NVIDIA. DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Lecture 19.2 - Data Preprocessing

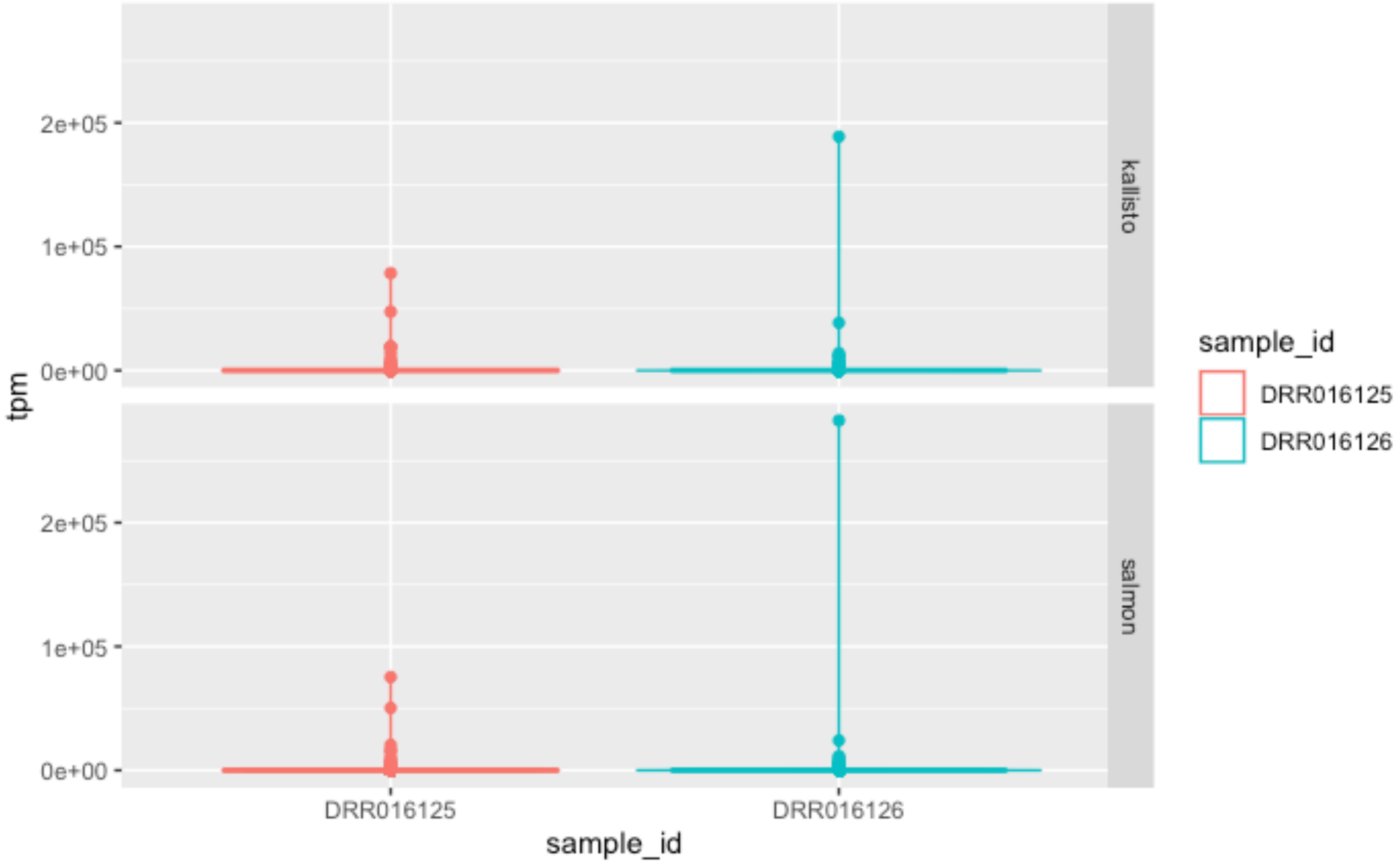
Data Distribution

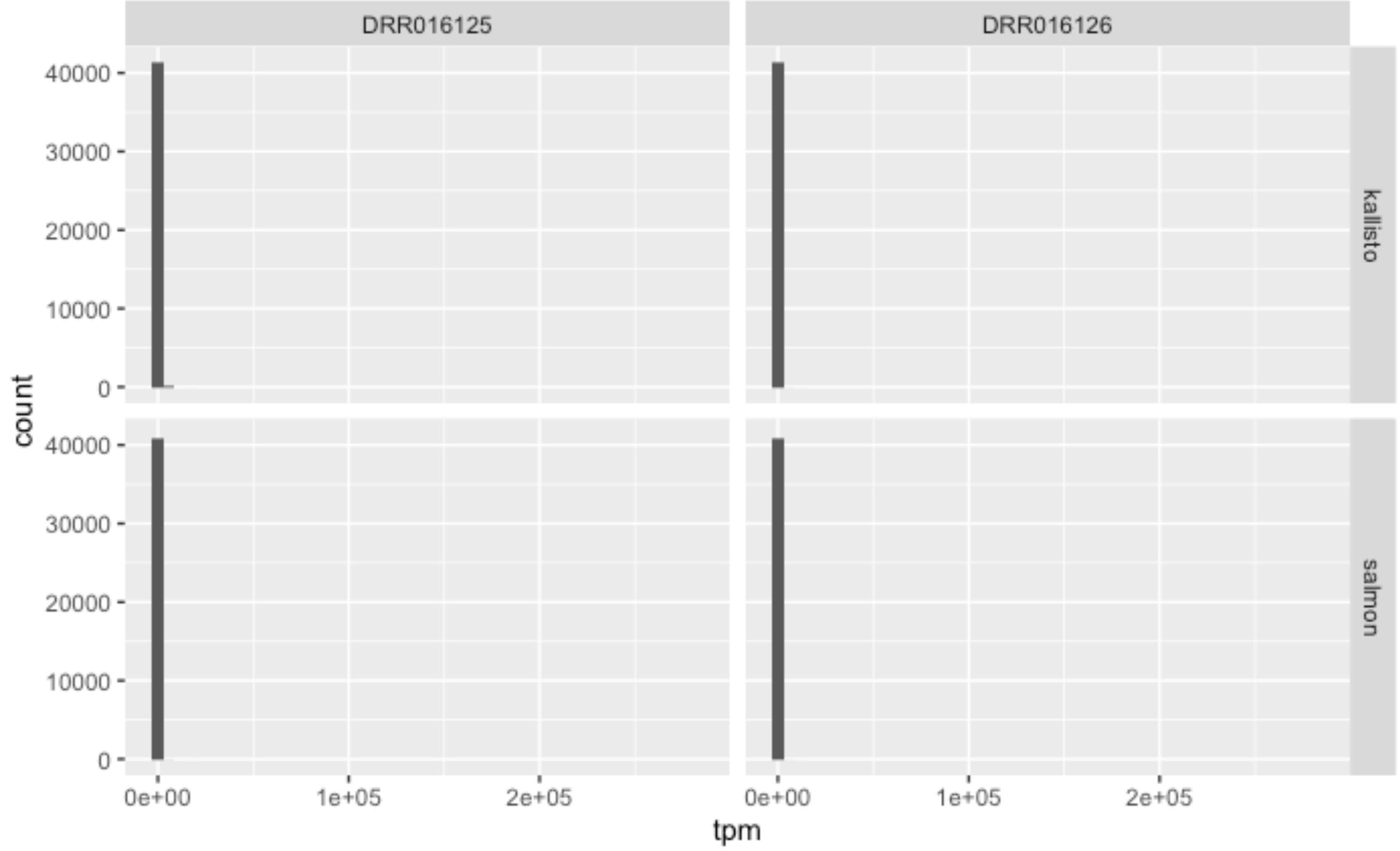
Gene expression values from RNAseq are extremely skewed to low values.

