Emily Moeller
2/26/2019

Homework 1

2. Exploring the dataset (10 points): Process the data file using whatever programming language you have chosen and answer the following questions:

(a) How many probes are included in the dataset? 22283

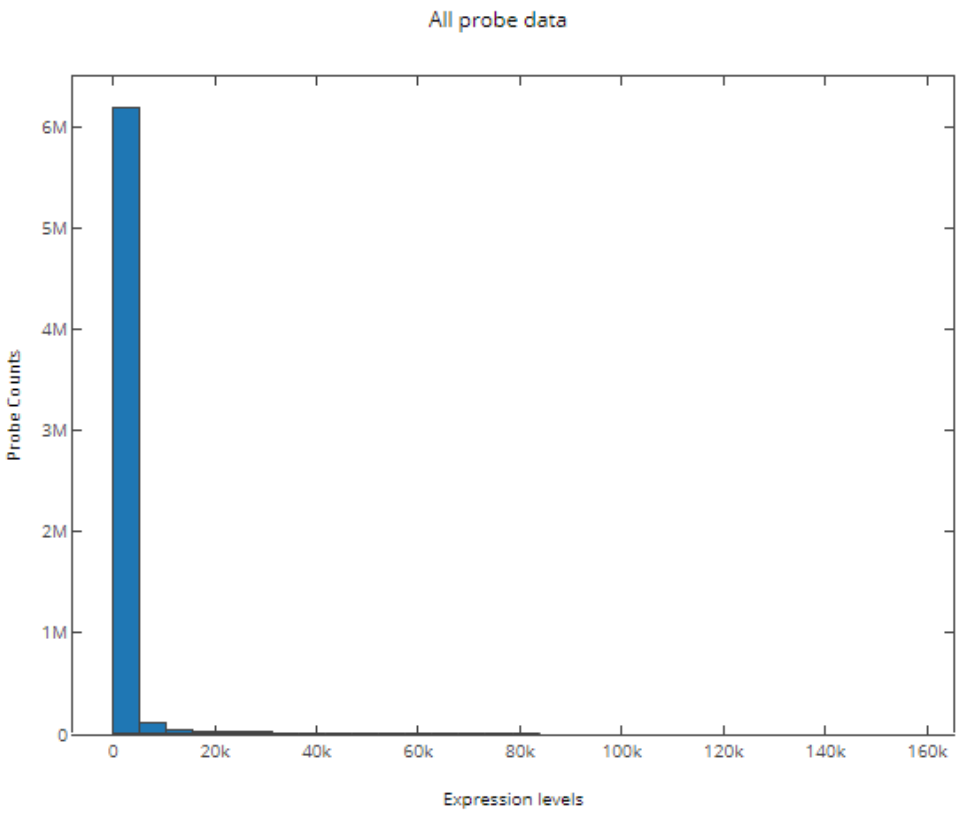
(b) How many patient samples are in the dataset? We will divide patients into two groups based on their relapse status for our analysis. How many patients were relapse free (relapse=0, i.e. no metastases)? How many patients had relapses (relapse=1)? Total number of patients= 286 Relapse free=179 Relapse=107

(c) How many unique genes are represented by the probes in the dataset? Note: we would typically average multiple probes mapping to the same gene, but to keep the lab relatively simple, we will analyze each probe independently

Unique genes=13212

3. Data processing and normalization (30 points):

(a) Plot a histogram (x-axis: expression levels, y-axis: probe counts) of the complete dataset. Describe the distribution—what is the overall shape? Replace any values <= 0 with a value of 1, and log-transform (log 10) the entire data matrix. Plot the log transformed data.



