Data Science for Bioinformatics - Week 01

Palle Villesen Not

August 32, 2318

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Q: Try and knit this this document now!

Press the "knit button" or press ctrl+shift+k

A little more on R markdown

You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

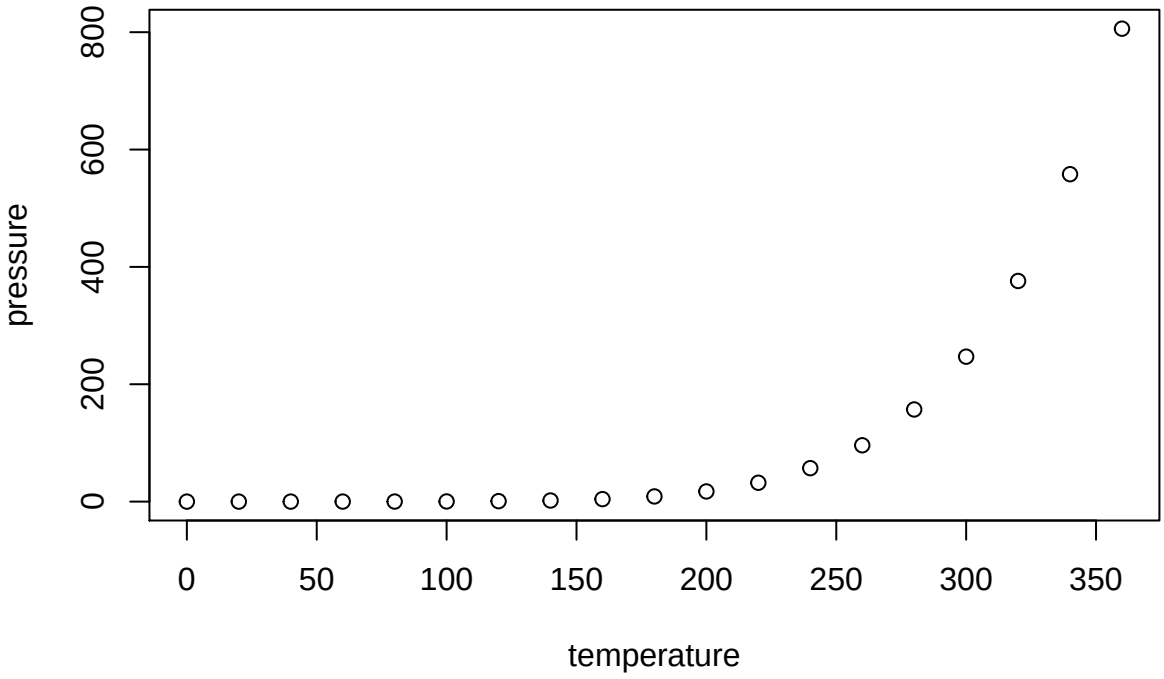
You insert these chunks of code by pressing ctrl+alt+i or by the >Code>Insert chunk menu

Q: Insert a chunk of code that uses the head() function to inspect the cars dataset

Including Plots

You can also embed plots, for example:

```
plot(pressure) # sad, no colors
```



R for data science exercises: Data visualization

First you should work through two online tutorials that will introduce you to ggplot2

Either you can do all the exercises in this document (to save your answers) or a new one.

The tutorial is from the free book "R for data science" written by the R Overlord Hadley Wickham

URL: http://r4ds.had.co.nz/index.html

- Read Welcome
- Read 1. Introduction
- Read 2. Introduction
- Go through 3. Data visualization (this takes some time but will introduce you to ggplot2)
- Read 27. Rmarkdown

Real data

```
# install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------- tidyverse 1.2.1 --
## <U+221A> ggplot2 3.0.0     <U+221A> purrr   0.2.4
## <U+221A> tibble  1.4.2     <U+221A> dplyr   0.7.4
## <U+221A> tidyr   0.8.1     <U+221A> stringr 1.3.1
## <U+221A> readr   1.1.1     <U+221A> forcats 0.3.0
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

By using the knowledge from above we will now work on a dataset

We first load the dataset into a variable we call "mammals"

