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)
})

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)
}
```

```
glm_interaction <- function(object, inc_cfd){

  residual <- object[['data']]

  confounder_num <- ncol(object[['confounder']])
  for(i in 1:confounder_num){
    sub_predictions <- object[['cfd_matrices']][[i]] %*% object[['column_factor']]
    residual <- residual - sub_predictions[object[['confounder']][,i], ]
  }
```

```
  column_factor <- object[['column_factor']]
  train_indicator <- object[['train_indicator']]

  confounder <- object[['confounder']][, inc_cfd]
  unique_cfd <- unique(confounder)

  interaction_indicator <- rep(0, nrow(confounder))
  for(k in 1:nrow(unique_cfd)){
    selected <- apply(confounder, 1, function(x) all(x == unique_cfd[k,]))
    interaction_indicator[selected] <- k
  }

  unique_ita <- unique(interaction_indicator)
  coeff_matrix <- matrix(0, nrow = length(unique_ita), ncol = nrow(column_factor))
  pval_matrix <- matrix(0, nrow = length(unique_ita), ncol = nrow(column_factor))

  for(i in unique_ita) {

    ids <- which(interaction_indicator == i);

    st_idx <- 1; ed_idx <- 1
    nonzero_num <- length(ids) * ncol(column_factor);
    outcomes = rep(0,nonzero_num);
    features = matrix(0, nrow = nonzero_num, ncol = nrow(column_factor))

    for(k in ids){
      ed_idx = st_idx + ncol(column_factor) - 1;
      features[st_idx:ed_idx, ] = t(column_factor);
      outcomes[st_idx:ed_idx] = residual[k,];
      st_idx = ed_idx + 1
    }

    data <- data.frame(response = outcomes, features)
    fit <- glm(response ~ . - 1, family = gaussian(), data = data)
    coeff_matrix[i,] <- unname(coefficients(fit))
    pval_matrix[i,] <- coef(summary(fit))[,4]
  }
  return(list(unique_cfd, coeff_matrix, pval_matrix))
}
```

```
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/insider_brainspan_fitted_object.RData")
# load("~/data/Results/brainspan/insider_brainspan_R23_fitted_object.RData")
attach(object) # attach it for easy syntax
str(object) # show the structure of our result
```

```
## List of 9
##  $ data           : num [1:524, 1:43411] 5.23 4.66 4.35 4.84 4.39 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:524] "V2" "V3" "V4" "V5" ...
##   .. ..$ : chr [1:43411] "ENSG00000000003" "ENSG00000000005" "ENSG00000000419" "ENSG00000000457" ...
```

```
##  $ confounder     : num [1:524, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:524] "V2" "V3" "V4" "V5" ...
##   .. ..$ : chr [1:2] "preriod_id" "sid"
##  $ trainset       : num [1:524, 1:43411] 5.23 0 0 0 4.39 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:524] "V2" "V3" "V4" "V5" ...
##   .. ..$ : chr [1:43411] "ENSG00000000003" "ENSG00000000005" "ENSG00000000419" "ENSG00000000457" ...
##  $ testset        : num [1:524, 1:43411] 0 4.66 4.35 4.84 0 ...
##  $ train_indicator: int [1:524, 1:43411] 1 0 0 0 1 1 1 1 1 1 ...
##  $ params         :List of 4
##   ..$ global_tol : num 1e-10
##   ..$ sub_tol    : num 1e-05
##   ..$ tuning_iter: num 100
##   ..$ max_iter   : num 50000
##  $ cfd_matrices   :List of 2
##   ..$ factor0: num [1:13, 1:19] -0.6989 0.347 0.0491 0.2013 0.1966 ...
##   ..$ factor1: num [1:26, 1:19] -1.18 -1.06 0.79 2.98 -1.2 ...
##  $ column_factor  : num [1:19, 1:43411] -0.00716 0.02072 0 0.00857 0.00791 ...
##  $ test_rmse      : num 4.66e-310
##  - attr(*, "class")= chr "insider"
```

```r
stage_factor <- cfd_matrices[[1]]
tissue_factor <- cfd_matrices[[2]]
# interactions <- cfd_matrices[[3]]

# 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(data), 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.

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

```
##                                  [,1]        [,2]       [,3]
## Early fetal_2                7.746964  0.09639597 -2.8537136
## Early fetal_3                8.285853  1.94231041 -6.0739432
## Early mid-fetal_4            8.354576  2.38083760 -3.0360041
## Early mid-fetal_5            6.241265  0.67605865  4.9439051
## Late mid-fetal_6             4.298411  1.65966519  9.4115303
## Late fetal_7                 1.251700 -8.41390699  3.3085385
## Neonatal and early infancy_8 -4.216844 -4.98698266  0.6482326
## Late infancy_9               -6.183211  3.63431613 -1.4479388
## Early childhood_10           -5.654813 -8.19647401  0.3832644
## Middle and late childhood_11 -7.523024  1.32828576 -4.5693528
## Adolescence_12               -6.917620 -3.09416958  1.8918560
## Young adulthood_13           -7.028660  1.07289911  1.2381415
## Middle adulthood_14          -8.127014  5.20730137  0.7595117
```

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

```
## Column_id: 16
```

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

```r
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")
```

