STA426 Week 5 Exercise

Franziska Lampart (franziska.lampart@uzh.ch)

October 22, 2016

Part 1

Extracting Information from Ensembl

Using the biomaRT library for R, a small set of annotations was extracted from the *Gallus gallus* genome from the BioMart database "Ensemble Genes".

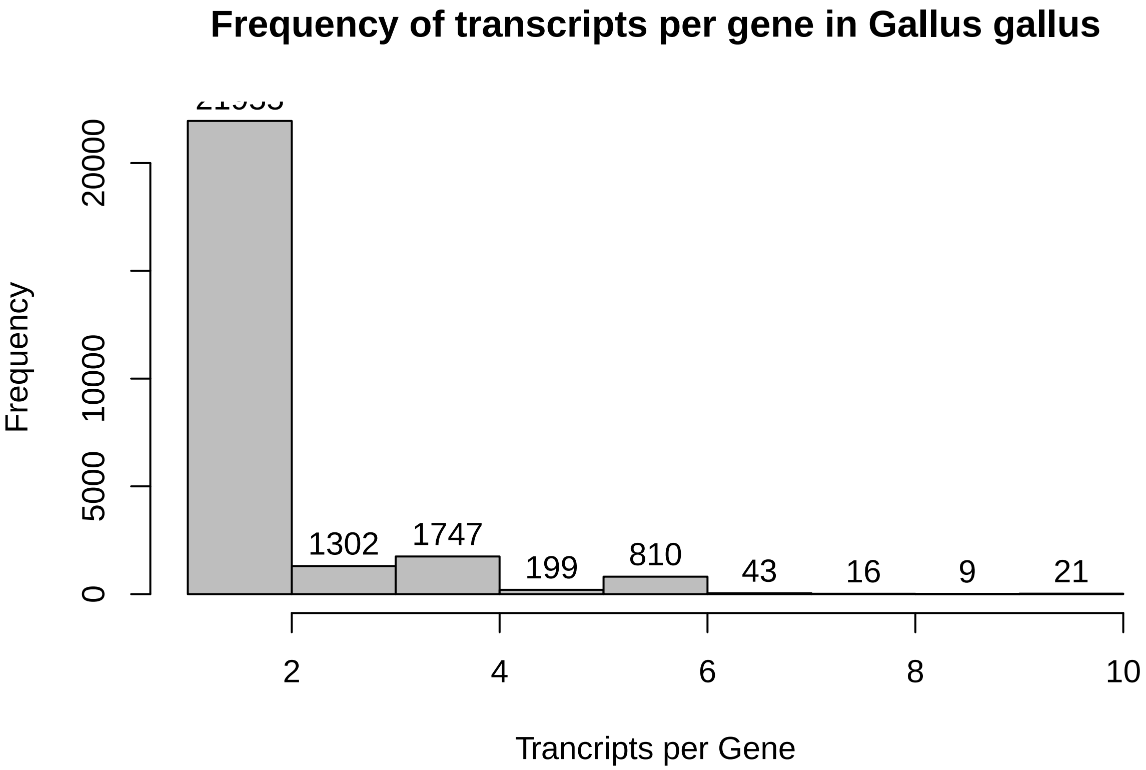
```r
ensembl = useMart("ensembl")
ensembl = useDataset("ggallus_gene_ensembl", mart=ensembl)

Ggallus_annot <- getBM(attributes=c('ensembl_gene_id','transcript_count',
                                    'percentage_gc_content'), mart = ensembl)
Ggallus_trancripts <- getBM(attributes=c('ensembl_transcript_id',
                                         'transcript_length'), mart = ensembl)
```

With the arguments attributes the annotations of interest were filtered from the data set. There are 1409 attributes available to search for. In the present exercise the attributes for the *Gene_ID*, *Transcript_Count*, *GC_Content*, *Transcript_Length* and *Transcript_ID* where used.

Analysing and Plotting the data

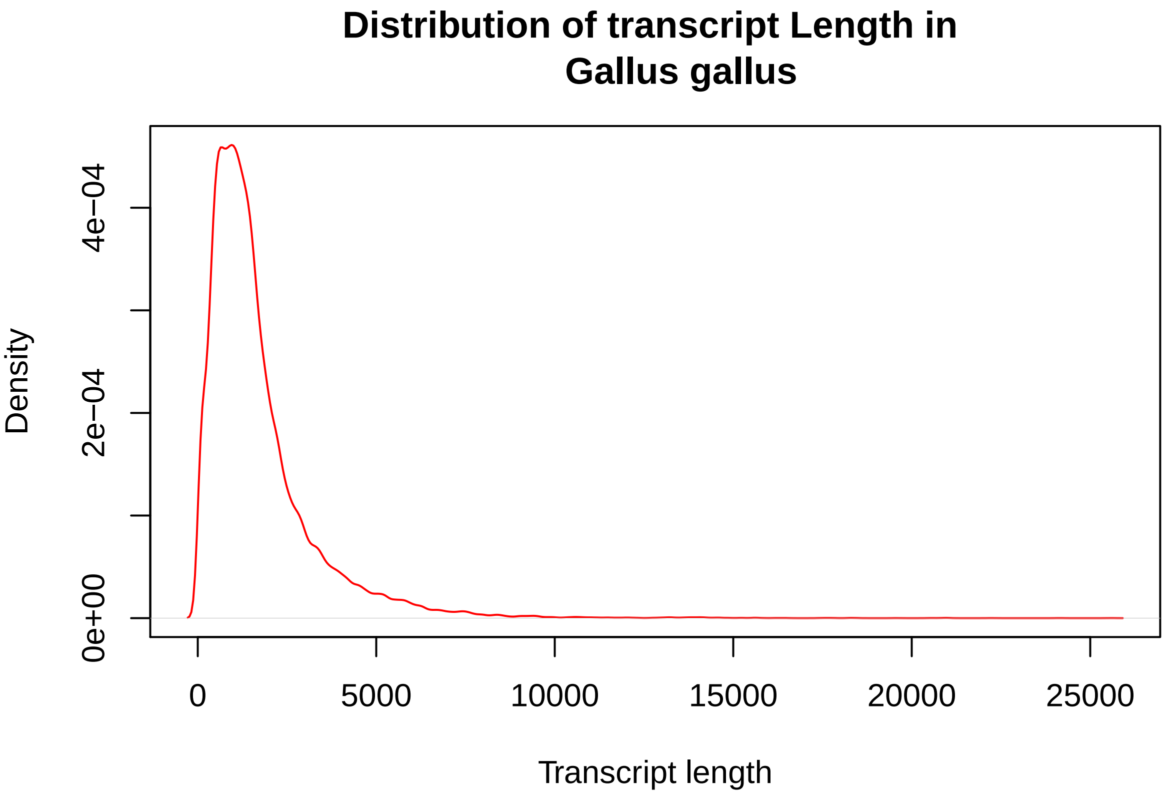
Number of Genes and Transcripts

There are 26102 genes annotated for *Gallus gallus* with a total number of 48760 transcripts. The number of transcripts per gene is visualized below;



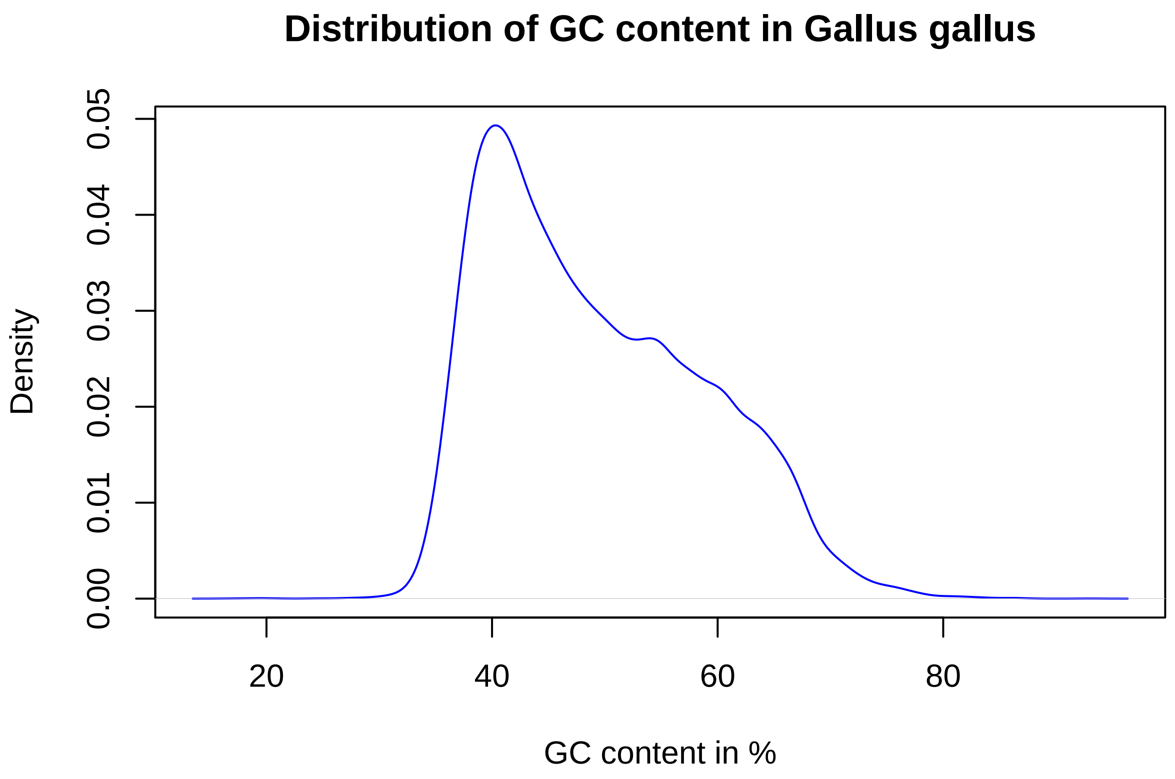
Distribution of Transcript length

The distribution of transcript length is based on the transcripts including the 5'- and 3'-UTR, with the shortest transcript length of 36 and a maximum length of 25593. The average transcript length is 1702.

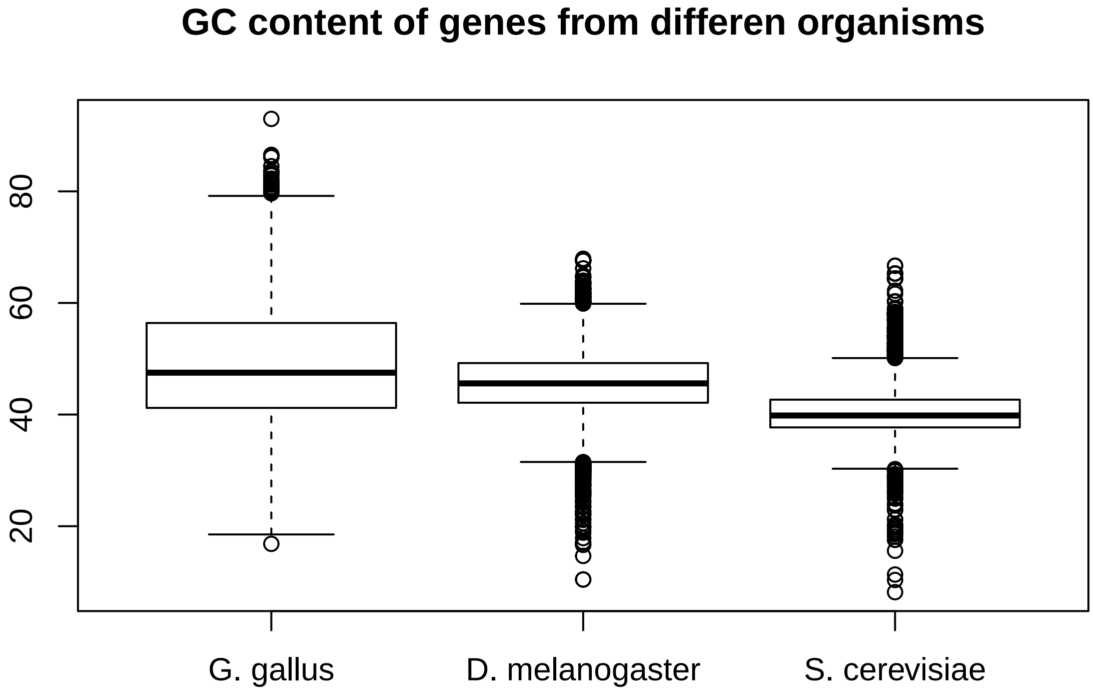


GC content of the genes

Below the density distribution of the GC content of all annotated genes is shown. The average GC content is 49.27



With this library it is possible to quickly extract different annotation information of variable organisms, which facilitates the comparison. As an example comparing the GC content of three different species.



Appendix

Session Info

```
## R version 3.2.3 (2015-12-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04 LTS
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] biomaRt_2.26.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.7          IRanges_2.4.8        XML_3.98-1.4
##  [4] digest_0.6.10        assertthat_0.1       bitops_1.0-6
##  [7] DBI_0.5-1            stats4_3.2.3         formatR_1.4
## [10] magrittr_1.5         RSQLite_1.0.0        evaluate_0.10
## [13] stringi_1.1.2        S4Vectors_0.8.11     rmarkdown_1.1
## [16] tools_3.2.3          stringr_1.1.0        Biobase_2.30.0
## [19] RCurl_1.95-4.8       parallel_3.2.3       yaml_2.1.13
## [22] BiocGenerics_0.16.1  AnnotationDbi_1.32.3 htmltools_0.3.5
## [25] knitr_1.14           tibble_1.2
```