Brainspan glm interaction analysis

ZHAO Kai

```
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"
```

```
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 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.

```
summary <- table(dic[, c(9, 11)])
colnames(summary) <- r_names
rownames(summary) <- structure[, 1]
print(summary)
```

```
##          Period
## sid       Early fetal_2 Early fetal_3 Early mid-fetal_4 Early mid-fetal_5
```

```
##   Ocx                2            0                 0                 0
##   M1C-S1C            2            0                 0                 2
##   AMY                2            3                 3                 3
##   MGE                2            0                 0                 0
##   STC                1            2                 2                 4
##   URL                2            0                 0                 0
##   CGE                2            0                 0                 0
##   DTH                2            3                 0                 0
##   MFC                2            2                 3                 4
##   DFC                2            3                 3                 4
##   OFC                2            3                 3                 3
##   LGE                2            0                 0                 0
##   ITC                1            3                 3                 2
##   HIP                2            3                 3                 3
##   VFC                1            3                 3                 4
##   PCx                2            0                 0                 0
##   TCx                1            0                 0                 0
##   A1C                0            3                 3                 4
##   V1C                0            3                 3                 4
##   STR                0            3                 3                 4
##   M1C                0            3                 3                 1
##   IPC                0            3                 3                 4
##   S1C                0            3                 3                 1
##   CB                 0            1                 2                 0
##   CBC                0            1                 0                 2
##   MD                 0            0                 1                 4
##          Period
## sid       Late mid-fetal_6 Late fetal_7 Neonatal and early infancy_8
##   Ocx                    0            0                           0
##   M1C-S1C                1            0                           0
##   AMY                    1            2                           3
##   MGE                    0            0                           0
##   STC                    2            3                           3
##   URL                    0            0                           0
##   CGE                    0            0                           0
##   DTH                    0            0                           0
##   MFC                    2            2                           2
##   DFC                    2            3                           2
##   OFC                    1            2                           2
##   LGE                    0            0                           0
##   ITC                    2            2                           3
##   HIP                    2            2                           2
##   VFC                    2            3                           2
##   PCx                    0            0                           0
##   TCx                    0            0                           0
##   A1C                    1            3                           2
##   V1C                    2            3                           2
##   STR                    2            2                           2
##   M1C                    1            2                           2
##   IPC                    2            2                           1
##   S1C                    1            2                           1
##   CB                     0            0                           0
##   CBC                    2            3                           2
##   MD                     1            2                           2
```

```
##          Period
## sid       Late infancy_9 Early childhood_10 Middle and late childhood_11
##   Ocx                  0                  0                           0
##   M1C-S1C              0                  0                           0
##   AMY                  0                  4                           3
##   MGE                  0                  0                           0
##   STC                  1                  5                           3
##   URL                  0                  0                           0
##   CGE                  0                  0                           0
##   DTH                  0                  0                           0
##   MFC                  1                  3                           3
##   DFC                  1                  4                           3
##   OFC                  1                  3                           2
##   LGE                  0                  0                           0
##   ITC                  1                  4                           3
##   HIP                  0                  3                           3
##   VFC                  0                  5                           3
##   PCx                  0                  0                           0
##   TCx                  0                  0                           0
##   A1C                  0                  3                           3
##   V1C                  1                  4                           3
##   STR                  0                  3                           1
##   M1C                  0                  3                           2
##   IPC                  1                  4                           3
##   S1C                  1                  3                           2
##   CB                   0                  0                           0
##   CBC                  1                  5                           3
##   MD                   1                  4                           1
##          Period
## sid       Adolescence_12 Young adulthood_13 Middle adulthood_14
##   Ocx                  0                  0                    0
##   M1C-S1C              0                  0                    0
##   AMY                  3                  5                    1
##   MGE                  0                  0                    0
##   STC                  4                  5                    1
##   URL                  0                  0                    0
##   CGE                  0                  0                    0
##   DTH                  0                  0                    0
##   MFC                  3                  5                    0
##   DFC                  3                  4                    1
##   OFC                  3                  5                    1
##   LGE                  0                  0                    0
##   ITC                  4                  5                    1
##   HIP                  3                  5                    1
##   VFC                  3                  5                    1
##   PCx                  0                  0                    0
##   TCx                  0                  0                    0
##   A1C                  3                  5                    1
##   V1C                  3                  4                    1
##   STR                  2                  5                    1
##   M1C                  3                  5                    1
##   IPC                  4                  5                    1
##   S1C                  3                  5                    1
##   CB                   0                  0                    0
```

```
##   CBC                 4                   5                   1
##   MD                  2                   5                   1
```

```
# interactions <- glm_interaction(object, c(1, 2))
load("~/data/multidimensional_datasets/brainspan_genes_matrix_csv/glm_interaction.RData")

unique_cfd <- interactions[[1]]
colnames(unique_cfd) <- c("period", "structure_id")

defination <- data.frame(stage_names = r_names[unique_cfd[, 1]], structure_names = structure[,
    1][unique_cfd[, 2]])

coeff_matrix <- interactions[[2]]
pval_matrix <- interactions[[3]]

significant_pvals <- apply(pval_matrix, 1, function(x) {
    sum(x <= 0.05)
})

cat("number of significant pvalues for each combination: ", significant_pvals, "\n")
```

```
## number of significant pvalues for each combination:  14 16 19 18 19 17 17 17 16 16 19 16 19 17 18 15
```

```
# 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],])
```