Also, many values are zeros.

Such distribution in data would be hindrance to many analytical tools, which often assumes symmetrical distribution and/or ”normality”.

TPM (Transcripts Per Kilobase Million) is used to measure gene or transcript expression levels.





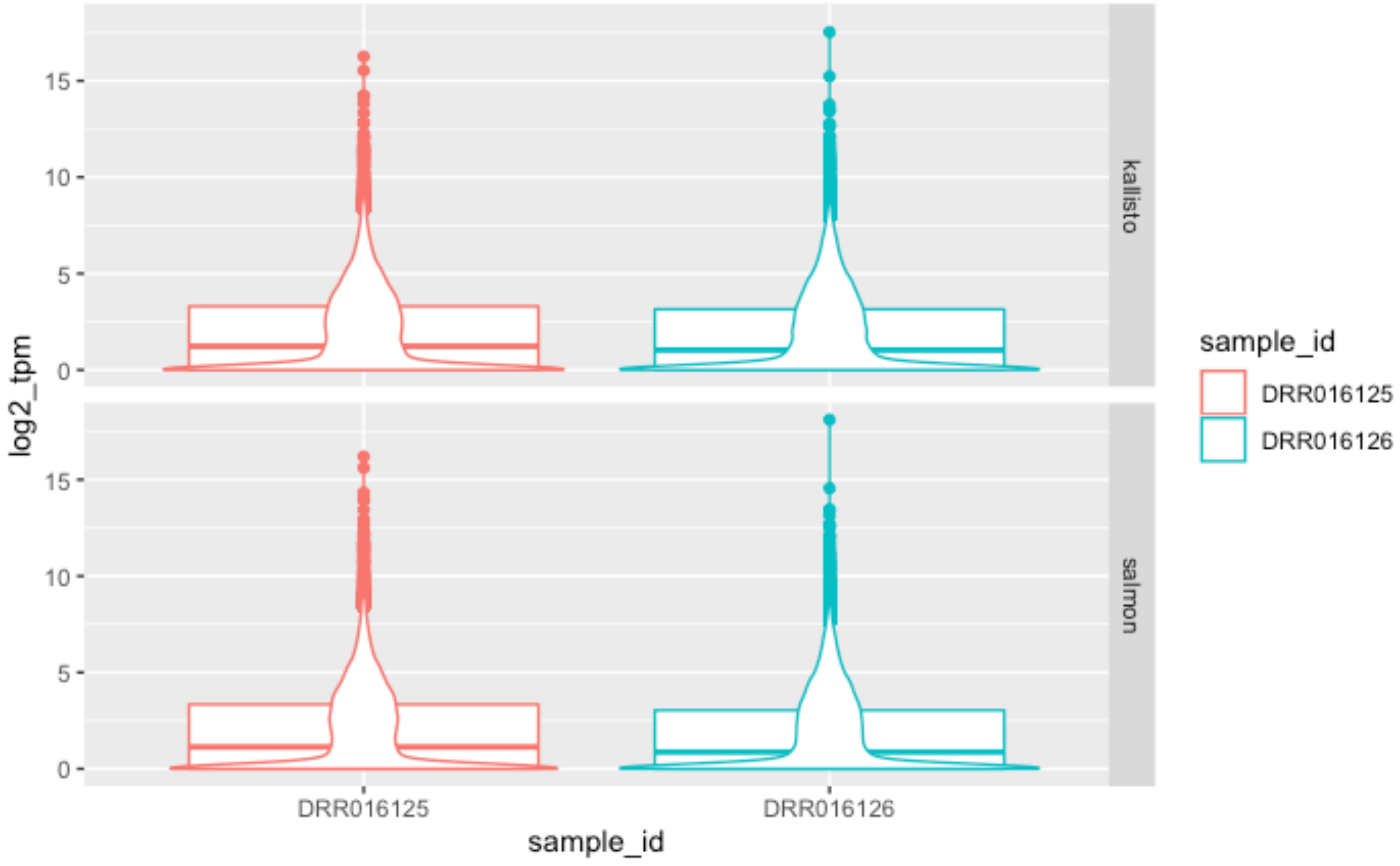
Data transformation

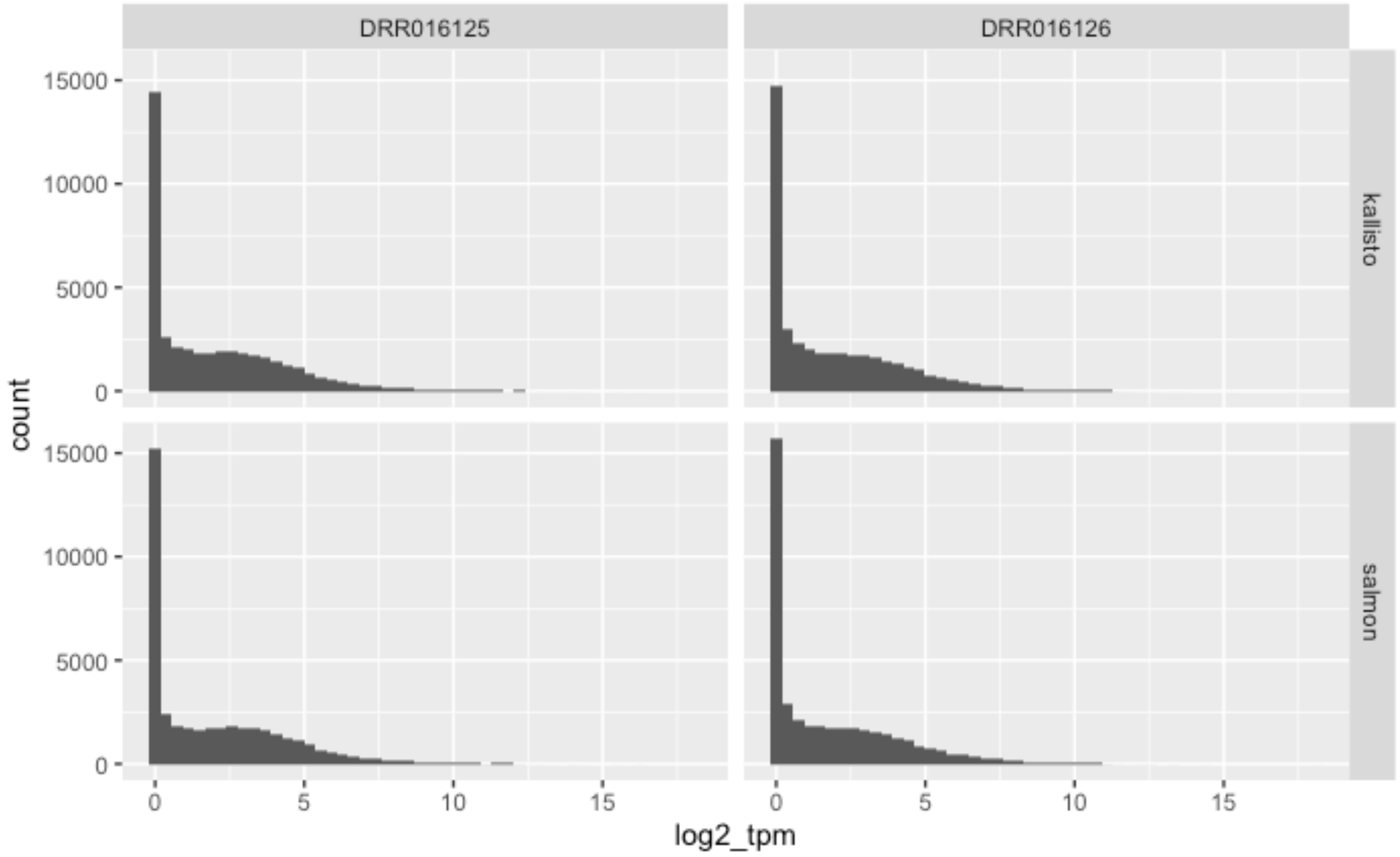
Some kind of transformation may help adjust data distribution.

Such distribution should be monotonic so that the rank distribution does not change.

Examples:

- Log2 transformation
- sqrt

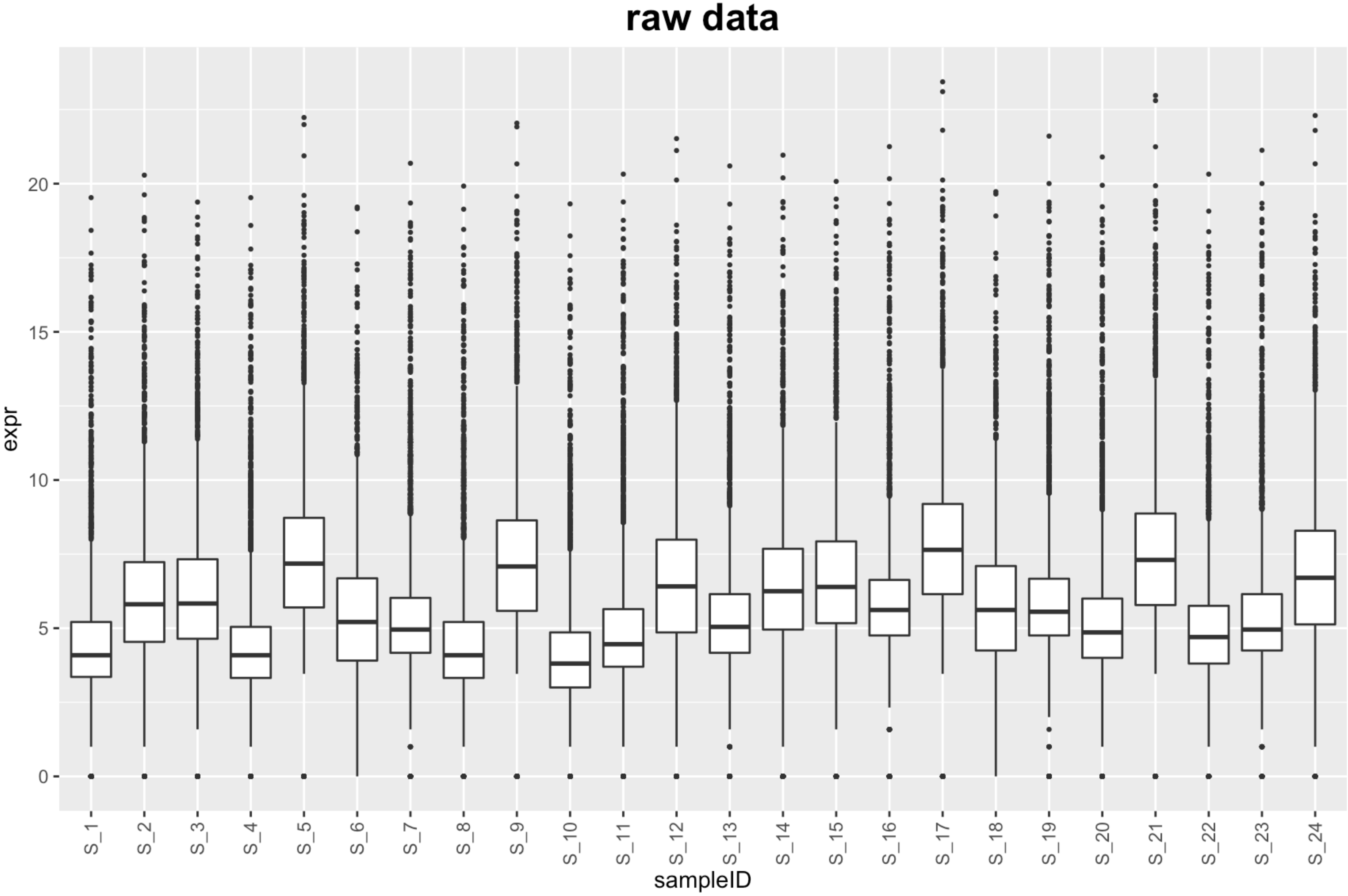




Sample to Sample Variation

Variability due to noise and artifacts, often attributed to differences in samples, RNA quality, sample preparation, etc.

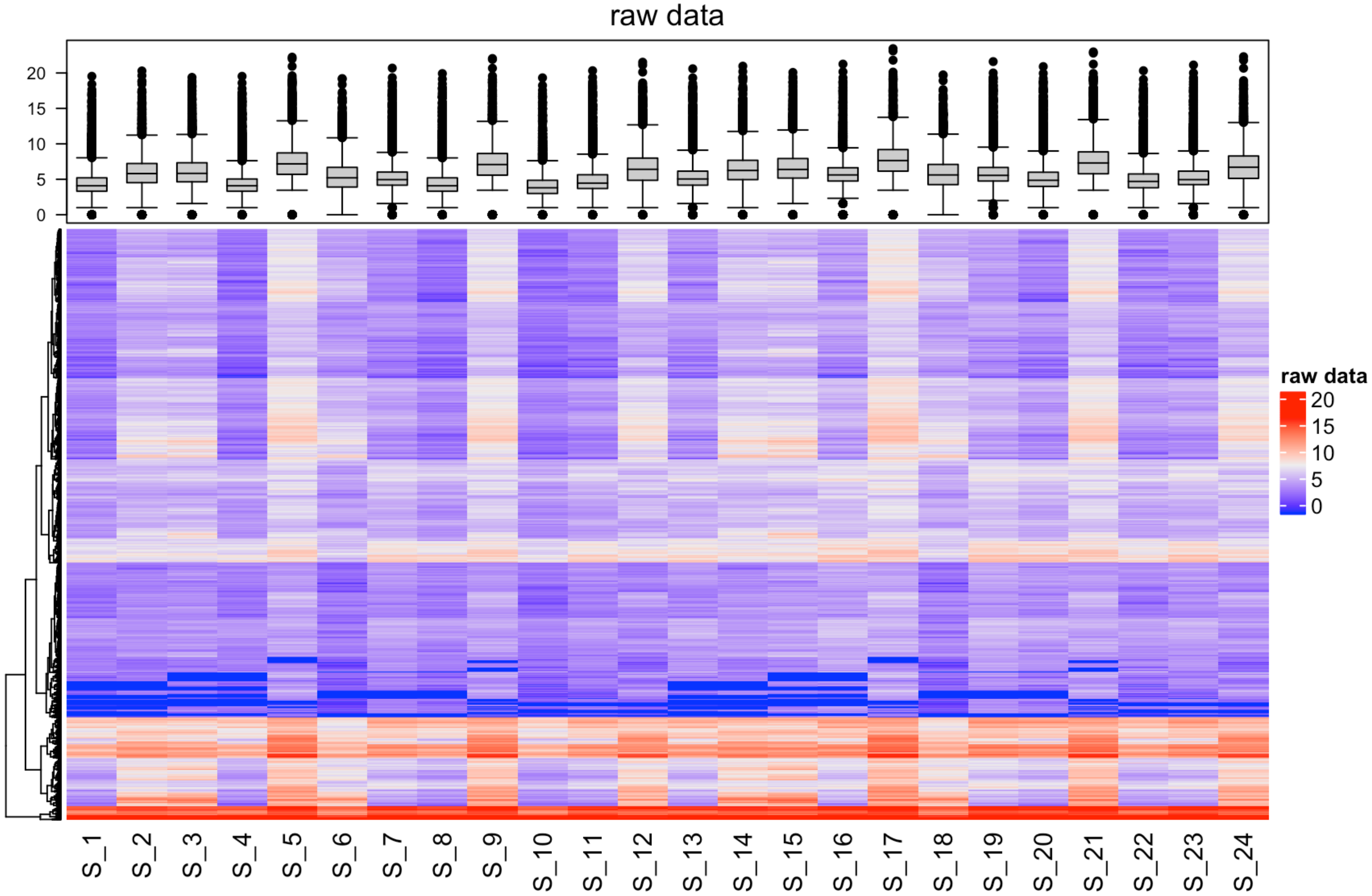
Needs to correct these biases caused by non-biological conditions for down stream data analysis



Sample to Sample Variation

Variability due to noise and artifacts, often attributed to differences in samples, RNA quality, sample preparation, etc.

Needs to correct these biases caused by non-biological conditions for down stream data analysis



Preprocessing

Mean centered

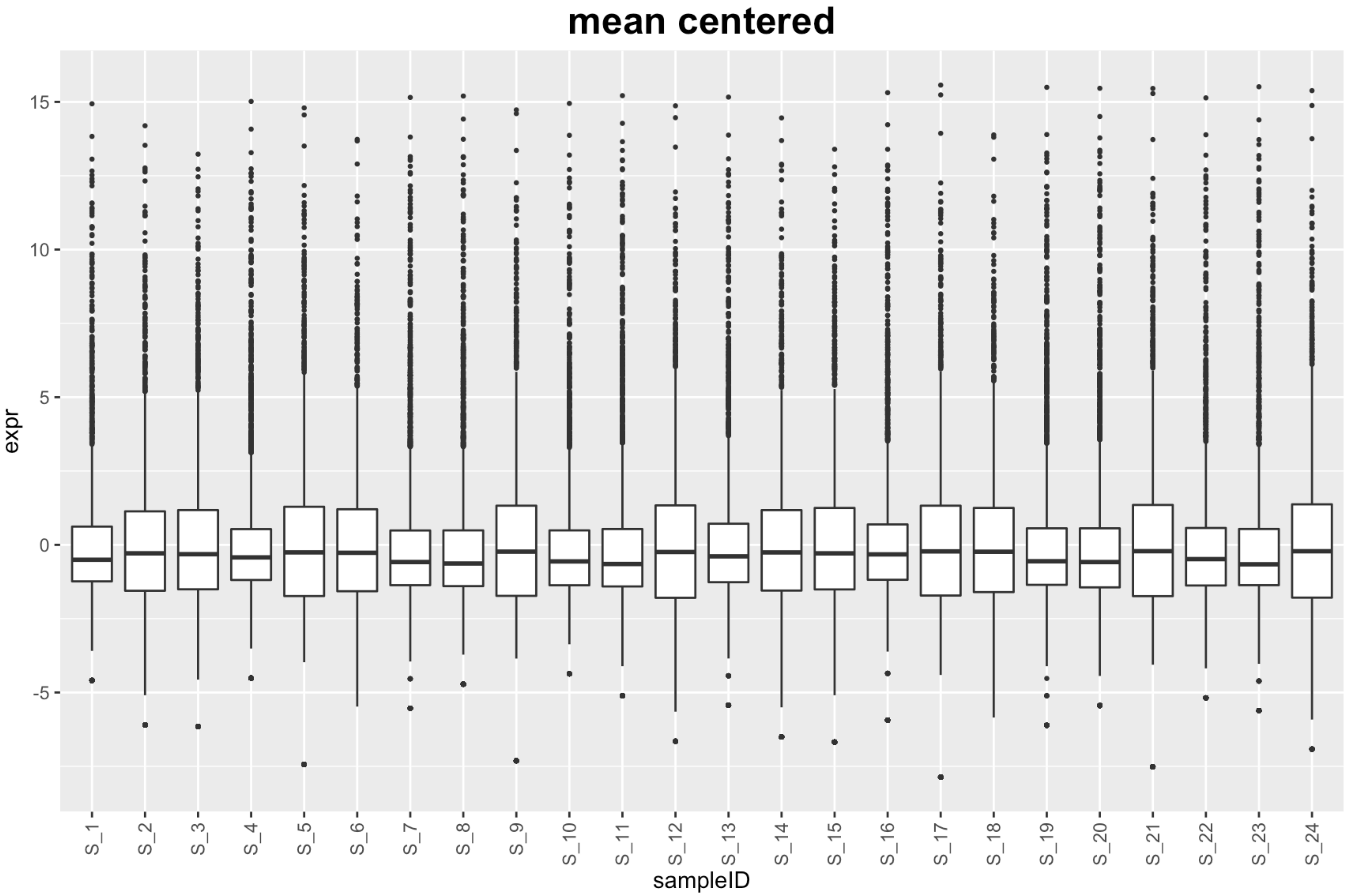
Set mean of each sample to ”0”

Let X_{ij} be an expression of gene *i* in sample *j*