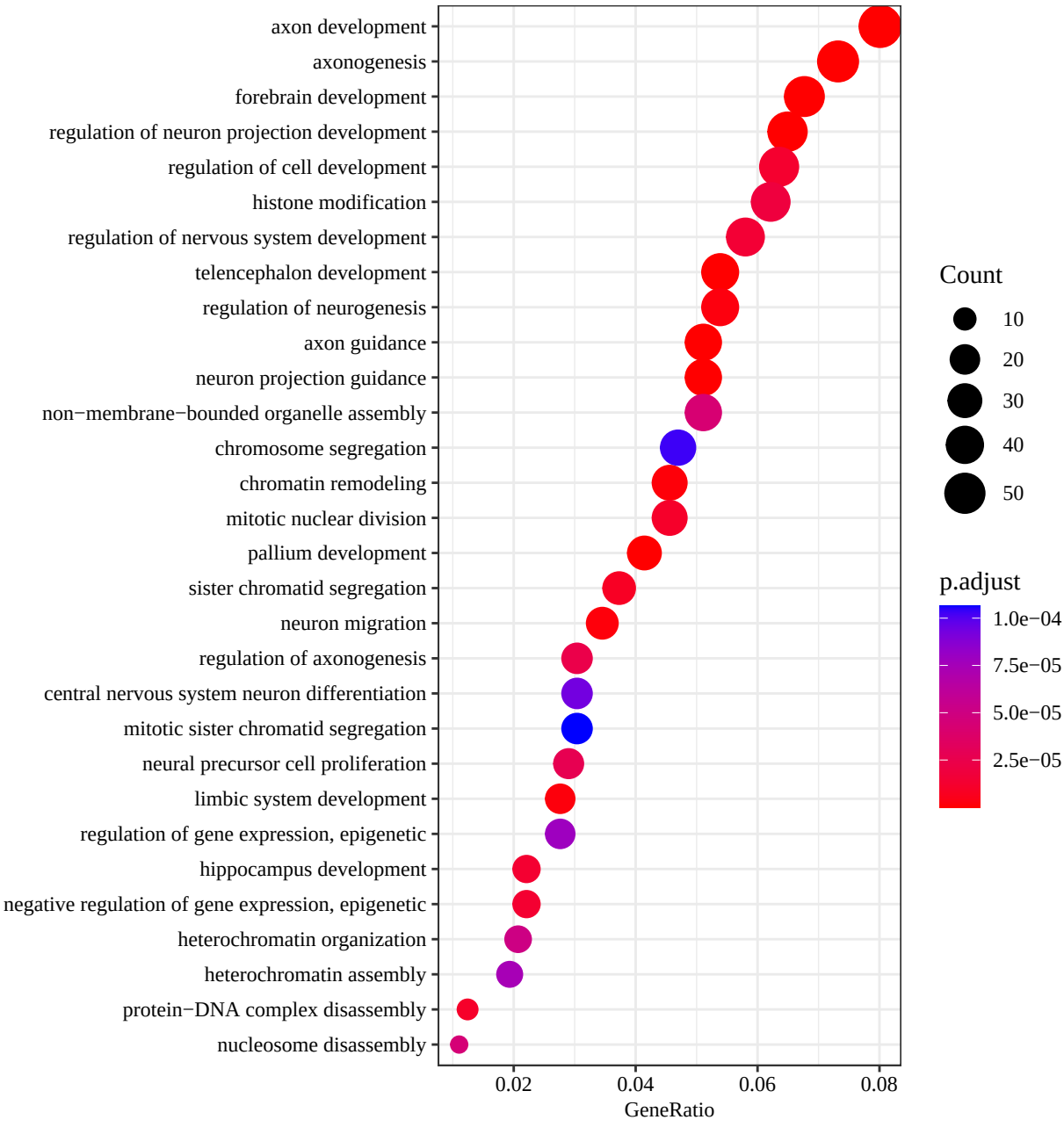

```
# axis.text.x = element_text(size=10,angle = 45, vjust = 1, hjust=1))
```

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

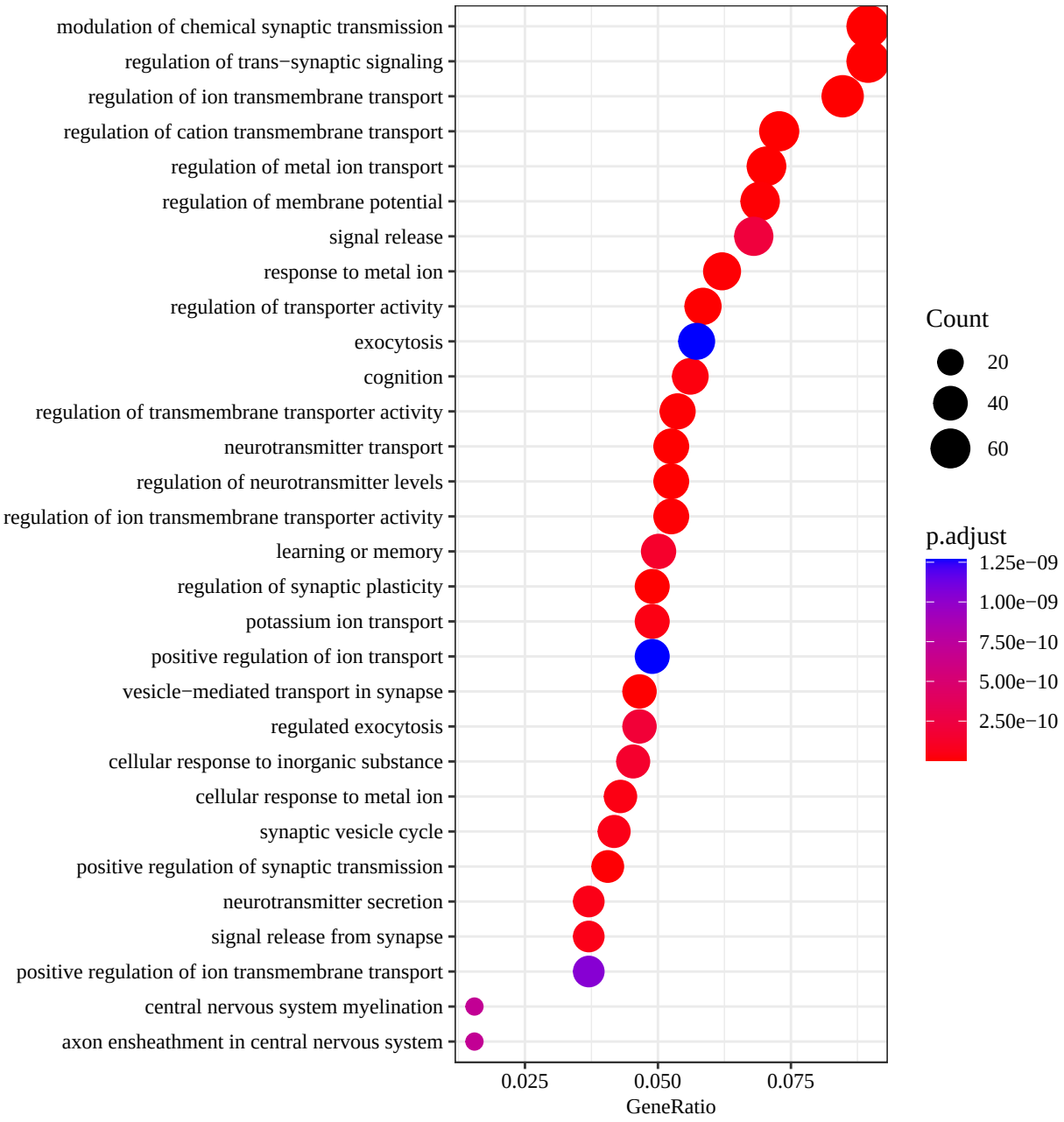
```
cat("Column_id:", metagene_id, "\n")
```

```
## Column_id: 16
```

```
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, label_format = 60) +
    theme(text = element_text(family = "Times New Roman"))
```




```
# 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, label_format = 60) +
    theme(text = element_text(family = "Times New Roman"))
```



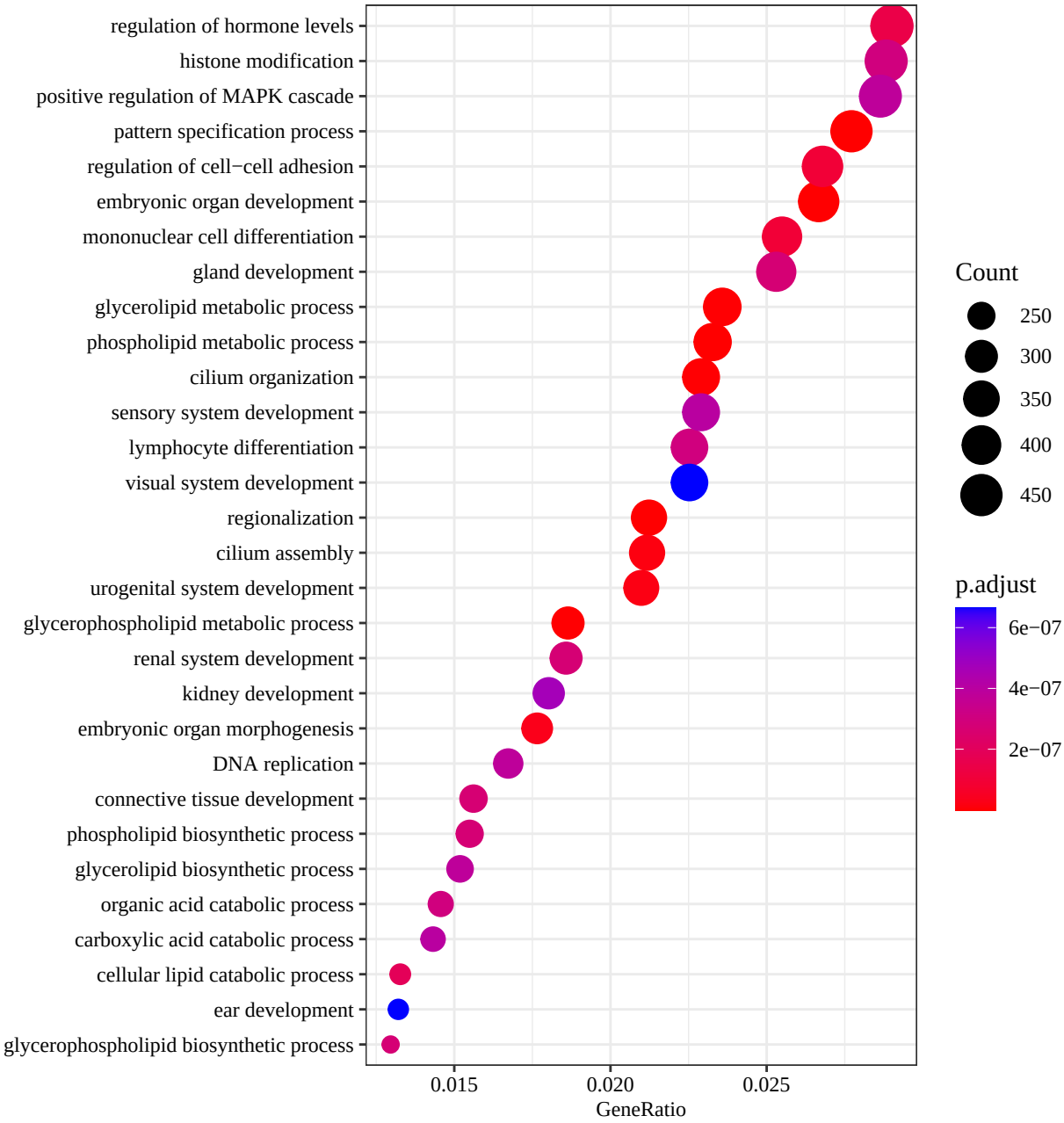
Then, we explore the metagene with the least variance.

```
# metagene_id <- ord[length(ord)]
metagene_id <- which.min(apply(tissue_factor, 2, function(x) truncated_var(x)))
cat("Column_id:", metagene_id, "\n")
```