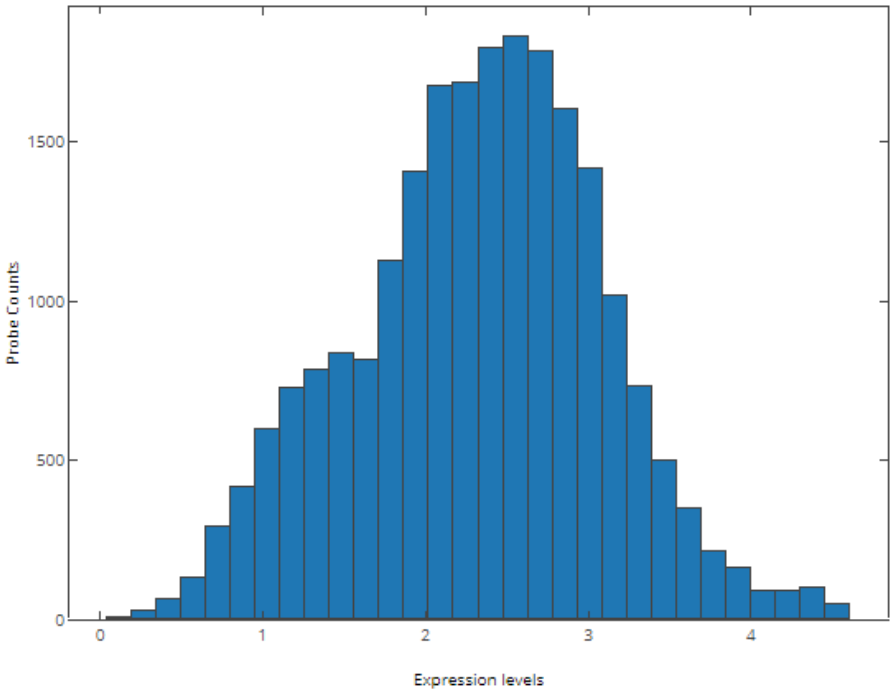
The data is left skewed.

It's a normal distribution.

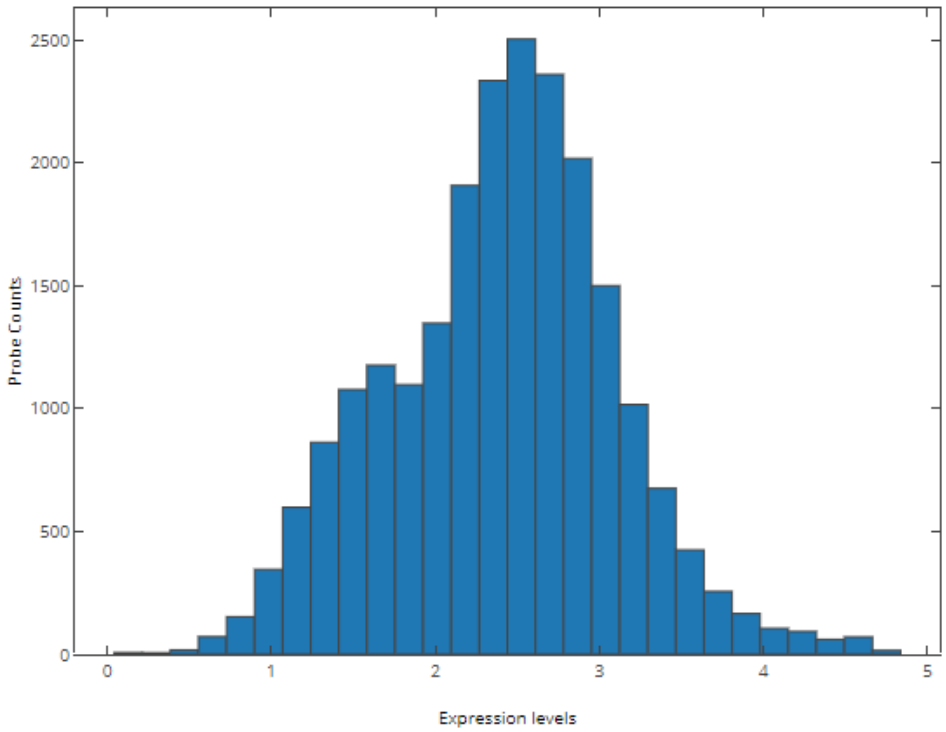
(b) Plot individual histograms of the log-transformed data for the first four arrays (GSM36777-GSM36780). How do the distributions compare from sample to sample?

