EEB416_Exam_Question_11

Tom McBrien

November 4, 2015

Loading Data

```
exons <- read.delim("Homo_sapiens.GRCh38.82.abinitio.exons.gtf", header = FALSE)
transcripts <- read.delim("Homo_sapiens.GRCh38.82.abinitio.transcripts.gtf", header = FALSE)
```

Adding Column Names

```
names(exons) <- c("Chromosome", "Source", "Type", "Start", "End")
names(transcripts) <- c("Chromosome", "Source", "Type", "Start", "End")
```

Making a Length Column

```
exons$length <- (exons$End - exons$Start)
transcripts$length <- (transcripts$End - transcripts$Start)
```

Checking Progress So Far

```
head(exons)
```

```
##   Chromosome  Source Type Start   End NA NA NA NA length
## 1          1 ensembl exon 12190 12227  .  +  . NA     37
## 2          1 ensembl exon 12613 12721  .  +  . NA    108
## 3          1 ensembl exon 14051 14149  .  +  . NA     98
## 4          1 ensembl exon 24738 24886  .  -  . NA    148
## 5          1 ensembl exon 18501 18554  .  -  . NA     53
## 6          1 ensembl exon 18268 18379  .  -  . NA    111
```

```
head(transcripts)
```

```
##   Chromosome  Source       Type  Start    End NA NA NA NA length
## 1          1 ensembl transcript  12190  14149  .  +  . NA   1959
## 2          1 ensembl transcript  14696  24886  .  -  . NA  10190
## 3          1 ensembl transcript  51913 106974  .  +  . NA  55061
## 4          1 ensembl transcript 131158 134833  .  +  . NA   3675
## 5          1 ensembl transcript 171728 184662  .  -  . NA  12934
## 6          1 ensembl transcript 185217 195411  .  -  . NA  10194
```

Looking at Mean Values by Chromosome

```
exon_means <- aggregate(exons[ ,10], list(exons$Chromosome), mean)
transcript_means <- aggregate(transcripts[ ,10], list(transcripts$Chromosome), mean)
```

Progress Check

```
head(exon_means)
```

```
##   Group.1        x
## 1       1 171.8068
## 2      10 172.4843
## 3      11 178.0485
## 4      12 166.4534
## 5      13 176.7690
## 6      14 177.2929
```

```
head(transcript_means)
```

```
##   Group.1        x
## 1       1 38640.75
## 2      10 37649.01
## 3      11 37019.33
## 4      12 44291.21
## 5      13 46893.00
## 6      14 43523.42
```

Making a Standard Error Equation

```
std.err <- function(x) {sd(x)/sqrt(length(x))}
```

Calculating Standard Error

```
exon_SE <- aggregate(exons[ ,10], list(exons$Chromosome), std.err)
transcripts_SE <- aggregate(transcripts[ ,10], list(transcripts$Chromosome), std.err)
```

Combining and Cleaning

```
exons_merged <- merge(exon_means, exon_SE, by.x = "Group.1",
    by.y = "Group.1")
transcripts_merged <- merge(transcript_means, transcripts_SE,
    by.x = "Group.1", by.y = "Group.1")
both_merged <- merge(exons_merged, transcripts_merged, by.x = "Group.1",
```

```
    by.y = "Group.1")
names(both_merged) <- c("chromosome", "exon_length", "standard error",
    "transcript length", "standard error")
both_merged_chromosomes_only <- both_merged[-23:-383, ]  #getting rid of weird non-chromosome rows
```