```
library(tidyverse)

mammals = read_csv(file = "../datasets/dataset.01.rsbl20150010supp1.csv")
```

```
## Parsed with column specification:
## cols(
##   gene = col_character(),
##   Species = col_character(),
##   labs = col_character(),
##   chrMark = col_character(),
##   chr = col_character(),
##   dN = col_double(),
##   dS = col_double(),
##   dNdS = col_double(),
##   RPKM = col_double(),
##   Tau = col_double(),
##   GC3 = col_double()
## )
```

Q: Use dim(), names(), head() and summary() to inspect the dataset

```
mammals %>% dim()
```

```
## [1] 17327    11
```

```
mammals %>% names()
```

```
##  [1] "gene"    "Species" "labs"    "chrMark" "chr"     "dN"      "dS"
##  [8] "dNdS"    "RPKM"    "Tau"     "GC3"
```

```
mammals %>% summary()
```

```
##      gene             Species             labs
##  Length:17327       Length:17327       Length:17327
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##    chrMark              chr                  dN               dS
##  Length:17327       Length:17327       Min.   :  1.000   Min.   :  2.000
##  Class :character   Class :character   1st Qu.:  2.047   1st Qu.:  5.245
##  Mode  :character   Mode  :character   Median :  5.021   Median : 12.413
##                                        Mean   : 11.319   Mean   : 25.070
##                                        3rd Qu.: 12.105   3rd Qu.: 30.520
##                                        Max.   :305.379   Max.   :792.537
##       dNdS              RPKM               Tau              GC3
##  Min.   :0.01007   Min.   :     0.01   Min.   :0.07375   Min.   :0.2051
```

```
##  1st Qu.:0.24280   1st Qu.:    27.22   1st Qu.:0.50532   1st Qu.:0.4509
##  Median :0.44009   Median :    76.59   Median :0.68844   Median :0.5754
##  Mean   :0.56606   Mean   :   183.21   Mean   :0.67017   Mean   :0.5767
##  3rd Qu.:0.75126   3rd Qu.:   171.48   3rd Qu.:0.84840   3rd Qu.:0.6985
##  Max.   :6.28740   Max.   :127494.16   Max.   :1.00000   Max.   :0.9452
```

How many rows and columns are in the dataset? 12

Which columns are text? gene Species labs chrMark chr? chrtype

Which columns are numbers? dN dS dNdS RPKM Tau GC3

Identify how each column of the data corresponds to the variables described in the methods section of the paper

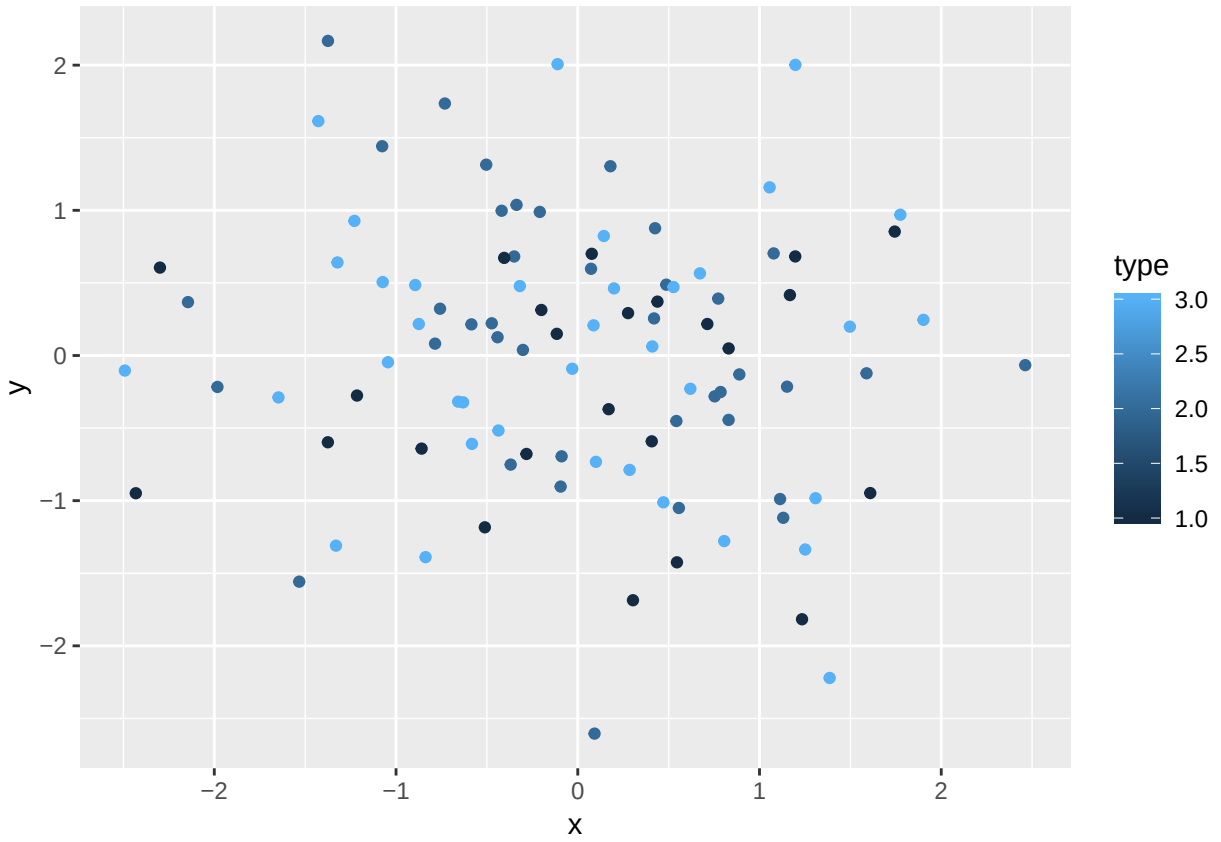
1.Expression Level ———> RPKM
2.Expression specificity —> Tau
3.GC content 3rd position -> GC3
4.Species ————————> Species
5.dN/dS ————————-> dNdS
6.Chromosome type ———> chrMark

Hint for keeping and saving a plot

```
plotdata = data.frame(x=rnorm(100), y=rnorm(100), type=sample(x = 1:3, size = 100, replace = T))

plot1    = ggplot(data = plotdata) +
  geom_point(mapping = aes(x = x, y = y, color=type))

plot(plot1)
```




```
ggsave(filename = "plot.week.01.first.plot.png", plot = plot1)
```

```
## Saving 6.5 x 4.5 in image
```

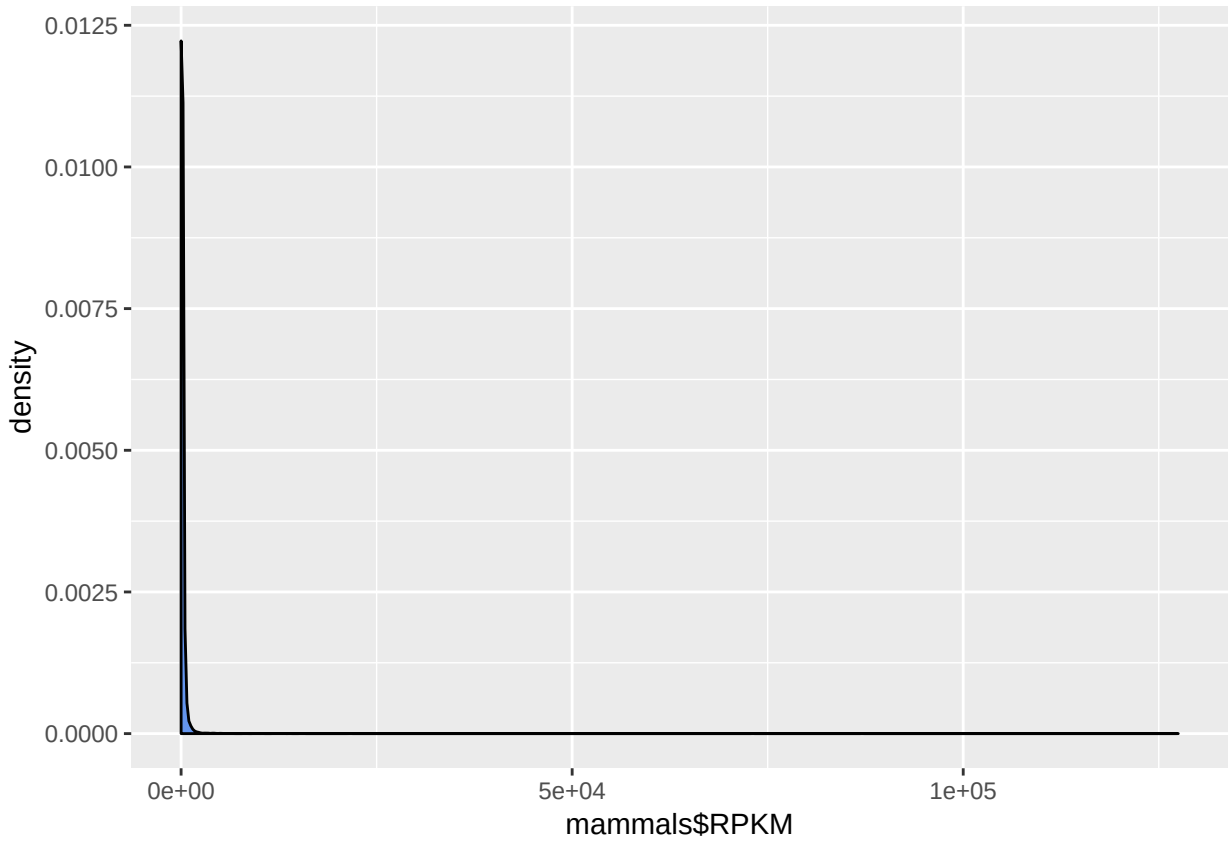
```
ggsave(filename = "plot.week.01.first.plot.pdf", plot = plot1)
```

