Seminar III: R/Bioconductor

August-December 2009

Leonardo Collado Torres

Bachelor in Genomic Sciences (LCG),
UNAM, Cuernavaca, Mexico

lcollado@lcg.unam.mx

http://www.lcg.unam.mx/~lcollado/

September 10, 2009

Assistants: Alejandro Reyes areyes@lcg.unam.mx, José Reyes jreyes@lcg.unam.mx and Víctor Moreno jmoreno@lcg.unam.mx

Note: Questions through the forum please. Those who are not from the sixth LCG generation send us an email so we can register you on the forum.

Note: I changed the original homework due to some unexpected errors with Uniprot using biomaRt and because the ENSEMBL tutorial was outdated.

Abstract

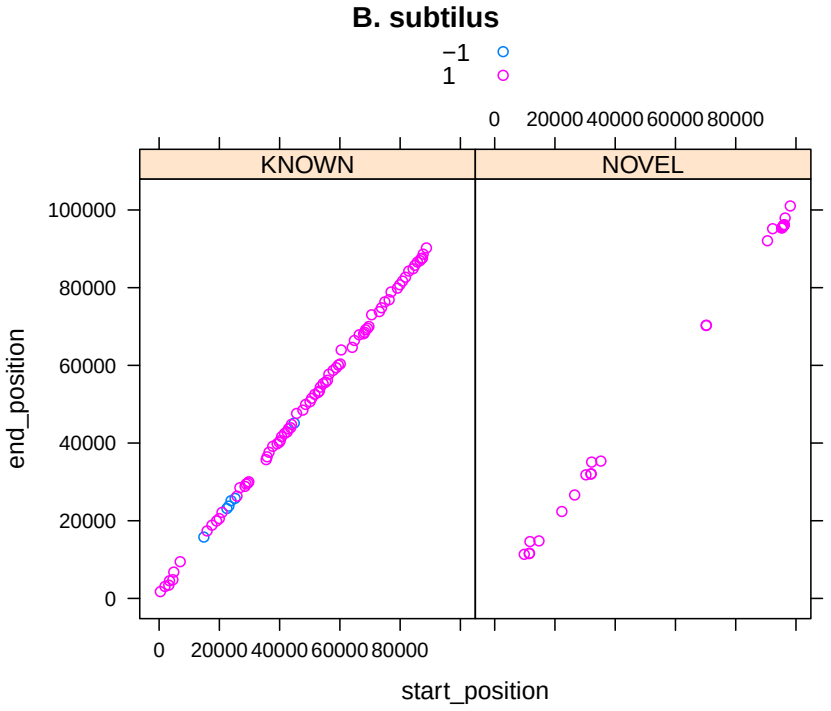
With the following exercises you'll tune your skills with packages such as biomaRt that enable you to download public data sets.

1 biomaRt

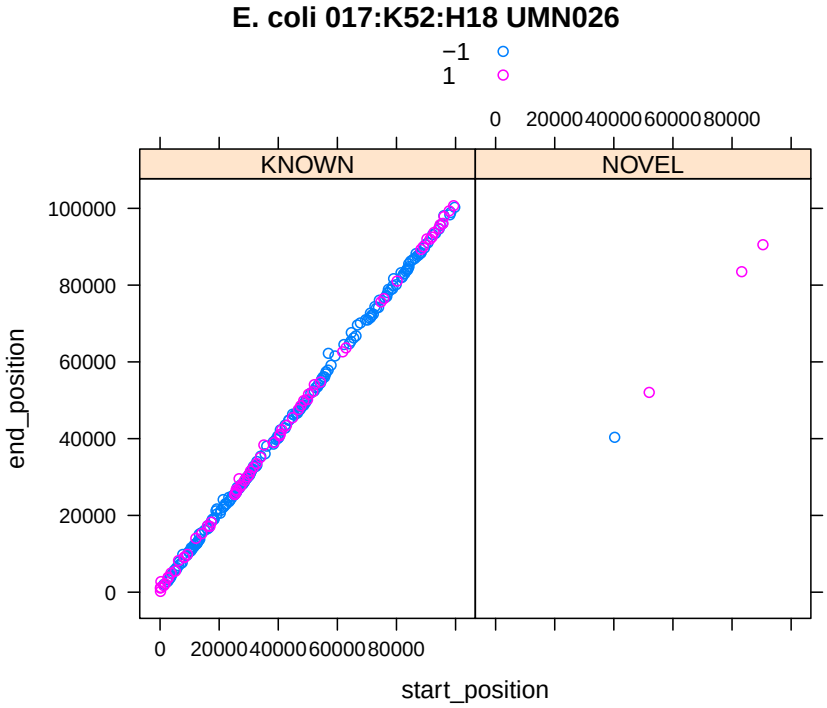
1. On the class we made the following xyplot. Reproduce it for Bacillus anthracis Sterne and Escherichia coli O17:K52:H18 UMN026

```
> library(biomaRt)
> library(lattice)
```

