Gene set enrichment analysis (GSEA)

J. Shah chimeric mouse collaboration

Kim Dill-McFarland, kadm@uw.edu

version May 13, 2020

Contents

Background

The purpose of this workflow is to determine which Broad Molecular Signature Database (MSigDB) terms are enriched in significant and modules from WCGNA analysis.

Setup

Load packages

```
# Data manipulation and figures
library(tidyverse)

#Print pretty tables to Rmd
library(knitr)
library(kableExtra)

`%notin%` <- Negate(`%in%`)
```

Set seed

```
set.seed(4389)
```

Load data

```
#Gene results
Shah_contrast_gene_pval <-
  read_csv("results/gene_level/Shah_contrast_gene_pval.csv")
#Module results
mods.net <- read_csv("results/module_Shah_contrast_deepSplit3_minMod50/Shah_contrast_genes_in_mod.csv")

#Genome
library(org.Mm.eg.db)
#Script for running GSEA
source("scripts/GSEA_enricher.R")
```

Gene set enrichment analysis (GSEA)

Gene ontology (GO) provides gene functions and annotations from a variety of sources. Specifics on the gene sets below can be found at http://software.broadinstitute.org/gsea/msigdb/collections.jsp.

Gene set descriptions

- Hallmark gene sets (H)
  - Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression.
- Basic gene sets (C5)
  - Gene sets that contain genes annotated by the same GO term. Includes:
  - BP: biological process
  - CC: cellular component
  - MF: molecular function
- Curated gene sets (C2)
  - Gene sets curated from various sources including online pathway databases, the biomedical literature, and knowledge of domain experts. Includes:
  - CGP: chemical and genetic perturbations
  - CP: Canonical pathways
    - BIOCARTA: BioCarta gene sets
    - KEGG: KEGG gene sets
    - PID: PID gene sets
    - REACTOME: Reactome gene sets
- Immunologic signatures (C7)
  - Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology.

GSEA gene-level

List genes with FDR < 0.05

```
gene.list <- Shah_contrast_gene_pval %>%
  filter(adj.P.Val <= 0.05) %>%
  dplyr::select(geneName) %>% unlist(use.names = FALSE)
```

```
#Run custom function on genes
enrich.fxn(gene.list = gene.list,
           category = "H", genome = org.Mm.eg.db,
           basename="contrast_signif_genes")

#### Other gene sets not run ####
#enrich.fxn(gene.list = gene.list,
#           category = "C5", genome = org.Mm.eg.db,
#           basename="contrast_signif_genes")

#enrich.fxn(gene.list = gene.list,
#           category = "C2", subcategory = "CGP", genome = org.Mm.eg.db,
#           basename="contrast_signif_genes")

#enrich.fxn(gene.list = gene.list,
#           category = "C2", subcategory = "CP:KEGG", genome = org.Mm.eg.db,
#           basename="contrast_signif_genes")

#enrich.fxn(gene.list = gene.list,
#           category = "C7", genome = org.Mm.eg.db,
#           basename="contrast_signif_genes")
```

GSEA gene-level summary

Significant enrichments, FDR ≤ 0.2.

