Introduction to Data Science - Fall 2019

Mid-term assignment

Mads Emil Marker Jungersen, studienummer: 201906249

Oct 11, 2019

Introduction

In this report, I will provide a step by step analysis of the "mammals" dataset. By doing so I will reproduce a fraction of the analysis that is behind the paper published in 2015 and entitled.

"Gene expression, chromosome heterogeneity and the fast-X effect in mammals", published by Nguyen et al in the journal *Biology Letters* in 2015:

https://doi.org/10.1098/rsbl.2015.0010

This will be the occasion to mobilize all the know how in R programming and rmarkdown, that I have accumulated in first period of the course.

Importing on the mammals dataset

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

Q1: Make a table summarizing the information available for each species

Make a new dataset **mammals_overview**. The dataset **mammals_overview** should contain as columns:

- The species name
- Variable chrMark ("A" for Autosomes and "X" for X chromsomes)
- Sample size n_genes for each sub group defined by the variable chrMark

Then **format** the dataset so you can include it as a table for your report. The table should provide for each species an overview of the number of genes (sample size) availablefor autosomes and X genes:

Firstly we group our dataset by 'Species' and 'chrMark', because these are the two variables that we are interested in. We combine the 'group_by()' function with the 'summarise()' function to create a new variable 'Sample_space' counting the genes in 'Species' and 'chrMark'. We display the top 5 rows of our new dataset using the 'head()' function, and we combine the 'head()' function with the 'kable()' function to format our display of the dataframe in a nice way.

```r
mammals_overview <- mammals %>%
  group_by(Species, chrMark) %>%
  summarise(Sample_size = n())

mammals_overview %>%
  head %>%
   kable(col.names = c("Species", "Gene Location", "Sample size"))
```

Species	Gene Location	Sample size
Chimp	A	1876
Chimp	X	32
Gori	A	1758
Gori	X	38
Human	A	1843
Human	X	37

We see that the variable 'Gene Location' in our dataset **mammals_overview** actually contains the two variables ("A" for Autosomes and "X" for X chromsomes). We can 'fix' this by using the 'pivot_wider()' function as follows:

```r
mammals_overview_wider <- mammals_overview %>%
  pivot_wider(names_from = chrMark, values_from = Sample_size)

mammals_overview_wider %>%
  kable(col.names = c("Species", "Sample Size for A", "Sample size for X"))
```

Species	Sample Size for A	Sample size for X
Chimp	1876	32
Gori	1758	38
Human	1843	37
Maca	4006	98
Mouse	4347	107
Oran	3115	70

Q2: Reproducing the main figure of the Biology Letters article

In the past week, we have made summaries of the dataset by grouping genes per Species and chromosome. So you can reuse your code - or borrow the solution- to build a dataset **mammalsmeans** containing means of gene expression and dN/dS for these subgroups.

Then "ggplot()" **mammalsmeans** to do the scatter plot reproducing the figure of the original paper.

And relative to the original publication, you can even improving it by making the size of every data point (a chromsome) reflect how many genes are on the chromsome.

