Neuro-Genomics
Exercise 1 - Principles of Sequencing Data Analysis

The purpose of this exercise is to learn the principles of sequencing data analysis. We will first discuss the importance of sequencing data to modern biology and the expected data structure (Part 1), then we will use R to inspect actual sequencing data (Part 2). We will use descriptive statistics to explore aspects of the data and consider some of the ways to analyze it (Part 3). Next, we will detect differentially expressed genes (Part 4). Then we will analyze time-course sequencing experiments and use the Fourier Transform to detect rhythmic patterns in gene expression (Part 5). Finally, we will integrate a few of the concepts to detect genes with variable expression levels (Part 6).

General instructions to all sections:

a) We expect some prior knowledge in R and in programming in general. Please consult us if this is not the case. In addition, vectorization in R is an important concept to be familiar with:
Vectorized Operations | R Programming for Data Science
b) The symbol ## marks tasks to perform
The symbol # marks comments related to the way the task should be performed
c) For each computational task write the lines of code that you use below the actual task and paste the output as well
d) For tasks that don't require code write your answer below the actual task. Trivial tasks, like installation of packages, don't require a reply.
e) If you don't know the R functions to perform specific tasks, use the web to find the appropriate R functions. For example, R Documentation and manuals are a great source.
f) This exercise can also be done in Matlab. The 'pasilla' data of Part 2 can be found here.
g) Solving most of this exercise in python is also possible, but it is less recommended since not all the functions we use here have one-to-one equivalents in python.

Part 1 - General introduction to sequencing data

Over the last decade next-generation sequencing of RNA has revolutionized molecular biology by starting to provide the molecular basis behind the function of tissues. The idea is simple, we can collect any tissue that we are interested in, extract the RNA from it, and then sequence the extracted RNA using a sequencing machine. [Note: typically only the mRNA are sequenced, using poly(A) enrichment of the mRNA. However, non-poly(A) transcripts can also be sequenced, for example mature microRNA sequencing, and therefore below we will use the term 'RNA' in the context of sequencing].
The output is millions of sequencing fragments ("reads") of the RNA molecules. These reads can then be aligned against the genome and the genes from which they were transcribed can be detected. Therefore, the result is a comprehensive estimate of the expression levels (i.e.

number of RNA copies) for all the genes in the tissue. Here we won't focus on the raw sequencing data (the reads), nor will we discuss the alignment process. Instead, we will start exploring the sequencing data after it was processed into a vector with the following structure:

NameOfGene	Counts (the number of reads that align to the specific gene)
Gene1	23
Gene2	7
…	
GeneN	452

Where N is the number of genes in the organism in question.
This expression vector typically represents the information (RNA profile) of one tissue sample for one particular experimental condition. However, the sample may be only a small part of a tissue, several tissues combined or even a small organism (for example, fish larvae). The sample can also be from cells grown outside of a tissue (that is, cultured cells or cell lines).

Read about the leading next-generation sequencing technology, focus on Figure 3 and the process of 'bridge amplification', which is explained in the Glossary section:
An Introduction to Next-Generation Sequencing Technology
(note: Figure 3 starts from double stranded genomic DNA, but the process is exactly the same for RNA as well, the only difference is that the starting point is double stranded cDNA)
In Illumina sequencing, both the library preparation and the bridge amplification steps contain PCR amplification. However, PCR can induce distortions, read:
Sources of PCR-induced distortions in high-throughput sequencing data sets
Focus on the abstract section, and the beginning of each one of the following results sections: 'Perfect PCR', 'PCR bias' and 'Stochastic amplification of low copy number amplicons'.
Can you explain why we can only get an **estimation** of the expression levels and not the actual number of RNA molecules for each gene?

