Brainspan Data Analysis

ZHAO Kai

Here is a brief introduction of analyzing the results of our proposed approach on Brainspan dataset.

Preparations

```
suppressPackageStartupMessages({
  require(ggplot2)
  require(formatR)
  require(knitr)
  require(cluster)
  require(factoextra)
  require(dplyr)
  require(RColorBrewer)
  require(clusterProfiler)
  require(org.Hs.eg.db)
  require(enrichplot)
  require(stringr)
  require(forcats)
  require(DOSE)
  require(ggplot2)
  require(hrbrthemes)
  require(viridis)
  require(reshape2)
  require(gridExtra)
  require(extrafont)
})
```

```
## Warning: package 'S4Vectors' was built under R version 3.6.3
```

```
truncated_var <- function(x){
    remove_idx <- c(which.max(x), which.min(x))
    var(x[-remove_idx])
}
wrap_labal <- function(x, width = 60){
    str_wrap(x, width=60)
}
```

```
opts_chunk$set(tidy.opts=list(width.cutoff=80),tidy=TRUE)

setwd("~/data/multidimensional_datasets/brainspan_genes_matrix_csv/")
# load results for brain span
load("~/data/Results/brainspan/brainspan_l1l2_penalty_32_June05.RData")
attach(fitted_obj) # attach it for easy syntax
str(fitted_obj) # show the structure of our result
```

```
## List of 6
##  $ iter         : int 28130
##  $ predictions  : num [1:524, 1:43411] 4.7 4.66 4.6 4.74 4.54 ...
##  $ stage_factor : num [1:13, 1:32] 4.48 1.45 4.28 3.37 2.84 ...
##  $ tissue_factor: num [1:26, 1:32] 0.0944 0.0453 1.3931 1.621 0.39 ...
##  $ column_factor: num [1:32, 1:43411] -0.00275 -0.000446 0.026263 0.010472 0.005311 ...
##  $ optimal_rmse : num 0.328
```

```
# read meta information
dic <- read.csv("~/data/Results/brainspan/dictionary.csv", stringsAsFactors = F)
# obtain ensemble genes included in our study
load("brainspan_dataset_annotated_fitered.RData")
gene_id <- data.frame(ensembl_gene_id = colnames(dataset)[-c(1:2)], stringsAsFactors =F)
# match the included genes with meta information
row_meta <- read.csv('rows_metadata.csv', stringsAsFactors = F)
meta <- inner_join(gene_id, row_meta, by = "ensembl_gene_id")

# prepare struture and stage names for naming corresponding latent factors
structure <- unique(dic[,c(6, 9)])
structure <- structure[order(structure[,2]),]
stage <- unique(dic[,c(11, 12)])
r_names <- apply(stage, 1, function(x) paste0(x[2], "_", trimws(x[1])))

# name tissue_factor and stage_factor
rownames(tissue_factor) <- structure[,1]
rownames(stage_factor) <- r_names
```

Explore development trajectory across entire lifespan

Here we explore the development trajectory with the latent representations of development stage factors. In the below exmaple, we selected the metagene with the largest variation across entire lifespan, visualized the trajectory of this metagene across all development stages, and explore the pathway enriched for top 2.5% genes that contribute to this metagene.

In this part of analysis, I only demonstrate with the metagens with greatest and smallest variance. In order to expand, analysis of other metagenes with a single for loop is fine to generate results.

```
# compute the variance for each metagene
matagene_var <- apply(stage_factor, 2, function(x) truncated_var(x))
ord <- order(matagene_var, decreasing = TRUE)
stage_factor[, ord[1:3]]
```

```
##                                 [,1]        [,2]         [,3]
## Early fetal_2                -1.552885  0.84247060 -0.250373525
## Early fetal_3                -1.801252 -1.11877431  0.002251351
## Early mid-fetal_4            -1.575359 -2.48568505 -0.431472003
## Early mid-fetal_5            -1.002906 -0.78830934 -2.327002953
## Late mid-fetal_6             -1.067641 -0.62085805  1.277955020
## Late fetal_7                  3.185649  3.88688700  5.825538299
## Neonatal and early infancy_8  3.346559  5.81230610 -1.554020093
## Late infancy_9                4.800958  3.15825263 -3.498539655
## Early childhood_10            4.805639  3.31020786  0.778813970
## Middle and late childhood_11  5.768576  0.05768701 -2.885860597
## Adolescence_12                5.493569 -2.58191830  4.919778218
## Young adulthood_13            6.457557 -3.01822699 -3.033868138
```

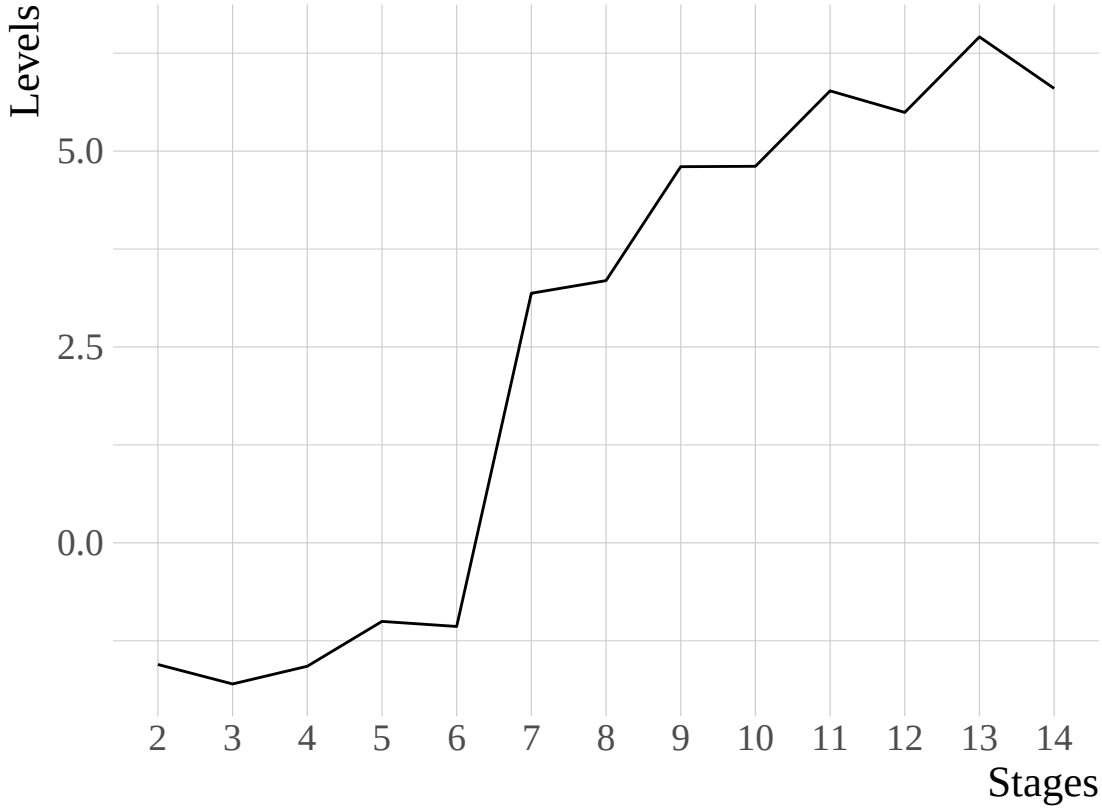
```
## Middle adulthood_14           5.799758 -4.11316549  0.823232639
```

```
# use the most variably metagene as an example
metagene_id <- ord[1]
cat("Column_id:", metagene_id, "\n")
```

```
## Column_id: 9
```

