Class 13 Unix Lab

Cindy Tran

3/1/2022

4. Inspect our Sequences

> Q. What does the star character accomplish here?

It applies the command to anything that ends with what is after the star.

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

68180

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

It won't put the output of the command into mm-first.fa

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

It would add the data to an existing file called mm-first.fa, but since we don't have a file called that, it would lead to an error.

6. Running More BLAST Jobs

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

Grep -c '>' mm-second.fa

10. Using RStudio Online to Read Your Output

```
tsv <- readr::read_delim("mm-second.x.zebrafish.tsv", col_names = c("qseqid", "sseqid", "pident", "leng
```

```
## Rows: 57616 Columns: 12
```

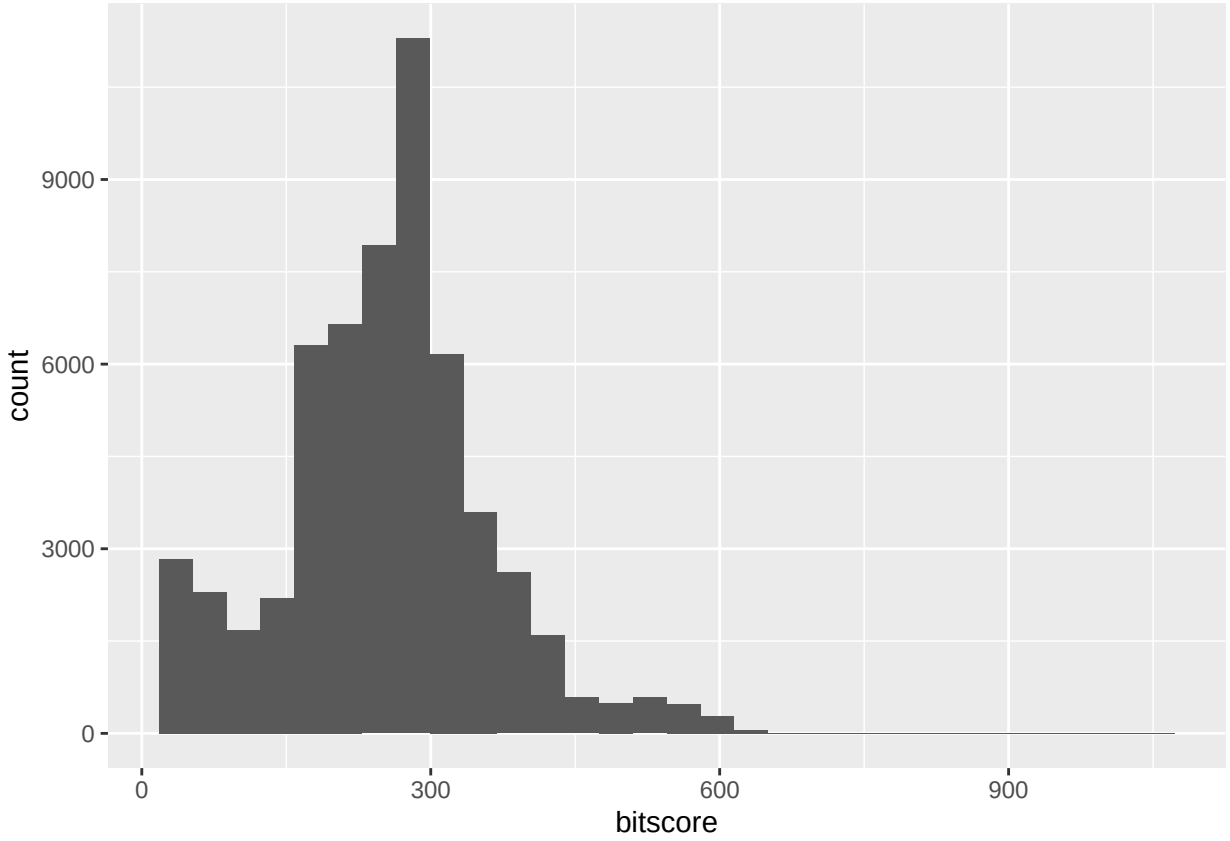
```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (2): qseqid, sseqid
## dbl (10): pident, length, mismatch, gapopen, qstart, qend, sstart, send, eva...

##
## 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.
```

