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)
```

```
##
```

```
## clusterProfiler v4.10.0  For help: https://yulab-smu.top/biomedical-knowledge-mining-book/
##
## If you use clusterProfiler in published research, please cite:
## T Wu, E Hu, S Xu, M Chen, P Guo, Z Dai, T Feng, L Zhou, W Tang, L Zhan, X Fu, S Liu, X Bo, and G Yu. clusterPr
ofiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021, 2(3):100141
```

```
##
## Attaching package: 'clusterProfiler'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(pathview)
```

```
##
```

```
## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3). Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products. For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.0     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter()   masks clusterProfiler::filter(), stats::filter()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ purrr::simplify() masks clusterProfiler::simplify()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
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))
```

```
## Warning: None of the genes or compounds mapped to the pathway!
## Argument gene.idtype or cpd.idtype may be wrong.
```

```
## Loading required namespace: org.Mm.eg.db
```

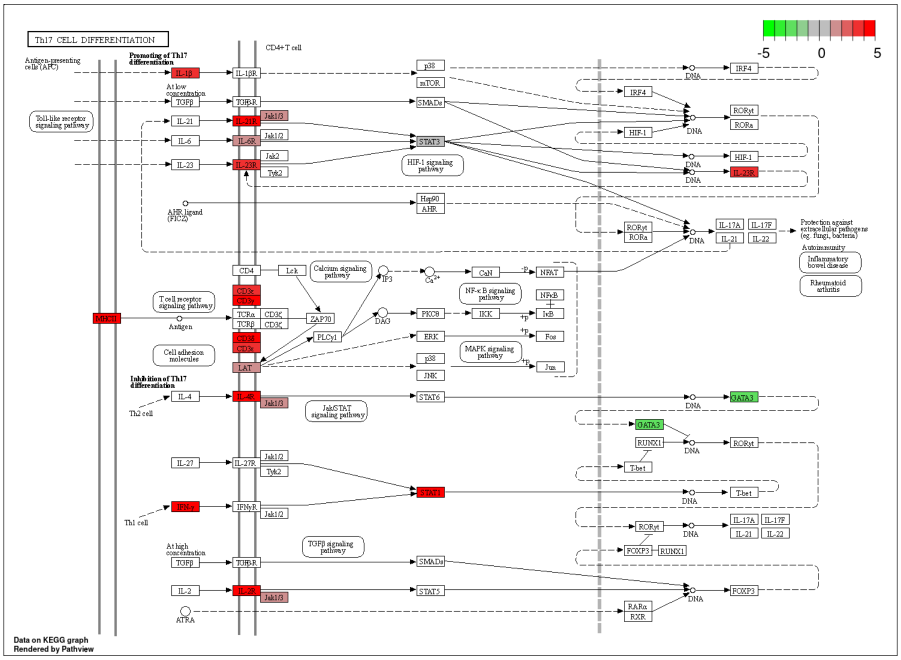
```
##
```

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

```
## Info: Working in directory /Users/sawle01/Documents/training/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(msigdb)

msigdb.mm <- getMsigdb(org = 'mm', id = 'EZID', version = '2023.1')
```

```
## see ?msigdb and browseVignettes('msigdb') for documentation
```

```
## loading from cache
```

```
hallmarks = subsetCollection(msigdb.mm, 'h')
msigdb_ids = geneIds(hallmarks)

term2gene <- enframe(msigdb_ids, name = "gs_name", value = "entrez") %>%
  unnest(entrez)
```

> 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, log2FoldChange) %>%
  mutate(rank = -log10(pvalue) * sign(log2FoldChange)) %>%
  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,
                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 prera
nked stats (0% of the list).
## The order of those tied genes will be arbitrary, which may produce unexpected results.
```

```
## Warning in preparePathwaysAndStats(pathways, stats, minSize, maxSize,
## gseaParam, : There are duplicate gene names, fgsea 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(.x, digits=3))) %>%
  mutate(across(c("pvalue", "p.adjust", "qvalue"), scales::scientific))
```