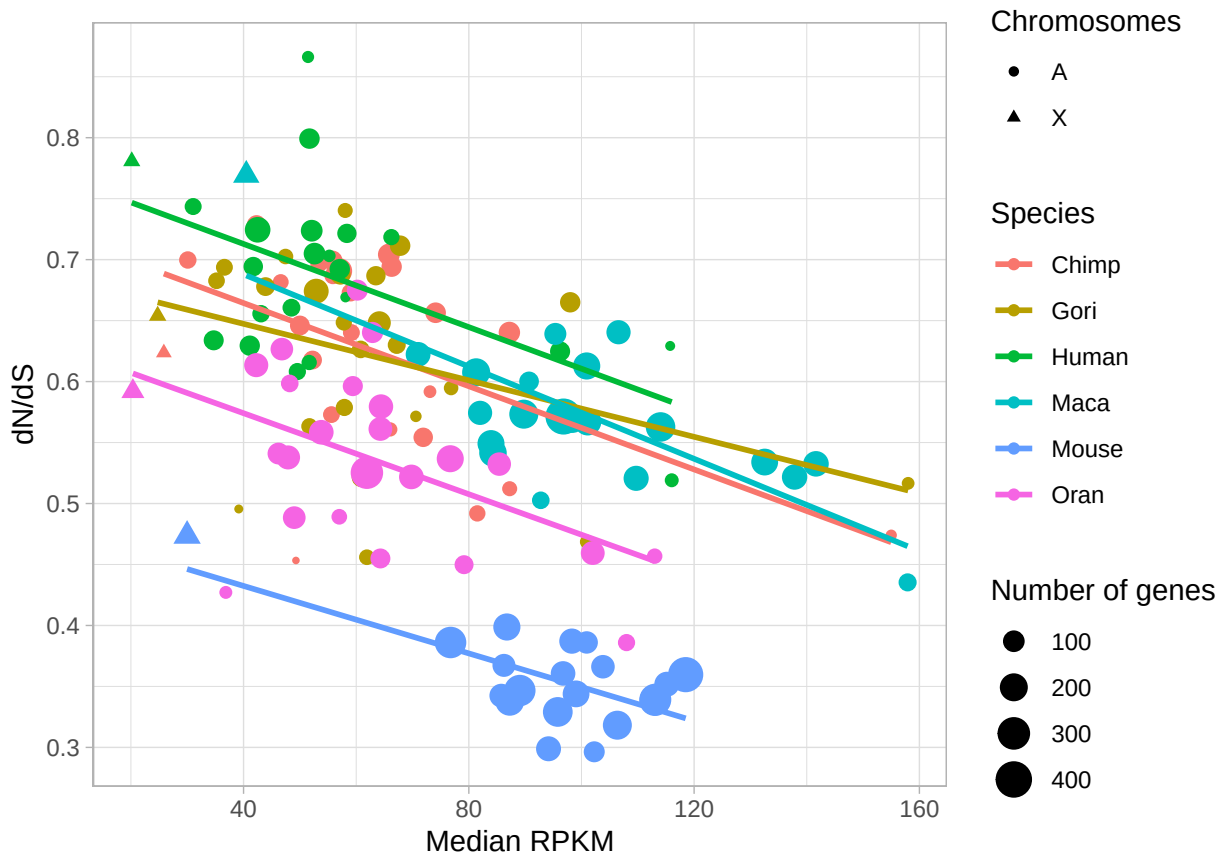
We start off by creating the **mammalsmeans** dataset by first grouping the mammals dataset by 'Species', 'chr' and 'chrMark'. We are using the 'summarise()' function to calculate the mean of dN and dS, the median of RPKM, the ratio between mean_dN and mean_dS and finally the number of genes using the 'n()' function. Afterwards we display the top 5 rows of our dataset using the 'head()' function formatted with the 'kable()' function.

```
mammalsmeans <- mammals %>%
  group_by(Species, chr, chrMark) %>%
  summarise(mean_dN = mean(dN),
            mean_dS = mean(dS),
            median_RPKM = median(RPKM),
            mean_dN_dS = mean_dN/mean_dS,
            n_genes = n())

mammalsmeans %>%
  head %>%
  kable(digits = 2)
```

Species	chr	chrMark	mean_dN	mean_dS	median_RPKM	mean_dN_dS	n_genes
Chimp	1	A	4.38	6.34	57.00	0.69	193
Chimp	10	A	4.71	6.78	66.30	0.69	107
Chimp	11	A	3.66	5.23	55.80	0.70	102
Chimp	12	A	3.88	6.29	52.25	0.62	88
Chimp	13	A	4.06	7.24	66.05	0.56	46
Chimp	14	A	4.03	5.87	55.85	0.69	66

