Exercise RNAseq Workflow: Task 3

Franca Eichenberger

Monday, October 24, 2016

Choose an organism and extract its annotated transcripts

I characterize the *Otolemur garnettii* genome from the BioMart database “Esemble Genes”.

```
otolemur<- useMart("ENSEMBL_MART_ENSEMBL", dataset ="ogarnettii_gene_ensembl",
                   host = "www.ensembl.org")
otolemur_annotation<-getBM(attributes = c("ensembl_gene_id","chromosome_name",
                                          "transcript_count"), mart=otolemur)
head(otolemur_annotation)
```

```
##      ensembl_gene_id chromosome_name transcript_count
## 1 ENSOGAG00000033513      GL873711.1                1
## 2 ENSOGAG00000004848      GL873711.1                1
## 3 ENSOGAG00000004855      GL873711.1                1
## 4 ENSOGAG00000034757      GL873711.1                2
## 5 ENSOGAG00000032358      GL873711.1                1
## 6 ENSOGAG00000034129      GL873711.1                1
```

```
otolemur_transcript<- getBM(attributes = c("ensembl_transcript_id",
                                           "transcript_length"), mart = otolemur)
head(otolemur_transcript)
```

```
##   ensembl_transcript_id transcript_length
## 1    ENSOGAT00000033717               468
## 2    ENSOGAT00000004853              3600
## 3    ENSOGAT00000004856              1137
## 4    ENSOGAT00000034967               858
## 5    ENSOGAT00000032533               882
## 6    ENSOGAT00000034336               513
```

I created two queries: Anannotation query including the *gene id*, *chromosomal location* and *transcript count per gene* from the otolemur genome and a separate Transcript query with the *transcript ID* and *transcript length*.

Visually explore the data

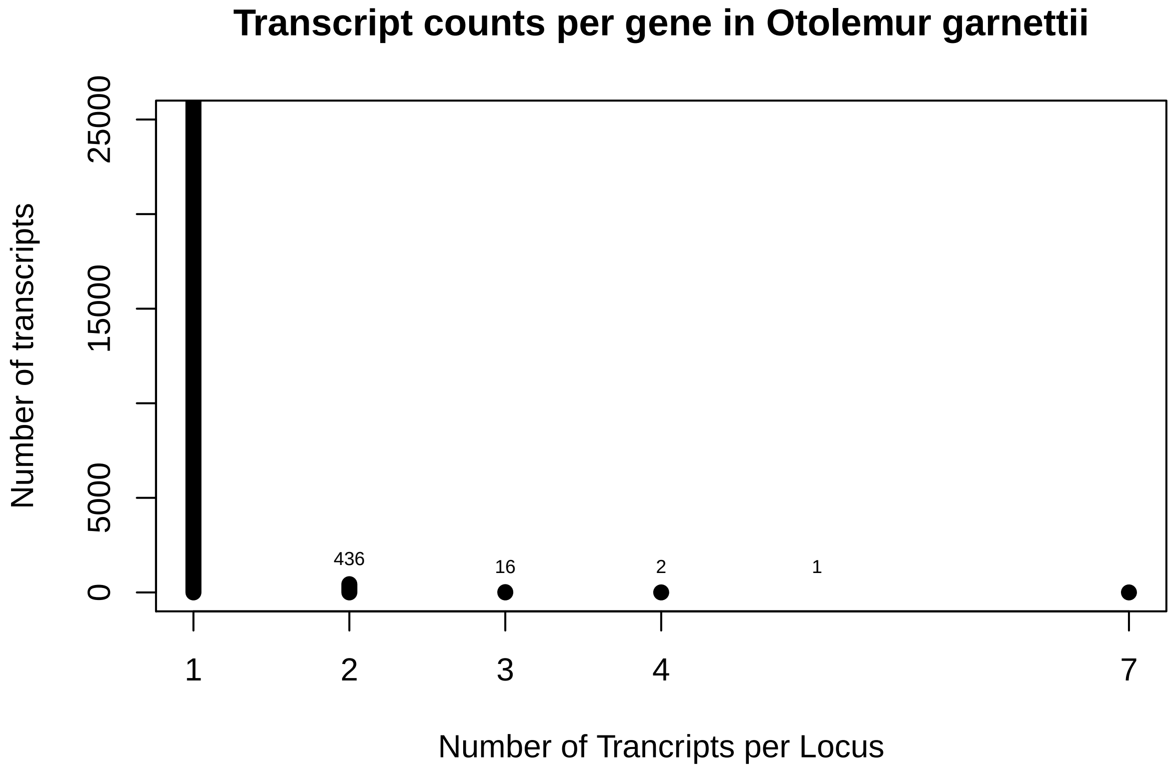
Number of Genes and Transcripts

Cumulative frequency distribution with the number of transcripts per locus:

```
paste( "Total number of genes = ",nrow(otolemur_annotation))
```

```
## [1] "Total number of genes =  28085"
```

```r
plot(table(otolemur_annotation[,3]),xlab = "Number of Trancripts per Locus",
     ylab = "Number of transcripts",
     type = "h", lwd =8,main = "Transcript counts per gene in Otolemur garnettii",
     ylim=c(0,25000))
text(table(otolemur_annotation[,3]),label= table(otolemur_annotation[,3]),
     col='black', pos = 3, cex = 0.6)
```



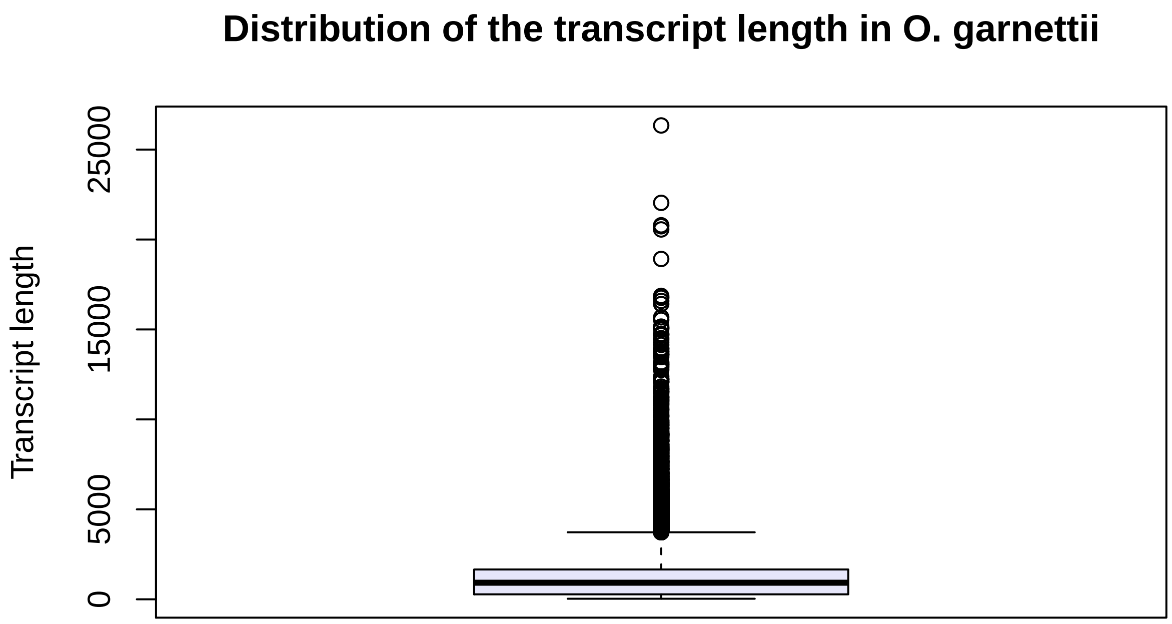
Locus 1 has by far the highest number of transcripts, but Loci 2, 3, 4 and 7 also show small numbers of transcripts.

Distribution of Transcript length

```r
paste( "Total number of transcripts = ",nrow(otolemur_transcript))
```

```
## [1] "Total number of transcripts =  28565"
```

```r
boxplot(otolemur_transcript[,2],
        main = "Distribution of the transcript length in O. garnettii",
        ylab = "Transcript length", col = "lavender")
```

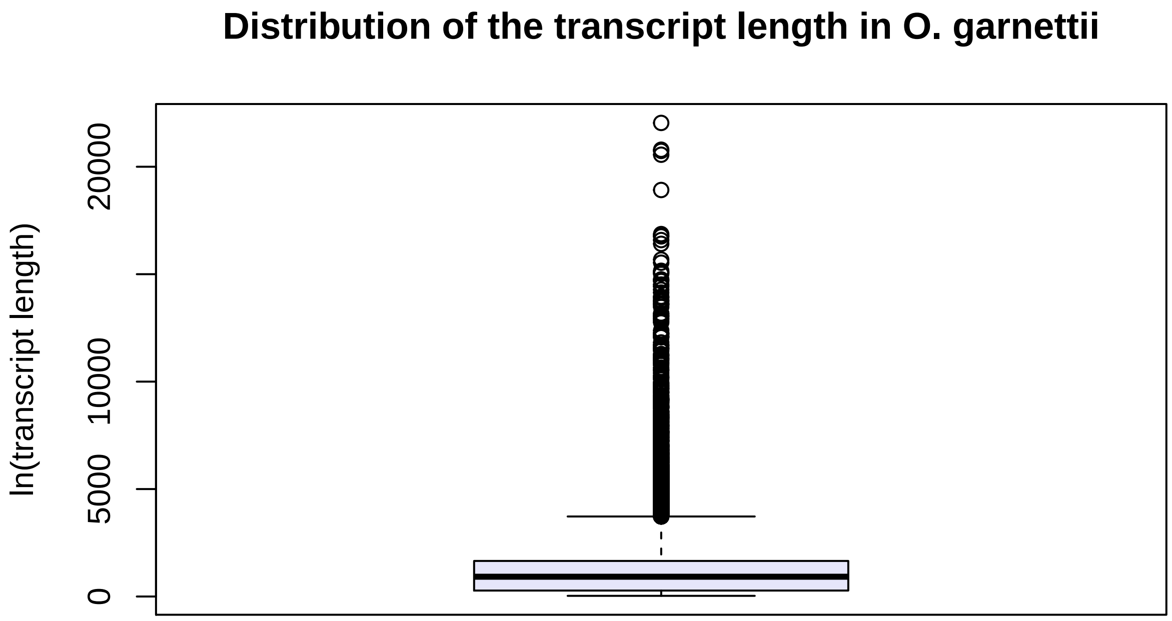


```
summary(otolemur_transcript[,2]) ## to get a summary of the transcript lengths
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       32     277     921    1250    1657   26340
```

