

# Class 16: Remote AWS Instance

Hyejeong Choi (PID: A16837133)

Read in the `mm-second.x.zebrafish.tsv` results:

```
results <- read.table("results.tsv")
```

Change the column names:

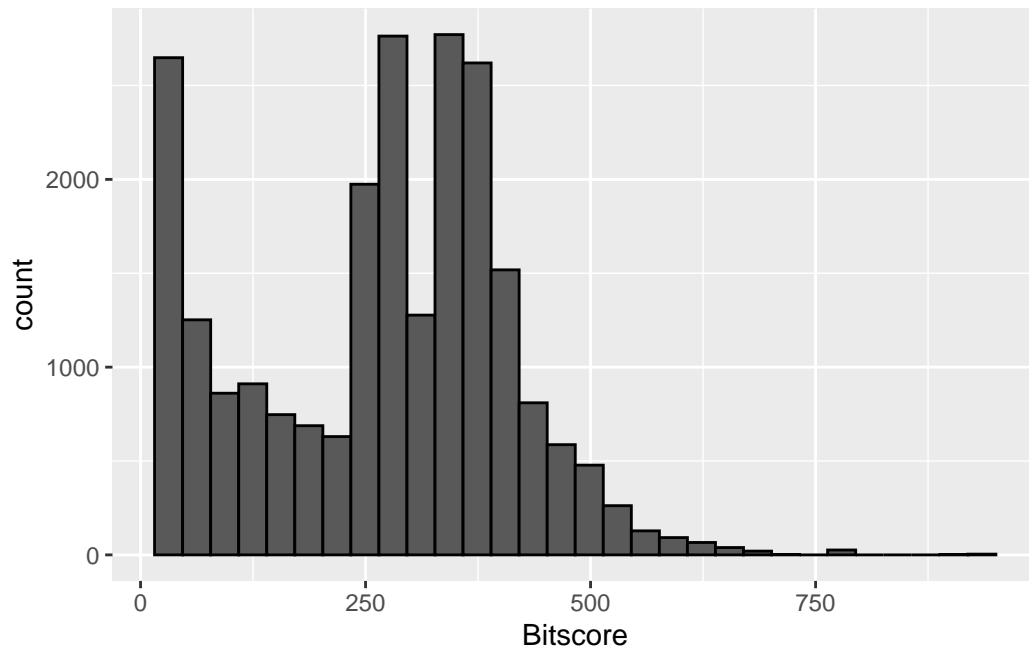
```
colnames(results) <- c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstar"
```

Make a histogram plot of the bitscores:

```
library(ggplot2)

ggplot(results) +
  aes(results$bitscore) +
  geom_histogram(color='black', bins=30) +
  labs(x='Bitscore')
```

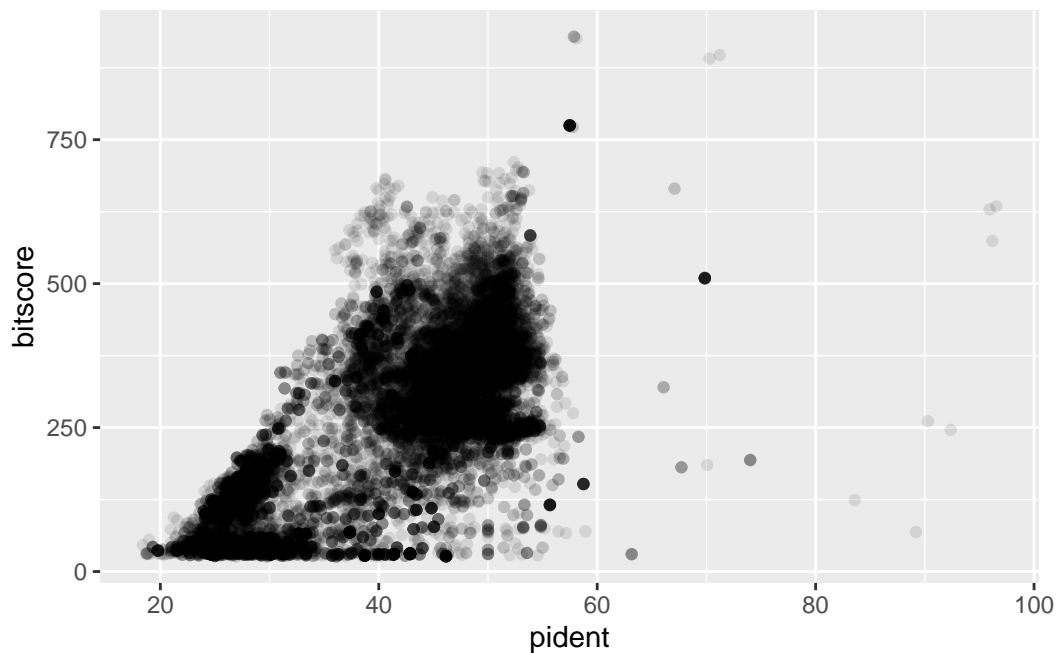
Warning: Use of ``results$bitscore`` is discouraged.  
i Use ``bitscore`` instead.