To compute the plot we pipe our newly created dataset **mammalsmeans** with the 'ggplot()' function using the aesthetics x = median_RPKM, y = mean_dN_dS and color = Species. On top of the 'ggplot()' function we add a layer using 'geom_point()' using the aesthetics size = n_genes to get size of the points as a function of the number of genes and shape = chrMark to differentiate between autosomes and x chromosomes. We also add a layer of 'geom_smooth()' using the linear method without showing confidence intervals around the line. Lastly we are formatting the layout with 'theme_light()' function and editing the labels for both the x and y axis and also the name of the legends.

```
mammalsmeans %>%
  ggplot(mapping = aes(x= median_RPKM, y = mean_dN_dS, color = Species))+
  geom_point(mapping = aes(size = n_genes, shape = chrMark), show.legend = T)+
  geom_smooth(method = "lm", se = F)+
  theme_light()+
  labs(x= "Median RPKM", y="dN/dS", shape = "Chromosomes", size = "Number of genes")+
  NULL
```



Q3: Comparing Human versus Chimpanzee rates of evolution

Background and motivation

Since Humans and Chimpanzees parted their way from their common ancestor, they have evolved independently. Remember: The dN/dS ratio measures how fast genes of their respective genomes have evolved. It makes sense to make such comparison chromsome by chromsome as we have an identical chromsome structure (the only difference at the chromosome level between human and chimpanzees, is one fusion of chromosomes called 2A and 2B in chimpanzees and correspond as fused entities to chr2 in humans).

Q3.1: Build the right summary dataset

We want to make a scatter plot as a visual that contrasts mean rate of evolution for genes chromsome by chromosome in both human vs. chimpanzees. To do so you need to reformat the data as a new dataset called *human_chimp_df* so you can ggplot it. We outline the steps on how to do so below.

Steps to answer Q3:

- To build *human_chimp_df* , you need to only consider these two species only and filter out the remaining species.
- You also should exclude chromosomes "2, 2A and 2B" . All other chromosomes are conserved between humans and chimpamnzees (check that the genes number are quite close on each chromosome)
- then you should "broaden your dataset" (see class examples week 05) so you can get a dataset where for each chromosome, you have matching columns with mean dN/dS on human and chimpanzee.

We start by using the 'filter()' function to fisrt only look at the species Human and Chimpanzee and to filter out the chromosomes '2', '2A' and '2B' because these chromosomes differ in humans and chimpanzees. Afterwards we select the variables that we are interested in, we do that by selecting everything except what is in between chrMark and median_RPKM (both inclusive) using the 'select()' function. In the end we use the 'pivot_wider()' function to spread out human and chimpanzee with values from mean_dN_dS and n_genes. As always we display the top 5 rows formatted with the 'kable()' function.