```
## Saving 6.5 x 4.5 in image
```

Q: Vizualize the distribution of gene expression for all species

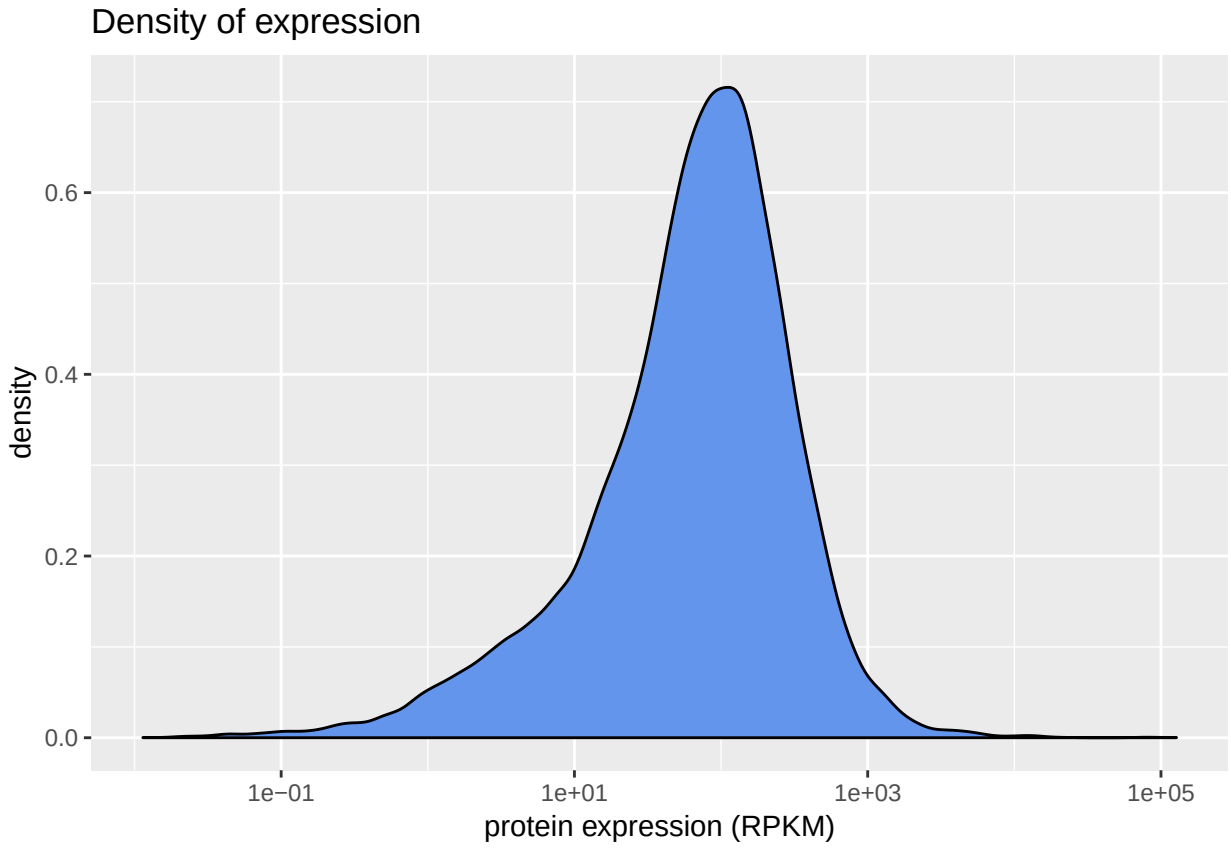
So basically, make a new code chunk and make a plot that shows the distribution of RPKM for all species in the same plot.

```
plotden1 = ggplot(mammals, aes(x = mammals$RPKM)) + geom_density(fill = "cornflowerblue")
plotden1
```



Q: Vizualize the distribution of gene expression for all species on a log scale

```
plotden1 + scale_x_log10() + labs(title = "Density of expression", x = "protein expression (RPKM)")
```



HINT: ?scale_x_log10()

A little trick to distinguish chromosome X from the other chromosomes (more on this later in the course)

We use a little simple trick to make a new variable that tells you if the chromosome is an autosome or sex chromosome.

