Class 13 lab report

Bangyan Hu (PID: A15540189)

11/9/2021

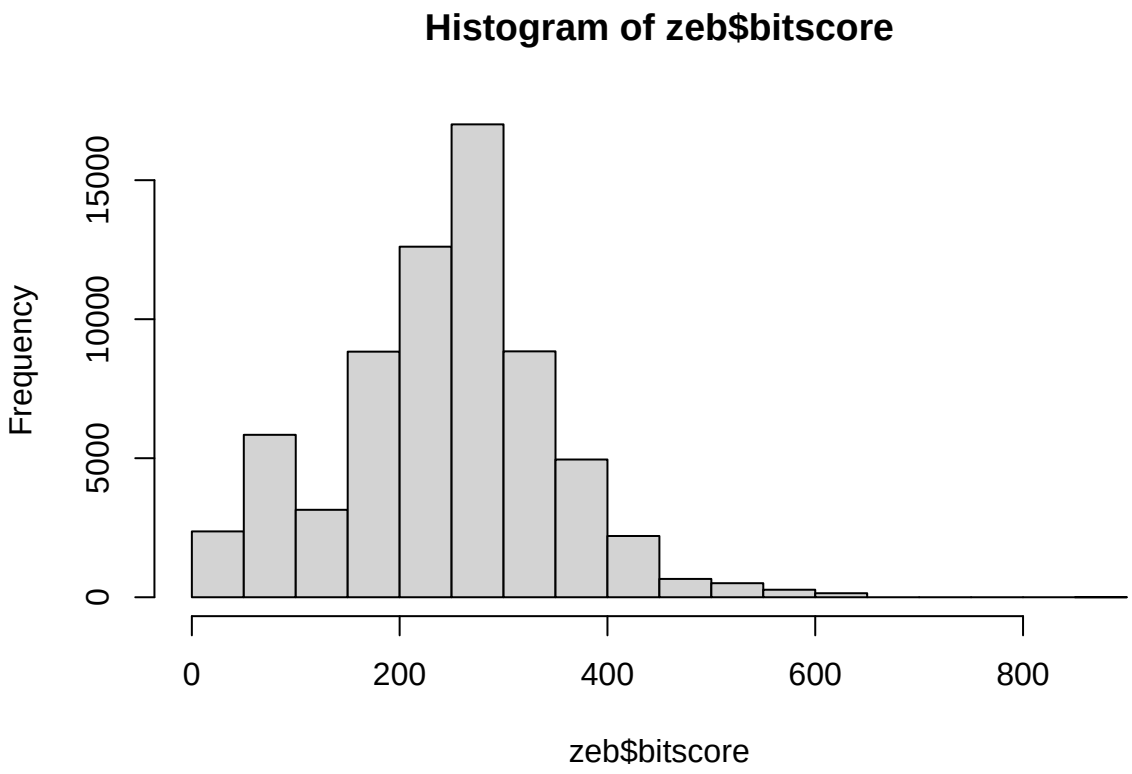
Read my mm-second.x.zebrafish.tsv.

```
zeb <- read.delim(file="mm-second.x.zebrafish.tsv")
colnames(zeb) = c("qseqid", "sseqid", "pident", "length",
                  "mismatch", "gapopen", "qstart", "qend",
                  "sstart", "send", "evalue", "bitscore")
head(zeb)
```

```
##        qseqid         sseqid pident length mismatch gapopen qstart qend sstart
## 1 YP_220551.1    NP_059332.1 44.509    346      188       3      1  344      1
## 2 YP_220551.1    NP_059341.1 24.540    163      112       3    112  263    231
## 3 YP_220551.1    NP_059340.1 26.804     97       65       2     98  188    200
## 4 YP_220552.1    NP_059333.1 88.132    514       61       0      1  514      1
## 5 YP_220552.1 XP_021326074.1 31.818     66       32       2    427  482     16
## 6 YP_220552.1 NP_001373511.1 31.818     66       32       2    427  482     48
##   send   evalue bitscore
## 1  344 8.62e-92    279.0
## 2  393 5.14e-06     49.7
## 3  296 1.00e-01     35.8
## 4  514 0.00e+00    877.0
## 5   78 6.70e+00     29.3
## 6  110 7.50e+00     29.6
```

