Matthew Markman

HW #1 CSCI 5461

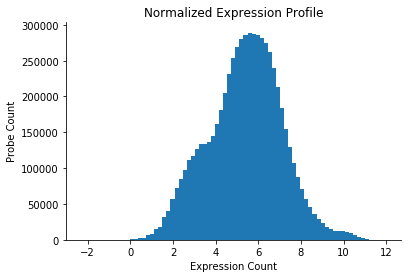
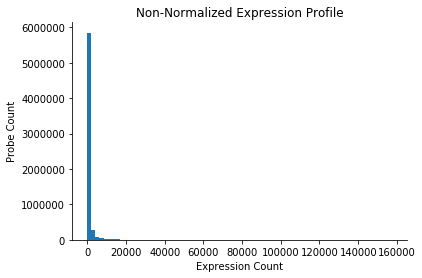
**Note: I used the pandas script for problems 2-4 and the R script for problem 5.**

2a) There are 22283 probes in the dataset.

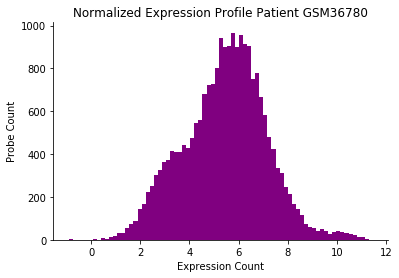
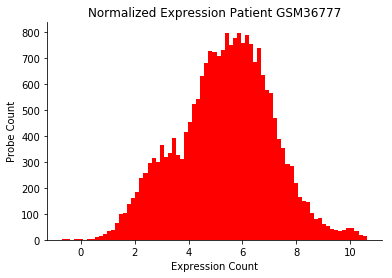
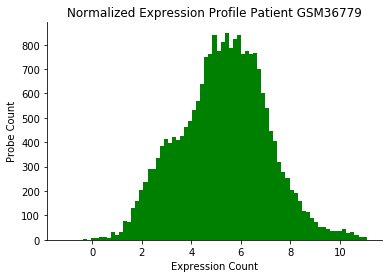
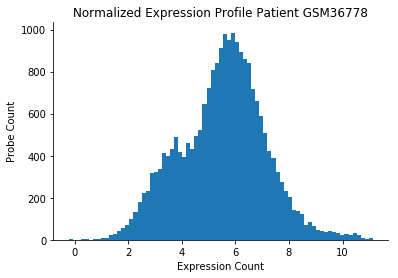
2b) There are 286 patients. Of those patients, 107 had relapses and 179 did not.

2c) There are 13211 unique genes represented by the 22283 probes in the dataset.

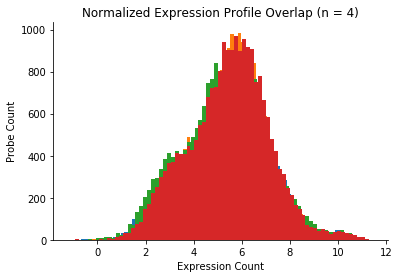
3a) The non-log normalized and log normalized plots are shown below:



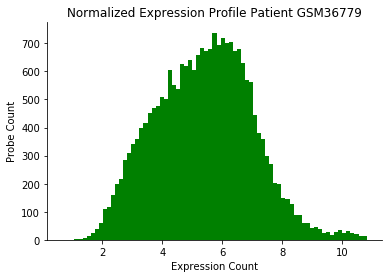
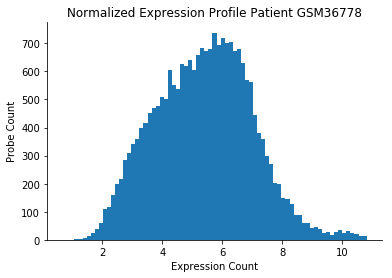
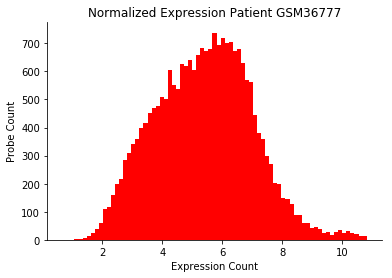
3b) We can see that although the overall shapes of the log transformed histograms of the first 4 samples are similar, there are some differences in overall shape. We can especially see this when we overlap the plots:



Overlapped plot:



3c) Plots of first four samples post-quantile normalization:



4a) Top 10 differentially expressed probes/gene names for T-test and Wilcoxon rank sum test are:

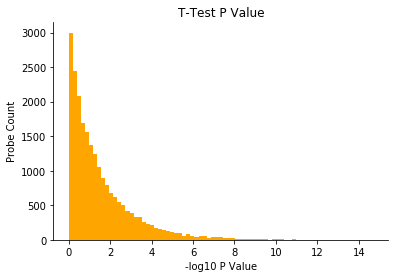
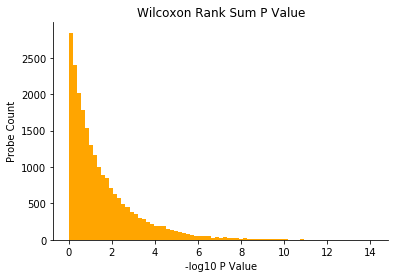
|  |  |
| --- | --- |
| **T-test- (gene name, probe name)** | **P-Value** |
| (ACBD3, 202324\_s\_at) | 4.180574e-07 |
| (WFDC1, 219478\_at) | 1.583293e-06 |
| (ABCC5, 209380\_s\_at) | 2.300614e-06 |
| (RACGAP1, 222077\_s\_at) | 3.744510e-06 |
| (CLINT1, 201769\_at) | 4.873324e-06 |
| (FBXO7, 201178\_at) | 5.807846e-06 |
| (ZFP36L2, 201369\_s\_at | 6.064825e-06 |
| (SHC1, 214853\_s\_at) | 6.439620e-06 |
| (ERP29, 201216\_at) | 9.569786e-06 |
| (NEK2, 204641\_at) | 9.983478e-06 |

|  |  |
| --- | --- |
| **Wilcoxon Test- (gene name, probe name)** | **P-Value** |
| (ACBD3, 202324\_s\_at) | 7.349828e-07 |
| (WFDC1, 219478\_at) | 1.768041e-06 |
| (BLZF1, 32088\_at) | 2.446633e-06 |
| (CLINT1, 201769\_at) | 3.595689e-06 |
| (ZFP36L2, 201369\_s\_at) | 4.676234e-06 |
| (RACGAP1, 222077\_s\_at) | 7.067712e-06 |
| (NEK2, 204641\_at) | 7.166448e-06 |
| (LACTB2, 218701\_at | 8.060662e-06 |
| (SHC1, 214853\_s\_at) | 8.116456e-06 |
| (SEC24A, 212900\_at) | 9.092105e-06 |

NEK2 stands for Never in Mitosis Related Kinase 2. It is a known oncogene associated with a wide variety of cancers and is especially associated with tumor progression. NEK2 has numerous roles in cell cycle mediation, so it stands to reason that an overexpression of NEK2 may cause cancer, potentially though premature separation of centrosomes in mitosis ([Fang and Zhang, 2017](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4889274/)).

Rac GTPase-activating protein (RACGAP1) plays a key role in mediating cytokinesis and transformation. It has also been shown that RacGAP1 expression was an independent predictive marker for lymph node metastasis. RACGAP1 works in tandem to upregulate RHO, which may be a driving force in metastasis ([Imaoka et al., 2015](https://academic.oup.com/carcin/article/36/3/346/315177))

4b) There are 3259 DE probes according to the t test and 3292 according to the rank sum test. Histograms of –log10 p values for both tests are shown below:



4c) I found 2740 overlapping probes. The top 10 overlapping genes/probes are: ACBD3, WFDC1, CLINT1, RACGAP1, NEK2, ZFP36L2, SHC1, FBX07, ERP29, and ABCC5. These may not be in order. I used a weighting system of the lowest ten values of rank\_ttest + rank\_wilcoxon.

5a) I found 2 differentially expressed probes with the Bonferroni correction on Wilcoxon rank sum p values.

5b) Using BH (assuming that the probes are independent), I found 175 differentially expressed genes (FDR <= 0.05).

5c) Here is my plot:

