Lab 1

2. Exploring the dataset

(a) How many genes are included in the dataset?

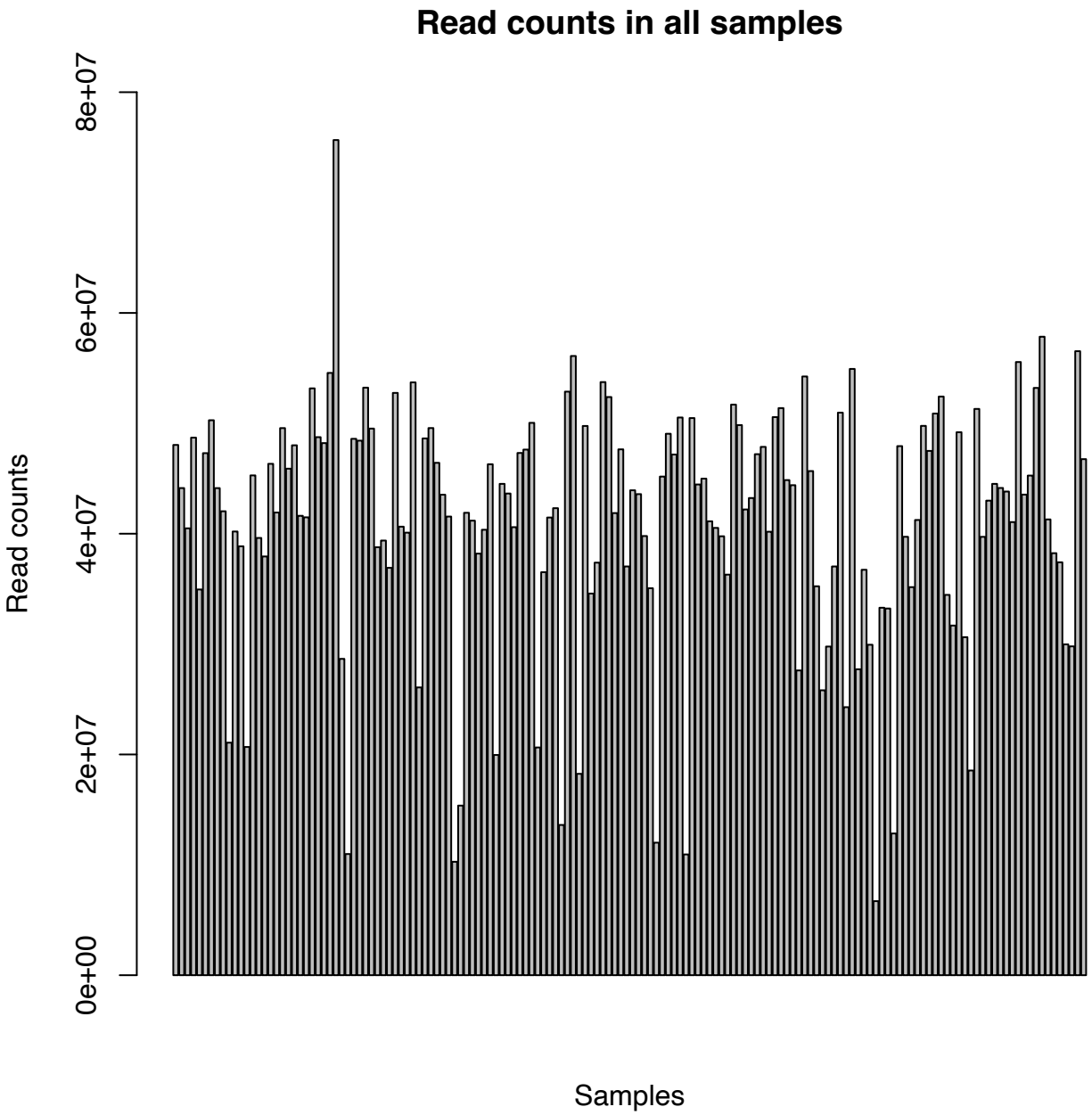
For count data direct from the RNA-seq, there are 23,368 genes. For dataset from DESeq2 analysis, there are 22,660 genes. The variation in gene counts between these two datasets maybe because during DESeq2 analysis some genes with very low abundant reads were removed.