Description	Overlap genes	FDR
HALLMARK_ALLOGRAFT_REJECTION	12	0.067906

GSEA module-level

List genes in each module.

```
gsea.temp <- data.frame()

for(i in 0:max(mods.net$module)){
  gene.list <- mods.net %>%
    filter(module == i) %>%
    dplyr::select(geneName) %>% unlist(use.names = FALSE)

  #Run custom function on genes
  enrich.fxn(gene.list = gene.list,
           category = "H", genome = org.Mm.eg.db,
           basename="module_genes")

  #### Other gene sets not run ####
  #enrich.fxn(gene.list = gene.list,
  #         category = "C5", genome = org.Mm.eg.db,
  #         basename="module_genes")

  #enrich.fxn(gene.list = gene.list,
  #         category = "C2", subcategory = "CGP", genome = org.Mm.eg.db,
  #         basename="module_genes")

  #enrich.fxn(gene.list = gene.list,
  #         category = "C2", subcategory = "CP:KEGG",
  #         genome = org.Mm.eg.db,
  #         basename="module_genes")

  #enrich.fxn(gene.list = gene.list,
  #         category = "C7", genome = org.Mm.eg.db,
  #         basename="module_genes")

  for(file in list.files(path="results/GSEA/",
                         pattern="module_genes",
                         full.names = TRUE)){
    gsea.temp <- bind_rows(gsea.temp, read_csv(file))

  }

  #Add module ID column
  if("module" %in% colnames(gsea.temp)){
    gsea.temp <- gsea.temp %>%
                 mutate(module = ifelse(is.na(module), i, module))
  } else{
    gsea.temp <- gsea.temp %>%
                 mutate(module = i)
  }

}

#Save
for(db in unique(gsea.temp$category)){
  if(db != "C2"){
    gsea.sub <- gsea.temp %>%
                filter(category == db)

    filename <- paste("results/GSEA/GSEA_module_genes_",
                      db, ".csv", sep="")
    write_csv(gsea.sub, filename)

  } else{
    gsea.sub <- gsea.temp %>%
                filter(category == db)
    for(db.sub in unique(gsea.sub$subcategory)){
      gsea.sub <- gsea.sub %>%
                filter(subcategory == db.sub)

      filename <- paste("results/GSEA/GSEA_module_genes_",
                      db,"_", db.sub, ".csv", sep="") %>%
        gsub(":",".",.)
      write_csv(gsea.sub, filename)
      }
  }
}
```

GSEA module-level summary

Significant enrichments, FDR ≤ 0.2.

Module	Description	Overlap genes	FDR
1	HALLMARK_OXIDATIVE_PHOSPHORYLATION	31	0.0000000
	HALLMARK_MYC_TARGETS_V1	23	0.0002332
	HALLMARK_ADIPOGENESIS	17	0.0439996
	HALLMARK_MTORC1_SIGNALING	17	0.0439996
	HALLMARK_FATTY_ACID_METABOLISM	15	0.0439996
2	HALLMARK_MITOTIC_SPINDLE	27	0.0000000
3	HALLMARK_OXIDATIVE_PHOSPHORYLATION	37	0.0000000
	HALLMARK_MYC_TARGETS_V1	23	0.0000143
	HALLMARK_DNA_REPAIR	15	0.0047204
	HALLMARK_INTERFERON_ALPHA_RESPONSE	10	0.0411901
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	9	0.1875940
4	HALLMARK_PROTEIN_SECRETION	7	0.0464935
	HALLMARK_PI3K_AKT_MTOR_SIGNALING	7	0.0464935
	HALLMARK_UNFOLDED_PROTEIN_RESPONSE	7	0.0470334
5	HALLMARK_KRAS_SIGNALING_UP	18	0.0000879
	HALLMARK_INFLAMMATORY_RESPONSE	11	0.1577737
	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	11	0.1577737
9	HALLMARK_INTERFERON_GAMMA_RESPONSE	8	0.0042739
11	HALLMARK_ALLOGRAFT_REJECTION	7	0.0802182
13	HALLMARK_APOPTOSIS	4	0.1593613
14	HALLMARK_COMPLEMENT	4	0.0744725
	HALLMARK_TNFA_SIGNALING_VIA_NFKB	4	0.0744725
	HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	2	0.1190851
	HALLMARK_ADIPOGENESIS	3	0.1521123
	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	3	0.1521123
	HALLMARK_MTORC1_SIGNALING	3	0.1521123
	HALLMARK_IL6_JAK_STAT3_SIGNALING	2	0.1521123
15	HALLMARK_ESTROGEN_RESPONSE_LATE	5	0.0135542
	HALLMARK_ESTROGEN_RESPONSE_EARLY	4	0.0541025
16	HALLMARK_DNA_REPAIR	4	0.0548914
	HALLMARK_INTERFERON_ALPHA_RESPONSE	3	0.0768660
17	HALLMARK_TNFA_SIGNALING_VIA_NFKB	9	0.0000220
	HALLMARK_INTERFERON_GAMMA_RESPONSE	7	0.0016156
	HALLMARK_INFLAMMATORY_RESPONSE	6	0.0066402
	HALLMARK_IL2_STAT5_SIGNALING	5	0.0339295
	HALLMARK_COMPLEMENT	4	0.1371110

