Homework 3 - BIOS 7649

Dominic Adducci

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'kableExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Question 1: T-statistics

Part A

For each gene, calculate the fold change between the knock-out and wildtype groups. List the top 10 genes that show the largest fold change (positive or negative).

There are 8 mice in the control group (C57B1/6 strain) and 8 mice in the experimental group (apo AI knockout). In order to calculate the fold change, either positive or negative between the two groups, the averages for each group needs to be calculated for each gene.

Table 1: 10 Largest Absolute Fold Changes

Gene Name	Difference
ApoAI,lipid-Img	-4.74924672295693
EST,HighlysimilartoA	-4.57282566300963
CATECHOLO-METHYLTRAN	-2.77224890674299
EST,WeaklysimilartoC	-1.54043051734832
ESTs,Highlysimilarto	-1.51471758262052
est	-1.46613543694816
similartoyeaststerol	-1.43245389359317
ApoCIII,lipid-Img	-1.39887354565181
psoriasis-associated	-1.2567134509252
Cy3RT	1.19328632324561

Table 1 illustrates the largest absolute difference in fold change. The difference was found by subtracting the averaged expression of the control group from the averaged expression of the experimental group. This means that positive values indicate increased expression in the experimental group, and negative valuees indicate increased expression in the control group.

Part B

Obtain the p-values from a two sided t-test for differential expression. How many genes are significant at the 0.01 level? List the top 10 genes that have the largest t-statistic and their corresponding p-value.

Table 2: 10 Largest t-statistics

Gene Name	t-statistic	p-value
ESTs,Weaklysimilarto	4.127	0.001245
estrogenrec	3.958	0.001589
	3.916	0.001609
MDB1430	3.720	0.002520
5'similartogb:X02747	3.687	0.002662
BLANK	3.632	0.004920
Meox2	3.582	0.004355
BLANK	3.564	0.003754
	3.549	0.003213
	3.542	0.004997

Part C: Alternative 't-statistics'

Part C.i

Calculate the modified 't-statistic' and corresponding p-value using the **samr** package in R used in Homework 2. How many genes are significant at the 0.01 level? List the top 10 genes that have the largest 'penalized' t-statistics.

Part C.ii

Calculate the 'moderated' t-statistic and corresponding p-value using the **limma** package from BioConductor. How many genes are significant at the 0.01 level? List the top 10 genes that have the largest 'penalized' t-statistics.

The 'moderated' t-statistic is

Table 3: 10 Largest modified t-statistics - Limma

Gene Name	Modified t-statistic	p-value	Adj. p-value (BH)
ApoAI,lipid-Img	-23.977	4.74e-15	3.03e-11
EST,HighlysimilartoA	-12.963	1.56e-10	4.99e-07
CATECHOLO-METHYLTRAN	-12.440	3.05e-10	6.48e-07
EST,WeaklysimilartoC	-11.750	7.61e-10	1.21e-06
ApoCIII,lipid-Img	-9.831	1.23e-08	1.57e-05
ESTs,Highlysimilarto	-9.013	4.54e-08	4.23e-05
est	-9.000	4.64e-08	4.23e-05
similartoyeaststerol	-7.440	7.04e-07	0.000562
EST,WeaklysimilartoF	-4.554	0.000249	0.176959
	-3.961	0.000925	0.528486

Part D

Compare and contrast the results for the four methods for ranking genes. Explain the difference in how the different t-statistics are calculated.