Chloe Schwieger (schwi415@umn.edu)
Maitrayee Deka (deka0031@umn.edu)

CSCI 5461 Homework 1

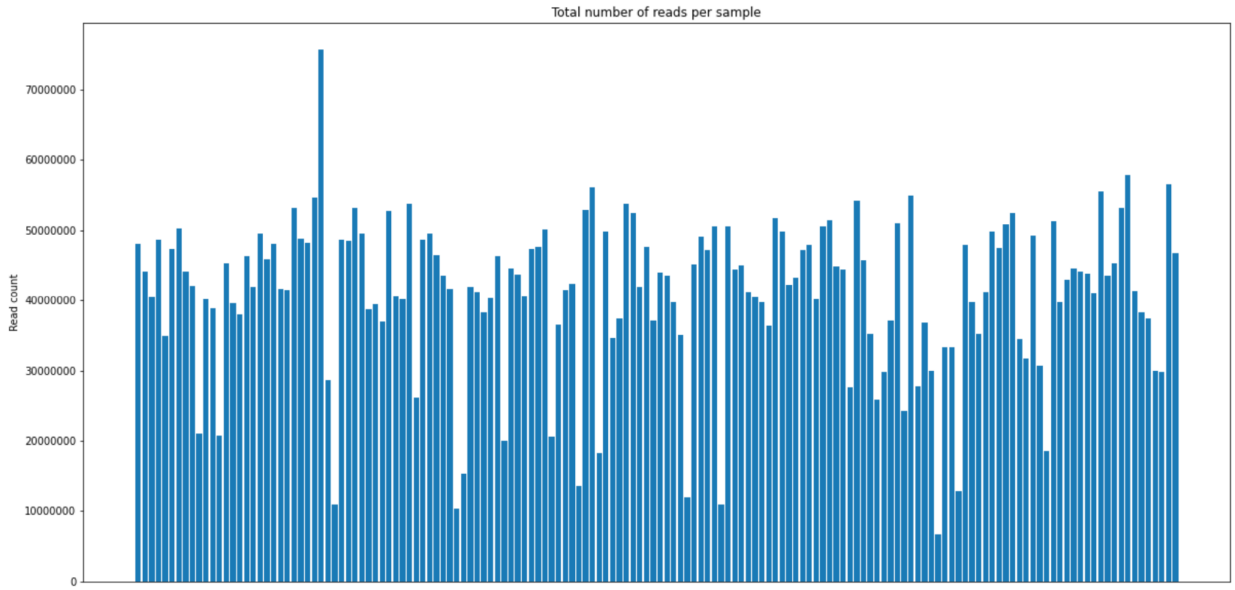
1. **Acquiring the gene expression data**
    a. Additional Question: After observing some issues when processing data using The Cancer Genome Atlas (TCGA) pipeline, they used an alternative method. Instead, Rsubread was utilized; it is a package that implements Subread algorithms and produces data with fewer zero-expression genes and more consistent expression levels across replicate samples when compared to the TCGA pipeline. The study got the raw sequencing data for 9264 tumor samples and 741 normal samples across 24 cancer types from the TCGA database and reprocessed the data using the Subread algorithm. Then, they calculated the normalized values of fragments per kilobase of exon per million reads mapped (FPKM) and transcripts per million (TPM)

2. **Exploring the dataset (10 points):**
    a. There are 23368 genes in the dataset.
    b. There are 154 patient samples in the dataset. Of these, 76 were in the short-term group, and 78 were in the long-term group.
    c. Additional Question: This research aims to identify the genetic alterations and core pathways that contribute to the development of glioblastoma. This is part of The Cancer Genome Atlas (TCGA) project, whose goal is to create a library of major cancer-causing genome alterations through integrated multi-dimensional analyses.