The plot below show the trajectory of the selected metagene cross all development stages.

```
loadfonts(quiet = T)
result <- data.frame(stage = r_names, levels = stage_factor[, metagene_id], stringsAsFactors = F)
result$stage <- factor(r_names, levels = r_names)
# ggplot(data = result, aes(x = stage, y = levels, group = 1)) +
# geom_line(linetype = 'dashed') + geom_point() + xlab('Stages') +
# ylab('Levels') + theme(plot.title = element_text(size=12, face = 'bold',
# hjust = 0.5), axis.title.y = element_text(size=10),
# text=element_text(size=10, family='Times New Roman'), axis.text.x =
# element_text(size=10,angle = 45, vjust = 1, hjust=1))
ggplot(data = result, aes(x = stage, y = levels, group = 1)) + scale_color_viridis(discrete = T) +
    scale_x_discrete(labels = 2:14) + geom_line() + theme_ipsum(base_family = "Times New Roman",
    base_size = 14, axis_title_size = 16) + xlab("Stages") + ylab("Levels")
```



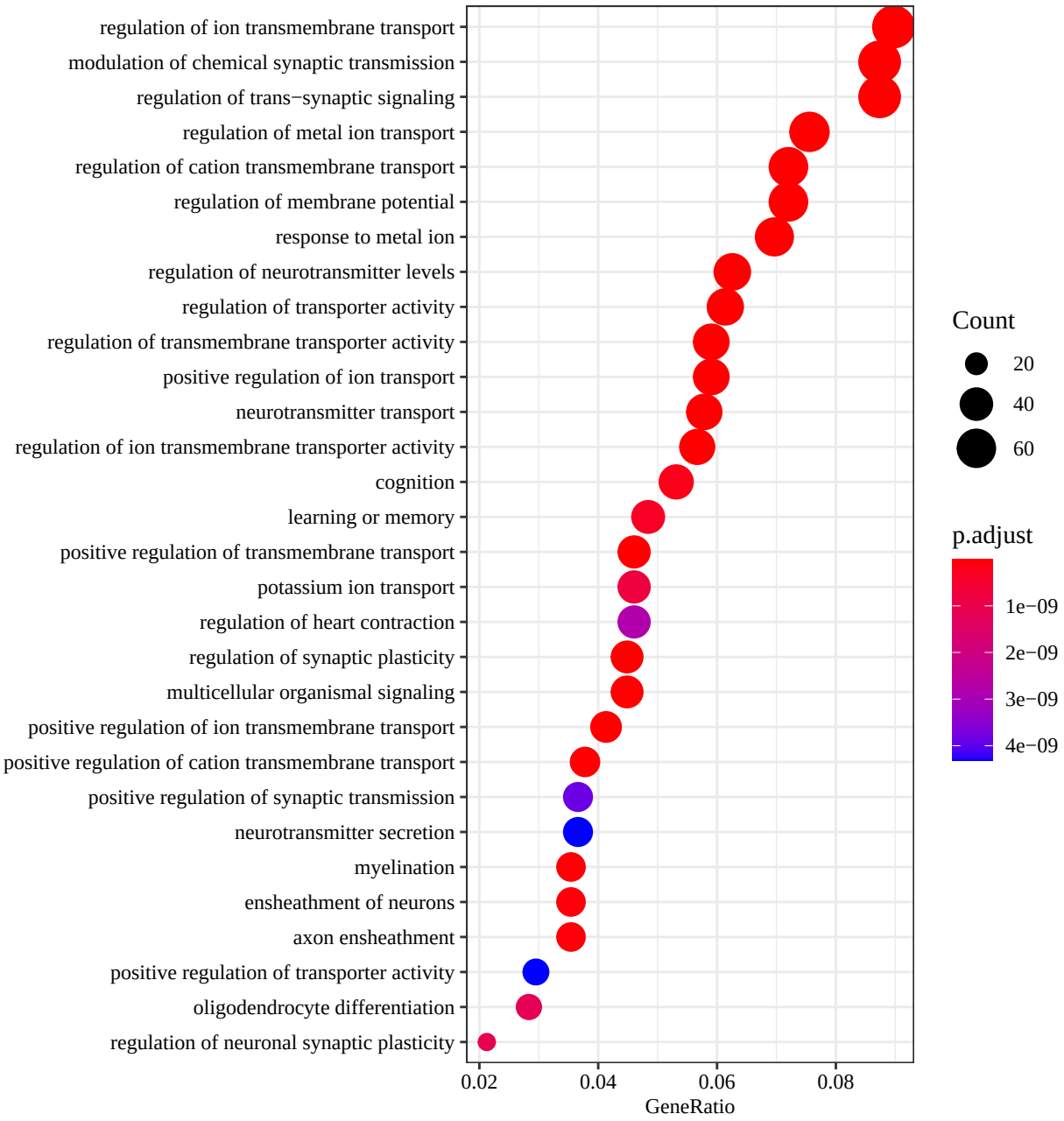

```
# axis.text.x = element_text(size=10,angle = 45, vjust = 1, hjust=1))
```

Then, we investigated the pathway enriched for top 2.5% genes that up-regulates and down-regulates this metagene.

```
cat("Column_id:", metagene_id, "\n")
```

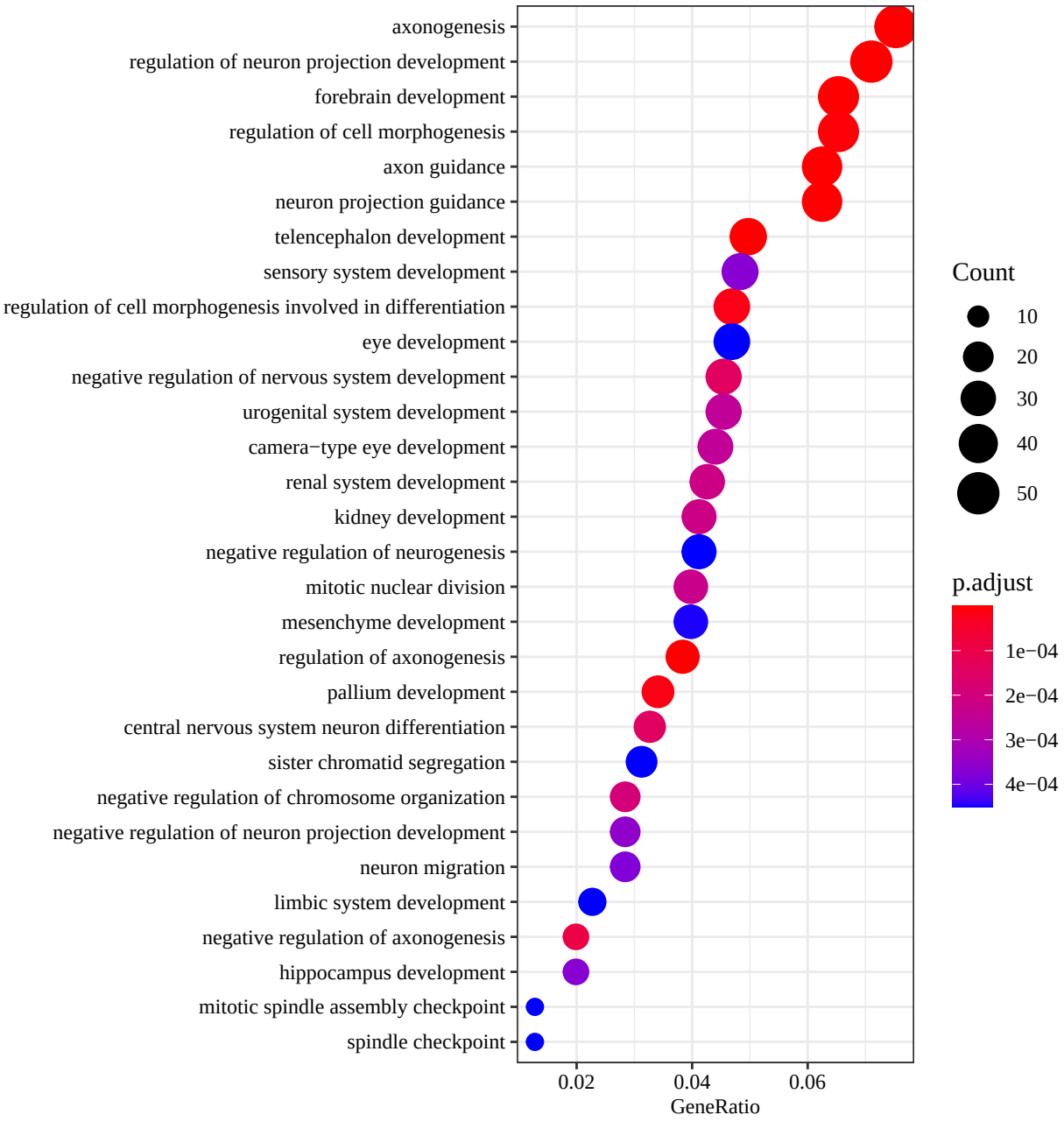
```
## Column_id: 9
```

```
cutoffs <- quantile(column_factor[metagene_id, ], probs = seq(0, 1, 0.025))

# up-regulation, select the highest quantile
selected <- (column_factor[metagene_id, ] >= cutoffs[length(cutoffs) - 1])  #
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(upreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) str_wrap(x,
    width = 60)) + theme(text = element_text(family = "Times New Roman"))
```




```
# the object from enrichGO can be converted to data frame with the following.
# result <- data.frame(upreg) save(upreg, file = paste0('metagene',
# metagene_id, 'upreg_dev_pathway.RData'))
```

```
# down-regulation, select the lowest quantile
selected <- (column_factor[metagene_id, ] <= cutoffs[2])
downreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(downreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```




```
# save(upreg, file = paste0('metagene', metagene_id,
# 'downreg_dev_pathway.RData'))
```

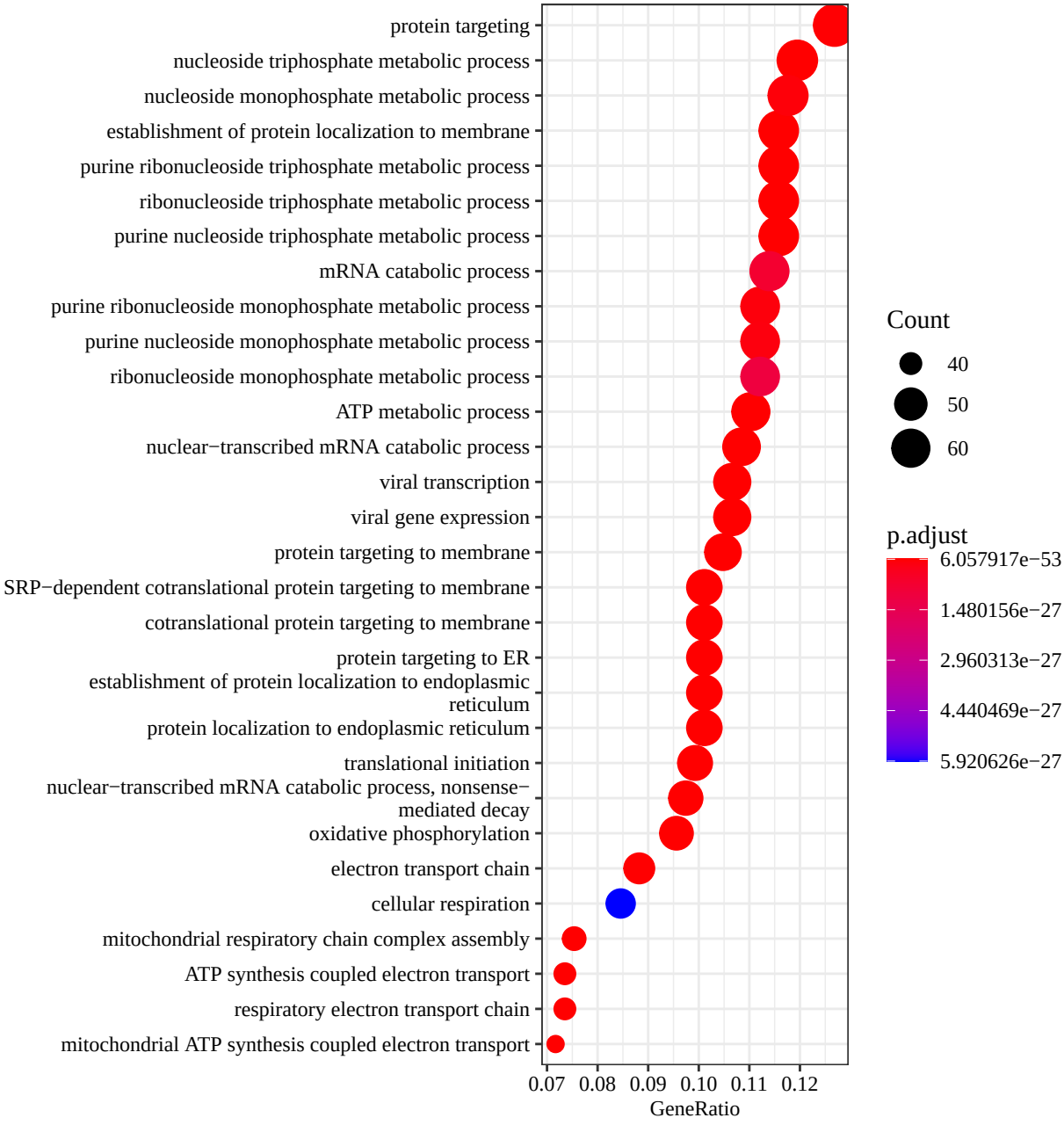
Then, we explore the metagene with the least variance.

```
metagene_id <- ord[length(ord)]
```

```
cat("Column_id:", metagene_id, "\n")
```

```
## Column_id: 17
```

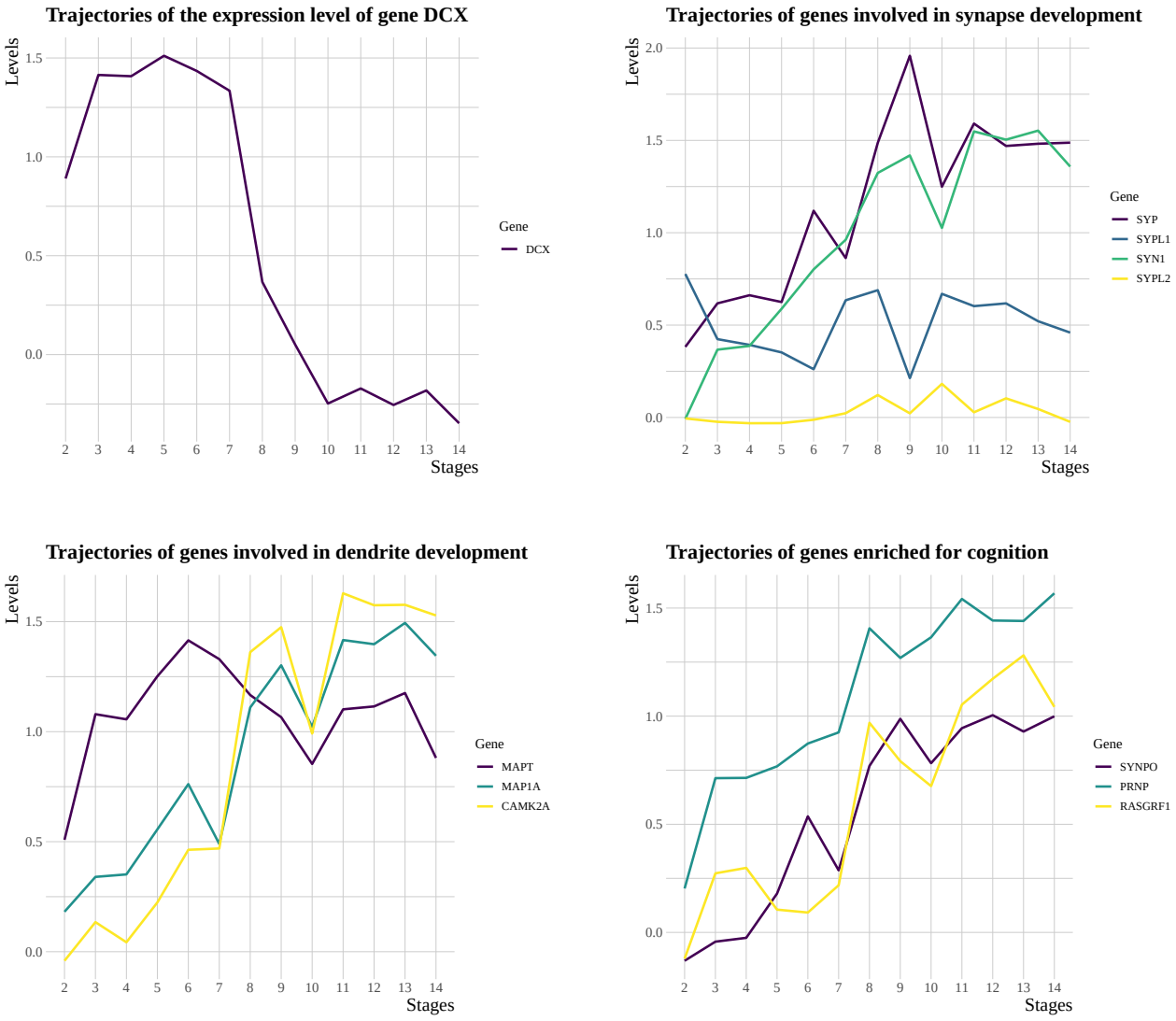
```
cutoffs <- quantile(column_factor[metagene_id, ], probs = seq(0, 1, 0.025))

# up-regulation, select the highest quantile
selected <- (column_factor[metagene_id, ] >= cutoffs[length(cutoffs) - 1])
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(upreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```



```r
# save(upreg, file = paste0('metagene', metagene_id,
# 'upreg_dev_pathway.RData'))

# down-regulation, select the lowest quantile selected <-
# (column_factor[metagene_id,] <= cutoffs[2]) downreg <- enrichGO(gene =
# unique(meta[selected,5]), OrgDb = 'org.Hs.eg.db', ont = 'BP', readable =
# TRUE) dot_plot(downreg, font = 8, showCategory=50) save(upreg, file =
# paste0('metagene', metagene_id, 'downreg_dev_pathway.RData'))
```

```r
# DCX expressed in neuronal progenitor cells and immature migrating neurons
# SYP, SYPL1, SYN1 for synapse development MAP1A, MAPT, CAMK2A for dendrite
# development
selected_genes <- c("DCX", "SYP", "SYPL1", "SYN1", "SYPL2", "MAPT", "MAP1A", "CAMK2A",
    "NRXN1", "POU2F3", "SYNPO", "PRNP", "RASGRF1")
col_ids <- sapply(selected_genes, function(x) which(meta[[4]] == x))

# result <- as.data.frame(downreg) cognition_genes <-
# unlist(strsplit(result[23,8], split = '/')) cognition_ids <-
# sapply(cognition_genes, function(x) which(meta[[4]] == x))

stage_profiles <- stage_factor %*% column_factor
selected <- stage_profiles[, col_ids]
colnames(selected) <- selected_genes
rownames(selected) <- r_names
result <- melt(selected)
colnames(result) <- c("Stage", "Gene", "Levels")
result$Stage <- factor(r_names, levels = r_names)

par(mfrow = c(2, 2))

p1 <- ggplot(data = result[result$Gene == "DCX", ], aes(x = Stage, y = Levels, group = Gene,
    color = Gene)) + scale_color_viridis(discrete = TRUE) + geom_line(size = 1) +
    ggtitle("Trajectories of the expression level of gene DCX") + theme_ipsum(base_family = "Times New
    base_size = 12, plot_title_face = "bold", axis_title_size = 16, plot_title_size = 18) +
    scale_x_discrete(labels = 2:14) + xlab("Stages") + ylab("Levels")

p2 <- ggplot(data = result[result$Gene %in% c("SYP", "SYN1", "SYPL1", "SYPL2"), ],
    aes(x = Stage, y = Levels, group = Gene, color = Gene)) + scale_color_viridis(discrete = TRUE) +
    geom_line(size = 1) + ggtitle("Trajectories of genes involved in synapse development") +
    theme_ipsum(base_family = "Times New Roman", base_size = 12, plot_title_face = "bold",
        axis_title_size = 16, plot_title_size = 18) + scale_x_discrete(labels = 2:14) +
    xlab("Stages") + ylab("Levels")

p3 <- ggplot(data = result[result$Gene %in% c("MAP1A", "MAPT", "CAMK2A"), ], aes(x = Stage,
    y = Levels, group = Gene, color = Gene)) + scale_color_viridis(discrete = TRUE) +
    geom_line(size = 1) + ggtitle("Trajectories of genes involved in dendrite development") +
    theme_ipsum(base_family = "Times New Roman", base_size = 12, plot_title_face = "bold",
        axis_title_size = 16, plot_title_size = 18) + scale_x_discrete(labels = 2:14) +
    xlab("Stages") + ylab("Levels")

p4 <- ggplot(data = result[result$Gene %in% c("SYNPO", "PRNP", "RASGRF1"), ], aes(x = Stage,
    y = Levels, group = Gene, color = Gene)) + scale_color_viridis(discrete = TRUE) +
    geom_line(size = 1) + ggtitle("Trajectories of genes enriched for cognition") +
    theme_ipsum(base_family = "Times New Roman", base_size = 12, plot_title_face = "bold",
```

```
        axis_title_size = 16, plot_title_size = 18) + scale_x_discrete(labels = 2:14) +
    xlab("Stages") + ylab("Levels")
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



Explore pathways that contribute to the brain structure development

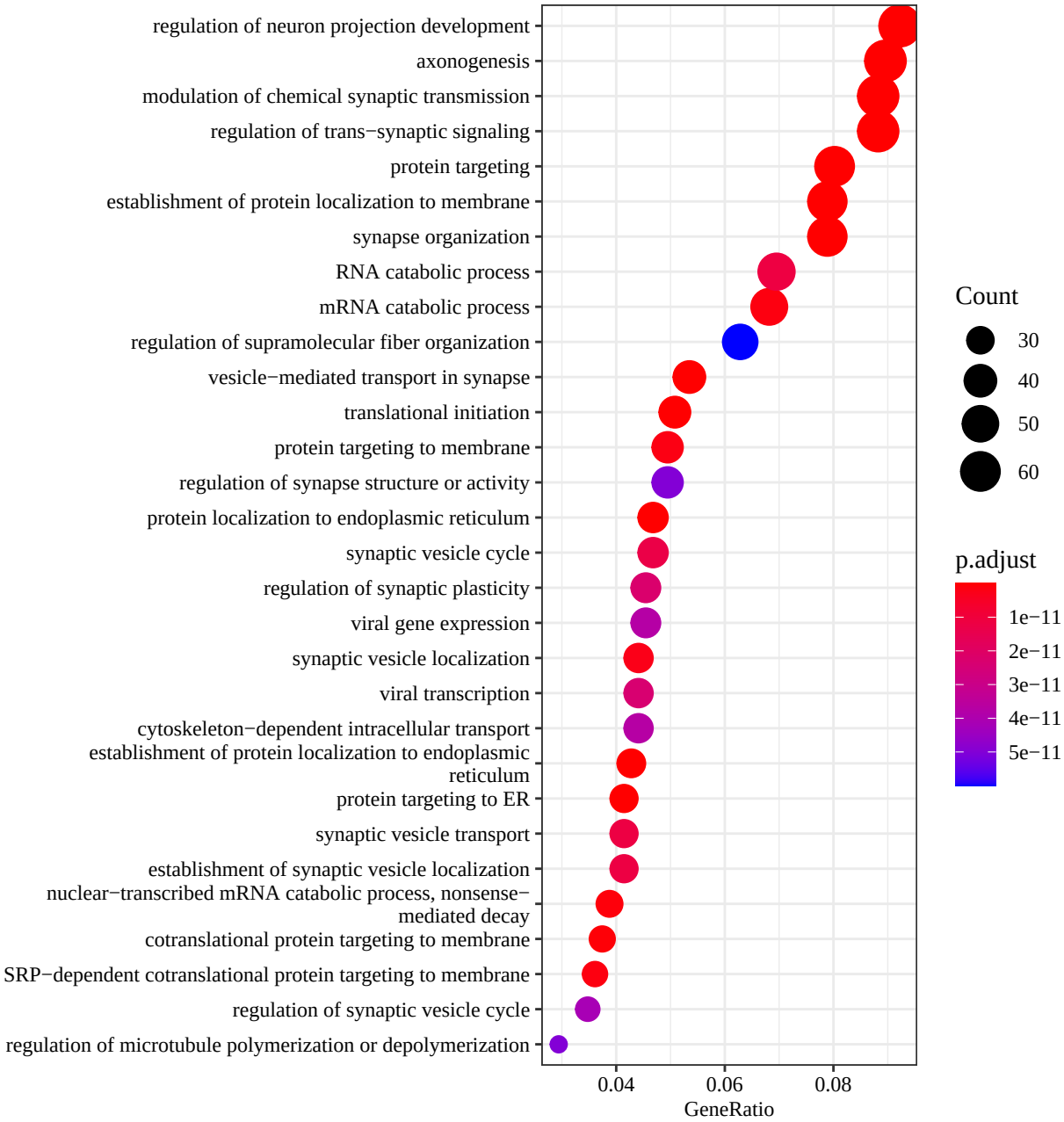
First, we obtain general expression profiles for different tissues, and analyze the functional pathways for each tissue.

In this part of analysis, I only demonstrate with the second tissue. In order to expand, analysis of other metagenes with a single for loop is fine to generate results for all tissues.

```
# tissue_matrix <- tissue_factor[,-c(14)] %*% column_factor[-c(14),]
tissue_matrix <- tissue_factor %*% column_factor
rownames(tissue_matrix) <- rownames(tissue_factor)

