Problem Set 3

Comp 169 Problem Set 3 Due Friday, October 27th, 2023

This homework assignment deals with single cell RNA sequencing (scRNAseq). Single cell RNA sequencing is frequently used by biologists interested in cell types, developmental processes, and tissue heterogeneity. mRNA is collected from individual cells, sequenced, and mapped to a reference genome to generate a cells-by-genes matrix with "counts" representing the number of transcripts in each cell mapped to each gene.

Single cell gene expression is generally noisy, sparse, and high dimensional. In many experiments the counts matrices can exceed 100,000 cells x 20,000 genes.

Here, we have provided a small (toy) data set of real single cell gene expression measurements obtained from mouse olfactory epithelium, the tissue in the sinonasal cavity that is responsible for the sense of smell. The olfactory epithelium contains diverse neuronal and non-neuronal cell types, including gland cells and sustentacular (aka supporting epithelial) cells.

We have provided the count data for expression in each cell - the cells are each designated by a unique barcode represented by a short DNA sequence represented as a character string. So the cell names are strings like "AGTGTCAGTAATCACC".

In addition to the expression counts, we have provided you with each cell's type (neural, gland, or sus(tentacular)), lists of a few "marker genes" whose high expression levels are associated with each cell type, and a set of two-dimensional coordinates for plotting the cells. All of these materials have been collated into an R object of type `list`, and stored in an `.rds` file.

Recall that RDS files are representations of single R objects that can be saved for easy importing and moving between systems. The functions `saveRDS()` and `readRDS` may be helpful with these. Typically, one combines several variables in a list and saves it in an RDS file. The functions `saveRDS()` and `readRDS` may be helpful in working with RDS files. Recall also that lists are essentially like vectors, except that any element in the list can be of any object type, even another list. Elements in a list can be designated using the `$` operator (for named lists only), or using `[[n]]` where `n` is the index of the element you wish to select.

1. Bootstrapping and filtering zero-inflated data (30 points) 1a (3 points) First, read in the object in the file `PS3_Q12_materials.rds` and verify the class of the thing you loaded.

Although the individual objects from the list are not restored as separate named items, the names associated with the list elements remain. What are the names of the variables that were stored in the RDS file?

