CMSC 3375 Homework 1

Isabelle Hu

2020-10-08

Download some RNA-seq data from the TCGA head and neckcancer project (TCGA-HNSC), explore the data, and create a histogram of the counts (from at leastone) of the samples.

```
knitr::opts_chunk$set(echo = TRUE)
library(TCGAbiolinks)
library(rmarkdown)
```

```
DataDirectory <- paste0("./GDC/", gsub("-","_","HNSC"))
FileNameData <- paste0(DataDirectory, "_","HTSeq_Counts",".rda")
```

```
# query TCGA-HNSC RNA-seq data
query1 <- GDCquery(project = "TCGA-HNSC",
                   data.category = "Transcriptome Profiling",
                   data.type = "Gene Expression Quantification",
                   workflow.type = "HTSeq - Counts")
```

```
## --------------------------------------
## o GDCquery: Searching in GDC database
## --------------------------------------
## Genome of reference: hg38
## --------------------------------------------
## oo Accessing GDC. This might take a while...
## --------------------------------------------
## ooo Project: TCGA-HNSC
## --------------------
## oo Filtering results
## --------------------
## ooo By data.type
## ooo By workflow.type
## ----------------
## oo Checking data
## ----------------
## ooo Check if there are duplicated cases
## ooo Check if there results for the query
## ------------------
```

```
## o Preparing output
## -------------------
```

```
# get cases for tumor primary (TP)
samplesDown <- getResults(query1, cols=c("cases"))
a.cases.TP <- TCGAquery_SampleTypes(barcode = samplesDown, typesample = "TP")
a.cases.sample <- a.cases.TP[1:20]
```

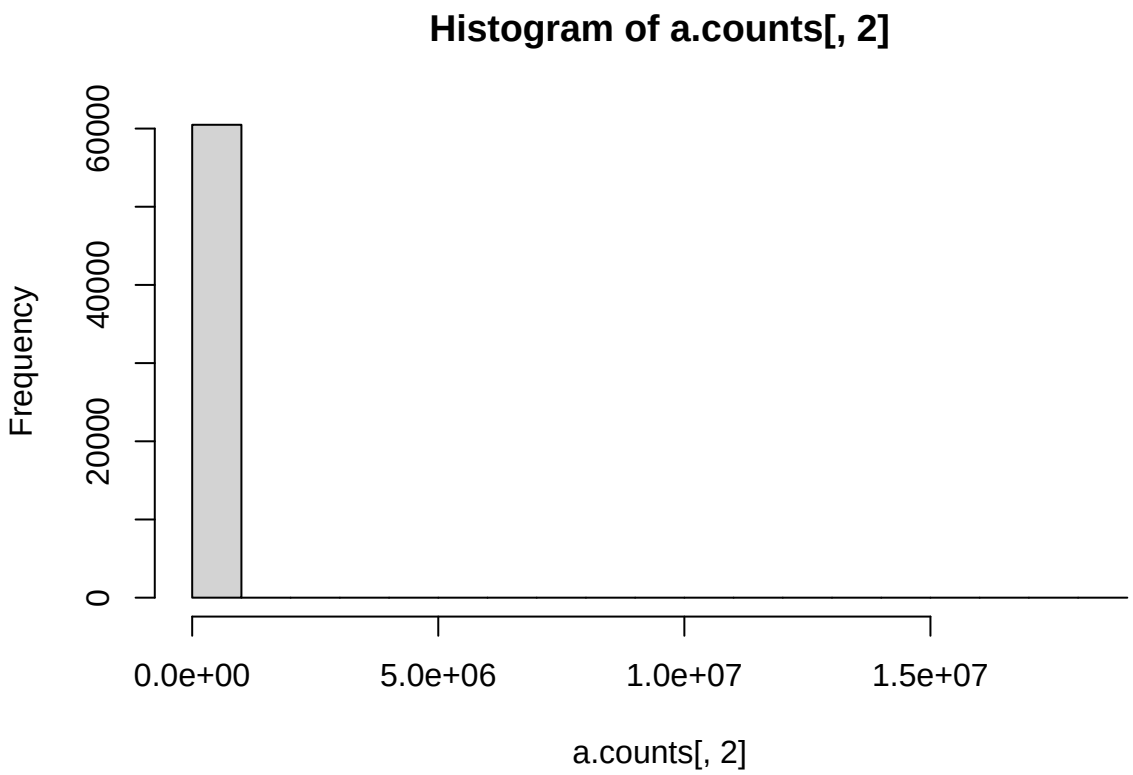
```
# query and download sample data
query2 <- GDCquery(project = "TCGA-HNSC",
                   data.category = "Transcriptome Profiling",
                   data.type = "Gene Expression Quantification",
                   workflow.type = "HTSeq - Counts",
                   barcode = a.cases.sample)
```

```
## ----------------------------------------
## o GDCquery: Searching in GDC database
## ----------------------------------------
## Genome of reference: hg38
## --------------------------------------------
## oo Accessing GDC. This might take a while...
## --------------------------------------------
## ooo Project: TCGA-HNSC
## --------------------
## oo Filtering results
## --------------------
## ooo By data.type
## ooo By workflow.type
## ooo By barcode
## ----------------
## oo Checking data
## ----------------
## ooo Check if there are duplicated cases
## ooo Check if there results for the query
## -------------------
## o Preparing output
## -------------------
```

```
GDCdownload(query = query2, directory = DataDirectory)
```