Making Table

```
length_means_table <- as.table(as.matrix(both_merged_chromosomes_only))
length_means_table
```

```
##     chromosome exon_length standard error transcript length standard error
## 1   1          171.8068    1.167457       38640.75           850.2083
## 2   10         172.4843    1.719529       37649.01          1110.7922
## 3   11         178.0485    2.043632       37019.33          1203.5705
## 4   12         166.4534    1.506568       44291.21          1335.9496
## 5   13         176.7690    3.595995       46893.00          1770.8841
## 6   14         177.2929    2.670241       43523.42          1707.7169
## 7   15         172.8530    2.128271       40051.73          1336.0761
## 8   16         172.8620    1.868135       29383.70           957.2974
## 9   17         171.8774    1.624757       30397.38           970.2432
## 10  18         174.1781    2.454048       44128.22          1690.4753
## 11  19         196.7540    2.598789       22339.40           697.7406
## 12  2          170.2351    1.650216       45416.00          1031.0711
## 13  20         168.6432    2.233778       36547.01          1466.0287
## 14  21         175.6853    3.111300       39109.13          2187.2548
## 15  22         176.2492    2.924695       26710.05          1266.4161
## 16  3          167.4443    1.585111       50669.65          1299.0419
## 17  4          180.7103    2.076310       51478.51          1482.6656
## 18  5          179.4528    2.148415       48089.70          1304.0569
## 19  6          175.6421    1.756959       43271.90          1220.2318
## 20  7          175.2980    1.922756       39159.98          1163.9141
## 21  8          172.3565    2.188355       44538.81          1317.8275
## 22  9          174.1220    2.012749       42599.17          1306.4864
## 384 X          193.3924    2.600181       49383.01          1804.1818
## 385 Y          191.0617    4.477689       41082.24          4011.7817
```

Subsetting for X and Y Chromosome Lengths

```
X_exon_lengths <- subset(exons, exons$Chromosome == "X")
Y_exon_lengths <- subset(exons, exons$Chromosome == "Y")
X_Y_exon_lengths <- subset(exons, exons$Chromosome == "X" |
    exons$Chromosome == "Y")
# subsetting to get rid of non-sex chromosomes
X_Y_exon_lengths_adjusted <- subset(X_Y_exon_lengths, X_Y_exon_lengths$length <
    2500)
# plot was skewed by large lengths that made up just
# 30/13349, or about 0.2% of sequences, so I took them
# out.  To see plot with these sequences, just
```

```
# substitute X_Y_exon_lengths into ggplot equation.
head(X_Y_exon_lengths)  #checking to make sure 'X' chromosome was taken
```

```
##        Chromosome  Source Type  Start    End NA NA.1 NA.2 NA.3 length
## 136820          X ensembl exon  11292  11409  .    +     .   NA    117
## 136821          X ensembl exon  13940  14129  .    +     .   NA    189
## 136822          X ensembl exon  17035  17156  .    +     .   NA    121
## 136823          X ensembl exon  18773  18965  .    +     .   NA    192
## 136824          X ensembl exon 229749 229870  .    +     .   NA    121
## 136825          X ensembl exon 253743 253851  .    +     .   NA    108
```

```
tail(X_Y_exon_lengths)  #checking for 'Y'
```

```
##        Chromosome  Source Type    Start      End NA NA.1 NA.2 NA.3 length
## 315395          Y ensembl exon 56769937 56770302  .    +     .   NA    365
## 315396          Y ensembl exon 56827497 56827796  .    -     .   NA    299
## 315397          Y ensembl exon 56826457 56826885  .    -     .   NA    428
## 315398          Y ensembl exon 56828687 56829061  .    -     .   NA    374
## 315399          Y ensembl exon 56834270 56834620  .    -     .   NA    350
## 315400          Y ensembl exon 56872387 56872556  .    -     .   NA    169
```

