Homework 2 Report

Part 1: Data Preparation

Genes in RNAseq gene expression profiles: 20530

Patient samples in RNAseq gene expression profiles: 227

Genes in microarray gene expression profiles: 12042

Patient samples in microarray gene expression profiles: 229

Part 2: t-Test and Wilcoxon Rank-Sum Statistics

Most significant differentially expressed genes and their p-values

t-test on RNAseq data
CYTH3 : 3.0098008945298933e-06
RGS14 : 1.2377496908085309e-05
ATP1A2 : 3.021467534308624e-05
TSPAN9 : 3.119161855029564e-05
FBXO4 : 3.4374730812896994e-05
CHUK : 5.300175074303698e-05
C12orf5 : 5.387038194521066e-05
EIF4E3 : 5.995904107751949e-05
LOC728606 : 6.696139097648087e-05
BTRC : 7.700893153314288e-05

Wilcoxon rank-sum test on RNAseq data
ATP1A2 : 6.241557588492144e-06
CYTH3 : 1.0018718466041045e-05
RGS14 : 3.239566952400099e-05
TEX261 : 4.9985317509497084e-05
BTRC : 5.4310032540826994e-05
SMTNL2 : 5.672391473591208e-05
ASAP3 : 5.672615542462553e-05
LOC728606 : 5.92997496100092e-05
SLC1A6 : 7.527573519735624e-05
CAMK2G : 7.638880907942947e-05

t-test on microarray data
AXL : 4.4899418978368735e-06
ADORA3 : 1.2833732198030451e-05
TLR7 : 8.843656271801246e-05
GALNT10 : 9.688612244176527e-05
TLR4 : 9.815673544814246e-05
PGDS : 0.00010448770763780476
FZD10 : 0.00020901161908830075
LILRA2 : 0.0002123795280329791
TREM2 : 0.00027548044981792394
KCTD14 : 0.0002890463115687482

Wilcoxon rank-sum test on microarray data
GALNT10 : 3.28225657697374e-05
AXL : 5.649231736014718e-05
ADORA3 : 0.00011938638214227499
TBX2 : 0.0001255667744448675
TLR7 : 0.0001287688931261546
SYNC1 : 0.00014000731521566636
FZD10 : 0.0001840030296751166
PLXNC1 : 0.00020135298207223356
TLR4 : 0.0002446822821735065
NDRG3 : 0.000254747347297092

Number of statistically significant differentially expressed genes (using alpha=0.05)

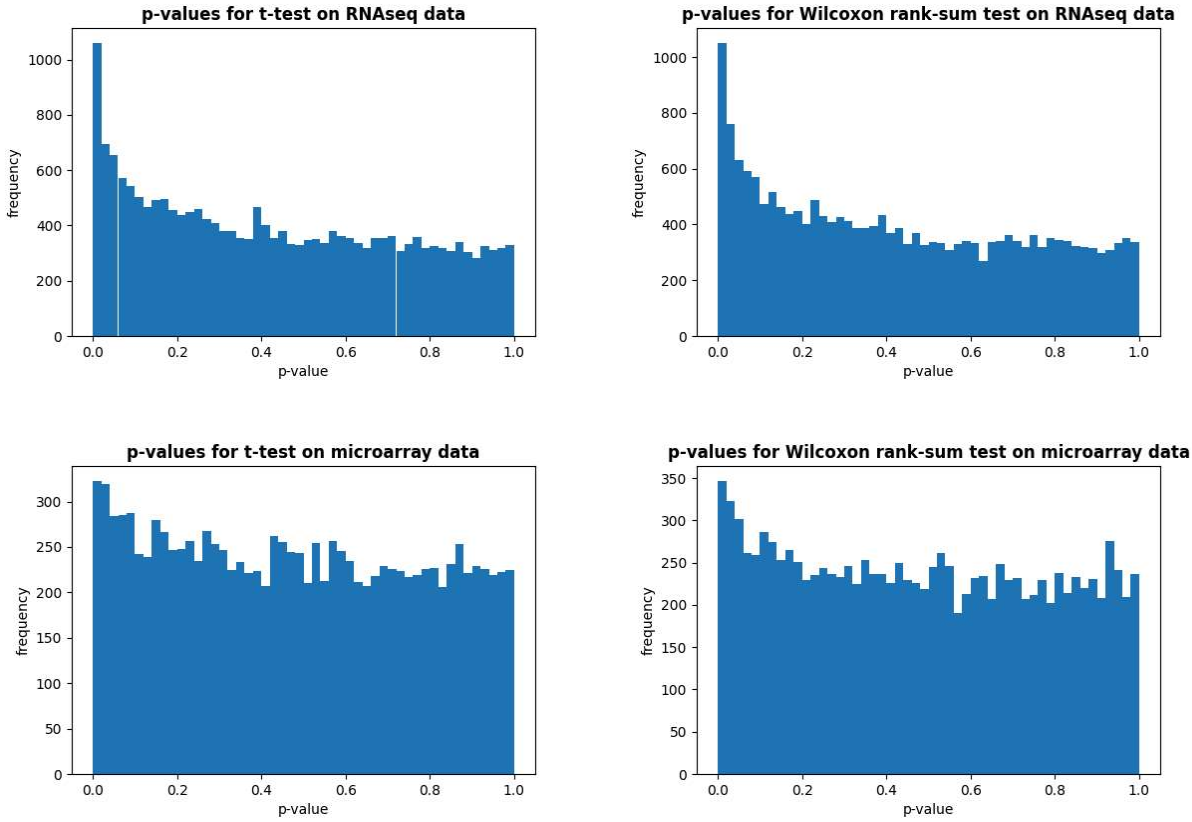
t-test on RNAseq data: 2100

Wilcoxon rank-sum test on RNAseq data: 2151

t-test on microarray data: 795

Wilcoxon rank-sum test on microarray data: 830

Histograms of the p-values of all genes in each test



The histograms represent the distribution of p-values for all of the genes analyzed for each test. More p-values closer to zero indicates that differentially expressed genes are statistically significant.

