QUESTION11

Trevor Schaff

November 5, 2015

Reading in the data and formatting the dataset:

```
data_exon <- read.csv("data/Homo_sapiens.GRCh38.82.abinitio_exons.gtf", sep="\t")
data_transcript <- read.csv("data/Homo_sapiens.GRCh38.82.abinitio_transcripts.gtf", sep="\t")
names(data_exon)[1]<-"chromosome"
names(data_exon)[4]<-"start"
names(data_exon)[5]<-"end"
names(data_transcript)[1]<-"chromosome"
names(data_transcript)[4]<-"start"
names(data_transcript)[5]<-"end"
```

Calculating lengths

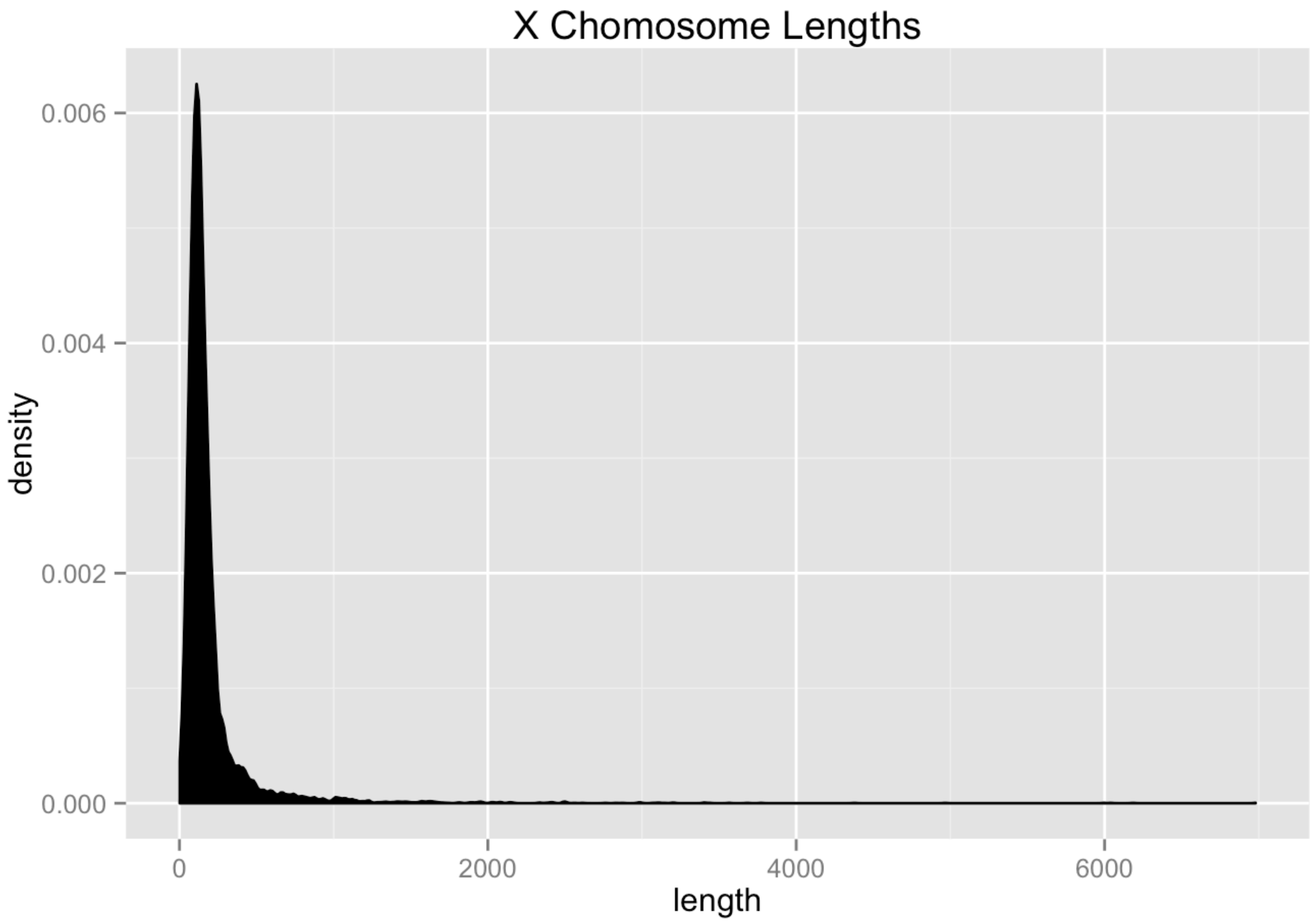
```
data_exon$length <- (data_exon$end - data_exon$start)
data_transcript$length <- (data_transcript$end - data_transcript$start)

#Only keeping chromosomes 1-22 and X and Y
data_exon_clean <- data_exon[data_exon$chromosome == "1" | data_exon$chromosome == "2"| data_exon$chromosome == "3" | data_exon$chromosome == "4" | data_exon$chromosome == "5" | data_exon$chromosome == "6" | data_exon$chromosome == "7" | data_exon$chromosome == "8" | data_exon$chromosome == "9" | data_exon$chromosome == "10" | data_exon$chromosome == "11" | data_exon$chromosome == "12" | data_exon$chromosome == "13" | data_exon$chromosome == "14" | data_exon$chromosome == "15" | data_exon$chromosome == "16" | data_exon$chromosome == "17" | data_exon$chromosome == "18" | data_exon$chromosome == "19" | data_exon$chromosome == "20" | data_exon$chromosome == "21" | data_exon$chromosome == "22" | data_exon$chromosome == "X" | data_exon$chromosome == "Y",]