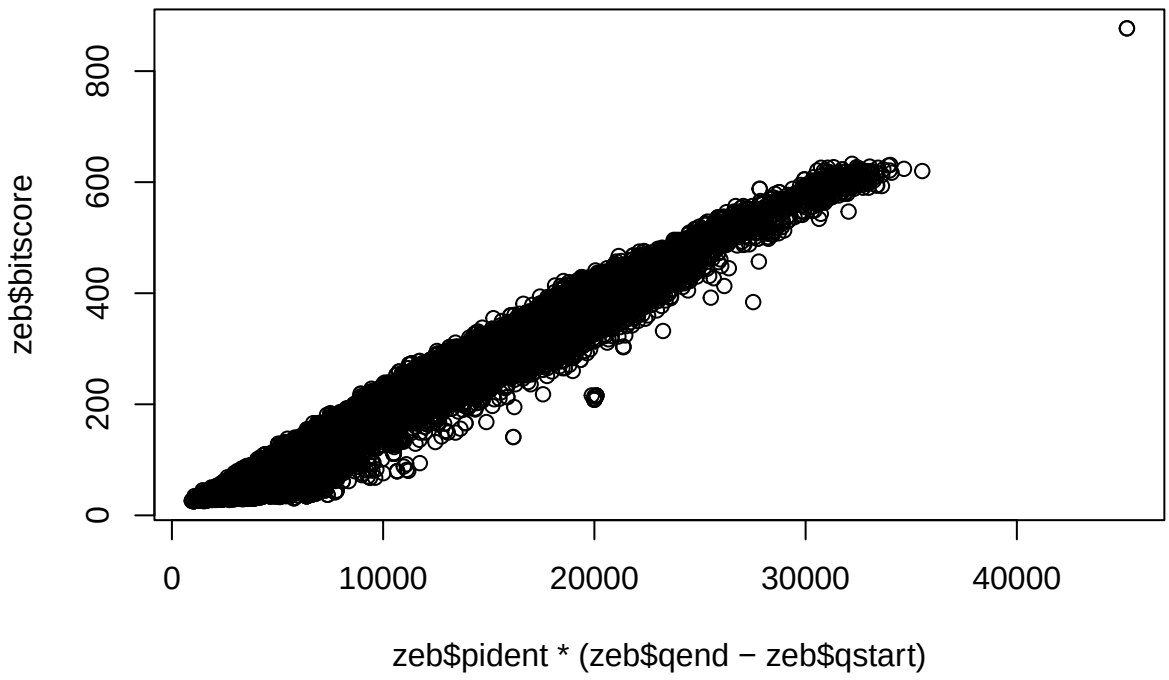
Make a histogram of the $bitscore values.

```
hist(zeb$bitscore, breaks=30)
```



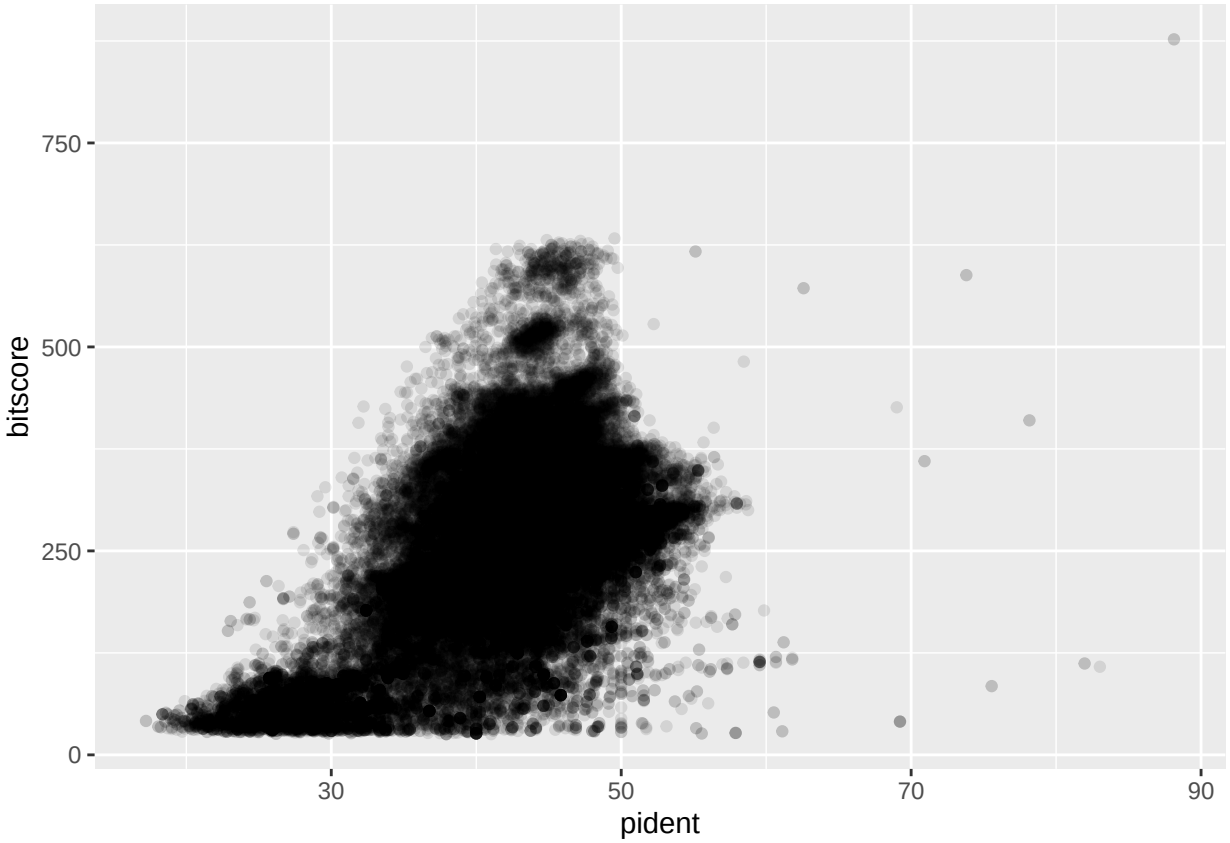
Bitscores are only somewhat related to pident; they take into account not only the percent identity but the length of the alignment.

```
## Asuming your blast results are stored in an object called 'b'
plot(zeb$pident  * (zeb$qend - zeb$qstart), zeb$bitscore)
```



