Class 16

Jiayi Zhou (PID:A17856751)

2025-11-20

Read the BLAST TSV and set column names

```
b <- read.table("mm-second.x.zebrafish.tsv",
                sep = "\t",
                header = FALSE,
                stringsAsFactors = FALSE)
```

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

```
head(b)
```

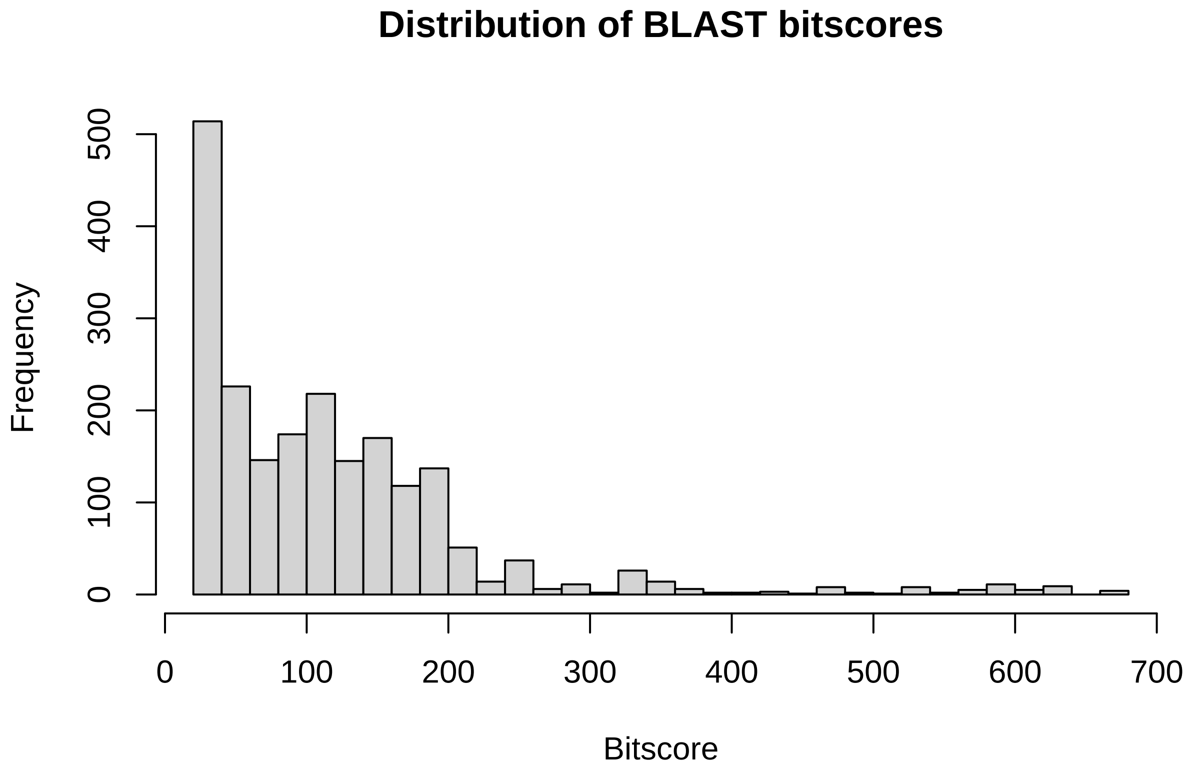
```
##        qseqid         sseqid pident length mismatch gapopen qstart qend sstart
## 1 NP_034603.2 XP_002663941.1 44.444     90       46       1    644  733    295
## 2 NP_034603.2 XP_021335885.1 41.667     96       46       2    644  733    295
## 3 NP_034603.2 XP_021329952.2 43.243     74       41       1    660  733    217
## 4 NP_036084.2 XP_073799717.1 28.230    209      140       6     17  220      1
## 5 NP_036084.2 NP_001156326.1 25.854    205      143       3     27  224     24
## 6 NP_036084.2 XP_073785644.1 24.265    136       97       1     87  216     18
##   send   evalue bitscore
## 1  380 6.07e-20     94.4
## 2  386 4.43e-18     88.6
## 3  289 2.21e-09     61.2
## 4  204 6.60e-10     58.5
## 5  226 4.99e-09     56.2
## 6  153 2.22e-07     50.4
```

```
str(b)
```

```
## 'data.frame':    2078 obs. of  12 variables:
##  $ qseqid  : chr  "NP_034603.2" "NP_034603.2" "NP_034603.2" "NP_036084.2" ...
##  $ sseqid  : chr  "XP_002663941.1" "XP_021335885.1" "XP_021329952.2" "XP_073799717.1" ...
##  $ pident  : num  44.4 41.7 43.2 28.2 25.9 ...
##  $ length  : int  90 96 74 209 205 136 143 141 141 141 ...
##  $ mismatch: int  46 46 41 140 143 97 98 99 99 99 ...
##  $ gapopen : int  1 2 1 6 3 1 3 3 3 3 ...
##  $ qstart  : int  644 644 660 17 27 87 87 87 87 87 ...
##  $ qend    : int  733 733 733 220 224 216 224 224 224 224 ...
##  $ sstart  : int  295 295 217 1 24 18 112 26 26 26 ...
##  $ send    : int  380 386 289 204 226 153 252 163 163 163 ...
##  $ evalue  : num  6.07e-20 4.43e-18 2.21e-09 6.60e-10 4.99e-09 ...
##  $ bitscore: num  94.4 88.6 61.2 58.5 56.2 50.4 48.5 47 47 47 ...
```