data_transcript_clean <- data_transcript[data_transcript$chromosome == "1" | data_transcript$chromosome == "2"| data_transcript$chromosome == "3" | data_transcript$chromosome == "4" | data_transcript$chromosome == "5" | data_transcript$chromosome == "6" | data_transcript$chromosome == "7" | data_transcript$chromosome == "8" | data_transcript$chromosome == "9" | data_transcript$chromosome == "10" | data_transcript$chromosome == "11" | data_transcript$chromosome == "12" | data_transcript$chromosome == "13" | data_transcript$chromosome == "14" | data_transcript$chromosome == "15" | data_transcript$chromosome == "16" | data_transcript$chromosome == "17" | data_transcript$chromosome == "18" | data_transcript$chromosome == "19" | data_transcript$chromosome == "20" | data_transcript$chromosome == "21" | data_transcript$chromosome == "22" | data_transcript$chromosome == "X" | data_transcript$chromosome == "Y",]
```

```r
#install.packages("plyr")
library(plyr)
#Standard error function
std <- function(x) sd(x)/sqrt(length(x))

#Calculating statistics
exon_data <- ddply(data_exon_clean,~chromosome,summarise,mean=format(round(mean(length),2)),sd=format(round(sd(length),2)),se=format(round(std(length),2)))
transcript_data <- ddply(data_transcript_clean,~chromosome,summarise,mean=format(round(mean(length),2)),sd=format(round(sd(length),2)),se=format(round(std(length),2)))

names(exon_data)[2]<-"exon_length"
names(exon_data)[3]<-"std_dev_exon"
names(exon_data)[4]<-"std_err_exon"
names(transcript_data)[2]<-"transcript_length"
names(transcript_data)[3]<-"std_dev_transcript"
names(transcript_data)[4]<-"std_err_transcript"

recom<-merge(exon_data, transcript_data, by.x="chromosome", by.y="chromosome")

recom2 <- recom[,c("chromosome", "exon_length", "std_err_exon", "transcript_length",
"std_err_transcript")]

output_table <- as.table(as.matrix(recom2))
output_table
```

