Differential Expression of Genes in Ductal and Lobular Neoplasms between Black and White Females with the TP53 Mutation

Christiana Leone and Michael Zeosky

University of Texas at Austin  
BIO 321G

# 

# 

# Methods

The primary goal of this study is to utilize TCGA-BRCA data to analyze the differences in gene expressions between black and white females who are found to have a mutated TP53 gene and are affected with ductal and lobular neoplasms. The Cancer Genomics Atlas (TCGA) is a program that has collected and molecularly characterized a vast number and range of cancers in order to try and determine what mutations may have caused them through genomic analysis. The Cancer Genomics Atlas Breast Invasive Carcinoma (TCGA-BRCA) collection is a project of this program, specifically analyzing genomic data relating to breast cancer. This data can be analyzed through the Genomic Data Commons (GDC) portal, which was utilized in this project.(Grossman et al. 2016) The GDC allowed us to isolate the data we were sampling through a set of filters. We utilized these filters to specify for only females from the TCGA-BRCA project that were affected in their breast with ductal and lobular neoplasms and were found to have a mutation in their TP53 gene. By then filtering by race, this sample was split between black and white females into the two groups that would be compared.

* Control: White females found in TCGA-BRCA with ductal and lobular neoplasms and a mutation in TP53. N = 23
* Experimental: Black females found in TCGA-BRCA with ductal and lobular neoplasms and a mutation in TP53. N = 22

This genomic data was then retrieved and downloaded from the GDC utilizing the TCGAbiolinks tool. (Colaprico & Silva, 2017) This data contained 43 subjects, originally including 23 samples in our control group and 22 samples in our experimental group. Utilizing Rstudio, this was then filtered to remove two individuals from group 1 with extremely low RD and one individual from group 2 who had a duplicate ID.(Rstudio Team, 2020) After filtering those individuals out, we were left with 21 samples for each group. Finally, we analyzed these genomes using the DESeq2 package in order to determine which genes had significant differential expression between the two groups. (Love et al. 2014) We set parameters as cutoffs to determine significance, where the absolute log2 fold change values had to be greater than 1.5 and the p value needed to be lower than 0.01. Furthermore, we created supplementary information which led us to further filtering out 4 sample outliers – 2 samples from each group. This left us with 19 samples for each group. We then reran DESeq and recreated visualizations while having 19 samples from each group – we created a table with the most significant genes and their corresponding values as Figure 1; a volcano plot to see which genes are significantly up-regulated and down-regulated as Figure 2; two pca plots to show repetition after further filtration to show good samples as Figures 3 and 4; a boxplot to visualize gene expression within our most significant gene between our two groups, black and white females as Figure 5; and a histogram expressing the frequency of loci in each grouping as Figure 6.

# Results

There were 1,031 loci that were found to have significantly different expression levels between our two groups of the 42,357 loci analyzed, meaning they met the absolute log fold change as well as adjusted p-value requirements for significance. Our gene that saw the most significant difference in expression levels between the two groups was MTND1P23, which had a log fold change of 7.8925037 and a p-value of 1.014609e-24, shown in Figure 1. Figure 1 displays a list of the 9 most significant genes in the TGCA-BRCA genome, including the gene names with their corresponding log fold change values and adjusted p-values.

Using information from our log2FoldChange and our -log10 p-adjusted value, a volcano plot was created as Figure 2. There were 1,031 genes that were significant – 585 up-regulated and 446 down-regulated. There were 19 control samples (white females in group 1) compared to 19 experimental samples (black females in group 2) after filtering out 4 samples in group 1 and 3 samples in group 2. MTND1P23, MTND4P24, MTCO1P40, IGLVI-70, and RPL10P9 are the most differentially expressed genes with an average adjusted p-value of 3.455501e-8. The MTND1P32 gene shows the highest significance with an adjusted p-value of 1.014609e-24 and proved to be the most up-regulated.

Next, we created a pca graph which converts the correlation or lack thereof among all of the cells into a 2D graph (Statquest, 2017). *After* the filtration of two individuals from group 1 with extremely low RD and one individual from group 2 who had a duplicate ID, and *prior* to finding 4 more sample outliers that were later removed, Figure 3 was created. The pca plot in Figure 3 shows some scattering on the left side, along with 4 sample outliers. In order to investigate deeper, we removed the 4 outlier samples, which had 2 samples from each group. After filtering the 4 sample outliers, we were left with the pca plot in Figure 4. Figure 4 shows much less scattering. There is some difference; the upper left side is tending to be the red group and bottom right side is tending to be the blue group, but overall, after filtering, there are no obvious trends or clustering.

