Assignment 2

May 15, 2017

0. Load the data with percent overlap/non-overlap per chromosome

```
df <- read.table(file="coverage.txt")
```

Part 1

(a)What are the first five genomic nucleotides from the first exon of this transcript?

AAAGG

The DICER1 mRNA molecule should have the same sequence as the sense DNA genomic sequences (substituting Ts by Us). As AK002007 is transcribed on the reversed strand and the default genomic sequence presented by the browser is the antisense one, we have to reverse it. Therefore, the first five nucleotides of the exon in the AK002007 cDNA are AAAGG.

(b) Look at the raw mRNA sequence of AK002007, from the database it actually comes from. What are the first five nucleotides?

When we check the raw mRNA sequence of AK002007, it can be seen that the first five nucleotides are GAAGC.

(c) How do you explain the discrepancy (maximum 5 lines)?

What is going on?

The sequencing process generated a truncated version of the mature mRNA starting at one of the last exons of the DICER1 gene. This truncated molecule bears the last 11 nucleotides of exon 21/27 ("gaagcaaaaag") and continues to hold the nucleotides in exon 22/27. Nevertheless, the aligner has mismapped these first nucleotides. Instead, the first 7 letters have been ignored and the remaining 4 letters have been mapped to the end of the intronic region between exons 21 and 22.

How could that happen?

This could be due to the aligner penalizing opening the intron gap just to align the leading 11 letters. This is only allowed because unfortunately the last 4 letters of exon 21 are identical to the last 4 letters in the intron, therefore providing to the aligner with the freedom to choose where to put these 4 bases

Part 2

(a)

```
## Warning in order(as.numeric(substring(df$V1, 4))): NAs introduced by
## coercion
```

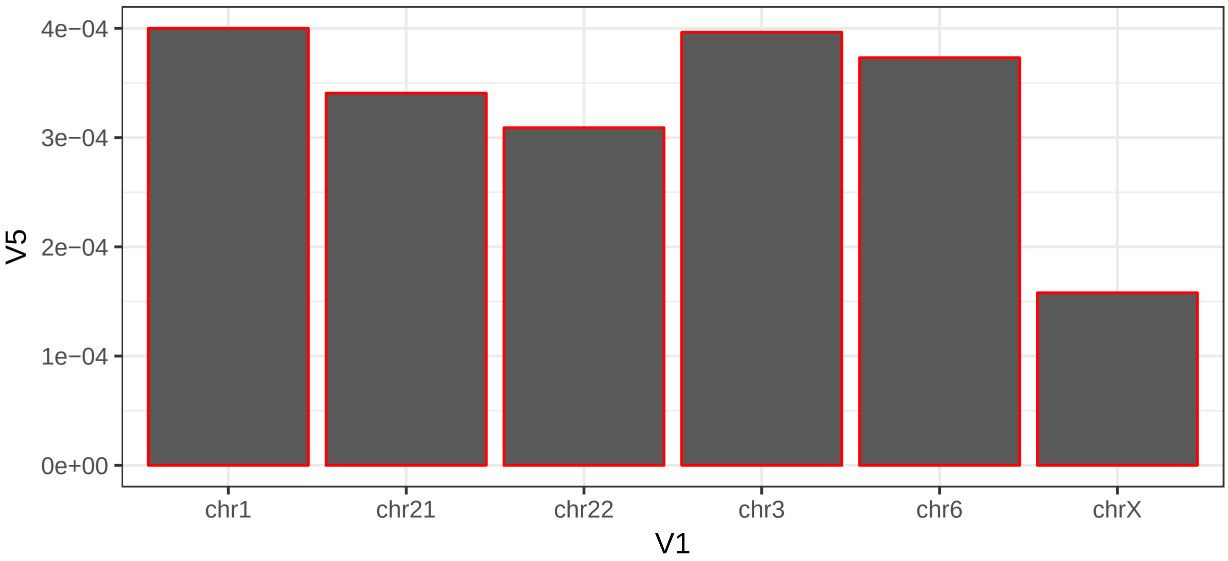
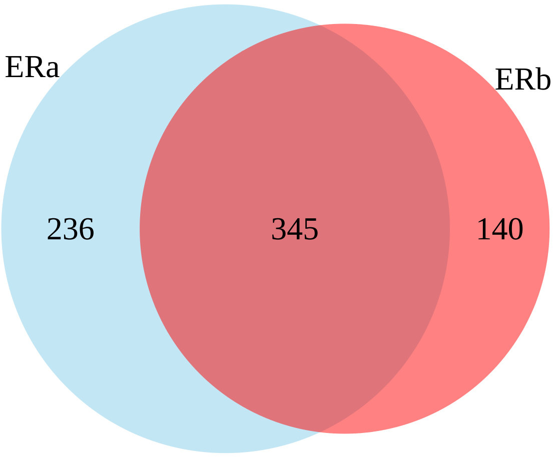


Figure 1. **A** *Histogram showing the distribution of the exon counts. Even though most of the genes contain less than 60 exons, as many as 150 may be found in some of them.* **B** *Detail for genes with max. 20 exons. The mode can be visualized at 3-5 exons per gene (max found at 4). The number of exons per gene decreases steadily beyond it.*



```
## (polygon[GRID.polygon.42], polygon[GRID.polygon.43], polygon[GRID.polygon.44], polygon[GRID.polygon.4
```

8. Appendix

Command line entries

bedtools genomecov -i ERa_hg18.bed -g genome.txt > coverage.txt bedtools intersect -a ERa_hg18.bed -b ERb_hg18.bed -c > AtoBoverlap.bed bedtools intersect -a ERb_hg18.bed -b ERa_hg18.bed -c > BtoAoverlap.bed