The histogram shows that the bitscores are clustered around 250-375. The histogram is also skewed to the left, and very few bitscores have high values.

Make a plot for bitscores and percent identity.

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



The plot shows that bitscore and percent identity have a weak, positive correlation.

```
ggplot(results, aes((results$pident * (results$qend - results$qstart)), bitscore)) + geom_point()
```

Warning: Use of `results\$pident` is discouraged.  
i Use `pident` instead.

Warning: Use of `results\$qend` is discouraged.  
i Use `qend` instead.

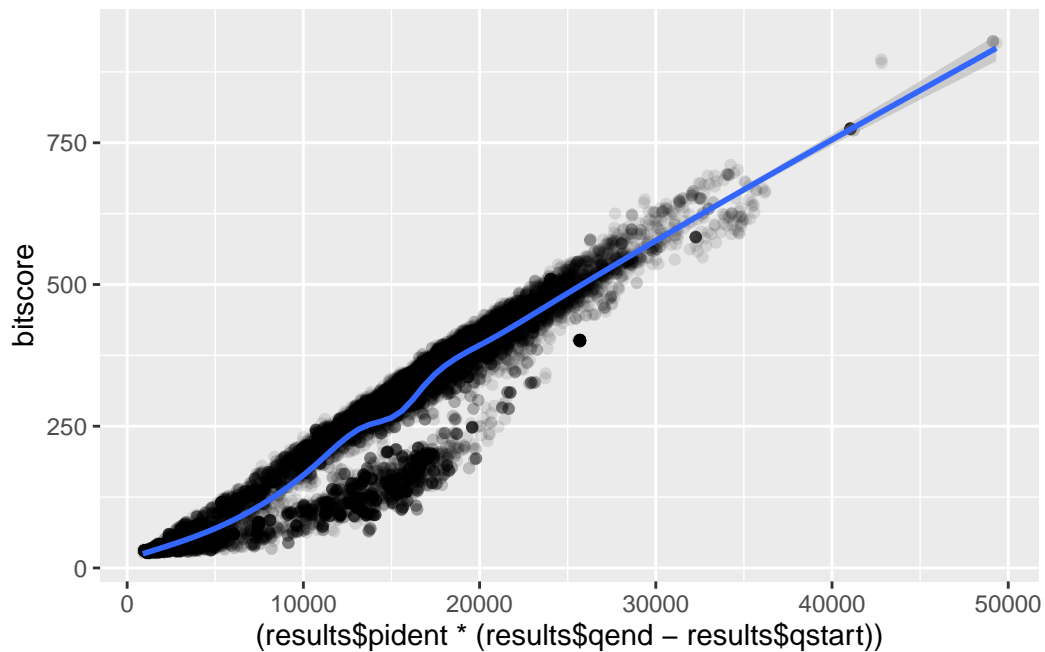
Warning: Use of `results\$qstart` is discouraged.  
i Use `qstart` instead.

Warning: Use of `results\$pident` is discouraged.  
i Use `pident` instead.

Warning: Use of `results\$qend` is discouraged.  
i Use `qend` instead.

Warning: Use of `results\$qstart` is discouraged.  
i Use `qstart` instead.

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



Q. Note the addition of the `-r` option here: What is its purpose? Also what about the `*`, what is its purpose here?

The `-r` is used to connect the remote computer and your local computer so that the indicated files and directories in the code can be transferred. The `*` is used to read the files that all have a similar pattern in the file name that you chose.

Q. There are lots of very useful options that you can supply to `rsync` including the `-P` and `-z` and `--exclude` options. Can you determine what they do?

The `-z` compresses the files. The `-P` integrates `--partial` and `--progress`. The `--exclude` allows you to leave out files and directories that you do not want to include.