Class16

Class16

Loaded data with column names

```
library(readr)
colnam= c("qseqid", "sseqid", "pident", "length", "mismatch", "gapopen", "qstart", "qend",
tsv=read_tsv("second.x.zebrafish.tsv",col_names = colnam)
```

```
Rows: 28789 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.
```

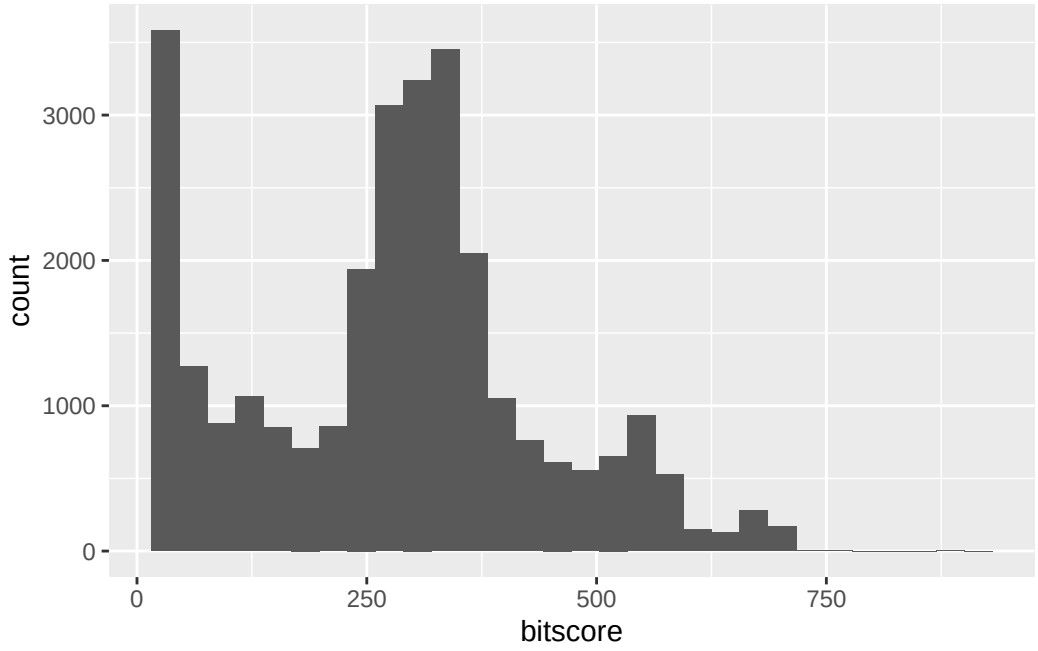
```
head(tsv)
```

```
# A tibble: 6 x 12
  qseqid      sseqid    pident length mismatch gapopen qstart  qend sstart  send
  <chr>       <chr>      <dbl>  <dbl>    <dbl>   <dbl>  <dbl> <dbl>  <dbl> <dbl>
1 NP_598866.1 XP_00929~   46.2    273      130       6      4   267    420   684
2 NP_598866.1 NP_00131~   46.2    273      130       6      4   267    476   740
3 NP_598866.1 XP_00929~   46.2    273      130       6      4   267    475   739
4 NP_598866.1 NP_00118~   33.1    127       76       5      4   126    338   459
5 NP_598866.1 NP_00100~   30.4    125       82       4      4   126    344   465
6 NP_598866.1 NP_00100~   30.6     62       41       2     53   113     43   103
# i 2 more variables: evalue <dbl>, bitscore <dbl>
```