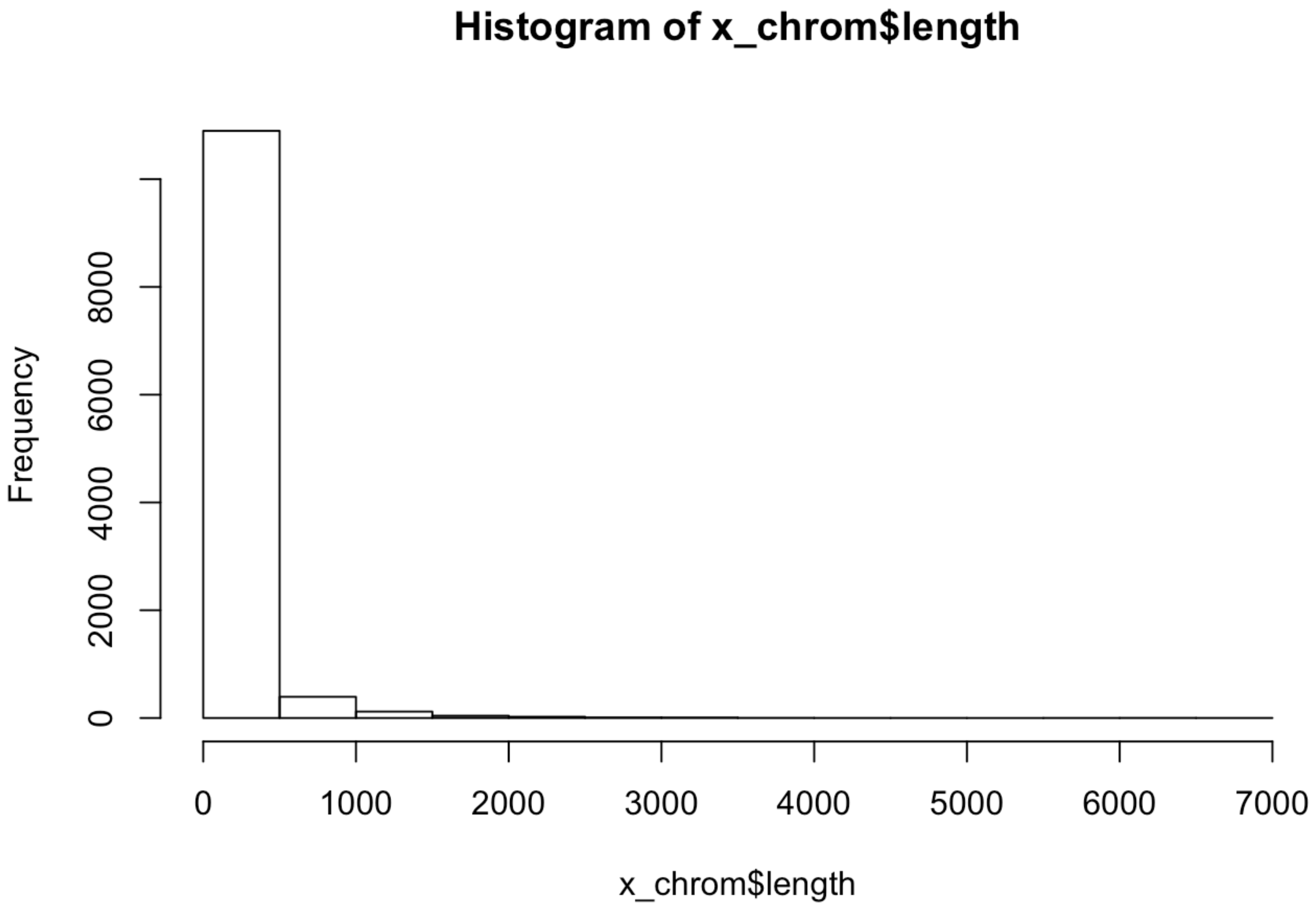
```
##   chromosome exon_length std_err_exon transcript_length std_err_transcript
## A 1          171.81      1.17         38650.07          850.37
## B 10         172.48      1.72         37649.01          1110.79
## C 11         178.05      2.04         37019.33          1203.57
## D 12         166.45      1.51         44291.21          1335.95
## E 13         176.77      3.6          46893             1770.88
## F 14         177.29      2.67         43523.42          1707.72
## G 15         172.85      2.13         40051.73          1336.08
## H 16         172.86      1.87         29383.7           957.3
## I 17         171.88      1.62         30397.38          970.24
## J 18         174.18      2.45         44128.22          1690.48
## K 19         196.75      2.6          22339.4           697.74
## L 2          170.24      1.65         45416             1031.07
## M 20         168.64      2.23         36547.01          1466.03
## N 21         175.69      3.11         39109.13          2187.25
## O 22         176.25      2.92         26710.05          1266.42
## P 3          167.44      1.59         50669.65          1299.04
## Q 4          180.71      2.08         51478.51          1482.67
## R 5          179.45      2.15         48089.7           1304.06
## S 6          175.64      1.76         43271.9           1220.23
## T 7          175.3       1.92         39159.98          1163.91
## U 8          172.36      2.19         44538.81          1317.83
## V 9          174.12      2.01         42599.17          1306.49
## W X          193.39      2.6          49383.01          1804.18
## X Y          191.06      4.48         41082.24          4011.78
```