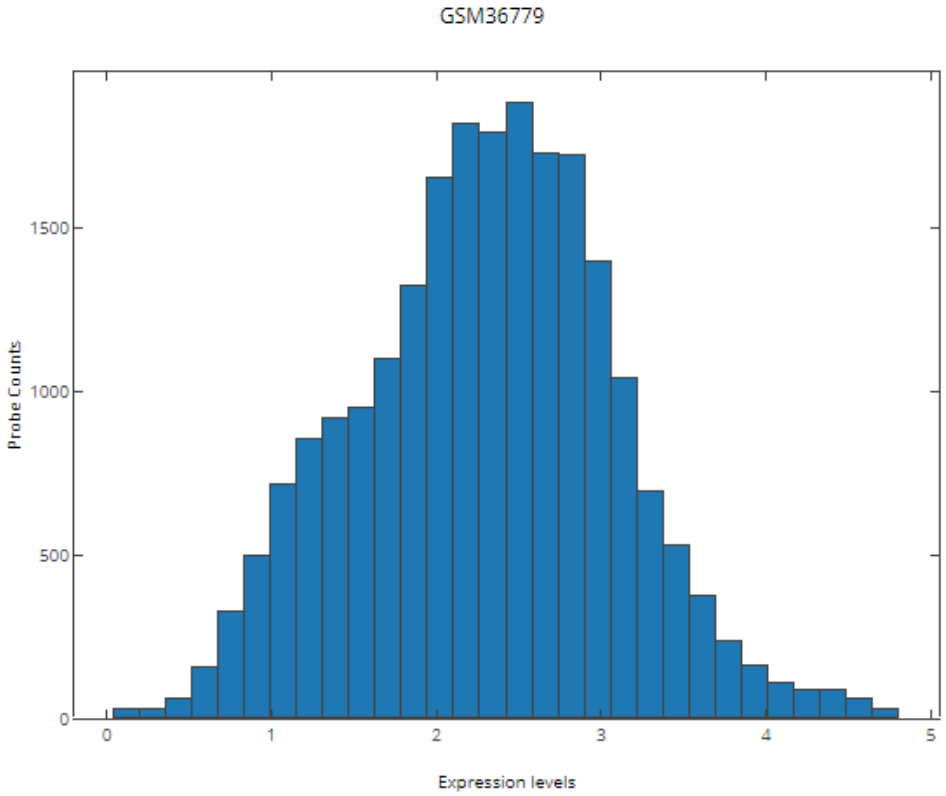
It's similar. These distributions are roughly normal, with more or the data being left skewed.