Figure 5 shows the boxplot presented from our data. The label “false” is representing the white females and the label “true” is representing the black/African American females. In Figure 5, we can see that the read depths are higher for the black/African Americans distribution. We can also see the median and mean are higher for the black patients.

Figure 6 displays a histogram of the frequency of loci in each group – black and white patients. Figure 6 did not display any outliers and served as supplemental information.

\*\*Figures 1-6 are displayed below.

Figure 1. Table of Significant Genes for TGCA-BRCA

| **Gene Name** | **Log Fold Change** | **P-Value (adjusted)** |
| --- | --- | --- |
| MTND1P23 | 7.8925037 | 1.014609e-24 |
| MTND4P24 | 4.8592839 | 6.446453e-16 |
| MTCO1P40 | 5.2355553 | 3.263060e-13 |
| IGLVI-70 | 6.9935529 | 9.514380e-11 |
| RPL10P9 | 3.5893494 | 1.726796e-07 |
| IGHV7-4-1 | 5.3123443 | 4.902610e-07 |
| AC119673.2 | -3.3176708 | 7.059671e-07 |
| AC010894.2 | 4.3998724 | 7.551571e-07 |
| WFDC21P | 3.8023196 | 8.149079e-07 |

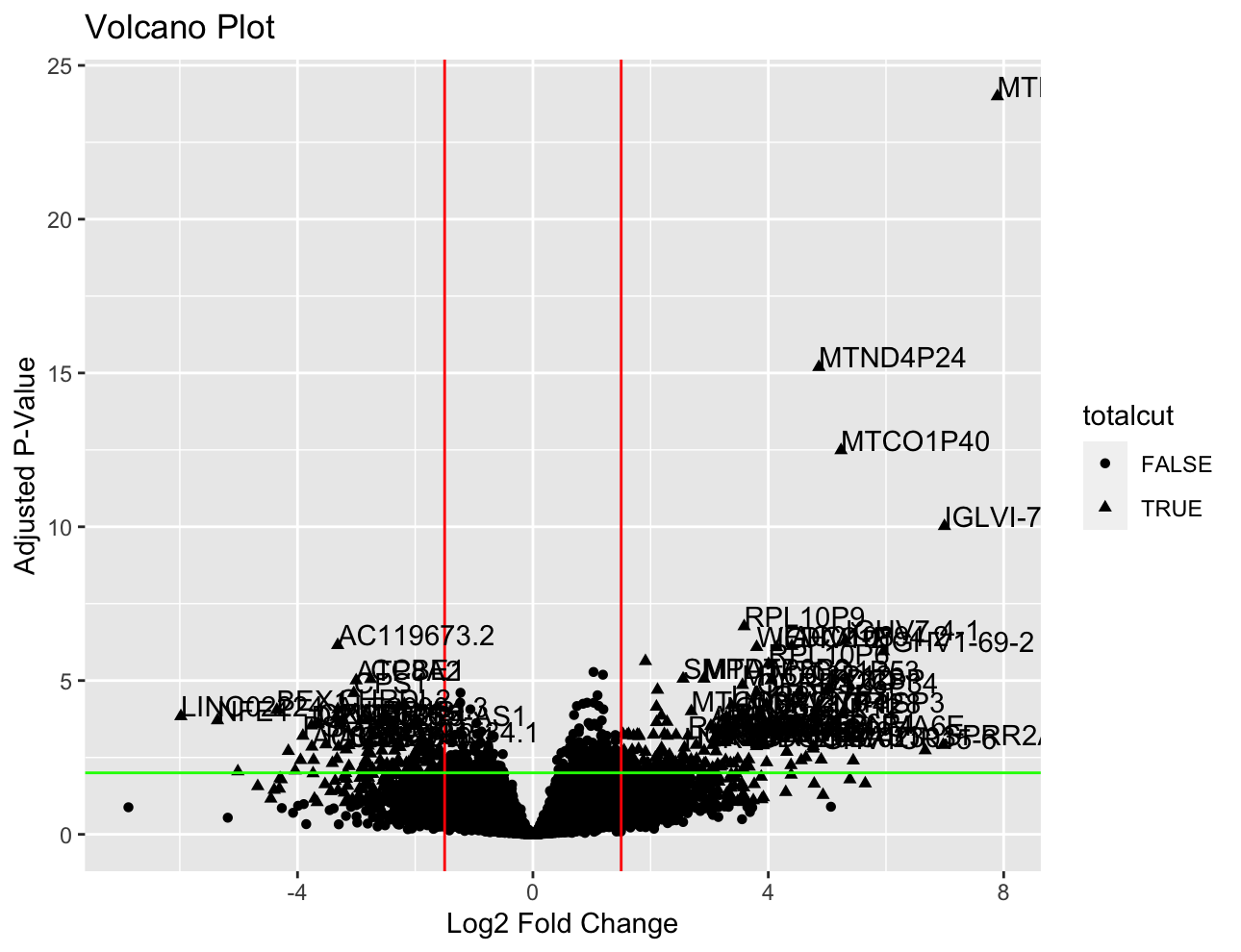
Figure 2. Volcano Plot of Gene Expression Between White and Black Females

