Lab 15

Q1. What does the star character accomplish here? Ask Barry, or your class neighbor, if you are not sure!

Q1A. The star character allows us to select all files that contain ".faa.gz."

Q2. How many sequences are in this mouse.1.protein.faa file? Hint: Try using grep to figure this out…

Q2A. There are 641 proteins.

Q3. What happens if you run the above command without the > mm-first.fa part?

Q3A. It prints out what head mouse.1.protein.faa printed out along with ">XP_017169522.1 cation channel sperm-associated protein subunit epsilon isoform X4 [Mus musculus]."

Q4. What happens if you were to use two '>' symbols (i.e. » mm-first.fa)?

Q4A. The output of "head -11 mouse.1.protein.faa" is appended to "mm-first.fa."

Q5.How would you determine how many sequences are in the mm-second.fa file?

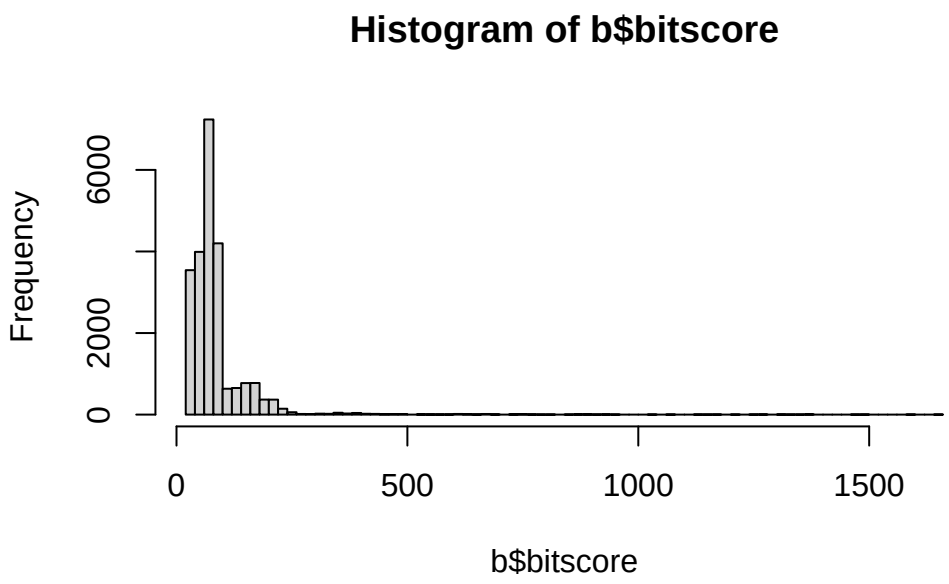
Q5A. I would type "grep mm-second.fa"

```
library(readr)
b <- read_tsv("mm-second.x.zebrafish.tsv", col_names=FALSE)
```

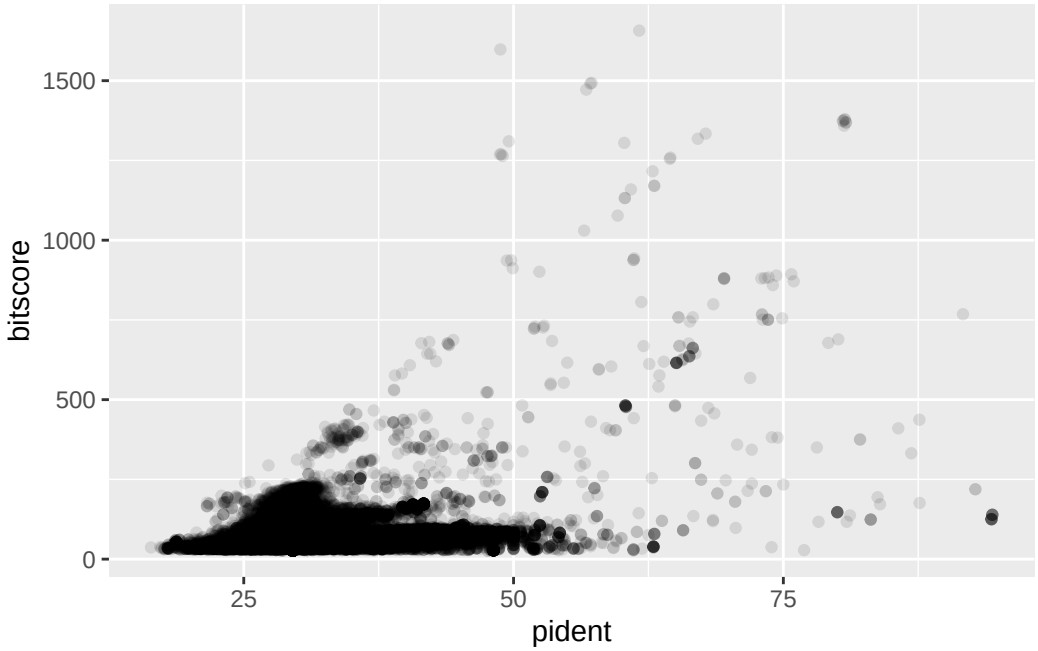
```
Rows: 23118 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: "\t"
chr  (2): X1, X2
dbl (10): X3, X4, X5, X6, X7, X8, X9, X10, X11, X12

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
names(b) <- c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart", "qe
View(b)
```

```
hist(b$bitscore, breaks=100)
```

