Exercise 1 *(Reading assignment)* © R Olayo Alarcon, CL Müller

As outlined in the lectures, a substantial part of the course material is covered by the online book "Modern Statistics for Modern Biology" by Susan Holmes and Wolfgang Huber.
The book is available at `https://www.huber.embl.de/msmb/`. Familiarize yourself with the overall structure of the book (also referred to as MSMB) and read the chapters Introduction and Chapter 1.

a) Check Figure 1 in the Introduction. Do you think modern biology still fits this framework? Think about arguments for and against the described paradigm.

b) Review Tukey's approach to scientific discovery and the term EDA. What does EDA stand for? Given your previous experience with data, what are some of the tools you already know that could qualify as EDA tools?

c) What does the term **large**-p-**small**-n problem mean? Give examples of biological data from the lecture where we encountered these types of problems.

d) Install and familiarize yourself with RStudio and recap some of *base* R functionality. `https://www.dataquest.io/blog/tutorial-getting-started-with-r-and-rstudio/`

Exercise 2 *(Generative models in R)* © R Olayo Alarcon, CL Müller, MSMB

In MSMB Chapter 1, you have familiarized yourself with (or recapitulated) the concept of generative models for discrete data in R. In particular, we learned about the Poisson, the Binomial, and the Multinomial distribution.

a) Check again the lecture slides and recapitulate what types of biological data you have encountered so far. Give two biological data examples and a corresponding scientific question where you think that any of the above probability distributions is useful.

b) Do MSMB Exercise 1.1 (below the text for completeness).
R can generate numbers from all known distributions. We now know how to generate random discrete data using the specialized R functions tailored for each type of distribution. We use the functions that start with an `r` as in `rXXXX`, where `XXXX` could be `pois`, `binom`, `multinom`. If we need a theoretical computation of a probability under one of these models, we use the functions `dXXXX`, such as `dbinom`, which computes the probabilities of events in the discrete binomial distribution, and `dnorm`, which computes the probability density function for the continuous normal distribution. When computing tail probabilities such as $P(X > a)$, it is convenient to use the cumulative distribution functions, which are called `pXXXX`. Find two other discrete distributions that could replace the `XXXX` above.

Exercise 3 *(C. elegans nucleotide frequencies)* © R Olayo Alarcon, CL Müller, MSMB

In MSMB Chapter 1, we covered the multinomial distribution. We are now using the distribution to model nucleotide frequencies in a real genome, the mitochondrial sequence of the worm C. elegans. This exercise is adapted from MSMB Exercise 1.8.
This is our opportunity to use Bioconductor for the first time. Since Bioconductor's package management is more tightly controlled than CRAN's, we need to use a special install function (from the BiocManager package) to install Bioconductor packages:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(c("Biostrings", "BSgenome.Celegans.UCSC.ce2"))
```

After that, we can load the genome sequence package as we load any other R packages.
The question we want to ask is whether the mitochondrial sequence of C. elegans is consistent with a model of equally likely nucleotides?

a) Explore the nucleotide frequencies of chromosome M by using a dedicated function in the Biostrings package from Bioconductor.

b) Test whether the C. elegans data is consistent with the uniform model (all nucleotide frequencies the same) using a simulation.

Exercise 4 *(Displaying GC content in S. aureus)* © R Olayo Alarcon, CL Müller, MSMB

In the lecture, we learned about nucleotide frequencies in genomes. Now let's look at the GC content in a real bacterial genome in MSMB Exercise 2.4.
Display GC content in a running window along the sequence of Staphylococcus Aureus. Read in a fasta file sequence from a file. This file is avalable in the data folder associated with the MSMB book.

```
staph = readDNAStringSet("../data/staphsequence.ffn.txt", "fasta")
```

a) Look at the complete staph object and then display the first three sequences in the set.

b) Find the GC content in tsequence windows of width 100.

c) Display the GC content in a sliding window as a fraction.

d) How could we visualize the overall trends of these proportions along the sequence?