Look at the data in each variable (you can use `head` to see the first few rows, and/or designate the first few columns in a matrix (e.g. matr[1:5,1:5]) to get a sense of what is in each). Show your code.

```
file_path <- "/Users/jiataizhang/Desktop/PS3_Q12_materials.rds"
data <- readRDS(file_path)
class_of_data <- class(data)
cat("Class of the object:", class_of_data, "\n")
```

```
## Class of the object: list
```

```r
variable_names <- names(data)
cat("Names of variables in the loaded object:", variable_names, "\n")
```

```
## Names of variables in the loaded object: Counts_Matrix Plot_Coords
```

```r
head(data$Counts_Matrix)
```

```
##        AAACCTGGTCTCACCT AAACGGGAGCCCTAAT AAACGGGAGTGTGAAT AAACGGGTCCTTCAAT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                0                3
## Lypla1                0                0                2                1
## Tcea1                 0                0                0                5
## Rgs20                 0                0                0                0
##        AAACGGGTCTGCTGTC AAAGCAACAGGCTGAA AAATGCCGTATCAGTC AACACGTAGCGTTTAC
## Xkr4                  0                1                0                0
## Gm1992                0                0                0                0
## Mrpl15                1                0                0                0
## Lypla1                0                0                0                0
## Tcea1                 1                0                0                1
## Rgs20                 0                0                0                0
##        AACCATGCAGTCGTGC AACCATGGTAGAAAGG AACCGCGGTCAGAATA AACTCAGGTTCTGTTT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                0                0
## Lypla1                0                0                0                0
## Tcea1                 0                2                2                0
## Rgs20                 0                0                0                0
##        AACTGGTAGTAGGCCA AACTTTCTCTGCGTAA AAGCCGCAGTCCATAC AAGGCAGGTTCAACCA
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                2                0                0
## Lypla1                0                2                2                0
## Tcea1                 1                0                0                1
## Rgs20                 0                0                0                0
##        AAGGCAGTCGACAGCC AATCGGTTCCTACAGA ACACCAACATCATCCC ACACCAAGTCCTCCAT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                2                0                0                1
## Lypla1                0                0                0                0
## Tcea1                 0                0                2                3
## Rgs20                 0                0                0                0
##        ACACCAATCCGTAGTA ACACTGACAGCTGGCT ACAGCTATCTTGAGGT ACATGGTGTAACGCGA
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                1                0                0                0
## Lypla1                0                0                0                0
## Tcea1                 0                1                1                0
## Rgs20                 0                0                0                0
##        ACATGGTTCCCGACTT ACCTTTAAGACGCACA ACGAGCCGTGTTCTTT ACGAGGAGTGAGGGTT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
```

```
## Mrpl15                0                 0                 0                 1
## Lypla1                0                 0                 2                 0
## Tcea1                 0                 5                 0                 0
## Rgs20                 0                 0                 0                 0
##        ACGATACAGTTCGCAT ACGATACCAGTCAGCC ACGATACTCTGTTGAG ACGCCGAGTGCATCTA
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 1                 0                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 1                 0                 0                 2
## Rgs20                 0                 0                 0                 0
##        ACGGAGAGTTCGTGAT ACTGTCCGTTACTGAC ACTTACTGTATCAGTC AGAGCGATCCAAAGTC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 1                 1                 1                 0
## Rgs20                 0                 0                 1                 0
##        AGAGCTTTCTAGCACA AGAGTGGGTTAGGGTG AGCAGCCGTTGGGACA AGCGGTCAGAGGGATA
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                1                 0                 0                 0
## Lypla1                1                 0                 0                 0
## Tcea1                 3                 1                 1                 0
## Rgs20                 1                 0                 0                 0
##        AGCGGTCCAGATCCAT AGCGTATTCCCAAGAT AGCGTCGCATCGATGT AGCTCCTAGTCATCCA
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                2                 0                 1                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 0                 1                 1                 3
## Rgs20                 0                 0                 0                 0
##        AGCTCCTGTCTCTCTG AGCTCTCCAGCCAGAA AGCTCTCTCGTATCAG AGGCCGTCAGCTTAAC
## Xkr4                  0                 1                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 2                 3                 1
## Rgs20                 0                 0                 0                 0
##        AGGTCATTCGATGAGG AGGTCCGAGCTAGTCT AGTGAGGCACCCATGG AGTGGGACAGCAGTTT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                1                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 2                 0                 0                 1
## Rgs20                 0                 0                 0                 0
##        AGTGTCAGTAATCACC AGTGTCAGTTCCACTC AGTGTCATCCACGCAG ATCCACCTCATTCACT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 1                 0
## Tcea1                 0                 0                 0                 0
## Rgs20                 1                 0                 0                 0
##        ATCCGAACATGGGAAC ATCGAGTAGACTACAA ATCGAGTGTGCACTTA ATGAGGGAGCCAGAAC
```

```
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 1                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                0                 1                 4                 0
## Rgs20                0                 0                 0                 0
##        ATGGGAGAGGCTACGA ATTACTCGTAGCTCCG ATTGGACGTGTCGCTG ATTGGTGCACAACTGT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 1                 3
## Rgs20                0                 0                 0                 0
##        ATTGGTGGTCCCTACT ATTGGTGTCAAGGTAA ATTGGTGTCTTGTACT ATTTCTGAGTTGTAGA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               2                 0                 1                 0
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 0                 0
## Rgs20                0                 0                 0                 0
##        CAACCAATCGACGGAA CAACCAATCTGGCGTG CAACCTCGTGAAATCA CAACTAGAGCGATAGC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               1                 0                 0                 0
## Lypla1               0                 0                 1                 0
## Tcea1                4                 0                 1                 0
## Rgs20                1                 0                 0                 0
##        CAAGGCCGTAATCACC CACACCTAGTCCGTAT CACACCTAGTGAATTG CACACCTCAAGCCTAT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 1                 1
## Lypla1               0                 0                 0                 0
## Tcea1                0                 1                 1                 4
## Rgs20                0                 0                 0                 0
##        CACACCTGTGGAAAGA CACAGGCCAGCGATCC CACAGTAAGTCATGCT CACATAGTCGGAAATA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 1                 1
## Lypla1               0                 0                 1                 0
## Tcea1                0                 0                 3                 0
## Rgs20                0                 0                 0                 0
##        CACCAGGAGCAATCTC CACCTTGCATGGTTGT CAGAATCCAGGACGTA CAGAATCTCGGGAGTA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 0                 1
## Rgs20                0                 0                 0                 0
##        CAGAGAGGTTTACTCT CAGATCACATCCGCGA CAGATCATCTCGTTTA CAGCATAAGCCGATTT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 2                 0                 0
## Lypla1               1                 0                 0                 0
## Tcea1                3                 0                 0                 0
```

```
## Rgs20                0                0                0                0
##        CAGCATATCAGTACGT CAGCATATCCCTAACC CAGCATATCCTTAATC CAGCTGGCAGCGTAAG
## Xkr4                 0                0                0                0
## Gm1992               0                0                2                0
## Mrpl15               0                0                1                1
## Lypla1               0                0                1                0
## Tcea1                0                0                1                1
## Rgs20                0                0                0                0
##        CAGGTGCCAGGTCTCG CAGTAACTCGGCGCAT CAGTCCTGTCACTGGC CAGTCCTTCATGGTCA
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               2                0                1                0
## Lypla1               0                0                0                0
## Tcea1                0                2                0                0
## Rgs20                0                0                0                0
##        CAGTCCTTCGTCACGG CATATGGGTCGCGAAA CATATTCCACACAGAG CATCAAGGTGAAATCA
## Xkr4                 0                0                0                1
## Gm1992               0                0                0                0
## Mrpl15               1                0                1                0
## Lypla1               0                0                0                0
## Tcea1                0                2                0                0
## Rgs20                0                0                0                0
##        CATCCACCAAATACAG CATCCACTCGAGCCCA CATCGAAGTTAAGACA CATGACACAAGGTTCT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                1                0                2                0
## Rgs20                1                0                0                0
##        CATTATCAGCCAGTTT CATTATCTCTGCCAGG CATTATCTCTTGTCAT CCACCTAAGAACAACT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                1                0                0
## Lypla1               0                1                0                0
## Tcea1                1                1                1                1
## Rgs20                0                0                0                0
##        CCACGGAGTACCGTAT CCACTACGTAAGTTCC CCATTCGAGTAAGTAC CCATTCGAGTAGGTGC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                1                0                1
## Rgs20                0                0                0                0
##        CCCAATCAGCACCGCT CCCAATCGTCTTGATG CCCATACTCACATAGC CCGTGGACAGTTTACG
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                0                0                3
## Rgs20                0                0                0                0
##        CCGTGGATCATTGCGA CCGTTCACAATGGATA CCGTTCATCAAACAAG CCTACACGTACCGGCT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
```

```
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 1                 0
## Rgs20                0                 0                 0                 0
##        CCTAGCTGTGTGTGCC CCTCAGTAGAAAGTGG CCTTCCCCAACACCTA CCTTCCCCAAGACACG
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                2                 0                 1                 1
## Rgs20                0                 0                 0                 0
##        CCTTCCCGTGCCTTGG CCTTTCTAGAGACGAA CGACTTCAGCGAGAAA CGAGAAGCAGCCACCA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               1                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 0                 1
## Rgs20                0                 0                 0                 0
##        CGAGAAGGTCGTCTTC CGAGCCAGTGCGGTAA CGAGCCAGTGGGTATG CGAGCCAGTTGAACTC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 1                 0
## Tcea1                1                 0                 0                 0
## Rgs20                0                 0                 0                 0
##        CGATGGCTCGCCGTGA CGATTGATCGAGGTAG CGCTATCTCAACACGT CGCTTCACATGGTCTA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 1
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 0                 0
## Rgs20                0                 0                 0                 0
##        CGGACTGTCCCGACTT CGGCTAGGTAGTGAAT CGGGTCATCAAGCCTA CGTAGCGGTATATGGA
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 1                 0                 0
## Lypla1               0                 0                 1                 0
## Tcea1                0                 0                 0                 2
## Rgs20                0                 0                 0                 0
##        CGTAGGCGTGTGGCTC CGTAGGCTCTGGCGTG CGTCACTAGGTGCTTT CGTCACTCAGGTCTCG
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 1                 0                 0
## Lypla1               1                 0                 1                 0
## Tcea1                0                 0                 1                 2
## Rgs20                0                 0                 0                 0
##        CGTCACTGTACCGGCT CGTCAGGAGACCTAGG CGTCAGGCACGTGAGA CGTCTACGTTGGACCC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 1
## Lypla1               0                 2                 0                 0
## Tcea1                0                 4                 0                 0
## Rgs20                0                 1                 0                 0
##        CGTGAGCAGTAGTGCG CTAGCCTAGATATGGT CTCACACCAAGCCGCT CTCAGAACAAACTGCT
## Xkr4                 0                 0                 0                 0
```

```
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 2                 0                 1                 1
## Rgs20                 0                 0                 0                 0
##        CTCCTAGAGCGTAATA CTCCTAGCACACATGT CTCGGGACATCGGTTA CTCGGGAGTGCACTTA
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 1                 1                 0
## Rgs20                 0                 0                 0                 0
##        CTCGTACAGGCATGGT CTCGTACGTACCGCTG CTCGTACGTCATCCCT CTCGTCAAGAGAACAG
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 0                 0                 0                 0
## Rgs20                 0                 0                 0                 0
##        CTCTAATCATGTCGAT CTCTAATTCAGTACGT CTGAAGTTCGAGGTAG CTGATAGCATGCCTTC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 1                 0
## Tcea1                 0                 1                 0                 0
## Rgs20                 0                 0                 0                 0
##        CTGGTCTGTTGACGTT CTGTGCTTCCCTCAGT CTTAACTAGGAATCGC CTTAACTTCGGATGTT
## Xkr4                  1                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 0
## Lypla1                0                 0                 0                 1
## Tcea1                 0                 1                 0                 0
## Rgs20                 0                 0                 0                 1
##        CTTACCGGTCCTGCTT CTTAGGAAGACAAAGG CTTAGGAGTTGGACCC CTTGGCTTCAGGCAAG
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 1                 1                 0
## Rgs20                 0                 0                 0                 0
##        CTTGGCTTCGACGGAA GAACCTAAGAGTGAGA GAACCTACAGGGTACA GAACCTATCAGGCAAG
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                2                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 0                 1                 0
## Rgs20                 0                 0                 0                 0
##        GAACGGAGTTACAGAA GAATGAACAGTCTTCC GAATGAAGTGTTAAGA GACACGCTCAACACGT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 1
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 2                 5                 0
## Rgs20                 0                 0                 0                 0
```

```
##        GACAGAGGTTGTGGCC GACCAATCATGGGACA GACCTGGAGGATTCGG GACCTGGCAGATGGCA
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                0                0
## Lypla1                0                0                0                0
## Tcea1                 0                1                4                1
## Rgs20                 0                0                1                0
##        GACGGCTCATGACATC GACGGCTGTTGAGGTG GACGTTAAGCTGAAAT GAGCAGAAGGCCCGTT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                2                0                0
## Lypla1                0                0                0                0
## Tcea1                 1                3                0                0
## Rgs20                 0                1                0                0
##        GAGCAGACAGATAATG GAGGTGAAGTTGCAGG GAGGTGACAACTGGCC GAGGTGAGTCCGAGTC
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                1                0
## Lypla1                1                0                0                0
## Tcea1                 1                1                0                1
## Rgs20                 0                0                0                0
##        GAGGTGATCTTGCAAG GATCAGTAGCTTCGCG GATCAGTCACTAGTAC GATCAGTGTAGCCTAT
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                0                2
## Lypla1                0                0                0                1
## Tcea1                 0                0                2                3
## Rgs20                 0                0                1                0
##        GATCGATGTGTATGGG GATCTAGTCTTGGGTA GATGAGGTCATCATTC GATGCTAGTAGGACAC
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                1                0
## Lypla1                1                0                1                0
## Tcea1                 2                1                0                2
## Rgs20                 0                0                0                0
##        GCAAACTCAAGTTGTC GCAATCAGTCGCTTCT GCACTCTGTCTGCAAT GCAGCCAAGTGCGATG
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                1                0
## Lypla1                0                0                0                0
## Tcea1                 0                0                1                2
## Rgs20                 0                0                0                0
##        GCAGCCAGTAGGACAC GCAGTTACAGCCAGAA GCATACAAGACCCACC GCATACACAAGTCATC
## Xkr4                  0                0                0                0
## Gm1992                0                0                0                0
## Mrpl15                1                0                0                0
## Lypla1                0                0                0                0
## Tcea1                 1                0                1                0
## Rgs20                 2                0                0                0
##        GCATGCGTCAGATAAG GCATGTACATTAGCCA GCGCAGTAGACCTTTG GCGCAGTAGCATGGCA
## Xkr4                  0                0                1                0
## Gm1992                0                0                0                0
## Mrpl15                0                0                1                0
## Lypla1                0                0                0                0
```

```
## Tcea1                1                0                0                0
## Rgs20                0                0                0                0
##        GCGCAGTAGTACGCGA GCGCAGTGTCAAGCGA GCGCGATGTAGCCTCG GCGGGTTCACCATCCT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                1                0
## Lypla1               0                1                0                1
## Tcea1                0                0                0                5
## Rgs20                0                0                0                0
##        GCTGCGAGTCGGCATC GCTGCGATCTAGCACA GCTTCCAAGATCCGAG GCTTCCATCGGTTAAC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               2                0                1                0
## Lypla1               0                0                0                0
## Tcea1                1                1                2                0
## Rgs20                0                0                0                0
##        GGAAAGCCAGATCGGA GGAAAGCGTGATGCCC GGAACTTCACAGGCCT GGAACTTCACGAAGCA
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                1                0
## Tcea1                0                1                0                0
## Rgs20                0                0                0                0
##        GGAACTTGTGACTACT GGAATAATCGGATGGA GGACATTTCTAACCGA GGATGTTTCAACGGCC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                0                0                0
## Rgs20                0                1                0                0
##        GGATGTTTCTCCTATA GGATTACAGATGTGTA GGATTACCAACGCACC GGCCGATTCAGTGTTG
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                0                0                0
## Rgs20                0                0                0                0
##        GGCCGATTCGCGTTTC GGCGTGTGTAAACCTC GGCTGGTCATGCGCAC GGCTGGTGTGTGCGTC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                0                1                0
## Lypla1               0                0                0                0
## Tcea1                0                1                1                0
## Rgs20                0                0                0                1
##        GGGACCTGTCCAACTA GGGATGAAGCAATCTC GGGATGAGTCGTTGTA GGGCACTCAATGGAAT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                1                0                0
## Rgs20                0                0                0                0
##        GGGTCTGTCTGCCCTA GGTGAAGTCTGGTATG GGTGCGTCATGACGGA GGTGTTAAGAGATGAG
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
```

```
## Mrpl15                1                 0                 0                 0
## Lypla1                1                 0                 0                 0
## Tcea1                 3                 0                 0                 0
## Rgs20                 0                 0                 0                 0
##        GTAACGTTCCAAGCCG GTACGTAGTGTGACGA GTACTCCTCAGTTGAC GTACTTTAGGCTATCT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 1                 0
## Tcea1                 0                 0                 1                 1
## Rgs20                 0                 0                 1                 0
##        GTAGTCACACCATCCT GTAGTCAGTAACGTTC GTATCTTGTTGATTCG GTCAAGTGTCTGGTCG
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 0                 0                 0
## Rgs20                 0                 0                 0                 0
##        GTCAAGTTCAGGTTCA GTCACAATCCGGGTGT GTCACGGGTTCTGGTA GTCCTCACACGAAAGC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                1                 0                 0                 0
## Tcea1                 0                 1                 0                 0
## Rgs20                 0                 0                 0                 0
##        GTCGGGTCAACACCTA GTCGGGTTCTTGAGAC GTCTTCGCAACTGCGC GTCTTCGGTGATAAAC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                1                 0                 0                 0
## Lypla1                1                 0                 0                 0
## Tcea1                 1                 0                 0                 1
## Rgs20                 0                 0                 0                 0
##        GTGAAGGTCACCTCGT GTGCAGCAGAGCTTCT GTGCAGCAGGAGTTGC GTGCATAGTCAAAGAT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 0                 6                 2
## Rgs20                 0                 0                 0                 0
##        GTGCTTCAGGCAAAGA GTGGGTCCAATCCAAC GTGTGCGGTCTTCTCG GTGTGCGGTGTGACCC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 0                 3                 0
## Rgs20                 0                 0                 0                 0
##        GTGTGCGTCCATTCTA GTGTTAGCACTCTGTC GTTACAGTCTGCTGCT GTTCTCGTCCGAAGAG
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 0                 5                 2
## Rgs20                 0                 0                 0                 0
##        GTTTCTACAGCTTAAC TAAGAGATCTCTAAGG TAAGAGATCTCTTATG TAAGCGTTCGGCTTGG
```

```
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 2                 0
## Rgs20                0                 0                 0                 0
##        TAAGTGCAGATCGATA TACAGTGAGTTAAGTG TACAGTGTCACTTCAT TACCTATAGATCCCAT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               2                 0                 0                 1
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 1                 4
## Rgs20                0                 0                 0                 0
##        TACCTTAAGGGAACGG TACGGATGTCCGAACC TACGGGCAGTAAGTAC TACGGGCCATCCTTGC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 0                 0
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 1                 1
## Rgs20                0                 0                 0                 0
##        TACGGGCCATGCGCAC TACGGGCGTAGATTAG TACGGGCTCCATGAAC TACTCATCAGTGAGTG
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 1                 0
## Lypla1               0                 0                 0                 0
## Tcea1                1                 0                 0                 0
## Rgs20                0                 0                 0                 0
##        TACTCGCTCAGTGTTG TACTTACCACATGTGT TACTTGTAGTGAATTG TACTTGTTCGAACTGT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               2                 0                 2                 0
## Lypla1               1                 0                 0                 1
## Tcea1                4                 0                 0                 3
## Rgs20                1                 0                 1                 0
##        TAGCCGGGTCTCCACT TAGCCGGTCGGAATCT TAGGCATGTTCTGAAC TATCAGGCAAACCCAT
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               1                 0                 0                 1
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 0                 0
## Rgs20                1                 0                 0                 0
##        TATTACCAGTGAACGC TATTACCCAATACGCT TCAACGATCCATGAGT TCAATCTGTCAATGTC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               1                 0                 0                 0
## Lypla1               1                 0                 0                 0
## Tcea1                2                 1                 1                 1
## Rgs20                0                 0                 0                 0
##        TCACAAGTCTCTAAGG TCACGAAAGCAGACTG TCACGAATCAGCTTAG TCAGATGGTACTCAAC
## Xkr4                 0                 0                 0                 0
## Gm1992               0                 0                 0                 0
## Mrpl15               0                 0                 1                 0
## Lypla1               0                 0                 0                 0
## Tcea1                0                 0                 0                 1
```

```
## Rgs20                0                0                0                0
##        TCAGGTAAGCGTCAAG TCATTACCAGTATGCT TCATTTGCAATGGATA TCATTTGTCCTCTAGC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                1                0                0
## Lypla1               0                1                0                0
## Tcea1                1                0                0                0
## Rgs20                0                0                0                0
##        TCCACACAGATCCCAT TCCACACAGTACGTTC TCCACACGTGATGTGG TCCCGATCAATCAGAA
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                1                0
## Tcea1                2                0                1                0
## Rgs20                0                0                0                0
##        TCGCGAGAGGTGCTAG TCGGTAAAGATCACGG TCTATTGCACATAACC TCTATTGCATTAACCG
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                0                0                0
## Rgs20                0                0                0                0
##        TCTATTGTCCTAAGTG TCTCATACAGGCGATA TCTGAGAGTTGTCTTT TCTGAGATCTTTACAC
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                1                0                0
## Lypla1               0                1                1                0
## Tcea1                1                0                0                0
## Rgs20                0                0                0                0
##        TCTTCGGAGTCTTGCA TCTTTCCCAGTAAGAT TGAAAGAGTGTATGGG TGACGGCGTGATGTGG
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                0                0                0
## Lypla1               0                0                0                0
## Tcea1                0                2                1                1
## Rgs20                0                0                0                0
##        TGACTAGTCCGTCAAA TGAGAGGGTTCGTCTC TGAGCATAGCTTCGCG TGAGCATGTCCTGCTT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
## Lypla1               0                1                0                0
## Tcea1                0                1                0                0
## Rgs20                0                0                0                0
##        TGCCAAATCAACCAAC TGCCCATGTTGCGCAC TGCCCTAAGGCTCTTA TGCGCAGAGCCCTAAT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               1                0                0                0
## Lypla1               1                0                0                0
## Tcea1                0                0                0                0
## Rgs20                0                0                0                0
##        TGCGGGTGTAGGGACT TGCGTGGTCTGATTCT TGCTACCTCCCTAATT TGCTGCTCAGACAAAT
## Xkr4                 0                0                0                0
## Gm1992               0                0                0                0
## Mrpl15               0                0                0                0
```

```
## Lypla1                0                 0                 0                 0
## Tcea1                 1                 0                 1                 0
## Rgs20                 0                 0                 0                 0
##        TGCTGCTTCGGACAAG TGGACGCAGGAGCGTT TGGGCGTAGTGTCTCA TGTATTCCACCTCGTT
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 2                 0
## Tcea1                 2                 0                 1                 0
## Rgs20                 0                 0                 0                 0
##        TGTCCCAAGTACCGGA TGTGGTATCAAGGTAA TTAACTCAGAGCCTAG TTAGGACCATTGGGCC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 1                 1                 1
## Rgs20                 0                 0                 0                 0
##        TTAGGCACAGTATGCT TTCGAAGGTCCATGAT TTCTACAAGAGTGAGA TTCTACAGTGTGTGCC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 1                 1
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 0                 0                 0
## Rgs20                 0                 0                 0                 0
##        TTCTCCTTCAACGAAA TTCTTAGCATCATCCC TTCTTAGGTTTGTGTG TTGCGTCTCACCAGGC
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 0                 0                 0
## Rgs20                 0                 0                 0                 0
##        TTGGAACGTACTTCTT TTGGCAACATGGTTGT TTGTAGGAGAGTTGGC TTGTAGGCAGTCACTA
## Xkr4                  0                 0                 0                 0
## Gm1992                0                 0                 0                 0
## Mrpl15                0                 0                 0                 0
## Lypla1                0                 0                 0                 0
## Tcea1                 0                 1                 0                 1
## Rgs20                 0                 0                 0                 0
##        TTTCCTCCACAACGTT TTTGTCAGTCGCGAAA
## Xkr4                  0                 0
## Gm1992                0                 0
## Mrpl15                1                 0
## Lypla1                0                 1
## Tcea1                 0                 0
## Rgs20                 0                 0
```

```
head(data$Plot_Coords)
```

```
##                          V1         V2
## AAACCTGGTCTCACCT -5.155221 -2.7345632
## AAACGGGAGCCCTAAT -4.794701 -2.7925018
## AAACGGGAGTGTGAAT -5.461591 -2.0370508
## AAACGGGTCCTTCAAT -1.601274 -0.2528715
## AAACGGGTCTGCTGTC  3.731006  5.0285312
```

```
## AAAGCAACAGGCTGAA  1.271016  2.6031148
```

1b (5 points) Write a function that takes in a numeric vector and uses bootstrapping to find the 95% confidence interval for the median value in that vector.

Then, write code that calls your function to report the 95% confidence interval for the gene 'Mt1' across all the cells in the data set. Write additional code that uses the same function to find the 95% confidence interval for the median expression level of the gene 'Rgs5.'

Does either of these confidence intervals include zero? Write your answer below your code.

```
library(boot)

calculate_median_ci <- function(data_vector, n_bootstrap = 10000, alpha = 0.05) {
  if (!is.numeric(data_vector)) {
    stop("Input data_vector must be a numeric vector.")
  }
  medians <- replicate(n_bootstrap, median(sample(data_vector, replace = TRUE)))
  medians <- medians[!is.na(medians)]
  ci <- quantile(medians, probs = c(alpha / 2, 1 - alpha / 2), na.rm = TRUE)
  return(ci)
}




gene_Mt1 <- as.numeric(data$Counts_Matrix['Mt1', ])
gene_Rgs5 <- as.numeric(data$Counts_Matrix['Rgs5', ])
ci_Mt1 <- calculate_median_ci(gene_Mt1)
ci_Rgs5 <- calculate_median_ci(gene_Rgs5)


ci_Mt1
```

```
##  2.5% 97.5%
##    15    20
```

```
ci_Rgs5
```

```
##  2.5% 97.5%
##     0     1
```

written answer goes here CF for Rgs5 includes zero.

1c (5 points) Most single-cell laboratory protocols attempt to physically remove any dead or dying cells from single cell suspensions prior to single-cell RNA sequencing. However, no method is infallible. Because of this it is necessary to computationally identify and exclude probable dead or dying cells from single cell analyses after sequencing. One characteristic of dead/dying cells is low RNA content; another is a high proportion of mitochondrial RNA.

UMI is an acronym for Unique Molecular Identifier. These correspond roughly to the individual transcripts in the cell. There may be many different transcripts or copies of a given gene in each cell. On the other hand, some genes will not be detected at all. You can compute the total number of UMIs or transcripts per cell by summing the 'counts' (which refer to numbers of UMIs mapped to each gene) in each cell.

Without writing a loop, create a vector containing the total number of UMIs detected per cell for all 358 cells. Ensure that the entries of the vector, which correspond to the cells, are named by their cellular barcodes (e.g., "AGGCCGTCAGCTTAAC"). Show your code and the mean, median, and range of the total UMI counts in this vector.

```
total_UMI_counts <- rowSums(data$Counts_Matrix)
names(total_UMI_counts) <- rownames(data$Counts_Matrix)
mean_UMI_count <- mean(total_UMI_counts)
median_UMI_count <- median(total_UMI_counts)
range_UMI_count <- range(total_UMI_counts)
cat("Mean UMI count:", mean_UMI_count, "\n")
```

```
## Mean UMI count: 189.1368
```

```
cat("Median UMI count:", median_UMI_count, "\n")
```

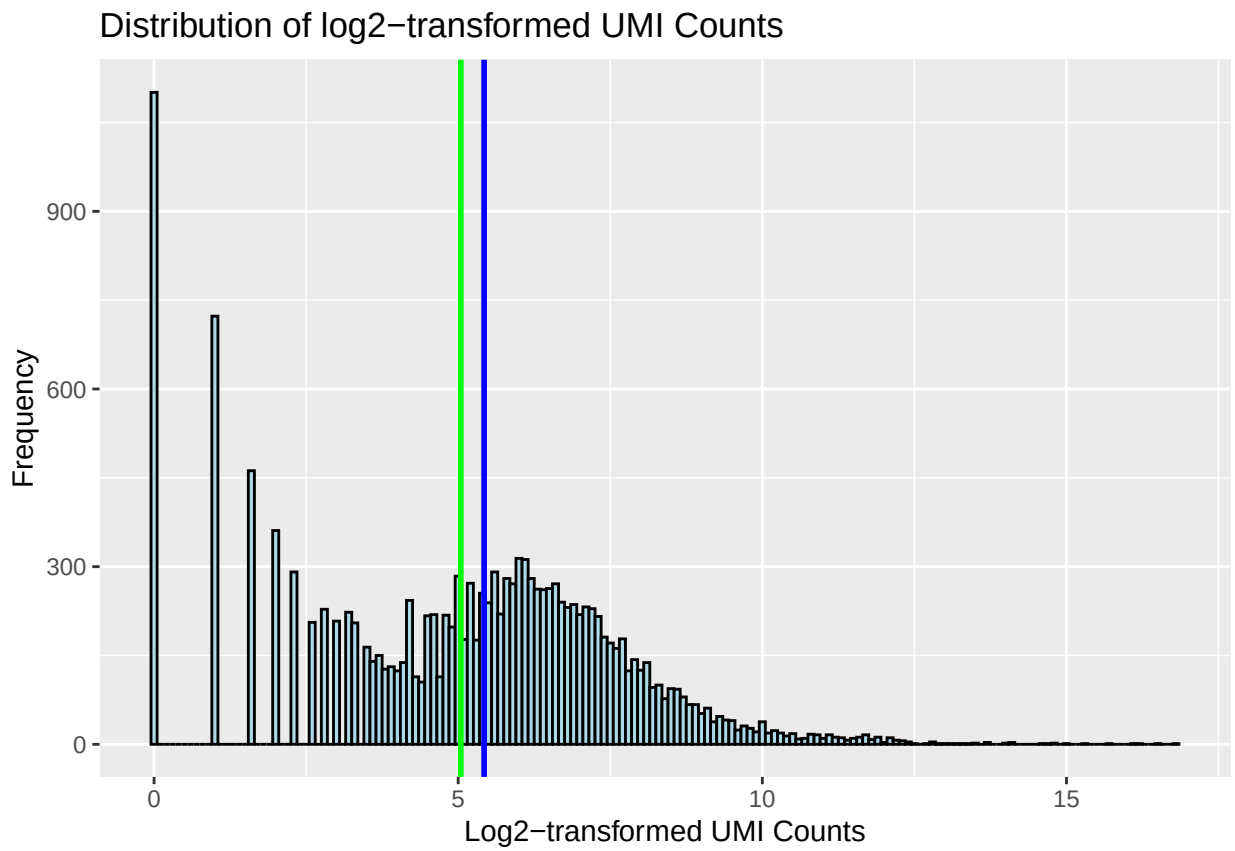
```
## Median UMI count: 43
```

```
cat("Range of UMI counts:", range_UMI_count[1], "to", range_UMI_count[2], "\n")
```

```
## Range of UMI counts: 1 to 113636
```

1d (4 points). Log2-transform the vector of total UMIs and generate a histogram of the distribution using ggplot, with the argument `bins = 50` to increase the number of bins shown. Emphasize the mean by overlaying a solid green vertical line there, and the median as a solid blue vertical line. What do the relative locations of the mean and median suggest about the skew of the data?

```
library(ggplot2)
log2_UMI_counts <- log2(total_UMI_counts)
histogram_plot <- ggplot(data.frame(UMI_counts = log2_UMI_counts), aes(x = UMI_counts)) +
  geom_histogram(binwidth = 0.1, fill = "lightblue", color = "black", bins = 50) +
  geom_vline(xintercept = mean(log2_UMI_counts), color = "green", linetype = "solid", linewidth = 1) +
  geom_vline(xintercept = median(log2_UMI_counts), color = "blue", linetype = "solid", linewidth = 1) +
  labs(title = "Distribution of log2-transformed UMI Counts",
       x = "Log2-transformed UMI Counts",
       y = "Frequency")
print(histogram_plot)
```

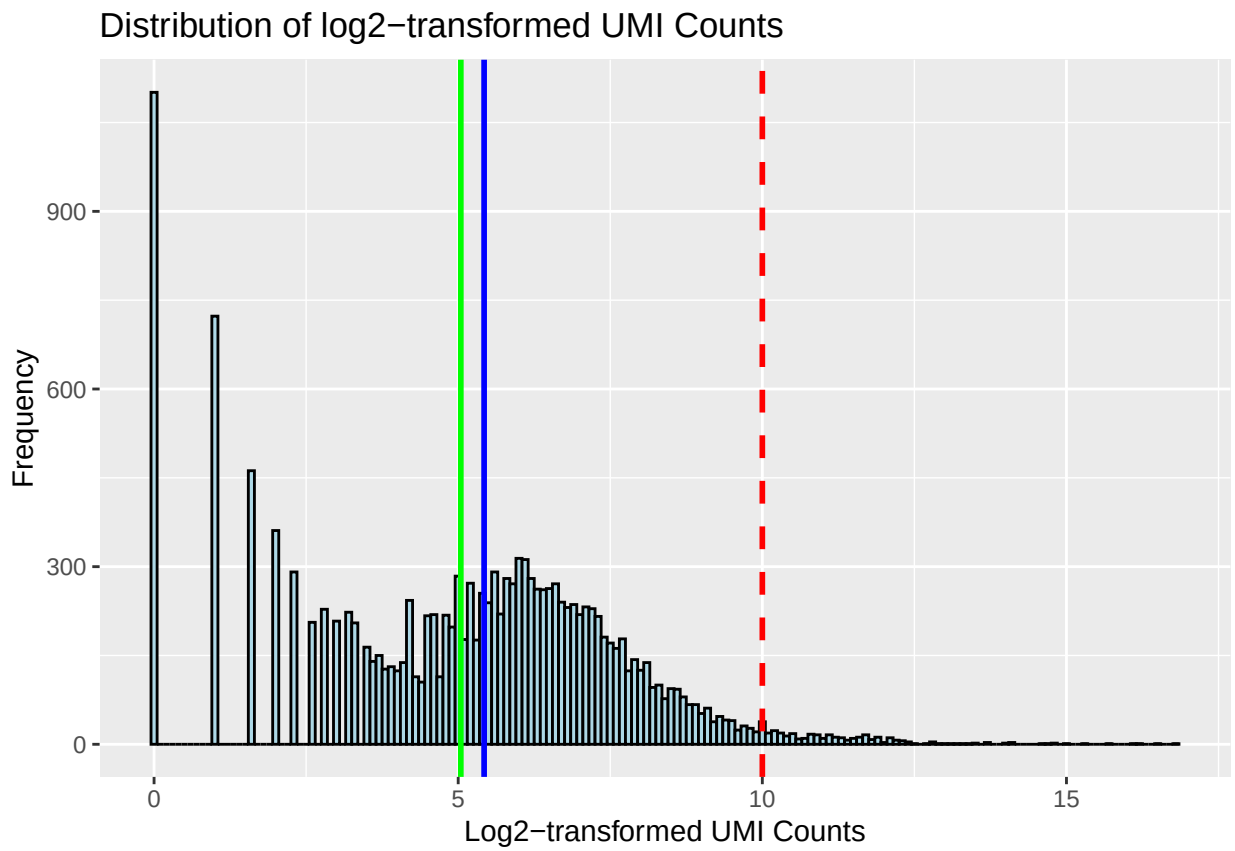


written answer goes here

The relative locations of the mean and median in the histogram suggest that the data may be right-skewed. If the mean is greater than the median, it indicates that there is a longer tail on the right side of the distribution, which is a characteristic of right-skewed data.

1e (4 points) Based on your histogram, visually identify a cutoff point below which you might consider the cells to be dead or dying. In practice, choosing this cutoff is sometimes arbitrary, but use your best judgement. Reproduce the histogram from problem 1d, but identify this threshold on your histogram with a dashed red line. What fraction of cells is below the cutoff you chose? Compute this value as well.

```
library(ggplot2)
log2_UMI_counts <- log2(total_UMI_counts)
cutoff_value <- 10
histogram_plot <- ggplot(data.frame(UMI_counts = log2_UMI_counts), aes(x = UMI_counts)) +
  geom_histogram(binwidth = 0.1, fill = "lightblue", color = "black", bins = 50) +
  geom_vline(xintercept = mean(log2_UMI_counts), color = "green", linetype = "solid", linewidth = 1) +
  geom_vline(xintercept = median(log2_UMI_counts), color = "blue", linetype = "solid", linewidth = 1) +
  geom_vline(xintercept = cutoff_value, color = "red", linetype = "dashed", linewidth = 1) +
  labs(title = "Distribution of log2-transformed UMI Counts",
       x = "Log2-transformed UMI Counts",
       y = "Frequency")
print(histogram_plot)
```




```
fraction_below_cutoff <- sum(log2_UMI_counts < cutoff_value) / length(log2_UMI_counts)
cat("Fraction of cells below the cutoff:", fraction_below_cutoff, "\n")
```

```
## Fraction of cells below the cutoff: 0.9777188
```

1f (5 points) Another QC metric of interest is the "mitochondrial ratio," the ratio of total counts mapping just to mitochondrial genes compared to the sum of the total counts for all genes. Assume for the purposes of this problem that mitochondrial genes are exactly those whose names start with the letters 'mt-'. Using the grep function in R, write a function that takes as input a counts matrix whose rows are designated by gene names and columns by cell barcodes, and returns a vector containing the mitochondrial ratio for all cells. Be sure to verify that totalUMIs > 0 for all cells, so that you do not try to divide by zero; if there are any cells with a total of zero UMIs, you may either print an error message or return a ratio of zero for those cells.

```
calculate_mitochondrial_ratio <- function(counts_matrix) {
  mt_gene_indices <- grep("^mt-", rownames(counts_matrix))
  total_counts_per_cell <- colSums(counts_matrix)
  mitochondrial_ratios <- numeric(length(total_counts_per_cell))
  for (i in 1:length(total_counts_per_cell)) {
    if (total_counts_per_cell[i] > 0) {
      mitochondrial_counts <- sum(counts_matrix[mt_gene_indices, i])
      mitochondrial_ratios[i] <- mitochondrial_counts / total_counts_per_cell[i]
    } else {
      mitochondrial_ratochondrial_ratios[i] <- 0
    }
```

```
  }
  return(mitochondrial_ratios)
}
```

1g (4 points) We typically say that a mitochondrial ratio above 0.2 suggests that the RNA is of poor quality (basically, you're seeing a lot of mitochondrial transcripts from dying cells).

Show how to use the function you wrote above to count the number of cells with either a mitochondrial ratio greater than 0.2, or a total UMI count below the cutoff you selected above. How many such cells are there? What fraction of the total does that account for?

```
mitochondrial_ratios <- calculate_mitochondrial_ratio(data$Counts_Matrix)

mitochondrial_ratio_cutoff <- 0.2
total_UMI_cutoff <- 10

poor_quality_cells <- which(mitochondrial_ratios > mitochondrial_ratio_cutoff | total_UMI_counts < total
```

```
## Warning in mitochondrial_ratios > mitochondrial_ratio_cutoff | total_UMI_counts
## < : longer object length is not a multiple of shorter object length
```

```
num_poor_quality_cells <- length(poor_quality_cells)

fraction_poor_quality_cells <- num_poor_quality_cells / length(total_UMI_counts)


cat("Number of cells with poor quality RNA:", num_poor_quality_cells, "\n")
```

```
## Number of cells with poor quality RNA: 6414
```

```
cat("Fraction of total cells with poor quality RNA:", fraction_poor_quality_cells, "\n")
```

```
## Fraction of total cells with poor quality RNA: 0.4253316
```

2. More QC and filtering (18 points) Technical limitations of single cell RNA capture lead to false zeroes in gene expression measurements, referred to as "drop-out" by the scRNAseq community. Additionally, many genes are expressed in a cell type- or state-specific manner; not all genes are expressed in all cells. Together these phenomena result in gene expression data that are incredibly sparse. Many genes expression values are disproportionately zero or near-zero across all cell types. Thus, many genes may be unimportant for downstream analyses and only serve to add noise to the data matrix.

2a (4 points) Create a data frame with gene names as row names, and two columns. The first column (call it `mean`) should contain the mean expression of the gene across all the cells. The second column (call it `zero_rate`) should contain the percentage of cells where each gene is undetected (has a `count` of zero). Use head() to print the first six rows of this data frame.

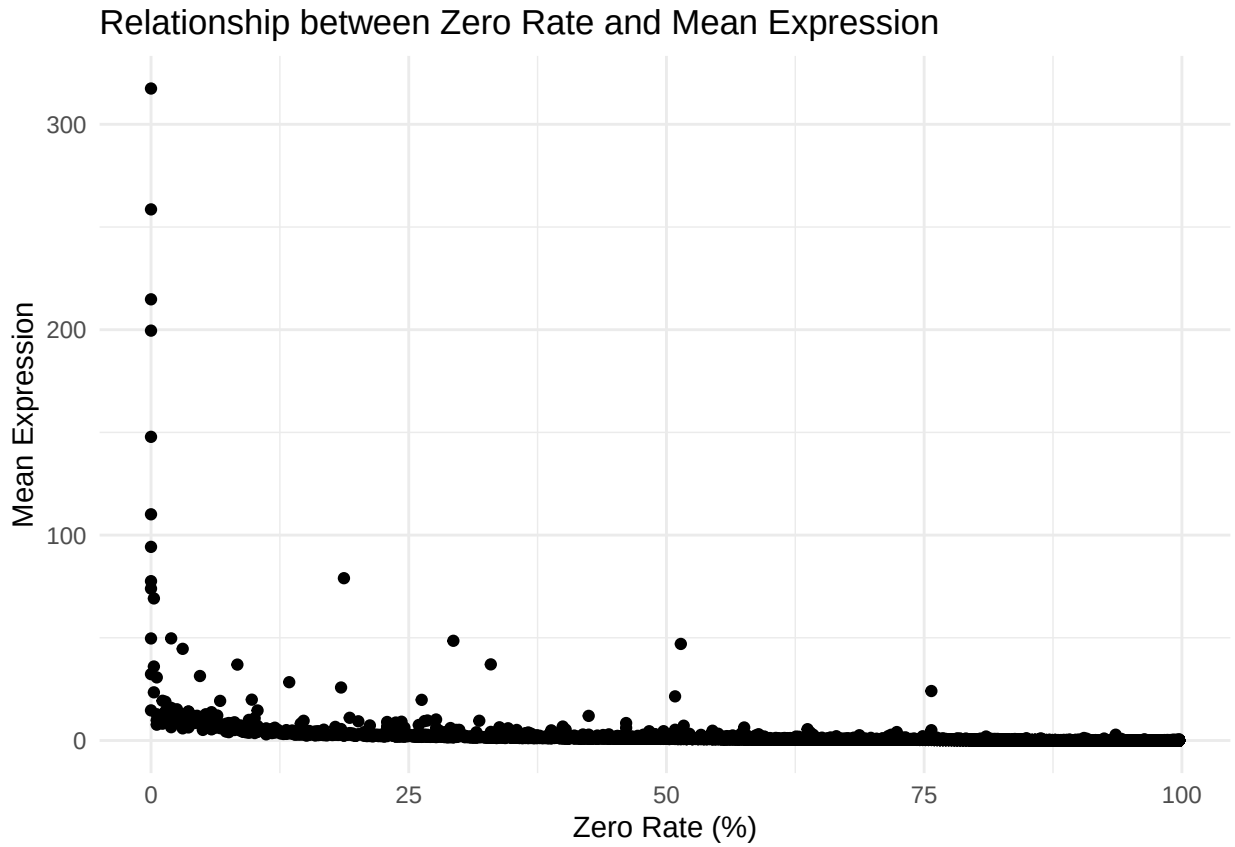
```
gene_means <- rowMeans(data$Counts_Matrix)
gene_zero_rates <- rowMeans(data$Counts_Matrix == 0) * 100
gene_summary <- data.frame(mean = gene_means, zero_rate = gene_zero_rates)
rownames(gene_summary) <- rownames(data$Counts_Matrix)
head(gene_summary)
```

```
##                mean zero_rate
## Xkr4   0.013966480  98.60335
## Gm1992 0.005586592  99.72067
## Mrpl15 0.245810056  79.60894
## Lypla1 0.145251397  87.15084
## Tcea1  0.703910615  56.98324
## Rgs20  0.050279330  95.25140
```

2b (4 points) Using this data frame, generate a scatter plot in `ggplot` to display the relationship between the zero-rate and mean expression.

