CSCI 5461 HW1

Daniel Chang

March 2023

2: Exploring the dataset

a: How many genes are included in the dataset?

There are 23,368 genes.

b: How many patient samples are in the dataset?

There are 154 patients in the dataset. There are 78 long-term patients and 76 short term patients.

3: Data processing and normalization

a: Total reads per sample

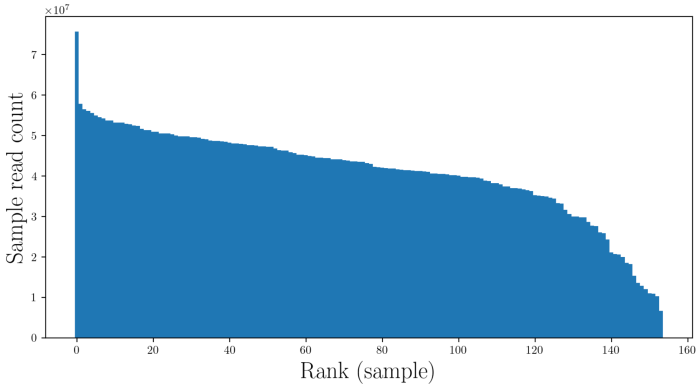


Figure 1: Rank ordered sample read counts

Sample read counts are roughly normally distributed, and range from 10 to 70 million.

b: Sample read count after total count normalization

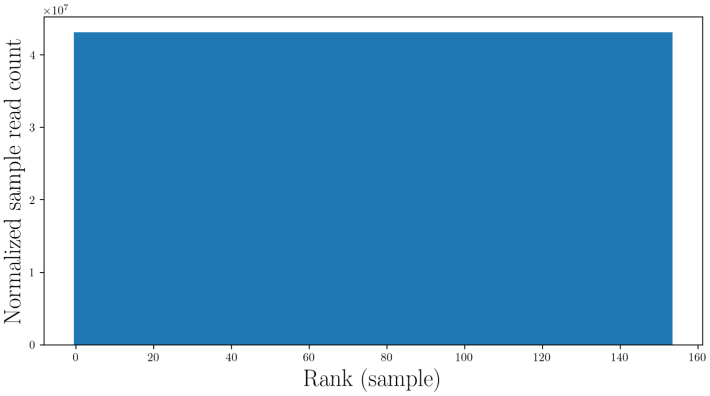


Figure 2: Rank ordered normalized sample read counts

Sample read counts are all normalized to the median read count: 43,114,673.5.

c: Log-transformed normalized count data

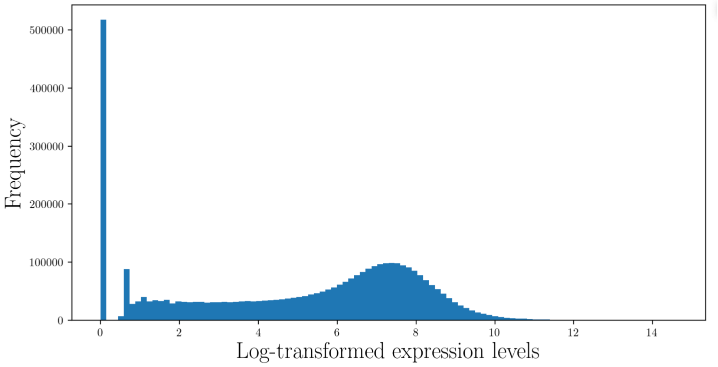
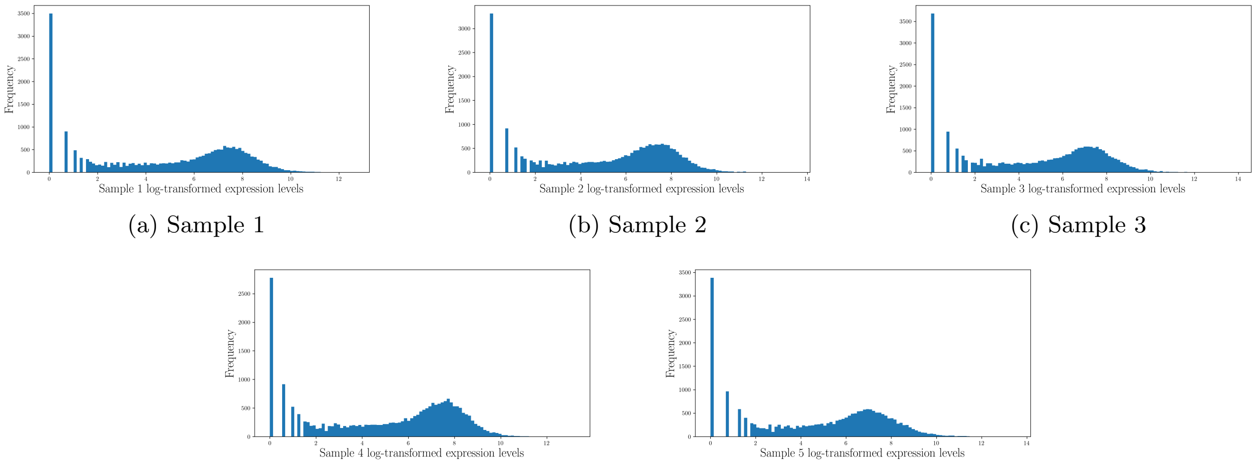


Figure 3: Log transformed data histogram

d: Log-transformed normalized count data per sample



(a) Sample 1

(b) Sample 2

(c) Sample 3

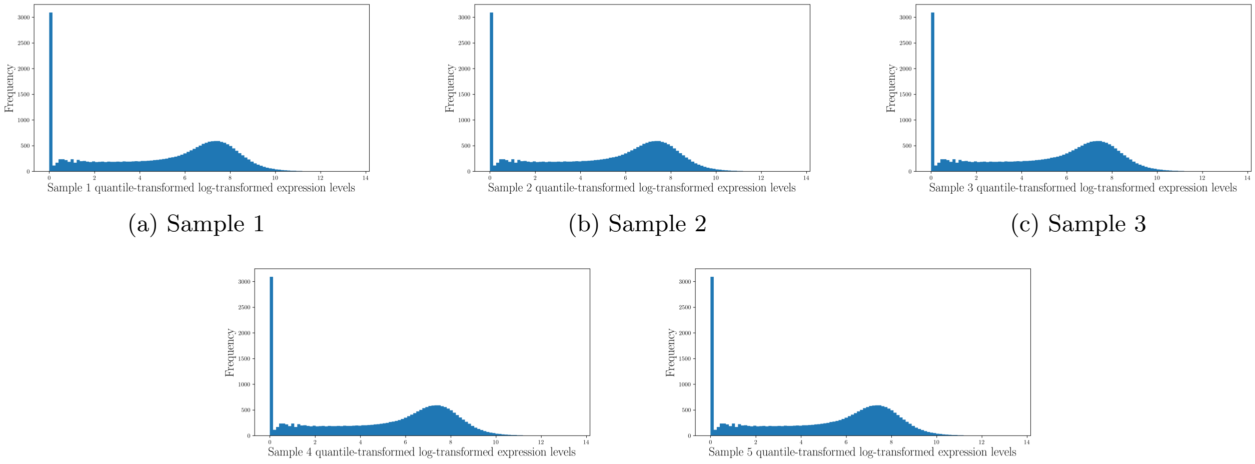
(d) Sample 4

(e) Sample 5

Figure 4: Log transformed data for first five samples

The distributions of the first five samples are roughly the same. However, they are “bumpy”, and the distributions are slightly shifted to the left or right in each sample relative to the overall distribution.

e: Quantile normalization



(a) Sample 1

(b) Sample 2

(c) Sample 3

(d) Sample 4

(e) Sample 5

Figure 5: Quantile normalized log transformed data for first five samples

After quantile normalization, the distribution of expression levels for each sample is more consistent and smoother.

4: Analysis of differential expression

a: Top 10 significant genes

Gene name	p value	Wilcoxon rank-sum statistic
KCTD3	0.000004	4.620313
PRPF40A	0.000019	4.278804
CDC73	0.000020	4.267962
ATP6V1G2-DDX39B	0.000052	4.047517
TMEM191A	0.000068	-3.982468
PIK3CA	0.000143	3.803582
AACS	0.000215	-3.700587
WFDC2	0.000234	-3.678904
SCYL3	0.000317	3.601206
ITPKA	0.000326	-3.593978

KCTD3 encodes for “a member of the potassium channel tetramerization-domain containing (KCTD) protein family”. KCTD3 may be related to GBM, as it “has been demonstrated to interact and increase stability and cell surface expression of hyperpolarization-activated cyclic nucleotide-gated channel 3 (HCN3)”, which has been “indicated to protect cells from apoptosis driven by HIF-1a and p53” [1].

AACS is a protein coding gene “predicted to enable acetoacetate-CoA ligase activity” and “involved in positive regulation of insulin secretion”. Downregulation of AACS may be associated with accelerated GBM tumor growth, as Acetoacetyl-CoA may facilitate ketone body synthesis, which allows non-transformed cells to utilize ketone bodies as an important energy source. This may mitigate tumor growth, as “malignant cells often highly depend on glycolysis for energy generation” [2].

b: Number of significant genes

There are 1,592 significant genes ($p < 0.05$)

c: Comparison with DESeq2

At an uncorrected $p < 0.05$ cutoff, there are 1,015 genes that are significant in both methods.

5: Multiple hypothesis correction

a: Bonferroni correction

After Bonferroni correction, there are 20 significant genes ($p < 0.05$).

b: Benjamini-Hochberg procedure

After BH, there are 264 significant genes ($FDR < 0.05$, assuming independent tests).

c: Rank ordered p-values

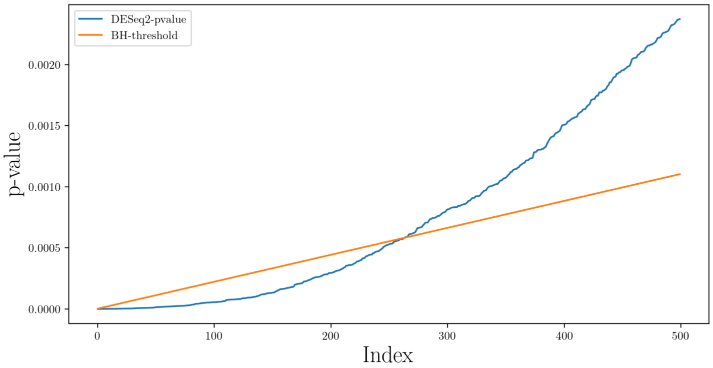


Figure 6: DESeq2 p-values and BH-thresholds

References

[1] Annapaola Angrisani, Annamaria Di Fiore, Enrico De Smaele, and Marta Moretti. The emerging role of the KCTD proteins in cancer. *Cell Communication and Signaling*, 19(1), May 2021.

[2] Gabriele D Maurer, Daniel P Brucker, Oliver Bähr, Patrick N Harter, Elke Hattingen, Stefan Walenta, Wolfgang Mueller-Klieser, Joachim P Steinbach, and Johannes Rieger. Differential utilization of ketone bodies by neurons and glioma cell lines: a rationale for ketogenic diet as experimental glioma therapy. *BMC Cancer*, 11(1), July 2011.