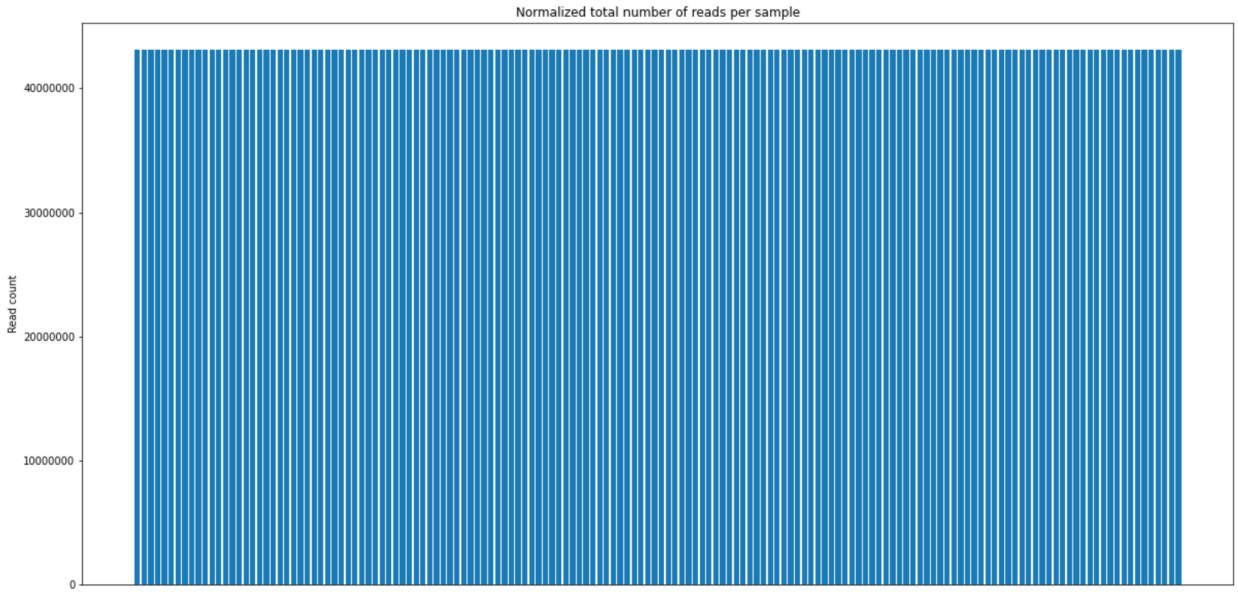
    They looked at the DNA copy number, gene expression and DNA methylation aberrations in both treated and untreated glioblastoma samples. 206 samples were qualified for the aforementioned analysis, but only 143 passed further quality control for resequencing. They looked for patterns in alterations via microarrays of genomic copy number alterations to find unreported focal alterations, like that of homozygous deletions with *NF1* and *PARK2* and the amplification of *AKT3*. Then, of the 143 samples, 91 were paired with healthy samples to examine somatic hypermutations of 601 genes. Of which, they noted many non-silent mutations that differed between treated and untreated glioblastoma as well as some genes with significant mutations, *p53* mutations, for example, were a common event in primary glioblastoma.

Overall, the creation of multi-dimensional data sets using strictly chosen samples has allowed for the discovery and verification of glioblastoma genes, core pathways, and possible interventions, showing the value of the TCGA project.