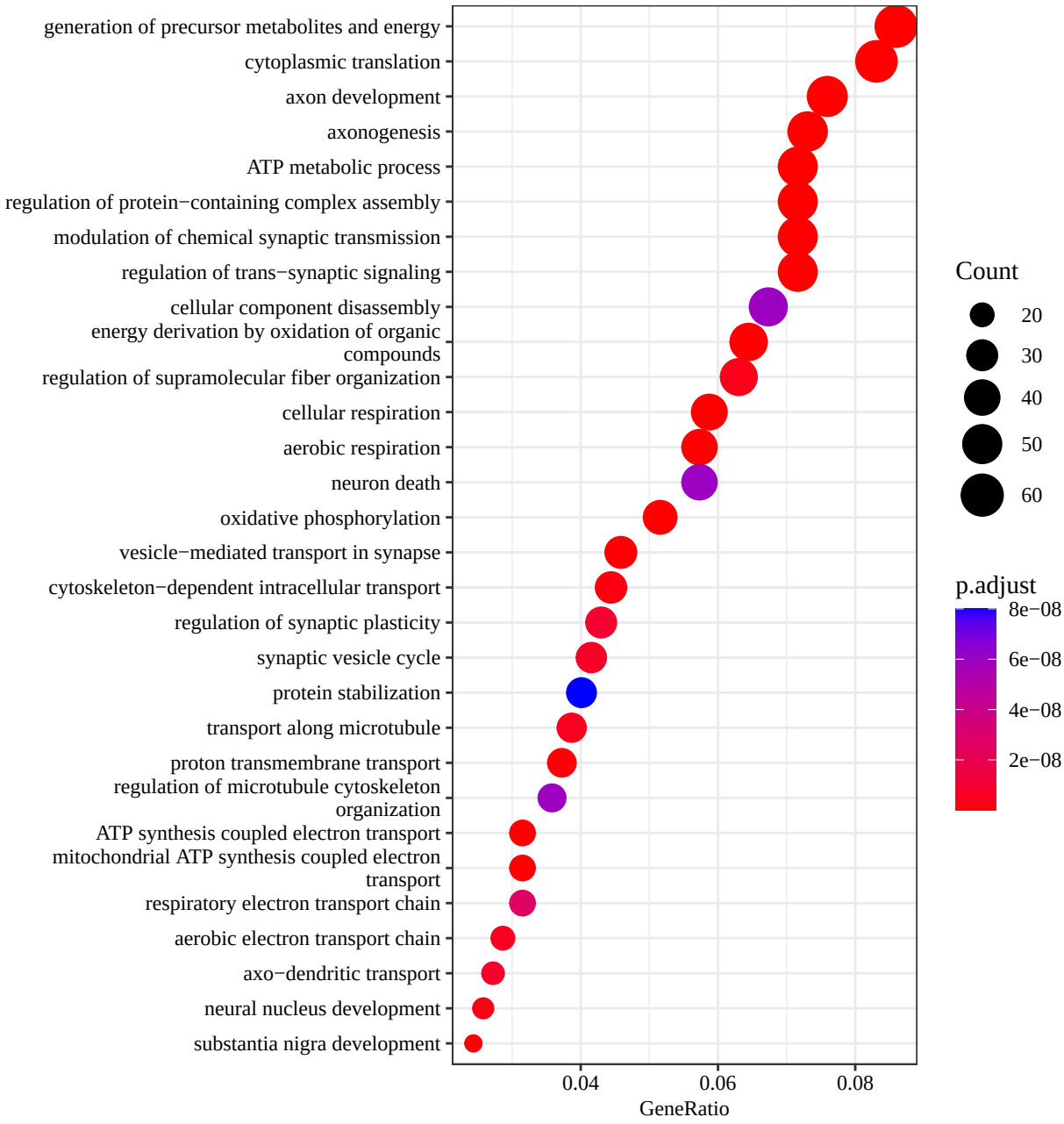
```
## Column_id: 14
```

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

# up-regulation, select the highest quantile
selected <- (column_factor[metagene_id, ] >= cutoffs[2])
upreg <- enrichGO(gene = unique(meta[selected, 5]), OrgDb = "org.Hs.eg.db", ont = "BP",
    readable = TRUE)
dotplot(upreg, font = 9, showCategory = 30, label_format = 50) +
```