(b) How many patient samples are in the dataset? We have divided patients into two “Groups” based on their survival time. How many patients survived short- (<= 1 year) and long-term (>1 year)?

There are a total of 154 patient samples in the dataset, and 76 of them are short-term patients, while 78 of them are long-term.

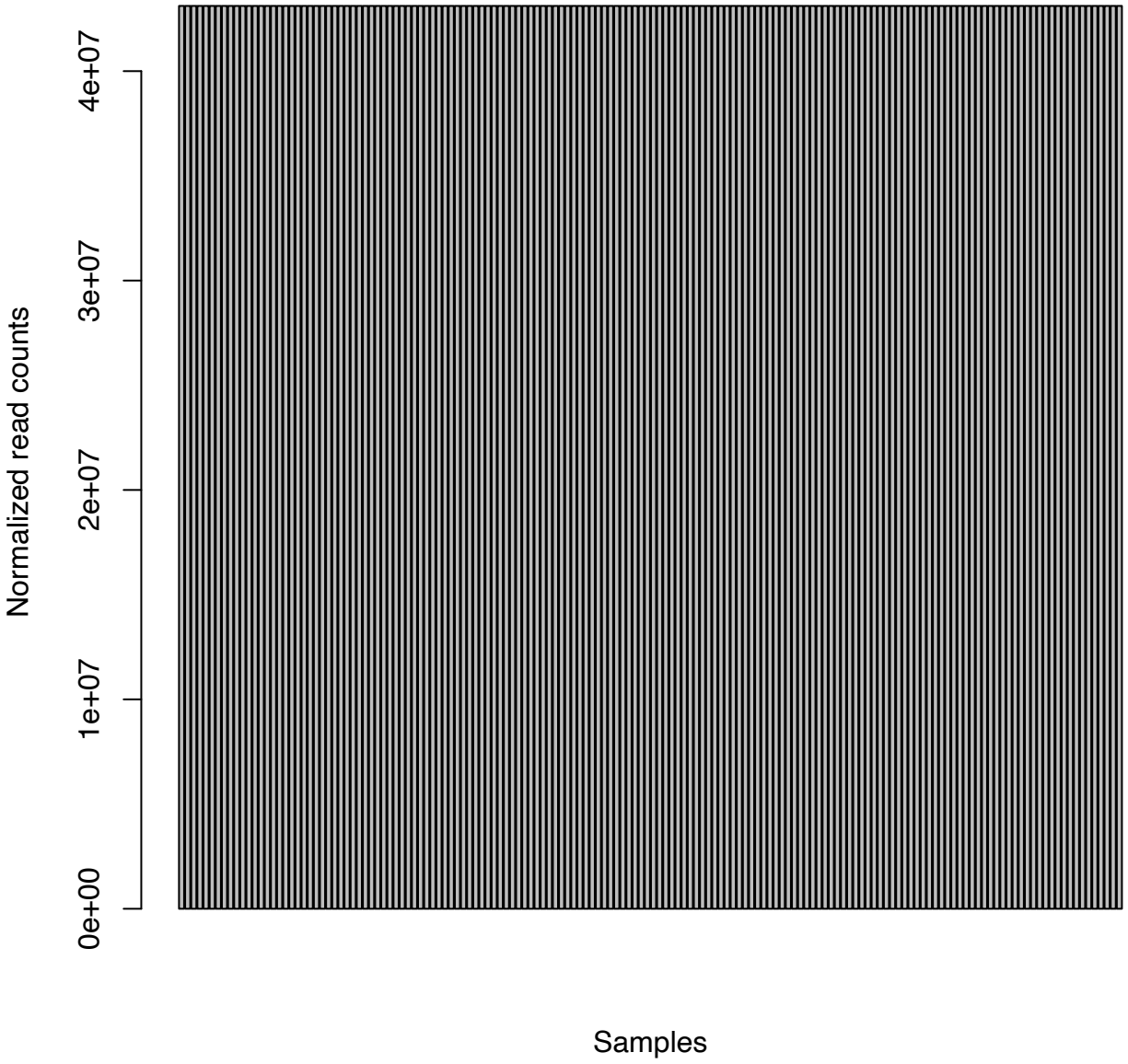
3. Data processing and normalization

(a) First, compute the total number of reads per sample and plot a bar graph that summarizes these values (one value per sample). Describe what you observe.