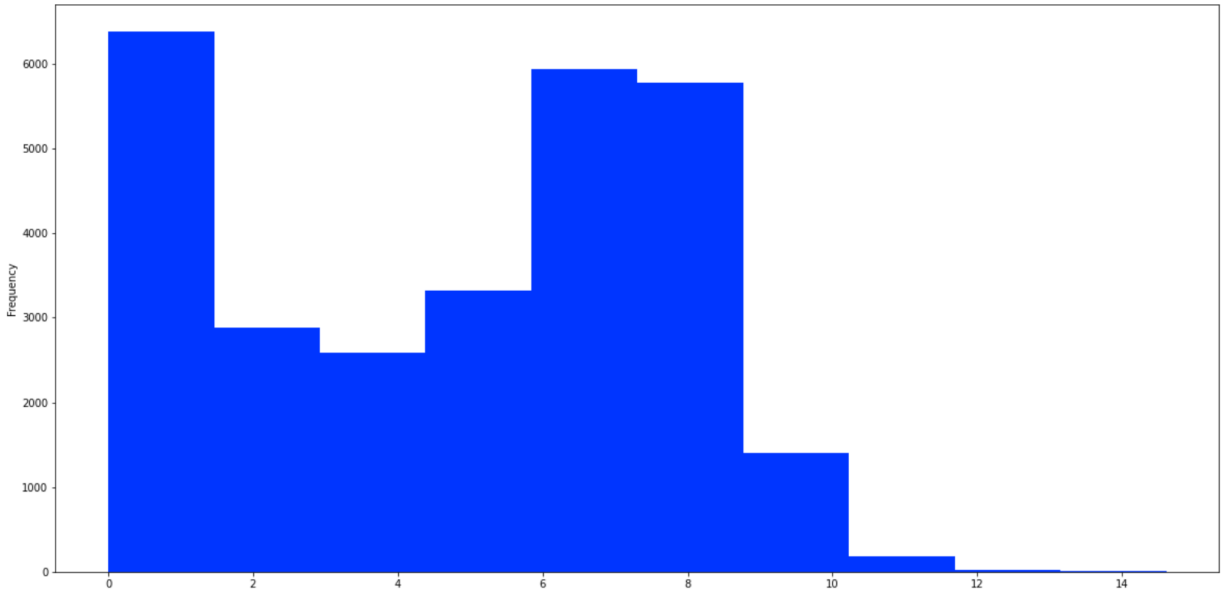
3. **Data processing and normalization (30 points):**
    a. From the plot of total reads per sample, we can observe that the values range from a minimum of approximately 6 million to a maximum of about 75 million. On average, the total number of reads appears to hover around 40 million.



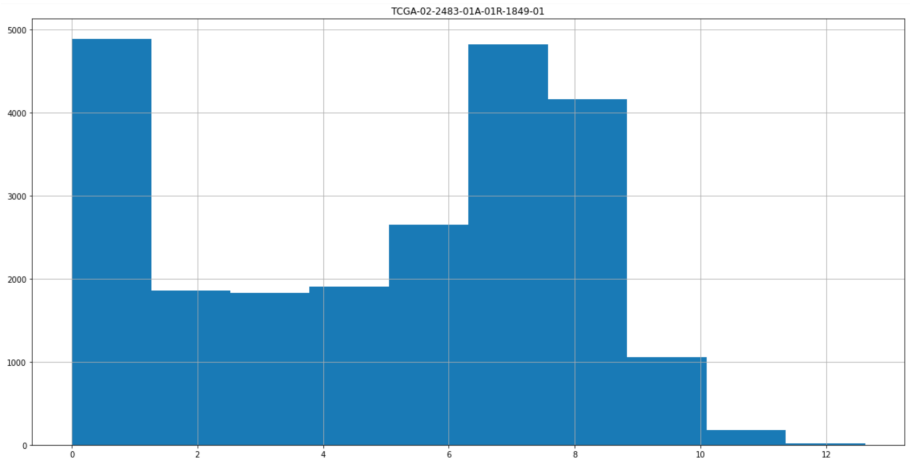
    b. After performing total count normalization, we can see that the total number of reads per sample now appears to be equal across all samples, as shown below.