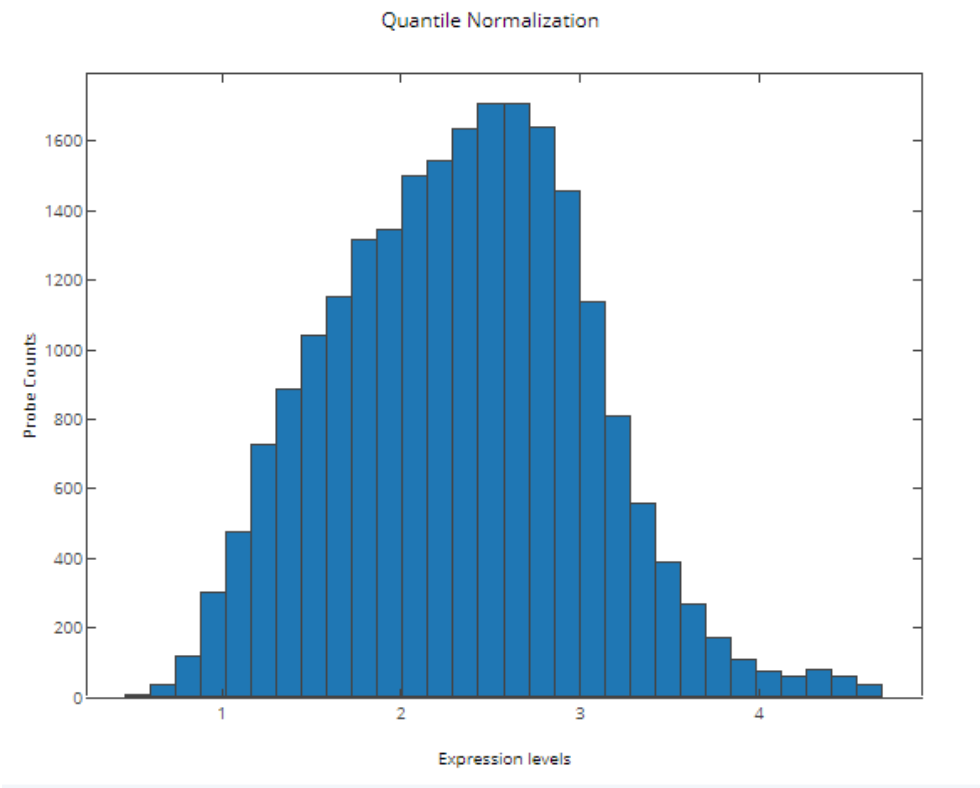
GSM36777



GSM36778




GSM36779
Probe Counts
1500
1000
500
0
0
1
2
3
4
5
Expression levels



(c) Perform quantile normalization on your log-transformed data (from part b) across all arrays (samples) such that each has the same empirical distribution. Use the mean of each probe across all samples as the reference distribution for this normalization. Plot a histogram of the normalized data for each of the first four samples (GSM36777GSM36780). Use the normalized data for the remaining problems.

The Graph will be the same for all samples when using the reference distribution.

4. Analysis of differential expression (30 points): Use the t-test and Wilcoxon rank-sum statistics to identify differentially expressed probes with a per-probe significance level of $p < 0.05$. You should divide the gene expression data into two groups (metastasis vs. nonmetastasis), using the relapse variable in the clinical data, and test each probe independently.

(a) For each test, list the top 10 probes, the corresponding gene names, and the p-values associated with them. Pick 1-2 of these genes and discuss what is known about their function and how it might relate to cancer metastasis. There are many databases such as GeneCards (http://www.genecards.org), NCBI (http://www.ncbi.nlm.nih.gov/gene), and WikiGenes (http://www.wikigenes.org) that you can use to learn more about a gene.

T-test

1. "202324_s_at" "ACBD3" 4.170738884875756e-07

2. "219478_at" "WFDC1" 1.5727299347442704e-06
3. "209380_s_at" "ABCC5" 2.2948029114418756e-06
4. "222077_s_at" "RACGAP1" 3.74116888551574e-06
5. "201769_at" "CLINT1" 4.859122437770904e-06
6. "201178_at" "FBXO7" 5.805919016072501e-06
7. "201369_s_at" "ZFP36L2" [6.0681838701536915e-06
8. "214853_s_at" "SHC1" 6.4504691092506845e-06
9. "201216_at" "ERP29" 9.58034467241886e-06
10. "204641_at" "NEK2" 9.963297255393795e-06

Unfortunately, there is a bug in my code that is causing my p values for the Wilcoxon rank-sum to be way off.

