorta ENSMUSG00000037563               1                1               1
##   PERSISTENCE
## 1           3
## 2           3
## 3           3
## 4           3
## 5           3
## 6           3
```

To underscore the utility of our method at picking up DEGs that would otherwise be overlooked, we look at persistently DEGs (if there are any!) that would not be picked up by Mann-Whitney.

```
## Find and visualize some persistently DEGs
## that are not Mann-Whitney significant
# Identify those in MOCHIS but not Mann-Whitney
set1 <- paste0(og_transcript_3_18$TRANSCRIPT,"_",og_transcript_3_18$TISSUE)
set2 <- paste0(tissue_transcript_3_18$TRANSCRIPT,"_",tissue_transcript_3_18$TISSUE)
mochis_unique_3_18 <- tissue_transcript_3_18[which(!(set2 %in% set1)),]
rownames(mochis_unique_3_18) <- 1:nrow(mochis_unique_3_18)

set1 <- paste0(og_transcript_18_24$TRANSCRIPT,"_",og_transcript_18_24$TISSUE)
set2 <- paste0(tissue_transcript_18_24$TRANSCRIPT,"_",tissue_transcript_18_24$TISSUE)
mochis_unique_18_24 <- tissue_transcript_18_24[which(!(set2 %in% set1)),]
rownames(mochis_unique_18_24) <- 1:nrow(mochis_unique_18_24)

set1 <- paste0(og_transcript_24_3$TRANSCRIPT,"_",og_transcript_24_3$TISSUE)
set2 <- paste0(tissue_transcript_24_3$TRANSCRIPT,"_",tissue_transcript_24_3$TISSUE)
```

```r
mochis_unique_24_3 <- tissue_transcript_24_3[which(!(set2 %in% set1)),]
rownames(mochis_unique_24_3) <- 1:nrow(mochis_unique_24_3)

# Find all unique tissue-transcript pairs in results
tissue_transcript_combined <- rbind(mochis_unique_3_18 %>% select(c("TISSUE", "TRANSCRIPT")),
                                    mochis_unique_18_24 %>% select(c("TISSUE", "TRANSCRIPT")),
                                    mochis_unique_24_3 %>% select(c("TISSUE", "TRANSCRIPT"))) %>%
  distinct()


# Compute the persistent DEGs
age3m_vs_age18m <- c()
age18m_vs_age24m <- c()
age3m_vs_age24m <- c()

for (i in 1:nrow(tissue_transcript_combined)) {
  if (nrow(mochis_unique_24_3 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age3m_vs_age24m <- c(age3m_vs_age24m, 1)
  } else {
    age3m_vs_age24m <- c(age3m_vs_age24m, 0)
  }
  if (nrow(mochis_unique_3_18 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age3m_vs_age18m <- c(age3m_vs_age18m, 1)
  } else {
    age3m_vs_age18m <- c(age3m_vs_age18m, 0)
  }
  if (nrow(mochis_unique_18_24 %>%
           subset(TISSUE == tissue_transcript_combined[i,]$TISSUE &
                  TRANSCRIPT == tissue_transcript_combined[i,]$TRANSCRIPT)) > 0) {
    age18m_vs_age24m <- c(age18m_vs_age24m, 1)
  } else {
    age18m_vs_age24m <- c(age18m_vs_age24m, 0)
  }
}

tissue_transcript_combined$`AGE3M_VS_AGE18M` <- age3m_vs_age18m
tissue_transcript_combined$`AGE18M_VS_AGE24M` <- age18m_vs_age24m
tissue_transcript_combined$`AGE3M_VS_AGE24M` <- age3m_vs_age24m
tissue_transcript_combined$PERSISTENCE <- (age3m_vs_age18m + age18m_vs_age24m + age3m_vs_age24m)

# How many persistently DEGs are there?
print(paste0("There are ",
             table(tissue_transcript_combined$PERSISTENCE)[3],
             " persistently DEGs not previously detected by Mann-Whitney."))
```

```
## [1] "There are 41 persistently DEGs not previously detected by Mann-Whitney."
```

```r
# Look closely at persistently DEGs
curated_persistent_degs <- tissue_transcript_combined %>% subset(PERSISTENCE == 3)
rownames(curated_persistent_degs) <- 1:nrow(curated_persistent_degs)
```

```r
# Distribution of persistently DEGs
table(curated_persistent_degs$TISSUE)
```

