HW #4 due Thursday, April 8
(Note: the HW has four pages)

Below are the un-normalized counts for a fictional RNA-seq experiment with only four genes.

	Condition 1	Condition 1	Condition 2	Condition 2
Gene A (1 kb)	1,000	2,000	4,000	3,000
Gene B (20 kb)	5,000	10,000	20,000	15,000
Gene C (5 kb)	1,500	3,000	12,000	9,000
Gene D (5 kb)	1,000	2,000	4,000	3,000
Total	8,500	17,000	40,000	30,000

1. (10 pts) Compute RPKM.

	Condition 1	Condition 1	Condition 2	Condition 2
Gene A (1 kb)				
Gene B (20 kb)				
Gene C (5 kb)				
Gene D (5 kb)				

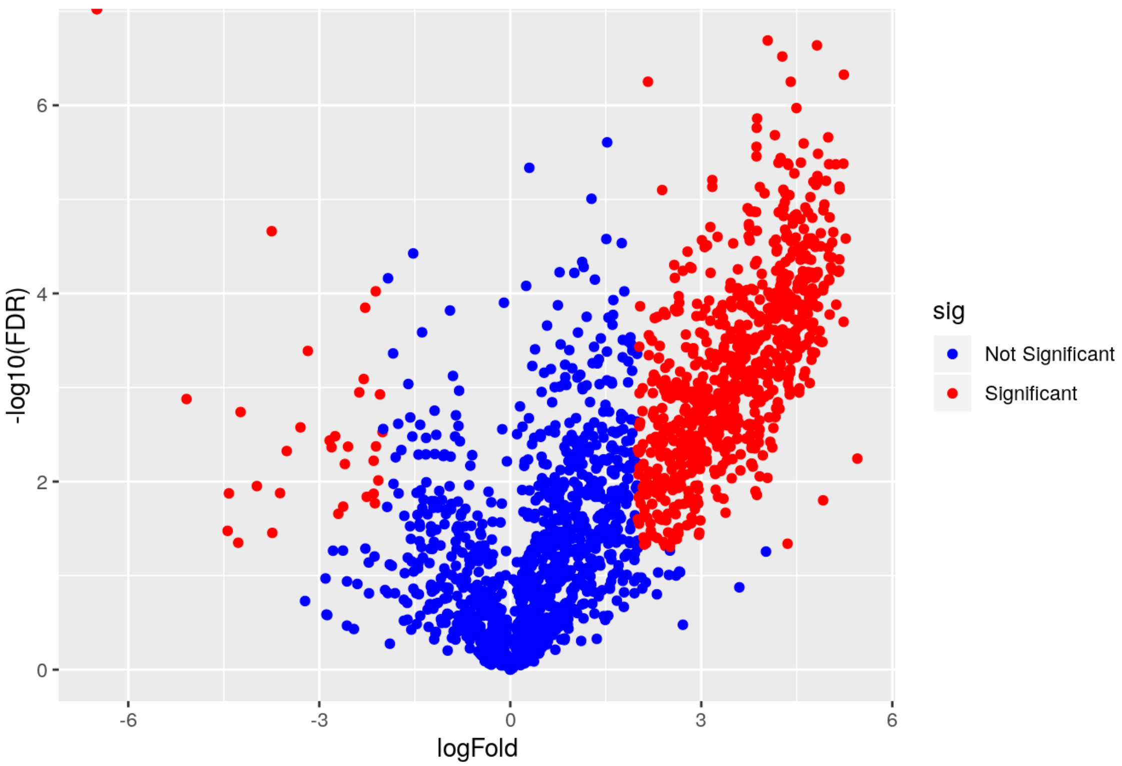
2. (10 pts) Compute TPM.

	Condition 1	Condition 1	Condition 2	Condition 2
Gene A (1 kb)				
Gene B (20 kb)				
Gene C (5 kb)				
Gene D (5 kb)				

3. (2 pts) Which (if any) of the four genes are expressed more under condition 2 than condition 1? (To answer this question, you do not need to do a statistical test.)

4. (2 pts) Which of the four genes has the highest expression for condition 1? (To answer this question, you do not need to do a statistical test.)

5. (2 pts) Below is a Volcano Plot for a real RNA-seq experiment with many more genes than the example above. “Fold” is the gene expression under condition 1 divided by the gene expression under condition 2. Let N1 be the number of genes that are expressed significantly more under condition 1 than condition 2, and let N2 be the number of genes that are expressed significantly more under condition 2 than condition 1. Based on this plot, are N1 and N2 about equal, is N1 > N2, is N1 < N2, or is there not enough information to know?



Below are ten un-adjusted p-values.

0.2, 0.6, 0.35, 0.33, 0.0001, 0.008, 0.88, 0.9, 0. 7, 0.62

6. (10 pts) Compute the ten adjusted p-values by the False Discovery Rate (Benjamini-Hochberg) procedure.

7. (2 pts) If instead of the FDR we had used the Bonferroni correction, how many of these ten p-values would be less than significance level $\alpha = 0.05$ after doing the Bonferroni correction?

An RNA-seq experiment is done to test for genes that are differentially expressed between conditions 1 and 2. The False Discovery Rate is set at 0.05.

8. (2 pts) After doing the FDR adjustment, 800 genes are significantly differentially expressed. About how many of these do we expect to be false positives? (You can give a numeric answer or say there is not enough information to know).

9. (2 pts) After doing the FDR adjustment, 11,200 genes are not significantly differentially expressed. About how many of these do we expect to be false negatives? (You can give a numeric answer or say there is not enough information to know).

10. (3 pts) Imagine you have done an RNA-seq experiment similar to the example at the beginning of this HW, except you have data for all ~20,000 genes. Similar to this example, you find there is one gene that clearly has much higher expression than all the other genes under condition 1.

    A scientist disputes your finding. This scientist argues that there is a bias in RNA-seq experiments that is not taken into account by RPKM or TPM. This scientist claims that in addition to gene expression, the read count in an RNA-seq experiment also depends on the GC-content of the reads. In particular, reads with GC-content near 50% are counted more relative to reads with much lower or much higher GC-content.

    Describe a method to test this scientist's claim. Assume you have data for many RNA-seq experiments, the DNA sequence of the reference genome, and the ability to write computer code. (Note: there is more than one correct response to this question.)