תרגיל בית 1

מגישים:

- **עדי אויזרוביץ 211872114**
- **יובל אלמוג 321582629**

Part 1 - General introduction to sequencing data

Explain why we can only get an estimation of the expression levels and not the actual number of RNA molecules for each gene:

Answer

In molecular biology, measuring gene expression typically involves quantifying the RNA produced from each gene, which serves as an indicator of gene activity. However, this process only provides an *estimation* of RNA levels, not an exact count of RNA molecules. Here's why:

1. **PCR Bias and Amplification Inefficiencies** In techniques like RT-qPCR (quantitative PCR), even though cDNA is amplified, PCR efficiency can vary between reactions and between different sequences, leading to *PCR bias* and amplification errors. These variations affect the final readout, making it difficult to directly correlate with the exact starting number of RNA molecules.
2. **Sampling and Stochastic Effects**: In cells, particularly in small samples, the number of RNA molecules for any given gene can vary significantly due to biological noise and small number effects. This makes exact quantification difficult because a sample may not perfectly represent the broader population of RNA molecules.
3. **RNA Degradation**: RNA is inherently less stable than DNA and can degrade easily during extraction, handling, or storage.

Part 2 - Explore sequencing data using R

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

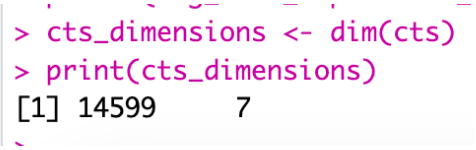
The first 10 rows of the cts matrix contain **gene expression count data** for each sample. Each row represents a gene , and each column corresponds to a sample. The values indicate the **number of RNA-seq reads** mapped to each gene in each sample, gene's expression level in that sample.

The **gene expression count data :**
Specific gene identifiers in the rows and sample labels in the columns.

The values represent the **raw read counts** for each gene in each sample:

- Rows are genes (gene IDs like FBgn0000003).
- Columns are samples with labels indicating the treatment conditions, such as treated or untreated.
- The values reflect how often reads from the RNA-seq data mapped to each gene in a given sample, used in differential expression levels.

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



```
> cts_dimensions <- dim(cts)
> print(cts_dimensions)
[1] 14599     7
```

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

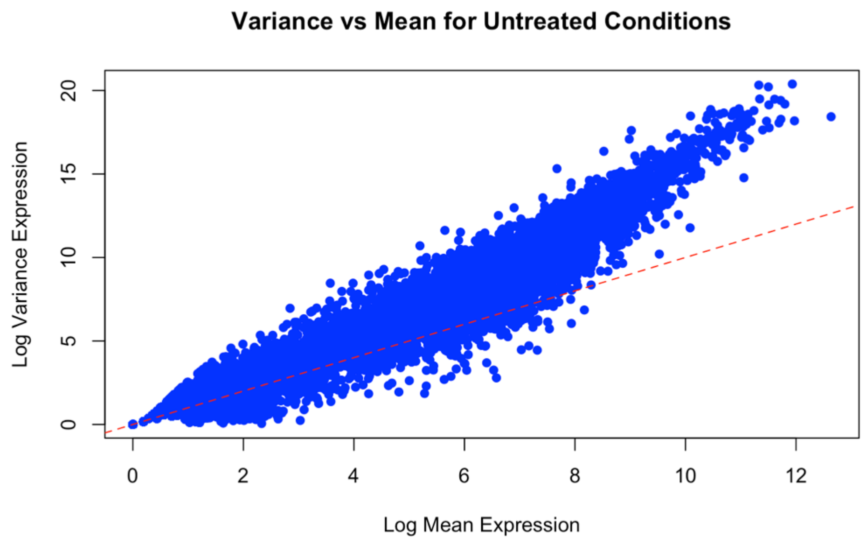
No, the sum of reads (i.e., the total count across all genes) is typically **not the same for each sample** in RNA-seq data. Variability in total read counts across samples can occur due to differences in sequencing depth, library preparation, or other experimental factors.

## Create a normalized version of the cts matrix

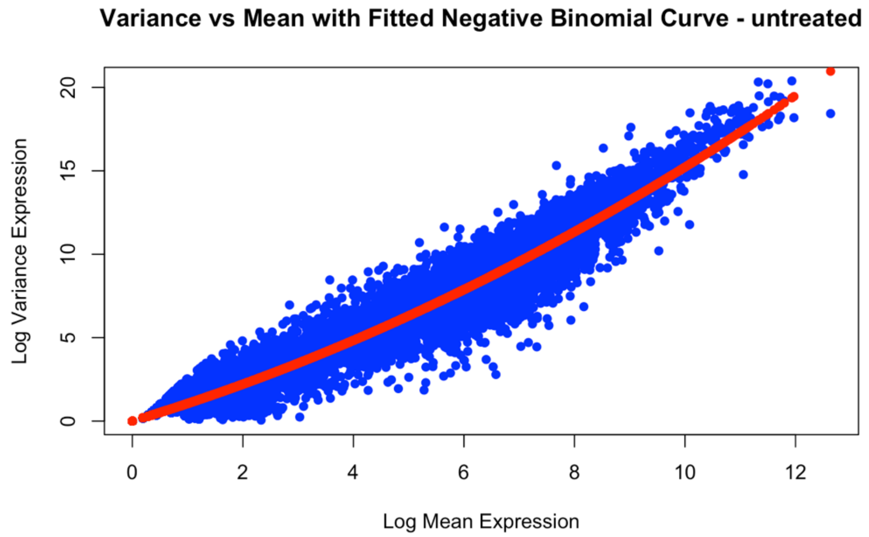
To normalize the cts matrix so that each column has the same total read count as the first column, you can calculate a scaling factor for each column. This factor will be the ratio of the total reads in the first column to the total reads in each other column. Then, multiply each column by its corresponding factor.

Part 3 - Basic statistics of sequencing data

## Does the data fit Poisson Distribution?


Variance vs Mean for Untreated Conditions

## Does the data fit a Negative Binominal Distribution?

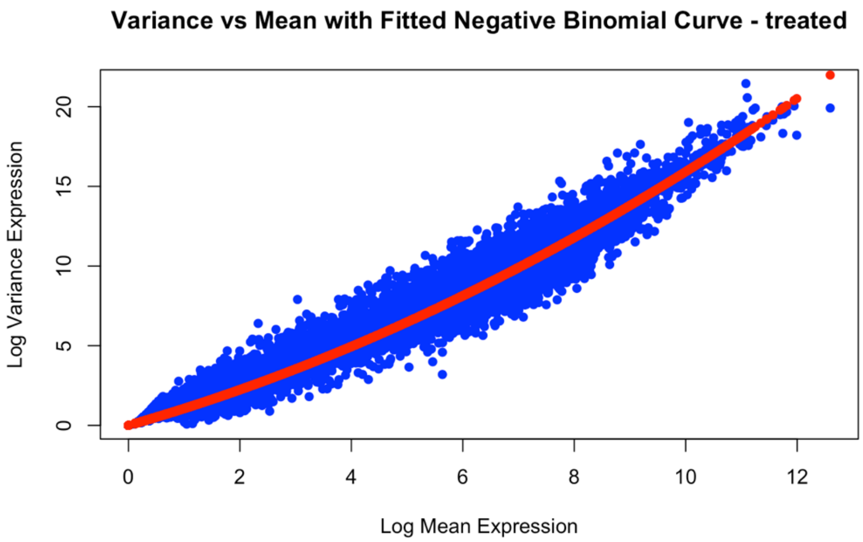

Variance vs Mean with Fitted Negative Binomial Curve - untreated

## What is the value of the dispersion parameter?

```
a <- coef(curved_fit)
a = 0.05220069
```

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

a = 0.05921764



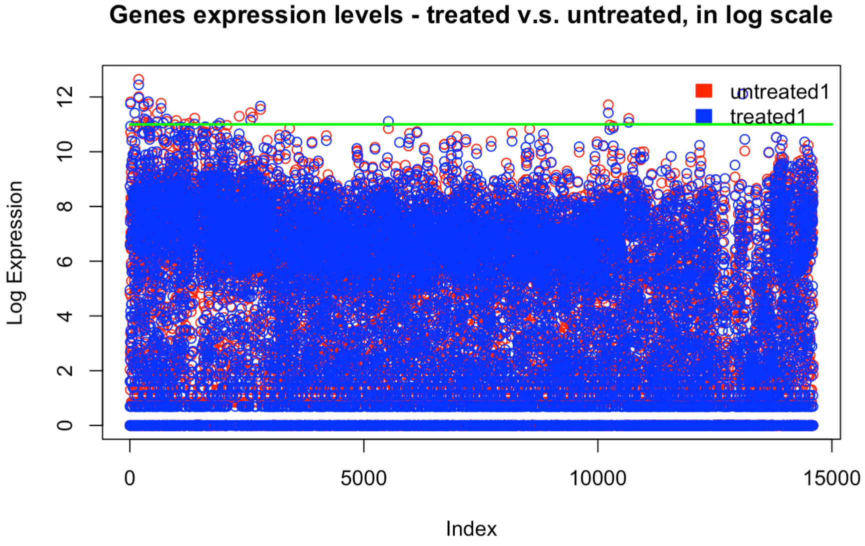
## Judging by the results obtained, do you think that the different untreated and treated samples in this experiment are technical or biological repeats?

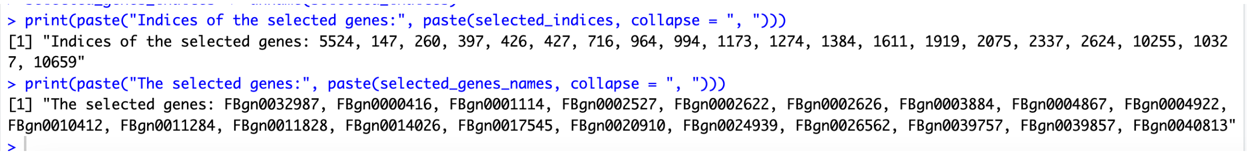
“When comparing samples of different conditions we usually have multiple replicates of each condition. Those replicates need to be independent for statistical inference to be valid. Such replicates are called “biological” replicates because they come from independent animals, dishes, or cultures. In contrast, splitting a sample in two and running it through the sequencer twice would be a “technical” replicate. In general, there is more variance associated with biological replicates than technical replicates.
In other words, the Poisson process in each sample has a slightly different expected count parameter. This is the source of the “extra” variance (overdispersion) we observe in sequencing data. In the framework of the NB distribution, it is accounted for by allowing Gamma-distributed uncertainty about the expected counts (the Poisson rate) for each gene. Conversely, if we were to deal with technical replicates, there should be no overdispersion and a simple Poisson model would be adequate.”

In this experiment the different samples are biological repeats. That’s because we can see that our data destribute like a Poisson Distribution but with an extra room for varriaty. The NB distribution can be defined as a **Poisson-Gamma mixture distribution**. This means that the NB distribution is a weighted mixture of Poisson distributions where the rate parameter λ (the expected counts).

Part 4 - Detect differentially expressed genes

## Using visual inspection, detect at least one gene that has different expression levels in the first treated sample compared to the first untreated sample

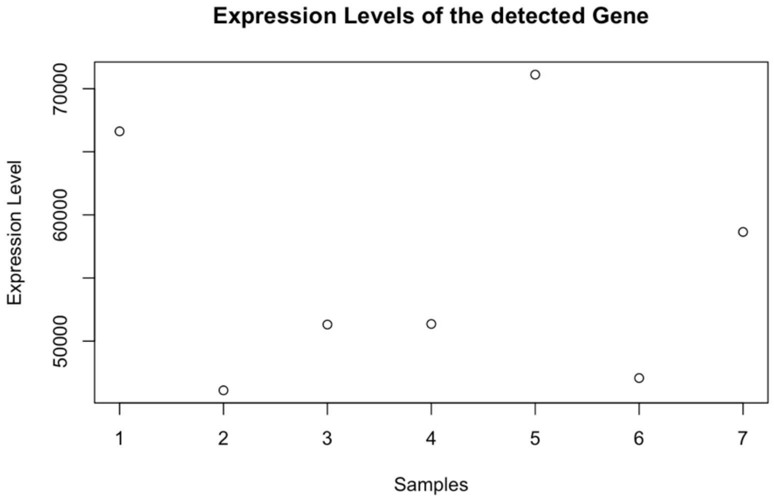




```
> print(paste("Indices of the selected genes:", paste(selected_indices, collapse = ", ")))
[1] "Indices of the selected genes: 5524, 147, 260, 397, 426, 427, 716, 964, 994, 1173, 1274, 1384, 1611, 1919, 2075, 2337, 2624, 10255, 1032
7, 10659"
> print(paste("The selected genes:", paste(selected_genes_names, collapse = ", ")))
[1] "The selected genes: FBgn0032987, FBgn0000416, FBgn0001114, FBgn0002527, FBgn0002622, FBgn0002626, FBgn0003884, FBgn0004867, FBgn0004922,
FBgn0010412, FBgn0011284, FBgn0011828, FBgn0014026, FBgn0017545, FBgn0020910, FBgn0024939, FBgn0026562, FBgn0039757, FBgn0039857, FBgn0040813"
>
```

## 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?

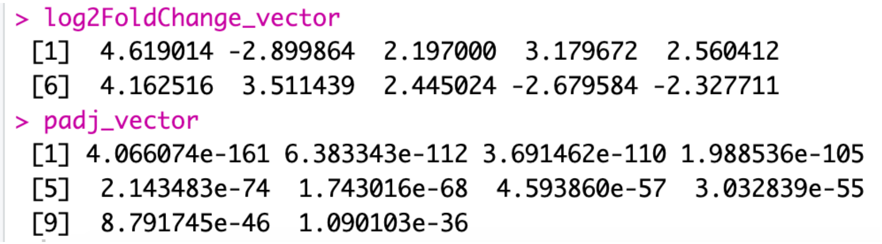
We will choose the first one of them: at index 5524. Gene: FBgn0032987.
The expression of the gene appears consistently different in the treated conditions compared to the untreated conditions:



## Does the data we have fit the basic assumptions of DEseq?

Yes, our data is indeed divides like negative binomial.

## Display only the column ‘log2FoldChange’ (the log transformed fold change between the two conditions) and the column ‘padj’ (the adjusted p-value) of these 10 genes. Is the gene that you detected using visual inspection among them?



```
> log2FoldChange_vector
 [1]  4.619014 -2.899864  2.197000  3.179672  2.560412
 [6]  4.162516  3.511439  2.445024 -2.679584 -2.327711
> padj_vector
 [1] 4.066074e-161 6.383343e-112 3.691462e-110 1.988536e-105
 [5]  2.143483e-74  1.743016e-68  4.593860e-57  3.032839e-55
 [9]  8.791745e-46  1.090103e-36
```

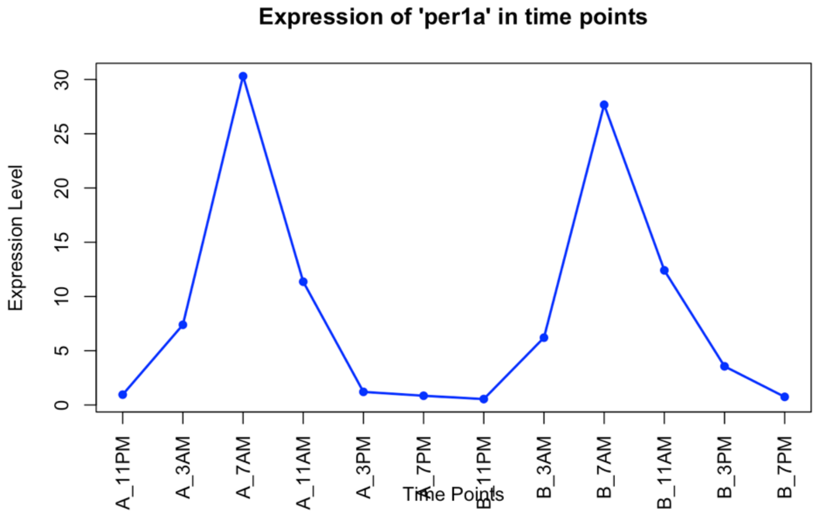
The visually inspected gene (FBgn0032987) is NOT among of this 10 genes.

Part 5 - Detect rhythmical patterns using Fourier transform

## Plot the expression of the gene 'per1a' at all the time points. Does the expression of this gene seem circadian?

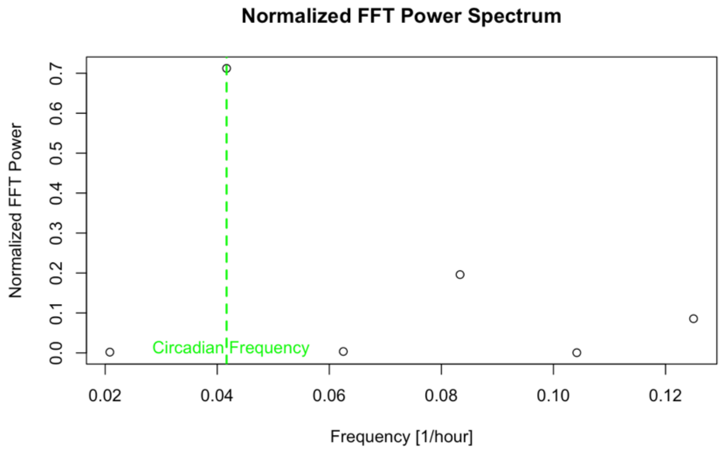
Per1 (Period1) is a core circadian clock gene that plays a crucial role in maintaining the body’s circadian rhythms. In addition, Per1 is responsive to light-dark cycles.

So yes, the expression of this gene is circadian, as you can see the peraiodic oscillations behavior peaks over time in the expression samples (peak at 7 am in both days):



## 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.

As we can see in the resulting plot, clearly the power in the circadian freuency (1/24 = 0.041667) is the highest:



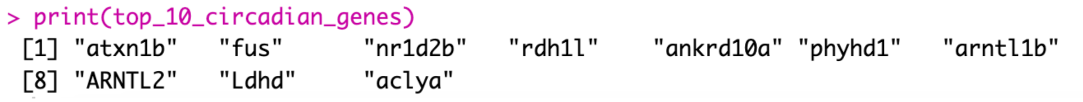
## Assume that you want to improve the design of this experiment to better detect the circadian frequency using the FFT. Which option do you think is better:

We think that option (b) is the right way to improve the experiment to better detect the circadian frequency using FFT. Because option (a) shorten the time step, that’s impact the frequency resolution. The nyquist frequency increases and the resolution gets better. But in option (b) the frequency resolution gets significantly better because the total number of samples (N) increases.
Also the additional cycales provides more data to robustly confirm the periodicity and reduce noise.
Prioritizing extended duration is more impactful for circadian rhythm detection.

## 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).

Frequency domain is excellent for identifying and quantifying oscillatory patterns globally and efficiently, while the time domain provides more biologically interpretable parameters and is better suited for specific hypothesis testing.

## 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:



```
> print(top_10_circadian_genes)
 [1] "atxn1b"   "fus"      "nr1d2b"   "rdh1l"    "ankrd10a" "phyhd1"   "arntl1b"
 [8] "ARNTL2"   "Ldhd"     "aclya"
```

arntl1b: These gene belong to the ARNTL (BMAL) family, which is a key component of the circadian clock. ARNTL interacts with clock to regulate transcription of circadian target genes.

Part 6 - Detecting genes with variable expression levels

Circadian genes are generally expected to be variable genes because Circadian genes follow an expression pattern in a 24-hour period, with peaks and at specific times. This leads to higher variability in their expression levels across time points compared to non-circadian genes. It Variable expression levels validates and confirms that that the Circadian gene is truly circadian.