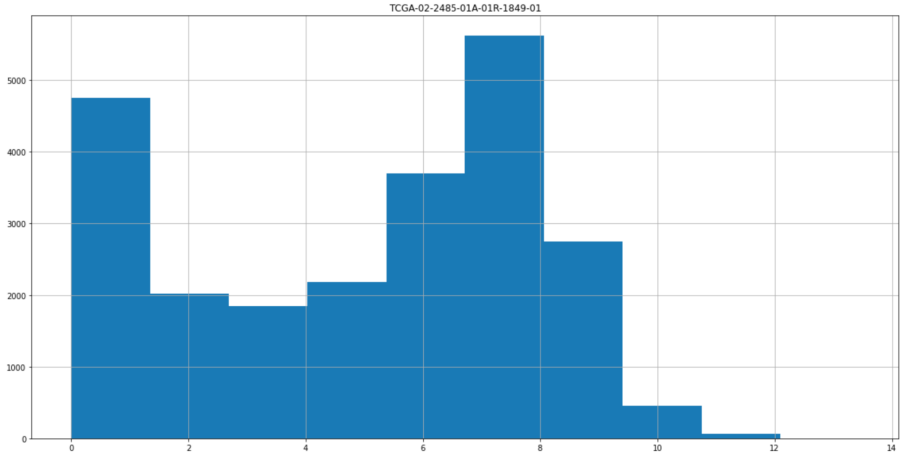


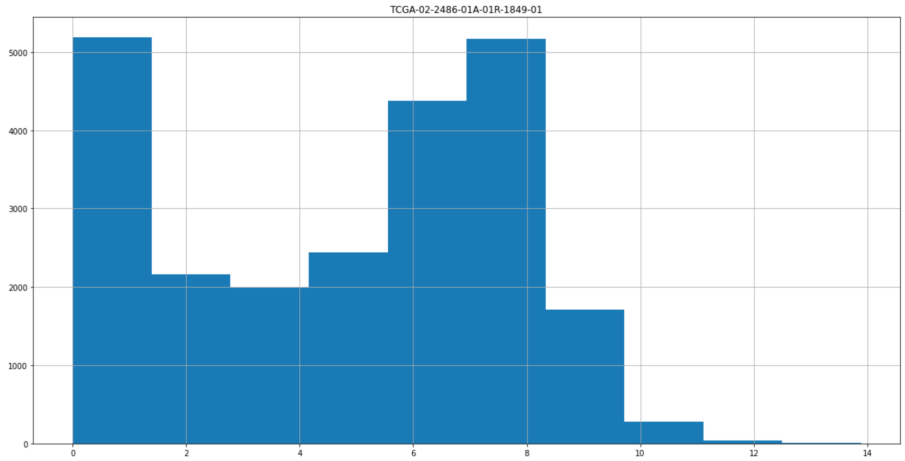
c. Log-transformation Plot

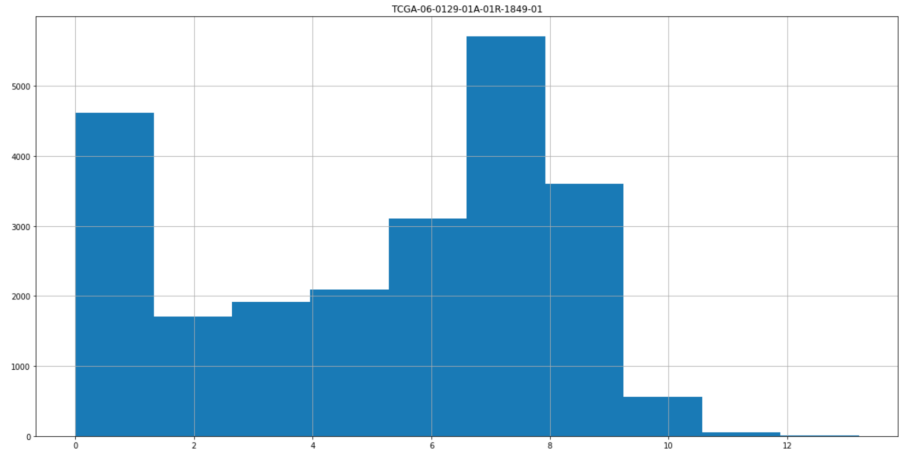


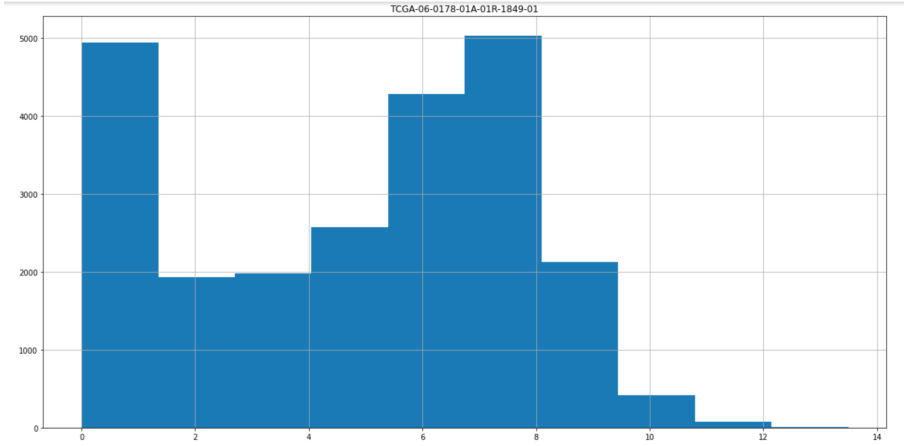
d. Overall, there is a similar distribution of log transformed data between the 5 plots. There are many counts in the general range of 0-1 as the 6-8 region with a steady number of counts between the two regions. There are also very few counts above 10 or 11. This holds true for all 5 samples.