Make a histogram of bitscore

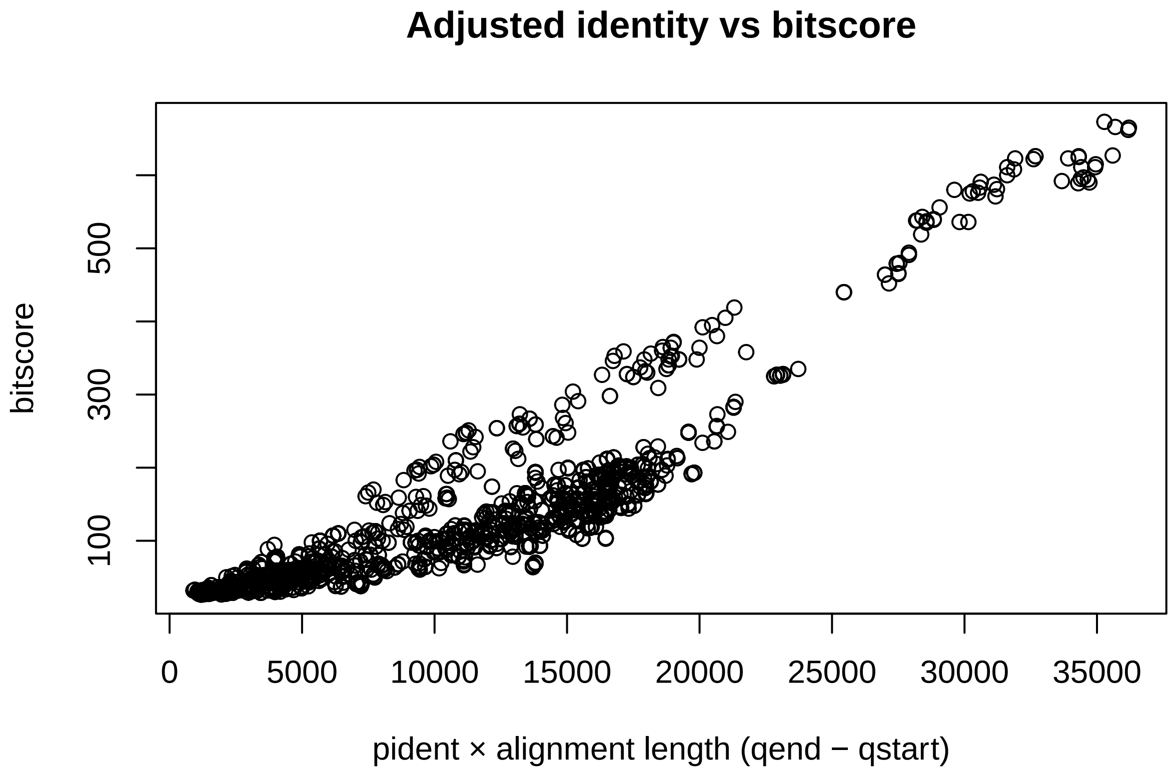
```
hist(b$bitscore,
     breaks = 30,
     main = "Distribution of BLAST bitscores",
     xlab = "Bitscore")
```



Larger bitscores are better.

```
adj_identity <- b$pident * (b$qend - b$qstart)

plot(adj_identity, b$bitscore,
     xlab = "pident × alignment length (qend - qstart)",
     ylab = "bitscore",
     main = "Adjusted identity vs bitscore")
```

