Comparing Normalisation methods

Giosuè Moscato

2022-11-18

0. Loading the data

The data that will be used for this analysis comes from the second round of sequencing. This is a preview of the matrix:

```
##                 RT11B RT12B RT13B RT14B RT15B RT16B RT17B RT18B RT19B RT20B
## hsa-miR-10a-5p    182   794   314  3168  4960   548   407  1130  2608   503
## hsa-miR-10b-5p    310  1718   621  5921 10043  1157   813  2897  4105  1389
## hsa-miR-181a-5p    86   301   114  2086  1828   325   169    86   460   581
## hsa-miR-185-3p      0     0     0     9     2     0     0     0     0     0
## hsa-miR-191-5p     80   260   145 16938  1635   231   108  1110  5058  1119
## hsa-miR-221-3p      1     3     0   178    15     5     7     0    23     9
```

The initial data has the following dimensions (rows = miRNAs, columns = samples):

```
## [1] 728  48
```

Furthermore we will use a second matrix containing metadata for each sample

```
##    Sample.Id Batch  Type Surgery IDH1   MGMT Recurrence Death
## 48     RT01B     6 preRT     STR   wt    met         no    no
## 49     RT02B     6 preRT     STR   wt    met        yes    no
## 50     RT03B     6 preRT     STR   wt no met        yes   yes
## 51     RT04B     6 preRT    <NA> <NA>   <NA>       <NA>  <NA>
## 52     RT05B     6 preRT     GTR   wt no met         no    no
## 53     RT06B     6 preRT    <NA> <NA>   <NA>       <NA>  <NA>
```

1. Setting cut-off

Selection of a subset of samples using a quantile cut-off value, the upper quartile.

For this analysis we will subset those samples with the value of the 3rd quartile > 0. The use of this cut-off will remove those samples that have a number of zero miRNA counts greater than 25% of all the miRNAs that were sequenced across all the samples.

The subsetted data has the following dimensions (rows, columns):

```
## [1] 728  34
```

2. Normalisation methods

In this part we are going to normalise our data using different methods of normalisation. (I followed this article in the choose of some normalisation methods and test for the qualitative assessment of normalized data (Optimization of miRNA-seq data preprocessing))

Several normalization methods were evaluated, including (1) cpm, (2) total count scaling, (3) upper quartile scaling (UQ), (4) TMM, (5) RLE, (6) DESeq, (7) MIXnorm , and (8) PoissonSeq. Each of these methods is described briefly.

1. cpm

Count-per-million—the simplest form of normalization, whereby each library is adjusted for differences in sequencing depth. The counts can then be adjusted to reads per million to facilitate comparison between samples.

2.Total count scaling

Total count scaling—After scaling each sample to its library size, they can be rescaled to a common value across all samples. The baseline reference can be chosen to be the sample with the median library size. If $s_{baseline}$ is the size of the reference library, and s_i is the sum of all reads of the any given library, then the normalization factor is as follows:

$$d_i = \frac{s_{baseline}}{s_i}$$

and the counts for the normalized samples would be

$$x_i' = d_i x_i$$

where x_i is the raw count for a specific target.

3. UQ normalisation

Upper-quartile scaling—In RNA-seq experiments, the predominance of zero and low-gene counts has led to the suggestion of a modified quantile-normalization method: the upper quartile of expressed miRNAs is used instead as a linear scaling factor. This method has been shown to yield better concordance with qPCR results than linear total counts scaling for RNA-seq data (1). It is expected that in miRNA-seq experiments, the 75^{th} percentile of the data will also be found at only 1 or 2 copies/library.

4. TMM

Trimmed mean of M—Normalization by total count scaling makes intuitive sense because it gives us the proportion of counts for a specific target across all samples. If a miRNA is present in the same proportion across all samples, it will be deemed as non-differentially expressed. However, this method does not take into consideration the potentially different RNA composition across the samples. TMM, proposed by Robinson et al. for RNA-seq data normalization, calculates a linear scaling factor, d_i, for sample *i*, based on a weighted mean after trimming the data by log fold-changes *(M)* relative to a reference sample and by absolute intensity (A) (2). TMM normalization takes into account the composition of the RNA population being sampled, which is neglected in total count scaling. This method is implemented in the R Bioconductor package edgeR, with default trimming of M-value by 30% and A-values by 5%.

5. RLE normalisation

Relative Log Expression—Similar to TMM, this normalization method is based on the hypothesis that the most genes are not DE. For a given sample, the RLE scaling factor is calculated as the median of the ratio, for each gene, of its read counts over its geometric mean across all samples. By assuming most genes are not DE, the median of the ratio for a given sample is used as a correction factor to all read counts to fulfill this hypothesis (3). This normalization method is included in the DESeq and DESeq2 Bioconductor packages.

6. DeSeq2

DESeq—To perform differential expression analysis using count data, Anders and Huber proposed modeling the data with the negative binomial distribution, and incorporating data-driven prior distributions to estimate the dispersion and fold changes (4). As a data preprocessing step, the authors introduced the size factor—a scaling factor—to bring the count values across all the samples to a common scale. The size factor for a given library is defined as the median of the ratios of observed counts to the geometric mean of each corresponding target over all samples. This method is implemented in the R Bioconductor package DESeq.

7. MIXnorm

MIXnorm is a new normalization method, labeled MIXnorm, for FFPE RNA-seq data (formalin-fixed paraffin-embedded). Though a number of normalization methods are available for RNA-seq data, none has been specifically designed for FFPE samples, of which a prominent feature is sparsity (i.e. excessive zero or small counts), caused by RNA degradation in such samples. MIXnorm relies on a two-component mixture model, which models non-expressed genes by zero-inflated Poisson distributions and models expressed genes by truncated normal distributions. link for further information.

The reason why I decided to test this normalisation method is that looking at the sctructure of our data we can notice the same features of FFPE samples, sparsity and excessive zero or small counts.

8. PoissonSeq

PoissonSeq (PS) models RNA-seq data by a Poisson log-linear model. Further information available here

3. Batch effect correction

For the Batch effect correction we will use the Combat_seq function implemented in the R package *sva*. ComBat allows users to adjust for batch effects in datasets where the batch covariate is known, using methodology described in Johnson et al. 2007. It uses either parametric or non-parametric empirical Bayes frameworks for adjusting data for batch effects. Users are returned an expression matrix that has been corrected for batch effects. The input data are assumed to be cleaned and normalized before batch effect removal.

ComBat_seq is an improved model from ComBat using negative binomial regression, which specifically targets RNA-Seq count data.

