Getting Data and Databases - HW

Kelsey Martin

2024-03-07

Assignment 6

is due on March 8th, 2024

Instructions:

Please turn in the answers to this assignment as a knitted R markdown document.

- For this, don't forget to comment out all package installations after you install them! You don't want to keep installing things over and over!

Answers will be graded for correctness, completeness, and how well the instructions are followed.

Turn in your assignment on canvas

```
@ These are the learning objectives associated with each question
```

QUESTION 1 (5 pts)

```
@ Students will display that they understand how to obtain data from the SRA.
```

Go to the SRA website.

This time, search for an organism of your interest, along with a tissue or condition of your interest. You can further narrow it down to a type of NGS platform (i.e., RNAseq) if you'd like.

Find something that you are interested in and that seems like a good candidate. Don't get anything too big either.

In your search results, **click on your sample of interest**. You should be pulled to a new page. Here, you will see a bunch of information on your sample. Under **Runs**, go ahead and click on your sample accession number. This will pull you to the SRA browser were you can see **Metadata, Analaysis, Reads, Data Access, and FASTA/FASTQ download tabs.**

Go under FASTA/FASTQ and download the FASTQ file.

Go ahead and put your FASTQ files in a new directory called seq_data. Make sure all file paths are correct!

Now read your file in.

- info about my files: This is ChIP-seq data that was used in a study on LSR2, which is a pleiotropic transcription factor that is differentially regulated between smooth and rough morphotypes of [Mycobacterium abscessus]. The one ending in 57 is the WT Rough morphotype, and the one ending in 52 is the WT Smooth morphotype.

```
#if (!require("BiocManager", quietly = TRUE))
#    install.packages("BiocManager")

#BiocManager::install("ShortRead")
library(ShortRead)
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
## Loading required package: BiocParallel
```

```
## Loading required package: Biostrings
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:utils':
##
##     findMatches
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
##
## Attaching package: 'IRanges'

## The following object is masked from 'package:grDevices':
##
##     windows

## Loading required package: XVector

## Loading required package: GenomeInfoDb

##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:base':
##
##     strsplit

## Loading required package: Rsamtools

## Loading required package: GenomicRanges

## Loading required package: GenomicAlignments

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'
```

```
## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians
```

```
## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
#THe data I got is from ChIP-seq, which I hope will still work for this... All the other files I was lo
seq1 <- readFastq("./seq_data/SRR25491352.fastq.gz", pattern=character(0))



# returns sequence information
sread(seq1)
```

```
## DNAStringSet object of length 3861616:
##           width seq
##       [1]   151 ATGGACGGTGTTTACGCCACTGGTTGGTCGA...GAACTCCAGTCACACTTGAATCACGTATGC
##       [2]   151 TACTACTCTACTCAGTGACTACTTGTCAAAC...GGGAAAGAGTGTAGATCTCGGTGGTCGCCG
##       [3]   151 ATCTGCACGATCTGCTGGTCACCGAGAAGGT...CCCGCCGGAGGTGGAGATCGGAAGAGCACA
##       [4]   151 CCACCTCCGGCGGGCACGGCTCACCCGCCAC...CAGATCATGCAGATAGATCGGAAGAGCGTC
##       [5]   151 GATGCGGGAAGACCGTTCTGCGTTACGAGGA...GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
##       ...   ... ...
## [3861612]   151 AGTCCGGCCCAGCTGCCCCTCCCGAGGAGGA...TTAAAAAGGGGGGGGGGGGGGGGGGGGGG
## [3861613]   151 GATCGGAAGAGCACACGTCTGAACTCCAGTC...GGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## [3861614]   151 GGGGGGGGGGGGGGGGGGGGGGGGGGGGGG...GGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## [3861615]   151 GAACATTTCCTTCTTGATCACCTGGAACTGC...TCGGAAGAGCACACGTCTGAACTCCAGTCT
## [3861616]   151 CCACGCGTTGCCGCACGCCCGCATCCTGATG...TCGGAAGAGCGTCGTGTAGGGAAAGAGTGT
```

```
# returns the Phred quality score
quality(seq1) # Phred-scaled quality score represents how confident we are in the assignment of each ba
```

```
## class: FastqQuality
## quality:
## BStringSet object of length 3861616:
##           width seq
##       [1]   151 ????????????????????????????????...????????????????????????????????
##       [2]   151 ????????????????????????????????...????????????????????????????????
##       [3]   151 ????????????????????????????????...????????????????????????????????
```

```
##       [4]   151 ????????????????????????????????...????????????????????????????
##       [5]   151 ????????????????????????????????...????????????????????????????
##       ...   ... ...
## [3861612]   151 ????????????????????????????????...????????????????????????????
## [3861613]   151 ????????????????????????????????...????????????????????????????
## [3861614]   151 ????????????????????????????????...????????????????????????????
## [3861615]   151 ????????????????????????????????...????????????????????????????
## [3861616]   151 ????????????????????????????????...????????????????????????????
```

```
# returns the read ID and length
id(seq1)
```

```
## BStringSet object of length 3861616:
##           width seq
##       [1]    70 SRR25491352.1.1 FS10000891:8:BR...29:1:1101:1070:1000 length=151
##       [2]    70 SRR25491352.1.2 FS10000891:8:BR...29:1:1101:1070:1000 length=151
##       [3]    70 SRR25491352.2.1 FS10000891:8:BR...29:1:1101:1080:1000 length=151
##       [4]    70 SRR25491352.2.2 FS10000891:8:BR...29:1:1101:1080:1000 length=151
##       [5]    70 SRR25491352.3.1 FS10000891:8:BR...29:1:1101:1150:1000 length=151
##       ...   ... ...
## [3861612]    77 SRR25491352.1930806.2 FS1000089...9:1:1116:16310:4090 length=151
## [3861613]    77 SRR25491352.1930807.1 FS1000089...9:1:1116:16320:4090 length=151
## [3861614]    77 SRR25491352.1930807.2 FS1000089...9:1:1116:16320:4090 length=151
## [3861615]    77 SRR25491352.1930808.1 FS1000089...9:1:1116:16350:4090 length=151
## [3861616]    77 SRR25491352.1930808.2 FS1000089...9:1:1116:16350:4090 length=151
```

go ahead and read in another sample file or two from the same BioProject. You can add your additional code below.

```
# again these file names are just examples of what i did, you will have to modify the paths.

seq2 <- readFastq("./seq_data/SRR25491357.fastq.gz", pattern=character(0))
```

QUESTION 2 (5 pts)

```
@ Students will do some basic analysis on their downloaded FASTQ file(s)
```

Using the samples you read in above, lets try out some more of shortread's functions to interrogate the data.

```
head(seq1)
```

```
## class: ShortReadQ
## length: 6 reads; width: 151 cycles
```

```
head(seq2)
```

```
## class: ShortReadQ
## length: 6 reads; width: 151 cycles
```

```r
summary(seq1)
```

```
##     Length      Class       Mode
##    3861616 ShortReadQ         S4
```

```r
summary(seq2)
```

```
##     Length      Class       Mode
##    2676736 ShortReadQ         S4
```

```r
# Performs qualtiy assessment on short reads

qa_data <- qa("./seq_data/", pattern=character(0), type="fastq") # this will output some information ab
print(qa_data)
```

```
## class: FastqQA(10)
## QA elements (access with qa[["elt"]]):
##   readCounts: data.frame(2 3)
##   baseCalls: data.frame(2 5)
##   readQualityScore: data.frame(1024 4)
##   baseQuality: data.frame(190 3)
##   alignQuality: data.frame(2 3)
##   frequentSequences: data.frame(100 4)
##   sequenceDistribution: data.frame(185 4)
##   perCycle: list(2)
##     baseCall: data.frame(1208 4)
##     quality: data.frame(302 5)
##   perTile: list(2)
##     readCounts: data.frame(0 4)
##     medianReadQualityScore: data.frame(0 4)
##   adapterContamination: data.frame(2 1)
```

```r
#it says I access parts of the qa analysis with indexing using double square brackets

qa_data[["frequentSequences"]]
```

```
##
## 1   GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 2   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 3   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 4   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 5   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 6   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 7   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTATTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 8   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 9   GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 10  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 11  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 12  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 13  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCCTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 14  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

```
## 15  CATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 16  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGGATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 17  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 18  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 19  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGGAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 20  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTATTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 21  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 22  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGCCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 23  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 24  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTGTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 25  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 26  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 27  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTTATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 28  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGCATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 29  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCCTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 30  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTAATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 31  AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 32  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCCTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 33  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 34  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGTTTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 35  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTGGAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 36  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTAATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 37  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 38  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 39  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 40  GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTGAAAAGAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 41  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTATTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 42  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 43  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTGAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 44  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCGGGGGGTTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 45  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 46  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTGGAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 47  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTGTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 48  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 49  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 50  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTTATGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 51  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 52  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 53  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 54  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 55  AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 56  AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 57  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 58  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 59  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 60  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 61  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 62  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTGGGGGGGGGGGGGGGGTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 63  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 64  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 65  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 66  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 67  CATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 68  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAATTGTGGGGGGGGGGGGTTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
```

```
## 69  GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 70  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCACGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 71  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGGTTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 72  ATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 73  GATCGGAAGAGCACACGTCTGAACTCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 74  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 75  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCCCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 76  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 77  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAATTTTGGGGGGGGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 78  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 79  GAATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGG
## 80  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 81  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 82  AATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGG
## 83  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGGGGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 84  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 85  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTGGGGGGGGGGGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 86  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 87  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTGGGGGGGGGGGGGGGGTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 88  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 89  GATCGGAAGAGCACACGTTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 90  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 91  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAGAGGGGGGGGGGGGGGGGGGGGGGGGGG
## 92  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAGGGGGGGGGGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 93  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTGGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 94  GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTGGTATGCCGTCTTCTGCTTGAAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGG
## 95  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGGGGGGGGGGGGGGGGTTTTTAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 96  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGTTTGGGGGGGGGGGGGGGTTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 97  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAATTTTGGGGGGCGGGGGGTTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 98  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCGGGGGGTTTTTAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 99  GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGTTTTTAAGAGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
## 100 GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGTTGGGGGGGGCGGGGGGTTTTTAAAAAGGGGGGGGGGGGGGGGGGGGGGGGGGGG
##     count type                 lane
## 1   46116 read SRR25491352.fastq.gz
## 2    4121 read SRR25491352.fastq.gz
## 3    3184 read SRR25491352.fastq.gz
## 4    2839 read SRR25491352.fastq.gz
## 5    2597 read SRR25491352.fastq.gz
## 6    1318 read SRR25491352.fastq.gz
## 7    1134 read SRR25491352.fastq.gz
## 8     972 read SRR25491352.fastq.gz
## 9     846 read SRR25491352.fastq.gz
## 10    826 read SRR25491352.fastq.gz
## 11    677 read SRR25491352.fastq.gz
## 12    553 read SRR25491352.fastq.gz
## 13    533 read SRR25491352.fastq.gz
## 14    457 read SRR25491352.fastq.gz
## 15    442 read SRR25491352.fastq.gz
## 16    414 read SRR25491352.fastq.gz
## 17    408 read SRR25491352.fastq.gz
## 18    400 read SRR25491352.fastq.gz
## 19    393 read SRR25491352.fastq.gz
## 20    332 read SRR25491352.fastq.gz
## 21    321 read SRR25491352.fastq.gz
```