```
##
##      bladder-lumen        bone-marrow          diaphragm              heart
##                  1                  1                  2                  9
##    large-intestine        limb-muscle              liver      mammary-gland
##                  2                  2                  1                  2
## mesenteric-fat-pad       skin-of-body             thymus            trachea
##                  5                  2                  4                 10
```

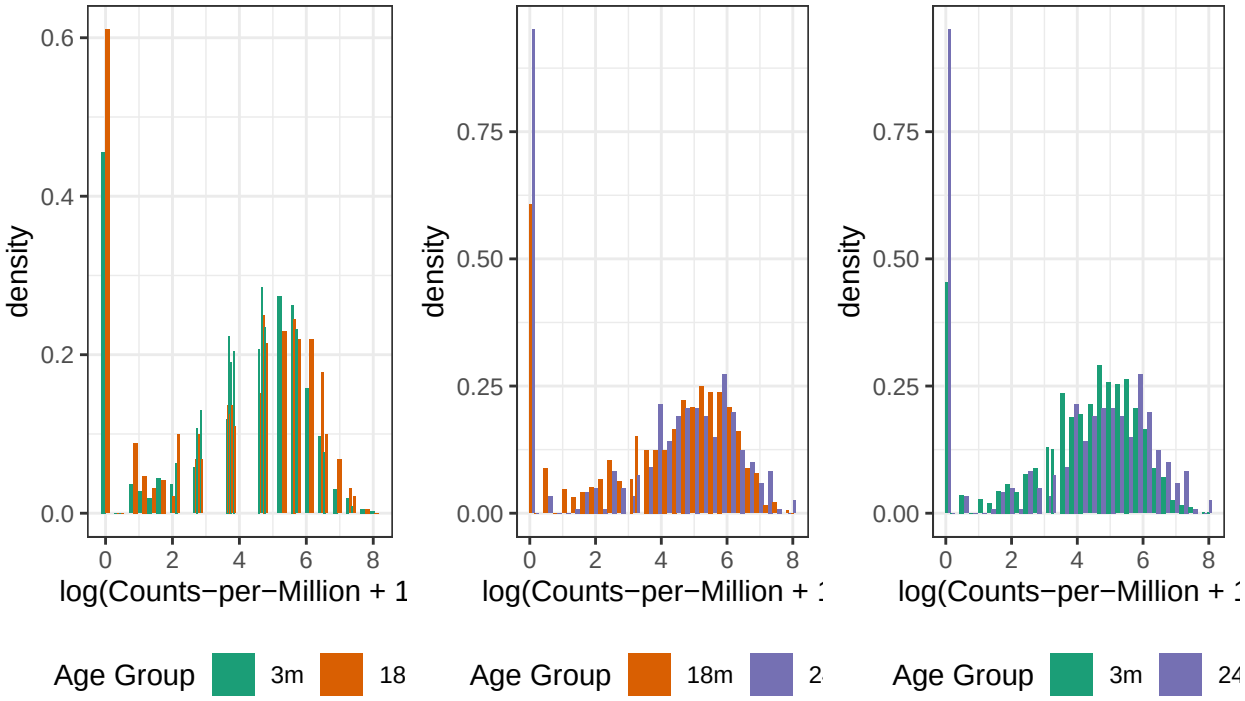
From the chunk above, we find that 41 tissue-specific genes are persistently DEGs. Moreover, most of them come from the trachea and the heart. Let us visualize some of these genes.

```r
## Visualizing MOCHIS-exclusive persistently DEGs
# Generate plots of gene expressions for persistent DEGs
plot_list <- list()
for (i in c(1,2,3,5,14,17,28,32)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_persistent_degs[i,]$TISSUE
  transcript <- curated_persistent_degs[i,]$TRANSCRIPT

  # Create local list
  this_deg_plot_list <- list()
  # Open tissue-specific data and select gene
  tissue_smartseq2_data <- readRDS(paste0(tab_mur_dir, "tissues/",tissue,"/local.rds"))

  smartseq2_sparse_mat <- tissue_smartseq2_data@assays$RNA
  this_gene_exp_level <- data.frame(
    TRANSCRIPT = smartseq2_sparse_mat@counts[transcript, 1:dim(smartseq2_sparse_mat@counts)[2]],
    AGE = tissue_smartseq2_data$age)

  # Separate out the 3m and 18m reads (counts/million)
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "18m")

  this_deg_plot_list[[1]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue)) +
    xlab("log(Counts-per-Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    scale_fill_manual(values = c("#1b9e77", "#d95f02")) +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))

