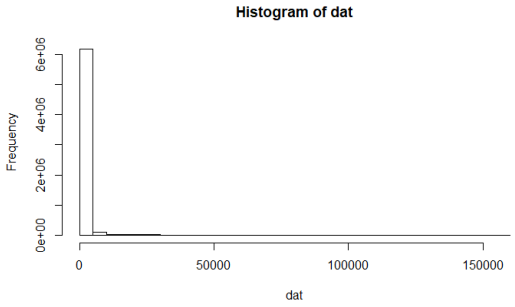
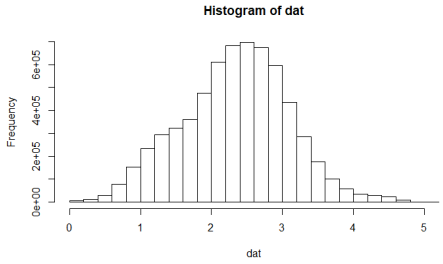
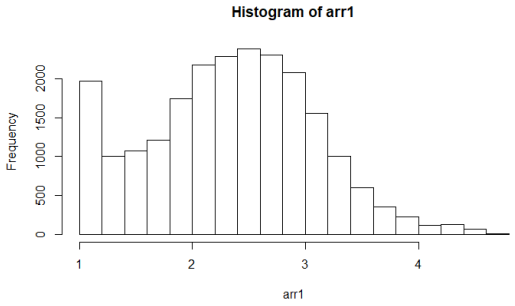
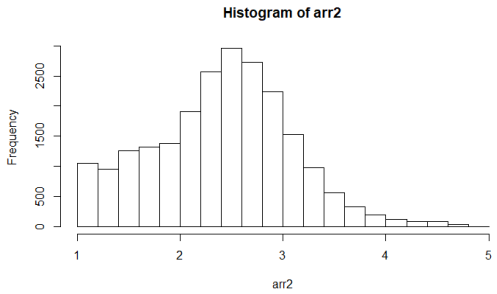
1. Cool done
2. Using the R programming language
    a. 22283
    b. There are 287 patients, 179 of which relapsed and 107 did not and one which probably died due the label of “NA”
    c. There would also be 22283 genes; one for each probe.
3. Data processing
    a. The distribution of the non-log transformed data is heavily distributed on the lower expression levels. Out of the 6,372,938 data points, 6,174,494 of them were in the range 0-5000. There was an extreme exponential drop-off all the way up to the most expressed gene which had a value of 157,291.8. After the log transforming, the data looks remarkably like a normal distribution. The plot on the left is the regular data and the plot on the right is the log transformed.

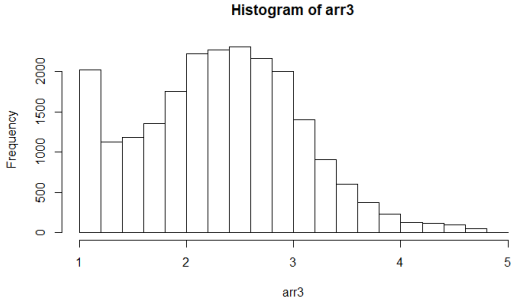


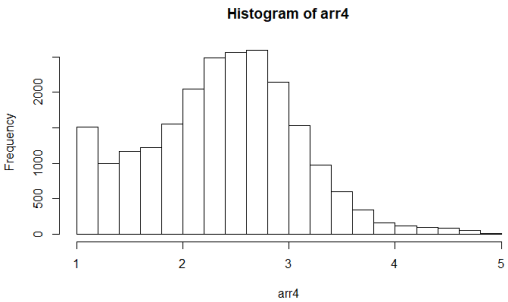


    b. All of the arrays are very similar. The one that sticks out the most would be array 2 which has more probe expressions in the 2-3 range