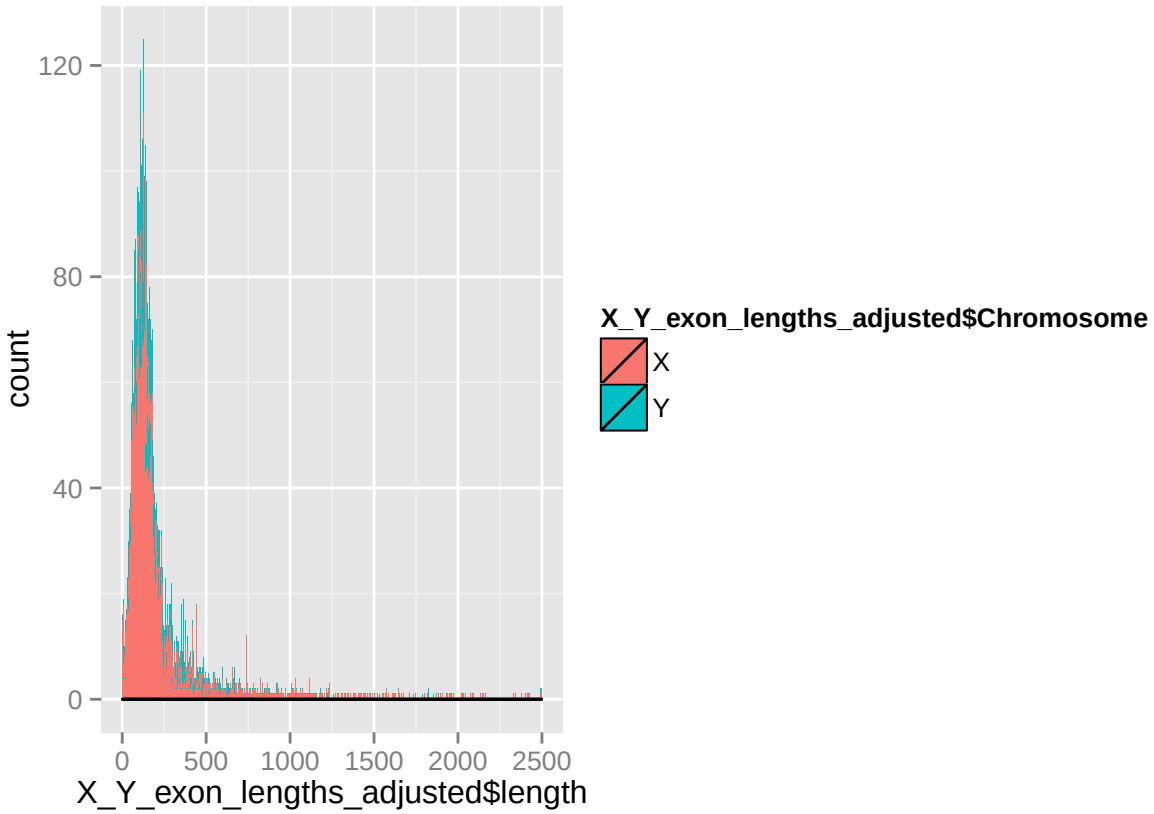
Frequency Histogram Of Exon Lengths

```
install.packages("ggplot2", repos = "http://cran.rstudio.com/")
```

```
## Installing package into '/Users/tmcbrien/Library/R/3.2/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/fl/6y5yynqs4p3fk90y_kj3mh9000y5rj/T//RtmpIkGklV/downloaded_packages
```

```
library(ggplot2)
ggplot(X_Y_exon_lengths_adjusted, aes(X_Y_exon_lengths_adjusted$length,
    fill = X_Y_exon_lengths_adjusted$Chromosome)) + geom_histogram(binwidth = 1) +
    geom_density(alpha = 0.1)
```



Loading .bed File and Manipulating To Compute Chromosome Lengths

```
exons_and_transcripts <- read.delim("Homo_sapiens.GRCh38.82.abinitio.bed",
    header = FALSE)
names(exons_and_transcripts) <- c("Chromosome", "Type", "Start",
    "End")
start_position <- aggregate(exons_and_transcripts$Start, by = list(exons_and_transcripts$Chromosome),
    min)  #finding smallest value in start for each chromosome
end_position <- aggregate(exons_and_transcripts$Start, by = list(exons_and_transcripts$Chromosome),
    max)
start_and_end <- merge(start_position, end_position, by.x = "Group.1",
    by.y = "Group.1")
names(start_and_end) <- c("chromosome", "start", "end")
start_and_end$chromosomelength <- (start_and_end$end - start_and_end$start)  #making new row of chromos
start_and_end_chromosomes_only <- start_and_end[-23:-383, ]  #getting rid of weird non-chromosome rows
head(start_and_end_chromosomes_only)  #checking
```

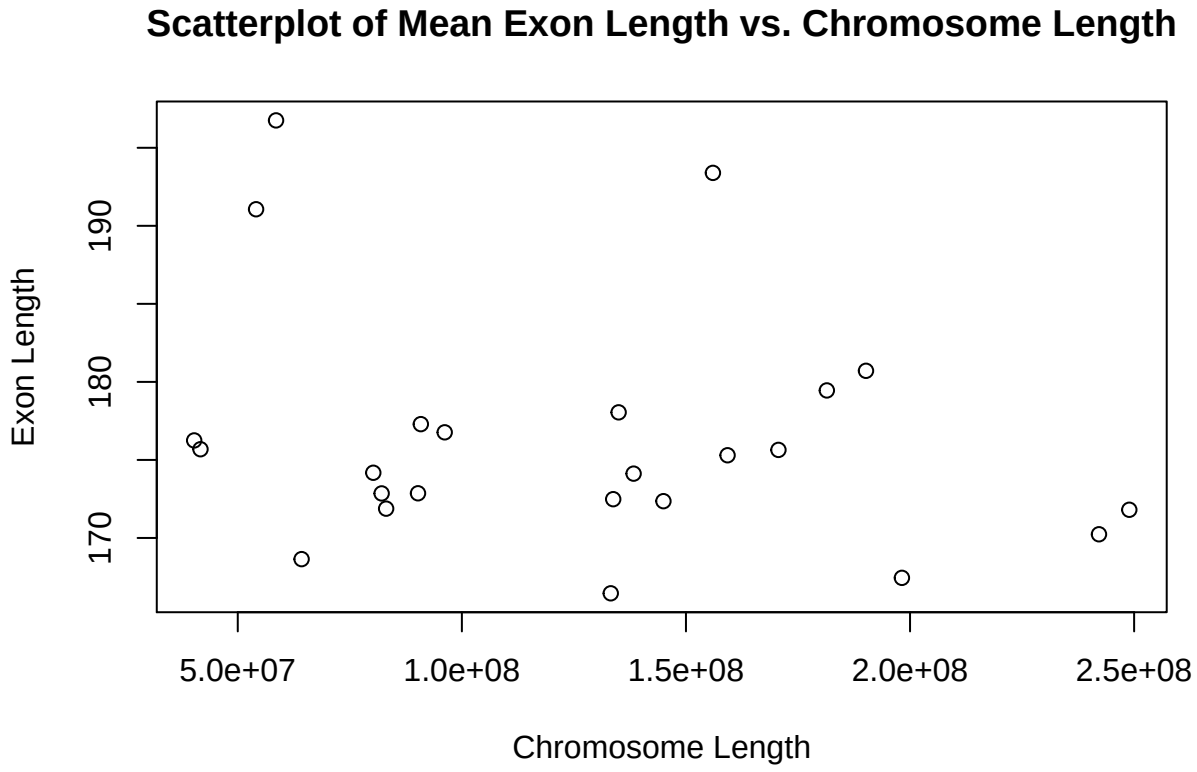
```
##   chromosome    start       end chromosomelength
## 1          1    12189 248936621        248924432
## 2         10    11839 133778493        133766654
## 3         11    61991 135041534        134979543
## 4         12    12739 133235851        133223112
## 5         13 18174009 114346290         96172281
## 6         14 16030433 106874971         90844538
```

Combining Exon Means and Chromosome Lengths to Compare

```
exon_means_for_merge <- exon_means[-23:-383, ]
names(exon_means_for_merge) <- c("chromosome", "mean_exon_length")
exon_means_chromosome_lengths <- merge(start_and_end_chromosomes_only,
    exon_means_for_merge, by.x = "chromosome", by.y = "chromosome")
```

Plotting and Comparing Chromosome Length and Exon Length

```
plot(exon_means_chromosome_lengths$chromosomelength, exon_means_chromosome_lengths$mean_exon_length,
    main = "Scatterplot of Mean Exon Length vs. Chromosome Length",
    xlab = "Chromosome Length", ylab = "Exon Length")
abline(lm(exon_means_chromosome_lengths$chromosomelength ~
    exon_means_chromosome_lengths$mean_exon_length), col = "blue")  #adding trendline
```




```
cor(exon_means_chromosome_lengths$chromosomelength, exon_means_chromosome_lengths$mean_exon_length)
```

```
## [1] -0.2443966
```

Because the correlation coefficient is -0.2443966, there does not appear to be a significant relationship.