Lab14

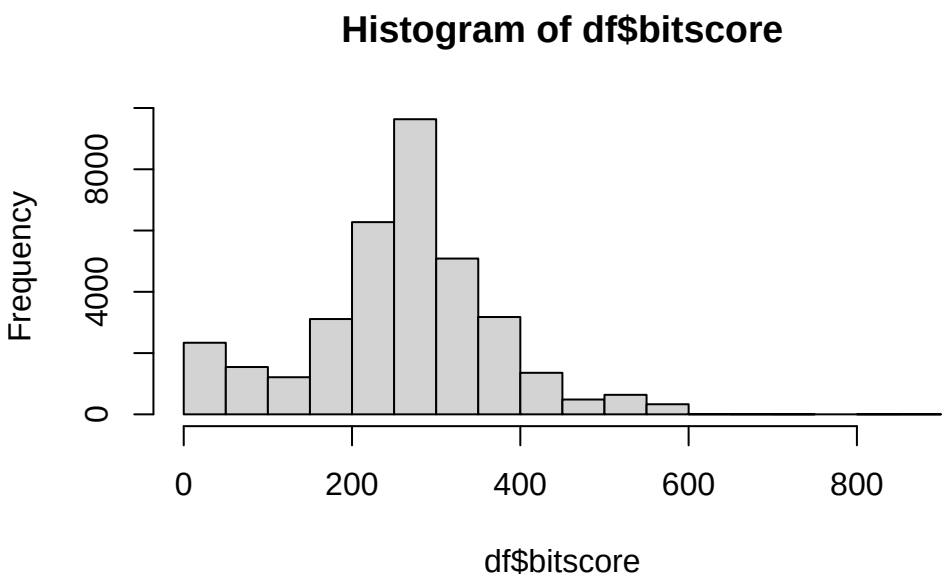
Runqi Zhang

Table of contents

```
# read tsv file using read.table()
df <- read.table("mm-second.x.zebrafish.tsv")

# set column names
colnames(df) <- c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart",
```

Make a histogram of the $bitscore values. You may want to set the optional breaks to be a larger number (e.g. breaks=30).

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

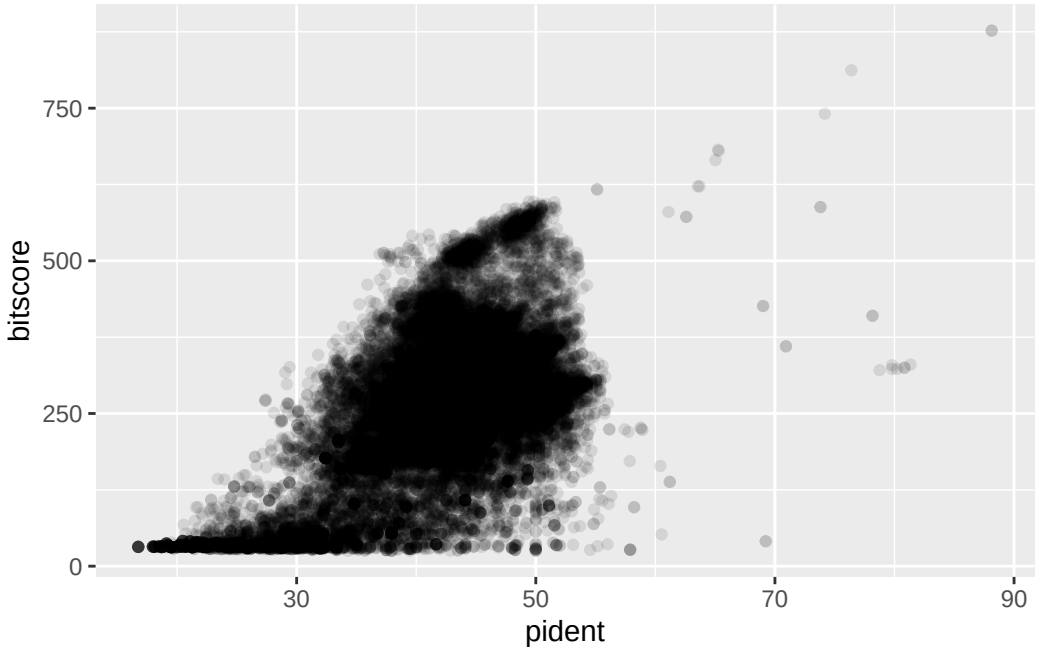


Is there a straightforward relationship between percent identity $(pident) and bitscore$(bitscore) for the alignments we generated?

```
cor(df$pident, df$bitscore)
```

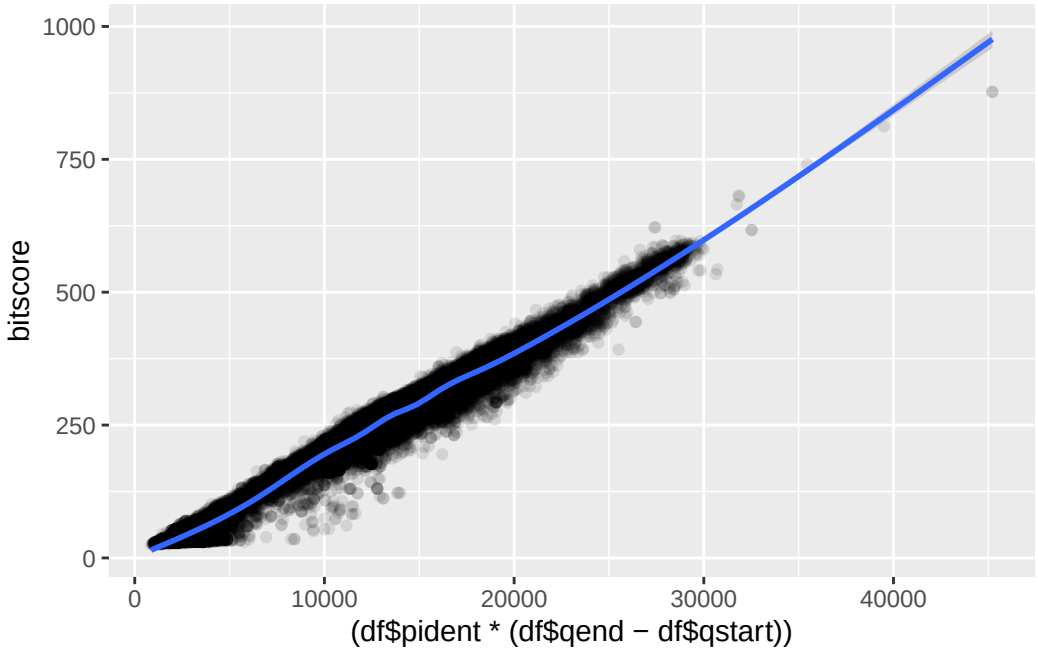
```
[1] 0.5815839
```

```
# create a ggplot
library(ggplot2)
ggplot(df, aes(pident, bitscore))+
    geom_point(alpha=0.1)
```




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

```
Warning: Use of `df$pident` is discouraged. Use `pident` instead.

Warning: Use of `df$qend` is discouraged. Use `qend` instead.

Warning: Use of `df$qstart` is discouraged. Use `qstart` instead.

Warning: Use of `df$pident` is discouraged. Use `pident` instead.

Warning: Use of `df$qend` is discouraged. Use `qend` instead.

Warning: Use of `df$qstart` is discouraged. Use `qstart` instead.

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


1000
750
500
250
0
bitscore
0
10000
20000
30000
40000
(df$pident * (df$qend − df$qstart))