Part 2 - Explore sequencing data using R

In this section we will use a dataset from an experiment performed on Drosophila melanogaster cell cultures. The experiment was designed to investigate the effect of RNAi knock-down of the splicing factor pasilla (see Conservation of an RNA regulatory map between Drosophila and mammals). Several samples were collected from untreated fly cells (i.e. control samples), and several samples were collected from fly cells that were treated with RNAi. The dataset is stored in the package 'pasilla'.
In most sequencing experiments, including this one, more than one sample is examined. This allows the detection of genes that have different expression levels between the tested samples (more about that in the next section). However, when testing more than one sample, how can we reliably compare the different expression vectors? In technical terms, the different expression vectors need to be normalized. The easiest normalization method is to account for the fact that different samples can produce different total number of sequencing reads.
To illustrate this point consider the following example: say that RNA was separately extracted

from two samples, one of diseased tissue and one of control. The same amount of RNA was loaded into the sequencing machine and sequenced. In principle, the sequencing machine should get a fixed amount X of RNA material as input and produce a fixed amount Y of sequencing reads (X and Y varies between the different types of machines). In practice however the total number of reads varies between different machine runs. Suppose that the first sample in our example generated 20 million reads and the second one generated 10 million reads. One might falsely conclude that many of the genes in the first sample have higher expression levels compared to the second sample. However, a simple normalization which enforces the constraint that the total number of reads to be the same for each run will allow a more valid comparison of the two samples.

```
## Install the package 'pasilla'
# copy and paste the lines below into R (if prompted - choose ‘Update all’)
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("pasilla")

## Use the installed library
# copy and paste the line below into R
library("pasilla")

## Load the actual data from the pasilla package
# copy and paste the lines below into R
pasCts <- system.file("extdata",
                      "pasilla_gene_counts.tsv",
                      package="pasilla", mustWork=TRUE)
pasAnno <- system.file("extdata",
                       "pasilla_sample_annotation.csv",
                       package="pasilla", mustWork=TRUE)
cts <- as.matrix(read.csv(pasCts,sep="\t",row.names="gene_id"))
coldata <- read.csv(pasAnno, row.names=1)
coldata <- coldata[,c("condition","type")]
rownames(coldata) <- sub("fb", "", rownames(coldata))
cts <- cts[, rownames(coldata)]

## We will use the matrix cts. Examine the first 10 lines of this matrix - what kind of information is in the matrix?

## Define a variable that will hold the dimensions of this matrix and print the matrix dimensions


## Is the sum of reads the same for each one of the samples (the different columns)?
```

Create a normalized version of the cts matrix
multiply each one of the columns (except the first one) by a factor so that the total number of reads per column will be equal to the first column

Make sure that in the normalized matrix the sum of reads is the same in all samples

Part 3 - Basic statistics of sequencing data

A typical sequencing experiment, like the one that we are examining, is normally designed to address the following question - which genes have different expression levels when comparing two (or more) experimental conditions. However, detecting differentially expressed genes is not trivial. Consider the following example: say that we have two normalized expression vectors, one of condition A and one of condition B, and gene X has expression level 150 in condition A and 200 in condition B. How can we know if gene X is differentially expressed between the two conditions? One can conclude from the increase in counts that gene X is influenced by the conditions tested. However, another likely explanation is that the increase in counts is due to either technical variations (noise introduced by the sequencing procedure) or biological variations (for example, difference in expression levels due to the inherent stochasticity of gene expression). Therefore, technical repeats (samples generated from the same biologically material) and/or biological repeats (samples generated from biologically distinct material) are needed in each sequencing experiment. Repeats are essential, but unfortunately the number of repeats that can be performed is usually limited by experimental and financial considerations. Given that usually only a handful of repeats are performed for each condition, it's practically impossible to deduce the technical and biological variations in the expression level of each gene from the data alone. Therefore we need to model the data (here 'data' is defined as the expression levels of each gene in one experimental condition but with several repeats) using well-known distributions. Knowing the distribution will provide the expected variability in the expression level of each gene, and allow comparison of different experimental conditions. However, if we use a well-known distribution that doesn't really model the data, then our estimation about the expected variability will be wrong. In this section we will model the data using two well-known distributions.

Does the data fit a Poisson distribution?
recall that the variance of the Poisson distribution is equal to its mean
first calculate the variance and the mean in the expression of each gene for the untreated conditions
then log transform the variance and mean (because the resulting values span multiple orders of magnitude). Tip: add 1 to all the values to avoid log of 0.
plot the variance vs the mean. Add the line x=y to your plot. The line represents a perfect Poisson distribution

Does the data fit a Negative binomial distribution?
In Negative binomial distribution the relationship between the variance and the mean is: variance = mean + a*mean^2 where 'a' is the dispersion parameter, a measure of how dispersed are the data compared to the Poisson distribution
use the nls function to fit a curve of variance = mean + a*mean^2 to the log transformed variance and mean. Tip - set the initial guess of the value of a to 0.
plot the resulting fitted curve on the variance vs mean plot

What is the value of the dispersion parameter?

