Clustering Analysis from RNAseq data (Heaney et al. 2021)

Sanjeev V Namjoshi (snamjoshi87@utexas.edu)

February 23rd, 2017 (revised June 15th, 2019)

This file contains all the data used for clustering the filtered count data.

Load packages

```
library(ggplot2)
library(reshape2)
library(magrittr)
library(RDAVIDWebService)
library(org.Mm.eg.db)
```

Load functions

The `functionalAnnotationChart()` function a list of genes, and ontology (term), and an output file name. It returns the Functional Annotation Chart from DAVID.

```
functionalAnnotationChart <- function(genes, term, filename) {

  require(RDAVIDWebService)
  require(org.Mm.eg.db)

  david <- DAVIDWebService$new(email = "<REDACTED>", url = "https://david.ncifcrf.gov/webservice/servic
  egenes <- select(org.Mm.eg.db, genes, "ENTREZID", "SYMBOL")[,2]

  chart <- addList(david, egenes,
                   idType = "ENTREZ_GENE_ID",
                   listName = "Functional Annotation Chart",
                   listType = "Gene")

  setAnnotationCategories(david, term)
  getFunctionalAnnotationChartFile(david, fileName = filename)

  return(paste("Ouput DAVID analysis file:", filename))
}
```

The `uniqueFilteredGoTerms()` function a list of downregulated genes and a list of upregulated genes from DAVID results, and a desired PValue and FDR cutoff. It first applies these cutoffs to both data sets. Then, it finds the unique GO terms between the remaining terms.

```
uniqueFilteredGoTerms <- function(downfileName, upfileName, PValue, FDR) {
  downBP <- read.table(downfileName, sep = "\t", header = TRUE, stringsAsFactors = FALSE, quote = "")
  downBP <- downBP[ ,c("Term", "PValue", "FDR")]
```

```
  downBP <- downBP[downBP$PValue < PValue | downBP$FDR < FDR, ]

  upBP <- read.table(upfileName, sep = "\t", header = TRUE, stringsAsFactors = FALSE, quote = "")
  upBP <- upBP[ , c("Term", "PValue", "FDR")]
  upBP <- upBP[upBP$PValue < PValue | upBP$FDR < FDR, ]

  sharedBP <- intersect(downBP$Term, upBP$Term)
  downOnlyBP <- setdiff(downBP$Term, upBP$Term)
  upOnlyBP <- setdiff(upBP$Term, downBP$Term)

  downOnlyTableBP <- downBP[downBP$Term %in% downOnlyBP, c("Term", "PValue", "FDR")]
  upOnlyTableBP <- upBP[upBP$Term %in% upOnlyBP, c("Term", "PValue", "FDR")]

  downOnlyTableBP <- downOnlyTableBP[order(downOnlyTableBP$PValue), "Term"]
  upOnlyTableBP <- upOnlyTableBP[order(upOnlyTableBP$PValue), "Term"]

  return(list(downTable = downOnlyTableBP, upTable = upOnlyTableBP))
}
```

Load data and filter

```
fc <- read.csv("finalData.csv", header = TRUE, stringsAsFactors = FALSE)
row.names(fc) <- fc$Row.names
fc[ ,1] <- NULL
fc <- fc[ ,c("logFC.x", "logFC.y")]
names(fc) <- c("RO", "RO_RAP")

counts <- read.csv("finalCountData.csv", header = TRUE, stringsAsFactors = FALSE, row.names = 1)
countsAvg <- data.frame(row.names = row.names(counts),
                        SAL_WT = rowMeans(counts[ ,1:3]),
                        RO_WT = rowMeans(counts[ ,7:9]),
                        RO_RAP_WT = rowMeans(counts[ ,11:13]))
countsAvg.log <- apply(countsAvg, 2, log2) %>% as.data.frame()

### Here we apply a fold-change filter
roDownGenes <- fc[fc$RO < quantile(fc$RO, 0.25), ]
roUpGenes <- fc[fc$RO > quantile(fc$RO, 0.75), ]

roRapDownGenes <- fc[fc$RO_RAP < quantile(fc$RO_RAP, 0.25), ]
roRapUpGenes <- fc[fc$RO_RAP > quantile(fc$RO_RAP, 0.75), ]
```

Unique GO clustering by fold-change group

Code to generate DAVID Functional Annotation Charts for Ro and RoRap. For all DAVID charts, we used a p-value cutoff of 0.001 and FDR cutoff of 0.05.