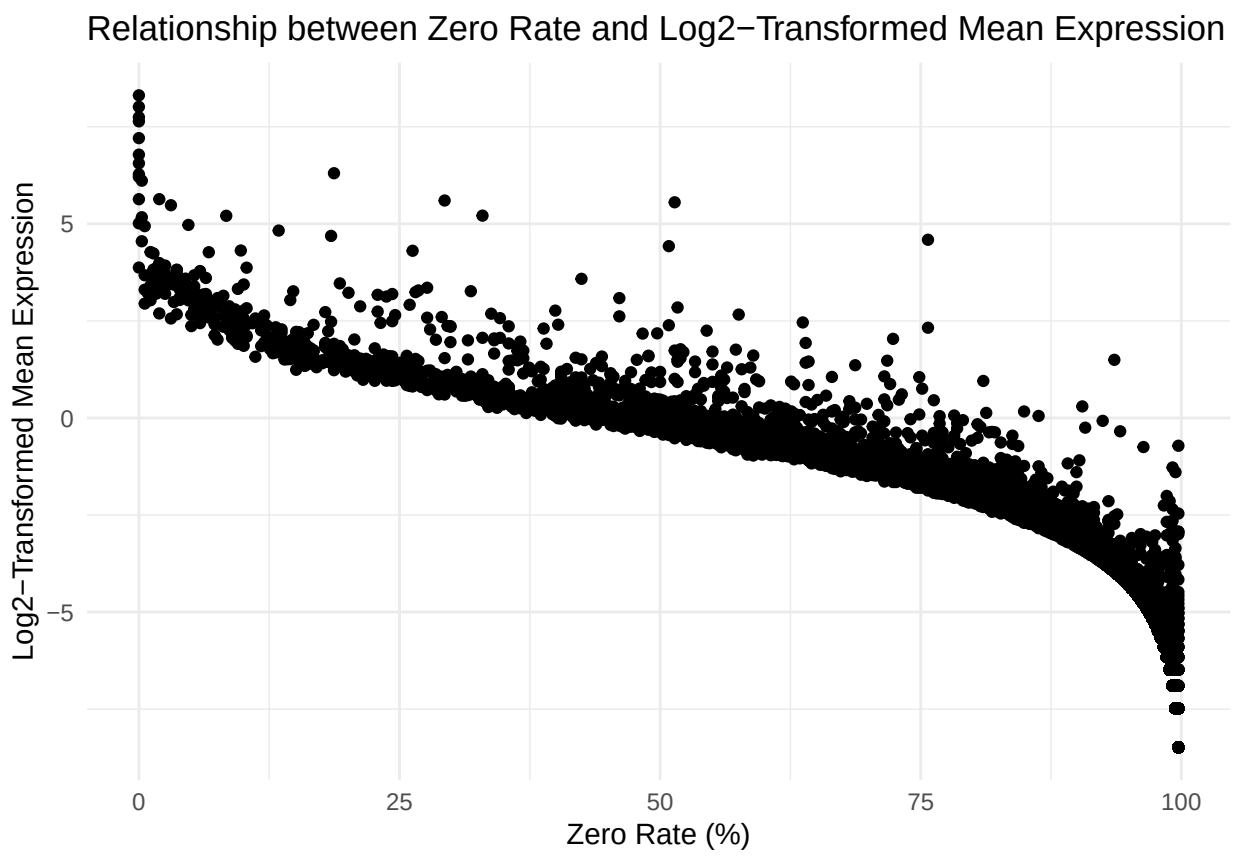
Does the relationship appear to be linear? What kind of relationship do you see if you log2-transform the mean expression values?

```
library(ggplot2)
scatter_plot <- ggplot(gene_summary, aes(x = zero_rate, y = mean)) +
  geom_point() +
  labs(title = "Relationship between Zero Rate and Mean Expression",
       x = "Zero Rate (%)",
       y = "Mean Expression") +
  theme_minimal()
print(scatter_plot)
```




```
gene_summary$log2_mean <- log2(gene_summary$mean)
scatter_plot_log2 <- ggplot(gene_summary, aes(x = zero_rate, y = log2_mean)) +
  geom_point() +
```

```
  labs(title = "Relationship between Zero Rate and Log2-Transformed Mean Expression",
       x = "Zero Rate (%)",
       y = "Log2-Transformed Mean Expression") +
  theme_minimal()
print(scatter_plot_log2)
```



written answer goes here log2-transforming the mean expression values linearizes the relationship, making it easier to understand and analyze.

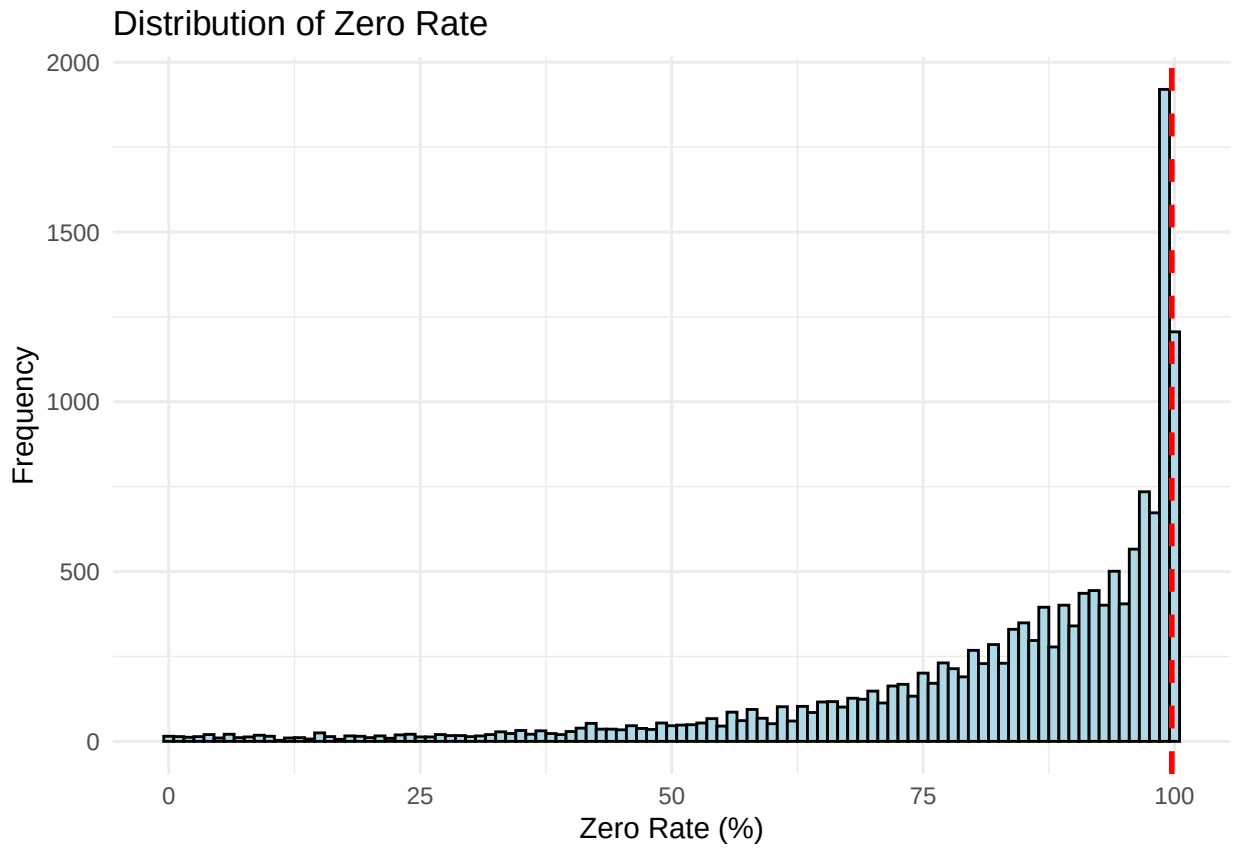
2c (5 points) Suppose you wanted to remove genes from the analysis if their zero-rate is above the 95th percentile. Using `ggplot`, generate a histogram showing the distribution of the zero-rate. Mark the 95th percentile with a dashed red line.

```
library(ggplot2)
zero_rate_hist <- ggplot(gene_summary, aes(x = zero_rate)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  geom_vline(xintercept = quantile(gene_summary$zero_rate, 0.95), color = "red", linetype = "dashed", si
  labs(title = "Distribution of Zero Rate",
       x = "Zero Rate (%)",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
```

```
## generated.
```

```
print(zero_rate_hist)
```



2d (5 points). Apply your total UMI threshold from **question 1e**, the mitochodrial ratio filter from **question 1g**, and the zero-rate threshold from **question 2c** to filter the original counts matrix. What are the dimensions of the new matrix?

```
total_UMI_threshold <- total_UMI_counts
mitochondrial_ratio_threshold <- mitochondrial_ratios
zero_rate_threshold <- gene_zero_rates

filtered_counts_matrix1 <- data$Counts_Matrix[total_UMI_counts >= total_UMI_threshold, ]

filtered_counts_matrix2 <- filtered_counts_matrix1[mitochondrial_ratios <= mitochondrial_ratio_threshol

filtered_counts_matrix3 <- filtered_counts_matrix2[gene_zero_rates, ]

new_matrix_dimensions <- dim(filtered_counts_matrix3)


new_matrix_dimensions
```

```
## [1] 15060   358
```

3. Using likelihood to predict cell types (25 points) Using a set of known marker genes for each cell type, you will now test a subset of the cells to determine the most likely cell type for each.

A "marker gene" is a gene that is expressed at a higher level in cells belonging to that cell type than in cells of the other types. Importantly, the expression levels of these marker genes typically follow a Poisson distribution near zero in cells that are **not** of the indicated cell type.

You will use these data to calculate the likelihoods of seeing the observed values of the marker genes. You will calculate three likelihood estimates for each cell – one per cell type. Since we are calculating the probability that a cell is **not** a given cell type, we will predict the cell type using the **minimum** likelihood.

Example:

- Suppose that for cell X, the probability of seeing the observed expression data, given the assumption that cell X is **not** a neuron (and is thus Poisson distributed), is .06
- the probability of seeing the observed expression data, given that cell X is **not** a sustentacular cell, is 0.15
- and the probability of seeing the observed expression data data, given that cell X is **not** a gland cell, is 0.0000001

Under these conditions, we would conclude that X is most likely a gland cell, because if it *were* a gland cell, seeing the observed data would be extremely unlikely.

(In order to use a straightforward maximum likelihood model, we could do our modeling in a positive direction instead, but the distributions of marker genes in their given cell types are quite variable, so we'd need to deal with much more complex models.)

We have 6 marker genes for each cell type. If the data were *not* derived from the indicated cell type, then the counts for the marker genes should fit a Poisson distribution. The lambda parameters for the Poisson distributions best fitting the data observed for those marker genes in the **other** two cell types (combined) are already provided for you in the RDS file containing the marker genes.

For this problem, you will determine the likelihood that each cell belongs to a certain cell type based on the marker genes. We are doing so via predicting the likelihood that a cell is *not* of a particular type by assessing how well it fits these Poisson distributions.