Histogram of bitscore

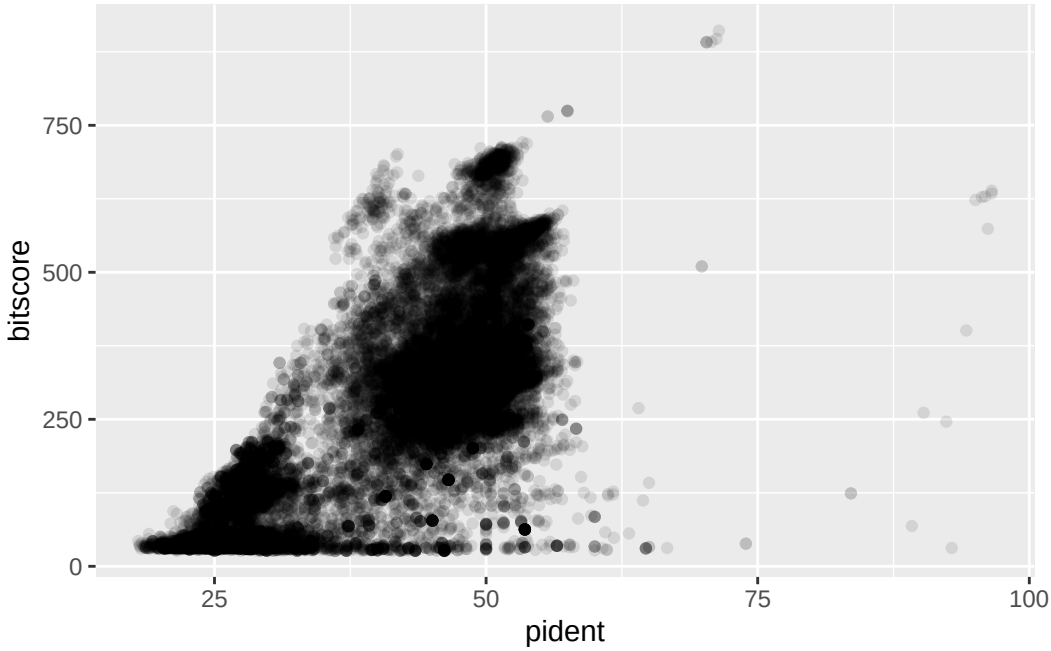
```
library(ggplot2)

ggplot(tsv) +aes(x=bitscore) + geom_histogram()
```

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



Scatterplots

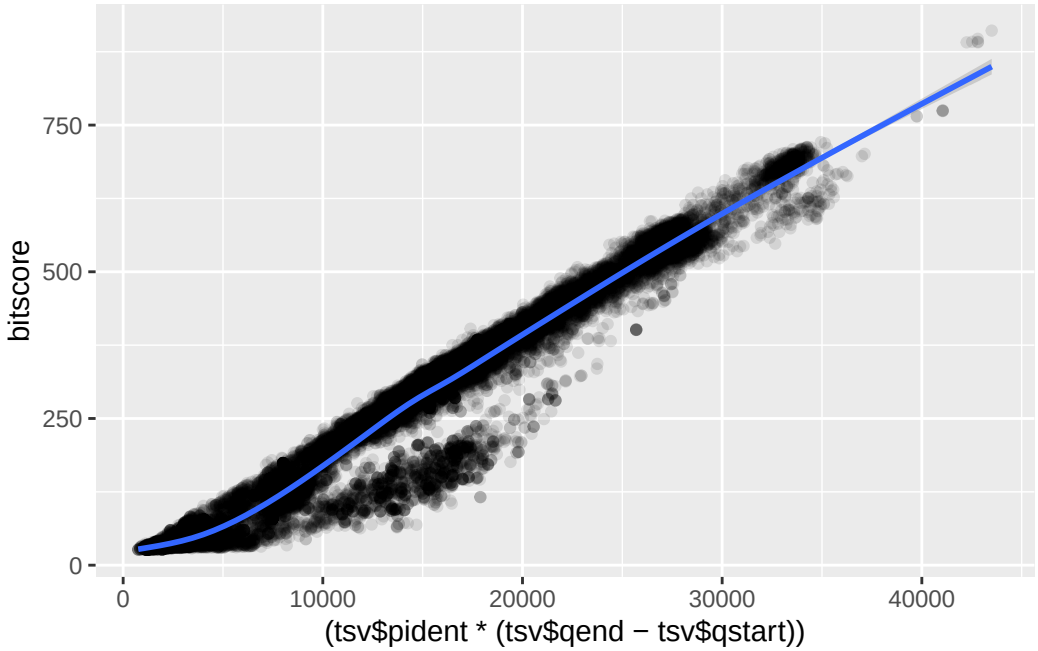
```
ggplot(tsv, aes(pident, bitscore)) + geom_point(alpha=0.1)
```



Taking into account the percent identity and the length of the alignment.

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

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


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