Perform the same analysis (including the plots) for the treated samples. What is the dispersion parameter?

Read about the statistics of sequencing experiments:
Why sequencing data is modeled as negative binomial
Judging by the results obtained, do you think that the different untreated and treated samples in this experiment are technical or biological repeats?

Part 4 - Detect differentially expressed genes

In this section we will first detect different expression levels between the two conditions (treated and untreated) using visual inspection. However, as discussed in the previous section, we can use the observation that the data fit a well-known distribution to estimate the expected variation in the expression level and to calculate the p-value that the gene is differentially expression between the two conditions. For that, we will use the R software DESeq, that will allow detection of all the differentially expressed genes between the treated and untreated conditions.

Using visual inspection, detect at least one gene that has different expression levels in the first treated sample compared to the first untreated sample
The first step is to plot the log of expression in one sample against the log of expression in the other sample (log transformation is used because the resulting values span multiple orders of magnitude)
Then use the obtained plot to visually locate gene(s) that has strikingly different expression level in the two conditions
Characterize the expression of the gene(s) you selected using a grid-like manner: for example, gene(s) that has expression levels higher than value X in the treated condition, and lower than expression Y in the untreated condition
Use the grid-like characterization to computationally detect the name of the gene(s)

Choose one of the genes you detected using visual inspection and plot the expression of this gene in all the conditions in this experiment. The x-axis should be a running index and the y-axis should be the expression of this gene in the three treated conditions and the four untreated

conditions, in the same order as in the ‘cts’ matrix. Is the expression of this gene consistently different between all the treated and untreated conditions?

read the abstract part of the manuscripts describing the R tool 'DEseq', which allows the detection of differentially expressed genes:
Differential expression analysis for sequence count data
Does the data we have fit the basic assumptions of DEseq?

```
## install the package 'DESeq2'
# copy and paste the lines below into R
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("DESeq2")

## use the installed library
# copy and paste the line below into R
library("DESeq2")

## use DESeq to detect the probability that each one of the genes is differentially expressed between the two conditions
# copy and paste the following lines into R
dds <- DESeqDataSetFromMatrix(countData = cts,
                              colData = coldata,
                              design = ~ condition)
dds <- DESeq(dds)
res <- results(dds)
res

## the matrix 'res' contains the p-value for all the genes - sort the genes according to the p-value, such that lower p-values will appear first

## detect 10 genes that are most different between the treated and untreated according to the obtained p-values. Display only the column ‘log2FoldChange’ (the log transformed fold change between the two conditions) and the column ‘padj’ (the adjusted p-value, see: Multiple comparisons problem) of these 10 genes. Is the gene that you detected using visual inspection among them?
```

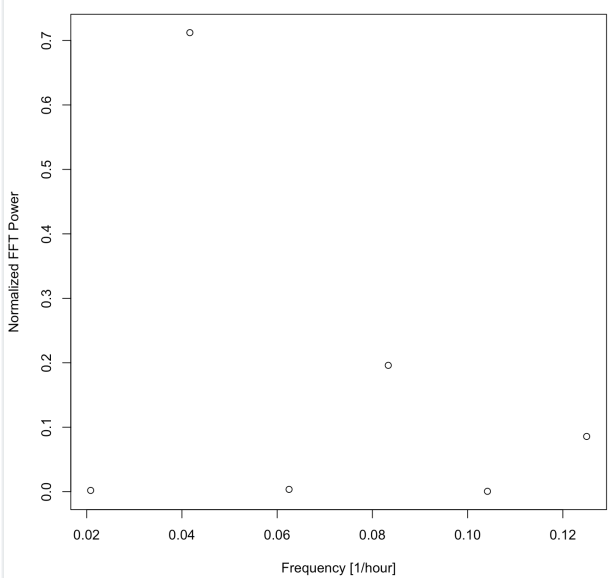
Part 5 - Detect rhythmical patterns using Fourier transform