TCGA-02-2483-01A-01R-1849-01
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12

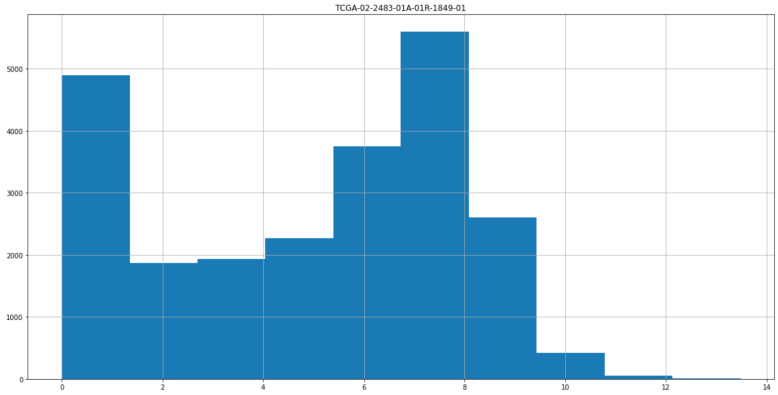

TCGA-02-2485-01A-01R-1849-01
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12
14

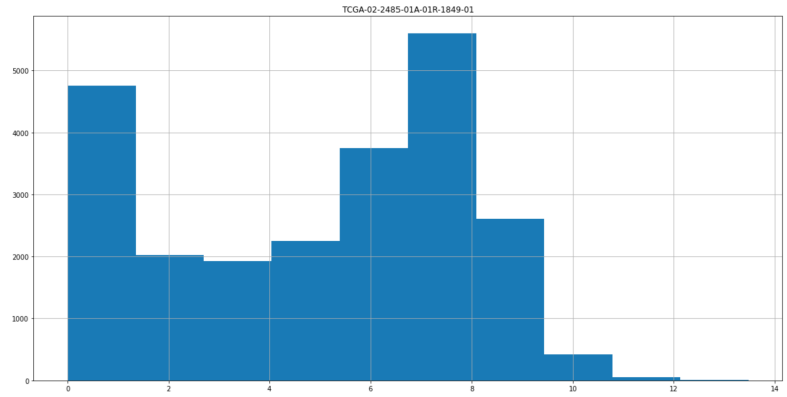

TCGA-02-2486-01A-01R-1849-01
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12
14

