Gene set enrichment analysis

Kim Dill-McFarland, kadm@uw.edu

version July 13, 2021

Contents

Overview

In this workshop, we introduce gene set analysis relevant to RNA-sequencing data. In it, we cover:

- Broad Molecular Signatures Database (MSigDB) gene sets
- hypergeometric enrichment with `clusterProfiler`
- gene set enrichment analysis (GSEA) with `fgsea`

During the workshop, we will build an R script together, which will be posted as 'live_notes' after the workshop here.

Prior to the workshop

Please install R, RStudio, and the following packages. See the setup instructions for more details.

```
#Data manipulation
library(tidyverse)
#Broad gene set database
library(msigdbr)
#Hypergeo enrichment
library(clusterProfiler)
```

```
#GSEA
library(fgsea)
```

Load data

Briefly, these data are from RNA-sequencing of human monocytes alone (`MEDIA`) or infected with *M. tuberculosis* in vitro (`TB`). Expression data are in an EList object containing expression (`E`), sample/patient metadata (`targets`), and gene metadata (`genes`).

```
#RNAseq expression data
load("data/RSTR_data_clean_subset.RData")
class(dat)
```

```
## [1] "EList"
## attr(,"package")
## [1] "limma"
```

```
names(dat)
```

```
## Loading required package: limma
```

```
## [1] "genes"   "targets" "E"
```

Additionally we have the results of a linear mixed effect model for each gene in the expression data. For this workshop, these data a filtered to just the TB-MEDIA variable.

```
#Linear model results
model.results <- read_csv("data/RSTR.Mtb.model.subset.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   gene = col_character(),
##   variable = col_character(),
##   pval = col_double(),
##   FDR = col_double(),
##   ensembl_gene_id = col_character(),
##   CHR = col_character(),
##   start = col_double(),
##   end = col_double(),
##   entrezgene_id = col_character()
## )
```

```
class(model.results)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
head(model.results)
```

```
## # A tibble: 6 x 9
##   gene   variable       pval      FDR ensembl_gene_id CHR       start       end
##   <chr>  <chr>         <dbl>    <dbl> <chr>           <chr>     <dbl>     <dbl>
## 1 A1BG   conditionTB 1.2e- 1 1.51e- 1 ENSG00000121410 19     58345178  58353492
## 2 A2M    conditionTB 7   e- 4 1.19e- 3 ENSG00000175899 12      9067664   9116229
## 3 A2ML1  conditionTB 3.9e- 1 4.40e- 1 ENSG00000166535 12      8822621   8887001
## 4 A4GALT conditionTB 0       0        ENSG00000128274 22     42692121  42721298
## 5 AAAS   conditionTB 6.8e-13 2.21e-12 ENSG00000094914 12     53307456  53324864
## 6 AACS   conditionTB 7.4e- 3 1.12e- 2 ENSG00000081760 12    125065434 125143333
## # ... with 1 more variable: entrezgene_id <chr>
```

Gene sets

Definition: two or more genes with shared function, structure, localization, or any other defining similarity

Use: glean interpretable, biologically-relevant meaning from a list of genes of interest

Gene sets can be as general as "cytosolic" to as specific as "up-regulated in blood vessel cells in response to wound in *Roy et al 2007*". They can be defined by an individual research study (like the latter) up to generally agreed upon knowledge formed from an entire body of literature (like the former). They also vary in size (2 to thousands), are not exclusive (one gene is in many gene sets), and many are redundant (transmembrane, membrane, lipid bilayer…).

It is up to you, the researcher, to select what group of gene sets you want to test. More on this below. Keep in mind, this is definitely not a case for "more is better". For example, getting 200 significant gene sets from your 250 significant genes does not help you much.

Broad MSigDB

The Broad Institute has collated many useful gene sets in their Molecular Signatures Database (MSigDB). This includes the following used in this workshop (descriptions modified from Broad).

