Introduction to Bulk RNAseq data analysis

Gene Set Testing for RNA-seq - Solutions

- Exercise 1 - pathview
- Exercise 2 - GSEA

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

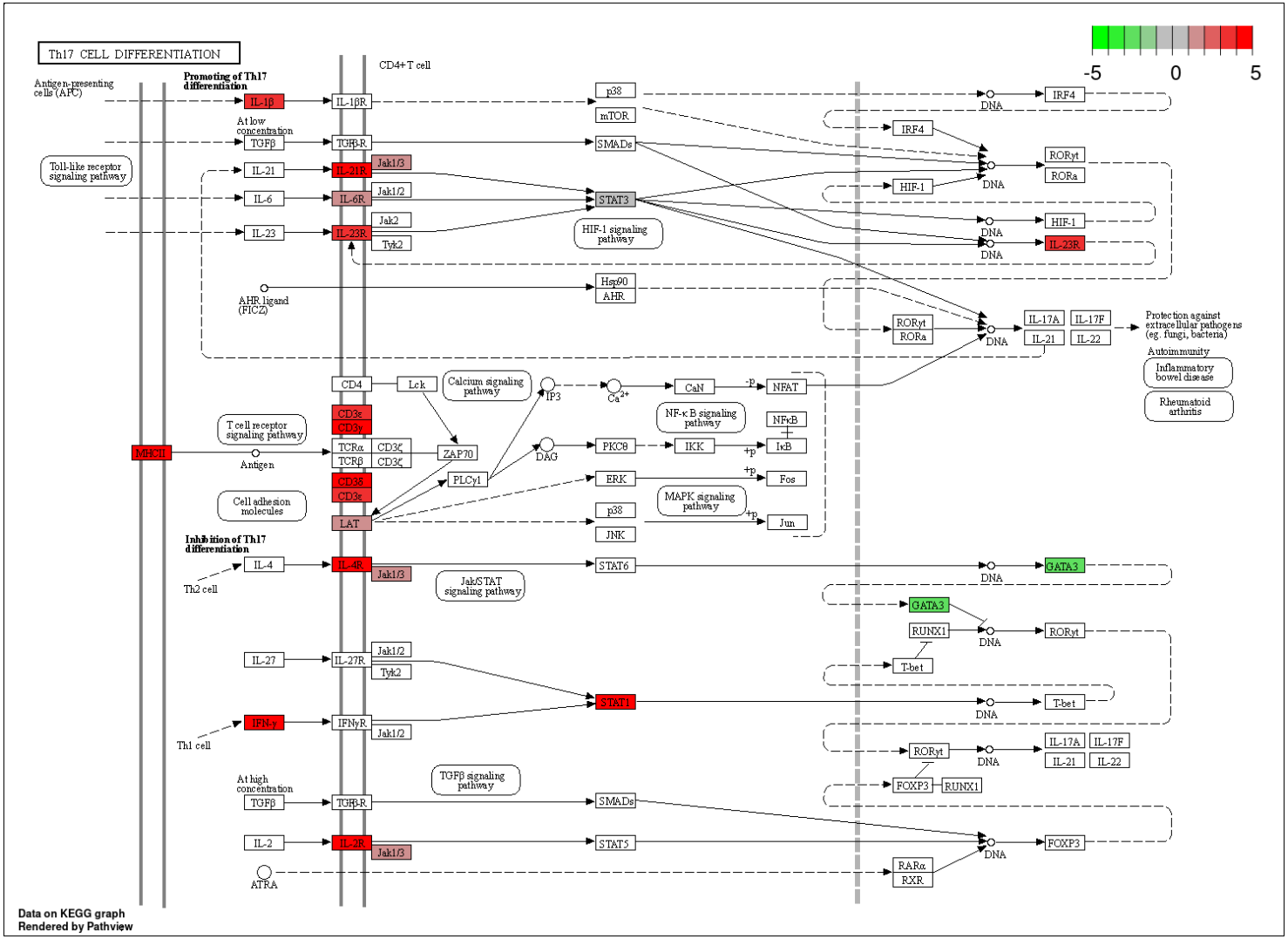
```
logFC <- shrink.d11 %>%
  drop_na(FDR, Entrez) %>%
  filter(FDR < 0.01) %>%
  pull(logFC, Entrez)

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 /Users/edward03/Bulk_RNAseq_Course_Base/Markdowns
```

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

mmu04659.pathview.png:



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, entrez_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(Entrez, pvalue, logFC) %>%
  mutate(rank = -log10(pvalue) * sign(logFC)) %>%
  arrange(desc(rank)) %>%
  pull(rank, Entrez)
```

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 fgseaMultilevel(...): 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))
```