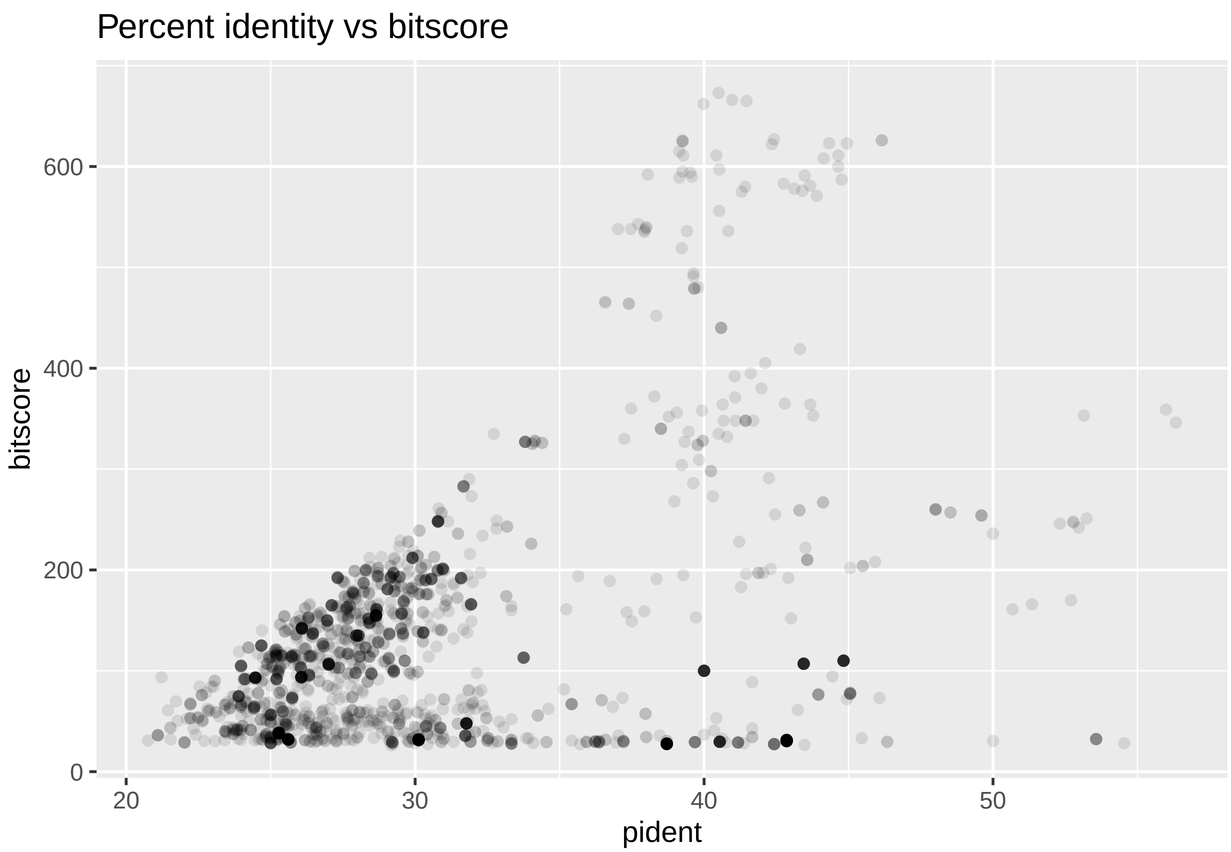


Using ggplot2

```r
library(ggplot2)
```

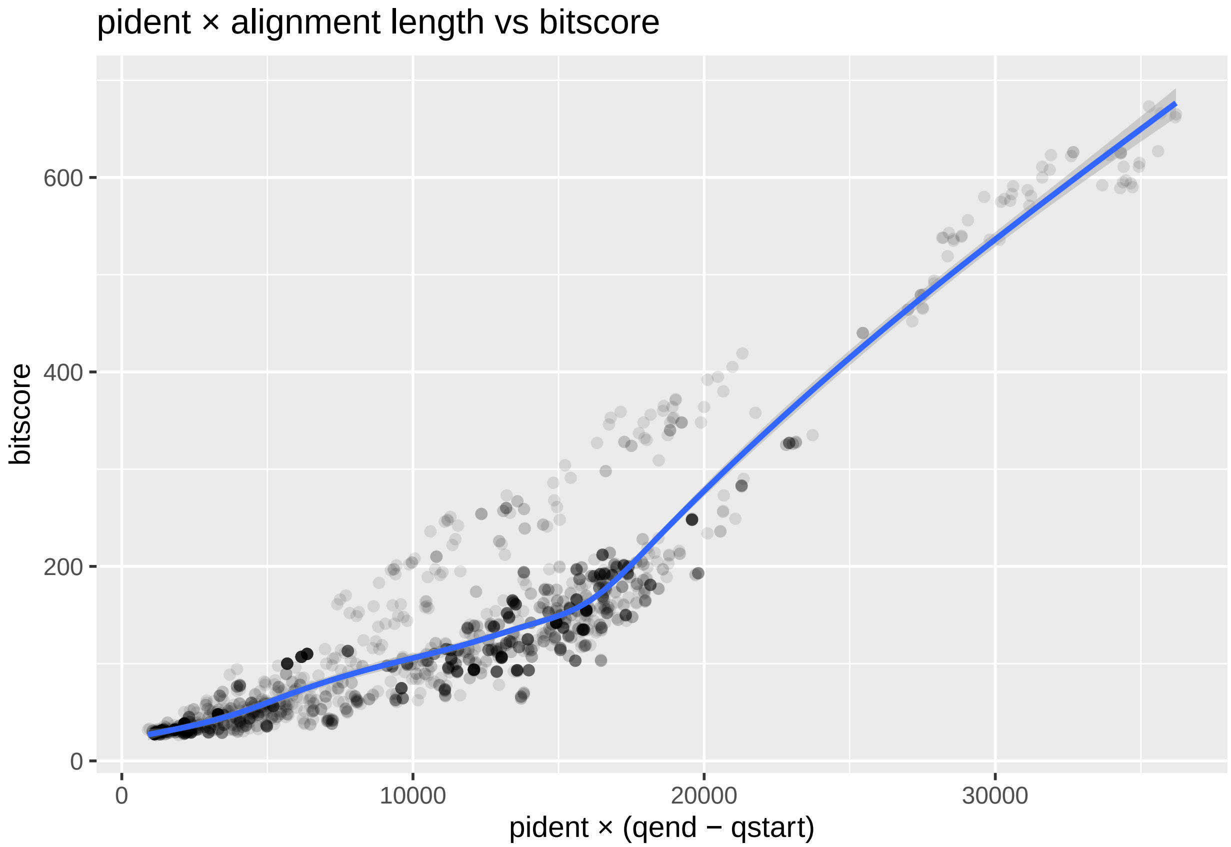
```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```r
ggplot(b, aes(x = pident, y = bitscore)) +
  geom_point(alpha = 0.1) +
  labs(title = "Percent identity vs bitscore")
```



```r
ggplot(b, aes(x = pident * (qend - qstart), y = bitscore)) +
  geom_point(alpha = 0.1) +
  geom_smooth() +
  labs(title = "pident × alignment length vs bitscore",
       x = "pident × (qend - qstart)",
       y = "bitscore")
```

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



Bitscore increases with both identity and alignment length. Two alignments with the same percent identity can have different bitscores if one covers more residues. So bitscore is only somewhat related to percent identity alone; it's more tightly linked to "how many residues are matching over how long a region."

Knit the R Markdown

Run `scp -i ~/Downloads/keyjz.pem -r ubuntu@ec2-44-252-77-168.us-west-2.compute.amazonaws.com:~/work .` in terminal.

`-r` purpose

- Means recursive.
- It copies directories and all their contents (subdirectories, files, etc.).
- Without `-r`, scp will fail on directories.

* Pupose

- `~/work/*` means "all files and directories inside ~/work".
- Copies everything in `~/work` from the remote machine into current local directory.

Using rsync

Run `scp -i ~/Downloads/keyjz.pem -r ubuntu@ec2-44-252-77-168.us-west-2.compute.amazonaws.com:~/work .` in terminal.

- `-a`: Archive mode – copy recursively and preserve permissions, timestamps, and other metadata so the directory is cloned as faithfully as possible.

- **`-z`**: Compress file data during transfer to reduce network bandwidth usage and often speed up transfers.
- **`-P`**: Show progress for each file and keep partially transferred files so interrupted transfers can resume.
- **`--exclude`**: Skip copying any files or directories whose names match the given pattern (e.g. `--exclude="*.psq"`).