```
    theme(text = element_text(family = "Times New Roman"))
```

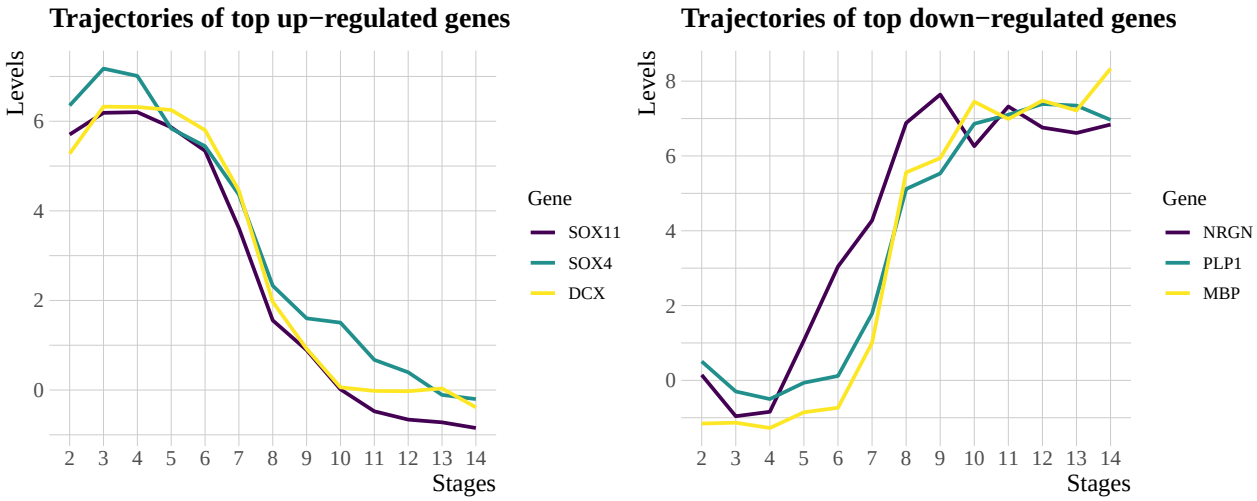



```
# save(upreg, file = paste0('metagene', metagene_id, 'upreg_dev_pathway.RData'))

# 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, label_format = 50) +
    theme(text = element_text(family = "Times New Roman"))
```




```
metagene_id <- ord[1]
gene_order <- order(column_factor[metagene_id, ], decreasing = TRUE)
selected_genes <- gene_id[gene_order[1:5], ]
col_ids <- sapply(selected_genes, function(x) which(meta[[1]] == x))

stage_profiles <- stage_factor %*% column_factor
selected <- stage_profiles[, col_ids]
colnames(selected) <- meta[[4]][col_ids]
rownames(selected) <- r_names
result1 <- melt(selected[, -c(1, 2)])
colnames(result1) <- c("Stage", "Gene", "Levels")
result1$Stage <- factor(r_names, levels = r_names)
```

```
selected_genes <- gene_id[gene_order[(length(gene_order) - 4):length(gene_order)], ]
col_ids <- sapply(selected_genes, function(x) which(meta[[1]] == x))

selected <- stage_profiles[, col_ids]
colnames(selected) <- meta[[4]][col_ids]
rownames(selected) <- r_names
result2 <- melt(selected[, -c(1, 4)])
colnames(result2) <- c("Stage", "Gene", "Levels")
result2$Stage <- factor(r_names, levels = r_names)

p1 <- ggplot(data = result1, aes(x = Stage, y = Levels, group = Gene, color = Gene)) +
    xlab("Stages") + ylab("Levels") + scale_color_viridis(discrete = TRUE) + geom_line(size = 1) +
    ggtitle("Trajectories of top up-regulated genes") + theme_ipsum(base_family = "Times New Roman",
    base_size = 12, axis_title_size = 14, plot_title_size = 16) + theme(plot.margin = unit(c(5,
    2, 5, 2), "pt")) + scale_x_discrete(labels = 2:14)

p2 <- ggplot(data = result2, aes(x = Stage, y = Levels, group = Gene, color = Gene)) +
    xlab("Stages") + ylab("Levels") + scale_color_viridis(discrete = TRUE) +
    geom_line(size = 1) + ggtitle("Trajectories of top down-regulated genes") +
    theme_ipsum(base_family = "Times New Roman", base_size = 12, axis_title_size = 14,
        plot_title_size = 16) + theme(plot.margin = unit(c(5, 2, 5, 2), "pt")) +
    scale_x_discrete(labels = 2:14)

grid.arrange(p1, p2, 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.