Making histograms of exon length for X and Y chromosomes:

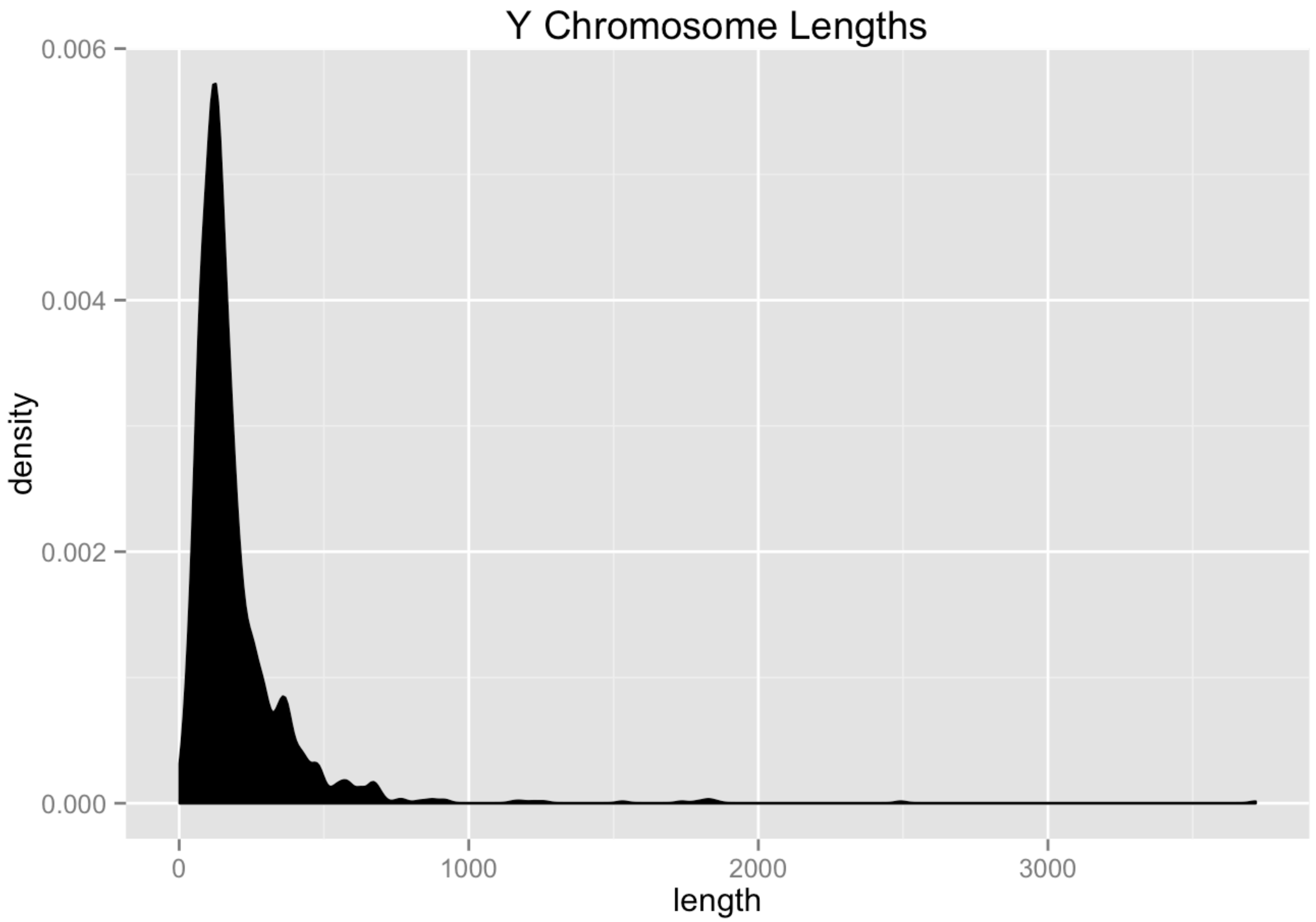
```
#install.packages("ggplot2")
library(ggplot2)
x_chrom <- subset(data_exon_clean, chromosome == "X", c(chromosome, length))

y_chrom <- subset(data_exon_clean, chromosome == "Y", c(chromosome, length))

ggplot(x_chrom) + geom_density(aes(x=length), fill="black") + ggtitle("X Chomosome Le
ngths")
```



