

Essential UNIX for Bioinformatics

AUTHOR

Nundini Varshney (PID: A16867985)

```
library(readr)
```

Warning: package 'readr' was built under R version 4.3.3

```
b <- read_tsv("mm-second.x.zebrafish.tsv")
```

Rows: 28788 Columns: 12

— Column specification —————

Delimiter: "\t"


chr (2): NP_598866.1, XP_009294521.1

dbl (10): 46.154, 273, 130, 6, 4, 267, 420, 684, 1.70e-63, 214

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.

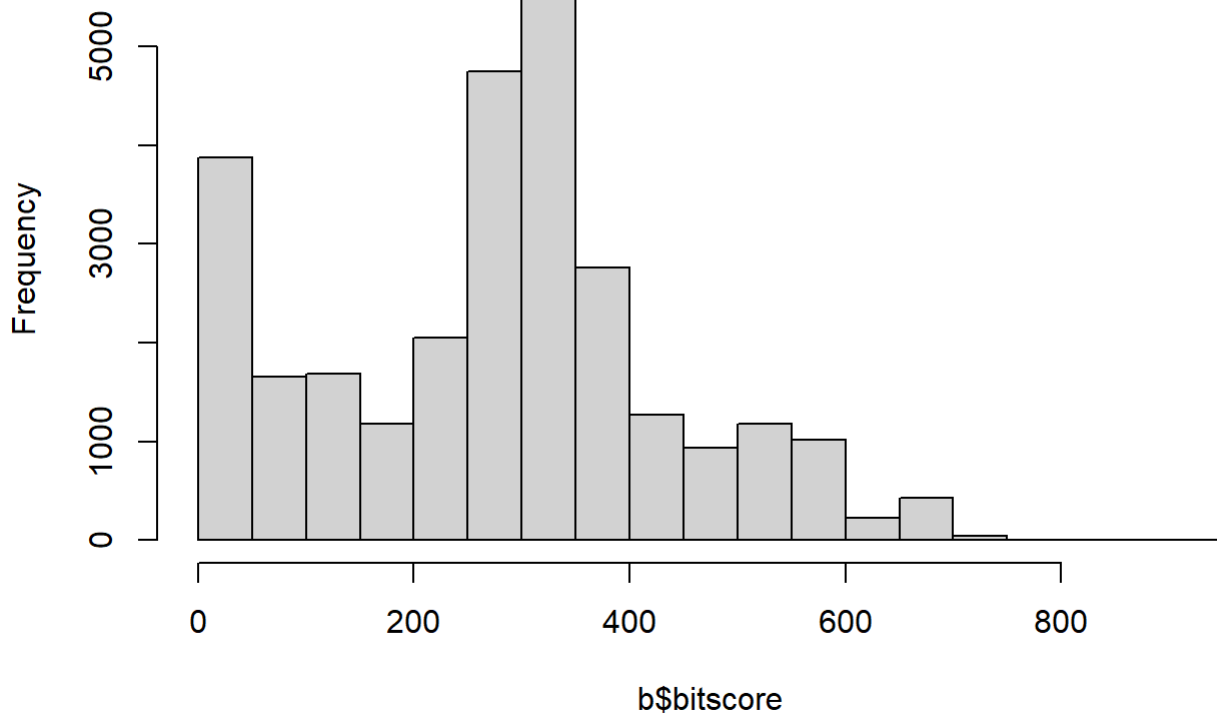
```
colnames(b) <- c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart", "qend",
```



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

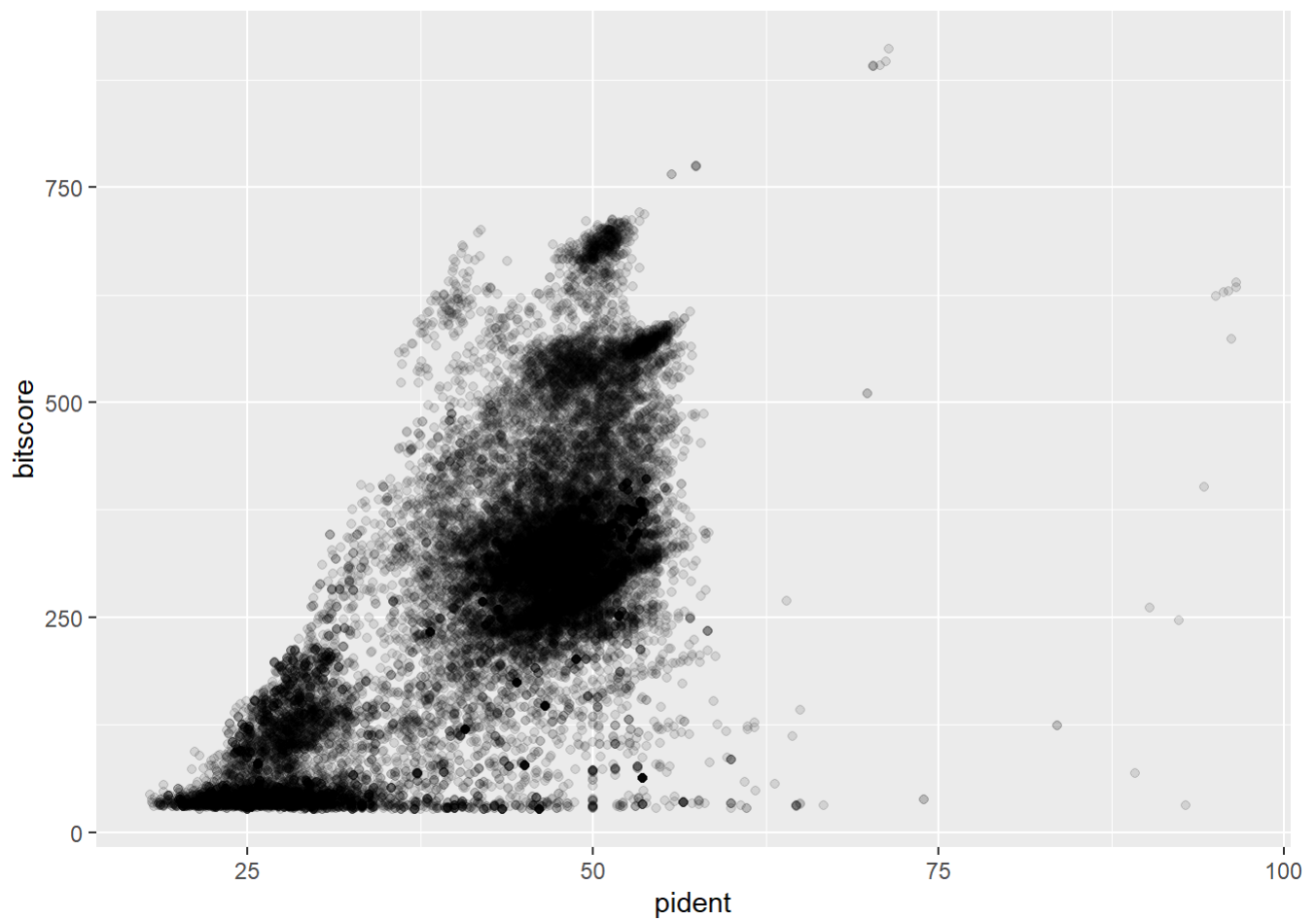
```
hist(b$bitscore)
```

Histogram of b\$bitscore



Plot of percent identity vs bitscore:

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



Plot of percent identity * length vs bitscore:

```
library(ggplot2)
ggplot(b, aes((b$pident * (b$qend - b$qstart)), bitscore)) + geom_point(alpha=0.1) + geom_smooth(
```

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