Then a normalized expression of gene *i* in sample *j*,

$$\hat{X}_{ij} = X_{ij} - u_j$$

where $u_j = \frac{\sum_i X_{ij}}{N}$



Preprocessing

Mean centered

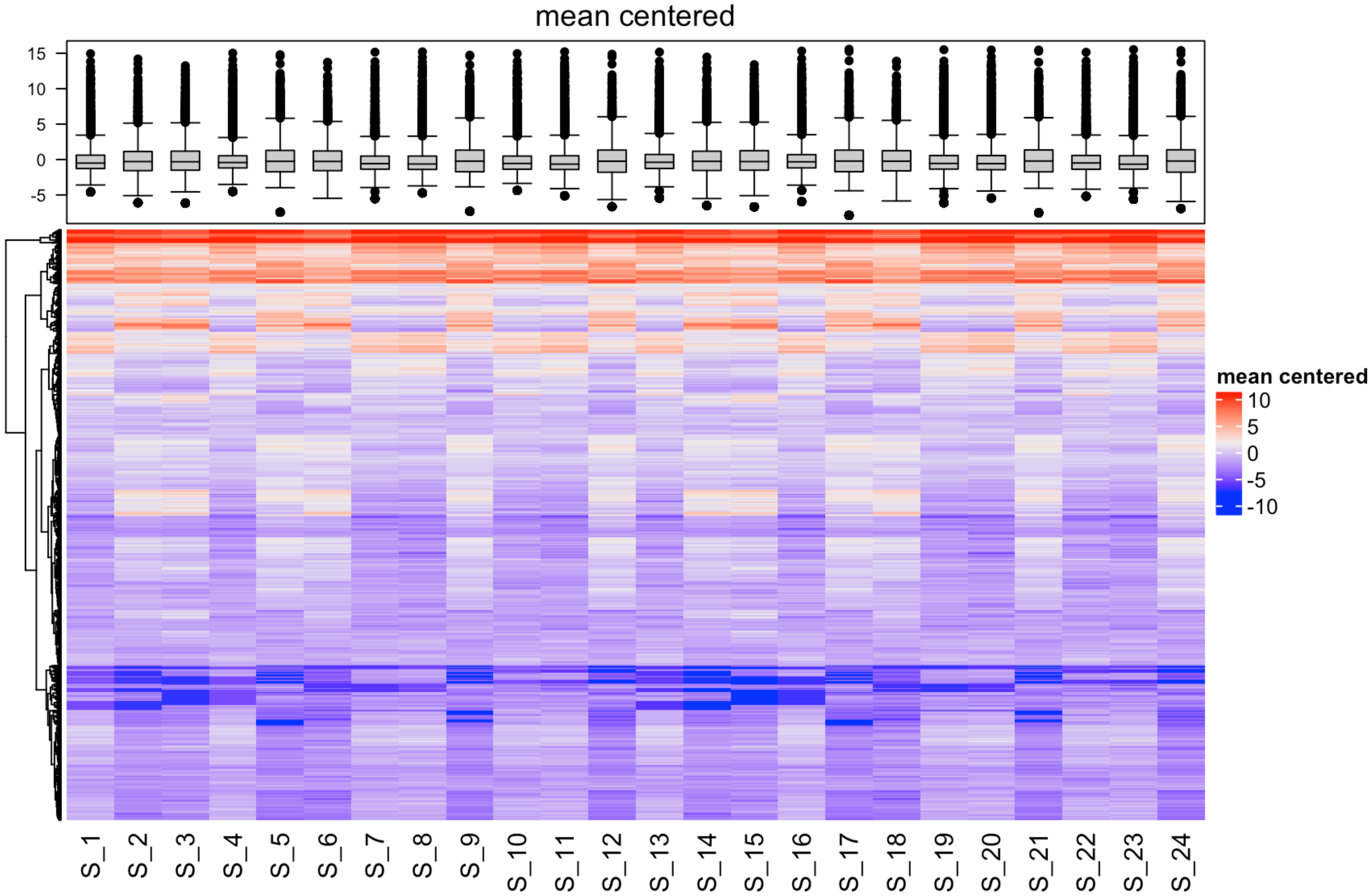
Set mean of each sample to ”0”

Let X_{ij} be an expression of gene *i* in sample *j*

Then a normalized expression of gene *i* in sample *j*,

$$\hat{X}_{ij} = X_{ij} - u_j$$

where $u_j = \frac{\sum_i X_{ij}}{N}$



Preprocessing

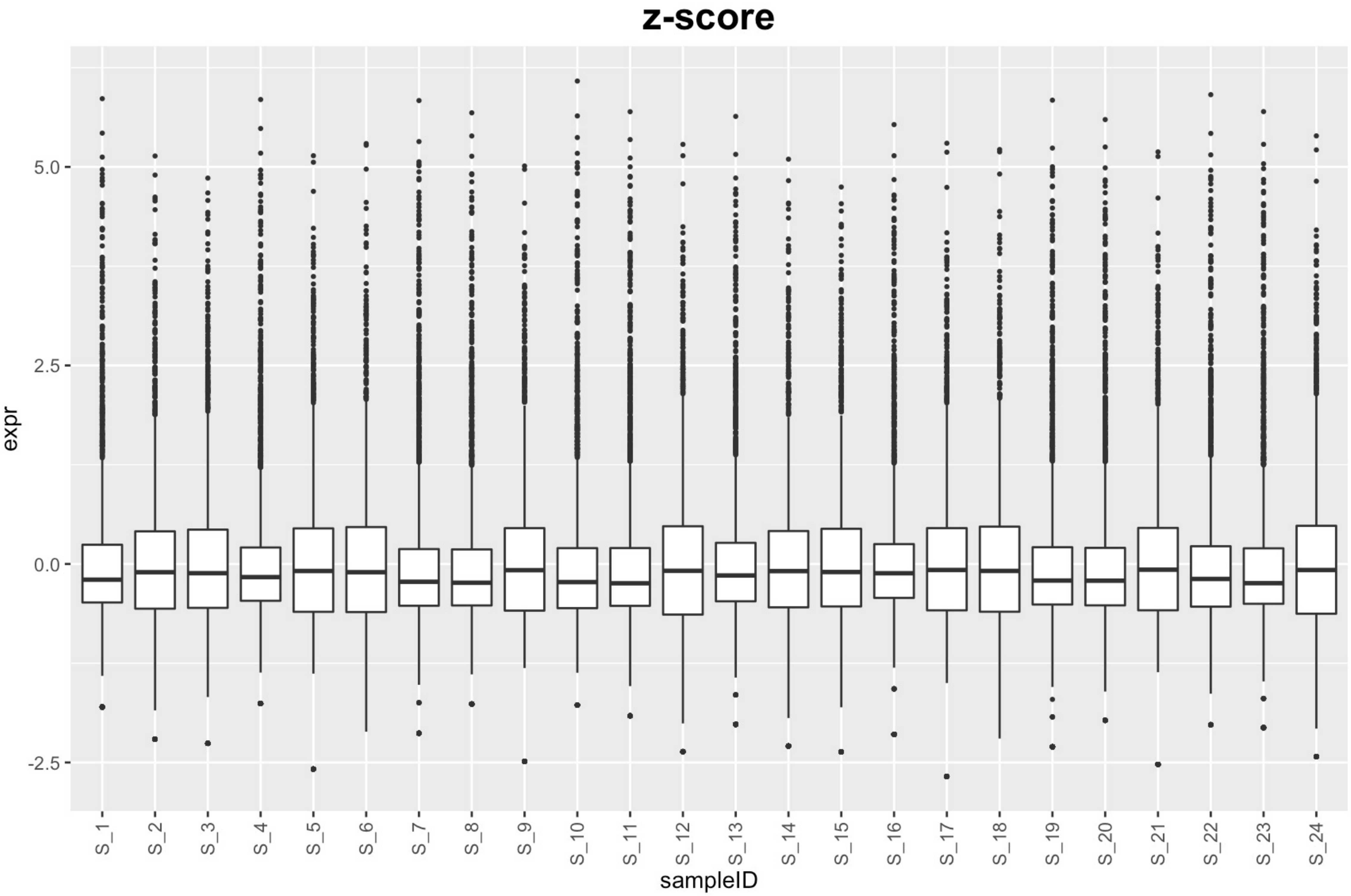
Z-core

a. k. a. Standardization

Normalized expression of gene *i* in sample *j*,

$$\hat{X}_{ij} = \frac{X_{ij} - u_j}{\sigma_j}$$

where $\sigma_j = E[(X_{ij} - u_j)^2]$



Preprocessing

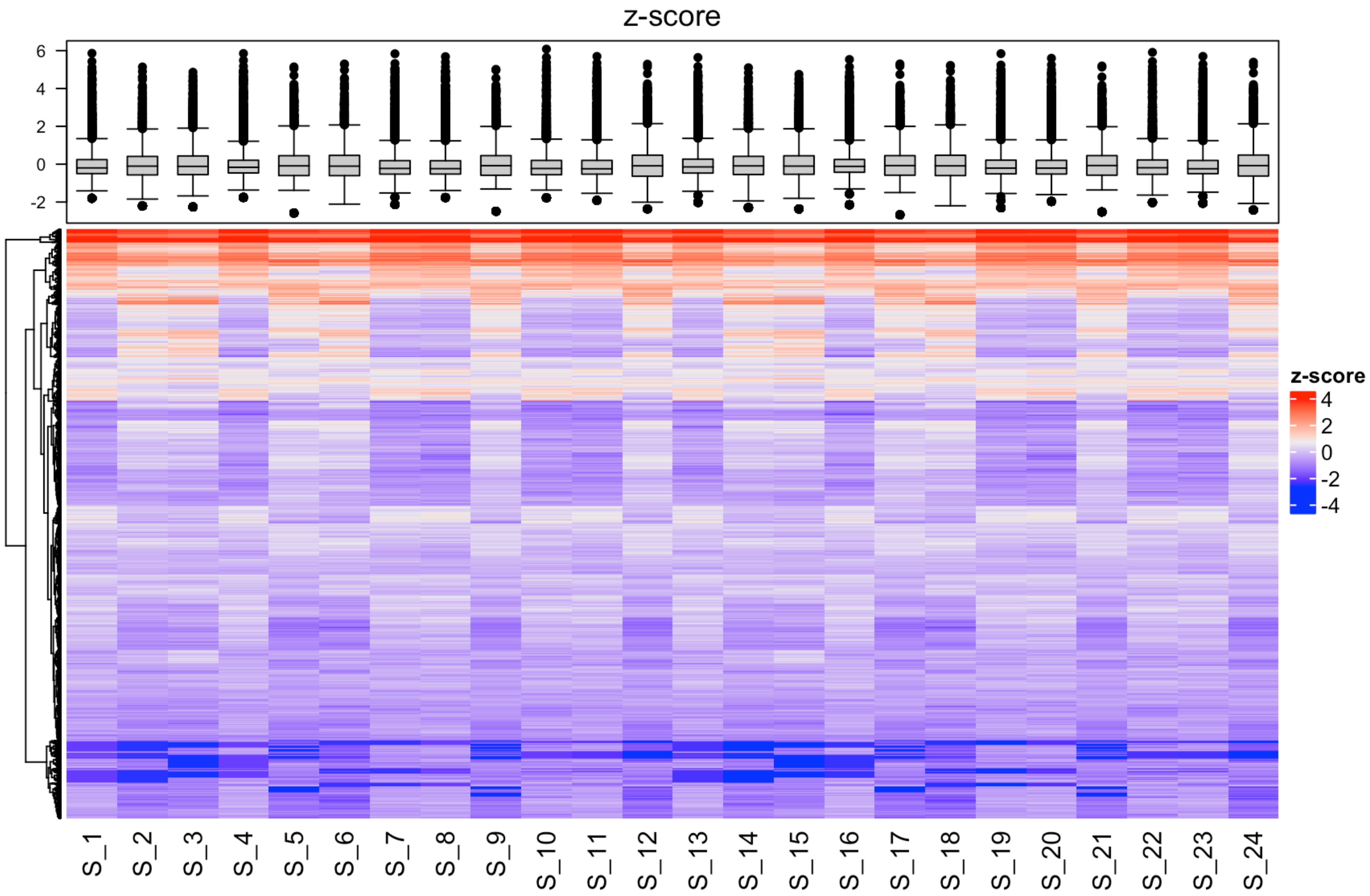
Z-core

a. k. a. Standardization

Normalized expression of gene *i* in sample *j*,

$$\hat{X}_{ij} = \frac{X_{ij} - u_j}{\sigma_j}$$

where $\sigma_j = E[(X_{ij} - u_j)^2]$



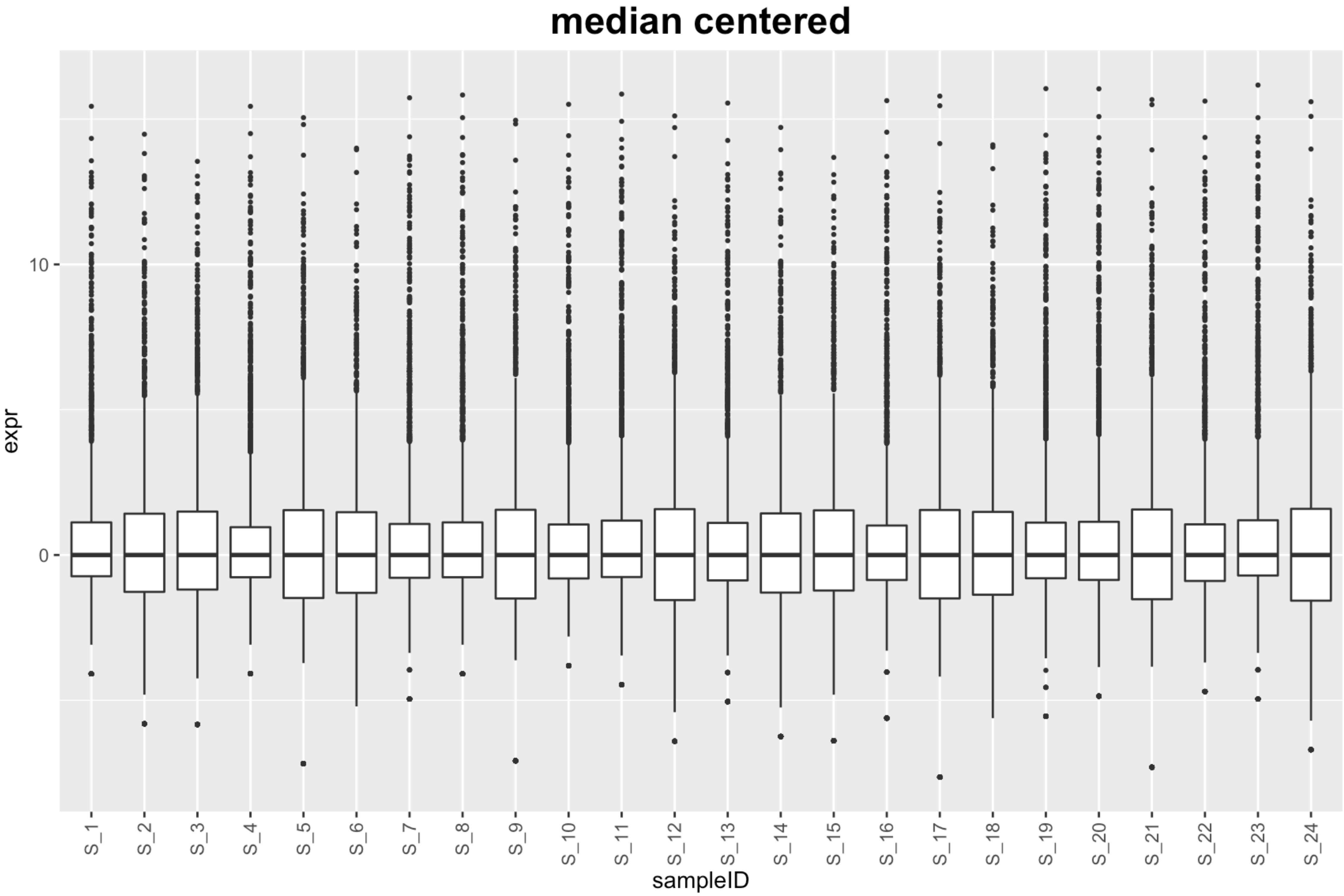
Preprocessing

Median centered

To find the median, we first arrange the observations in order from smallest to largest value.

If there is an odd number of observations, the median is the middle value.