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

```
library(ggplot2)
ggplot(tsv) +
  aes(bitscore) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



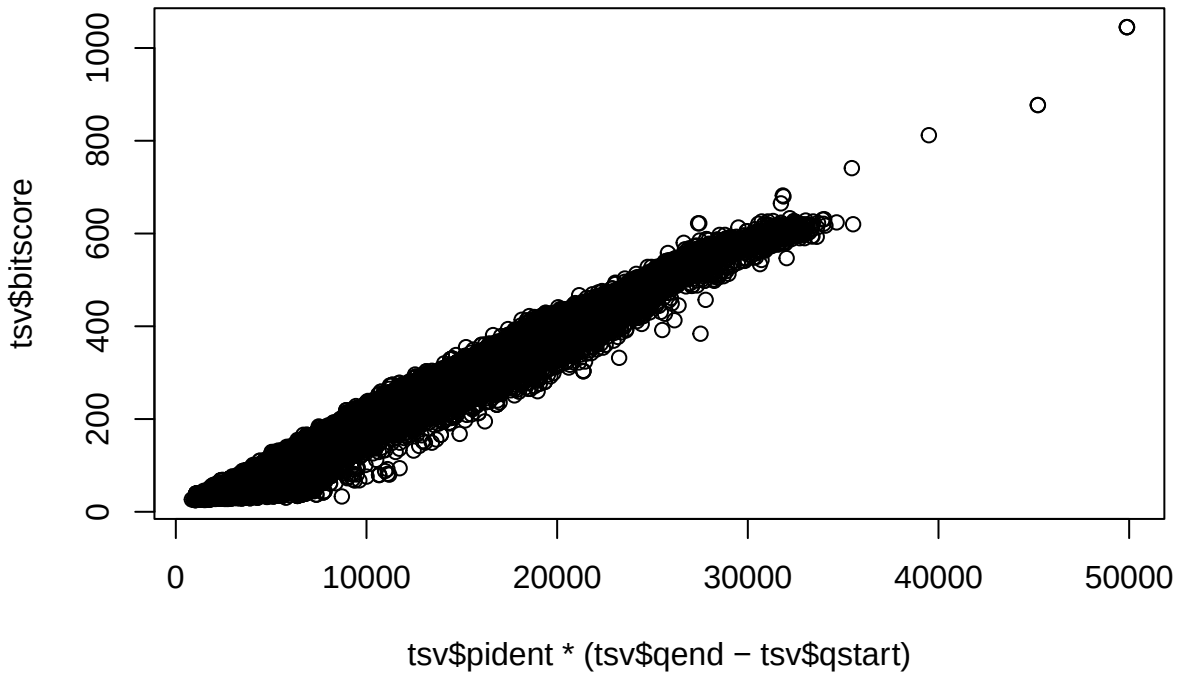
> Q. What do you notice here? Note that larger bitscores are better.

Most bitscores are around 300/

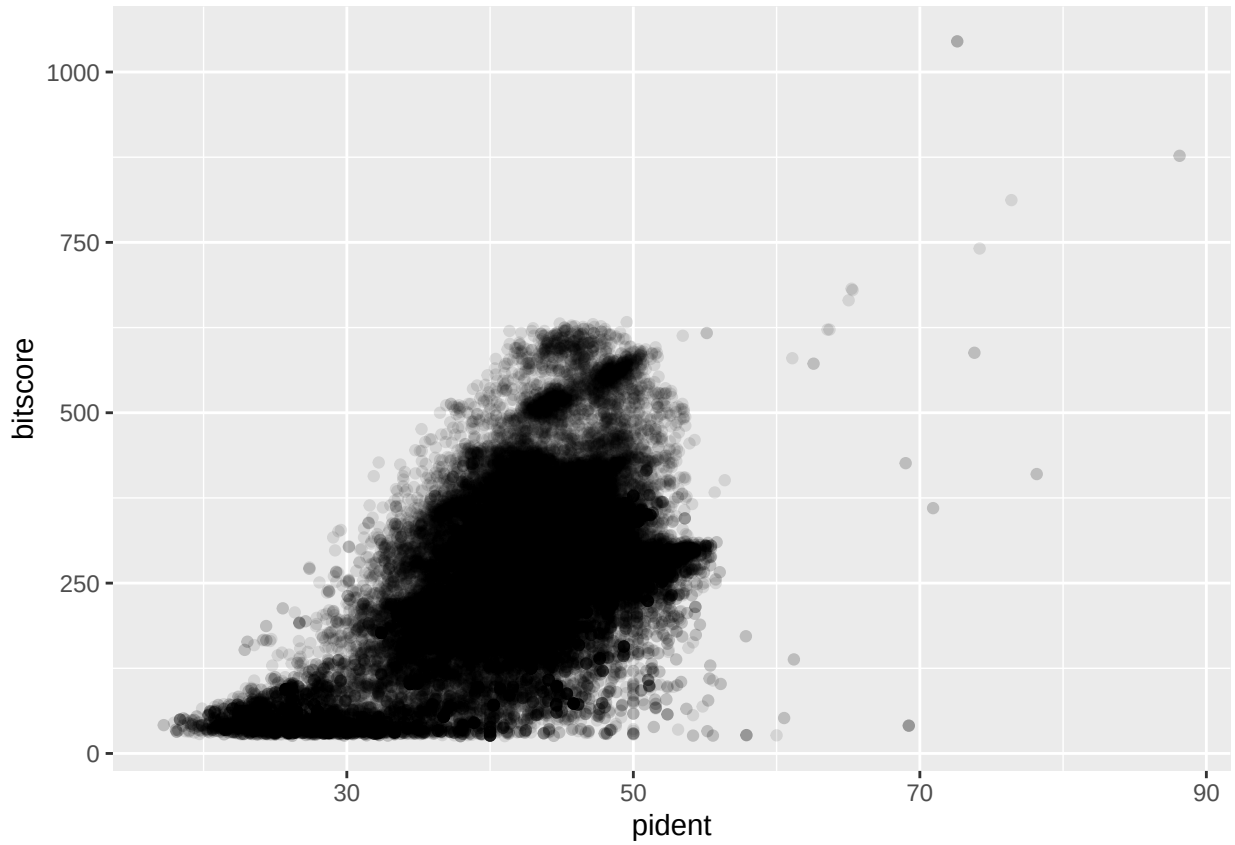
> Is there a straightforward relationship between percent identity (*pident*)*andbitscore*(bitscore) for the alignments we generated?

The answer is that bitscores are only somewhat related to pident; they take into account not only the percent identity but the length of the alignment.

