Hadi Ahmad
ahmad287 - 5833802
03/11/2025
CSCI 5461

CSCI 5461 Homework 2 Report

Files:

report.pdf	
DataProcessing.py	separates RNAseq and microarray data into groups 1 & 2.
MyHPTest.py	prints top 10 genes and p-values from SeqData or ArrayData.
MyBonferri.py	outputs significant genes and # of significant genes.
MyFDR.py	takes input file and # of genes; returns upper bound of FDR.

1-6)

How many genes are in the RNAseq gene expression profiles?

20,530

How many patient samples are in the RNAseq gene expression profiles?

308

How many genes are in the microarray gene expression profiles?

12,042

How many patient samples are in the microarray gene expression profiles?

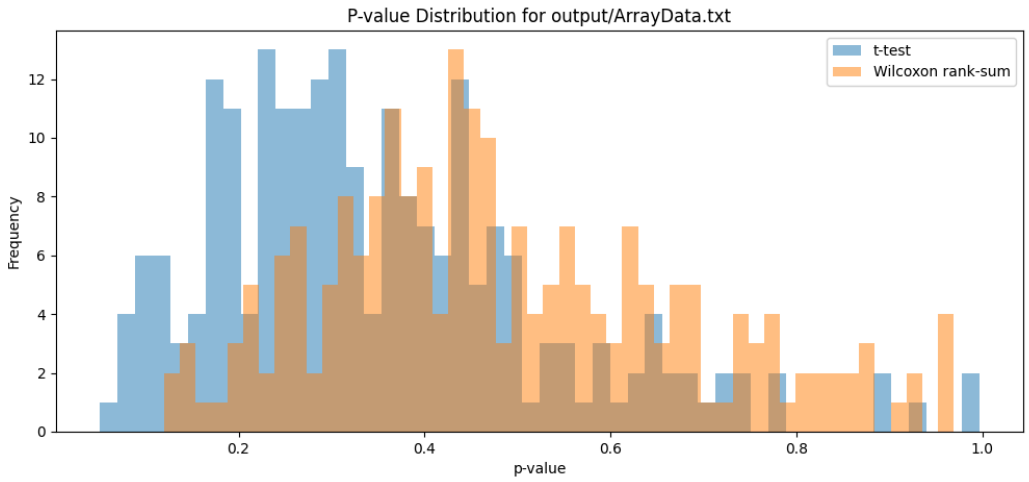
593

2)

Top 10 genes by t-test for ArrayData:
TCGA-29-1691-01: **0.050939020625269414**
TCGA-30-1860-01: **0.07263791948677291**
TCGA-13-1507-01: **0.07710136548960868**
TCGA-61-1738-01: **0.08027566969998572**
TCGA-61-1740-01: **0.08841363179281882**
TCGA-23-1109-01: **0.09074696109106231**
TCGA-13-0897-01: **0.091309702017386**
TCGA-13-2060-01: **0.09352778448357892**
TCGA-29-1695-01: **0.09452499755341554**
TCGA-13-0920-01: **0.09716428014388227**

Top 10 genes by Wilcoxon rank-sum test for ArrayData:
TCGA-29-1691-01: **0.11980952707156248**
TCGA-29-1695-01: **0.12240831560180228**
TCGA-30-1860-01: **0.14029621450508384**
TCGA-13-0897-01: **0.14565337078441495**
TCGA-61-1738-01: **0.15102174609296415**
TCGA-23-1109-01: **0.16328889905354138**
TCGA-13-1489-02: **0.17276960140466602**
TCGA-13-2060-01: **0.19875829538188883**
TCGA-24-1550-01: **0.20074361126993312**
TCGA-59-2348-01: **0.20076113867225098**

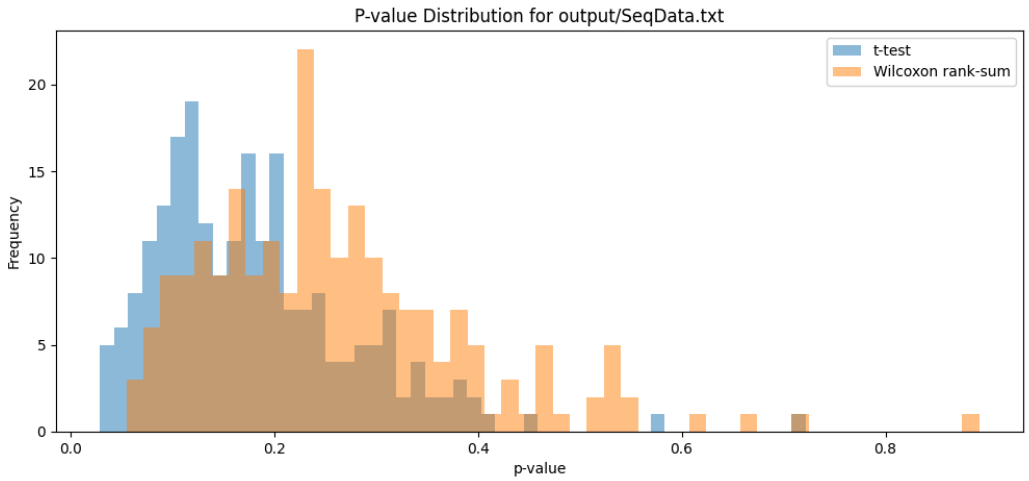
# of significant genes in t-test: **0**
# of significant genes in Wilcoxon rank-sum test: **0**