# use the second brain region as example
tissue_id <- 2
cat("Tissue name:", rownames(tissue_factor)[tissue_id], "\n")
```

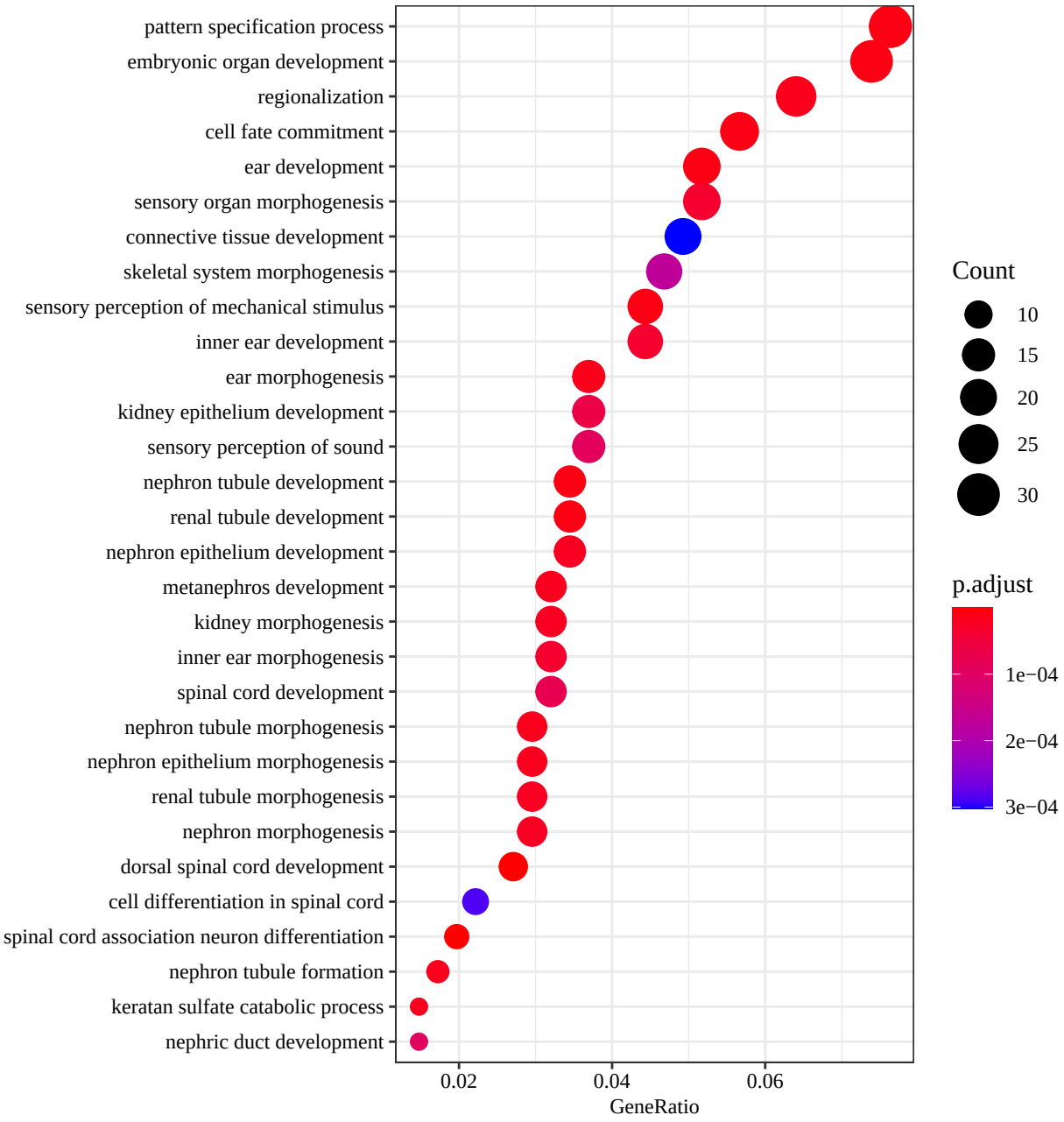
```
## Tissue name: M1C-S1C
```

```
cutoffs <- quantile(tissue_matrix[tissue_id, ], probs = seq(0, 1, 0.025))

# up_regulation, select the highest quantile
selected <- (tissue_matrix[tissue_id, ] >= cutoffs[length(cutoffs) - 1])
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
# result <- data.frame(upreg)
dotplot(upreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```




```
# down-regulation, select the lowest quantile
selected <- (tissue_matrix[tissue_id, ] <= cutoffs[2])
```

```
downreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(downreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```

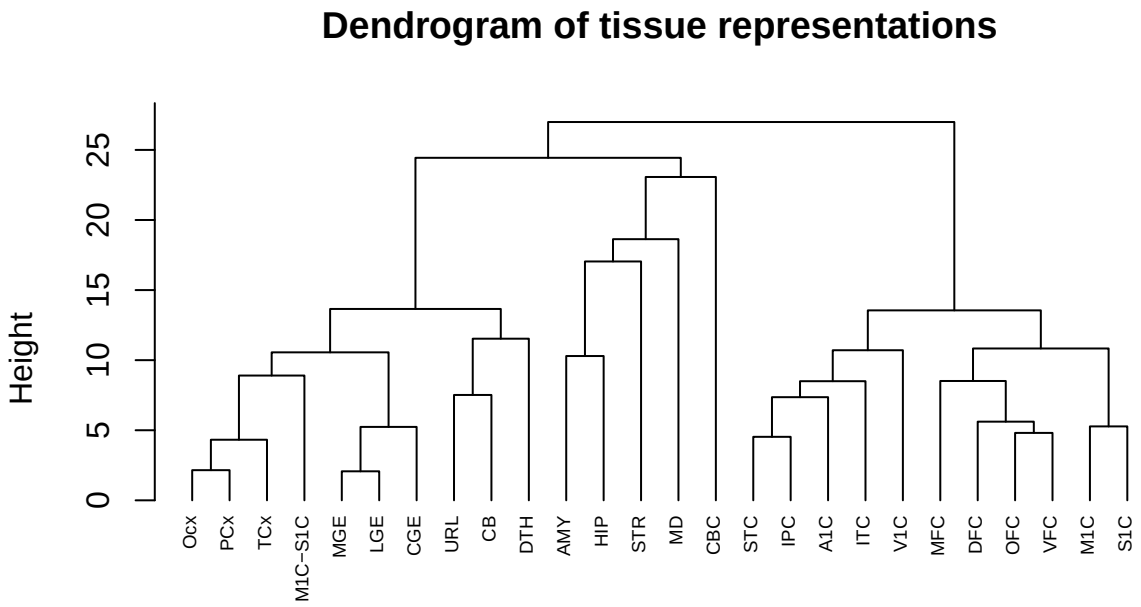



```
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")

# function to compute coefficient
ac <- function(x) {
    agnes(tissue_factor, method = x)$ac
}
ac_vec <- sapply(m, function(x) ac(x))
```

```
ac_vec
```

```
##   average    single  complete      ward
## 0.6288339 0.6156082 0.6351167 0.7110749
```