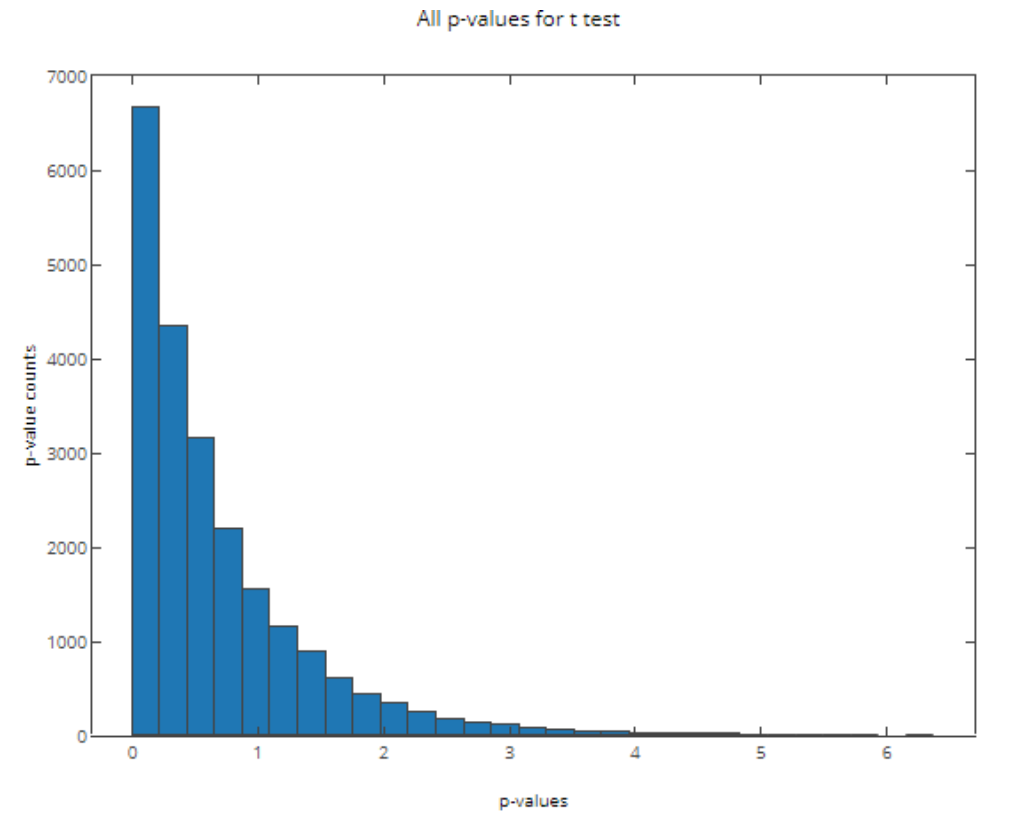
Wilcoxon rank-sum:

1. "214853_s_at" "SHC1" 1.1946492316730048e-100
2. "204619_s_at" "VCAN" 1.3131810819121096e-100
3. "201769_at" "CLINT1" 2.3522804693268963e-100
4. "202066_at" "PPFIA1" 2.796995597888782e-100
5. "212652_s_at" "SNX4" 6.139902578281166e-100
6. "220300_at" "RGS3" 8.405892039404566e-100
7. "200699_at" "KDELR2" 1.2032451860927034e-98
8. "201897_s_at" "CKS1B" 2.8812565359020423e-98
9. "212900_at" "SEC24A" 6.887602609992814e-98
10. "37512_at" "HSD17B6" 1.0001137594504825e-97

The gene ACBD3 plays an important role in the sorting and modification of proteins exported from the endoplasmic reticulum. It also is involved in the maintenance of the Golgi structure and function by its interaction with the vital membrane protein giantin. If this gene is not properly controlling its specified functions in the protein giantin and the proteins from the endoplasmic reticulum, then maybe it could play a role in cancer metastasis.