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
1	HALLMARK_OXIDATIVE_PHOSPHORYLATION	Genes encoding proteins involved in oxidative phosphorylation.	198	-0.488	-2.064	1.39e-08	1.74e-07	1.28e-07	4323
2	HALLMARK_INTERFERON_GAMMA_RESPONSE	Genes up-regulated in response to IFNG [GeneID=3458].	202	0.944	1.971	1.00e-10	1.67e-09	1.23e-09	824
3	HALLMARK_INTERFERON_ALPHA_RESPONSE	Genes up-regulated in response to alpha interferon proteins.	97	0.951	1.96	1.00e-10	1.67e-09	1.23e-09	619
4	HALLMARK_ALLOGRAFT_REJECTION	Genes up-regulated during transplant rejection.	196	0.901	1.876	1.00e-10	1.67e-09	1.23e-09	874
5	HALLMARK_IL6_JAK_STAT3_SIGNALING	Genes up-regulated by IL6 [GeneID=3569] via STAT3 [GeneID=6774], e.g., during acute phase response.	83	0.876	1.801	7.29e-06	5.21e-05	3.84e-05	1422
6	HALLMARK_INFLAMMATORY_RESPONSE	Genes defining inflammatory response.	195	0.826	1.719	6.27e-07	6.27e-06	4.62e-06	1146

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
7	HALLMARK_COMPLEMENT	Genes encoding components of the complement system, which is part of the innate immune system.	189	0.817	1.7	3.45e-06	2.87e-05	2.12e-05	1294
8	HALLMARK_TNFA_SIGNALING_VIA_NFKB	Genes regulated by NF-kB in response to TNF [GeneID=7124].	195	0.8	1.666	8.68e-06	5.43e-05	4.00e-05	1621
9	HALLMARK_APOPTOSIS	Genes mediating programmed cell death (apoptosis) by activation of caspases.	158	0.803	1.655	4.34e-05	2.41e-04	1.78e-04	1545
10	HALLMARK_IL2_STAT5_SIGNALING	Genes up-regulated by STAT5 in response to IL2 stimulation.	196	0.773	1.61	5.44e-05	2.72e-04	2.00e-04	1106

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(Entrez, pvalue, logFC) %>%
  mutate(rank = -log10(pvalue) * sign(logFC)) %>%
  arrange(desc(rank)) %>%
  pull(rank,Entrez)

# 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(...): There were 3 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(...): 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))
```

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
1	HALLMARK_INTERFERON_ALPHA_RESPONSE	Genes up-regulated in response to alpha interferon proteins.	97	0.936	1.763	1.00e-10	1.18e-09	9.21e-10	839
2	HALLMARK_INTERFERON_GAMMA_RESPONSE	Genes up-regulated in response to IFNG [GeneID=3458].	202	0.929	1.761	1.00e-10	1.18e-09	9.21e-10	899
3	HALLMARK_ALLOGRAFT_REJECTION	Genes up-regulated during transplant rejection.	196	0.913	1.731	1.00e-10	1.18e-09	9.21e-10	976
4	HALLMARK_IL6_JAK_STAT3_SIGNALING	Genes up-regulated by IL6 [GeneID=3569] via STAT3 [GeneID=6774], e.g., during acute phase response.	83	0.872	1.643	1.03e-05	8.09e-05	6.34e-05	1149
5	HALLMARK_INFLAMMATORY_RESPONSE	Genes defining inflammatory response.	195	0.864	1.638	1.00e-10	1.18e-09	9.21e-10	1322

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qvalue	rank
6	HALLMARK_COMPLEMENT	Genes encoding components of the complement system, which is part of the innate immune system.	189	0.82	1.553	5.59e-07	5.25e-06	4.12e-06	1143
7	HALLMARK_IL2_STAT5_SIGNALING	Genes up-regulated by STAT5 in response to IL2 stimulation.	196	0.778	1.476	5.65e-05	3.79e-04	2.97e-04	1577
8	HALLMARK_KRAS_SIGNALING_UP	Genes up-regulated by KRAS activation.	196	0.765	1.451	1.92e-04	1.13e-03	8.83e-04	1515
9	HALLMARK_TNFA_SIGNALING_VIA_NFKB	Genes regulated by NF-kB in response to TNF [GeneID=7124].	195	0.75	1.422	1.08e-03	5.63e-03	4.41e-03	1390
10	HALLMARK_APOPTOSIS	Genes mediating programmed cell death (apoptosis) by activation of caspases.	158	0.748	1.411	3.63e-03	1.71e-02	1.34e-02	1618