Figure 3. PCA Plot of Sample Data prior to Filtering (including outliers)

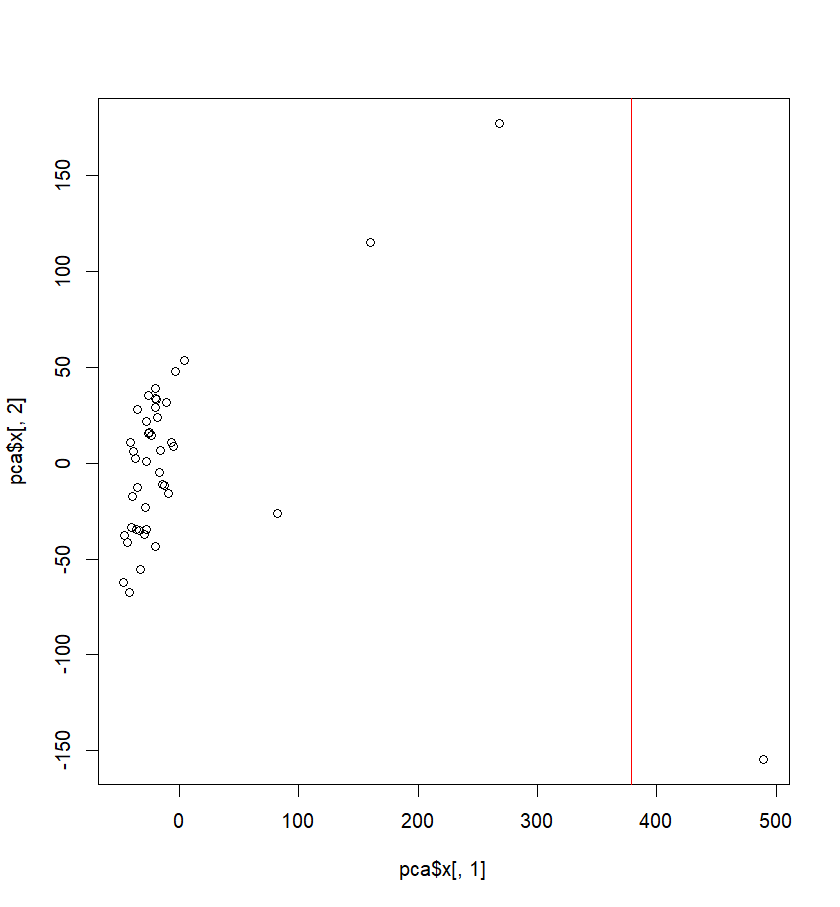


Figure 4. PCA Plot of Sample Data after Filtering (excluding outliers, control N=19, exp. N=19)