  # Separate out the 18m and 24m reads (counts/million)
  # [!] Recall that mammary-gland transcripts only have 3m, 18m and 21m age groups
  if (tissue == "mammary-gland") {
    to_plot <- this_gene_exp_level %>% subset(AGE == "18m" | AGE == "21m")
```

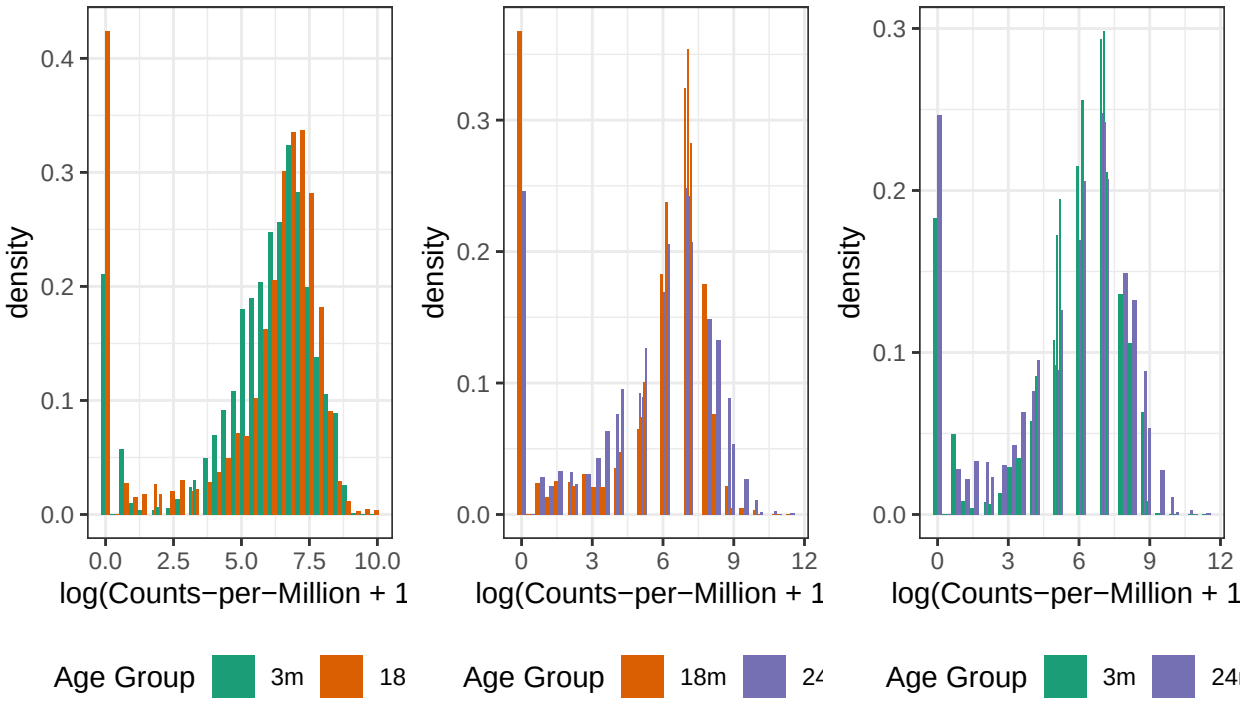
```
  this_deg_plot_list[[2]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue, "*")) +
    xlab("log(Counts-per-Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    scale_fill_manual(values = c("#d95f02", "#7570b3")) +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
} else {
  to_plot <- this_gene_exp_level %>% subset(AGE == "18m" | AGE == "24m")

  this_deg_plot_list[[2]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue)) +
    xlab("log(Counts-per-Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    scale_fill_manual(values = c("#d95f02", "#7570b3")) +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
}

# Separate out the 3m and 24m reads (counts/million)
# [!] Recall that mammary-gland transcripts only have 3m, 18m and 21m age groups
if (tissue == "mammary-gland") {
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "21m")

  this_deg_plot_list[[3]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
    geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                   position = position_dodge2()) +
    #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
    theme_bw() +
    ggtitle(paste0(transcript,"\nin ", tissue, "*")) +
    xlab("log(Counts-per-Million + 1)") +
    guides(fill=guide_legend("Age Group"))  +
    scale_fill_manual(values = c("#1b9e77", "#7570b3")) +
    theme(legend.position = "none",
          plot.title = element_text(hjust = 0.5,
                                    face = "plain",
                                    size = 10))
} else {
  to_plot <- this_gene_exp_level %>% subset(AGE == "3m" | AGE == "24m")

  this_deg_plot_list[[3]] <- ggplot(to_plot, aes(x = log(TRANSCRIPT+1))) +
```

```
      geom_histogram(aes(y = ..density.., fill = factor(AGE)),
                     position = position_dodge2()) +
      #geom_density(aes(fill = factor(AGE)), alpha = 0.5) +
      theme_bw() +
      ggtitle(paste0(transcript,"\nin ", tissue)) +
      xlab("log(Counts-per-Million + 1)") +
      guides(fill=guide_legend("Age Group"))  +
      scale_fill_manual(values = c("#1b9e77", "#7570b3")) +
      theme(legend.position = "none",
            plot.title = element_text(hjust = 0.5,
                                      face = "plain",
                                      size = 10))
  }

  plot_list[[i]] <- this_deg_plot_list
}

## Generate all pairwise plots for each persistent DEG
for (i in c(1,2,3,5,14,17,28,32)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_persistent_degs[i,]$TISSUE
  transcript <- curated_persistent_degs[i,]$TRANSCRIPT

  # Generate combined plot
  combined_plot <- ggarrange(plot_list[[i]][[1]] + ggtitle(""),
                             plot_list[[i]][[2]] + ggtitle(""),
                             plot_list[[i]][[3]] + ggtitle(""),
                             nrow = 1, ncol = 3,
                             legend = "bottom")
  # Annotate plot
  annotate_figure(combined_plot,
                  top = text_grob(paste0("Distribution of Transcript Counts\n(",
                                         transcript, " in ",  tissue, ")"),
                                  face = "bold", size = 12)) %>% plot()
}
```

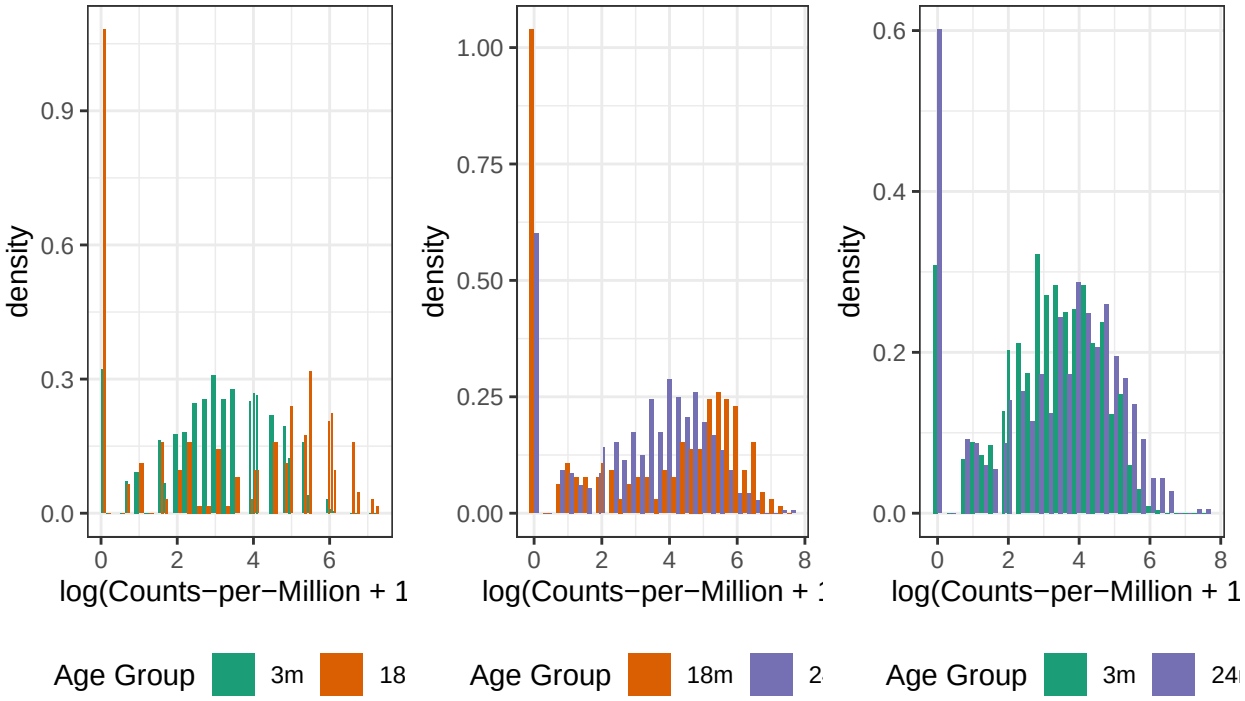
Distribution of Transcript Counts
(ENSMUSG00000041959 in bladder−lumen)



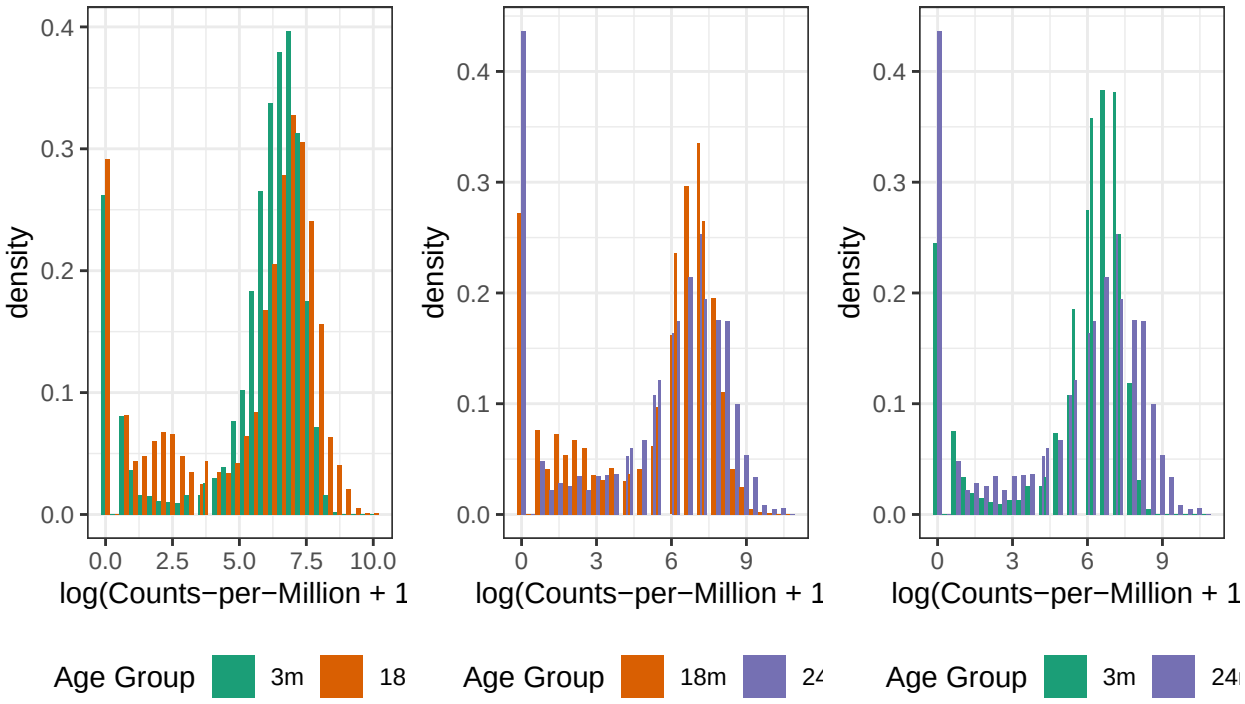