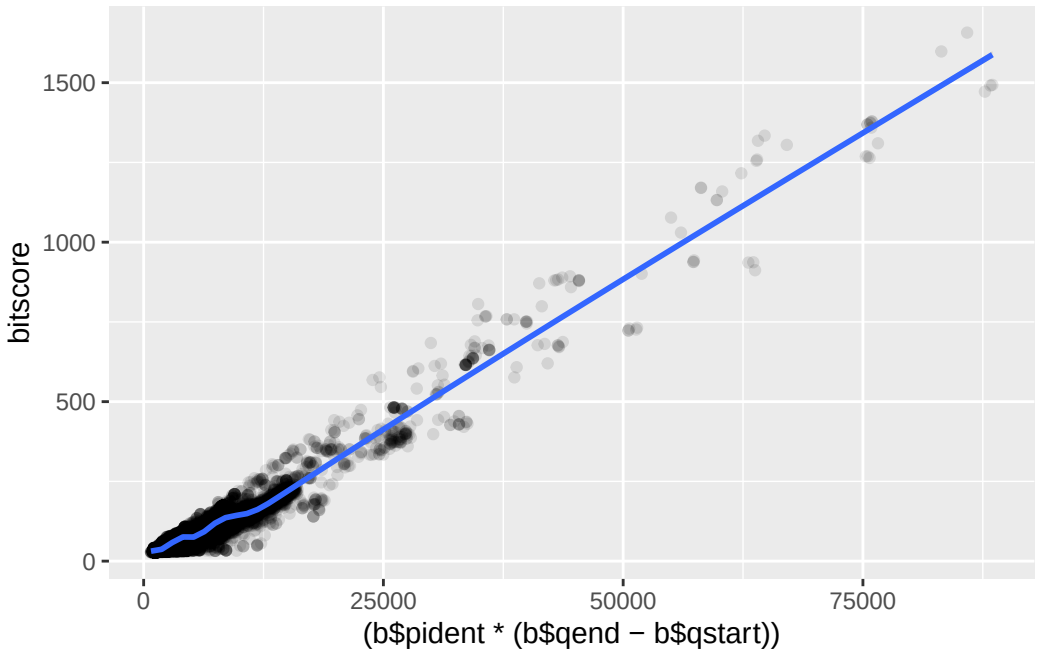



```
library(ggplot2)
ggplot(b, aes(pident, bitscore)) + geom_point(alpha=0.1)
```




```
ggplot(b, aes((b$pident * (b$qend - b$qstart)), bitscore)) + geom_point(alpha=0.1) + geom_
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Q6. Note the addition of the -r option here: What is it's purpose? Also what about the *, what is it's purpose here?

Q6A. The purpose of the -r option is to combine our local machine and virtual machine. The purpose of "*" is to indicate where to save our work on the virtual machine.