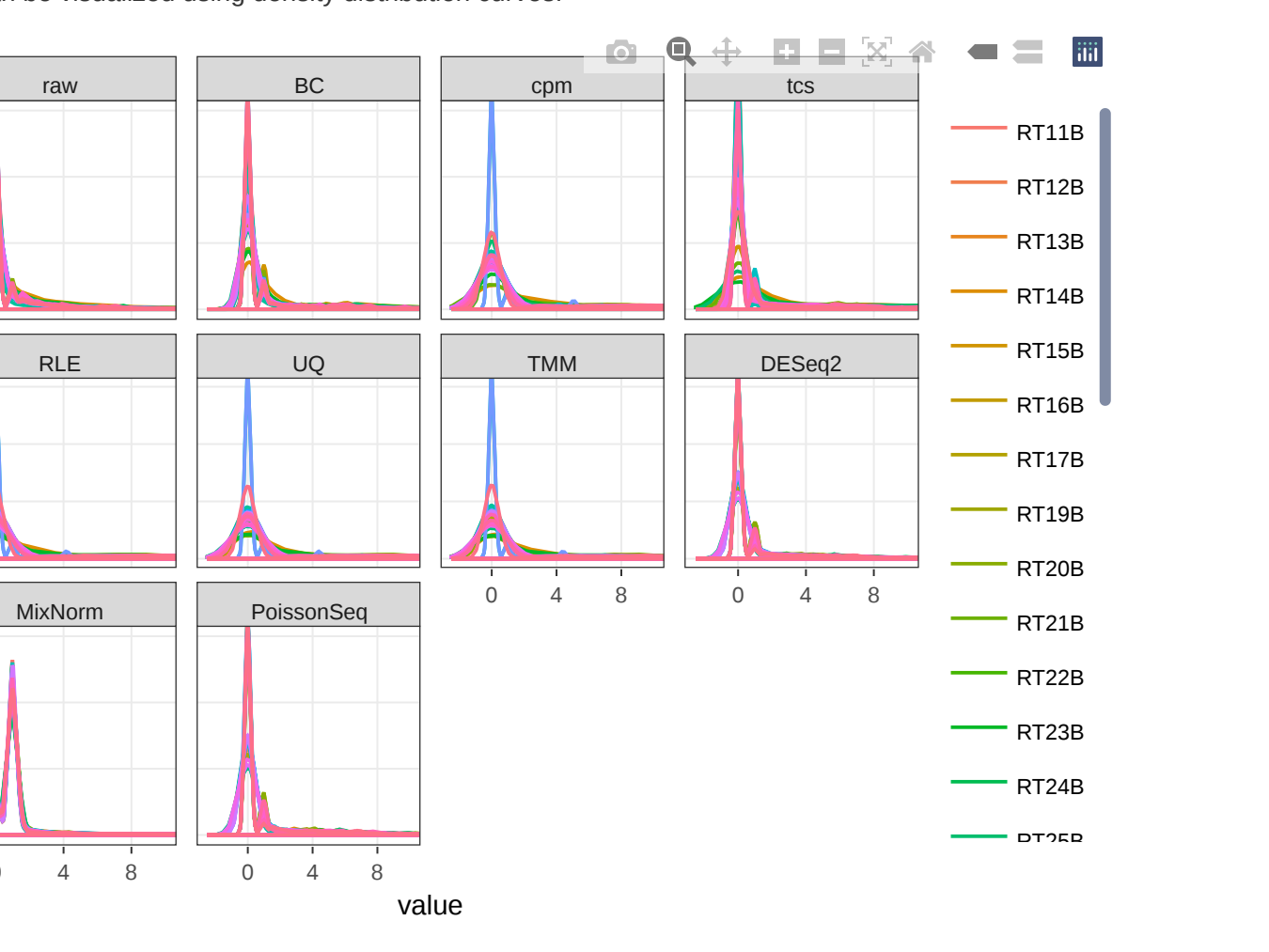
We wiil apply the Combat_seq function to the different-normalised data using the following matrix. The second column of this matrix indicate the number of batch for each sample:

```
##    Sample.Id Batch    Type Surgery IDH1   MGMT Recurrence Death
## 58     RT11B     7 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 59     RT12B     7 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 60     RT13B     7 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 61     RT14B     7  postRT     GTR   wt no met        yes    no
## 62     RT15B     7  postRT     GTR   wt    met        yes   yes
## 63     RT16B     7  postRT     STR   wt    met         no    no
## 64     RT17B     7  postRT     STR   wt    met         no   yes
## 66     RT19B     7  postRT     GTR   wt    met        yes   yes
## 67     RT20B     7  postRT     STR   wt no met        yes   yes
## 68     RT21B     8   preRT     STR   wt no met        yes   yes
## 69     RT22B     8   preRT     GTR   wt no met         no   yes
## 70     RT23B     8   preRT     GTR   wt    met        yes   yes
## 71     RT24B     8   preRT     GTR   wt no met        yes   yes
## 72     RT25B     8   preRT     STR   wt no met        yes   yes
## 73     RT26B     8   preRT     STR   wt    met         no   yes
## 76     RT29B     8 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 77     RT30B     8 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 78     RT31B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 79     RT32B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 80     RT33B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 81     RT34B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 82     RT35B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 83     RT36B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 84     RT37B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 85     RT38B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 86     RT39B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 87     RT40B     9 Healthy    <NA> <NA>   <NA>       <NA>  <NA>
## 89     RT42B    10   preRT     GTR  mut    met        yes    no
## 90     RT43B    10   preRT     GTR  mut    met        yes    no
## 91     RT44B    10   preRT     STR   wt no met        yes    no
## 92     RT45B    10   preRT     GTR  mut    met         no    no
## 93     RT46B    10   preRT     STR   wt no met        yes    no
## 94     RT47B    10   preRT     STR   wt    met        yes   yes
## 95     RT48B    10   preRT     STR   wt no met        yes    no
```