Distribution of Transcript Counts
(ENSMUSG00000008668 in bone−marrow)



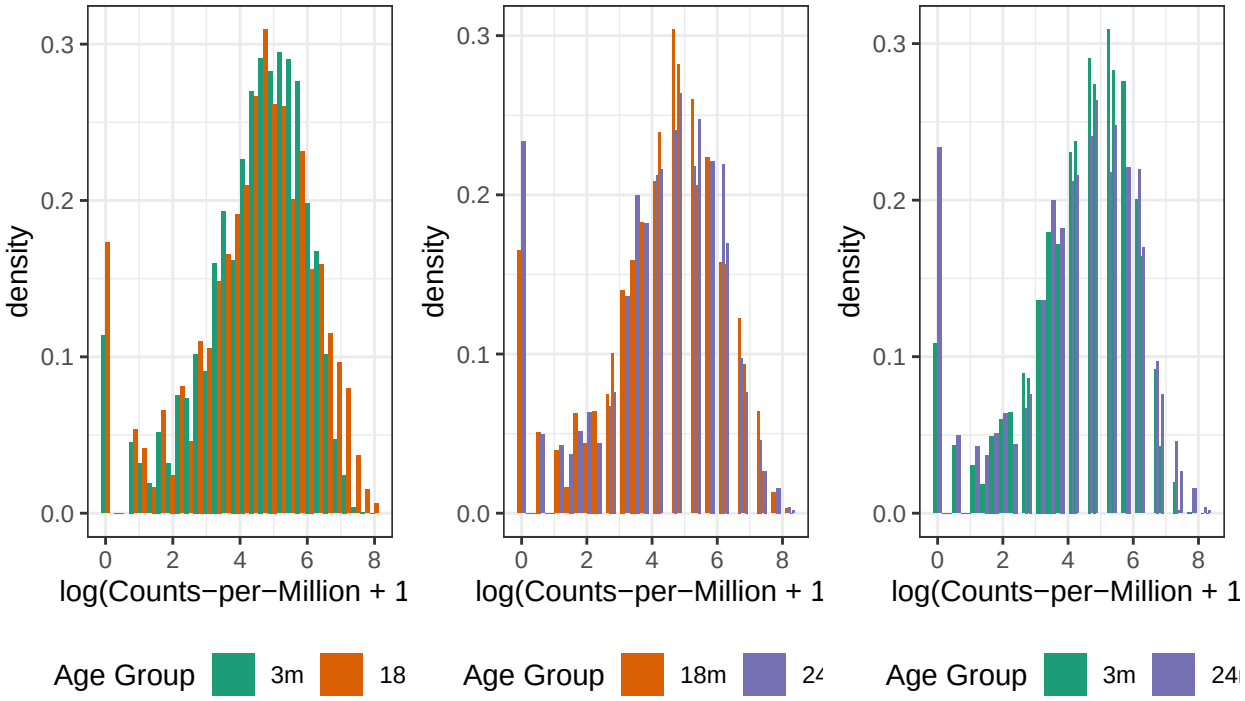
Distribution of Transcript Counts
(ENSMUSG00000057322 in diaphragm)



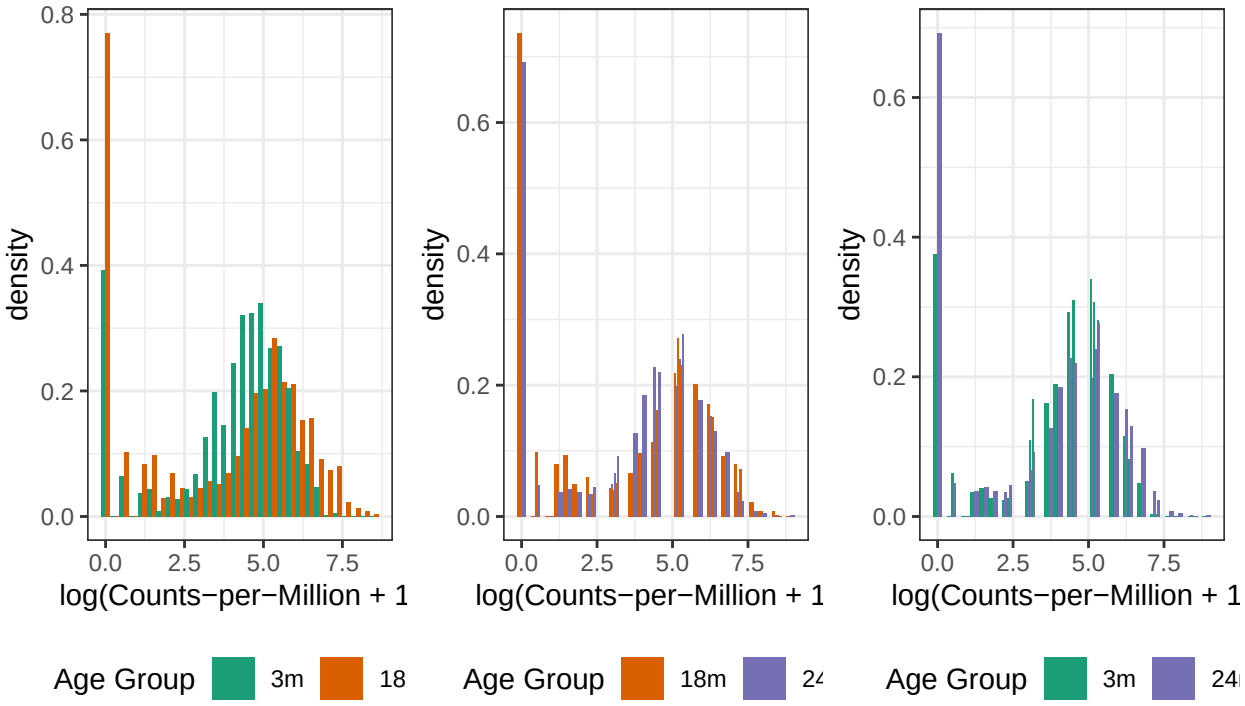
Distribution of Transcript Counts
(ENSMUSG00000030057 in heart)

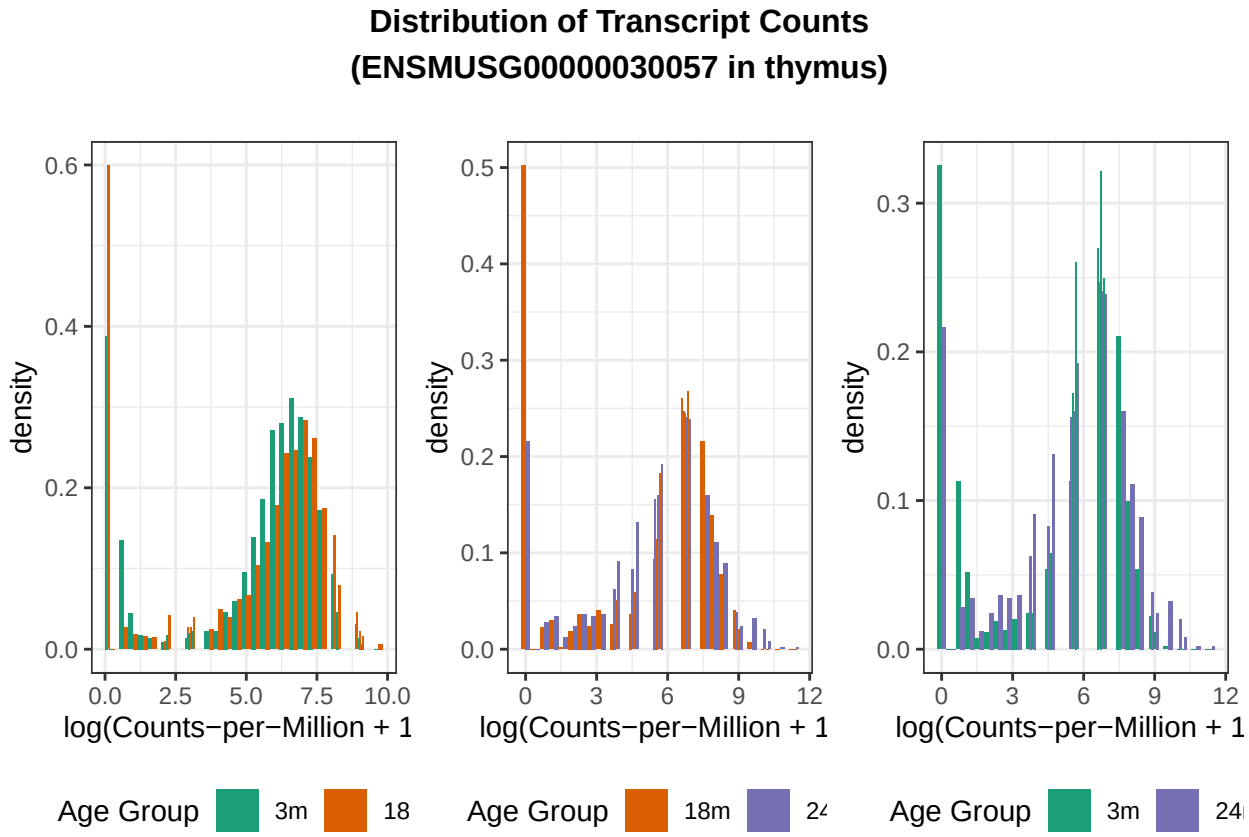


Distribution of Transcript Counts
(ENSMUSG00000008348 in large−intestine)



Distribution of Transcript Counts
(ENSMUSG00000038900 in limb−muscle)





Additionally, let us investigate the direction of dispersion changes for the 41 DEGs.