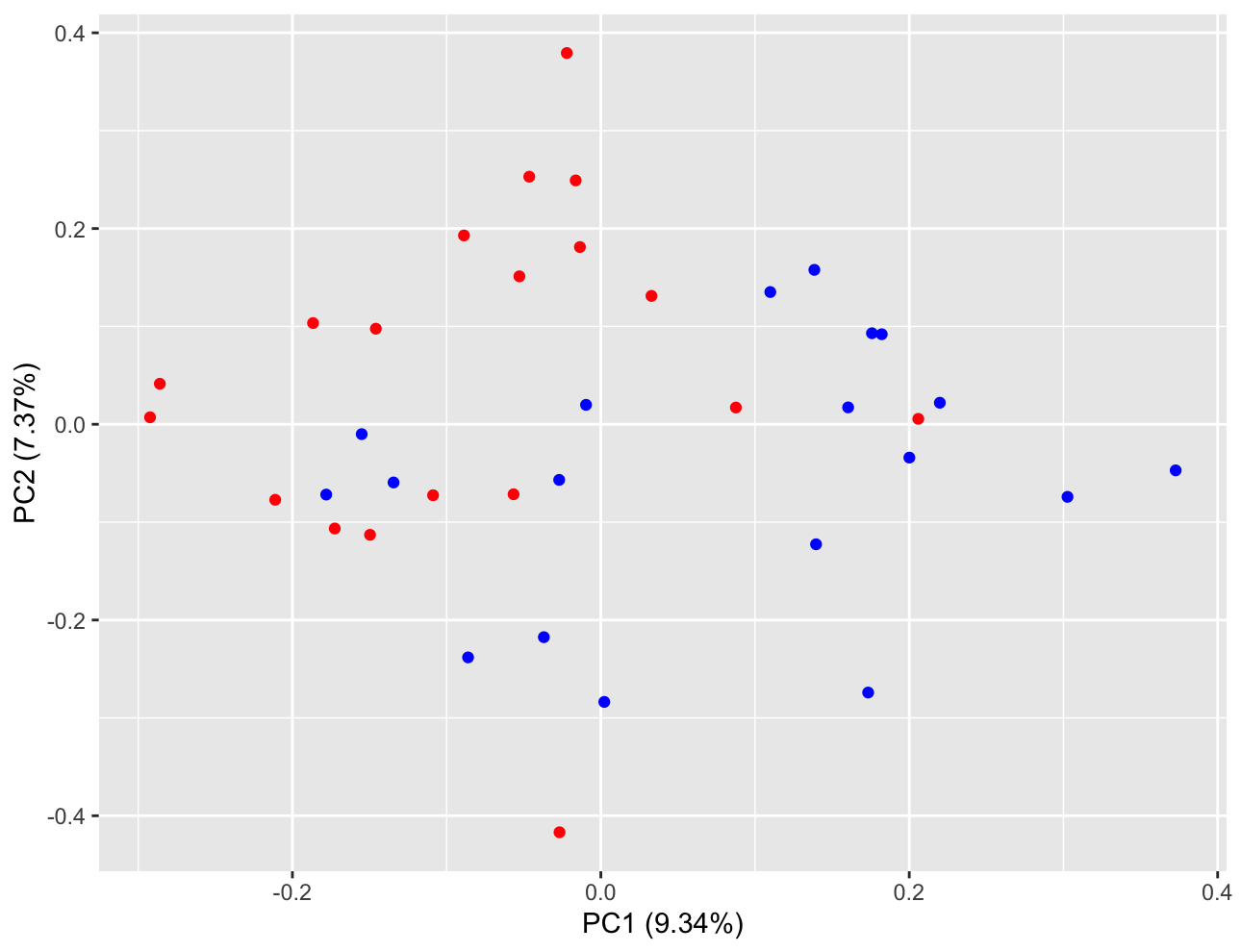


Figure 5: Boxplot of Normalized Read Depth Counts for MTND1P23