4. Global assessment of normalised and batch corrected data

1. Comparison of data distribution

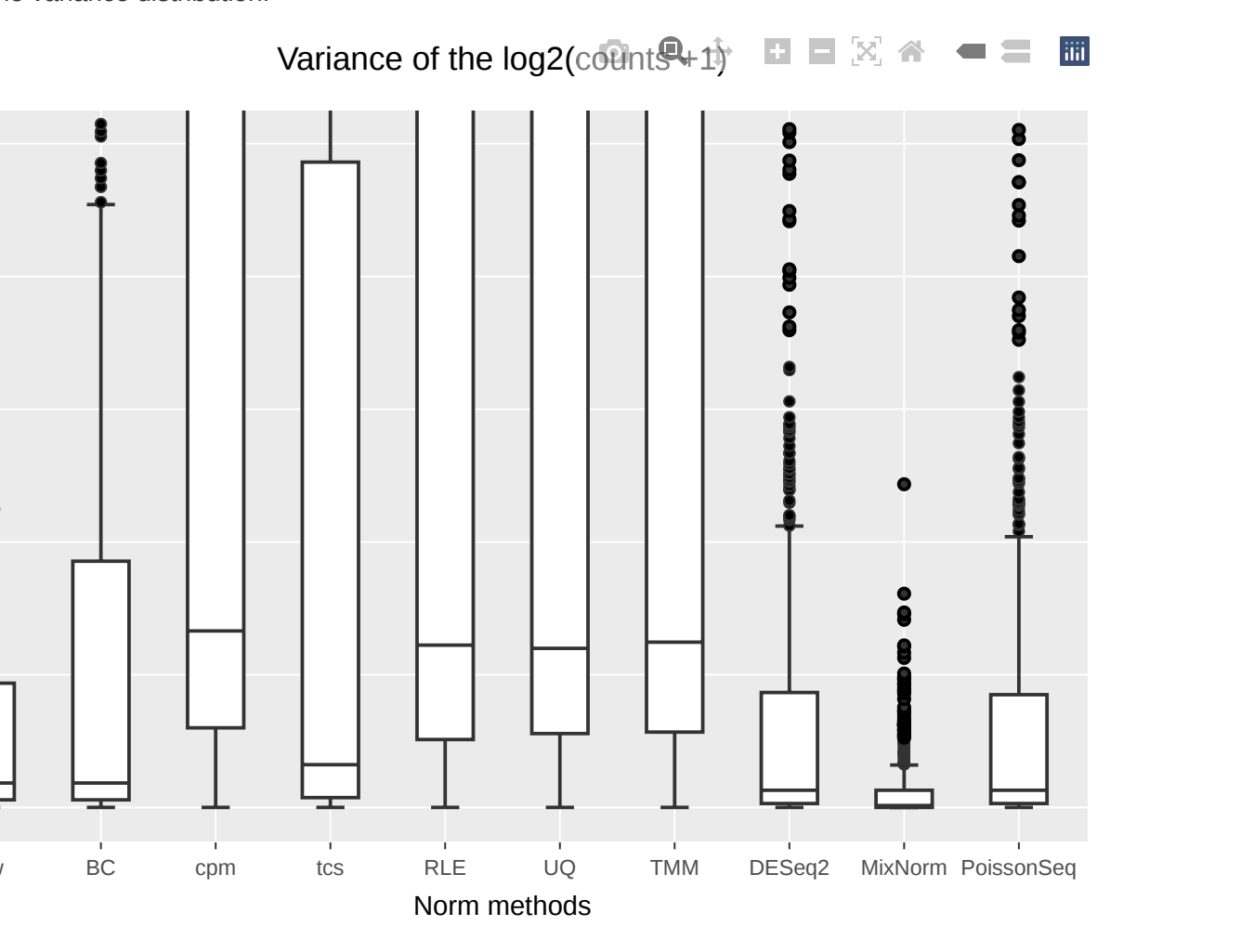
As an illustration of the different normalization methods, the absolute distribution of the miRNA count data following normalization and batch correction can be visualized using density distribution curves.



To avoid problems associated with zero values, the data were log2 transformed after the addition of +1 to all counts. From the density curves of the raw counts, it is evident that there are some inconsistencies between the distribution profiles of the samples. BC shows the distribution of the data that was just batch corrected (no norm method applied). Adjusting the data by total count scaling introduces more variability to the data, whereas all other methods resulted in more similar distribution across all samples. The only exceptions are the samples RT31B and RT37B that show a different distribution in the cpm,RLE,UQ, and TMM normalisation methods.

2. Variance comparison

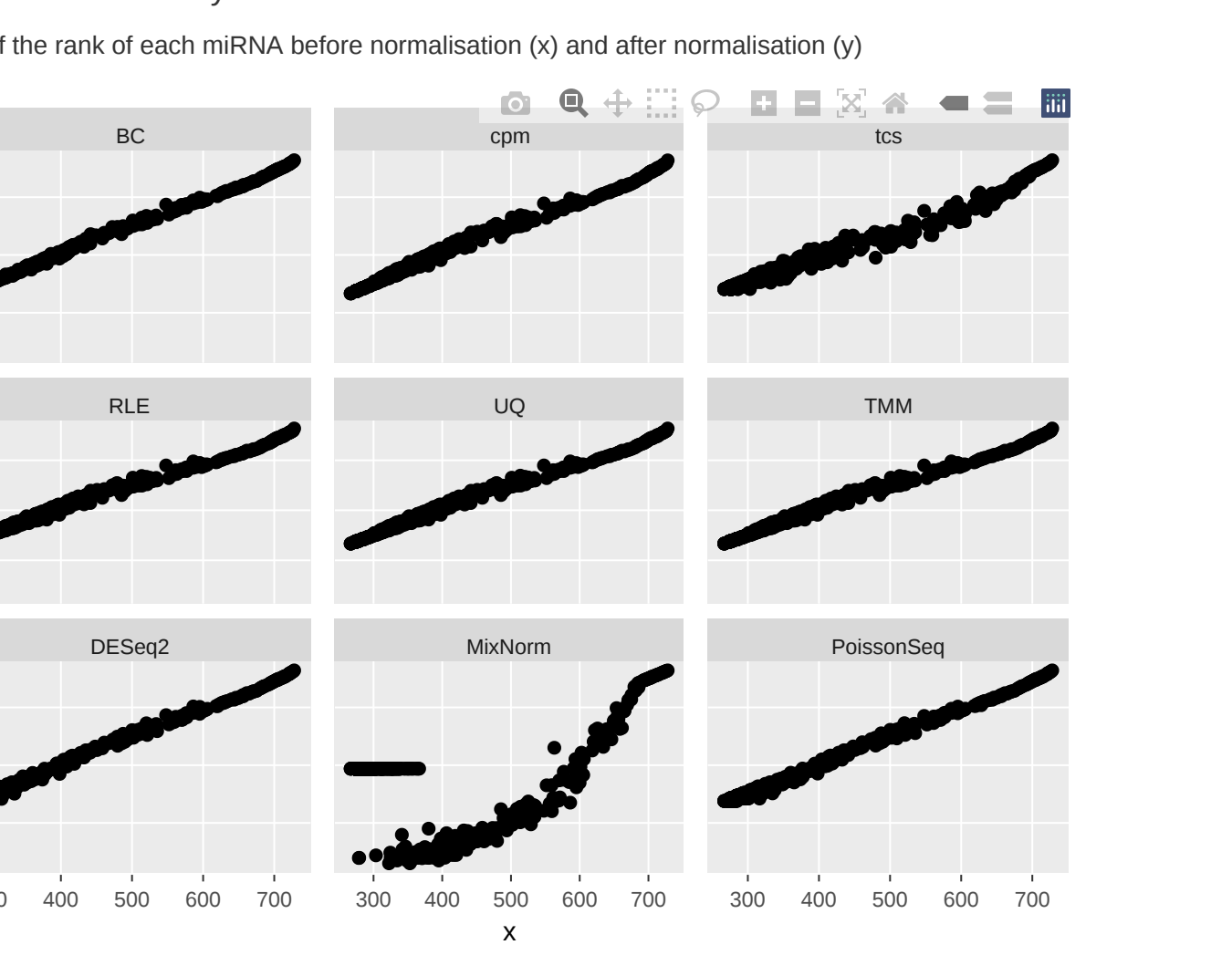
Boxplots of the variance distribution.



The variance of the log2 counts of all miRNAs was computed across the samples, and visualized using boxplots. The data are grouped according to the normalization method. A clear increase in variance is observed in the data normalized by cpm, total count scaling, RLE, UQ, and TMM while a decrease is seen in DeSeq2, MixNorm, or PoissonSeq normalized data.

3. Comparison of linearity of the ranks between raw and normalised data

Scatterplot of the rank of each miRNA before normalisation (x) and after normalisation (y)



here we measure the Pearson correlation between the ranks of each miRNA before and after normalisation:

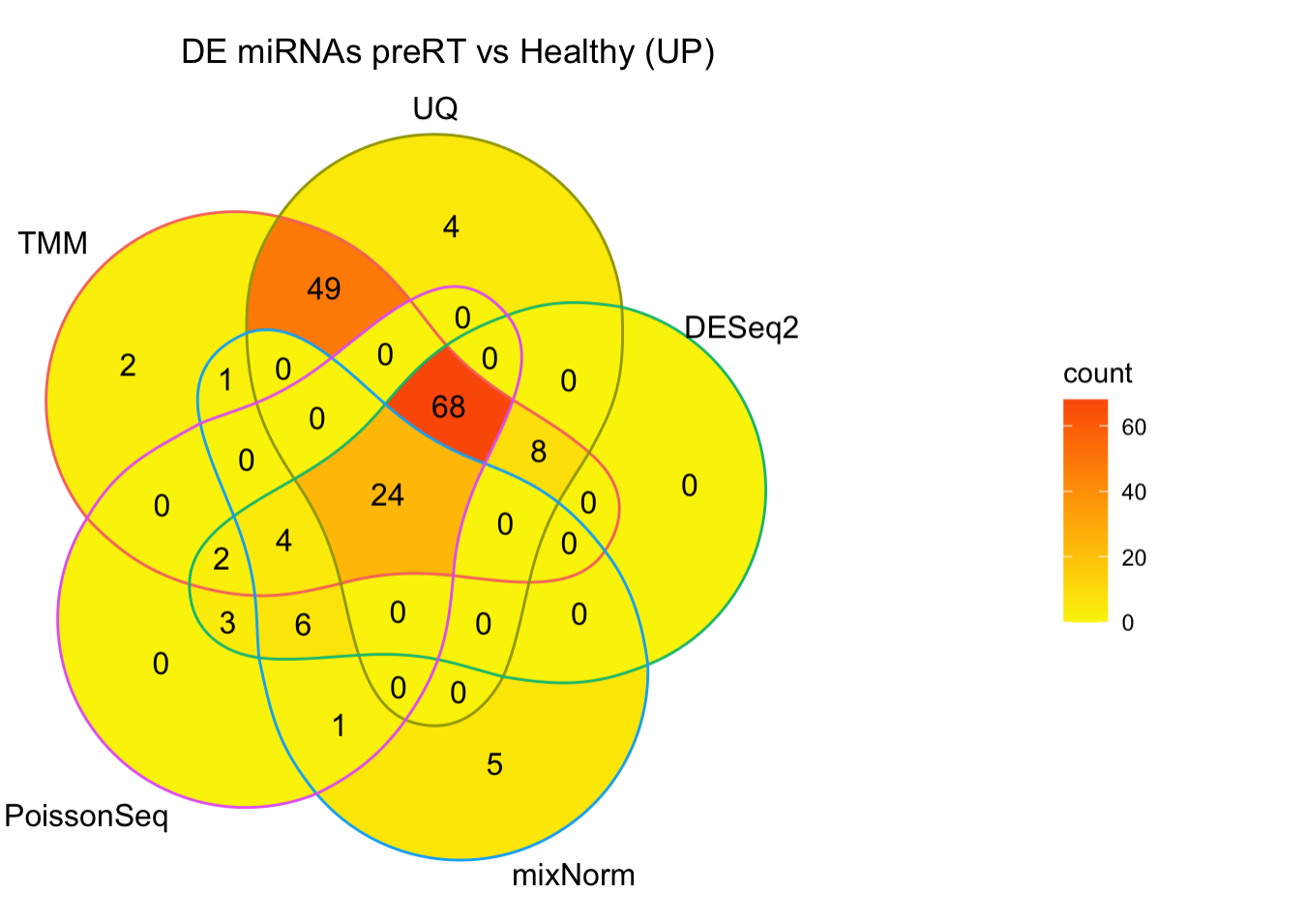
```
##         BC       cpm       tcs       RLE        UQ       TMM    DESeq2
##   99.90440  99.72721  99.13173  99.71686  99.71651  99.71361  99.82691
##    MixNorm PoissonSeq
##   34.71666   99.80020
```