```
hc3 <- agnes(tissue_factor, method = unname(m[which(ac_vec == max(ac_vec))]))
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of tissue representations")
```



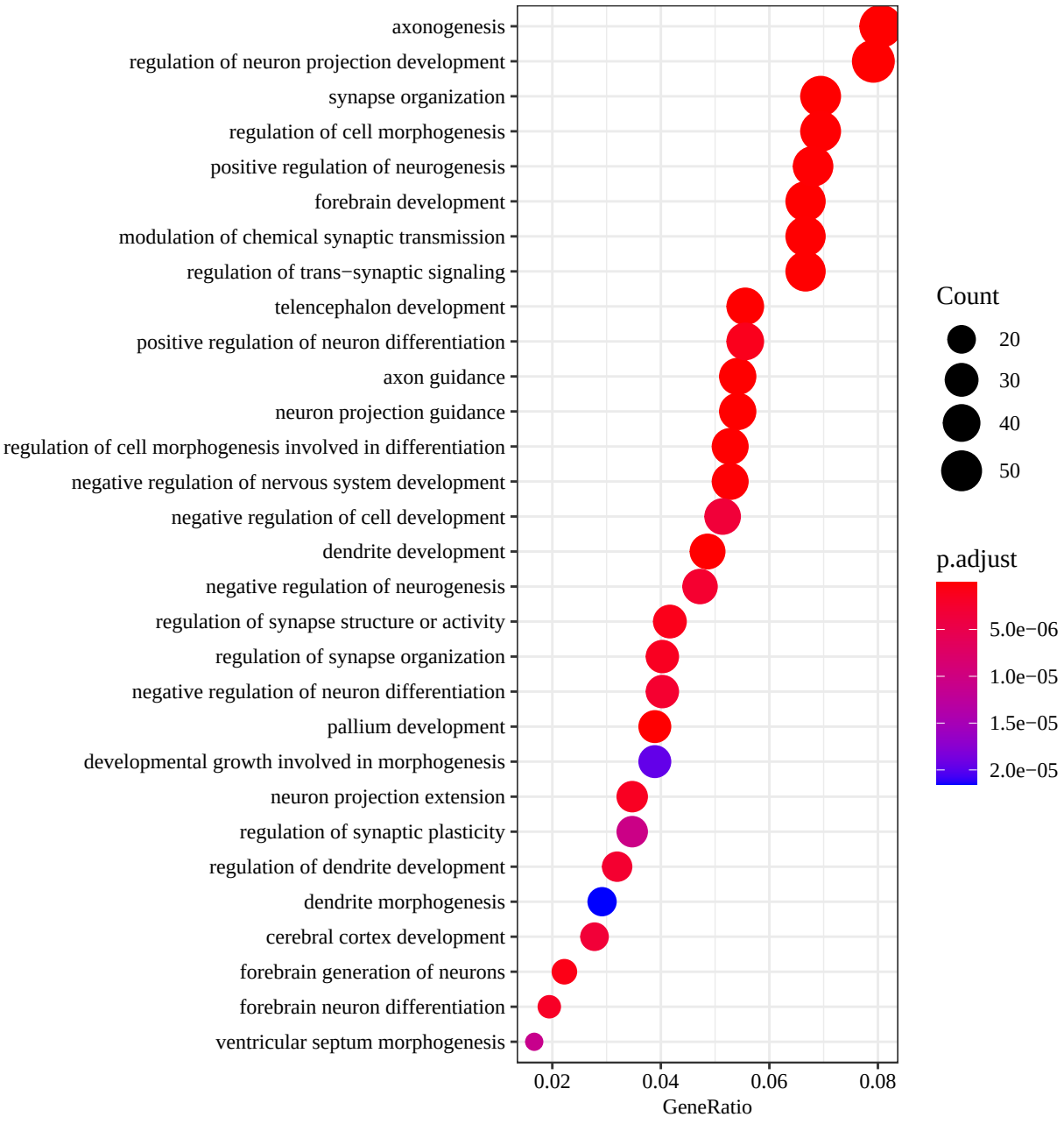
Second, we explore the pathways that contribute the most to the expression difference across different brain structures. This analysis can identify biological processes that partially contribute to the difference across different brain structures.

The second kind of analysis only can partial explain the pathways that signal differently across all tissues in our study, so expanding it with other two or three metagenes with high variance is enough.

```
# tiss_var <- apply(tissue_factor, 2, var)
tiss_var <- apply(tissue_factor, 2, function(x) truncated_var(x))

metagene_id <- which.max(tiss_var)
cat("Column_id:", metagene_id, "\n")
```

```
## Column_id: 26
```

```
cutoffs <- quantile(column_factor[metagene_id, ], probs = seq(0, 1, 0.025))

# up-regulation, select the highest quantile
selected <- (column_factor[metagene_id, ] >= cutoffs[length(cutoffs) - 1])
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(upreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```



Explore the interaction between development stages and brain regions

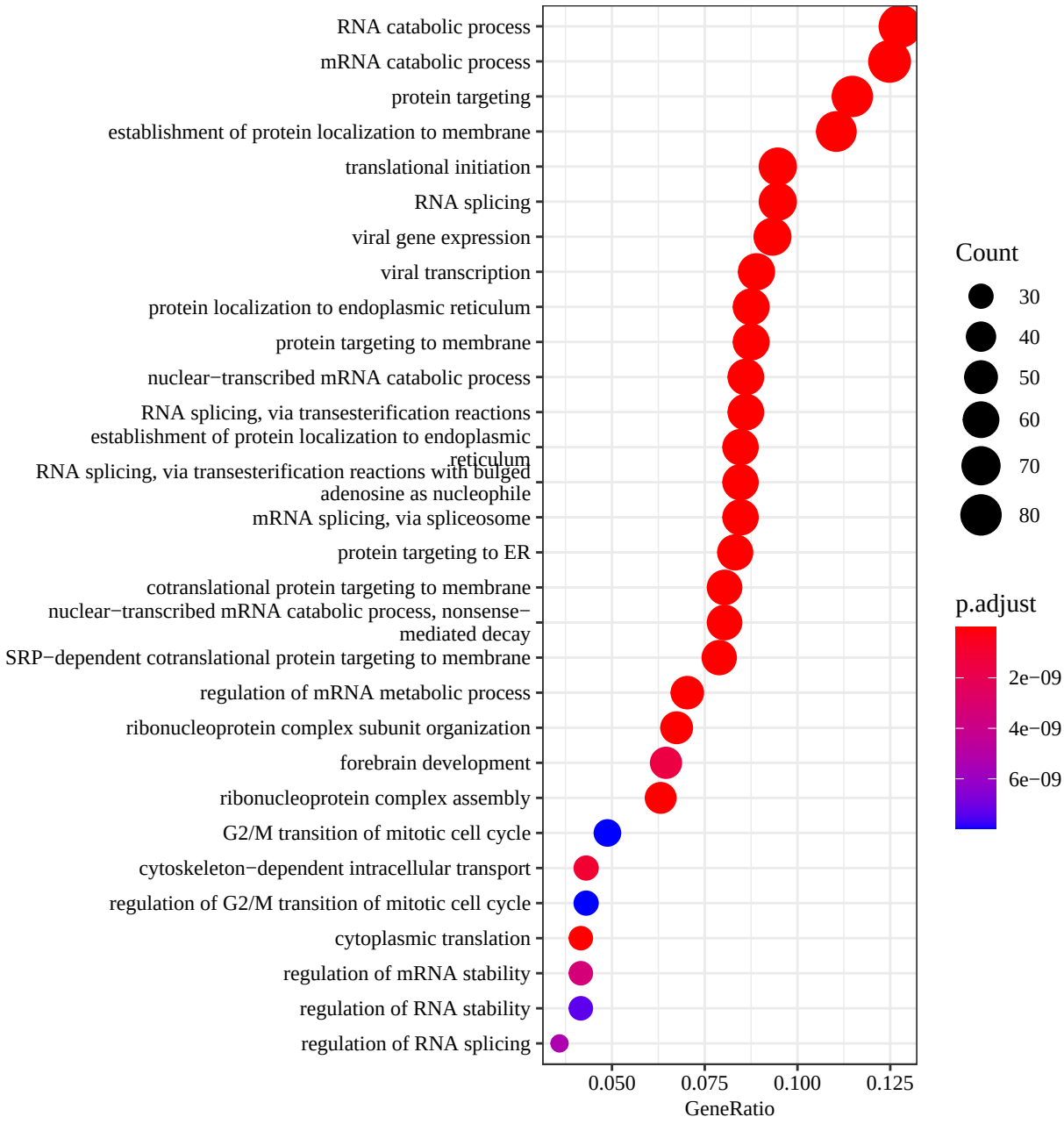
Exploring the interaction is an important feature of our approach, so if possible we may carry out analysis on all possible combinations between brain regions and development stages and select reasonable results for interpretation.

```
# calculate the correlation between different tissues and stages correlations
# <- cor(t(stage_factor[,-c(3,13)]), t(tissue_factor[,-c(3,13)]))
correlations <- cor(t(stage_factor), t(tissue_factor))
rownames(correlations) <- rownames(stage_factor)
colnames(correlations) <- rownames(tissue_factor)
# find the regions with correlations to each development stage
imp_regions <- t(apply(correlations, 1, function(x) colnames(correlations)[order(x,
```