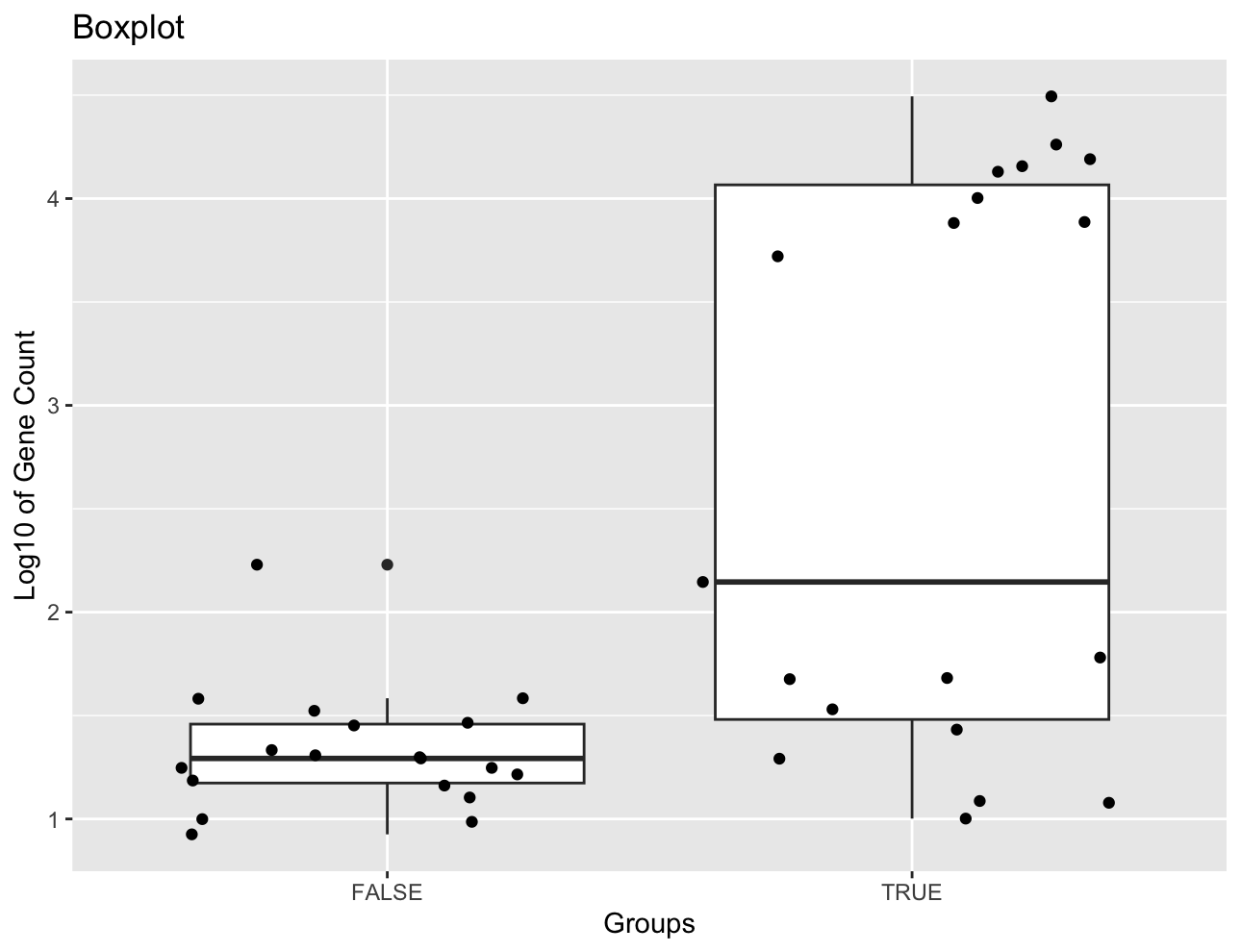
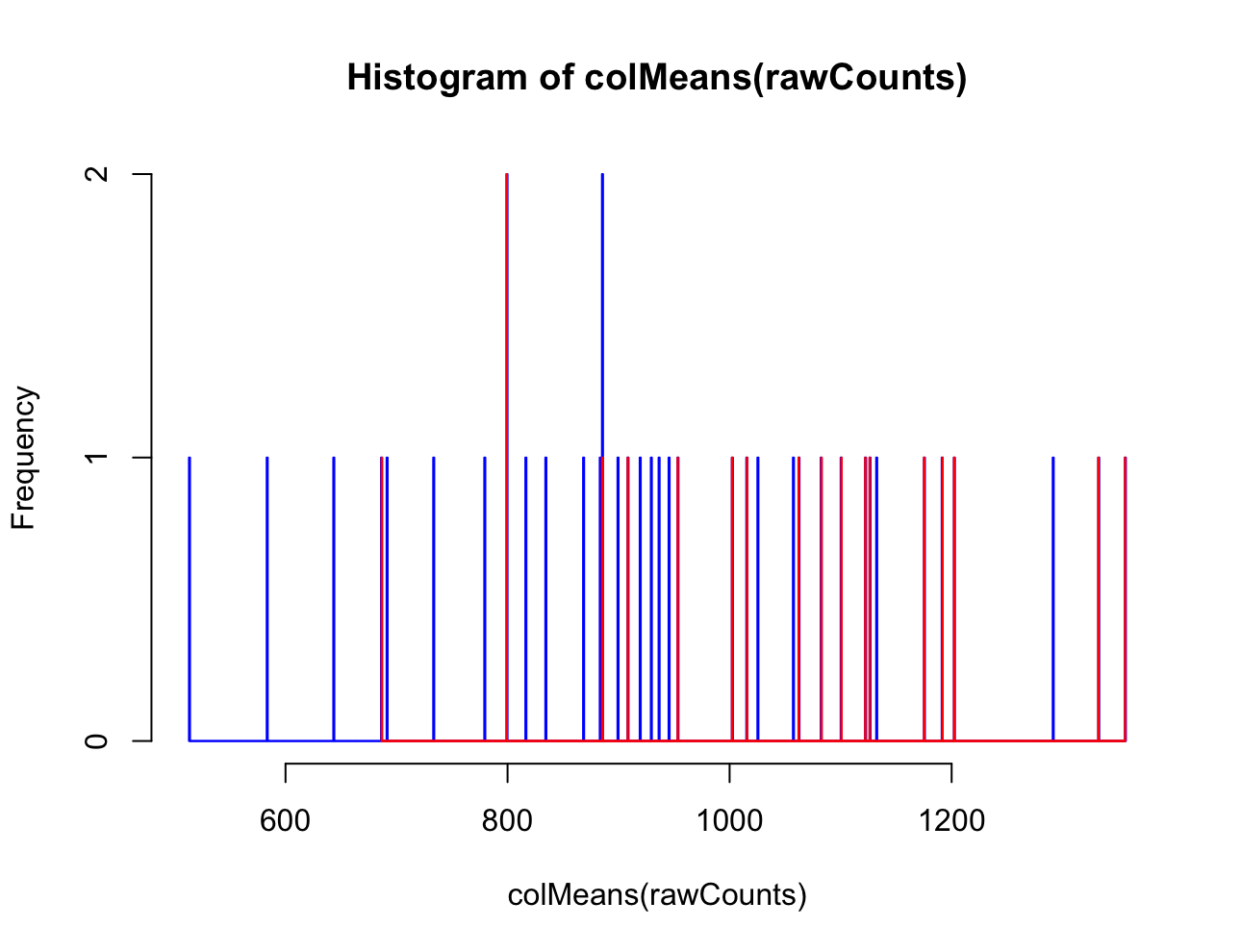