Or using ggplot

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

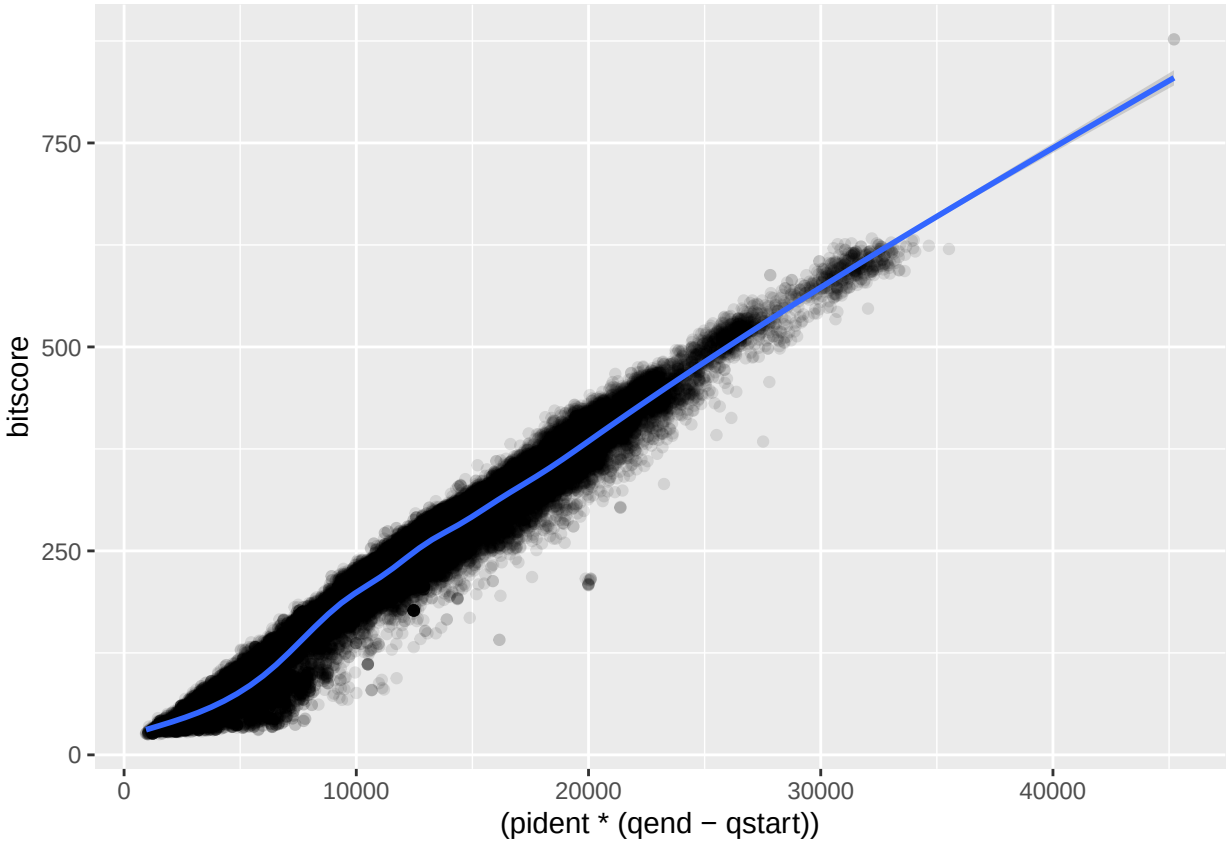


From this graph, we can see that bitscores are only somewhat related to pident.

Take into account not only the percent identity but the length of the alignment.

```
ggplot(zeb, aes((pident * (qend - qstart)), bitscore)) +
  geom_point(alpha=0.1) + geom_smooth()
```

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



There is about a straightforward relationship between bitscore and (percent identity x length).