```
plot(tsv$pident  * (tsv$qend - tsv$qstart), tsv$bitscore)
```

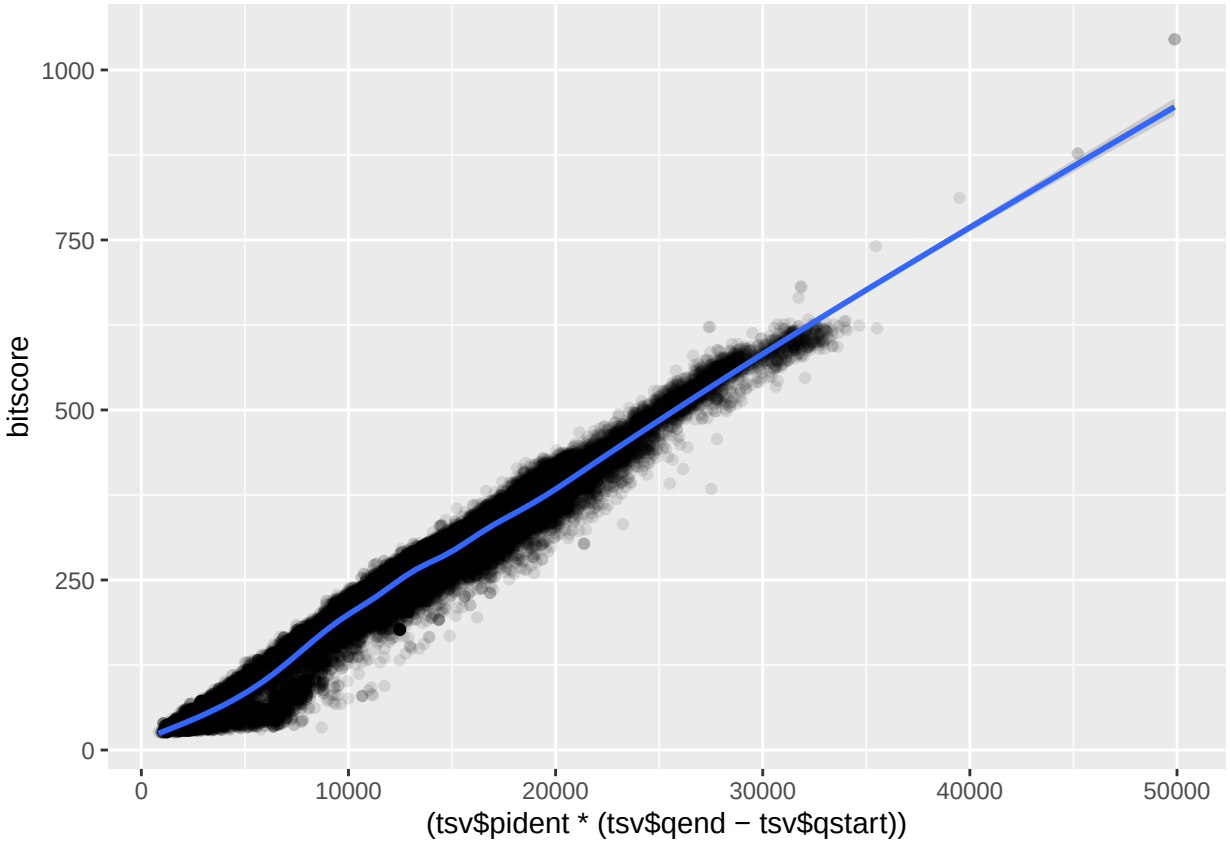



```
ggplot(tsv, aes(pident, bitscore)) + geom_point(alpha=0.1)
```




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

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



On your local machine scp -i ~/Downloads/barry_bioinf.pem -r ubuntu@YOUR_IP_ADDRESS:~/work/* . > Q. Note the addition of the -r option here: What is it's purpose? Also what about the *, what is it's purpose here?

-r means to copy the entire directory and * means to copy everything that is in the work directory.

Q. Why did it take longer to BLAST mm-second.fa than mm-first.fa?

Because it is a larger file

Q. When we plot e-values why do you often work in -log(evalue) units?

Because it makes relationships easier to see and smoother lines when there are large variances.

What is an advantage of rsync over scp?

Rsync allows you to sync both remote and local directories.

What is the advantage of using R (and other tools) on remote machines vs our local computer?

Remote machine can do things much faster.

Q. What is the disadvantage of remote vs local work?

It requires more setup to work remotely.