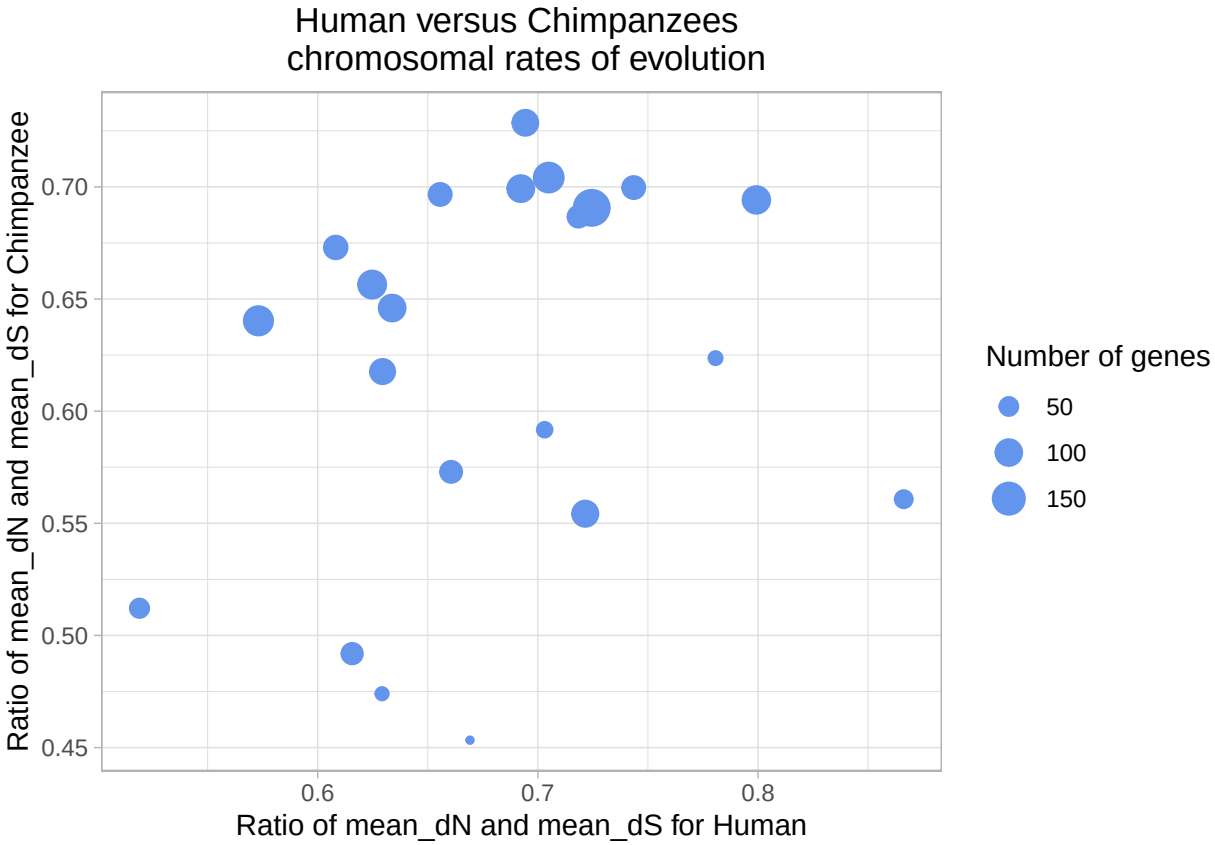
```
human_chimp_df <- mammalsmeans %>%
  filter(Species %in% c("Chimp", "Human"), !chr %in% c("2", "2A", "2B")) %>%
  select(-(chrMark:median_RPKM)) %>%
  pivot_wider(names_from = Species, values_from = c(mean_dN_dS,n_genes))

human_chimp_df %>%
  head() %>%
  kable(digits = 2)
```

chr	mean_dN_dS_Chimp	mean_dN_dS_Human	n_genes_Chimp	n_genes_Human
1	0.69	0.72	193	191
10	0.69	0.80	107	109
11	0.70	0.69	102	108
12	0.62	0.63	88	107
13	0.56	0.87	46	34
14	0.69	0.72	66	62

Q3.2: Make a scatterplot contrasting Human versus Chimpanzee chromosomal rates of evolution

To compute the scatterplot we are using the 'ggplot()' function with aesthetics x = mean_dN_dS_Human and y = mean_dN_dS_Chimp and size = n_genes_Chimp. We have used the size = n_genes_Chimp, because the number of genes in chimpanzees are almost the same as the number of genes in humans. Then we add a layer of points using 'geom_point()'. In the end we are changing the label of the legend.

```r
human_chimp_df %>%
  ggplot(mapping = aes(x = mean_dN_dS_Human, y = mean_dN_dS_Chimp,
                       size = n_genes_Chimp))+
  geom_point(color = "cornflowerblue")+
  labs(size = "Number of genes",
       x= "Ratio of mean_dN and mean_dS for Human",
       y="Ratio of mean_dN and mean_dS for Chimpanzee",
       title="Human versus Chimpanzees \n chromosomal rates of evolution")+
  theme_light()+
  theme(plot.title = element_text(hjust = 0.5))+ #centering title
  NULL
```