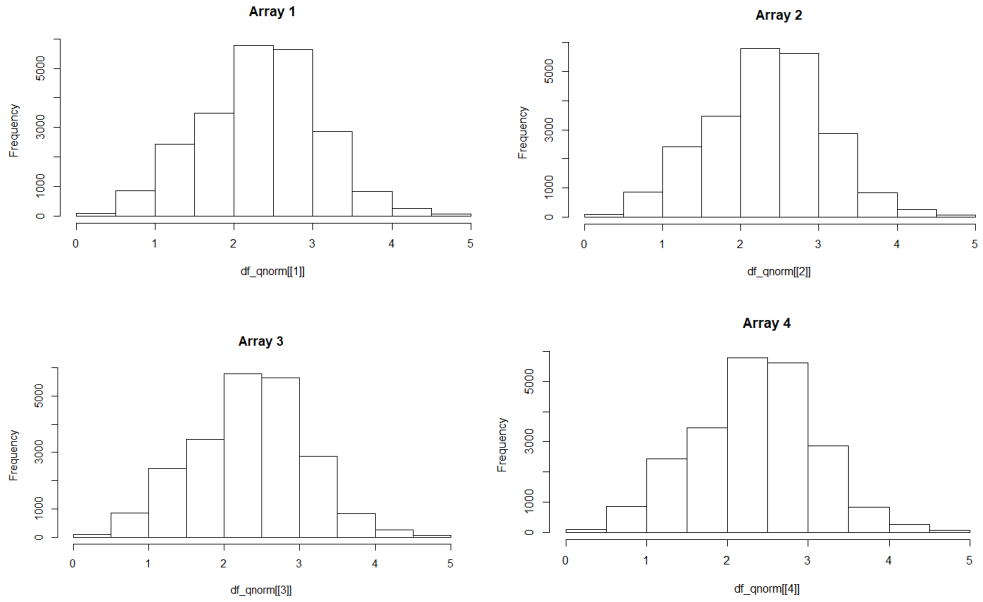








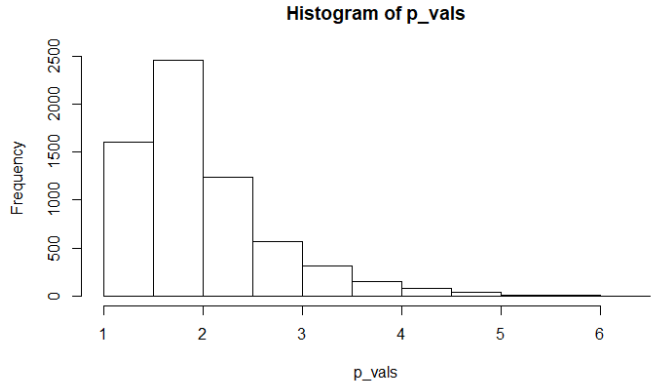
c. The normalized arrays are almost identical



4. Analysis of Differential Expression
   a. The t-test is on the top and the rank sum on the bottom

p-value	4.2659 e-07	8.4802 e-07	1.2217 e-06	2.1041 e-06	2.87055 e-06	2.9876 e-06	3.5879 e-06	4.0500 e-06	5.9785 e-06	6.5065 e-06
gene	ACBD3	WFDC1	CLINT1	RGS3	RACGAP1	NEK2	BOLA2	ZF36L2	ABCC5	FBXO7
p-value	7.2939 e-07	1.8074 e-06	2.3423 e-06	3.6213 e-06	4.6927 e-06	7.0922 e-06	7.1169 e-06	8.0884 e-06	8.1436 e-06	9.1858 e-06
gene	ACBD3	WFDC1	BLZF1	CLINT1	ZF36L2	RACGAP1	NEK2	LACTB2	SHC1	SEC24A

   a. The gene ACBD3 has the lowest p-value for each test. GeneCards says its function is “Among its related pathways are Clathrin derived vesicle budding and Vesicle-mediated transport.” And also, the internet told me that it gives instructions for making transporter proteins. Multiple studies say that it is a biomarker for aggressive and metastatic cancers, but the functional role is not completely elucidated as one study so eloquently put it.
   b. T-test selected 3197 genes and rank sum selected 3285 genes.



   c. 2679
5. Multiple Hypothesis Correction

a. The number of selected probes fell to 3
b. The independence assumption is that each gene expression level does not affect one another. Controlling for the FDR, there were 177 significant genes.
c.