```
## 22    285 read SRR25491352.fastq.gz
## 23    284 read SRR25491352.fastq.gz
## 24    274 read SRR25491352.fastq.gz
## 25    273 read SRR25491352.fastq.gz
## 26    267 read SRR25491352.fastq.gz
## 27    253 read SRR25491352.fastq.gz
## 28    252 read SRR25491352.fastq.gz
## 29    224 read SRR25491352.fastq.gz
## 30    215 read SRR25491352.fastq.gz
## 31    212 read SRR25491352.fastq.gz
## 32    201 read SRR25491352.fastq.gz
## 33    200 read SRR25491352.fastq.gz
## 34    190 read SRR25491352.fastq.gz
## 35    186 read SRR25491352.fastq.gz
## 36    175 read SRR25491352.fastq.gz
## 37    166 read SRR25491352.fastq.gz
## 38    159 read SRR25491352.fastq.gz
## 39    159 read SRR25491352.fastq.gz
## 40    150 read SRR25491352.fastq.gz
## 41    149 read SRR25491352.fastq.gz
## 42    148 read SRR25491352.fastq.gz
## 43    146 read SRR25491352.fastq.gz
## 44    135 read SRR25491352.fastq.gz
## 45    130 read SRR25491352.fastq.gz
## 46    128 read SRR25491352.fastq.gz
## 47    128 read SRR25491352.fastq.gz
## 48    127 read SRR25491352.fastq.gz
## 49    127 read SRR25491352.fastq.gz
## 50    123 read SRR25491352.fastq.gz
## 51   6062 read SRR25491357.fastq.gz
## 52    356 read SRR25491357.fastq.gz
## 53    353 read SRR25491357.fastq.gz
## 54    172 read SRR25491357.fastq.gz
## 55    143 read SRR25491357.fastq.gz
## 56    129 read SRR25491357.fastq.gz
## 57    105 read SRR25491357.fastq.gz
## 58    103 read SRR25491357.fastq.gz
## 59     88 read SRR25491357.fastq.gz
## 60     87 read SRR25491357.fastq.gz
## 61     83 read SRR25491357.fastq.gz
## 62     80 read SRR25491357.fastq.gz
## 63     65 read SRR25491357.fastq.gz
## 64     65 read SRR25491357.fastq.gz
## 65     62 read SRR25491357.fastq.gz
## 66     61 read SRR25491357.fastq.gz
## 67     56 read SRR25491357.fastq.gz
## 68     54 read SRR25491357.fastq.gz
## 69     46 read SRR25491357.fastq.gz
## 70     44 read SRR25491357.fastq.gz
## 71     44 read SRR25491357.fastq.gz
## 72     40 read SRR25491357.fastq.gz
## 73     39 read SRR25491357.fastq.gz
## 74     39 read SRR25491357.fastq.gz
## 75     37 read SRR25491357.fastq.gz
```

```
## 76     35 read SRR25491357.fastq.gz
## 77     34 read SRR25491357.fastq.gz
## 78     33 read SRR25491357.fastq.gz
## 79     29 read SRR25491357.fastq.gz
## 80     29 read SRR25491357.fastq.gz
## 81     27 read SRR25491357.fastq.gz
## 82     24 read SRR25491357.fastq.gz
## 83     24 read SRR25491357.fastq.gz
## 84     24 read SRR25491357.fastq.gz
## 85     23 read SRR25491357.fastq.gz
## 86     23 read SRR25491357.fastq.gz
## 87     22 read SRR25491357.fastq.gz
## 88     22 read SRR25491357.fastq.gz
## 89     21 read SRR25491357.fastq.gz
## 90     21 read SRR25491357.fastq.gz
## 91     20 read SRR25491357.fastq.gz
## 92     20 read SRR25491357.fastq.gz
## 93     19 read SRR25491357.fastq.gz
## 94     18 read SRR25491357.fastq.gz
## 95     16 read SRR25491357.fastq.gz
## 96     16 read SRR25491357.fastq.gz
## 97     15 read SRR25491357.fastq.gz
## 98     15 read SRR25491357.fastq.gz
## 99     15 read SRR25491357.fastq.gz
## 100    14 read SRR25491357.fastq.gz
```

Are there any interesting things about your data files that you see?