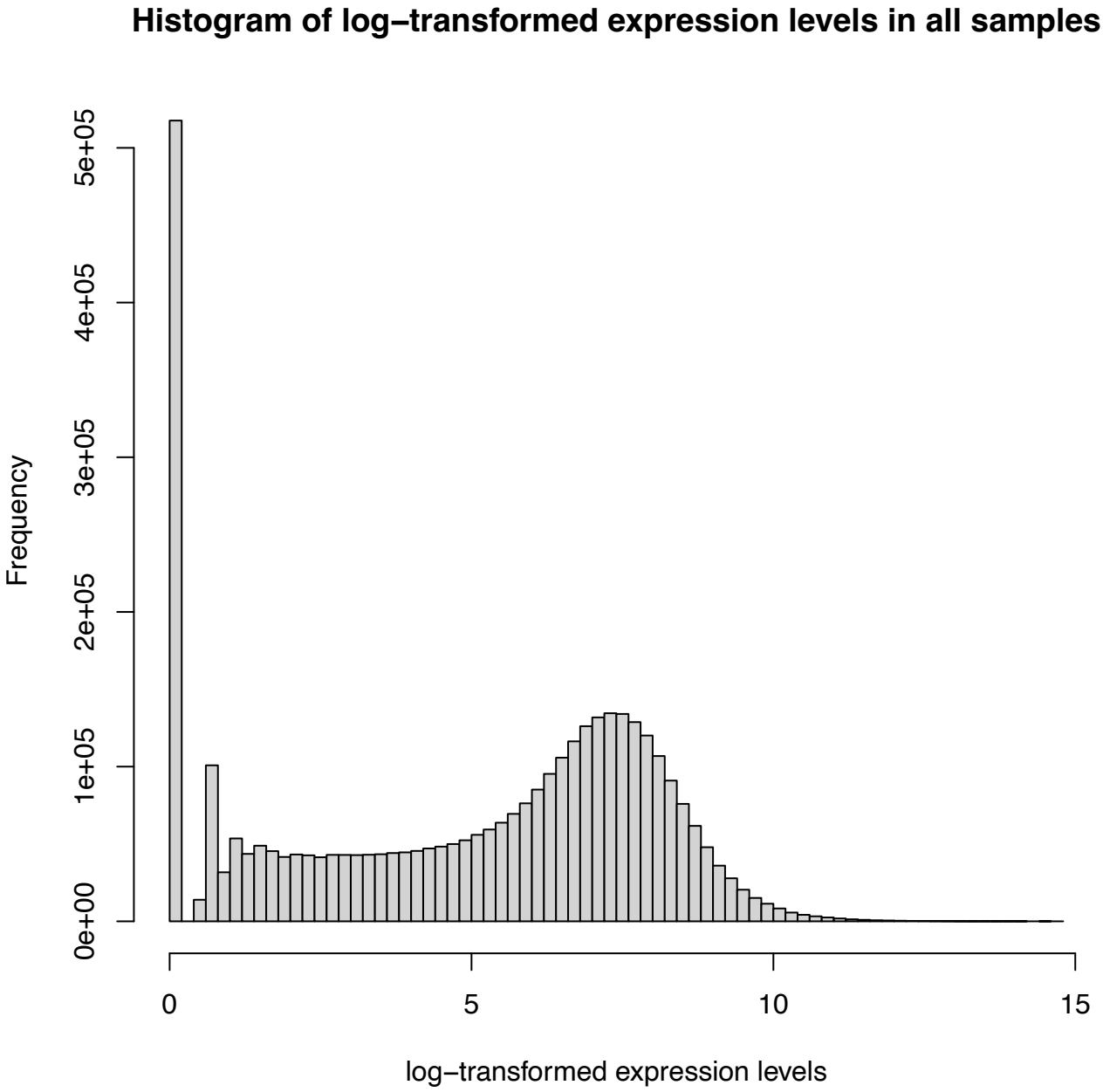
Based on the bar plot shown above, the read count varies a lot across all samples. The range of the read count of each sample is between. Although most of the samples have read counts around 5–6 × 10^7, there are still some extreme read counts in samples, ranging from 0.5 to 7.5 × 10^7. I assume this un-normalized and un-uniformed data will largely affect the downstream analysis since there are lots of variations in read counts among samples.

b) Perform a "total count" normalization for read-depth in each sample, and remake the histogram plot to verify that it produces the expected result.