In the previous sections we focused on differential genes expression (DGE), which is the classical type of analysis of RNA sequencing data. DGE can be framed as follows - given two experimental conditions A and B, each with one or more technical and biological repeats, detect a set of genes that have different expression levels in A compared to their expression levels in B. Naturally, this analysis (and any other kind of analysis for that matter), can't fit all experimental designs and questions.
In this section we will focus on time-course experiments. In these experiments a sample is collected at different points in time and the RNA is sequenced. This allows following the time-dependent behavior of gene expression on a genome-wide scale. These experiments are often technically challenging because a given sample can only be sequenced at one time point because extraction of the RNA for sequencing destroys the sample. This technical challenge is usually overcome using biological replicates: consider for example an experiment designed to follow the time-dependent gene expression following administration of a drug to a model animal - several animals can be treated with the drug simultaneously, and at each time point, one or more animals are sacrificed and the RNA from the tissue of interest is collected for sequencing.
The data analysis aspects of time-course experiments are also challenging. If the data contains some rhythmical patterns, then the problem can be addressed with Fast Fourier transform. This is the case with Circadian rhythms. While physiological examples of circadian rhythms have been known for hundreds of years, and the key genes that make up the core circadian clock have been known for almost 50 years, only over the last decade has next-generation sequencing of RNA revealed the full extent of circadian rhythms; it turns out that cellular circadian clocks are present in most (if not all) cell types, tissues and organisms. The cellular circadian clocks in turn cause circadian rhythms in hundreds of genes, which relate to multiple physiological pathways.
We will explore RNA sequencing data generated in zebrafish. This is a time-course experiment designed to identify genes with circadian expression patterns. We will start by visual exploration of the data in the time domain, and then move to the frequency domain to process the data.

Load the file CircadianRNAseq.csv which contains the processed RNA sequencing data into R. Examine the last 5 rows of the matrix, what is the time step (i.e. the time in hours between each measurement)?
download the file and store it locally in the computer
use the function 'read.csv' to read this csv file
use the function 'as.matrix' so that the dataset will be read as a matrix
Additional information regarding the data:
Each row represents one gene. The first column 'RefSeqID' gives the official gene annotation, and the last column 'GeneSymbol' gives the common name of the gene. Columns 2 to 13 contain the time-course measurements which were performed over two consecutive days ('A' means day 1 and 'B' is day 2). Each value in columns 2 to 13 represents the number of sequencing reads that were aligned against the gene in a given row. However, these values are not raw count data - these are normalized count values (therefore - no further normalization is needed).

Plot the expression of the gene 'per1a' at all the time points. The x-axis labels should be all the time points ("A_11PM"...). In the plot show both the data points and a line that connects them. Does the expression of this gene seem circadian? Read about the gene Per1 - should this gene be circadian?
first detect the location of this gene in the last column (‘GeneSymbol’) of the matrix
use xaxt="n" in the ‘plot’ function to remove the default running index in the x-axis
then use the function ‘axis’ to define the x-axis labels

We will convert the data from the time domain to the frequency domain to determine if the expression profile indeed corresponds to a 24 hr period. For the gene ‘per1a’, calculate, using the fast Fourier transform, the power (squared amplitude) of the different frequencies measured in this experiment. The aim is to generate the following plot (which clearly shows that the power in the circadian frequency 1/24 is the highest):



first use the function ‘as.numeric’ to convert the values in the matrix into numerical values
compute the fast discrete Fourier transform using the function ‘fft’
compute the power of the discrete Fourier transform by multiplying the resulting values of the ‘fft’ function by the corresponding conjugate values (recall that the discrete Fourier transform generate complex values, and therefore multiplying them by the conjugated values generate real numbers)
examine the vector of the discrete Fourier transform powers. The first value in this vector represents the power of frequency zero and it will be ignored. The rest of the vector is palindromic (the power spectrum is reflected around N/2) with only 6 unique values in locations 2:7 - these represent the powers in the FFT frequencies (that will be computed below), starting from the lowest frequency and moving to the highest. Normalize this subset of powers (which correspond to the FFT frequencies) by dividing each member with the sum of all powers.