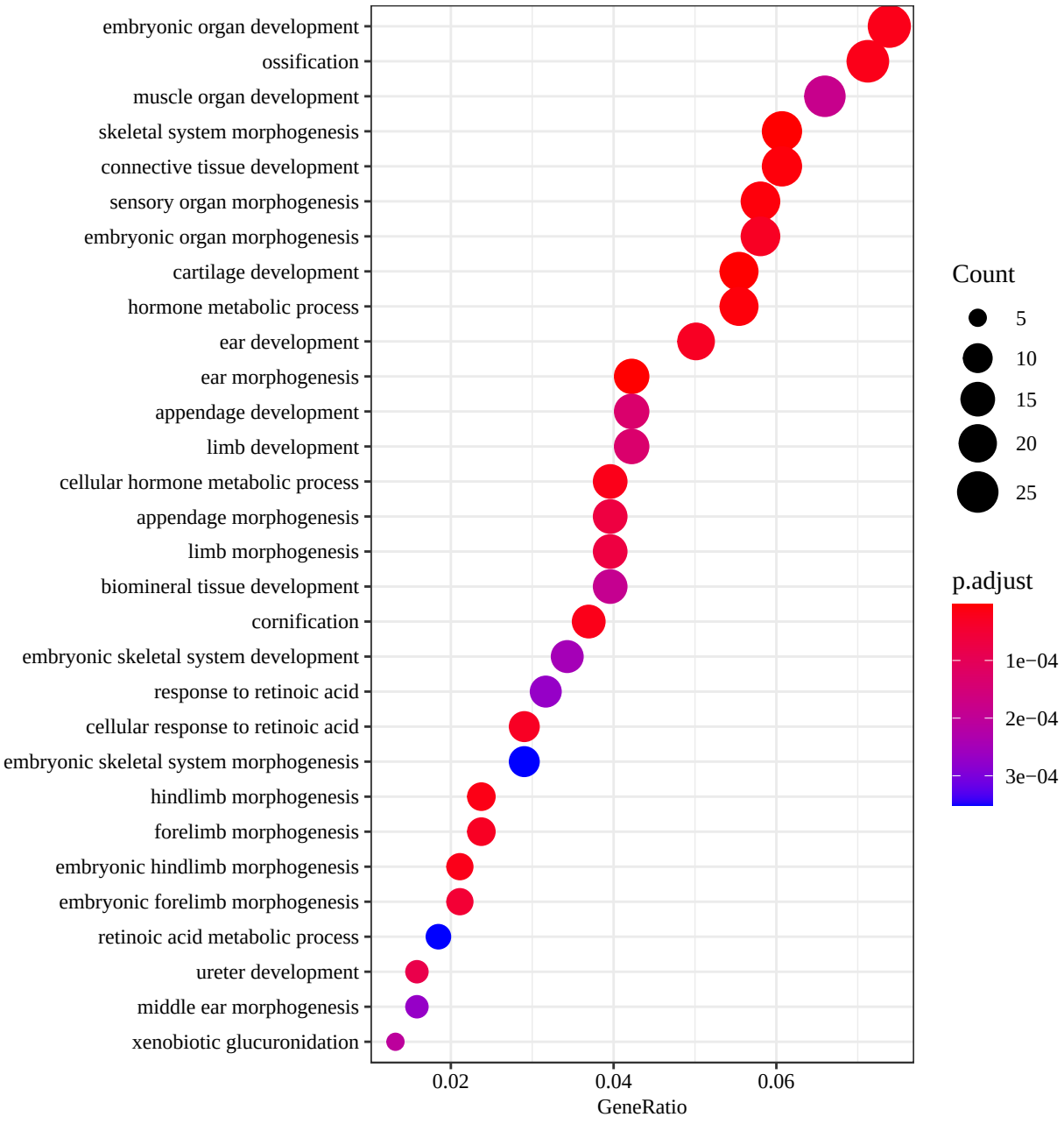
```
    decreasing = TRUE)[1:5])))
head(imp_regions)
```

```
##                   [,1]  [,2]      [,3]  [,4]  [,5]
## Early fetal_2     "LGE" "A1C"     "MGE" "CGE" "VFC"
## Early fetal_3     "VFC" "TCx"     "S1C" "IPC" "M1C"
## Early mid-fetal_4 "PCx" "M1C-S1C" "A1C" "MGE" "LGE"
## Early mid-fetal_5 "PCx" "TCx"     "MGE" "LGE" "Ocx"
## Late mid-fetal_6  "MGE" "VFC"     "LGE" "S1C" "ITC"
## Late fetal_7      "A1C" "TCx"     "MGE" "URL" "M1C-S1C"
```

```
# here we use this two as an example
stage_idx <- which(rownames(stage_factor) == "Early fetal_2")
tissue_idx <- which(rownames(tissue_factor) == "LGE")
# interaction_effect <- drop((stage_factor[stage_idx, -c(14)] *
# tissue_factor[tissue_idx, -c(14)]) %*% column_factor[-c(14), ])
interaction_effect <- drop((stage_factor[stage_idx, ] * tissue_factor[tissue_idx,
    ]) %*% column_factor)
cutoffs <- quantile(interaction_effect, probs = seq(0, 1, 0.025))

# up-regulation, select the highest quantile
selected <- (interaction_effect >= cutoffs[length(cutoffs) - 1])

# down-regulation, select the lowest quantile selected <- (interaction_effect
# <= cutoffs[2])
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)

dotplot(upreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```




```
# up-regulation, select the highest quantile
selected <- (interaction_effect <= cutoffs[2])

# down-regulation, select the lowest quantile selected <- (interaction_effect
# <= cutoffs[2])
downreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)

dotplot(downreg, font = 9, showCategory = 30) + scale_y_discrete(labels = function(x) wrap_labal(x)) +
    theme(text = element_text(family = "Times New Roman"))
```


embryonic organ development
ossification
muscle organ development
skeletal system morphogenesis
connective tissue development
sensory organ morphogenesis
embryonic organ morphogenesis
cartilage development
hormone metabolic process
ear development
ear morphogenesis
appendage development
limb development
cellular hormone metabolic process
appendage morphogenesis
limb morphogenesis
biomineral tissue development
cornification
embryonic skeletal system development
response to retinoic acid
cellular response to retinoic acid
embryonic skeletal system morphogenesis
hindlimb morphogenesis
forelimb morphogenesis
embryonic hindlimb morphogenesis
embryonic forelimb morphogenesis
retinoic acid metabolic process
ureter development
middle ear morphogenesis
xenobiotic glucuronidation
0.02
0.04
0.06
GeneRatio
Count
5
10
15
20
25
p.adjust
1e−04
2e−04
3e−04