GSEA module DE groups

Groups are:

1. Up in uninfected only
2. Down in uninfected only
3. Up in infected only
4. Down in infected only
5. Up in both
6. Down in both

Group genes in module 0

Group genes with FDR < 0.3 in 6 module groups of interest.

```
mod_0 <- Shah_contrast_gene_pval %>%
  filter(geneName %in% filter(mods.net, module == 0)$geneName &
           adj.P.Val <= 0.3)

mod_0_UI_up <- mod_0 %>%
  filter(group == "uninfected" & FC.group == "up") %>%
  dplyr::select(geneName) %>%
  distinct(geneName) %>%  unlist(use.names = FALSE)
mod_0_UI_down <- mod_0 %>%
  filter(group == "uninfected" & FC.group == "down") %>%
  dplyr::select(geneName) %>%
  distinct(geneName) %>%  unlist(use.names = FALSE)

mod_0_I_up <- mod_0 %>%
  filter(group == "infected" & FC.group == "up") %>%
  dplyr::select(geneName) %>%
  distinct(geneName) %>%  unlist(use.names = FALSE)
mod_0_I_down <- mod_0 %>%
  filter(group == "infected" & FC.group == "down") %>%
  dplyr::select(geneName) %>%
  distinct(geneName) %>%  unlist(use.names = FALSE)

mod_0_both_up <- intersect(mod_0_UI_up, mod_0_I_up)
mod_0_both_down <- intersect(mod_0_UI_down, mod_0_I_down)

mod_0_UI_up2 <- mod_0_UI_up[mod_0_UI_up %notin%
                          c(mod_0_UI_down,
                            mod_0_I_up, mod_0_I_down)]
mod_0_UI_down2 <- mod_0_UI_down[mod_0_UI_down %notin%
                           c(mod_0_UI_up,
                            mod_0_I_up, mod_0_I_down)]

mod_0_I_up2 <- mod_0_I_up[mod_0_I_up %notin%
                           c(mod_0_UI_up, mod_0_UI_down,
                            mod_0_I_down)]
mod_0_I_down2 <- mod_0_I_down[mod_0_I_down %notin%
                            c(mod_0_UI_up, mod_0_UI_down,
                            mod_0_I_up)]
```

Sort genes that have different fold change directions in uninfected vs. infected. If one infection group is significant at FDR < 0.1, use that for grouping.

```
#Check genes not in a group
all <- c(mod_0_UI_up, mod_0_UI_down, mod_0_I_up, mod_0_I_down)
all2 <- c(mod_0_both_up, mod_0_both_down,
          mod_0_UI_up2, mod_0_UI_down2,
          mod_0_I_up2, mod_0_I_down2)

missing <- mod_0 %>%
  filter(geneName %in% all[all %notin% all2]) %>%
  arrange(geneName, adj.P.Val)

#Sort my hand
mod_0_UI_up <- c(mod_0_UI_up2)
mod_0_UI_down <- c(mod_0_UI_down2)
mod_0_I_up <- c(mod_0_I_up2, "ENSMUSG00000024935","ENSMUSG00000027534")
```

```
mod_0_I_down <- c(mod_0_I_down2)
```

These leaves 8 module 0 genes which were not grouped into a module FC group because they showed similarly significant but different directions in uninfected and infected groups.

```
missing %>%
  filter(geneName %notin% c("ENSMUSG00000024935",
                            "ENSMUSG00000027534")) %>%
  dplyr::select(geneName, mgi_symbol, adj.P.Val, group, FC.group) %>%

 kable() %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE) %>%
  collapse_rows(columns = 1:2, valign = "top")
```

geneName	mgi_symbol	adj.P.Val	group	FC.group
ENSMUSG00000007646	Rad51c	0.2078087	infected	up
		0.2752807	uninfected	down
ENSMUSG00000020268	Lyrm7	0.1131500	infected	up
		0.1459106	uninfected	down
ENSMUSG00000027570	Col9a3	0.2141066	uninfected	down
		0.2998444	infected	up
ENSMUSG00000044005	Gls2	0.1811460	infected	down
		0.2133073	uninfected	up
ENSMUSG00000054716	Zfp771	0.2283433	infected	up
		0.2772522	uninfected	down
ENSMUSG00000090121	Abhd12b	0.1656325	infected	up
		0.2645692	uninfected	down
ENSMUSG00000092260	Zfp963	0.1865873	infected	up
		0.2736297	uninfected	down
ENSMUSG00000100937	1700020D05Rik	0.1450344	infected	up
		0.2386485	uninfected	down