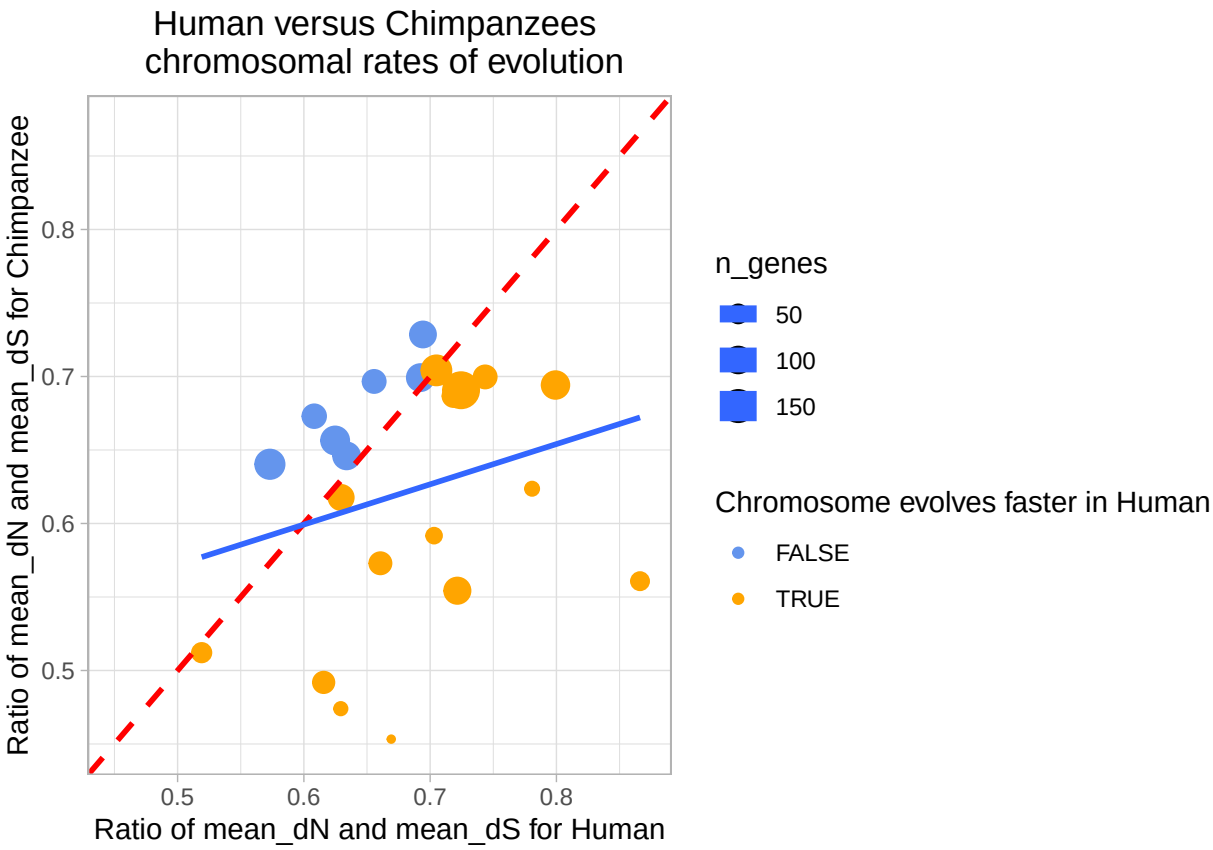
Q3.3: Write a few lines where you comment and interpret the results shown in the plot.

In order to better interpret my results in the scatterplot, I am going to make a few changes to the plot where I use the 'mutate()' function to create a new column containg True if the human chromosome evolves faster than the chimpanzee chromosome i.e. mean_dN_dS_Human > mean_dN_dS_Chimp, otherwise it will return False wich corresponds to the fact that the chimpanzee chromosome evolves faster than the human chromosome.

In the new scatterplot I have added a layer with the 'geom_abline()' function with a slope of 1 and an interception of 0. The reason for the line is to see how many chromosomes in chimpanzees evolve faster than the chromosomes in human. The points above the line means that the chromosome is evolving faster in chimpanzees, and points below the line corresponds to chromosomes that are evolving faster in human. In order to see it more clearly, I have also added a color aesthetic to the 'geom_point()' function, which corresponds to the new created column called 'human_faster_than_chimp'. I use the 'scale_color_manual()'