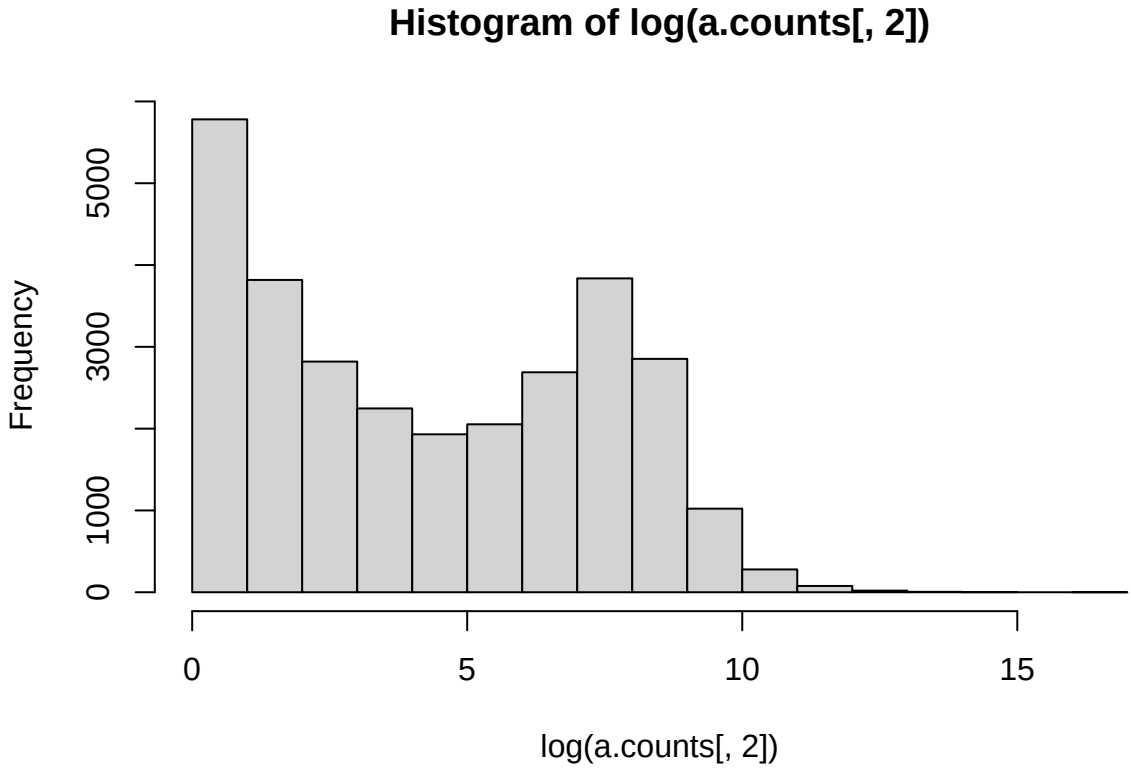
```
## Downloading data for project TCGA-HNSC
## Of the 20 files for download 20 already exist.
## All samples have been already downloaded
```

```r
# pick one patient to load counts data
a.counts <- read.table("./GDC/HNSC/TCGA-HNSC/harmonized/Transcriptome_Profiling/Gene_Expression_Quantifi
# histogram of counts
hist(a.counts[,2])
```



Most genes have zero or very low counts, overwhelming visualization of other gene counts. So I applied log transformation to the counts and replotted the histogram below.

```r
# histogram of ln(counts)
hist(log(a.counts[,2]))
```


Histogram of log(a.counts[, 2])
Frequency
0
1000
3000
5000
0
5
10
15
log(a.counts[, 2])