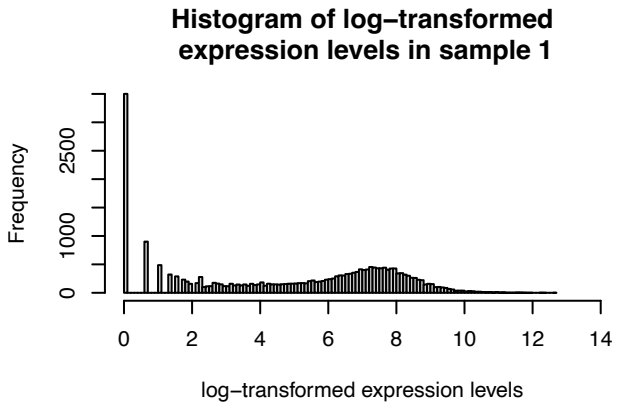
After normalization, the total read counts in each sample should all be the median read counts across all samples, which is calculated to be 43114673.5. Based on the bar plot above, all samples have exact same read counts. This aligns with what I expected.

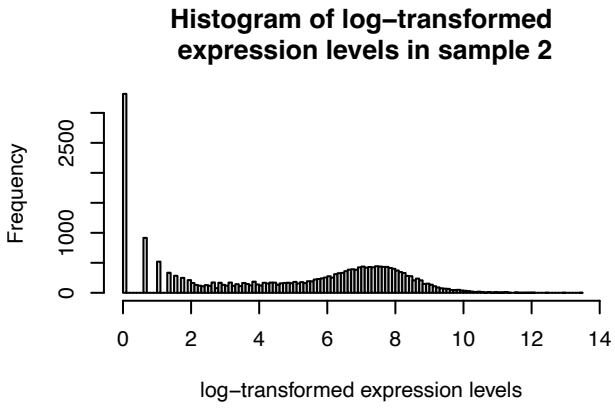
c) Log-transform the normalized count data of the complete dataset (all genes, all samples). Remake the histogram.

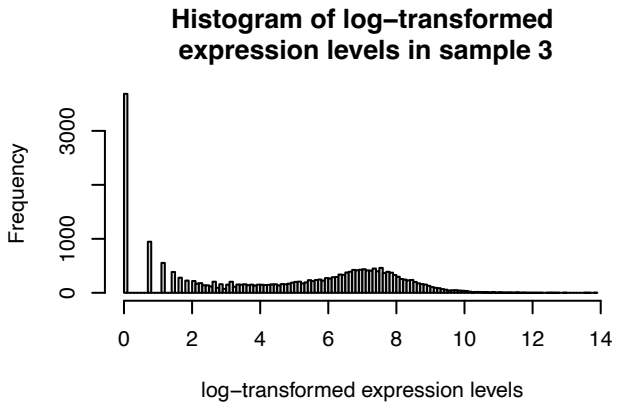


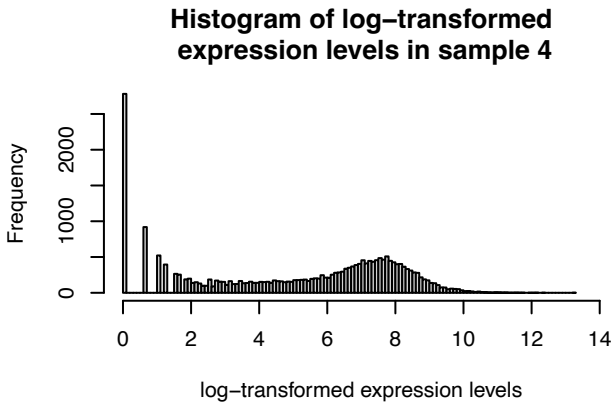
(d) Plot individual histograms of the log-transformed data for the first five samples. How do the distributions compare from sample to sample?

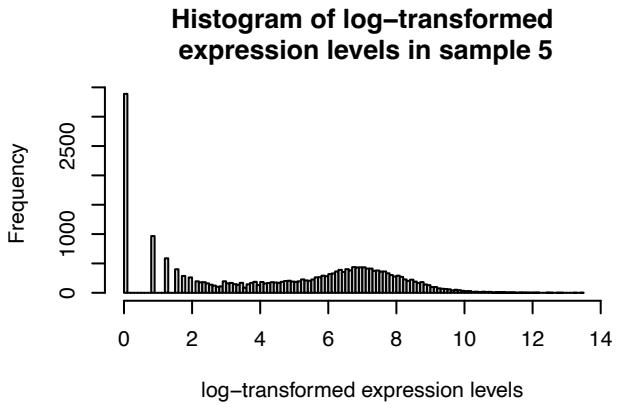
The distributions of five plots exhibit lots of similarities. There are many 0 values, but it is common in RNA-seq data. Many samples have log gene expression level ranging from 6 to 9, which is consistent across all 5 samples.





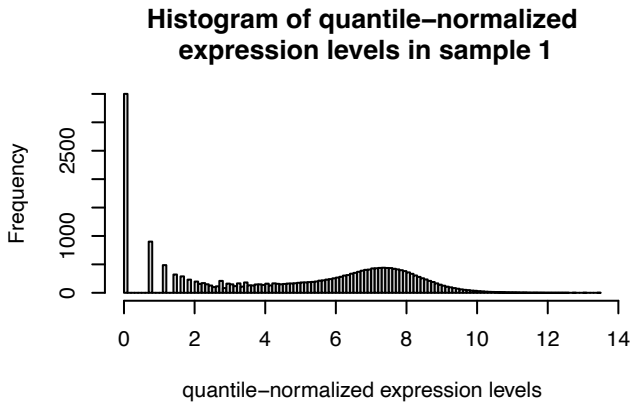


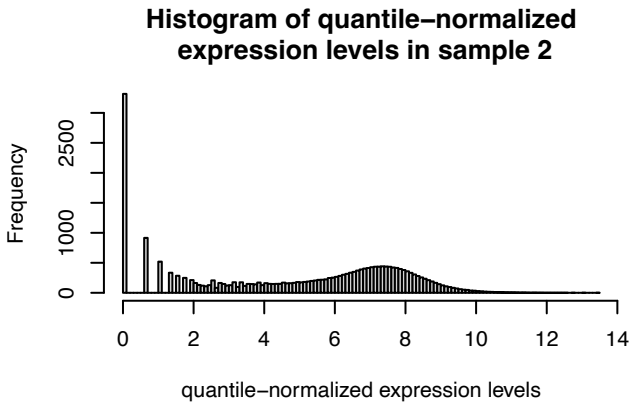


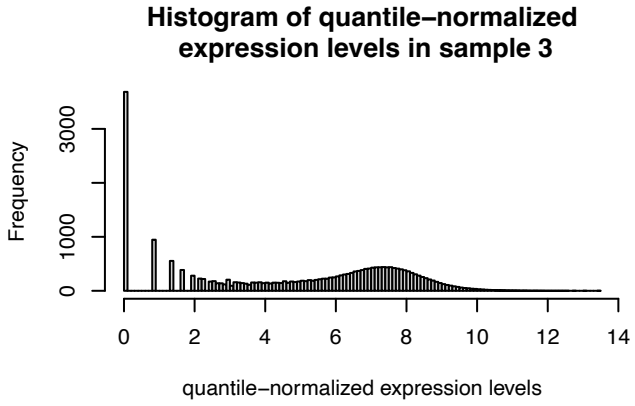


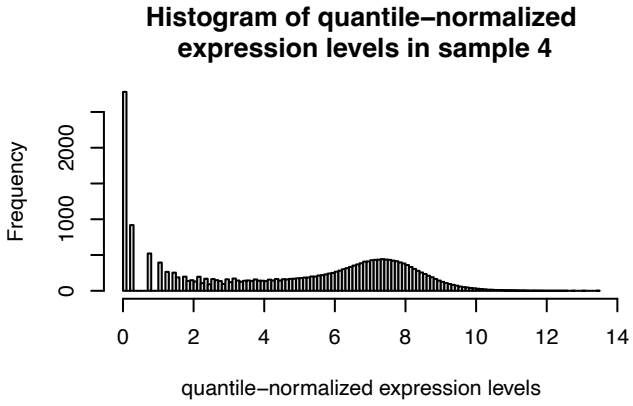
(e) Implement and perform quantile normalization on your log-transformed data (from part c) across all samples such that each has the same empirical distribution. Plot a histogram of the quantile-normalized data for each of the first five samples.

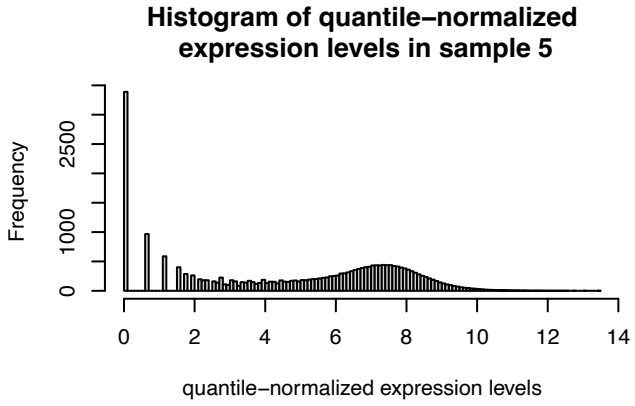
After quantile-normalization, the mean read counts for each sample will be exact the same. The five histograms below indicate the distributions of these five samples are very similar with each other, even identical.











4. Analysis of differential expression

(a) List the top 10 genes, 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 GBM.

The top 10 genes with lowest p-values are shown as followings:

	p_value
KCTD3	3.898632e-06
PRPF40A	1.940810e-05
CDC73	2.021023e-05
TMEM191A	7.526306e-05
PIK3CA	1.426170e-04
AACS	2.150961e-04
ATP6V1G2.DDX39B	2.151576e-04
WFDC2	2.213082e-04
NLRP10	2.213940e-04
SCYL3	3.123445e-04

KCTD3 belongs to the human family of Potassium (K+) Channel Tetramerization Domain (KCTD) proteins. So far only partial proteins of this family have been studied, but increasing evidences have shown that the proteins from this family have important biological functions, such as protein degradation, and therefore different members in this protein family can promote or prevent cancer development. KCTD3 was found to be associated with neurogenetic and neurodevelopmental disorders. KCTD3 has been found to increase stability and cell surface expression of hyperpolarization-activated cyclic nucleotide-gated channel 3 (HCN3), which has been found overexpressed in neuroblastoma tumors, and may serve functions such as protect tumor cells from apoptosis. It seems that the overexpressed KCTD3 can promote development of neuroblastoma tumors. In the dataset of GBM, the average gene expression level in long-term patients is lower than that in short-term patients (8.32 and 8.54, respectively, the code is displayed at the end of the R script). This result also confirms that KCTD3 can promote the progression of cancer, such as GBM.

PRPF40A (Pre-mRNA-processing factor 40 homolog A) is a human gene involving pre-mRNA splicing. It is found to be associated with the spliceosome, which is a complex molecular machinery that is responsible for removing intros from pre-mRNA. One study has demonstrated that in high-grade astrocytoma/glioblastomas tissues, a marked dysregulation in the expression levels of multiple components of the splicing machinery, including PRPF40A, compared to the control tissues. In the GBM dataset, I found that the average expression level of PRPF40A is higher in short-term patients (8.3) than that in long-term patients (8.15), which may also indicates that the dysregulation of PRPF40A is associated with GBM development.

b) Report the number of significant genes based on the Wilcoxon rank-sum test at an uncorrected $p < 0.05$ cutoff. Report the number of significant genes from the DESeq2 approach at an uncorrected $p < 0.05$.

The number of significant genes based on the Wilcoxon rank-sum test is1683. The number of significant genes based on provided DESeq2 report is 2591.

c) What is the overlap between the set of genes deemed significant at an uncorrected $p < 0.05$ cutoff by the two different approaches (DESeq2 and Wilcoxon rank-sum statistics)? Report the total number of genes that overlap and a list of those overlapping genes.

The number of overlap genes is: 1007. These genes are reported in the file named "overlap_gene.txt".