Top 10 genes by t-test for SeqData:
TCGA-24-1550-01: **0.029287491101010935**
TCGA-23-1109-01: **0.03139603505599098**
TCGA-13-0920-01: **0.03790885448690603**
TCGA-29-1691-01: **0.037975485600042574**
TCGA-09-0369-01: **0.039599244672068885**
TCGA-61-2102-01: **0.04339888411100968**
TCGA-25-1323-01: **0.04675915303543957**
TCGA-25-1626-01: **0.04771821072331441**
TCGA-61-1740-01: **0.05530234270980623**
TCGA-13-1507-01: **0.05574490491592756**

Top 10 genes by Wilcoxon rank-sum test for SeqData:
TCGA-24-1550-01: **0.05514354099781494**
TCGA-23-1109-01: **0.0705759767629625**
TCGA-13-0920-01: **0.0712308304379362**
TCGA-13-1507-01: **0.07321872154973821**
TCGA-09-0369-01: **0.07613094479358742**
TCGA-24-2262-01: **0.07885996481987545**
TCGA-61-2102-01: **0.08109822331520192**
TCGA-13-2060-01: **0.08255852095629077**
TCGA-29-1691-01: **0.0852140233140261**
TCGA-29-1761-01: **0.08861639450280388**

# of significant genes in t-test: **8**
# of significant genes in Wilcoxon rank-sum test: **0**



3)

0
No significant genes found after Bonferroni correction for SeqData
0
No significant genes found after Bonferroni correction for ArrayData

Using FDR with t-test to identify differentially expressed genes:

Upper bound of FDR for selecting 20 genes from ArrayData: **0.14001502028220658**
Upper bound of FDR for selecting 20 genes from SeqData: **0.07236414473346986**

Upper bound of FDR for selecting 50 genes from ArrayData: **0.21959569638239854**
Upper bound of FDR for selecting 50 genes from SeqData: **0.10682934098685029**

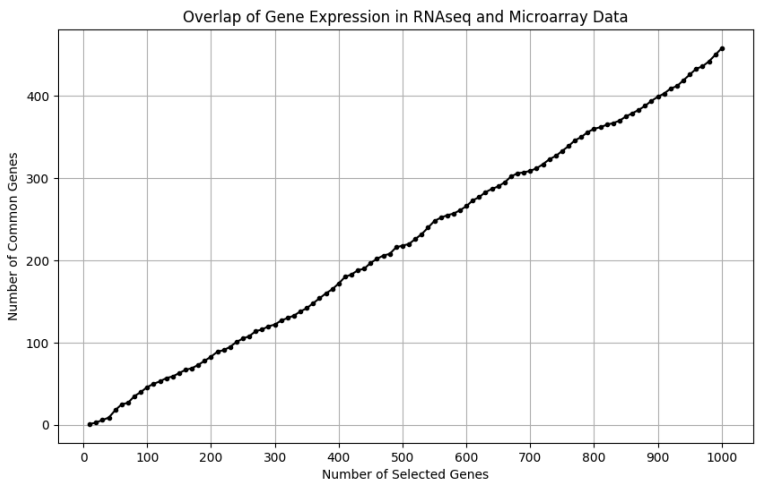
Upper bound of FDR for selecting 100 genes from ArrayData: **0.29782696085126126**
Upper bound of FDR for selecting 100 genes from SeqData: **0.15317791197234842**

Upper bound of FDR for selecting 200 genes from ArrayData: **0.6344106958776847**
Upper bound of FDR for selecting 200 genes from SeqData: **0.3120275237521143**

4)
I didn't really understand what this question was asking for so I took some liberty in how I went about it.

I found the mean of the confidence values of genes among all patients in SeqData and ArrayData and sorted those from highest to lowest.

I observed a linear relationship between the number of selected genes and the number of common genes. For the range of the graph, the slope is unchanged.



The consistent identification of top genes by both gene expression methods shows they are in agreement. The reason why the number of common genes is roughly half the number of selected genes doesn't diminish the quality of the data, as merging them removes this issue.