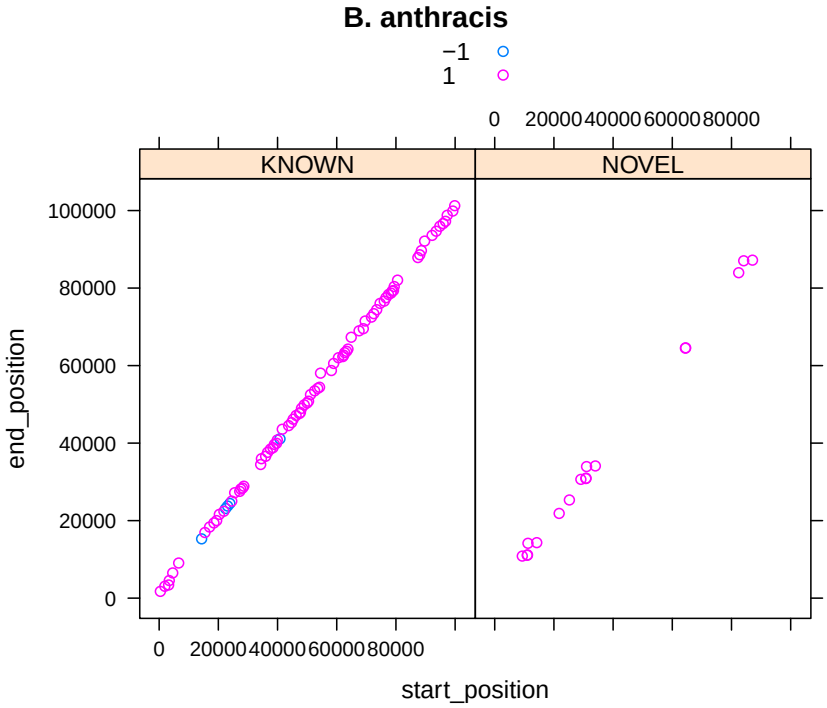
```
> bsub <- useMart("bacterial_mart_54", dataset = "bac_6_gene")
> res <- getBM(attributes = c("start_position", "end_position",
+     "strand", "status"), filters = c("start", "end"), values = list("1",
+     "100000"), mart = bsub)
> print(xyplot(end_position ~ start_position | status, group = strand,
+     data = res, auto.key = T, main = "B. subtilus"))
```




```
> coli <- useMart("bacterial_mart_54", dataset = "esc_32471_gene")
> res2 <- getBM(attributes = c("start_position", "end_position",
+     "strand", "status"), filters = c("start", "end"), values = list("1",
+     "100000"), mart = coli)
> print(xyplot(end_position ~ start_position | status, group = strand,
+     data = res2, auto.key = T, main = "E. coli 017:K52:H18 UMN026"))
```




```
> anthrax <- useMart("bacterial_mart_54", dataset = "bac_20000_gene")
> res3 <- getBM(attributes = c("start_position", "end_position",
+     "strand", "status"), filters = c("start", "end"), values = list("1",
+     "100000"), mart = anthrax)
> print(xyplot(end_position ~ start_position | status, group = strand,
+     data = res3, auto.key = T, main = "B. anthracis"))
```



2. Compare the three plots and the resulting data sets. Write your own conclusions

```
> dim(res)

[1] 110   4

> dim(res2)

[1] 262   4

> dim(res3)

[1] 101   4
```

B. subtilus and B. anthracis have a nearly the same number of genes on the first 100000bp with B. subtilus having 9 more and a total of 110. Both have a very strong bias for the + strand with the great majority being of known type. E. coli 017:K52:H18 UMN026 has only 4 novel genes but overall has 262 genes on the same subset, which more than doubles any of the other two bacteria. On the xyplot this fact is reflected by a higher density of points, and there seems to be no bias on the strand. Some parts have genes on both

strands, though we can clearly notice some segments where nearly all the genes are on one strand; specially on the - strand.