```
mammals = mammals %>%
  mutate(chrtype =  ifelse(chr!="X", "Autosome", "Chromosome X")) %>%
  arrange(chrtype) %>%
  {.} #???
```

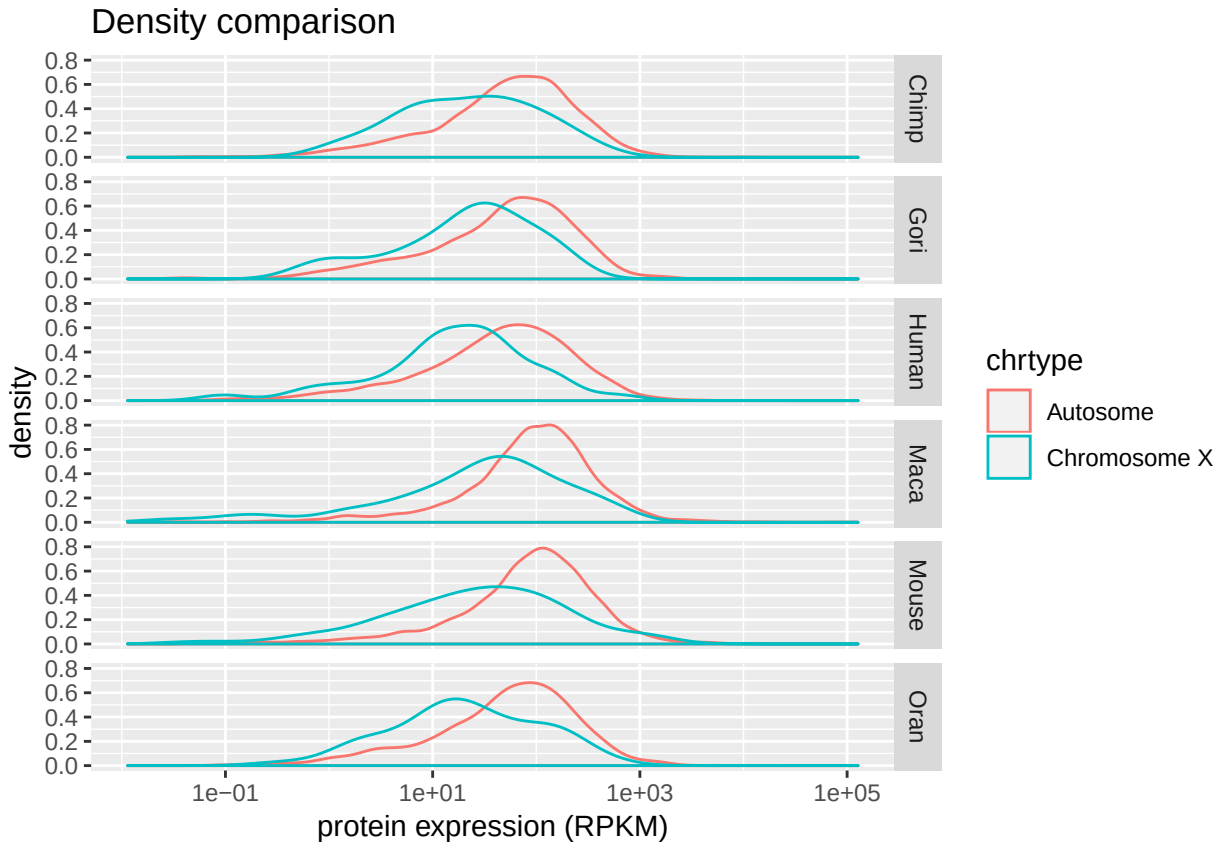
```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
head(mammals)
```

```
## # A tibble: 6 x 12
##   gene    Species labs  chrMark chr      dN    dS  dNdS   RPKM   Tau   GC3
##   <chr>   <chr>   <chr> <chr>   <chr> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1 ENSG00~ Chimp   Chim~ A       1      2.04  4.17 0.490 182    0.830 0.769
## 2 ENSG00~ Chimp   Chim~ A       1      2.08  2.61 0.797 258    0.698 0.910
## 3 ENSG00~ Chimp   Chim~ A       1      2.58  6.36 0.406 157    0.418 0.866
## 4 ENSG00~ Chimp   Chim~ A       1     10.3  14.7  0.701   4.58 0.734 0.837
## 5 ENSG00~ Chimp   Chim~ A       1      2.34  4.75 0.493 107    0.639 0.757
## 6 ENSG00~ Chimp   Chim~ A       1      2.19  9.43 0.232 249    0.333 0.859
## # ... with 1 more variable: chrtype <chr>
```

Q: Vizualize the distribution of gene expression for all species on a log scale, but plot each species in its own subplot

```
ggplot(mammals, aes(x = RPKM)) + scale_x_log10() + geom_density(aes(color = chrtype)) + facet_grid(Spec
```



HINT: ?facet_grid
HINT: ?facet_wrap

```
?facet_wrap
```

Q: do a scatter plot of gene expression (log scale) against dNdS, color by chromosome type, facet by species

```
ggplot(mammals, aes(x = RPKM, y = dNdS, color = chrtype)) + geom_point(size = 0.3) + scale_x_log10() + 
```


Scatterplot
Chimp
Gori
Human
Maca
Mouse
Oran
dNdS
0
2
4
6
1e−01
1e+01
1e+03
1e+05
RPKM
chrtype
Autosome
Chromosome X