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qv
1	HALLMARK_INTERFERON_ALPHA_RESPONSE	HALLMARK_INTERFERON_ALPHA_RESPONSE	162	0.957	1.962	1.00e-10	1.67e-09	1.
2	HALLMARK_INTERFERON_GAMMA_RESPONSE	HALLMARK_INTERFERON_GAMMA_RESPONSE	291	0.946	1.939	1.00e-10	1.67e-09	1.
3	HALLMARK_ALLOGRAFT_REJECTION	HALLMARK_ALLOGRAFT_REJECTION	294	0.914	1.874	1.00e-10	1.67e-09	1.
4	HALLMARK_OXIDATIVE_PHOSPHORYLATION	HALLMARK_OXIDATIVE_PHOSPHORYLATION	223	-0.498	-1.87	7.25e-09	7.25e-08	5.
5	HALLMARK_IL6_JAK_STAT3_SIGNALING	HALLMARK_IL6_JAK_STAT3_SIGNALING	113	0.879	1.794	8.90e-07	5.04e-06	3.
6	HALLMARK_IL2_STAT5_SIGNALING	HALLMARK_IL2_STAT5_SIGNALING	301	0.826	1.696	2.33e-09	2.91e-08	2.
7	HALLMARK_INFLAMMATORY_RESPONSE	HALLMARK_INFLAMMATORY_RESPONSE	299	0.814	1.672	2.75e-08	2.29e-07	1.
8	HALLMARK_TNFA_SIGNALING_VIA_NFKB	HALLMARK_TNFA_SIGNALING_VIA_NFKB	266	0.806	1.656	9.07e-07	5.04e-06	3.
9	HALLMARK_COMPLEMENT	HALLMARK_COMPLEMENT	299	0.793	1.627	8.84e-07	5.04e-06	3.
10	HALLMARK_UV_RESPONSE_UP	HALLMARK_UV_RESPONSE_UP	291	0.776	1.59	1.72e-05	8.61e-05	6.

> 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, log2FoldChange) %>%
  mutate(rank = -log10(pvalue) * sign(log2FoldChange)) %>%
  arrange(desc(rank)) %>%
  pull(rank, Entrez)

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

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

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

```
## Warning in fgseaMultilevel(pathways = pathways, stats = stats, minSize =
## minSize, : There were 2 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(.x, digits=3))) %>%
  mutate(across(c("pvalue", "p.adjust", "qvalue"), scales::scientific))
```

	ID	Description	setSize	enrichmentScore	NES	pvalue	p.adjust	qv
1	HALLMARK_INTERFERON_ALPHA_RESPONSE	HALLMARK_INTERFERON_ALPHA_RESPONSE	162	0.936	1.759	1.00e-10	1.20e-09	9.
2	HALLMARK_INTERFERON_GAMMA_RESPONSE	HALLMARK_INTERFERON_GAMMA_RESPONSE	291	0.927	1.736	1.00e-10	1.20e-09	9.
3	HALLMARK_ALLOGRAFT_REJECTION	HALLMARK_ALLOGRAFT_REJECTION	294	0.92	1.721	1.00e-10	1.20e-09	9.
4	HALLMARK_IL6_JAK_STAT3_SIGNALING	HALLMARK_IL6_JAK_STAT3_SIGNALING	113	0.874	1.639	2.77e-07	1.90e-06	1.
5	HALLMARK_INFLAMMATORY_RESPONSE	HALLMARK_INFLAMMATORY_RESPONSE	299	0.842	1.576	1.00e-10	1.20e-09	9.
6	HALLMARK_IL2_STAT5_SIGNALING	HALLMARK_IL2_STAT5_SIGNALING	301	0.82	1.534	2.42e-09	2.32e-08	1.
7	HALLMARK_COMPLEMENT	HALLMARK_COMPLEMENT	299	0.797	1.491	2.31e-07	1.85e-06	1.
8	HALLMARK_UV_RESPONSE_UP	HALLMARK_UV_RESPONSE_UP	291	0.771	1.444	2.78e-05	1.67e-04	1.
9	HALLMARK_KRAS_SIGNALING_UP	HALLMARK_KRAS_SIGNALING_UP	331	0.755	1.409	4.07e-05	2.17e-04	1.
10	HALLMARK_TNFA_SIGNALING_VIA_NFKB	HALLMARK_TNFA_SIGNALING_VIA_NFKB	266	0.739	1.383	1.06e-03	5.09e-03	3.