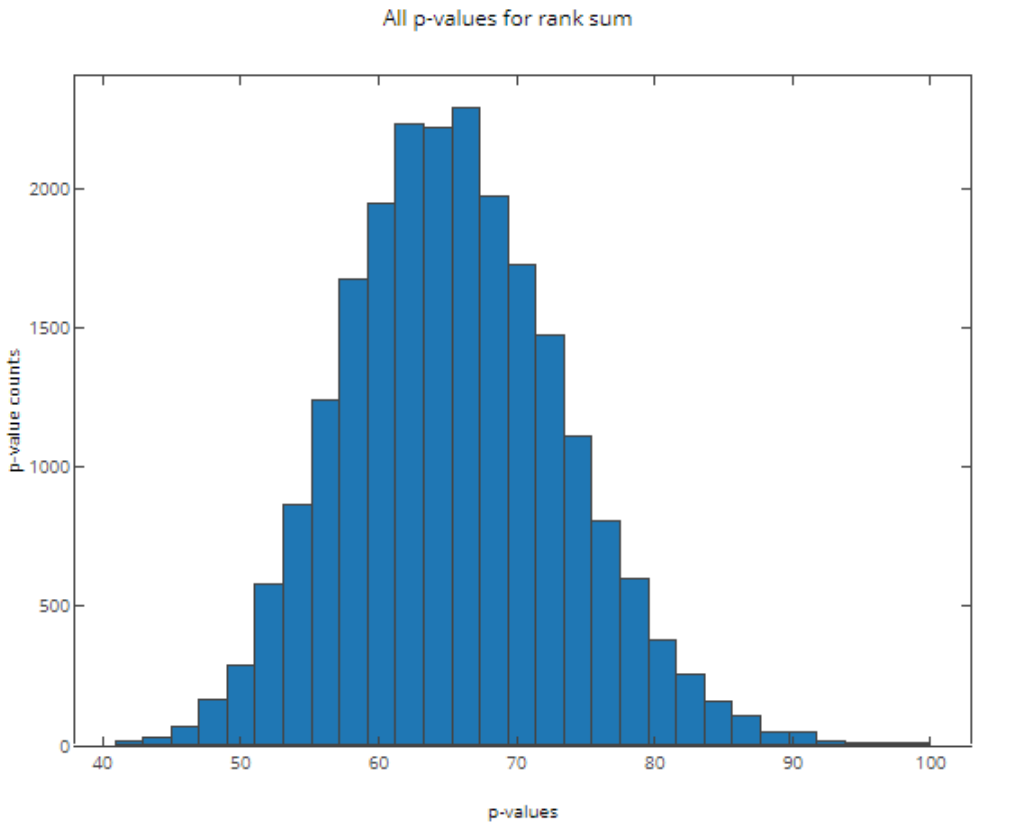
(b) Report the number of selected probes for each test and plot the histogram of the log transformed (negative log10) p-values of all probes.

T test: 3262 p-values of probes were < .05



Wilcoxon rank-sum: 22283 p-values of probes were < .05

The following graph is based on the incorrect data from the previous problem



(c) What is the overlap between the set of probes deemed significant by the two different approaches (t-test and Wilcoxon rank-sum statistics)? You do not need to report the entire list associated with each approach. You only need to report the total number of probes that overlap and a list of those overlapping probes.

5. Multiple hypothesis correction (30 points):

(a) Use Bonferroni correction with the rank-sum test to identify differentially expressed probes at a global significance level $p < 0.05$ and report the number of selected probes.

(b) Use the Benjamini-Hochberg step-up procedure to control the False Discovery Rate (FDR) with the rank-sum statistic to identify differentially expressed probes at an FDR < 0.05.Report the number of selected probes by the BH procedure (specify which independence assumption you are using for the BH procedure).

(c) Rank the probes by their p-values in ascending order and plot the p-values and the threshold used for adjusted p-values as a function of the index of the ranked probes. More specifically, with the x-axis as the index of the ranked probes from 1 to the number of probes, plot the first 500 probes (i on the x-axis, the p-value of the ith probe on the y-axis for each probe). As a separate color, plot the adjusted p-value threshold at each point (i on the x-axis, threshold for adjusted p-value on the y-axis), where i is the index of the genes and alpha=0.05.