function to set the colors to use in the color aesthetic. In the end i have set two identical x, and y-axises which makes it easier to see any correlation or trend. We also add a trend line, using the linear method in the 'geom_smooth()' function.

```
human_chimp_df %>%
  mutate(human_faster_than_chimp = mean_dN_dS_Human > mean_dN_dS_Chimp) %>%
  ggplot(mapping = aes(x = mean_dN_dS_Human, y = mean_dN_dS_Chimp,
                       size = n_genes_Chimp))+
  geom_point(mapping = aes(color = human_faster_than_chimp))+
  geom_abline(slope = 1, intercept = 0, color="red", linetype = 2, size = 1 )+
  geom_smooth(method = lm, se=FALSE)+
  labs(size = "n_genes",
       x= "Ratio of mean_dN and mean_dS for Human",
       y="Ratio of mean_dN and mean_dS for Chimpanzee",
       color = "Chromosome evolves faster in Human",
       title = "Human versus Chimpanzees \n chromosomal rates of evolution")+
  xlim(0.45,0.87)+
  ylim(0.45,0.87)+
  scale_color_manual(values = c("cornflowerblue","orange"))+
  theme_light()+
  theme(plot.title = element_text(hjust = 0.5))+ #centering title
  NULL
```



Is there a trend in the data ? How many chromosomes seem to evolve faster in chimpanzee (higher dN/dS) Looking at the newly created scatterplot it seems that there is an slightly upgoing trend, which means that if a chromosome is evolving fast in humans (high ratio of mean_dN/mean_dS) it also seems to be relative fast in chimpanzees. Also it is clear from the plot that 7 chromosomes is evolving faster

in chimpanzees while the remaining 15 chromosomes are evolving faster in humans. We get these numbers by counting the points on the plot, but we could also find the number by using the following R-code.

```
human_chimp_df %>%
  mutate(human_faster_than_chimp = mean_dN_dS_Human > mean_dN_dS_Chimp) %>%
  filter(human_faster_than_chimp == TRUE) %>%
  group_by(human_faster_than_chimp)%>%
  summarise(n())
```

```
## # A tibble: 1 x 2
##   human_faster_than_chimp `n()`
##   <lgl>                   <int>
## 1 TRUE                       15
```

Are the rates of evolution in both species quite correlated?

To get a feeling of how correlated the rates of evolution are, we are bulding a simple linear regression model, which actually is the same as the 'geom_smooth()' function with method 'lm' does, but in this case we are interested in some quantified value explaining this.

```
linear_model <- lm(data = human_chimp_df, mean_dN_dS_Chimp ~ mean_dN_dS_Human)
```

After building the model, we can call 'summary()' function:

```
summary(linear_model)
```

```
##
## Call:
## lm(formula = mean_dN_dS_Chimp ~ mean_dN_dS_Human, data = human_chimp_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16491 -0.05947  0.03891  0.06017  0.10348
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.4350     0.1579   2.755   0.0122 *
## mean_dN_dS_Human   0.2737     0.2306   1.187   0.2492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08244 on 20 degrees of freedom
## Multiple R-squared:  0.06581,    Adjusted R-squared:  0.0191
## F-statistic: 1.409 on 1 and 20 DF,  p-value: 0.2492
```

We are primarely interested in the R-squared value. In this case the R-squared value is 0.066 which means that linear model are able to explain around 6% of the variation in the plot. Eventhough the R-squared value is quite low, there still is some kind of correlation as also explained earlier, that if one chromosome is evolvinger fast in one of the species, it is also evolving relatively fast in the other species.

Is one species evolving faster since speciation?