3a (4 points). Begin by loading `PS3_Q34_materials.rds`, which contains two data.frames. Marker_genes contains the six marker genes for each cell type. Reduced_CountsMat contains the counts matrices for these 18 marker genes in a subset of 52 of the cells.

```
file_path_2 <- "/Users/jiataizhang/Desktop/PS3_Q34_materials.rds"
data_2 <- readRDS(file_path_2)
marker_genes <- data_2$Marker_Genes
reduced_countsmat <- data_2$Reduced_CountsMat

data_2
```

```
## $Marker_Genes
##    celltype   genesymbols lambda
## 1    neuron        Cartpt 0.3038
## 2    neuron       Sult1d1 0.2405
## 3    neuron       Jakmip1 0.4177
## 4    neuron       Fam213b 0.6392
## 5    neuron 1700012B09Rik 0.4367
## 6    neuron         Tstd1 0.1519
## 7     gland          Msln 0.0340
## 8     gland        Prss33 0.0255
```

```
## 9     gland        Phlda1 0.0298
## 10    gland         Kifc3 0.0511
## 11    gland      Map1lc3a 0.7872
## 12    gland        Tagln2 0.1021
## 13      sus         Aifm3 0.0923
## 14      sus        Mapk11 0.2731
## 15      sus        Unc13b 0.2731
## 16      sus        Notch2 0.1218
## 17      sus          Fut2 0.0258
## 18      sus       Cyp39a1 0.1292
##
## $Reduced_CountsMat
##               AAACCTGGTCTCACCT AAACGGGAGCCCTAAT AAACGGGTCCTTCAAT
## Cartpt                       0                0                3
## Sult1d1                      0                0                9
## Jakmip1                      0                0               25
## Fam213b                      1                1               20
## 1700012B09Rik                0                0               19
## Tstd1                        0                0                5
## Msln                         0                0                1
## Prss33                       0                0                0
## Phlda1                       0                0                0
## Kifc3                        0                0                2
## Map1lc3a                     1                0               16
## Tagln2                       0                0                0
## Aifm3                        0                0                2
## Mapk11                       1                0                1
## Unc13b                       1                1                3
## Notch2                       0                1                0
## Fut2                         1                0                0
## Cyp39a1                      0                1                1
##               AAACGGGTCTGCTGTC AAAGCAACAGGCTGAA AAATGCCGTATCAGTC
## Cartpt                      18                6               14
## Sult1d1                      7                4                7
## Jakmip1                      6                4                4
## Fam213b                      9               10                5
## 1700012B09Rik                7                4                6
## Tstd1                        0                2                3
## Msln                         1                0                0
## Prss33                       0                0                0
## Phlda1                       0                0                0
## Kifc3                        0                0                0
## Map1lc3a                     1                1                0
## Tagln2                       0                0                0
## Aifm3                        1                0                0
## Mapk11                       0                0                0
## Unc13b                       0                0                0
## Notch2                       0                0                1
## Fut2                         0                0                0
## Cyp39a1                      0                0                0
##               AACACGTAGCGTTTAC AACCATGGTAGAAAGG AACCGCGGTCAGAATA
## Cartpt                       0                0                0
## Sult1d1                      0                1                1
## Jakmip1                      0                3                0
```

```
## Fam213b                      1                 0                 0
## 1700012B09Rik                0                 0                 1
## Tstd1                        0                 0                 0
## Msln                        12                11                16
## Prss33                       9                 1                 8
## Phlda1                       3                26                 3
## Kifc3                        2                 6                 7
## Map1lc3a                    13                38                62
## Tagln2                       5                 1                 1
## Aifm3                        0                 0                 0
## Mapk11                       1                 0                 1
## Unc13b                       0                 1                 0
## Notch2                       0                 1                 0
## Fut2                         0                 0                 0
## Cyp39a1                      0                 0                 0
##               AACTCAGGTTCTGTTT AACTGGTAGTAGGCCA AACTTTCTCTGCGTAA
## Cartpt                       7                 1                 0
## Sult1d1                      8                 0                 1
## Jakmip1                      5                 0                 1
## Fam213b                      7                 0                 2
## 1700012B09Rik                6                 0                 0
## Tstd1                        0                 0                 1
## Msln                         0                14                29
## Prss33                       0                 0                23
## Phlda1                       0                 3                11
## Kifc3                        0                 4                22
## Map1lc3a                     0                21                78
## Tagln2                       0                 5                 8
## Aifm3                        0                 0                 0
## Mapk11                       0                 0                 1
## Unc13b                       2                 1                 1
## Notch2                       0                 0                 0
## Fut2                         0                 0                 0
## Cyp39a1                      0                 0                 0
##               AAGCCGCAGTCCATAC AAGGCAGGTTCAACCA AAGGCAGTCGACAGCC
## Cartpt                       0                 4                 5
## Sult1d1                      0                 4                 1
## Jakmip1                      1                11                 6
## Fam213b                      1                 4                 8
## 1700012B09Rik                0                 4                 5
## Tstd1                        0                 2                 2
## Msln                         8                 0                 0
## Prss33                       4                 0                 0
## Phlda1                       6                 0                 0
## Kifc3                        1                 0                 0
## Map1lc3a                    40                 0                 1
## Tagln2                       4                 0                 0
## Aifm3                        0                 0                 0
## Mapk11                       1                 0                 0
## Unc13b                       0                 0                 0
## Notch2                       0                 0                 0
## Fut2                         0                 0                 0
## Cyp39a1                      2                 0                 0
##               AATCGGTTCCTACAGA ACACCAACATCATCCC ACACCAAGTCCTCCAT
```

```
## Cartpt                      15                 5                 1
## Sult1d1                      3                 4                 5
## Jakmip1                     10                 6                17
## Fam213b                     13                10                15
## 1700012B09Rik               13                10                10
## Tstd1                        2                 0                 0
## Msln                         0                 0                 0
## Prss33                       1                 0                 0
## Phlda1                       0                 0                 0
## Kifc3                        0                 0                 1
## Map1lc3a                     1                 1                 9
## Tagln2                       0                 0                 2
## Aifm3                        0                 0                 2
## Mapk11                       0                 0                 1
## Unc13b                       0                 0                 0
## Notch2                       0                 0                 0
## Fut2                         0                 0                 0
## Cyp39a1                      0                 0                 1
##                ACACCAATCCGTAGTA ACACTGACAGCTGGCT ACAGCTATCTTGAGGT
## Cartpt                      11                11                 0
## Sult1d1                      3                 4                 5
## Jakmip1                      9                13                 5
## Fam213b                     16                 8                10
## 1700012B09Rik                5                10                 6
## Tstd1                        4                 3                 1
## Msln                         0                 0                 0
## Prss33                       0                 0                 0
## Phlda1                       0                 0                 0
## Kifc3                        0                 0                 0
## Map1lc3a                     0                 1                 1
## Tagln2                       0                 0                 0
## Aifm3                        0                 0                 0
## Mapk11                       0                 0                 0
## Unc13b                       0                 1                 0
## Notch2                       0                 0                 0
## Fut2                         0                 0                 0
## Cyp39a1                      0                 0                 0
##                ACATGGTTCCCGACTT ACCTTTAAGACGCACA ACGAGGAGTGAGGGTT
## Cartpt                       4                 0                 1
## Sult1d1                      7                 1                 0
## Jakmip1                      5                 1                 0
## Fam213b                      7                 0                 0
## 1700012B09Rik               11                 0                 0
## Tstd1                        2                 1                 1
## Msln                         0                53                16
## Prss33                       0                 8                 4
## Phlda1                       0                12                 7
## Kifc3                        0                10                 1
## Map1lc3a                     0                87                20
## Tagln2                       0                28                 0
## Aifm3                        0                 0                 0
## Mapk11                       0                 0                 0
## Unc13b                       0                 1                 0
## Notch2                       0                 0                 0
```

```
## Fut2                          0                 0                 0
## Cyp39a1                       0                 0                 0
##                ACGATACAGTTCGCAT ACGATACCAGTCAGCC ACGGAGAGTTCGTGAT
## Cartpt                        1                 1                 1
## Sult1d1                       0                 4                 0
## Jakmip1                       1                12                 0
## Fam213b                       0                19                 0
## 1700012B09Rik                 0                13                 0
## Tstd1                         0                 4                 0
## Msln                          0                 0                25
## Prss33                        0                 0                 1
## Phlda1                        0                 0                 8
## Kifc3                         0                 0                13
## Map1lc3a                      0                 1                45
## Tagln2                        0                 0                 4
## Aifm3                         4                 0                 0
## Mapk11                        2                 0                 1
## Unc13b                        5                 0                 0
## Notch2                        1                 0                 0
## Fut2                          0                 0                 0
## Cyp39a1                       1                 0                 0
##                ACTGTCCGTTACTGAC ACTTACTGTATCAGTC AGAGCGATCCAAAGTC
## Cartpt                       30                26                 0
## Sult1d1                      10                 6                 1
## Jakmip1                       9                14                 0
## Fam213b                       1                 5                 0
## 1700012B09Rik                 8                 8                 1
## Tstd1                         2                 1                 0
## Msln                          0                 0                 0
## Prss33                        0                 0                 0
## Phlda1                        0                 0                 0
## Kifc3                         0                 0                 0
## Map1lc3a                      3                 1                 0
## Tagln2                        0                 0                 1
## Aifm3                         0                 0                 1
## Mapk11                        0                 0                 1
## Unc13b                        0                 0                 1
## Notch2                        0                 0                 2
## Fut2                          0                 0                 0
## Cyp39a1                       0                 0                 0
##                AGAGCTTTCTAGCACA AGAGTGGGTTAGGGTG AGCGGTCCAGATCCAT
## Cartpt                        6                16                 0
## Sult1d1                       7                 5                 0
## Jakmip1                      21                14                 0
## Fam213b                      11                 7                 0
## 1700012B09Rik                16                18                 0
## Tstd1                         0                 1                 0
## Msln                          0                 0                 0
## Prss33                        0                 0                 0
## Phlda1                        0                 0                 0
## Kifc3                         1                 0                 1
## Map1lc3a                     12                 2                 0
## Tagln2                        2                 0                 0
## Aifm3                         1                 0                 0
```

```
## Mapk11                      0                 0                 2
## Unc13b                      1                 1                 1
## Notch2                      2                 0                 0
## Fut2                        0                 0                 1
## Cyp39a1                     0                 0                 0
##               AGCGTCGCATCGATGT AGCTCCTAGTCATCCA AGCTCTCCAGCCAGAA
## Cartpt                     14                 1                 5
## Sult1d1                     3                 1                 2
## Jakmip1                    10                 0                 8
## Fam213b                     4                 0                 6
## 1700012B09Rik               7                 1                10
## Tstd1                       0                 1                 0
## Msln                        0                16                 0
## Prss33                      0                 4                 0
## Phlda1                      0                 2                 0
## Kifc3                       0                 4                 0
## Map1lc3a                    2                29                 0
## Tagln2                      0                 1                 0
## Aifm3                       0                 0                 0
## Mapk11                      0                 0                 0
## Unc13b                      0                 0                 0
## Notch2                      0                 0                 0
## Fut2                        0                 0                 0
## Cyp39a1                     0                 0                 0
##               AGGCCGTCAGCTTAAC AGGTCATTCGATGAGG AGGTCCGAGCTAGTCT
## Cartpt                      8                 5                 3
## Sult1d1                     4                 6                 4
## Jakmip1                     5                 7                 4
## Fam213b                     4                 8                12
## 1700012B09Rik               7                 8                12
## Tstd1                       1                 3                 2
## Msln                        7                 0                 0
## Prss33                      1                 0                 0
## Phlda1                      4                 0                 0
## Kifc3                       4                 0                 0
## Map1lc3a                   14                 1                 2
## Tagln2                      1                 0                 0
## Aifm3                       0                 0                 0
## Mapk11                      0                 0                 1
## Unc13b                      1                 0                 0
## Notch2                      1                 0                 0
## Fut2                        0                 0                 0
## Cyp39a1                     0                 0                 0
##               AGTGAGGCACCCATGG AGTGGGACAGCAGTTT AGTGTCAGTAATCACC
## Cartpt                      8                 0                 2
## Sult1d1                     4                 0                 3
## Jakmip1                     3                 0                13
## Fam213b                    13                 2                11
## 1700012B09Rik               6                 0                 9
## Tstd1                       3                 0                 2
## Msln                        0                15                 1
## Prss33                      0                 3                 0
## Phlda1                      0                 3                 0
## Kifc3                       0                 3                 0
```

```
## Map1lc3a                    1                32                 1
## Tagln2                      0                 2                 0
## Aifm3                       0                 0                 0
## Mapk11                      0                 0                 0
## Unc13b                      0                 0                 0
## Notch2                      0                 0                 0
## Fut2                        0                 0                 0
## Cyp39a1                     0                 0                 0
##               AGTGTCAGTTCCACTC AGTGTCATCCACGCAG ATCCGAACATGGGAAC
## Cartpt                      7                20                 1
## Sult1d1                     7                10                 0
## Jakmip1                     8                12                 0
## Fam213b                    12                 5                 2
## 1700012B09Rik               8                 9                 1
## Tstd1                       2                 1                 0
## Msln                        0                 0                17
## Prss33                      0                 0                 3
## Phlda1                      0                 0                14
## Kifc3                       0                 0                 7
## Map1lc3a                    0                 0                20
## Tagln2                      0                 0                 0
## Aifm3                       0                 0                 0
## Mapk11                      0                 0                 1
## Unc13b                      0                 0                 1
## Notch2                      0                 0                 1
## Fut2                        0                 0                 0
## Cyp39a1                     0                 0                 0
##               ATCGAGTGTGCACTTA ATGAGGGAGCCAGAAC ATGGGAGAGGCTACGA
## Cartpt                     16                 1                11
## Sult1d1                     3                 0                 5
## Jakmip1                     7                 0                11
## Fam213b                    11                 1                 7
## 1700012B09Rik               7                 0                 6
## Tstd1                       0                 0                 1
## Msln                        0                20                 0
## Prss33                      0                 1                 0
## Phlda1                      0                16                 0
## Kifc3                       0                 5                 0
## Map1lc3a                    2                16                 0
## Tagln2                      4                 3                 1
## Aifm3                       1                 0                 0
## Mapk11                      0                 0                 0
## Unc13b                      1                 0                 1
## Notch2                      1                 0                 0
## Fut2                        0                 0                 0
## Cyp39a1                     0                 0                 0
##               ATTACTCGTAGCTCCG ATTGGACGTGTCGCTG ATTGGTGCACAACTGT
## Cartpt                      0                14                25
## Sult1d1                     3                 2                10
## Jakmip1                    11                 8                13
## Fam213b                     8                 9                14
## 1700012B09Rik              10                 2                 9
## Tstd1                       1                 4                 2
## Msln                        0                 0                 0
```

```
## Prss33                       0                0                0
## Phlda1                       0                0                0
## Kifc3                        0                0                0
## Map1lc3a                     0                0                0
## Tagln2                       0                0                0
## Aifm3                        0                0                0
## Mapk11                       2                0                0
## Unc13b                       0                0                0
## Notch2                       0                0                0
## Fut2                         0                0                0
## Cyp39a1                      0                0                0
##               ATTGGTGGTCCCTACT
## Cartpt                       0
## Sult1d1                      0
## Jakmip1                      1
## Fam213b                      2
## 1700012B09Rik                0
## Tstd1                        0
## Msln                         7
## Prss33                       6
## Phlda1                       2
## Kifc3                        7
## Map1lc3a                    34
## Tagln2                       2
## Aifm3                        0
## Mapk11                       1
## Unc13b                       1
## Notch2                       0
## Fut2                         0
## Cyp39a1                      0
##
## $Celltype_Groundtruth
##  [1] "sus"    "sus"    "neuron" "neuron" "neuron" "neuron" "gland"  "gland"
##  [9] "gland"  "neuron" "gland"  "gland"  "gland"  "neuron" "neuron" "neuron"
## [17] "neuron" "neuron" "neuron" "neuron" "neuron" "neuron" "gland"  "gland"
## [25] "sus"    "neuron" "gland"  "neuron" "neuron" "sus"    "neuron" "neuron"
## [33] "sus"    "neuron" "gland"  "neuron" "neuron" "neuron" "neuron" "neuron"
## [41] "gland"  "neuron" "neuron" "neuron" "gland"  "neuron" "gland"  "neuron"
## [49] "neuron" "neuron" "neuron" "gland"
```