Figure 6: Histogram of the Frequency of Loci in Each Group​​

**Analysis**

Figure 1 is an important part of this research because it showed the most significant difference in expression levels between the two groups of TGCA-BRCA samples are found in the MTND1P23 gene, alongside some other significant genes. The very low p-values show that it is incredibly unlikely that these genes expression rates were only different between the two groups due to chance. This figure can be interpreted by noting that these specific genes saw significantly higher expression rates in the black/African American samples than the white samples if they had positive log fold values, and significantly lower expression rates in black/African American samples in genes with negative log fold values.

Figure 2 displays a volcano plot which compares gene expression between black females (false) and white females (true). Figure 2 displayed that the MTND1P23 gene was the most up-regulated, which can be interpreted by understanding that the MTND1P23 gene is most significantly expressed in black females. Furthermore, with the filters we’ve applied, MTND1P23 is at least expressed a little bit in both groups.

Figures 3 and 4 are both included as supplementary information because it shows the difference between our boxplots with and without outliers and proves repetition to show that the samples look good. In the pca plot in Figure 3, 4 outliers are shown and more experiments should be pursued to further interpret this. In the pca plot in Figure 4, based on the gene expression, there is very little overall difference between our two groups (black and white), meaning the data is largely similar. There is some difference on the top left side tending to be the red group and bottom right side tending to be the blue group, but overall after filtering, there is no sample that sticks way out. We don’t know what it is, but something about those 4 sample outliers (that were later removed from our groups) are different from the rest of the samples. Furthermore, our PC1 was 9.34% and our PC2 was 7.37% which are both really low percentages of variance, meaning the variance is not explained well and there may be less of a difference between our white and black groups after all.