5. DE analysis

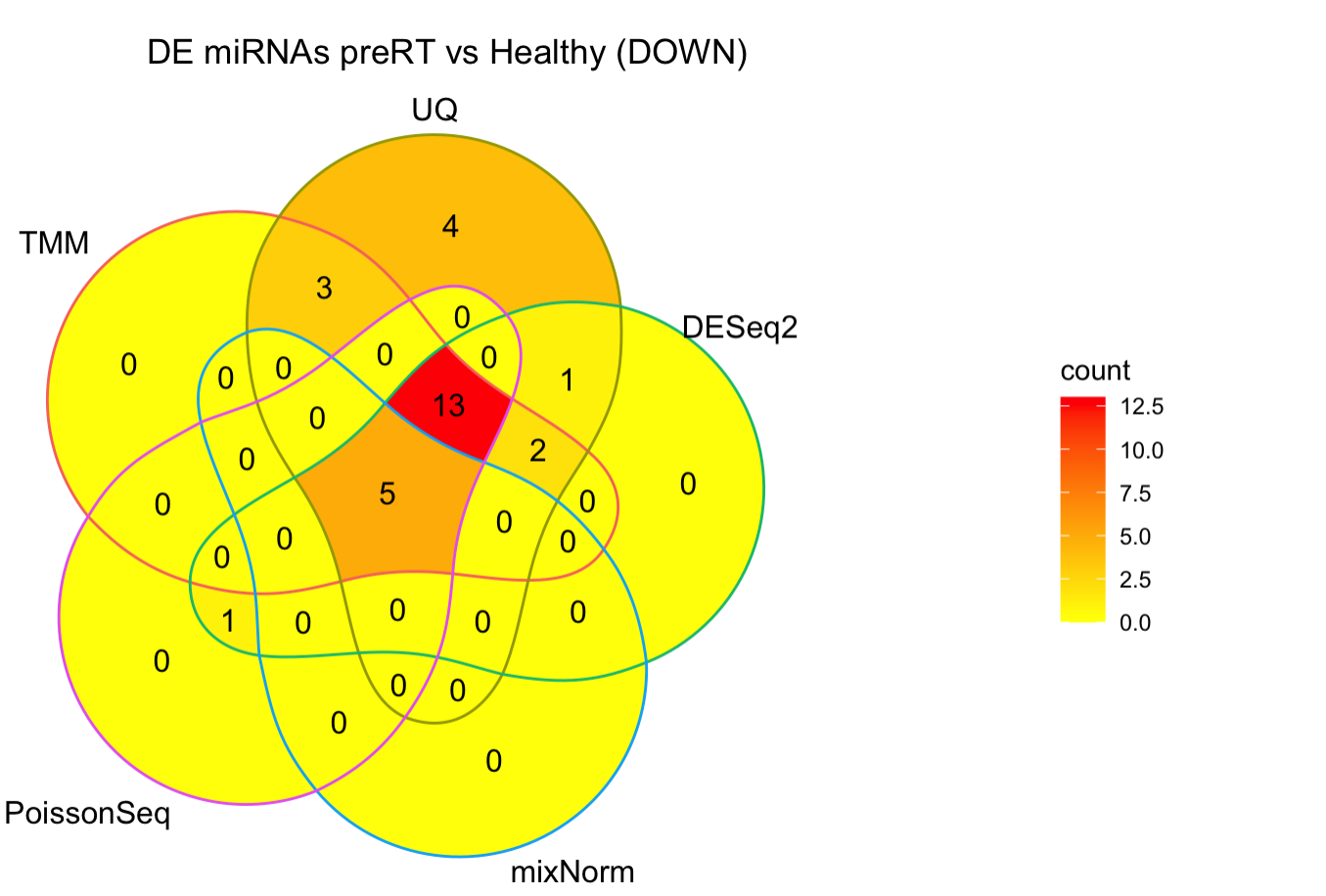
In this part we will look for Differential Expressed miRNAs between pre-RadioTherapy and Healthy samples and across a subset of different normalisation methods. For DE analysis we will use **DESeq** function from the *DESeq2* package with the following parameters:

. adjusted p-value < 0.05

. log2FoldChange > 0 (up-regulated) or < 0 (down-regulated)