```r
## Generate dataframe summarizing direction of dispersion changes
persistent_degs_direction_df <- data.frame(TISSUE = character(),
                                           TRANSCRIPT = character(),
                                           AGE3M_VS_AGE18M = character(),
                                           AGE18M_VS_AGE24M = character(),
                                           AGE3M_VS_AGE24M = character())

for (i in 1:nrow(curated_persistent_degs)) {
  # Get raw read counts data for that tissue and transcript
  tissue <- curated_persistent_degs[i,]$TISSUE
  transcript <- curated_persistent_degs[i,]$TRANSCRIPT

  results_df <- read.csv(paste0(tab_mur_dir,"tissues/",tissue,"/mochis_p_val_table.csv"))

  relevant_row <- results_df %>% subset(TRANSCRIPT == transcript)
  persistent_degs_direction_df <- rbind(persistent_degs_direction_df,
                                        data.frame(TISSUE = tissue,
                                                   TRANSCRIPT = transcript,
                                                   AGE3M_VS_AGE18M = ifelse(relevant_row$INV_3_18, "-",
                                                   AGE18M_VS_AGE24M = ifelse(relevant_row$INV_18_24, "-
                                                   AGE3M_VS_AGE24M = ifelse(relevant_row$INV_24_3, "-",
}

# Print tissue-wide counts for dispersion shift directions
kableExtra::kable(persistent_degs_direction_df) %>%
  kableExtra::kable_styling(position = "center",
                            latex_options="scale_down")
```

Summary of Findings. When investigating the direction of change in dispersion, we find that

1. Except for mammary gland, all tissues exhibit a positive change (increase) in dispersion from 3m to 18m.
2. There is heterogeneity in change in dispersion from 18m to 24m.
3. Across the 22 tissues, there is heterogeneity as to whether more genes experience positive change in dispersion from 3m to 24m than from 3m to 18m.

When investigating genes that are persistently differentiated over the mice lifecourse, we find that

4. There are 2174 persistently DEGs, of which 41 are *entirely* missed by Mann-Whitney test.
5. For the 41 persistently DEGs detected by MOCHIS, we find heterogeneous patterns of change in dispersion over time. In particular, it is *not always true* that the change in dispersion is unidirectional (usually thought to be positive as an organism ages).

TISSUE	TRANSCRIPT	AGE3M_VS_AGE18M	AGE18M_VS_AGE24M	AGE3M_VS_AGE24M
bladder-lumen	ENSMUSG00000041959	+	+	+
bone-marrow	ENSMUSG00000008668	+	+	+
diaphragm	ENSMUSG00000057322	+	-	+
diaphragm	ENSMUSG00000067288	+	+	+
heart	ENSMUSG00000030057	+	+	+
heart	ENSMUSG00000044533	+	+	+
heart	ENSMUSG00000004207	+	+	+
heart	ENSMUSG00000035530	+	+	+
heart	ENSMUSG00000025428	+	+	+
heart	ENSMUSG00000030432	+	+	+
heart	ENSMUSG00000027523	+	+	+
heart	ENSMUSG00000071866	+	+	+
heart	ENSMUSG00000059291	+	+	+
large-intestine	ENSMUSG00000008348	+	-	+
large-intestine	ENSMUSG00000020423	+	-	+
limb-muscle	ENSMUSG00000073702	+	-	+
limb-muscle	ENSMUSG00000038900	+	-	+
liver	ENSMUSG00000008348	+	+	+
mammary-gland	ENSMUSG00000029622	-	+	+
mammary-gland	ENSMUSG00000027523	-	+	-
mesenteric-fat-pad	ENSMUSG00000073702	+	+	+
mesenteric-fat-pad	ENSMUSG00000021250	+	+	+
mesenteric-fat-pad	ENSMUSG00000034994	+	+	+
mesenteric-fat-pad	ENSMUSG00000021127	+	+	+
mesenteric-fat-pad	ENSMUSG00000092341	+	+	+
skin-of-body	ENSMUSG00000004207	+	-	+
skin-of-body	ENSMUSG00000048076	+	+	+
thymus	ENSMUSG00000030057	+	+	+
thymus	ENSMUSG00000025362	+	+	+
thymus	ENSMUSG00000066551	+	+	+
thymus	ENSMUSG00000012405	+	+	+
trachea	ENSMUSG00000060743	+	-	+
trachea	ENSMUSG00000026234	+	-	+
trachea	ENSMUSG00000028788	+	-	+
trachea	ENSMUSG00000025428	+	-	+
trachea	ENSMUSG00000028367	+	-	+
trachea	ENSMUSG00000027447	+	-	+
trachea	ENSMUSG00000031812	+	-	+
trachea	ENSMUSG00000024190	+	-	+
trachea	ENSMUSG00000055302	+	-	+
trachea	ENSMUSG00000015656	+	-	+