Run GSEA

```
file.remove("results/GSEA/GSEA_module_groups_H.csv", showWarnings=FALSE)
#Add groups to modules
mods.net.groups <- mods.net %>%
  mutate(mod.group = ifelse(module %in% c(14,7,16,9), "both_up",
                    ifelse(module %in% c(5,11,17,10,2), "both_down",
                     ifelse(module %in% c(13,15), "UI_up",
                      ifelse(module %in% c(6,12), "UI_down",
                        ifelse(module %in% c(1,3), "I_up",
                          ifelse(module %in% c(4,8), "I_down",
                                 NA))))))) %>%
  filter(module != 0)

for(mod.group.name in c("both_up", "both_down",
              "UI_up", "UI_down",
              "I_up", "I_down")){

  from_mods <- mods.net.groups %>%
    filter(mod.group == mod.group.name) %>%
    dplyr::select(geneName) %>%
    distinct() %>% unlist(use.names = FALSE)

  from_mod0 <- get(paste("mod_0", mod.group.name, sep="_"))

  gene.list.all <- unique(c(from_mods, from_mod0))

  #Run custom function on genes
  enrich.fxn(gene.list = gene.list.all,
           category = "H", genome = org.Mm.eg.db,
           basename=paste(mod.group.name, length(gene.list.all), sep="_"))
}

#Combine results
files <- list.files(path="results/GSEA/",
                         pattern="down|up",
                         full.names = TRUE)

gsea.result3 <- data.frame()
for(i in 1:length(files)){
  group.name <- gsub("results/GSEA//GSEA_", "", files[i])
  group.name <- gsub("_H.csv", "", group.name)
  group.name <- gsub("UI_", "uninfected_", group.name)
  group.name <- gsub("I_", "infected_", group.name)

  gsea.temp <- read_csv(files[i]) %>%
    #Add module group name
    mutate(mod.group = group.name) %>%
    #Get number of genes from name
    separate(mod.group, into=c("mod.group","size.mod.group"),
             sep="_(?=[^_]+$)") %>%
    #Extract values from ratios
    separate(BgRatio, into=c("size.term","size.category"), sep="/") %>%
    separate(GeneRatio, into=c("size.overlap.term","size.overlap.category"),
             sep="/") %>%
    mutate_at(vars("size.term","size.category",
                   "size.overlap.term","size.overlap.category"),
                   as.numeric) %>%
    #Calculate k/K
    mutate("k/K"=size.overlap.term/size.term) %>%

    #Reorder variables
    dplyr::select(category, mod.group,
                  size.mod.group, size.overlap.category, size.category,
                  Description, size.overlap.term, size.term, `k/K`,
                  p.adjust, ENTREZIDs:ENSEMBLIDs) %>%
    arrange(p.adjust)

 gsea.result3 <- bind_rows(gsea.result3, gsea.temp)
}

#Save to disk
write_csv(gsea.result3, "results/GSEA/GSEA_module_groups_H.csv")
```

```r
file.remove(files, showWarnings=FALSE)
```

GSEA module group summary

Significant enrichments, FDR ≤ 0.2.