There seems to be an outlier with length >2500, removing the outlier gives a better overview of the data:

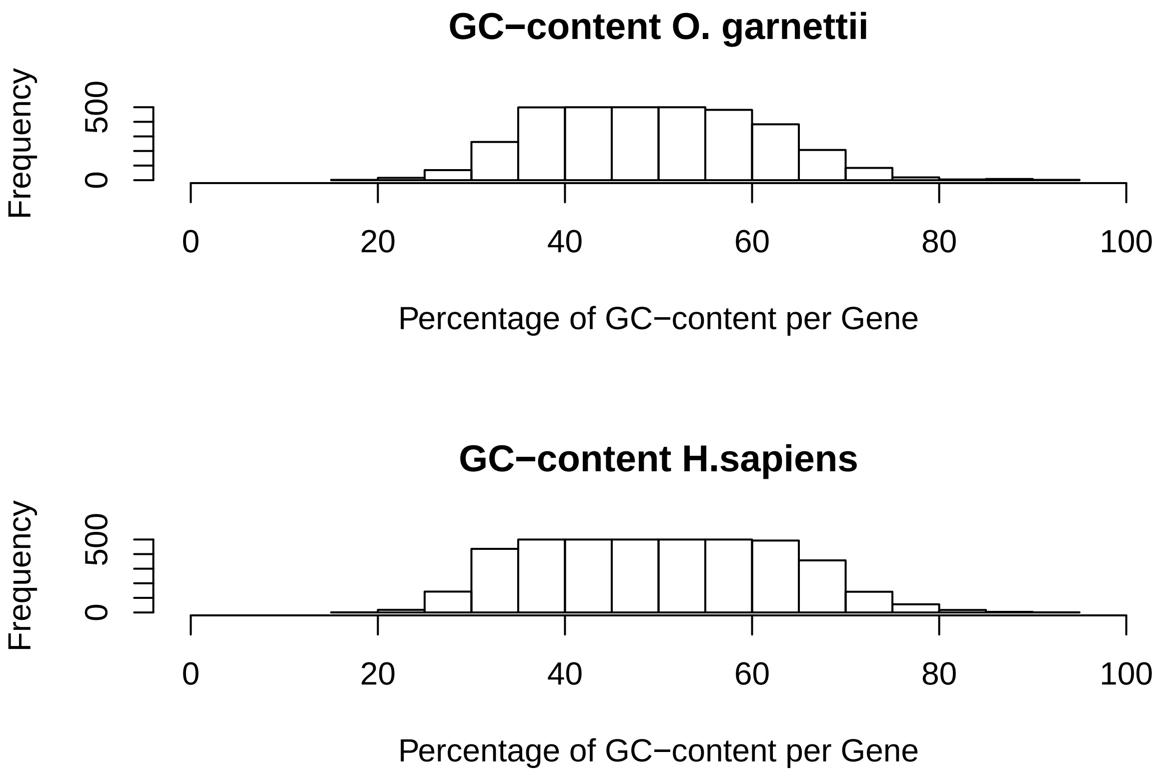
```
boxplot(otolemur_transcript[-which.max(otolemur_transcript[,2]),2],
        main = "Distribution of the transcript length in O. garnettii",
        ylab = "ln(transcript length)", col = "lavender", xlog = TRUE)
```



GC Content in *O.garnettii* compared to *H.sapiens*

```r
otolemur_gccontent <- getBM(attributes = c("percentage_gc_content"),
                            mart = otolemur)

sapiens<- useMart("ENSEMBL_MART_ENSEMBL", dataset ="hsapiens_gene_ensembl",
                  host = "www.ensembl.org")
sapiens_gccontent<- getBM(attributes = c("percentage_gc_content"), mart = sapiens)
par(mfrow = c(2,1))
hist(otolemur_gccontent[,1], main = "GC-content O. garnettii",
     xlab = "Percentage of GC-content per Gene", xlim = c(0,100))
hist(sapiens_gccontent[,1], main = "GC-content H.sapiens",
     xlab = "Percentage of GC-content per Gene", xlim = c(0,100) )
```



```r
par(mfrow = c(1,1))
```

The genome of Homo sapiens has a higher CG Content than Otolemur garnettii one.