If there is an even number of observations, the median is the average of the two middle values.



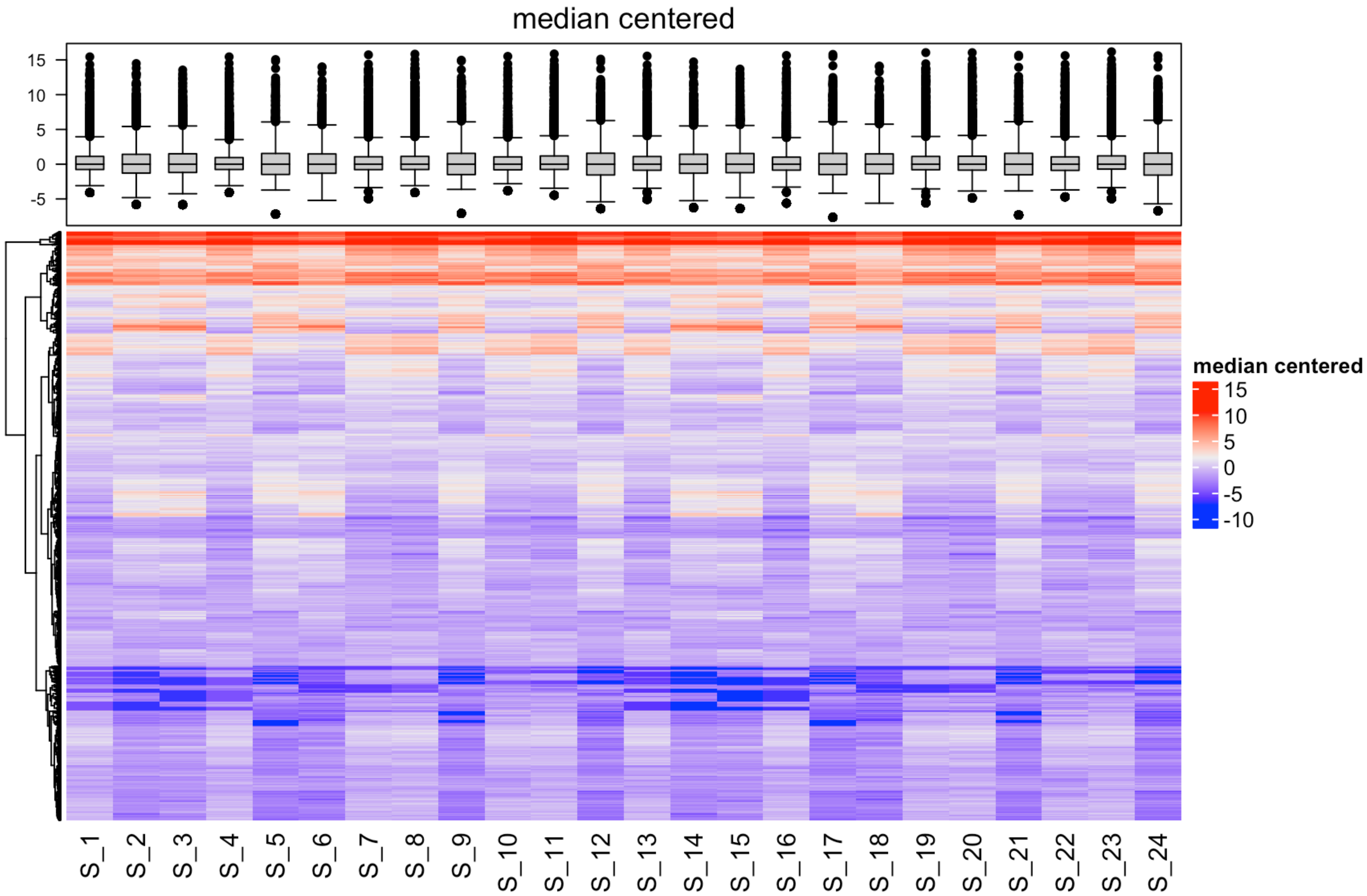
Preprocessing

Median centered

To find the median, we first arrange the observations in order from smallest to largest value.

If there is an odd number of observations, the median is the middle value.

If there is an even number of observations, the median is the average of the two middle values.



NVIDIA. DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You