Introduction to Bulk RNAseq data analysis

Gene Set Testing for RNA-seq - Solutions

Contents

Exercise 1 - pathview

Load the required packages and data for Day 11 if you have not already done so.

```
library(msigdbr)
library(clusterProfiler)
library(pathview)
library(tidyverse)

shrink.d11 <- readRDS("RObjects/Shrunk_Results.d11.rds")
```

1. Use `pathview` to export a figure for “mmu04659”or “mmu04658”, but this time only use genes that are statistically significant at FDR < 0.01

```
logFC <- shrink.d11 %>%
  drop_na(padj, Entrez) %>%
  filter(padj < 0.01) %>%
  pull(log2FoldChange, Entrez)

pathview(gene.data = logFC,
         pathway.id = "mmu04659",
         species = "mmu",
         limit = list(gene=5, cpd=1))
```

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

```
## Info: Working in directory /mnt/c/Users/hugot/Documents/BTF/course_materials/Bulk_RNAseq_Course_Base
```

```
## Info: Writing image file mmu04659.pathview.png
```

mmu04659.pathview.png:

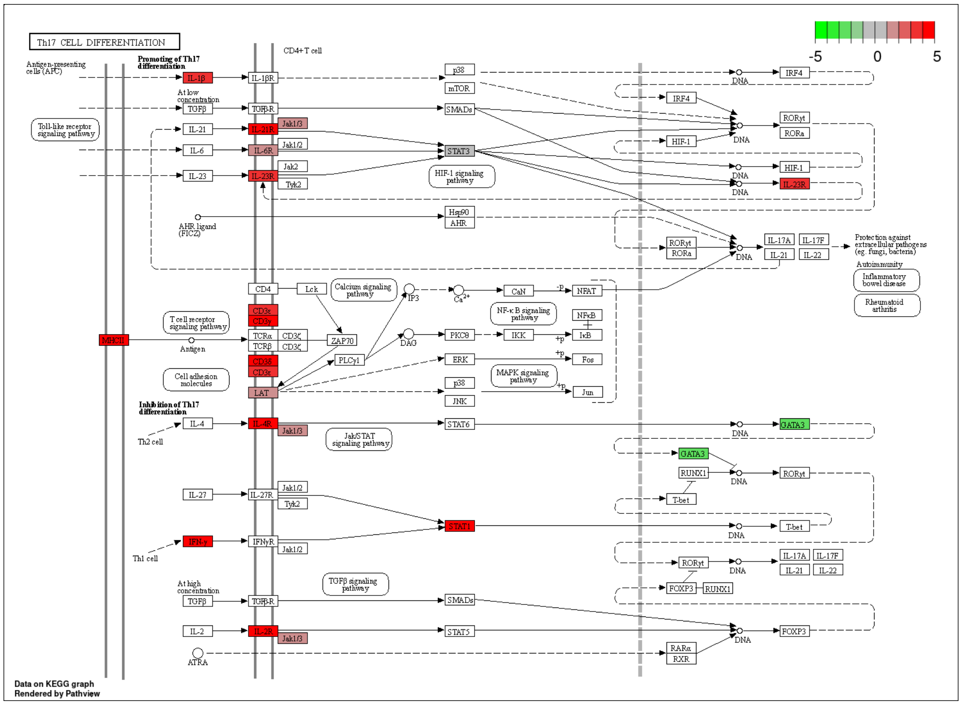

Figure 1: mmu04659 - Th17 cell differentiation

Exercise 2 - GSEA

> Another common way to rank the genes is to order by pvalue, but also, sorting so that upregulated genes are at the start and downregulated at the end - you can do this combining the sign of the fold change and the pvalue.

First load the pathway details if you have not already done so.

```
library(msigdbr)
term2gene <- msigdbr(species = "Mus musculus", category = "H") %>%
  dplyr::select(gs_name, ensembl_gene)
term2name <- msigdbr(species = "Mus musculus", category = "H") %>%
  dplyr::select(gs_name, gs_description) %>%
  distinct()
```

1. Rank the genes by statistical significance - you will need to create a new ranking value using `-log10({p value}) * sign({Fold Change})`.

```
# rank genes
rankedGenes.e11 <- shrink.d11 %>%
  drop_na(GeneID, pvalue, log2FoldChange) %>%
  mutate(rank = -log10(pvalue) * sign(log2FoldChange)) %>%
  arrange(desc(rank)) %>%
  pull(rank, GeneID)
```

2. Run GSEA using the new ranked genes and the Hallmark pathways.

```
# conduct analysis:
gseaRes.e11 <- GSEA(rankedGenes.e11,
                TERM2GENE = term2gene,
                TERM2NAME = term2name,
                pvalueCutoff = 1.00,
                minGSSize = 15,
                maxGSSize = 500)
```

```
## preparing geneSet collections...

## GSEA analysis...

## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize, gseaParam, : There are ties in
## The order of those tied genes will be arbitrary, which may produce unexpected results.

## Warning in fgseaMultilevel(pathways = pathways, stats = stats, minSize =
## minSize, : For some pathways, in reality P-values are less than 1e-10. You can
## set the ‘eps‘ argument to zero for better estimation.

## leading edge analysis...

## done...
```

View the results:

```
as_tibble(gseaRes.e11) %>%
  arrange(desc(abs(NES))) %>%
  top_n(10, wt=-p.adjust) %>%
  dplyr::select(-core_enrichment) %>%
  mutate(across(c("enrichmentScore", "NES"), round, digits=3)) %>%
  mutate(across(c("pvalue", "p.adjust", "qvalue"), scales::scientific))
```

3. Conduct the same analysis for the day 33 Infected vs Uninfected contrast.

```
# read d33 data in:
shrink.d33 <- readRDS("RObjects/Shrunk_Results.d33.rds")

# rank genes
rankedGenes.e33 <- shrink.d33 %>%
  drop_na(GeneID, pvalue, log2FoldChange) %>%
  mutate(rank = -log10(pvalue) * sign(log2FoldChange)) %>%
  arrange(desc(rank)) %>%
  pull(rank,GeneID)

# perform analysis
gseaRes.e33 <- GSEA(rankedGenes.e33,
                TERM2GENE = term2gene,
                TERM2NAME = term2name,
                pvalueCutoff = 1.00,
                minGSSize = 15,
                maxGSSize = 500)
```

```
## preparing geneSet collections...
```

```
## GSEA analysis...
```

```
## Warning in fgseaMultilevel(pathways = pathways, stats = stats, minSize =
## minSize, : There were 1 pathways for which P-values were not calculated
## properly due to unbalanced (positive and negative) gene-level statistic values.
## For such pathways pval, padj, NES, log2err are set to NA. You can try to
## increase the value of the argument nPermSimple (for example set it nPermSimple
## = 10000)
```

```
## Warning in fgseaMultilevel(pathways = pathways, stats = stats, minSize =
## minSize, : For some pathways, in reality P-values are less than 1e-10. You can
## set the ‘eps‘ argument to zero for better estimation.
```

```
## leading edge analysis...
```

```
## done...
```

View the results:

```
as_tibble(gseaRes.e33) %>%
  arrange(desc(abs(NES))) %>%
  top_n(10, wt=-p.adjust) %>%
  dplyr::select(-core_enrichment) %>%
  mutate(across(c("enrichmentScore", "NES"), round, digits=3)) %>%
  mutate(across(c("pvalue", "p.adjust", "qvalue"), scales::scientific))
```