- Hallmark: summarize and represent specific well-defined biological processes. Generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression
  - Total: 50
  - Examples: glycolysis, inflammatory response, apoptosis
- Curated (C2)
  - Canonical pathways: from pathway databases including KEGG, REACTOME, etc. Canonical representations of a biological process compiled by domain experts
  - Total: 2922
  - Examples: Caspase pathway, signaling by NOTCH1, DNA repair
- Gene ontology GO (C5)
  - Biological process: molecular-level activities performed by gene products
  - Total: 7481
  - Examples: viral life cycle, vitamin D biosynthetic process, mitochondrial calcium-ion transmembrane transport

Hallmark is the most general group and is usually where we start our analyses. It helps reduce the complexity of a large gene list down to easily understood terms. It also provides some context when looking at more specific gene sets as you can often group the other sets under a more general Hallmark term. However, because Hallmark is the most general, it is also the most incomplete with 50 gene sets representing just under 4400 genes.

Curated and gene ontology sets are more specific than Hallmark, though they contain sets with a range of specificities. They are particularly useful in adding details to a Hallmark term or defining testable hypotheses for follow-up experiments in the lab. They contain a lot more genes than Hallmark but many sets are redundant or overlapping. If there are many significant Hallmark terms, we often do not run these more specific sets as it gives too many results for interpretation. In contrast, when there are few significant Hallmark terms, these more granular gene sets may provide better insight.

There are several other sets in MSigDB that are not described here. Some of these are too specific (individual studies) to be of use in our analyses (C2 chemical and genetic perturbations or C7 immunologic signatures) while others cover areas not relevant to our experimental design (C1 chromosome position, C3 gene regulation, C4/C6 cancer-oriented). There are still more gene sets that are not included in MSigDB, though this database is a pretty comprehensive place to start.

In reality, we usually run all of Hallmark, canonical pathways, and biological process. Then, we consider how many significant gene sets each gives us, think further about the top hits, and perform additional analyses or experiments to see which will be highlighted in publication.

Next we will use gene sets to unravel how monocytes respond to *M. tuberculosis* infection.

Hypergeometric enrichment

This analysis is a simple enrichment. We ask if our list of significant genes is enriched in gene sets more than random chance allows.

Get gene set data

First, we extract the lists of genes in Broad Hallmark terms from the package `msigdbr`. We format it as a data frame so we can use the `tidyverse`.

```
#Hallmark
H <- as.data.frame(msigdbr(species = "Homo sapiens",
                           category = "H"))
```

Here, we see the start of the Hallmark data including

- gs_cat: gene set category, in this case H for Hallmark
- gs_name: short gene set name
- gene_symbol: HGNC symbol for each each in the gene set
- gs_description: sentence describing the gene set name further
- additional variables: other gene identifiers like ENSEMBL, ENTREZ, etc

```
head(H)
```