Upon analysis for the boxplot for Figure 5, we should not examine the two distributions to compare their medians, means, or standard deviations. Even though we can see there are more read depths in the true group and assume that we can interpret that as more expression, that may not be the case. When we added jitter points to help visualize what’s happening, we saw two different trends in the true (black) distribution which affect our calculations. This data can only be interpreted so far as there being a gene in one half of the black patients that is massively up-regulated, about 100x more than the other half of black patients, which is why DEseq is detecting it as significant. The two trends on the true distribution cannot be explained biologically and these RNA sequencing results need to be viewed skeptically because the p-value hides a lot of things. Upon further investigation, a possibility to explain this occurrence could be that there is a pre-existing gene that segregates within the black/African American population, causing the gene to behave a little differently about half of the time.

Figure 6 can be examined by there being no display of repetition, meaning we filtered appropriately and there was nothing wrong with our sampling data. Therefore, this proved that there were not any red flags and allowed us to confirm the accuracy of our sample data.

The goal of this study is to utilize TCGA-BRCA data to analyze the differences in gene expressions between black and white females who are found to have a mutated TP53 gene *and* are affected with ductal and lobular neoplasms. It was found that black/African American women express the MTND1P23 significantly more than white women, which may mean that there is a loss of function of the protein (Prelich, 2012). Based on another study, significant gene expression could be due to genetic makeup, exposure to harmful substances, or environment (Yedjou et al. 2019). Another study states “White women are slightly more likely to develop breast cancer than Black, Hispanic, and Asian women. But Black women are more likely to develop more aggressive, more advanced-stage breast cancer that is diagnosed at a young age. Black women are also more likely to die from breast cancer.”(“*Race/Ethnicity*”, 2023) Further research could focus on investigating the cause for the significant increase found in a large portion of the black women sampled. Throughout this study, it was not determined that either the white or black race of women are more likely to present ductal and lobular neoplasms than the other. In conclusion, while this study shed light on a difference in expression rates of the MTND1P23 gene between black and white females who are found to have a mutated TP53 gene and are affected with ductal and lobular neoplasms, there is still a lot of further investigation is required to understand what this data means, the trends that are examined, and how the information is placed within the context of the goals of this experiment.

# 

# Works Cited

Antonio Colaprico, Tiago Chedraoui Silva. *TCGAbiolinks*. Bioconductor, 2017. *DOI.org (Datacite)*, <https://doi.org/10.18129/B9.BIOC.TCGABIOLINKS>.

Grossman, Robert L., Heath, Allison P., Ferretti, Vincent, Varmus, Harold E., Lowy, Douglas R., Kibbe, Warren A., Staudt, Louis M. (2016) Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine375:12, 1109-1112

Love MI, Huber W, Anders S (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, **15**, 550. doi: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).

Prelich, Gregory. “Gene Overexpression: Uses, Mechanisms, and Interpretation.” *Genetics*, vol. 190, no. 3, Mar. 2012, pp. 841–54. *DOI.org (Crossref)*, https://doi.org/10.1534/genetics.111.136911.

*Race/Ethnicity*. (2023) https://www.breastcancer.org/risk/risk-factors/race-ethnicity. Accessed 21 Apr. 2023.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.

*StatQuest: PCA Main Ideas in Only 5 Minutes!!!* [*www.youtube.com*](http://www.youtube.com), (2017) https://www.youtube.com/watch?v=HMOI\_lkzW08. Accessed 21 Apr. 2023.

Yedjou, Clement G., et al. “Health and Racial Disparity in Breast Cancer.” *Breast Cancer Metastasis and Drug Resistance*, edited by Aamir Ahmad, vol. 1152, Springer International Publishing, 2019, pp. 31–49. *DOI.org (Crossref)*, https://doi.org/10.1007/978-3-030-20301-6\_3.

The results shown here are in whole or part based upon data generated by the TCGA Research Network: [https://www.cancer.gov/tcga](https://www.cancer.gov/ccg/research/genome-sequencing/tcga).