Module group	Genes in module group	Description	Genes in overlap (k)
both_down	1071	HALLMARK_MITOTIC_SPINDLE	38
		HALLMARK_INFLAMMATORY_RESPONSE	28
		HALLMARK_HYPOXIA	28
		HALLMARK_KRAS_SIGNALING_UP	27
		HALLMARK_ALLOGRAFT_REJECTION	26
both_up	517	HALLMARK_TNFA_SIGNALING_VIA_NFKB	14
		HALLMARK_XENOBIOTIC_METABOLISM	14
		HALLMARK_INTERFERON_GAMMA_RESPONSE	14
		HALLMARK_TGF_BETA_SIGNALING	6
infected_up	1199	HALLMARK_OXIDATIVE_PHOSPHORYLATION	71
		HALLMARK_MYC_TARGETS_V1	48
		HALLMARK_ADIPOGENESIS	32
		HALLMARK_INTERFERON_ALPHA_RESPONSE	19
		HALLMARK_DNA_REPAIR	25
		HALLMARK_INTERFERON_GAMMA_RESPONSE	28
		HALLMARK_MTORC1_SIGNALING	26

R session

```r
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.8.2     msigdbr_7.0.1          clusterProfiler_3.12.0
##  [4] org.Mm.eg.db_3.8.2     AnnotationDbi_1.46.1   IRanges_2.18.3
##  [7] S4Vectors_0.22.1       Biobase_2.44.0         BiocGenerics_0.30.0
## [10] kableExtra_1.1.0       knitr_1.28             forcats_0.5.0
## [13] stringr_1.4.0          dplyr_0.8.5            purrr_0.3.4
## [16] readr_1.3.1            tidyr_1.0.3            tibble_3.0.1
## [19] ggplot2_3.3.0          tidyverse_1.3.0
##
```

```
## loaded via a namespace (and not attached):
##  [1] fgsea_1.10.1       colorspace_1.4-1    ggridges_0.5.2
##  [4] ellipsis_0.3.0     qvalue_2.16.0       fs_1.4.1
##  [7] rstudioapi_0.11    farver_2.0.3        urltools_1.7.3
## [10] graphlayouts_0.7.0 ggrepel_0.8.2       bit64_0.9-7
## [13] fansi_0.4.1        lubridate_1.7.8     xml2_1.3.2
## [16] splines_3.6.1      GOSemSim_2.10.0     polyclip_1.10-0
## [19] jsonlite_1.6.1     broom_0.5.6         GO.db_3.8.2
## [22] dbplyr_1.4.3       ggforce_0.3.1       BiocManager_1.30.10
## [25] compiler_3.6.1     httr_1.4.1          rvcheck_0.1.8
## [28] backports_1.1.6    assertthat_0.2.1    Matrix_1.2-18
## [31] cli_2.0.2          tweenr_1.0.1        htmltools_0.4.0
## [34] prettyunits_1.1.1  tools_3.6.1         igraph_1.2.5
## [37] gtable_0.3.0       glue_1.4.0          reshape2_1.4.4
## [40] DO.db_2.9          fastmatch_1.1-0     Rcpp_1.0.4.6
## [43] enrichplot_1.4.0   cellranger_1.1.0    vctrs_0.2.4
## [46] nlme_3.1-147       ggraph_2.0.2        xfun_0.13
## [49] rvest_0.3.5        lifecycle_0.2.0     DOSE_3.10.2
## [52] europepmc_0.3      MASS_7.3-51.6       scales_1.1.0
## [55] tidygraph_1.1.2    hms_0.5.3           RColorBrewer_1.1-2
## [58] yaml_2.2.1         memoise_1.1.0       gridExtra_2.3
## [61] UpSetR_1.4.0       triebeard_0.3.0     stringi_1.4.6
## [64] RSQLite_2.2.0      BiocParallel_1.18.1 rlang_0.4.6
## [67] pkgconfig_2.0.3    evaluate_0.14       lattice_0.20-41
## [70] cowplot_1.0.0      bit_1.1-15.2        tidyselect_1.0.0
## [73] plyr_1.8.6         magrittr_1.5        R6_2.4.1
## [76] generics_0.0.2     DBI_1.1.0           pillar_1.4.4
## [79] haven_2.2.0        withr_2.2.0         modelr_0.1.7
## [82] crayon_1.3.4       rmarkdown_2.1       viridis_0.5.1
## [85] progress_1.2.2     grid_3.6.1          readxl_1.3.1
## [88] data.table_1.12.8  blob_1.2.1          reprex_0.3.0
## [91] digest_0.6.25      webshot_0.5.2       gridGraphics_0.5-0
## [94] munsell_0.5.0      viridisLite_0.3.0   ggplotify_0.0.5
```