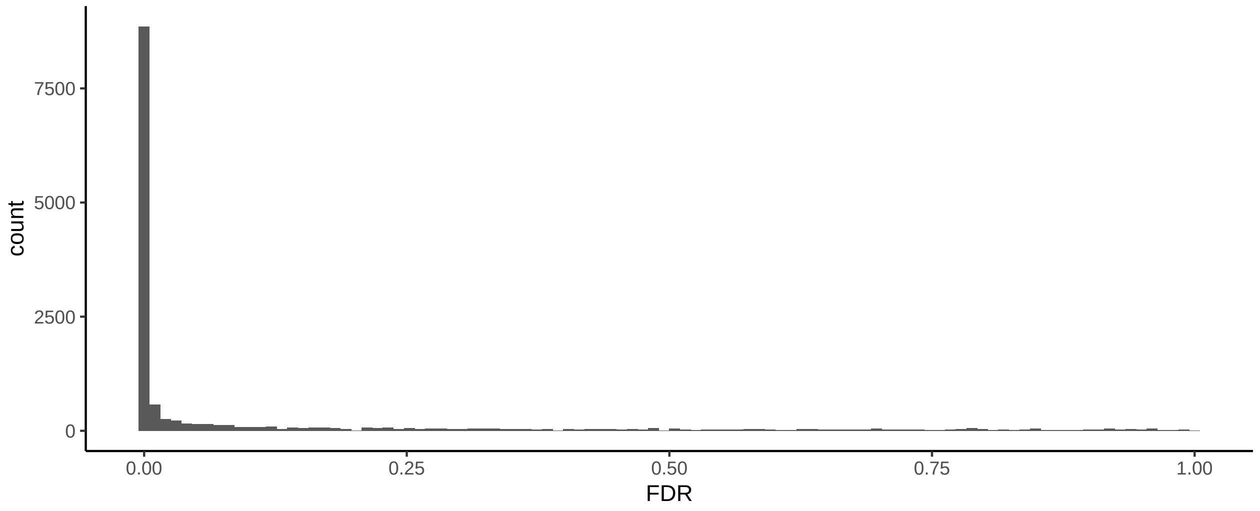
```
##   gs_cat gs_subcat               gs_name gene_symbol entrez_gene
## 1      H           HALLMARK_ADIPOGENESIS       ABCA1          19
## 2      H           HALLMARK_ADIPOGENESIS       ABCB8       11194
## 3      H           HALLMARK_ADIPOGENESIS       ACAA2       10449
## 4      H           HALLMARK_ADIPOGENESIS       ACADL          33
## 5      H           HALLMARK_ADIPOGENESIS       ACADM          34
## 6      H           HALLMARK_ADIPOGENESIS       ACADS          35
##      ensembl_gene human_gene_symbol human_entrez_gene human_ensembl_gene gs_id
## 1 ENSG00000165029             ABCA1                19    ENSG00000165029 M5905
## 2 ENSG00000197150             ABCB8             11194    ENSG00000197150 M5905
## 3 ENSG00000167315             ACAA2             10449    ENSG00000167315 M5905
## 4 ENSG00000115361             ACADL                33    ENSG00000115361 M5905
## 5 ENSG00000117054             ACADM                34    ENSG00000117054 M5905
## 6 ENSG00000122971             ACADS                35    ENSG00000122971 M5905
##   gs_pmid gs_geoid gs_exact_source gs_url
## 1
## 2
## 3
## 4
## 5
## 6
##                                                          gs_description
## 1 Genes up-regulated during adipocyte differentiation (adipogenesis).
## 2 Genes up-regulated during adipocyte differentiation (adipogenesis).
## 3 Genes up-regulated during adipocyte differentiation (adipogenesis).
## 4 Genes up-regulated during adipocyte differentiation (adipogenesis).
## 5 Genes up-regulated during adipocyte differentiation (adipogenesis).
## 6 Genes up-regulated during adipocyte differentiation (adipogenesis).
```

Define significant genes

Next, we must define our "significant genes". This type of enrichment is binary in that it only considers if a gene is significant or not. There is no granularity beyond yes/no. Thus, it is very important to consider our FDR cutoff as it is an arbitrary value that can dramatically impact results.

Let's plot our FDR values as a histogram to help us determine a cutoff. We see that the data are heavily skewed. This is not uncommon when you have a large biological perturbation like bacterial infection.

```
ggplot(model.results, aes(x=FDR)) +
  geom_histogram(bins=100) +
  theme_classic()
```



In fact, there is a large proportion of FDR values that are zero.

```
table(model.results$FDR == 0)
```

```
##
## FALSE  TRUE
## 10476  3502
```

These zeroes are actually a result of limitations in our linear modeling function. After a certain point (in this case, 1E-16), the function stops computing and saves the P-value as zero to save time.

Here, we will use FDR < 1E-16 to see the genes that change the most with Mtb infection. *This is not necessarily what I'd recommend for these data.* It's just for workshop purposes.

```
signif <- model.results %>%
  filter(FDR <= 1E-16)
```

Run enrichment

Now we can run the enrichment in `clusterProfiler`. We can use any gene ID in our dataset; it just has to match the gene IDs in the database.

Here, we run the enrichment using ENTREZ ID. Note that you must select ONLY the columns needed from the Hallmark database. `enricher( )` isn't smart enough to know which columns to use on its own.

```
#Select unique EMTREZ IDs for significant genes
signif.entrez <- unique(signif$entrezgene_id)
```

```
#Select matching ID column in database
H.entrez <- select(H, gs_name, entrez_gene)
#enrichment
enrich.H <- enricher(gene = signif.entrez, TERM2GENE = H.entrez)
```

Similarly, we could use ENSEMBL IDs.

```
#Select unique ENSEMBL IDs for significant genes
signif.ensembl <- unique(signif$ensembl_gene_id)
#Select matching ID column in database
H.ensembl <- select(H, gs_name, ensembl_gene)
#enrichment
enrich.H <- enricher(gene = signif.ensembl, TERM2GENE = H.ensembl)
```

Importantly, these will yield slightly different results, because the ID databases are not perfect. We must just accept that this is life and choose whichever ID we have / prefer. Since the expression data use ENSEMBL, we'll move forward with that.

```
length(signif.ensembl)
```

```
## [1] 3492
```

```
length(signif.entrez)
```

```
## [1] 3496
```

Extract enrichment results

`clusterProfiler` outputs results in an S4 object.

```
class(enrich.H)
```

```
## [1] "enrichResult"
## attr(,"package")
## [1] "DOSE"
```

We briefly mentioned this data type in intro R. It works very similarly to S3 (like the expression data) except you use `@` to extract data. For example, the enrichment results are in a data frame.

```
enrich.H@result
```

This includes

```
colnames(enrich.H@result)
```

```
## [1] "ID"          "Description" "GeneRatio"   "BgRatio"     "pvalue"
## [6] "p.adjust"    "qvalue"      "geneID"      "Count"
```

- ID: Unique name for gene set
- Description: Descriptive text for gene set, often the same as ID
- GeneRatio: Significant genes in gene set / Significant genes in all gene sets tested
- BgRatio: Total genes in gene set / Total genes in all gene sets tested
- pvalue: Significance
- p.adjust: FDR adjust significance
- qvalue: Q-value used in P-value calculation
- geneID: vector of significant gene IDs in gene set. Separated by /
- Count: Significant genes in gene set

These column names aren't the most informative and R doesn't understand the ratios as numbers.

```
class(enrich.H@result$GeneRatio)
```

```
## [1] "character"
```

So we'll do some additional formatting. We refer to a gene set as a "term" and the Hallmark database as a "category" for brevity.

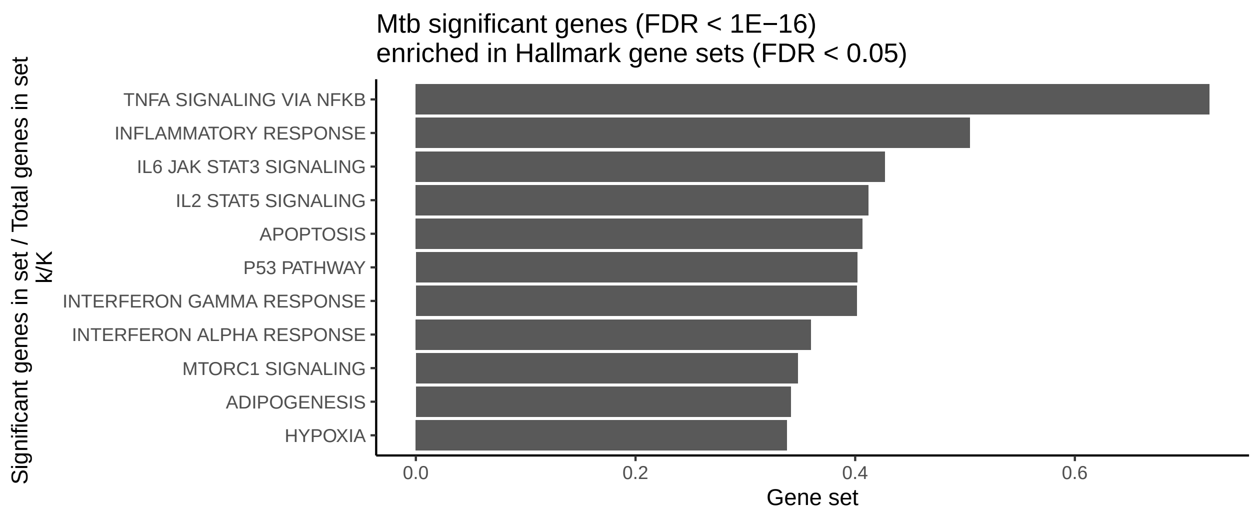
```
enrich.H.df <- enrich.H@result %>%
  #separate ratios into 2 columns of data
  separate(BgRatio, into=c("size.term","size.category"), sep="/") %>%
  separate(GeneRatio, into=c("size.overlap.term", "size.overlap.category"),
             sep="/") %>%
  #convert to numeric
  mutate_at(vars("size.term","size.category",
                  "size.overlap.term","size.overlap.category"),
              as.numeric) %>%
  #Calculate k/K
  mutate("k.K"=size.overlap.term/size.term)
```

Visualize significant enrichment

The most common visualization for enrichment is k/K as it is the ratio of significant genes relative to total genes in the gene set. We also subset to significant enrichment at a defined FDR. Here, let's use FDR < 0.05. This is actually pretty strict for enrichment and people often go up to FDR < 0.2.

```
enrich.H.df %>%
  filter(p.adjust <= 0.05) %>%
  #Beautify descriptions by removing _ and HALLMARK
  mutate(Description = gsub("HALLMARK_","", Description),
         Description = gsub("_"," ", Description)) %>%

ggplot(aes(x=reorder(Description, k.K), #Reorder gene sets by k/K values
           y=k.K)) +
  geom_col() +
  theme_classic() +
  #Some more customization to pretty it up
  #Flip x and y so long labels can be read
  coord_flip() +
  #fix labels
  labs(x="Significant genes in set / Total genes in set \nk/K",
      y="Gene set",
      title = "Mtb significant genes (FDR < 1E-16)\nenriched in Hallmark gene sets (FDR < 0.05)")
```



Exercises

Here's some helpful code to extract the other gene sets of interest.

```
#Canonical pathways
#Note that we have to call each database and combine them together
C2.CP <- as.data.frame(msigdbr(species = "Homo sapiens",
                               category = "C2",
                               subcategory = "CP:BIOCARTA")) %>%
        bind_rows(as.data.frame(msigdbr(species = "Homo sapiens",
                                category = "C2",
                                subcategory = "CP:KEGG"))) %>%
        bind_rows(as.data.frame(msigdbr(species = "Homo sapiens",
                                category = "C2",
                                subcategory = "CP:PID"))) %>%
        bind_rows(as.data.frame(msigdbr(species = "Homo sapiens",
                                category = "C2",
                                subcategory = "CP:REACTOME")))
#Gene ontology biological process
C5.BP <- as.data.frame(msigdbr(species = "Homo sapiens",
                               category = "C5", subcategory = "GO:BP"))
```

1. Run enrichment of the current significant gene list against the canonical pathways `C2.CP` database. Do you get more or fewer significant gene sets? Is this what you'd expect give what you know about C2?
2. Run Hallmark enrichment on genes significant at FDR < 0.01. Do you get more or fewer significant gene sets? Why might this be the case?
3. Further modify the final ggplot. Add color, change the FDR cutoff, do whatever your heart desires!