For the first cell in Reduced_CountsMat, with barcode 'AAACCTGGTCTCACCT', calculate the probability of seeing the observed data for gene 'Cartpt' (the first gene in the matrix Reduced_CountsMat) under the assumption that it was generated from a Poisson distribution with the lambda value associated with Cartpt in Marker_Genes. This is the probability that you would see the observed expression value for Cartpt if the cell is not a neuron.

```
observed_value <- 0

lambda_value <- 0.3038


probability <- dpois(observed_value, lambda_value)

probability
```

```
## [1] 0.7380085
```

3b (6 points).

Now, compute the probability above for all six neuron markers, and combine these probabilities to give a single probability of the observed data based on those six markers. That is, for this same cell, with barcode 'AAACCTGGTCTCACCT', compute the probability that you would see the observed counts for all six neuron markers, assuming that the observed counts for each marker came from a Poisson distribution with the given lambda value *for that marker*, and then combine them into one joint probability. This is easily done with just one call to dpois() if you have a vector of the observed counts for the marker genes and another of the lambdas for the corresponding genes in the same order. You need only one R command to combine the individual marker probabilities.

```
neuron_marker_genes <- c("Cartpt", "Sult1d1", "Jakmip1", "Fam213b", "1700012B09Rik", "Tstd1")
lambda_values_neuron <- marker_genes[marker_genes$celltype == "neuron" & marker_genes$genesymbols %in% n

observed_expression_neuron <- reduced_countsmat[neuron_marker_genes, "AAACCTGGTCTCACCT"]
individual_probabilities_neuron <- dpois(observed_expression_neuron, lambda_values_neuron)
joint_probability_neuron <- prod(individual_probabilities_neuron)
print(joint_probability_neuron)
```

```
## [1] 0.07155149
```

Do the same thing for sus cells.

```
sus_marker_genes <- c("Aifm3", "Mapk11", "Unc13b", "Notch2", "Fut2", "Cyp39a1")
lambda_values_sus <- marker_genes[marker_genes$celltype == "sus" & marker_genes$genesymbols %in% sus_ma
observed_expression_sus <- reduced_countsmat[sus_marker_genes, "AAACCTGGTCTCACCT"]
individual_probabilities_sus <- dpois(observed_expression_sus, lambda_values_sus)
joint_probability_sus <- prod(individual_probabilities_sus)
print(joint_probability_sus)
```

```
## [1] 0.0007704658
```

And again for gland cells.

```
gland_marker_genes <- c("Msln", "Prss33", "Phlda1", "Kifc3", "Map1lc3a", "Tagln2")
lambda_values_gland <- marker_genes[marker_genes$celltype == "gland" & marker_genes$genesymbols %in% gla

observed_expression_gland <- reduced_countsmat[gland_marker_genes, "AAACCTGGTCTCACCT"]
individual_probabilities_gland <- dpois(observed_expression_gland, lambda_values_gland)
joint_probability_gland <- prod(individual_probabilities_gland)
print(joint_probability_gland)
```

```
## [1] 0.2811202
```

3c (4 points). Now you have computed the probability of the data for the first cell given that it is not a neuron, given that it is not a gland cell, and given that it is not a sustentacular cell.

Given these values, which type of cell do you think it is?
The cell type with the lowest probability is "sus cell" with a probability of 0.0007704658. herefore, based on these probabilities, the first cell is most likely a sus cell.

3d (6 points). Now we'll generalize these computations to apply (hint!) to all the cell types. You may want to write helper functions to compute these values.

For each cell type (neuron, sus, gland), create a vector with 52 elements containing the joint probability of the data for each cell given the observed values of all the marker genes, assuming that the cell is not of the given type.

```
joint_prob <- function(barcode, cell_type, count_matrix, marker_data) {
  marker_genes <- marker_data$genesymbols[marker_data$celltype == cell_type]
  obs_cells <- count_matrix[marker_genes, barcode]
  lambda <- marker_data$lambda[marker_data$genesymbols %in% marker_genes]
  probs <- dpois(obs_cells, lambda)
  return(prod(probs))
}

help_function <- function(barcode, count_matrix, marker_data) {
  prob_neuron <- joint_prob(barcode, "neuron", count_matrix, marker_data)
  prob_sus <- joint_prob(barcode, "sus", count_matrix, marker_data)
  prob_gland <- joint_prob(barcode, "gland", count_matrix, marker_data)

  return(c(neuron = prob_neuron, sus = prob_sus, gland = prob_gland))
}

barcodes <- colnames(reduced_countsmat)

results <- sapply(barcodes, help_function, count_matrix = reduced_countsmat, marker_data = marker_genes)

results_df <- data.frame(results)
rownames(results_df) <- 1:nrow(results_df)

head(results_df)
```

```
##   AAACCTGGTCTCACCT AAACGGGAGCCCTAAT AAACGGGTCCTTCAAT AAACGGGTCTGCTGTC
## 1     0.0715514946      0.071551495    3.576327e-102     1.707143e-53
## 2     0.0007704658      0.001720766     2.042969e-07     3.695660e-02
## 3     0.2811202028      0.357114079     1.647553e-20     9.558087e-03
##   AAAGCAACAGGCTGAA AAATGCCGTATCAGTC AACACGTAGCGTTTAC AACCATGGTAGAAAGG
## 1     1.185451e-27     4.292550e-42     7.155149e-02     3.269936e-04
## 2     4.003965e-01     4.876829e-02     1.093483e-01     1.331862e-02
## 3     2.811202e-01     3.571141e-01     8.519188e-74    4.606751e-153
##   AACCGCGGTCAGAATA AACTCAGGTTCTGTTT AACTGGTAGTAGGCCA AACTTTCTCTGCGTAA
## 1     1.175656e-02     1.300171e-32     3.400711e-02     0.0003489497
## 2     1.093483e-01     1.493151e-02     1.093483e-01     0.0298630156
## 3    2.249412e-166     3.571141e-01     1.684472e-73     0.0000000000
##   AAGCCGCAGTCCATAC AAGGCAGGTTCAACCA AAGGCAGTCGACAGCC AATCGGTTCCTACAGA
## 1     2.988706e-02     1.139377e-27     4.519630e-24     2.848124e-63
## 2     9.126557e-04     4.003965e-01     4.003965e-01     4.003965e-01
## 3     5.359242e-96     3.571141e-01     2.811202e-01     7.168565e-03
##   ACACCAACATCATCCC ACACCAAGTCCTCCAT ACACCAATCCGTAGTA ACACTGACAGCTGGCT
## 1     5.413978e-34     1.484951e-53     1.537539e-51     4.223410e-53
## 2     4.003965e-01     6.017940e-05     4.003965e-01     1.093483e-01
## 3     2.811202e-01     3.042651e-11     3.571141e-01     2.811202e-01
##   ACAGCTATCTTGAGGT ACATGGTTCCCGACTT ACCTTTAAGACGCACA ACGAGGAGTGAGGGTT
## 1     3.651290e-25     1.070557e-35      0.001708124     5.165680e-03
```

```
## 2     4.003965e-01     4.003965e-01      0.109348281     4.003965e-01
## 3     2.811202e-01     3.571141e-01      0.000000000     6.964354e-82
##   ACGATACAGTTCGCAT ACGATACCAGTCAGCC ACGGAGAGTTCGTGAT ACTGTCCGTTACTGAC
## 1     1.420477e-02     3.483217e-59     3.400711e-02     5.834818e-81
## 2     8.995727e-15     4.003965e-01     1.093483e-01     4.003965e-01
## 3     3.571141e-01     2.811202e-01    3.639017e-174     2.903427e-02
##   ACTTACTGTATCAGTC AGAGCGATCCAAAGTC AGAGCTTTCTAGCACA AGAGTGGGTTAGGGTG
## 1     6.579647e-76     0.0117565587     3.677235e-72     7.301232e-73
## 2     4.003965e-01     0.0000204456     7.486490e-05     1.093483e-01
## 3     2.811202e-01     0.0364613474     1.124433e-14     1.106489e-01
##   AGCGGTCCAGATCCAT AGCGTCGCATCGATGT AGCTCCTAGTCATCCA AGCTCTCCAGCCAGAA
## 1     0.1119391342     3.162756e-41     5.425325e-04     1.056555e-29
## 2     0.0001052071     4.003965e-01     4.003965e-01     4.003965e-01
## 3     0.0182485294     1.106489e-01     1.354032e-90     3.571141e-01
##   AGGCCGTCAGCTTAAC AGGTCATTCGATGAGG AGGTCCGAGCTAGTCT AGTGAGGCACCCATGG
## 1     1.889292e-27     3.782326e-36     1.040737e-36     9.165184e-37
## 2     1.331862e-02     4.003965e-01     1.093483e-01     4.003965e-01
## 3     3.640858e-44     2.811202e-01     1.106489e-01     2.811202e-01
##   AGTGGGACAGCAGTTT AGTGTCAGTAATCACC AGTGTCAGTTCCACTC AGTGTCATCCACGCAG
## 1     2.286786e-02     7.587510e-41     4.137813e-45     4.661045e-69
## 2     4.003965e-01     4.003965e-01     4.003965e-01     4.003965e-01
## 3     6.503421e-92     9.558087e-03     3.571141e-01     3.571141e-01
##   ATCCGAACATGGGAAC ATCGAGTGTGCACTTA ATGAGGGAGCCAGAAC ATGGGAGAGGCTACGA
## 1     3.033866e-03     3.149793e-48     2.173734e-02     8.192013e-43
## 2     3.637315e-03     1.229309e-03     4.003965e-01     1.093483e-01
## 3    9.312398e-112     5.010013e-07    1.578456e-115     3.646135e-02
##   ATTACTCGTAGCTCCG ATTGGACGTGTCGCTG ATTGGTGCACAACTGT ATTGGTGGTCCCTACT
## 1     3.204274e-33     5.055009e-42     1.132019e-91     9.551904e-03
## 2     1.493151e-02     4.003965e-01     4.003965e-01     2.986302e-02
## 3     3.571141e-01     3.571141e-01     3.571141e-01     5.892581e-88
```