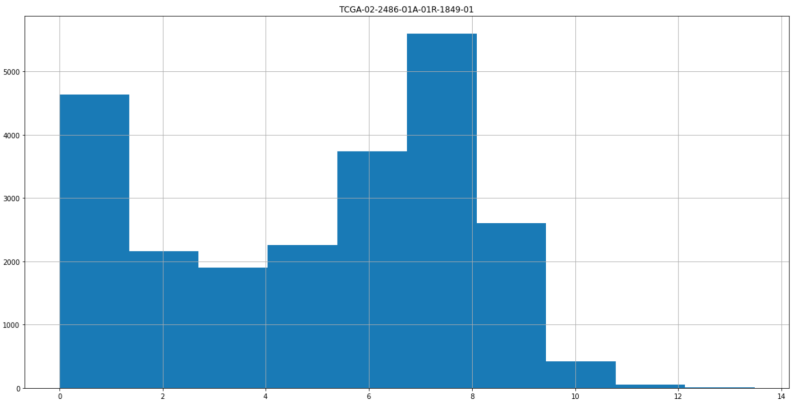


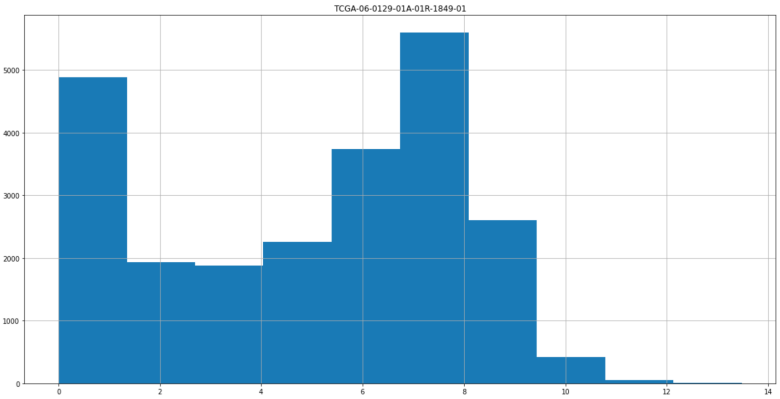


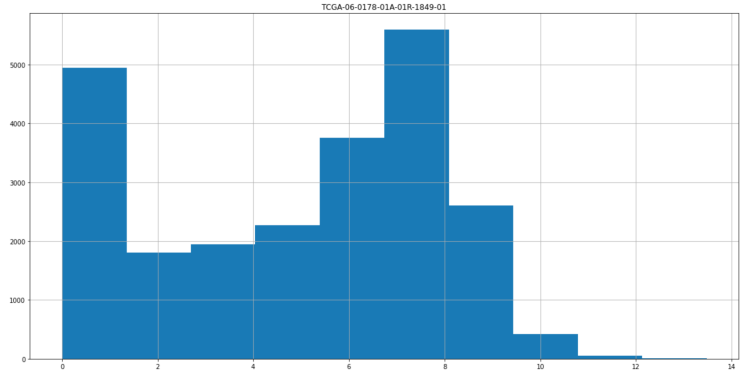
e. Quantile Normalization Plots











4. Analysis of differential expression (30 points):

a. Genes of Significance
    i. *PIK3CA*: The *PIK3CA* gene encodes a protein called phosphatidylinositol 3-kinase (PI3K), which is involved in various cellular processes such as cell growth, proliferation, and survival. Mutations in the *PIK3CA* gene can cause overactivation of the PI3K signaling pathway, leading to uncontrolled cell growth and proliferation. In glioblastoma, the *PIK3CA* gene mutations are often found in conjunction with other genetic alterations such as *TP53* and *EGFR* gene mutations. These mutations can interact with and contribute to the effects of each other, furthering the development and progression of glioblastoma.
    ii. *KCTD3*: *KCTD3* seems to act as a tumor suppressor gene in glioblastoma, thus its loss of function contributes to glioblastoma tumorigenesis. *KCTD3* has also been shown to regulate the activity of the NF-κB signaling pathway, which is involved in cell survival, inflammation, and immune response, and its dysregulation can promote glioblastoma progression.

	Gene Name	Statistic	pvalue
9071	KCTD3	4.611278	0.000004
16607	PRPF40A	4.276997	0.000019
3248	CDC73	4.249893	0.000021
20944	TMEM191A	-3.971626	0.000071
15858	PIK3CA	3.801775	0.000144
13	AACS	-3.713236	0.000205
22365	WFDC2	-3.695166	0.000220
18175	SCYL3	3.610241	0.000306
8858	ITPKA	-3.579523	0.000344
4706	DDX59	3.566875	0.000361

b. There are 1573 significant genes
c. There are 1010 overlapping genes, see overlapping_genes.csv for list
d. Additional Question: Of our top 10 most significant genes, 8 seem to have sufficient and detailed research into their function and connection to glioblastoma development and progression while 2 (*DDX9* and *TMEM191A*) seem promising, yet somewhat ambiguous. Other notable genes, including those from the papers, include *TP53, EGFR, PTEN, CDKN2A/B, CDK4, MGMT,* and *TERT*. These genes include anything from tumor suppressor genes, to different enzymes, kinases, receptors, and signaling molecules.

5. **Multiple hypothesis correction (30 points):**
    a. There are 20 significant genes in the DESeq2 results after using Bonferroni Correction
    b. There are 264 significant genes after applying the BH procedure. With Benjamini-Hotchberg, we are assuming that genes are independent - the probability of observing a significant result for one test does not increase or decrease the probability of observing a significant result for any other test. BH adjusts the p-values individually to control the FDR.
    c. BH Plot