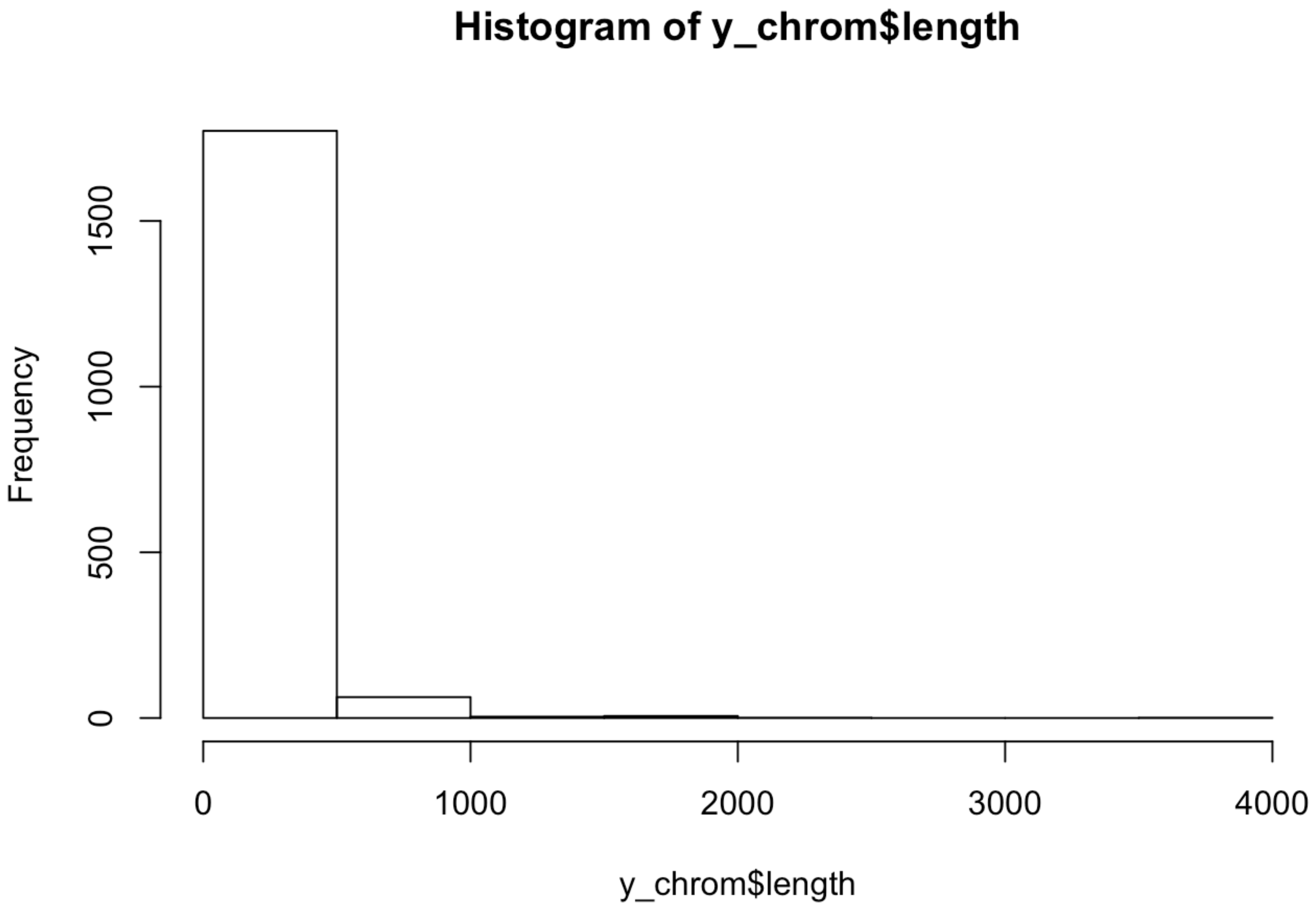
```
hist(x_chrom$length)
```



```
ggplot(y_chrom) + geom_density(aes(x=length), fill="black") + ggtitle("Y Chromosome Lengths")
```



```
hist(y_chrom$length)
```

