BioC for HTS - PDCB topic
Infrastructure and Input/Output 03

LCG Leonardo Collado Torres
lcollado@wintergenomics.com – lcollado@ibt.unam.mx

October 20th, 2010

Practice

Data

- We'll use some real data!
- We have 500k paired-end reads with an insert size around 450bp from Ecoli.
- These reads were aligned to the genome using Bowtie (first base == 1).

Long Exercise

1. Read the alignment file into R.
2. Plot the nucleotide frequency by cycle for:
   2.1 reads aligning to the plus strand
   2.2 reads aligning to the minus strand
3. Was there any difference?
4. Construct a GRanges object where you'll have 1 range for every pair of reads. You will probably need to check how Bowtie handles paired-end alignments and the definition of the Bowtie output. The Ecoli genome is 4639675 bp long.
5. Make a plot showing the distribution of the paired-end read sizes (boxplot, histogram, cumulative plot, ...). Is it near the expected value?

Long Exercise

6. How many of the reads overlap (ignore self and ignore redundant)?
7. If the reads overlap, make a plot of the distribution of the number of overlaps per read.
8. Calculate the coverage per base of the genome. Plot the distribution (ignore bases with a coverage of 0). What is the median coverage?
9. If you were to split the genome every time the coverage went to 0, how many fragments do we have? Plot the distribution of the sizes of these fragments.
10. If we were to consider all bases with a coverage less than the median coverage (the one we calculated above) as 0s, how many fragments do we have? Plot the distribution of their sizes.

Notes

- Try to use lattice for the plots.
- Use the full file for the exercise. However, if you just want to test your code or if your laptop doesn't have much RAM feel free to use the short version.

```
> library(ShortRead)
> aln <- readAligned(dirPath = ".",
+     pattern = "R017_8_iterative.map",
+     type = "Bowtie")
> print(object.size(aln), units = "Mb")
```

```
336.8 Mb
```

Notes

```
> alnShort <- readAligned(dirPath = ".",
+     pattern = "R017_8_iterative_short.map",
+     type = "Bowtie")
> print(object.size(alnShort), units = "Mb")
```

```
16.9 Mb
```

Session Information

```
> sessionInfo()

R version 2.12.0 Under development (unstable) (2010-09-08 r52880)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.utf8
 [2] LC_NUMERIC=C
 [3] LC_TIME=en_US.utf8
 [4] LC_COLLATE=en_US.utf8
 [5] LC_MONETARY=C
 [6] LC_MESSAGES=en_US.utf8
 [7] LC_PAPER=en_US.utf8
 [8] LC_NAME=C
 [9] LC_ADDRESS=C
[10] LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.utf8
[12] LC_IDENTIFICATION=C

attached base packages:
```

Session Information

```
[1] stats     graphics  grDevices
[4] utils     datasets  methods
[7] base

other attached packages:
[1] ShortRead_1.7.20
[2] Rsamtools_1.1.15
[3] lattice_0.19-11
[4] Biostrings_2.17.41
[5] GenomicRanges_1.1.25
[6] IRanges_1.7.34

loaded via a namespace (and not attached):
[1] Biobase_2.9.0 grid_2.12.0
[3] hwriter_1.2
```