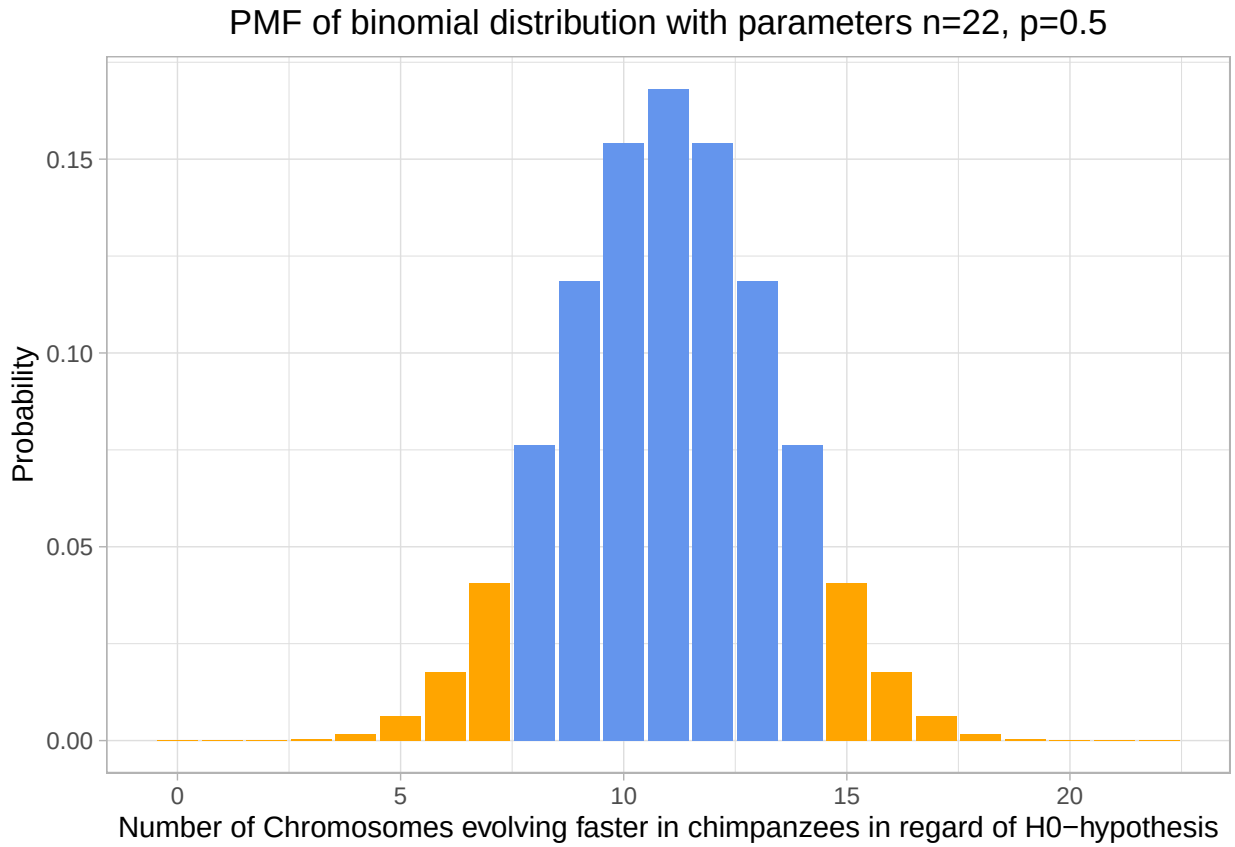
On a chromosome level it seems that humans are evolving a little bit faster than chimpanzees given the fact that 15 chromosomes are evolving faster in humans, while 'only' 7 chromosomes are evolving faster in chimpanzees.

Q3.4 Is the number of chromosomes evolving faster in chimpanzee higher than expected "by chance"

If species are evolving at the same rate "overall", we expect – by a symmetry argument – that about 1/2 of the chromosomes are faster in chimpanzees . . . so use a binomial distribution for modelling as stochastic variable the number of chromosomes that evolve faster in one species. Calculate the probability of observing "just by chance" a pattern as extreme or more extreme than the one you can see on the plot.