Gene set enrichment analysis (GSEA)

Additional resources

Groups

- Rladies Seattle Not just for ladies! A pro-actively inclusive R community with both in-person and online workshops, hangouts, etc.
- TidyTuesday A weekly plotting challenge

- R code club Dr. Pat Schloss opened his lab's coding club to remote participation.
- Seattle useR Group

Online

- R cheatsheets also available in RStudio under Help > Cheatsheets
- The Carpentries

R session

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/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] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] limma_3.44.3          fgsea_1.14.0          clusterProfiler_3.16.1
##  [4] msigdbr_7.4.1         forcats_0.5.1         stringr_1.4.0
##  [7] dplyr_1.0.7           purrr_0.3.4           readr_1.4.0
## [10] tidyr_1.1.3           tibble_3.1.2          ggplot2_3.3.5
## [13] tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##   [1] colorspace_2.0-2     ggridges_0.5.3       ellipsis_0.3.2
##   [4] qvalue_2.20.0        fs_1.5.0             rstudioapi_0.13
##   [7] farver_2.1.0         urltools_1.7.3       graphlayouts_0.7.1
##  [10] ggrepel_0.9.1        bit64_4.0.5          scatterpie_0.1.6
##  [13] AnnotationDbi_1.50.3 fansi_0.5.0          lubridate_1.7.10
##  [16] xml2_1.3.2           splines_4.0.2        cachem_1.0.5
##  [19] GOSemSim_2.14.2      knitr_1.33           polyclip_1.10-0
##  [22] jsonlite_1.7.2       broom_0.7.8          GO.db_3.11.4
##  [25] dbplyr_2.1.1         ggforce_0.3.3        BiocManager_1.30.16
##  [28] compiler_4.0.2       httr_1.4.2           rvcheck_0.1.8
##  [31] backports_1.2.1      assertthat_0.2.1     Matrix_1.3-4
##  [34] fastmap_1.1.0        cli_3.0.0            tweenr_1.0.2
##  [37] htmltools_0.5.1.1    prettyunits_1.1.1    tools_4.0.2
##  [40] igraph_1.2.6         gtable_0.3.0         glue_1.4.2
##  [43] reshape2_1.4.4       DO.db_2.9            fastmatch_1.1-0
##  [46] Rcpp_1.0.7           enrichplot_1.8.1     Biobase_2.48.0
##  [49] cellranger_1.1.0     vctrs_0.3.8          babelgene_21.4
##  [52] ggraph_2.0.5         xfun_0.24            rvest_1.0.0
##  [55] lifecycle_1.0.0      DOSE_3.14.0          europepmc_0.4
##  [58] MASS_7.3-54          scales_1.1.1         tidygraph_1.2.0
##  [61] hms_1.1.0            parallel_4.0.2       RColorBrewer_1.1-2
##  [64] yaml_2.2.1           memoise_2.0.0        gridExtra_2.3
##  [67] downloader_0.4       triebeard_0.3.0      stringi_1.6.2
##  [70] RSQLite_2.2.7        highr_0.9            S4Vectors_0.26.1
##  [73] BiocGenerics_0.34.0  BiocParallel_1.24.1  rlang_0.4.11
##  [76] pkgconfig_2.0.3      evaluate_0.14        lattice_0.20-44
##  [79] labeling_0.4.2       cowplot_1.1.1        bit_4.0.4
##  [82] tidyselect_1.1.1     plyr_1.8.6.9000      magrittr_2.0.1
##  [85] R6_2.5.0             IRanges_2.22.2       generics_0.1.0
##  [88] DBI_1.1.1            pillar_1.6.1         haven_2.4.1
##  [91] withr_2.4.2          modelr_0.1.8         crayon_1.4.1
##  [94] utf8_1.2.1           rmarkdown_2.9        viridis_0.6.1
##  [97] progress_1.2.2       grid_4.0.2           readxl_1.3.1
## [100] data.table_1.14.0    blob_1.2.1           reprex_2.0.0
## [103] digest_0.6.27        gridGraphics_0.5-1   stats4_4.0.2
## [106] munsell_0.5.0        viridisLite_0.4.0    ggplotify_0.0.7
```