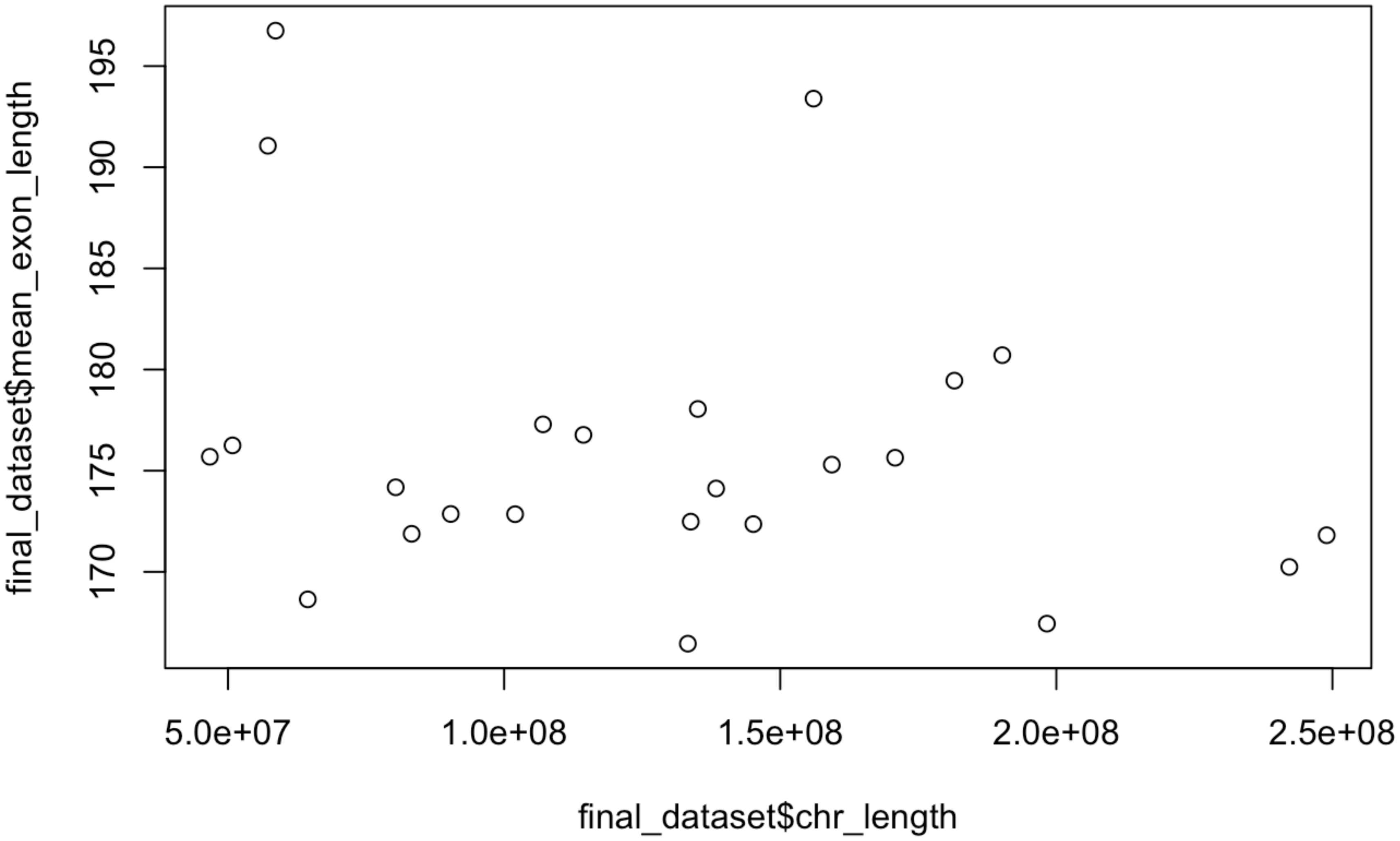


Comparing chromosome length vs mean exon length

```
chrlengths<- matrix(c(248956422,242193529,198295559,190214555,181538259,170805979,159
345973,145138636,138394717,133797422,135086622,133275309,114364328,107043718,10199118
9,90338345,83257441,80373285,58617616,64444167,46709983,50818468,156040895,57227415))

names(chrlengths)[1]<-"chrlengths"

chrcols <- matrix(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,"X","Y")
)

lengths_with_chrs<-transform(chrlengths, new.col = chrcols)
names(lengths_with_chrs)[1]<-"chr_length"
names(lengths_with_chrs)[2]<-"chromosome"
average_exon_lengths <-matrix(c(171.81,170.24,167.44,180.71,179.45,175.64,175.3,172.3
6,174.12,172.48,178.05,166.45,176.77,177.29,172.85,172.86,171.88,174.18,196.75,168.64
,175.69,176.25,193.39,191.06))

final_dataset <-transform(lengths_with_chrs, new.col = average_exon_lengths)
names(final_dataset)[1]<-"chr_length"
names(final_dataset)[3]<-"mean_exon_length"

cor(final_dataset$chr_length, final_dataset$mean_exon_length)
```

```
## [1] -0.2544491
```

```
plot(final_dataset$chr_length, final_dataset$mean_exon_length)
```



There is not a significant relationship between the length of a chromosome and its corresponding mean exon length. The correlation coefficient is -0.25.