PrediXcan Data 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(ape)
  require(qgraph)
})
```

```
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)
}

split_str <- function(s){
    # l <- unlist(strsplit(s, split=c('-|_')))
    l <- unlist(strsplit(s, split=c('_')))
    idx <- which(l == 'v7')
    len <- length(l)

    disease <- l[1]
    tissue <- paste(l[(idx + 1): len], collapse = '_')
    c(disease, tissue)
}
```

```
load("~/data/Results/prediXcan/predXcan_l1l2_penalty_18.RData")
attach(fitted_obj) # attach it for easy syntax
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     column_factor, disease_factor, tissue_factor
```

```
## The following objects are masked from fitted_obj (pos = 3):
```

```
##
##     column_factor, disease_factor, iter, optimal_rmse,
##     tissue_factor, trainset
```

```
str(fitted_obj) # show the structure of our result
```

```
## List of 6
##  $ iter          : int 50000
##  $ trainset      : num [1:221, 1:10949] 0 0 -1.24 0 0 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:221] "AD-2018_gtex_v7_Brain_Amygdala" "AD-2018_gtex_v7_Brain_Anterior_cingulate_co
##   .. ..$ : chr [1:10949] "A2MP1" "A3GALT2" "A4GALT" "A4GNT" ...
##  $ disease_factor: num [1:17, 1:18] 2.5 -2.469 0.672 1.513 1.688 ...
##  $ tissue_factor : num [1:13, 1:18] 1.21 1.36 1.49 1.47 1.51 ...
##  $ column_factor : num [1:18, 1:10949] -0.07003 0.00297 0.00625 -0.09145 0.03756 ...
##  $ optimal_rmse  : num 0.786
```

```
meta_info <- t(unname(sapply(rownames(trainset), function(s) split_str(s))))

confounders <- as.data.frame(meta_info, stringsAsFactors = T)
colnames(confounders) <- c('disease', 'tissue')

rownames(disease_factor) <- levels(confounders[[1]])
rownames(tissue_factor) <- levels(confounders[[2]])
colnames(column_factor) <- colnames(trainset)

gene_symbol <- sapply(colnames(trainset), function(x) unlist(strsplit(x, ".", fixed = T))[1])
gene_symbol <- unname(gene_symbol)
mapping <- bitr(gene_symbol, fromType = "SYMBOL",
                toType = c("ENTREZID", "SYMBOL"),
                OrgDb = org.Hs.eg.db)
```

```
## 'select()' returned 1:many mapping between keys and columns
```

```
## Warning in bitr(gene_symbol, fromType = "SYMBOL", toType =
## c("ENTREZID", : 16.76% of input gene IDs are fail to map...
```

```
mapping <- mapping[mapping[[2]] != 100505381,]
rownames(mapping) <- mapping[, 1]
```

Explore the cluster between psychiatric disorders

The dendogram is used to show the closeness between different psychiatric disorders.

```
scores <- t(cor(t(disease_factor), t(disease_factor)))
disease_names <- rownames(disease_factor)

# select top 3 correlated disease as an example
result <- t(apply(scores, 1, function(x) disease_names[order(x, decreasing = T)][2:4]))
print(result)
```

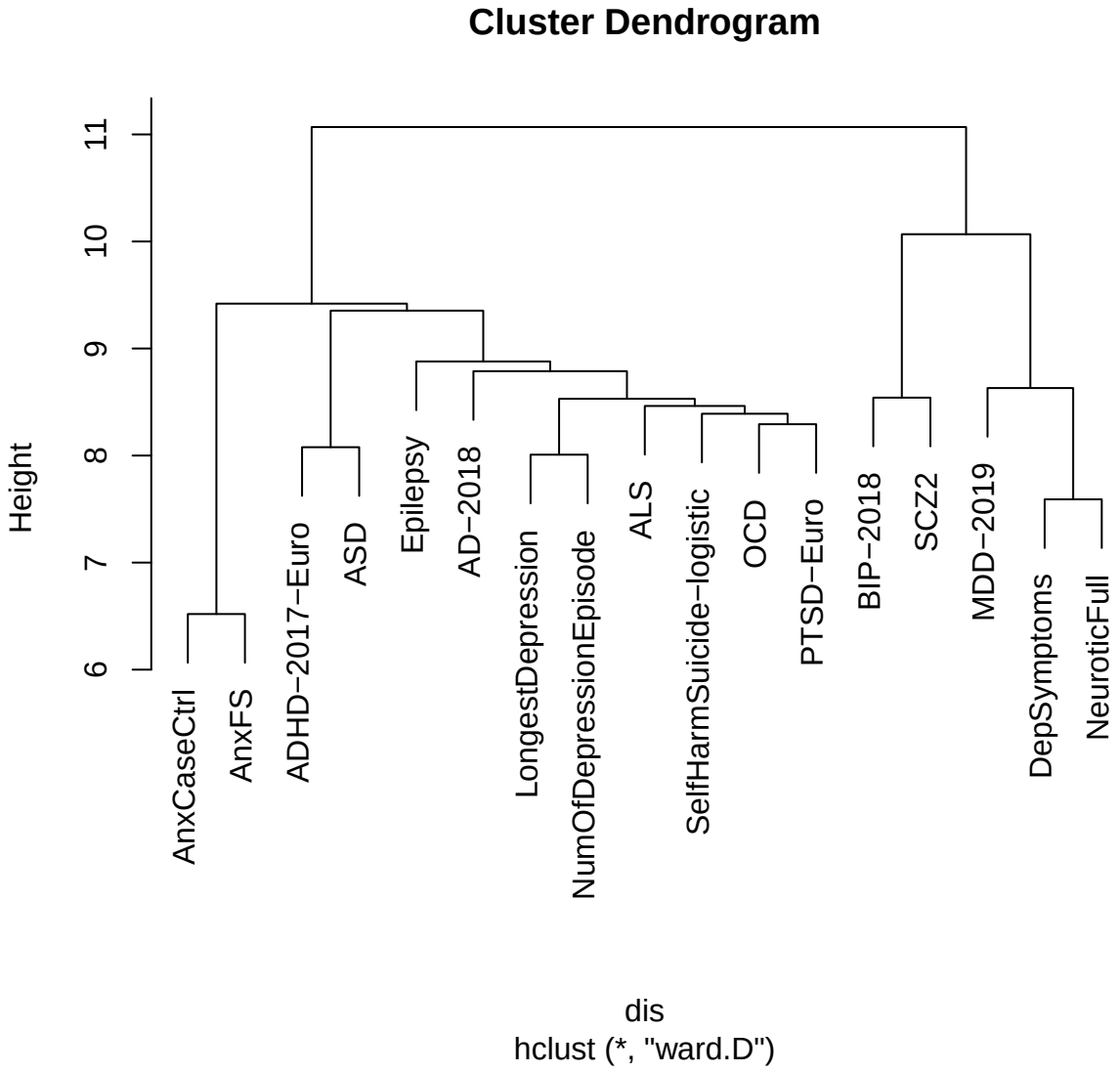
```
##                         [,1]                    [,2]
## AD-2018                 "ALS"                   "LongestDepression"
## ADHD-2017-Euro          "ASD"                   "MDD-2019"
## ALS                     "Epilepsy"              "AD-2018"
## AnxCaseCtrl             "AnxFS"                 "NeuroticFull"
```

```
## AnxFS                     "AnxCaseCtrl"             "NeuroticFull"
## ASD                       "ADHD-2017-Euro"          "SCZ2"
## BIP-2018                  "SCZ2"                    "MDD-2019"
## DepSymptoms               "NeuroticFull"            "MDD-2019"
## Epilepsy                  "ALS"                     "ADHD-2017-Euro"
## LongestDepression         "NumOfDepressionEpisode" "MDD-2019"
## MDD-2019                  "NeuroticFull"            "DepSymptoms"
## NeuroticFull              "DepSymptoms"             "MDD-2019"
## NumOfDepressionEpisode    "LongestDepression"       "NeuroticFull"
## OCD                       "SCZ2"                    "BIP-2018"
## PTSD-Euro                 "OCD"                     "ALS"
## SCZ2                      "BIP-2018"                "MDD-2019"
## SelfHarmSuicide-logistic "Epilepsy"                "PTSD-Euro"
##                           [,3]
## AD-2018                   "SelfHarmSuicide-logistic"
## ADHD-2017-Euro            "Epilepsy"
## ALS                       "PTSD-Euro"
## AnxCaseCtrl               "MDD-2019"
## AnxFS                     "DepSymptoms"
## ASD                       "MDD-2019"
## BIP-2018                  "OCD"
## DepSymptoms               "SCZ2"
## Epilepsy                  "SelfHarmSuicide-logistic"
## LongestDepression         "AD-2018"
## MDD-2019                  "SCZ2"
## NeuroticFull              "SCZ2"
## NumOfDepressionEpisode    "MDD-2019"
## OCD                       "PTSD-Euro"
## PTSD-Euro                 "SelfHarmSuicide-logistic"
## SCZ2                      "DepSymptoms"
## SelfHarmSuicide-logistic "AD-2018"
```

```
# use dendogram to visually show the relationship between between different diseases
dis <- dist(disease_factor, method = "euclidean") # distance matrix
fit <- hclust(dis, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit) # display dendogram
```



Explore association between different diseases

In this part, we explored the biogical processes (BPs) enriched by the pseudo expression profile of MDD. This analysis aims to discover the BPs involved in all brain regions included in our study.

However, there is no BPs found. In my opinion, MDD may affect only several not all brain regions, so the pattern related to MDD is covered by expression of other brain regions. Also, I tried other disease like SCZ, and the situation is also similar.

```
disease_matrix <- (disease_factor %*% column_factor)

# use MDD as an example
disease_id <- 16
cutoffs <- quantile(disease_matrix[disease_id,], probs = seq(0, 1, 0.025))
colnames(disease_matrix) <- colnames(trainset)

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

```
# cutoffs <- quantile(column_factor[metagene_id,], probs = seq(0, 1, 0.05))
# selected <- (column_factor[metagene_id,] >= cutoffs[length(cutoffs) - 1]) #

upreg <- enrichGO(gene = na.omit(mapping[gene_symbol[selected], 2]),
                  OrgDb = 'org.Hs.eg.db',
                  ont = "BP",
                  readable = TRUE)
if(nrow(upreg) > 0){
    dotplot(upreg, font = 9, showCategory=30) +
    scale_y_discrete(labels = function(x) str_wrap(x, width=60))
} else {
    cat("No BPs are enriched with the gene list!åå")
}
```

```
## No BPs are enriched with the gene list!åå
```

```
# 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 <- (disease_matrix[disease_id,] <= cutoffs[2])
downreg <- enrichGO(gene = na.omit(mapping[gene_symbol[selected], 2]),
                    OrgDb = 'org.Hs.eg.db',
                    ont = "BP",
                    readable = TRUE)
if(nrow(downreg) > 0){
    dotplot(downreg, font = 9, showCategory=30) +
    scale_y_discrete(labels = function(x) wrap_labal(x))
} else {
    cat("No BPs are enriched with the gene list!")
}
```

```
## No BPs are enriched with the gene list!
```

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

Explore the association between diseases and tissues and its mechanism

In the analyis below, we first explore the association between diseases and tissues. We list top three relevant brain regions for each disorder.

```
tissue_names <- rownames(tissue_factor)
scores <- cor(t(disease_factor), t(tissue_factor))

result <- t(apply(scores, 1, function(x) tissue_names[order(x, decreasing = T)][1:3]))
print(result)
```

```
##                          [,1]
## AD-2018                  "Brain_Caudate_basal_ganglia"
## ADHD-2017-Euro           "Brain_Putamen_basal_ganglia"
## ALS                      "Brain_Caudate_basal_ganglia"
## AnxCaseCtrl              "Brain_Cerebellum"
## AnxFS                    "Brain_Frontal_Cortex_BA9"
## ASD                      "Brain_Caudate_basal_ganglia"
```

```
## BIP-2018                 "Brain_Hypothalamus"
## DepSymptoms              "Brain_Spinal_cord_cervical_c-1"
## Epilepsy                 "Brain_Frontal_Cortex_BA9"
## LongestDepression        "Brain_Amygdala"
## MDD-2019                 "Brain_Putamen_basal_ganglia"
## NeuroticFull             "Brain_Spinal_cord_cervical_c-1"
## NumOfDepressionEpisode   "Brain_Hippocampus"
## OCD                      "Brain_Caudate_basal_ganglia"
## PTSD-Euro                "Brain_Hypothalamus"
## SCZ2                     "Brain_Substantia_nigra"
## SelfHarmSuicide-logistic "Brain_Putamen_basal_ganglia"
##                          [,2]
## AD-2018                  "Brain_Frontal_Cortex_BA9"
## ADHD-2017-Euro           "Brain_Frontal_Cortex_BA9"
## ALS                      "Brain_Frontal_Cortex_BA9"
## AnxCaseCtrl              "Brain_Substantia_nigra"
## AnxFS                    "Brain_Caudate_basal_ganglia"
## ASD                      "Brain_Amygdala"
## BIP-2018                 "Brain_Spinal_cord_cervical_c-1"
## DepSymptoms              "Brain_Substantia_nigra"
## Epilepsy                 "Brain_Substantia_nigra"
## LongestDepression        "Brain_Caudate_basal_ganglia"
## MDD-2019                 "Brain_Substantia_nigra"
## NeuroticFull             "Brain_Hippocampus"
## NumOfDepressionEpisode   "Brain_Caudate_basal_ganglia"
## OCD                      "Brain_Cerebellum"
## PTSD-Euro                "Brain_Cerebellum"
## SCZ2                     "Brain_Spinal_cord_cervical_c-1"
## SelfHarmSuicide-logistic "Brain_Cortex"
##                          [,3]
## AD-2018                  "Brain_Putamen_basal_ganglia"
## ADHD-2017-Euro           "Brain_Anterior_cingulate_cortex_BA24"
## ALS                      "Brain_Anterior_cingulate_cortex_BA24"
## AnxCaseCtrl              "Brain_Cerebellar_Hemisphere"
## AnxFS                    "Brain_Putamen_basal_ganglia"
## ASD                      "Brain_Hypothalamus"
## BIP-2018                 "Brain_Substantia_nigra"
## DepSymptoms              "Brain_Cerebellum"
## Epilepsy                 "Brain_Spinal_cord_cervical_c-1"
## LongestDepression        "Brain_Hippocampus"
## MDD-2019                 "Brain_Hypothalamus"
## NeuroticFull             "Brain_Cerebellar_Hemisphere"
## NumOfDepressionEpisode   "Brain_Putamen_basal_ganglia"
## OCD                      "Brain_Spinal_cord_cervical_c-1"
## PTSD-Euro                "Brain_Spinal_cord_cervical_c-1"
## SCZ2                     "Brain_Hypothalamus"
## SelfHarmSuicide-logistic "Brain_Amygdala"
```

Then, we try to reveal the BPs that enriched by the interaction between substantia nigra and depression symptoms. Here we are interested in the down regulated BPs.

```
disease_name <- "DepSymptoms"
tissue_name <- "Brain_Substantia_nigra"

interaction <- disease_factor[disease_name, ] * tissue_factor[tissue_name,]
```

```r
profile <- interaction %*% column_factor
cutoffs <- quantile(profile, probs = seq(0, 1, 0.025))
# selected <- (profile >= cutoffs[length(cutoffs)-1])
selected <- (profile <= cutoffs[2])

downreg <- enrichGO(gene = na.omit(mapping[gene_symbol[selected], 2]),
                    OrgDb = 'org.Hs.eg.db',
                    ont = "BP",
                    readable = TRUE)

if(nrow(downreg) > 0){
    dotplot(downreg, font = 9, showCategory=30) +
    scale_y_discrete(labels = function(x) wrap_labal(x))
} else {
    cat("No BPs are enriched with the gene list!")
}
```

```
## No BPs are enriched with the gene list!
```