3e (5 points). Negative log likelihood values are easier to read than actual probabilities. Take the natural log (using the default log function with no arguments) of these probability vectors and multiply by -1. Figure out which cell type has the maximum negative log likelihood score for each cell/barcode. This is the cell type to predict! Create a vector of the 52 predicted cell types.

```
neg_log_likelihood <- function(probs) {
  return(-1 * log(probs))
}

predict_cell_type <- function(neg_log_likelihoods) {
  max_score_index <- which.max(neg_log_likelihoods)
  cell_types <- names(neg_log_likelihoods)
  return(cell_types[max_score_index])
}

barcodes <- colnames(reduced_countsmat)
prob_results <- t(sapply(barcodes, help_function, count_matrix = reduced_countsmat, marker_data = marke

neg_log_likelihood_results <- apply(prob_results, 2, neg_log_likelihood)

predicted_cell_types <- apply(neg_log_likelihood_results, 1, predict_cell_type)

print(predicted_cell_types)
```

```
## AAACCTGGTCTCACCT AAACGGGAGCCCTAAT AAACGGGTCCTTCAAT AAACGGGTCTGCTGTC
##            "sus"            "sus"         "neuron"         "neuron"
## AAAGCAACAGGCTGAA AAATGCCGTATCAGTC AACACGTAGCGTTTAC AACCATGGTAGAAAGG
##         "neuron"         "neuron"          "gland"          "gland"
## AACCGCGGTCAGAATA AACTCAGGTTCTGTTT AACTGGTAGTAGGCCA AACTTTCTCTGCGTAA
##          "gland"         "neuron"          "gland"          "gland"
## AAGCCGCAGTCCATAC AAGGCAGGTTCAACCA AAGGCAGTCGACAGCC AATCGGTTCCTACAGA
##          "gland"         "neuron"         "neuron"         "neuron"
## ACACCAACATCATCCC ACACCAAGTCCTCCAT ACACCAATCCGTAGTA ACACTGACAGCTGGCT
##         "neuron"         "neuron"         "neuron"         "neuron"
## ACAGCTATCTTGAGGT ACATGGTTCCCGACTT ACCTTTAAGACGCACA ACGAGGAGTGAGGGTT
##         "neuron"         "neuron"          "gland"          "gland"
## ACGATACAGTTCGCAT ACGATACCAGTCAGCC ACGGAGAGTTCGTGAT ACTGTCCGTTACTGAC
##            "sus"         "neuron"          "gland"         "neuron"
## ACTTACTGTATCAGTC AGAGCGATCCAAAGTC AGAGCTTTCTAGCACA AGAGTGGGTTAGGGTG
##         "neuron"            "sus"         "neuron"         "neuron"
## AGCGGTCCAGATCCAT AGCGTCGCATCGATGT AGCTCCTAGTCATCCA AGCTCTCCAGCCAGAA
##            "sus"         "neuron"          "gland"         "neuron"
## AGGCCGTCAGCTTAAC AGGTCATTCGATGAGG AGGTCCGAGCTAGTCT AGTGAGGCACCCATGG
##          "gland"         "neuron"         "neuron"         "neuron"
## AGTGGGACAGCAGTTT AGTGTCAGTAATCACC AGTGTCAGTTCCACTC AGTGTCATCCACGCAG
##          "gland"         "neuron"         "neuron"         "neuron"
## ATCCGAACATGGGAAC ATCGAGTGTGCACTTA ATGAGGGAGCCAGAAC ATGGGAGAGGCTACGA
##          "gland"         "neuron"          "gland"         "neuron"
## ATTACTCGTAGCTCCG ATTGGACGTGTCGCTG ATTGGTGCACAACTGT ATTGGTGGTCCCTACT
##         "neuron"         "neuron"         "neuron"          "gland"
```

4. Visually comparing min likelihood to KNN predictors (27 points) 4a (5 points) Create a single data frame containing the maximum likelihood cell type prediction vector from above and the x and y coordinates for those particular cells, derived from the `Plot_Coordinates` provided, for all 358 cells that you loaded in the RDS file for Question 1.

Verify that you have the right cell barcode names corresponding to the predicted type for that cell and the corresponding plot values. The row names for each cell should be the cell's barcode.

Then, use that data frame to generate a scatterplot with `ggplot2` that colors each point according to its predicted cell type. How well do your predictions and these projected x and y plot locations separate the three types of cells?

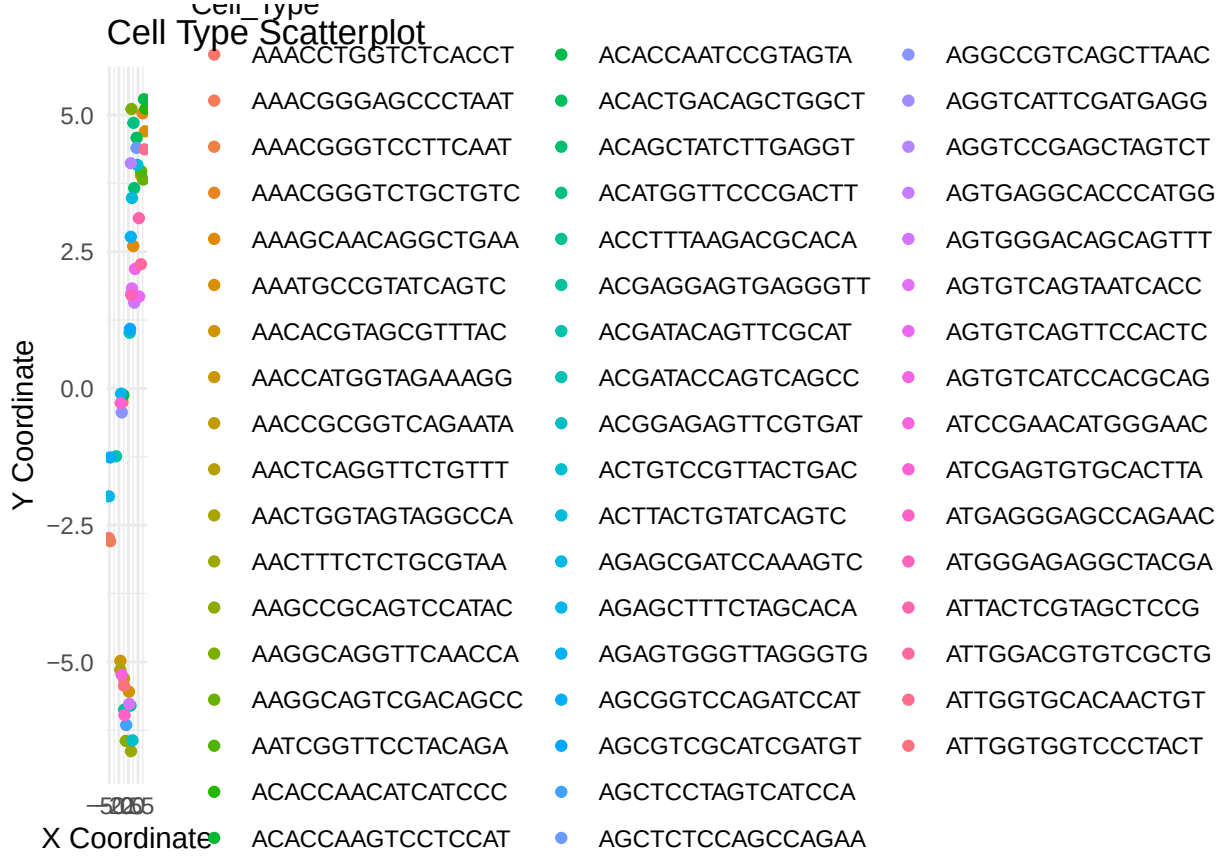
```
matching_barcodes <- barcodes
matching_data <- data$Plot_Coords[matching_barcodes, ]

data_frame <- data.frame(
  Cell_Barcode = matching_barcodes,
  Cell_Type =barcodes ,
  X = matching_data$V1,
  Y = matching_data$V2
)

library(ggplot2)


ggplot(data_frame, aes(x = X, y = Y, color = Cell_Type)) +
  geom_point() +
```

```
  labs(title = "Cell Type Scatterplot", x = "X Coordinate", y = "Y Coordinate") +
  theme_minimal()
```



4b (10 points)

Using the train() function in the caret package, assess knn as a predictor *on the RNA count data in Q3* using 5-fold cross validation, only for k=3. That is, do not tune the method to find the optimal k.

You will need to do something with your count data matrix so the genes can be used as "features". The ground truth cell type labels are also provided in the RDS file.

Save the output of this function in the variable `knn.fit`.

Use the predict() function with newdata set to the same training data set to determine how the cross-validation models performed on the training data itself.

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
##
##     melanoma
```

```
RNA_Counts <- data_2$Reduced_CountsMat
cell_labels <- marker_genes$celltype

data_df <- data.frame(RNA_Counts)

data_df$CellType <- factor(cell_labels)

ctrl <- trainControl(method = "cv", number = 5)

knn.fit <- train(CellType ~ ., data = data_df, method = "knn", trControl = ctrl, tuneGrid = data.frame(

print(knn.fit)
```

```
## k-Nearest Neighbors
##
## 18 samples
## 52 predictors
##  3 classes: 'gland', 'neuron', 'sus'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 15, 14, 15, 15, 13
## Resampling results:
##
##   Accuracy   Kappa
##   0.8933333  0.8375
##
## Tuning parameter 'k' was held constant at a value of 3
```

4c (7 points)

Use the confusionMatrix() function in the caret package to cross-tabulate your predicted results for knn prediction and for minimum likelihood.

```
library(caret)

knn_predictions <- predict(knn.fit, newdata = data_df)
confusion_knn <- confusionMatrix(knn_predictions, data_df$CellType)

print(confusion_knn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction gland neuron sus
##     gland      6      0   0
##     neuron     0      5   0
##     sus        0      1   6
##
## Overall Statistics
##
##                Accuracy : 0.9444
##                  95% CI : (0.7271, 0.9986)
```

```
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : 9.55e-08
##
##                   Kappa : 0.9167
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: gland Class: neuron Class: sus
## Sensitivity                1.0000        0.8333     1.0000
## Specificity                1.0000        1.0000     0.9167
## Pos Pred Value             1.0000        1.0000     0.8571
## Neg Pred Value             1.0000        0.9231     1.0000
## Prevalence                 0.3333        0.3333     0.3333
## Detection Rate             0.3333        0.2778     0.3333
## Detection Prevalence       0.3333        0.2778     0.3889
## Balanced Accuracy          1.0000        0.9167     0.9583
```

4d (5 points)

Knn typically does not perform well on unbalanced data sets. Did it do well here? What might account for what you are seeing?

write answer here

K-nearest neighbors (KNN) can perform poorly on unbalanced datasets, where some classes have many more samples than others. The dataset's class imbalance might lead to KNN struggling to correctly classify minority classes. To assess its performance, analyze the confusion matrix, consider class balance, explore resampling techniques, evaluate alternative algorithms, optimize the value of K, and use appropriate evaluation metrics like precision, recall, and F1-score. Ensemble methods may also be useful in improving classification performance on imbalanced datasets.