```
I looked at the frequent sequences that occurred in the runs, and a lot of them were really long strings
```

We are pretty limited by what we can do not together and within the scope of the class, so we wont delve deeper into the other shortread functions. But go ahead and check them out.

```
-   The 'trimtails' function looks pretty useful!
```

There are other packages like shortread including "Rqc".

Rqc is an optimized tool designed for quality control and assessment of high-throughput sequencing data. It performs parallel processing of entire files and produces a report which contains a set of high-resolution graphics.

QUESTION 3 (5 pts)

```
@ Students will gain an understanding of the different types of databases out there, how they differ, ar
```

Pick one **primary database** that you think could be relevant to your research or interests and describe what data you could obtain from there that would help you moving forward.

Now do the same with one **secondary database** type.

Finally, do the same for one **specialized database** type.

Place your responses here:

Primary Database (raw data/info)
- [GenBank](https://www.ncbi.nlm.nih.gov/genbank/): getting raw sequence information, comparing sequenc

Secondary Database (Derived info, analysis from primary info)
-[PATRIC](https://www.bv-brc.org/): This is a bioinformatics resource for bacterial human pathogens. It

Specialized Database
- [Mycobrowser](https://mycobrowser.epfl.ch/): This is a repository for annotated genes in mycobacteria
- [TB Genome Annotation Portal](https://orca2.tamu.edu/U19/): This is similar to Mycobrowser, but this

---

QUESTION 4 (5 pts)

@ Students will learn about a specific R function of their choice, and impliment an example of its usag

Choose an R function you think is interesting and would like to learn more about. Also mention the R package it can be found in.

```
#I am interested in using BiomaRt to retrieve gene information and annotations

if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("biomaRt")
```

```
## Bioconductor version 3.18 (BiocManager 1.30.22), R 4.3.2 (2023-10-31 ucrt)
```

```
## Old packages: 'GenomeInfoDb', 'S4Arrays'
```

Describe what the function does or how it can be used.

This package from BioConductor is a way to interact with gene information databases and get that informa
It is most well documented for use with the [ENSEMBL database](https://useast.ensembl.org/index.html).
It also has pretty good documentation on [BioConductor](https://bioconductor.org/packages/release/bioc/
and the [Github page](https://github.com/grimbough/biomaRt) is up to date.

Give a functional example of its use. I want to try it too.

```
library(biomaRt)
# Specify we are using Ensembl dataset
ensembl <- useMart("ensembl")
#specify which species dataset to use
ensembl <- useDataset("hsapiens_gene_ensembl", mart = ensembl)
#One thing I can do with it is get the sequence of a gene I'm interested. Here I am getting the coding
cbl_sequence <- getSequence(id = "ENSG00000110395", type = "ensembl_gene_id", mart = ensembl, seqType="

cbl_sequence
```

```
##
## 1
## 2                                                                                  ATGGCCGGCAACGTGAAGAAGAGCTCT
## 3
## 4
## 5 ATGGCCGGCAACGTGAAGAAGAGCTCTGGGGCCGGGGGCGGCAGCGGCTCCGGGGGCTCGGGTTCGGGTGGCCTGATTGGGCTCATGAAGGACGCCTTC
## 6
##   ensembl_gene_id
## 1 ENSG00000110395
## 2 ENSG00000110395
## 3 ENSG00000110395
## 4 ENSG00000110395
## 5 ENSG00000110395
## 6 ENSG00000110395
```

Also provide a link to some documentation describing this function. The **NAME** notation is used for creating a hyperlink. You can provide the link below by filling in the syntax: