

aKiley Huffman
QBIO-478
Spring 2025

QBIO 478 HW #4

1. (10 pts) Consider the data table below for an RNA-seq experiment with only 5 genes. There are three wild-type (WT) replicates and three mutant replicates. Compute the RPKM table.

	WT 1	WT 2	WT 3	Mutant 1	Mutant 2	Mutant 3
Gene A (1kb)	30,000	60,000	120,000	30,000	60,000	120,000
Gene B (10kb)	300,000	600,000	1,200,000	300,000	600,000	1,200,000
Gene C (5kb)	30,000	60,000	120,000	300,000	600,000	1,200,000
Gene D (2kb)	300,000	600,000	1,200,000	30,000	60,000	120,000
Gene E (20kb)	340,000	680,000	1,360,000	340,000	680,000	1,360,000
Total	1,000,000	2,000,000	4,000,000	1,000,000	2,000,000	4,000,000

$$RPKM = \frac{10^9 \cdot C}{N \cdot L} \text{ where:}$$

C = Number of reads mapped to the gene

N = Total number of mapped reads in the sample

L = Length of the gene in base pairs

Sample Calculation

Gene A (1000bp) for WT 1

$$RPKM = \frac{10^9 \cdot 30000}{10000 \cdot 1000} = 30$$

Table 1. RPKM Table. Values were calculated using the TPM formula on a calculator.

Gene	Length (kb)	WT 1	WT 2	WT 3	Mutant 1	Mutant 2	Mutant 3
A	1	30000	30000	30000	30000	30000	30000
B	10	30000	30000	30000	30000	30000	30000
C	5	6000	6000	6000	60000	60000	60000
D	2	150000	150000	150000	15000	15000	15000
E	20	17000	17000	17000	17000	17000	17000

2. (10 pts) Consider the same data table as problem #1. Compute the TPM table.

Normalize read counts by gene length: $\text{Reads per kb} = \frac{\text{Raw Count}}{\text{Gene Length in kb}}$

Compute TPM for each gene: $\text{TPMs} = \frac{\text{Reads per kb}}{\text{Sum of all reads per kb}} \cdot 10^6$

Sample Calculations for WT1

Gene	Raw Count	Length (kb)	Reads/kb	TPM
A	30000	1	30000	$(30,000/233000) \times 10^6 \approx 128,755.36$
B	300000	10	30000	$(30,000/233000) \times 10^6 \approx 128,755.36$
C	30000	5	6000	$(6,000/233000) \times 10^6 \approx 25,751.50$
D	300000	2	150000	$(150,000/233000) \times 10^6 \approx 643,347.05$
E	340000	20	17000	$(17,000/233000) \times 10^6 \approx 72,390.09$
Total	1000000		233000	

Table 2. TPM Table. Values were calculated using the TPM formula on a calculator. WT2 and WT3 have proportionally the same read distribution as WT1, so TPMs remain the same; this is also true for the mutants.

Gene	Length (kb)	WT 1	WT 2	WT 3	Mutant 1	Mutant 2	Mutant 3
A	1	128,755	128,755	128,755	25,751	25,751	25,751
B	10	128,755	128,755	128,755	25,751	25,751	25,751
C	5	25,751	25,751	25,751	128,755	128,755	128,755
D	2	643,347	643,347	643,347	6,434	6,434	6,434
E	20	72,390	72,390	72,390	72,390	72,390	72,390

3. (10 pts) Five statistical tests have been done and here are the five unadjusted p-values: 10^{-2} , 10^{-5} , 0.02, 0.35, and 0.83. Fix the significance level $\alpha = 0.05$.

a. If we don't do multiple tests correction, how many p-values will be significant?

If we don't do multiple tests correct, only 3 of the p-values will be significant (10^{-2} , 10^{-5} , and 0.02) since they are all less than the significance level $\alpha = 0.05$.

b. If we do the Bonferroni correction, how many p-values will be significant?

The Bonferroni adjusts the p-value threshold to control family-wise error rate:

$\alpha = \frac{0.05}{5} = 0.01$. After this, only 2 of the p-values will be significant (10^{-2} and 10^{-5}),

which are less than $\alpha = 0.01$.

c. Do the False Discovery Rate (FDR) computation for each p-value. How many of these adjusted p-values are less than $\alpha = 0.05$?

Sort the p-values in ascending order and assign ranks:

Rank (i)	p-value	BH Critical Value	Compare to given p-values
1	10^{-10}	$(1/5) \cdot 0.05 = 0.01$	$10^{-10} < 0.01$ TRUE
2	10^{-5}	$(2/5) \cdot 0.05 = 0.02$	$10^{-5} < 0.02$ TRUE
3	0.02	$(3/5) \cdot 0.05 = 0.03$	$0.02 < 0.03$ TRUE
4	0.35	$(4/5) \cdot 0.05 = 0.04$	$0.35 < 0.04$ FALSE
5	0.83	$1 \cdot 0.05 = 0.05$	$0.83 < 0.05$ FALSE

Only 3 of these adjusted p-values are less than $\alpha = 0.05$ using FDR (10^{-10} , 10^{-5} , and 0.02).

4. (10 pts) In an RNA-seq experiment using short-read Illumina sequencing technology, why do we fragment the mRNA? In an RNA-seq experiment using long-read sequencing technology, why would we not fragment the mRNA?

Why do we fragment in short-read sequencing

In an RNA-seq experiment using short-read Illumina sequencing technology, we fragment the mRNA because Illumina and other short-read platforms typically generate reads ~50–300 base pairs long. mRNAs are much longer (often several kilobases), so we must fragment them to capture different parts of the transcript. This allows us comprehensive coverage of long transcripts, accurate quantification by counting fragments aligned across the gene body, and reassembly of transcripts from overlapping short reads.

Why we do not fragment for long-read sequencing

In an RNA-seq experiment using long-read sequencing technology, we would not fragment the mRNA because these technologies can sequence full-length transcripts in one read (up to tens of kilobases). Fragmentation here would defeat the purpose because it would break apart full isoforms and make isoform reconstruction harder. We would also lose the long-range information necessary to distinguish alternative splicing and transcription start sites.

5. (10 pts) Many genes' expressions oscillate throughout the day, and repeat this oscillation every 24 hours. This phenomenon has been observed in many organisms, in humans it is estimated that as many as ~40% of genes have expressions that oscillate throughout the day. Possible causes are responses to sunlight or temperature, whether the individual is asleep or awake, the time since the last meal or drink, etc. To learn more about circadian rhythms here is a link: https://en.wikipedia.org/wiki/Circadian_rhythm

Say you want to do an RNA-seq experiment to determine which genes are differentially expressed between people with and without type 1 diabetes. Explain how circadian rhythms present a challenge for this experiment. Describe a possible solution to this challenge.

Explain how circadian rhythms present a challenge for this experiment

Circadian rhythms present a challenge for this experiment because gene expression levels fluctuate over a 24-hour cycle. Thus, if you collect samples from people at different times of day, you might mistakenly interpret normal circadian fluctuations as disease-specific changes. For example, a gene could appear "differentially expressed" between diabetics and controls, but it could be just because one group had samples drawn in the morning, and the other in the evening. This introduces confounding and reduces your experiment's validity.

Describe a possible solution to this challenge

A simple solution is to collect all samples at the same time of day across all participants (diabetics and controls). You could also record collection times and adjust for them as covariates in your analysis.

6. (20 pts) The Genotype-Tissue Expression (GTEx) Portal is a public resource from the Broad Institute. Here is a link: <https://www.gtexportal.org/home/>

In the upper right-hand corner search for the gene ACE2.

a. Click on the top option, “Bulk tissue gene expression for ACE2.” Slide your mouse over the figure for the different tissues. What is TPM for the Lung? What is TPM for the Small Intestine? What is TPM for the Testis?

- Lung TPM: 0.6802
- Small Intestine TPM: 15.56
- Testis TPM: 20.74

b. Next click on the lower option, “Exon expression for ACE2.” How many different isoforms are listed? How many exons in the isoform with the most exons (listed first)? How many exons in the isoform with the least exons (listed last)?

- Different isoforms listed: 11
- Exons in the isoform with the most exons: 19
- Exons in the isoform with the least exons: 3

****Note:** The genes were not listed in descending order from most exons to least exons. I had to check each isoform to identify which isoforms had the most exons and which had the least.

Now in the upper right-hand corner search for the gene PTPN11.

c. Click on the top option, “Bulk tissue gene expression for PTPN11.” Slide your mouse over the figure for the different tissues. What is TPM for the Brain – Spinal Cord? What is TPM for Whole Blood?

- Brain TPM - Spinal Cord: 93.83
- Whole Blood TPM: 3.613

d. Next click on the lower option, “Exon expression for PTPN11.” How many different isoforms are listed? How many exons in the isoform with the most exons (listed first)? How many exons in the isoform with the least exons (listed last)?

- Different isoforms listed: 16
- Exons in the isoform with the most exons: 17
- Exons in the isoform with the least exons: 2

****Not:** The genes were not listed in descending order from most exons to least exons. I had to check each isoform to identify which isoforms had the most exons and which had the least.

e. Explore the website. You can look at ACE2 or PTPN11 in more detail, or look at something else. Report something interesting you find.

While exploring the GTEx Portal, I discovered that ACE2 gene expression levels vary significantly across different human tissues, with the testis exhibiting the highest expression. This elevated expression in the testis is significantly higher than in other tissues, including the lungs, which are commonly associated with ACE2 due to its role as the entry receptor for SARS-CoV-2. This finding suggests that tissues beyond the respiratory system, such as the testis, may be more susceptible to SARS-CoV-2 infection than previously understood.

I also discovered an intriguing aspect of the ERICH3 gene: it exhibits distinct expression patterns across different tissues. Specifically, ERICH3 is highly expressed in various regions of the human brain, including the nucleus accumbens and frontal cortex.