Part 3: Multiple Testing Correction

Statistically significant differentially expressed genes using Bonferroni correction (alpha=0.05)

Bonferroni correction on RNAseq data: 0

Bonferroni correction on microarray data: 0

The Bonferroni correction is very conservative, especially when used on a large number of genes. Due to this, it is not surprising that no genes remain significant after the correction. Only a few p-values remain below 1 after the correction, with none of them being less than the significance level of 0.05.

Statistically significant differentially expressed genes using FDR on t-tests (alpha=0.05)

SeqData.txt using 20 genes
Significant genes after FDR correction: 2
CHCHD4 : 0.02514506847659216
GPR133 : 0.003984368278558147
Upper bound of FDR correction:
0.045000000000000005

ArrayData.txt using 20 genes
Significant genes after FDR correction: 0
Upper bound of FDR correction: 0.05

SeqData.txt using 50 genes
Significant genes after FDR correction: 1
GPR133 : 0.009960920696395368
Upper bound of FDR correction: 0.049

ArrayData.txt using 50 genes
Significant genes after FDR correction: 0
Upper bound of FDR correction: 0.05

SeqData.txt using 100 genes
Significant genes after FDR correction: 2
GPR133 : 0.009960920696395368
NEBL : 0.009960920696395368
Upper bound of FDR correction: 0.049

ArrayData.txt using 100 genes
Significant genes after FDR correction: 0
Upper bound of FDR correction: 0.05

SeqData.txt using 200 genes
Significant genes after FDR correction: 3
GPR133 : 0.013281227595193824
NEBL : 0.013281227595193824
C14orf72 : 0.013281227595193824
Upper bound of FDR correction: 0.04925

ArrayData.txt using 200 genes
Significant genes after FDR correction: 0
Upper bound of FDR correction: 0.05

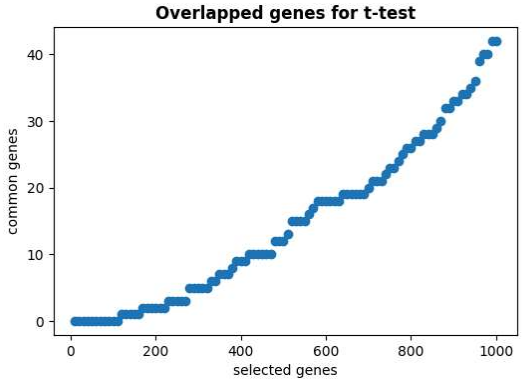
The results from using the false discovery rate (FDR) correction indicate that some RNAseq genes were identified as significant with significance level 0.05, but no genes from the microarray data were. It's possible that the signal in the microarray data is weaker, or the variance in measurements is higher. Either of these would lead to fewer significant genes.

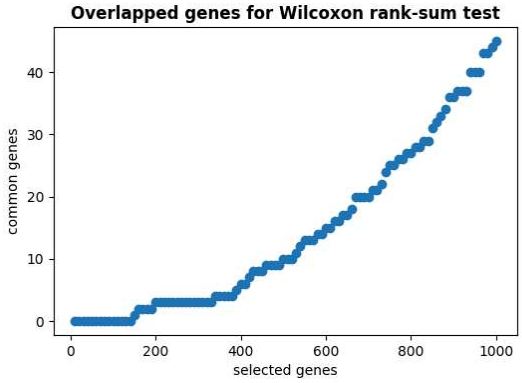
The upper bound of the FDR correction represents the maximum proportion of false positives among all the genes declared as significant after the multiple testing correction. A false positive in this case is a gene being identified as significant when in reality it is not.

Using a calculation from above as an example, when the multiple testing correction was applied to the RNAseq data using 200 genes, 3 genes were found to be significant and the upper bound of the FDR was 0.04925. This means that at most ~4.93% of these significant genes might be false positives.

For the microarray data, no significant genes were found, so the FDR upper bound remains at the significance level alpha=0.05.

Part 4: Overlapped genes





As the number of selected genes increases, the number of common genes also increases. This is intuitive, as more genes being selected allows for more opportunities of overlap.

The shape of the plots for the t-test and Wilcoxon rank-sum tests are very similar and non-linear. At first, the number of common genes increases slowly as the number of selected genes increases, but then the common genes begin to increase more rapidly.

When the most significant 1000 genes from each technology are compared, there are only about 40-50 genes overlapping. If the two technologies detected differentially expressed genes more similarly, the number of overlaps would closely match the number of selected genes and the plot of selected genes versus common genes would be nearly linear. Therefore, based on these plots, the differentially expressed genes detected by the two genomic technologies are not very similar.