5. Multiple hypothesis correction

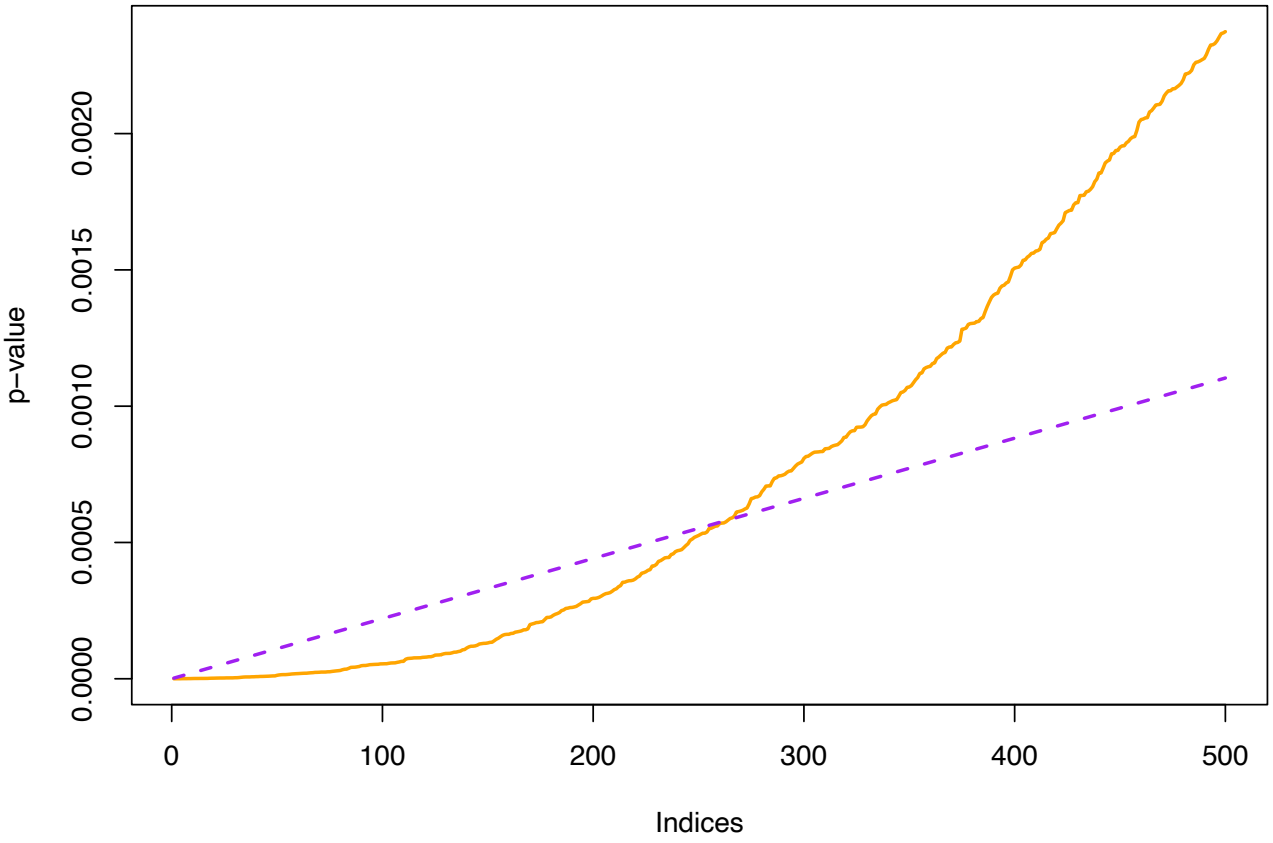
(a) Use Bonferroni correction with the DESeq2 results to identify differentially expressed genes at a global significance level (corrected $p < 0.05$) and report the number of significant genes.

The number of significant genes after Bonferroni correction with DESeq2 results is 20.

(b) Use the Benjamini-Hochberg step-up procedure to control the False Discovery Rate (FDR) with the DESeq2 statistics to identify differentially expressed genes at an FDR < 0.05. Report the number of significant genes after applying the BH procedure (specify which independence assumption you are using for the BH procedure).

The number of significant genes after Benjamini-Hochberg procedure with DESeq2 results is 264.

(c) Rank the genes 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 genes. More specifically, with the x-axis as the index of the ranked genes from 1 to the number of genes, plot the first 500 genes (i on the x-axis, the p-value of the ith gene on the y-axis). 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.


p−value
0.0000
0.0005
0.0010
0.0015
0.0020
0
100
200
300
400
500
Indices