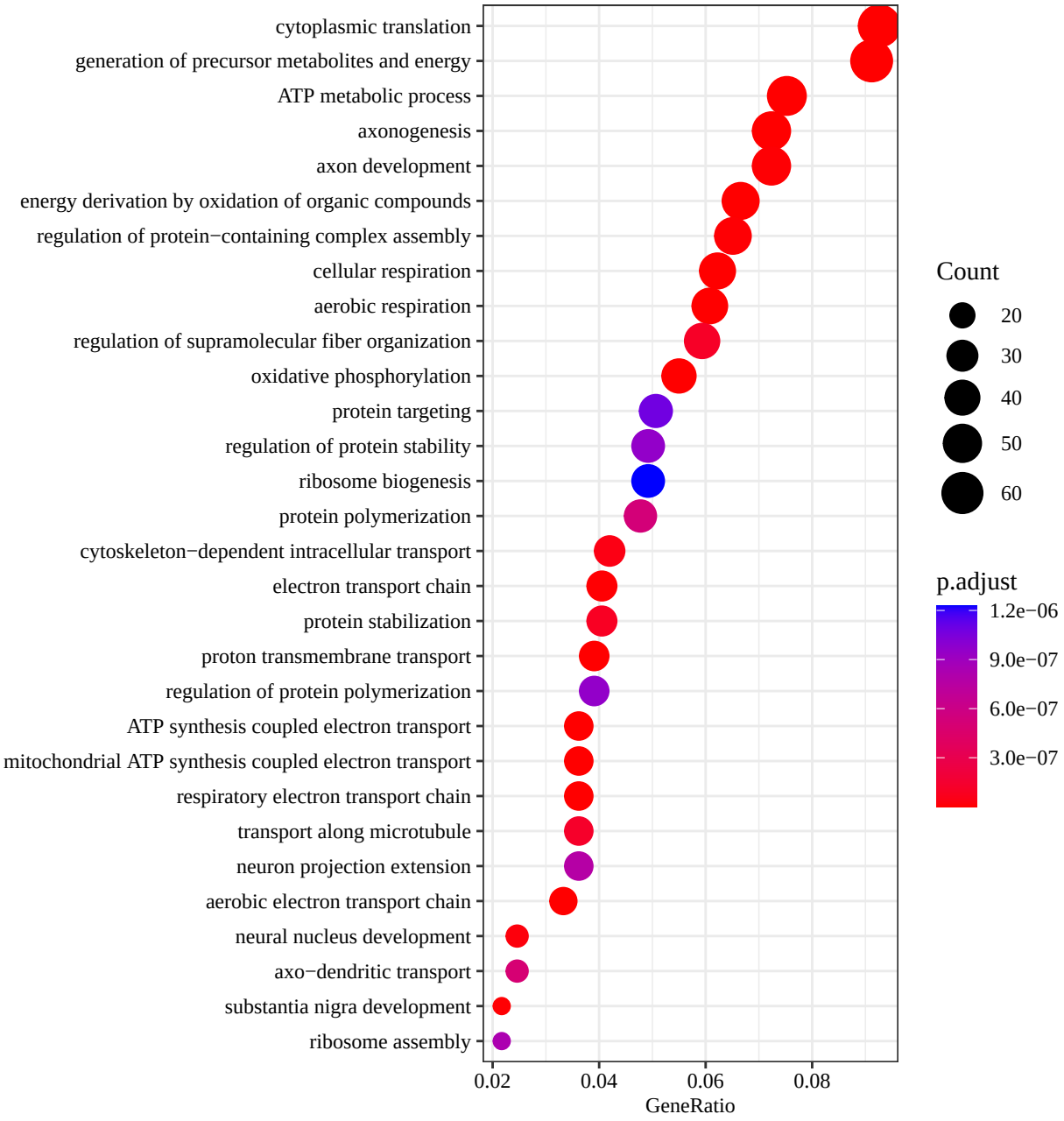
In this part of analysis, I only demonstrate with the second tissue. In order to expand, analysis of other metagenes with a single for loop is fine to generate results for all tissues.

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

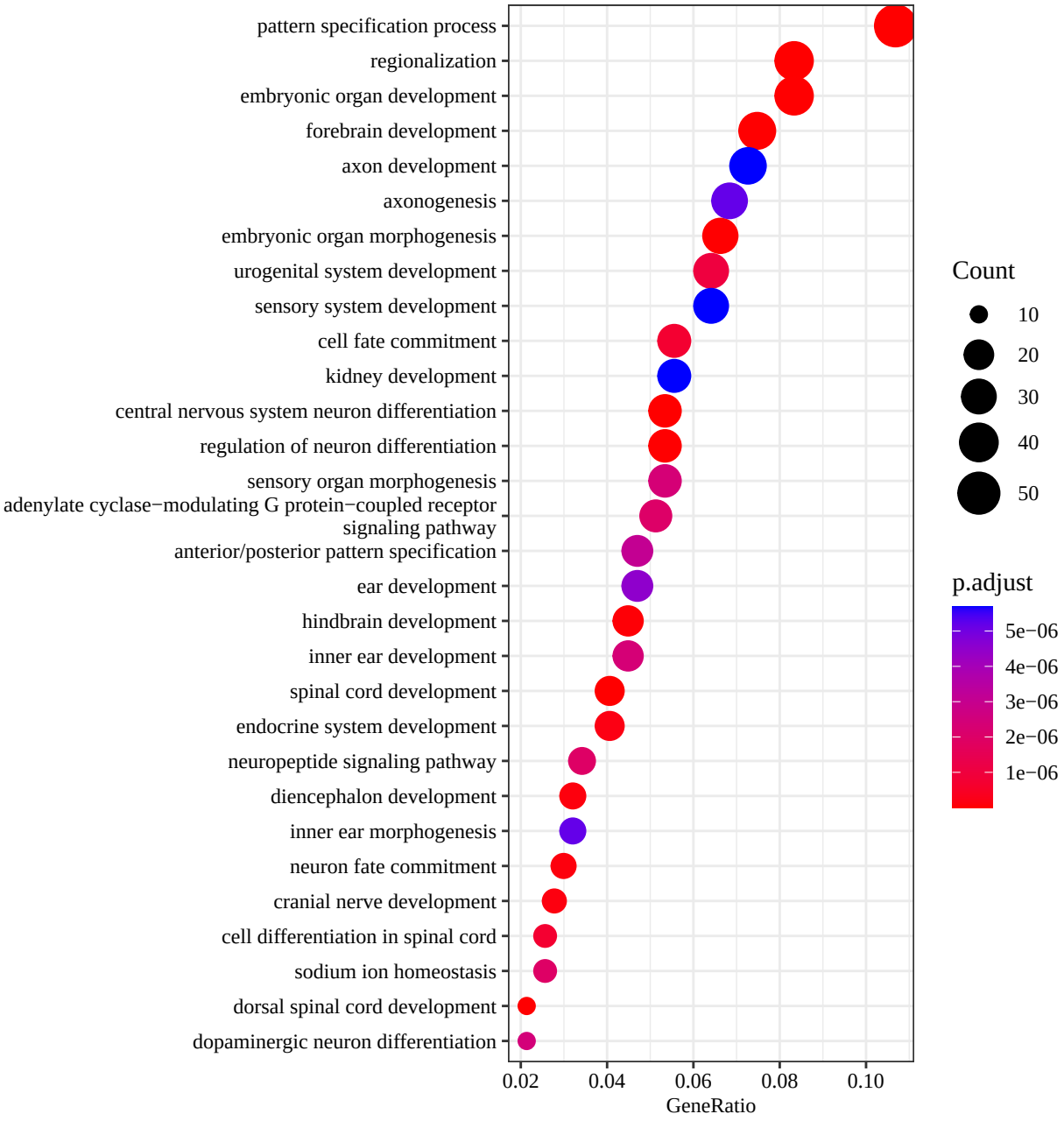
```r
# use the second brain region as example
tissue_id <- 2
cat("Tissue name:", rownames(tissue_factor)[tissue_id], "\n")
```