create a vector of the frequencies that are present in the discrete Fourier transform of this time-course experiment. Recall that in discrete Fourier transform the highest frequency is the Nyquist frequency, which is half of the sampling rate, namely 1/(2*DeltaT) whereas DeltaT is the time step (the sampling rate fs is one over the time step). Both the lowest frequency and the frequency step are fs/N or 1/(N*DeltaT) whereas N is the number of time points. Use the function 'seq' to create the vector of the frequencies.
finally, plot the powers against the corresponding frequencies

Assume that you want to improve the design of this experiment to better detect the circadian frequency using the FFT. Which option do you think is better:
(a) to design an experiment in which the time step will be shorter, for example 2 hours between each time measurement, while keeping the overall measured time 48 hours.
Or:
(b) to increase the overall number of days measured, for example 4 days instead of 2 days, while keeping the time step the same.

discuss the potential advantages of detecting circadian genes in the frequency domain versus analyzing the data in the time domain (for example - by trying to fit the expression pattern of a gene with a cosine with 24 hours period).

Process all the genes in the dataset and sort them according to the normalized FFT power in the circadian frequency of 1/24. Print the common names ('GeneSymbol') of the 10 genes with the highest normalized powers. Look them up using the web and find at least one known circadian gene among them to validate the analysis.
store all the circadian frequency powers in a vector
use the function ‘order’ to sort the vector and report the indexes of the sorted vector. Tip - use the ‘abs’ function to make sure that R stores the powers as real numbers. Also make sure that NA values (NA stands for ‘not available’ and will be generated if the fft was performed on a vector of zeros) are ordered last.

Part 6 - Detecting genes with variable expression levels

One of key tasks in RNA sequencing data analysis is revealing genes with aberration of expression levels between different treatments/conditions or between different time points. These genes are generally termed ‘variable genes’. Detection of variable genes is possible even without a robust statistical model of the biological and technical repeats and therefore it can be useful in cases where detection of differentially expressed genes (described previously) can’t be performed. Moreover, detection of variable genes is often used as a form of dimension reduction; instead of analyzing all genes in the experiment (usually >10,000 genes), focusing only on the variable genes can significantly reduce the dimensionality by an order of magnitude

or more. Lastly, detection of variable genes is an essential step in single-cells RNA sequencing, in which methods for detection of differentially expressed genes are not established yet.

How do we define and detect variable genes? As demonstrated in the previous parts, for a given gene, the variance in the expression levels is highly dependent on the average expression level. Thus, variable genes are often defined as genes with variance (between different treatments/conditions/time points etc) that is higher than the expected variance for genes with similar expression levels (see for example the methods section of: Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets). In this part we will again use the circadian dataset to detect variable genes and examine if and how they relate to circadian genes detection.

Identify the set of genes that was most variable across the circadian dataset, after controlling for the relationship between mean expression and variability. Calculate the mean and a variance for each gene across the different time points, and place genes into 20 bins based on their average expression, starting from a minimal average expression of 3 (log-transformed). Within each bin, z-normalize the variance of all genes within the bin, in order to identify outlier genes whose expression values were highly variable even when compared to genes with similar average expression. Print the common names ('GeneSymbol') of the the 40 genes with the highest z-values. Can you detect known circadian genes among them? Is it expected that circadian genes will also be variable genes?
Detailed instructions:
Create a numerical matrix from all the count data in CircadianRNAseq.csv, by applying the function 'as.numeric' only on the columns which contain numerical data. Note that the function 'as.numeric' returns a vector, which should be assigned back to a matrix form using the function 'matrix'
Calculate the variance and mean for each gene in all the time points
As before, log transform the variance and the mean (Tip: add 1 before applying the log transformation)
To allow binning of the data according to the mean expression levels, sort the vector of the mean in ascending order. Note that the vector of the variance and a vector which holds the original index information should be sorted in parallel, for example:
Index Mean Var
1 15 2.7
2 7 5.4
3 12 18
after sorting it should be:
Index Mean Var
2 7 5.4
3 12 18
1 15 2.7
Use this link to learn about sorting in R: How to Sort Data in R

Bin the mean expression values into 20 groups starting from the minimal value of 3 and ending with the maximal mean expression. Then, calculate the z-score of the variance of each gene, in each bin separately. The z-score will be the variance of each gene minus the mean variance in that bin, dividing by the standard deviation of the variance in that bin.
sort the resulting z-values in a descending order. Note that a vector which holds the index information should be sorted in parallel.