Note that in the event that DAVID servers are not available or something about their data processing has changed, the following files generated below are already available in this directory:

- `david_ro_down_1000_biological_process.txt`
- `david_ro_up_1000_biological_process.txt`
- `david_roRap_down_1000_biological_process.txt`

- david_roRap_up_1000_biological_process.txt

```
### Query DAVID for Ro and get unique GO terms
functionalAnnotationChart(roDownGenes,
                          term = "GOTERM_BP_ALL",
                          filename = "david_ro_down_1000_biological_process.txt")

functionalAnnotationChart(roUpGenes,
                          term = "GOTERM_BP_ALL",
                          filename = "david_ro_up_1000_biological_process.txt")

uniqueBpRo <- uniqueFilteredGoTerms(downfileName = "david_ro_down_1000_biological_process.txt",
                                    upfileName = "david_ro_up_1000_biological_process.txt",
                                    PValue = 0.001,
                                    FDR = 0.05)

### Query DAVID for Ro+Rap and get unique GO terms
functionalAnnotationChart(roRapDownGenes,
                          term = "GOTERM_BP_ALL",
                          filename = "david_roRap_down_1000_biological_process.txt")

functionalAnnotationChart(roRapUpGenes,
                          term = "GOTERM_BP_ALL",
                          filename = "david_roRap_up_1000_biological_process.txt")

uniqueBpRoRap <- uniqueFilteredGoTerms(downfileName = "david_roRap_down_1000_biological_process.txt",
                                       upfileName = "david_roRap_up_1000_biological_process.txt",
                                       PValue = 0.001,
                                       FDR = 0.05)
```

Session info:

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.1 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8           LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8          LC_NAME=en_US.UTF-8
##  [9] LC_ADDRESS=en_US.UTF-8        LC_TELEPHONE=en_US.UTF-8
## [11] LC_MEASUREMENT=en_US.UTF-8    LC_IDENTIFICATION=en_US.UTF-8
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] amap_0.8-14             gdata_2.17.0
##  [3] dplyr_0.5.0             Ckmeans.1d.dp_3.4.6-4
##  [5] factoextra_1.0.3        xlsx_0.5.7
##  [7] xlsxjars_0.6.1          rJava_0.9-8
```

```
##  [9] gridExtra_2.2.1         org.Mm.eg.db_3.4.0
## [11] RDAVIDWebService_1.12.0 GOstats_2.40.0
## [13] Category_2.40.0         Matrix_1.2-7.1
## [15] AnnotationDbi_1.36.0    IRanges_2.8.1
## [17] S4Vectors_0.12.0        Biobase_2.34.0
## [19] graph_1.52.0            BiocGenerics_0.20.0
## [21] magrittr_1.5            reshape2_1.4.2
## [23] ggplot2_2.2.0
##
## loaded via a namespace (and not attached):
##  [1] mclust_5.2             ggrepel_0.6.5          Rcpp_0.12.8
##  [4] mvtnorm_1.0-5          lattice_0.20-34        GO.db_3.4.0
##  [7] class_7.3-14           gtools_3.5.0           assertthat_0.1
## [10] rprojroot_1.1          digest_0.6.10          R6_2.2.0
## [13] plyr_1.8.4             backports_1.0.4        RSQLite_1.1
## [16] evaluate_0.10          diptest_0.75-7         lazyeval_0.2.0
## [19] annotate_1.52.0        kernlab_0.9-25         whisker_0.3-2
## [22] rmarkdown_1.2          labeling_0.3           splines_3.3.2
## [25] stringr_1.1.0          RCurl_1.95-4.8         munsell_0.4.3
## [28] htmltools_0.3.5        nnet_7.3-12            tibble_1.2
## [31] dendextend_1.3.0       XML_3.98-1.5           AnnotationForge_1.16.0
## [34] MASS_7.3-45            bitops_1.0-6           grid_3.3.2
## [37] RBGL_1.50.0            xtable_1.8-2           GSEABase_1.36.0
## [40] gtable_0.2.0           DBI_0.5-1              scales_0.4.1
## [43] stringi_1.1.2          genefilter_1.56.0      flexmix_2.3-13
## [46] robustbase_0.92-6      tools_3.3.2            fpc_2.1-10
## [49] trimcluster_0.1-2      DEoptimR_1.0-8         survival_2.40-1
## [52] yaml_2.1.14            colorspace_1.3-1       cluster_2.0.5
## [55] prabclus_2.2-6         memoise_1.0.0          knitr_1.15.1
## [58] modeltools_0.2-21
```