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

```r
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, label_format = 60) + theme(text = element_text(family = "Ti
```



```
# 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 = TRU
dotplot(downreg, font = 9, showCategory = 30, label_format = 60) +
    theme(text = element_text(family = "Times New Roman"))
```

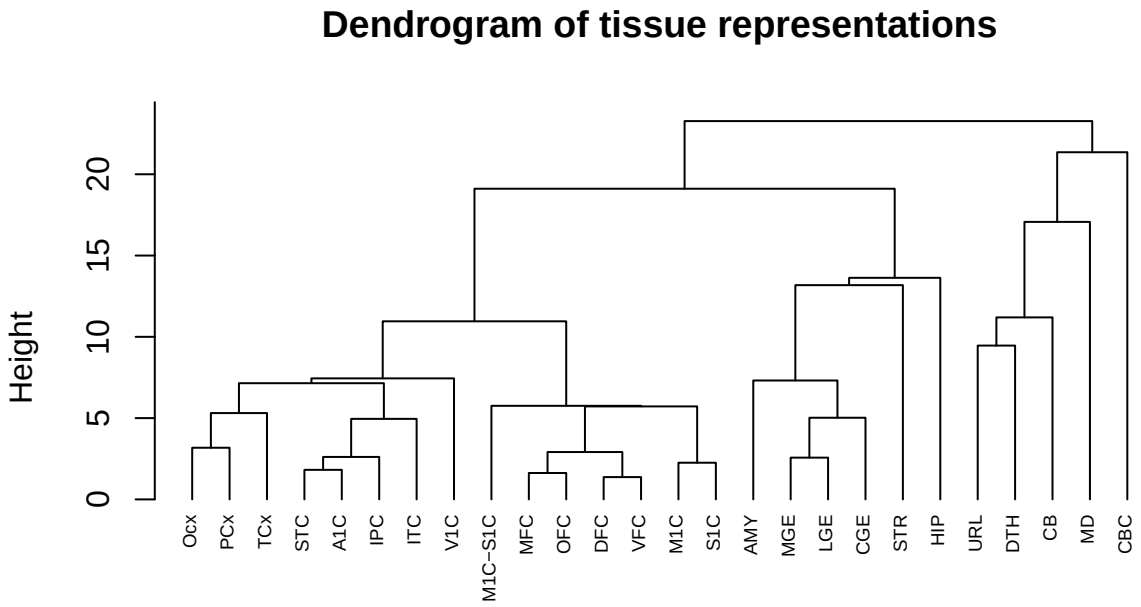



```
m <- c("average", "single", "complete", "ward")
names(m) <- c("average", "single", "complete", "ward")

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

```r
ac_vec <- sapply(m, function(x) ac(x))
ac_vec
```

```
##   average    single  complete      ward
## 0.6733423 0.6851574 0.6938096 0.7365267
```