We start by creating the H0-hypothesis, which is assuming that human and chimpanzees evolve at the same rate overall. In regard of H0, we expect that about 1/2 of the chromosomes evolves faster in chimpanzees. We use the binomial distribution to model a stochastic variable of the number of chromosomes evolving faster in chimpanzees. The parameters in the binomial distributin is 22 for the number of chromosomes and 0.5 as the probability of succes i.e a chromose is evolving faster in chimpanzees. From Q3.3 we saw that 7 chromosomes were evelving faster in Chimpanzees, whereas the remaining 15 chromoses were evolving faster in Humans. What we saw/observed in the plot corresponds to the orange bins in the below plot of the binomial distribution with parameters 22 and 0.5. The below Probability Mass function of the binomial distribution is created by first making a tibble with two columns. The first column 'n_faster' is a vector containing the numbers from 0 to 22, whereas the second column contains the probability of observing each of the numbers in the vector with the parameters 22 and 0.5 in the binomial distribution. After creating the tibble, i create a new column called 'obs' using the 'mutate()' function. The new column contains True if the number in the vector is less than or equal to 7 or greather than or equal to 15. I use the pipe operator to call the 'ggplot()' function on top of my tibble with aesthetics x=n_faster, y=SS and fill = obs. Then I add a layer of columns using 'geom_col()' function. In the end I format the plot using 'theme_light()', and controlling the colors used in my fill aesthetic using the 'scale_fill_manual()' function.

```r
chr_binom <- tibble(n_faster = c(0:22), SS = dbinom(c(0:22),22,0.5))
chr_binom %>%
  mutate(obs = n_faster <= 7 | n_faster >= 15)%>%
  ggplot(mapping = aes(x=n_faster, y=SS, fill=obs))+
  geom_col(show.legend = F)+
  scale_fill_manual(values = c("cornflowerblue", "orange"))+
  theme_light()+
  labs(x="Number of Chromosomes evolving faster in chimpanzees in regard of H0-hypothesis",
       y="Probability",
       title="PMF of binomial distribution with parameters n=22, p=0.5")+
  theme(plot.title = element_text(hjust = 0.5))+ #centering title
  NULL
```



We are now going to quntify how extreme that observation is relative to our H0-hypothesis. We do that by calculating the probability that observations coming from H0 could be just as extreme or more extreme as what we actually observed. Wich means that we based on our H0-hypothesis are calculating the probability of observing at least 7 chromosomes evolving faster in Chimpnzees, we do that by using the cumulative density function:

```r
pbinom(7,22,1/2)
```

```
## [1] 0.06690025
```

We see that the probability is around 7%, which means that there is a probability of around 7% of observing something just as extreme or more extreme than what we obersed, assuming the H0-hypothesis to be true. It furthermore means that obeserving 7 chromosomes evolving faster in chimpanzees are still within the 95% range of the distribution of our H0-hypothesis. Because of the fact that obeserving 7 chromosomes evolving

faster in chimpanzees are still within the 95% range of what we expect based on our H0-hypothesis, I don't believe there is any reason to reject our H0-hypothesis. Another way to calculate probability of observing "just by chance" a pattern as extreme or more extreme than the one you can see on the plot, is to calculate the probability of observing 15 or more chromosomes evolving faster in humans, which will return the same result:

```
1-pbinom(14,22,1/2)
```

```
## [1] 0.06690025
```

Reproduce the work

To be able to reproduce my work in this report, and to make sure people know which R environment I have used to do my analaysis, i run the following R code as the final remark.

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 17134)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.25      forcats_0.4.0   stringr_1.4.0   dplyr_0.8.3
##  [5] purrr_0.3.2     readr_1.3.1     tidyr_1.0.0     tibble_2.1.3
##  [9] ggplot2_3.2.1   tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_0.2.5 xfun_0.9         haven_2.1.1      lattice_0.20-38
##  [5] colorspace_1.4-1 vctrs_0.2.0      generics_0.0.2   htmltools_0.3.6
##  [9] yaml_2.2.0       utf8_1.1.4       rlang_0.4.0      pillar_1.4.2
## [13] glue_1.3.1       withr_2.1.2      modelr_0.1.5     readxl_1.3.1
## [17] lifecycle_0.1.0  munsell_0.5.0    gtable_0.3.0     cellranger_1.1.0
## [21] rvest_0.3.4      evaluate_0.14    labeling_0.3     fansi_0.4.0
## [25] highr_0.8        broom_0.5.2      Rcpp_1.0.2       scales_1.0.0
## [29] backports_1.1.4  jsonlite_1.6     hms_0.5.1        digest_0.6.21
## [33] stringi_1.4.3    grid_3.6.1       cli_1.1.0        tools_3.6.1
## [37] magrittr_1.5     lazyeval_0.2.2   crayon_1.3.4     pkgconfig_2.0.3
## [41] zeallot_0.1.0    xml2_1.2.2       lubridate_1.7.4  assertthat_0.2.1
## [45] rmarkdown_1.15   httr_1.4.1       rstudioapi_0.10  R6_2.4.0
## [49] nlme_3.1-140     compiler_3.6.1
```