```r
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")
```



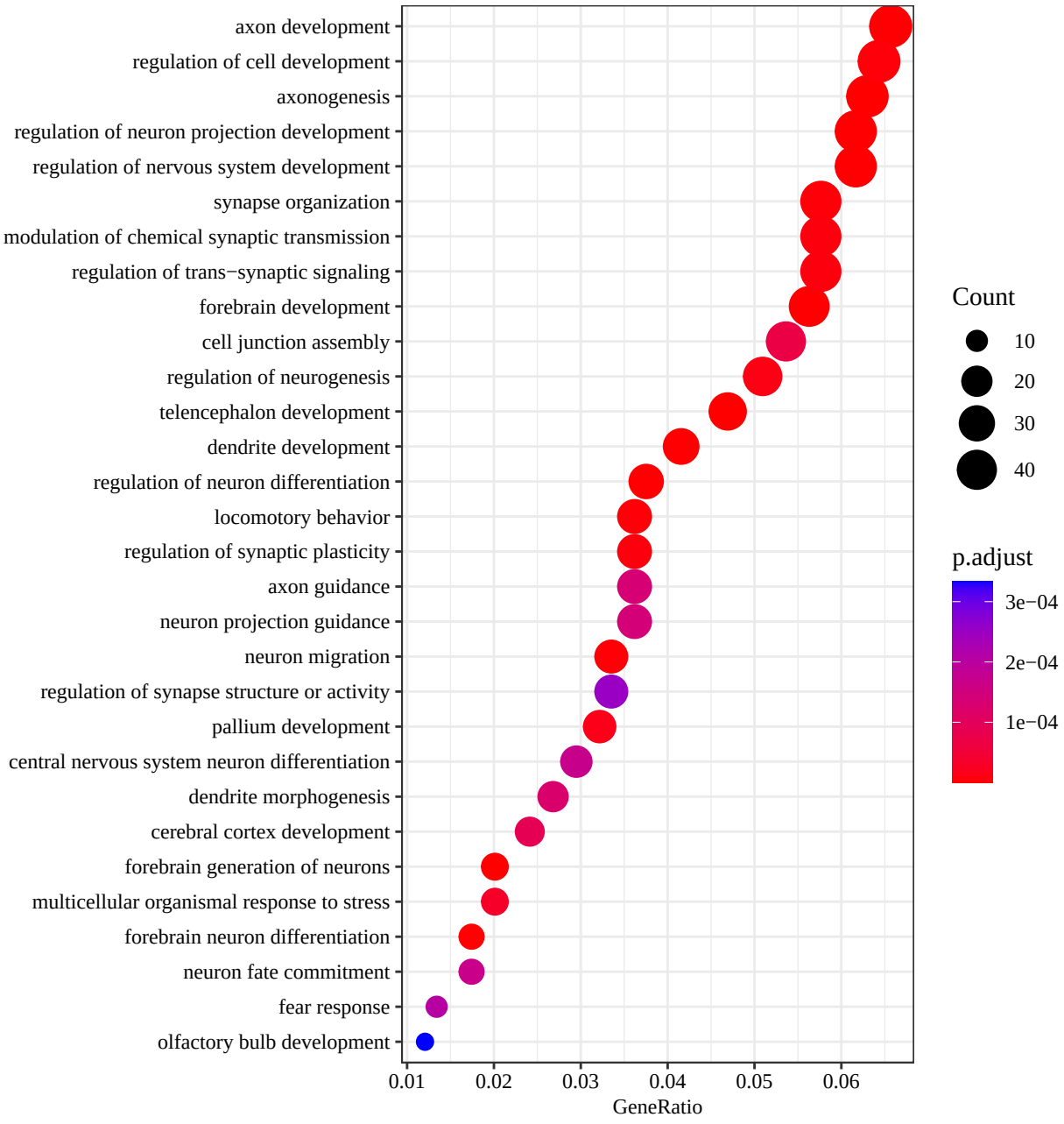
Second, we explore the pathways that contribute the most to the expression difference across different brain structures. This analysis can identify biological processes that partially contribute to the difference across different brain structures.

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.

```r
# 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: 12
```

```r
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, label_format = 60) + theme(text = element_text(family = "Ti
```



Explore the interaction between development stages and brain regions

\textcolor{red}{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.

```
table(dic[, c(9, 11)])
```

```
##     Period
## sid  2 3 4 5 6 7 8 9 10 11 12 13 14
##   1  2 0 0 0 0 0 0 0  0  0  0  0  0
##   2  2 0 0 2 1 0 0 0  0  0  0  0  0
##   3  2 3 3 3 1 2 3 0  4  3  3  5  1
```

```
##   4  2 0 0 0 0 0 0 0  0  0  0  0  0
##   5  1 2 2 4 2 3 3 1  5  3  4  5  1
##   6  2 0 0 0 0 0 0 0  0  0  0  0  0
##   7  2 0 0 0 0 0 0 0  0  0  0  0  0
##   8  2 3 0 0 0 0 0 0  0  0  0  0  0
##   9  2 2 3 4 2 2 2 1  3  3  3  5  0
##   10 2 3 3 4 2 3 2 1  4  3  3  4  1
##   11 2 3 3 3 1 2 2 1  3  2  3  5  1
##   12 2 0 0 0 0 0 0 0  0  0  0  0  0
##   13 1 3 3 2 2 2 3 1  4  3  4  5  1
##   14 2 3 3 3 2 2 2 0  3  3  3  5  1
##   15 1 3 3 4 2 3 2 0  5  3  3  5  1
##   16 2 0 0 0 0 0 0 0  0  0  0  0  0
##   17 1 0 0 0 0 0 0 0  0  0  0  0  0
##   18 0 3 3 4 1 3 2 0  3  3  3  5  1
##   19 0 3 3 4 2 3 2 1  4  3  3  4  1
##   20 0 3 3 4 2 2 2 0  3  1  2  5  1
##   21 0 3 3 1 1 2 2 0  3  2  3  5  1
##   22 0 3 3 4 2 2 1 1  4  3  4  5  1
##   23 0 3 3 1 1 2 1 1  3  2  3  5  1
##   24 0 1 2 0 0 0 0 0  0  0  0  0  0
##   25 0 1 0 2 2 3 2 1  5  3  4  5  1
##   26 0 0 1 4 1 2 2 1  4  1  2  5  1
```

```
# interactions <- glm_interaction(object, c(1, 2))

# unique_cfd <- interactions[[1]] coeff_matrix <- interactions[[2]] pval_matrix
# <- interactions[[3]]

# interaction_indicator <- rep(0, nrow(object[['confounder']])) for(k in
# 1:nrow(unique_cfd)){ selected <- apply(confounder, 1, function(x) all(x ==
# unique_cfd[k,])) interaction_indicator[selected] <- k }

# intersted_idx <- which(apply(pval_matrix, 1, function(x) sum(x < 1e-8)) ==
# 19)

# confound_values <- (confounder[which(interaction_indicator ==
# intersted_idx[3]), ])

# stage_id <- r_names[confound_values[1]] tissue_id <-
# structure[confound_values[2], 1]

# metagene_id <- which.max(coeff_matrix[intersted_idx[3],])

# # which.min(coeff_matrix[intersted_idx[3],])


# 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])

# # # 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'))
```