This first Venn diagram shows the logical relation between sets of DE miRNAs (up-regulated) across different normalisation methods. A total of 24 miRNAs are Differentially Expressed in 5 Normalisation methods as showed in the plot



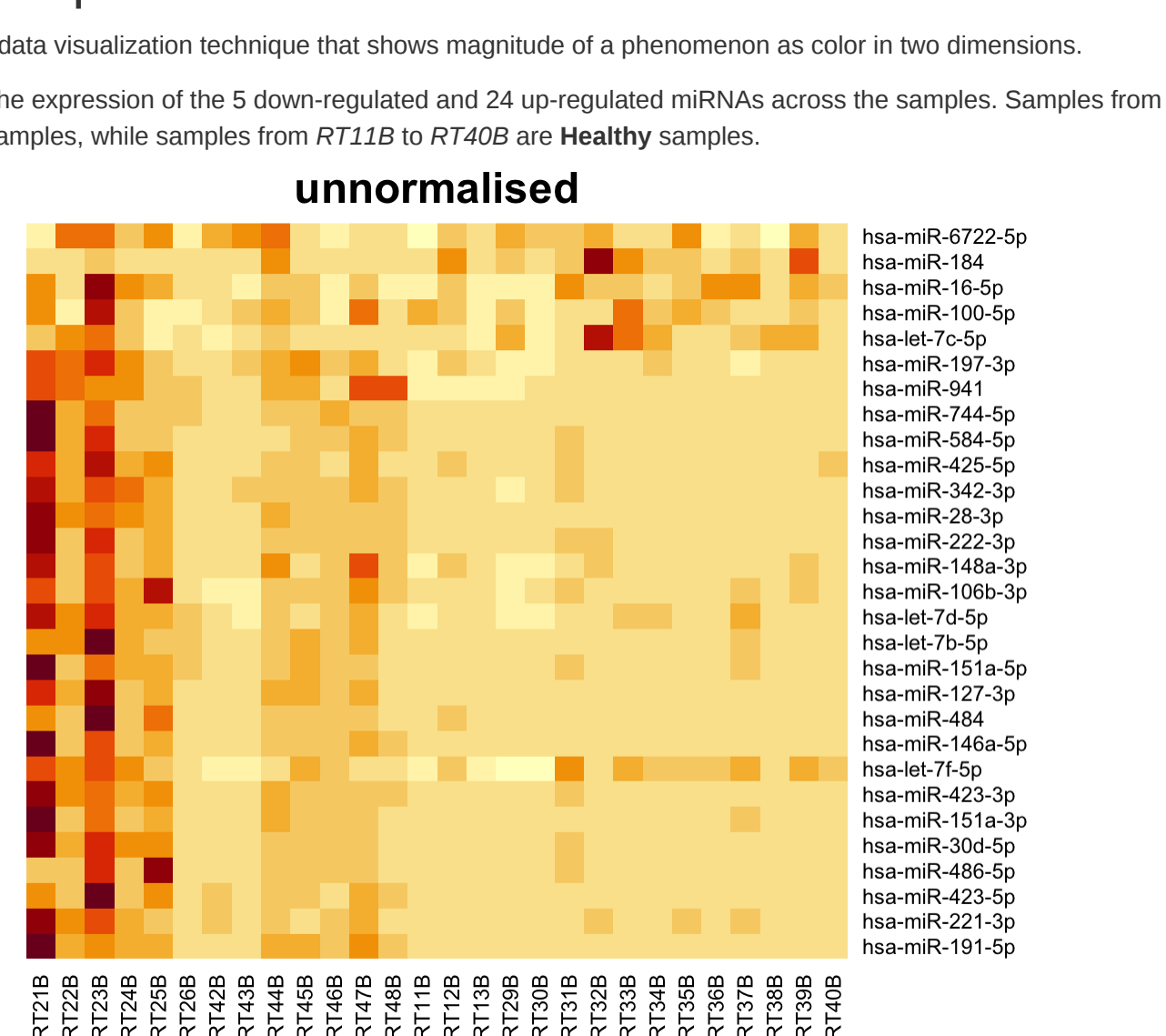
This second Venn diagram shows the logical relation between sets of DE miRNAs (down-regulated) across different normalisation methods. A total of 5 miRNAs are Differentially Expressed in 5 Normalisation methods as showed in the plot

6. Heatmaps

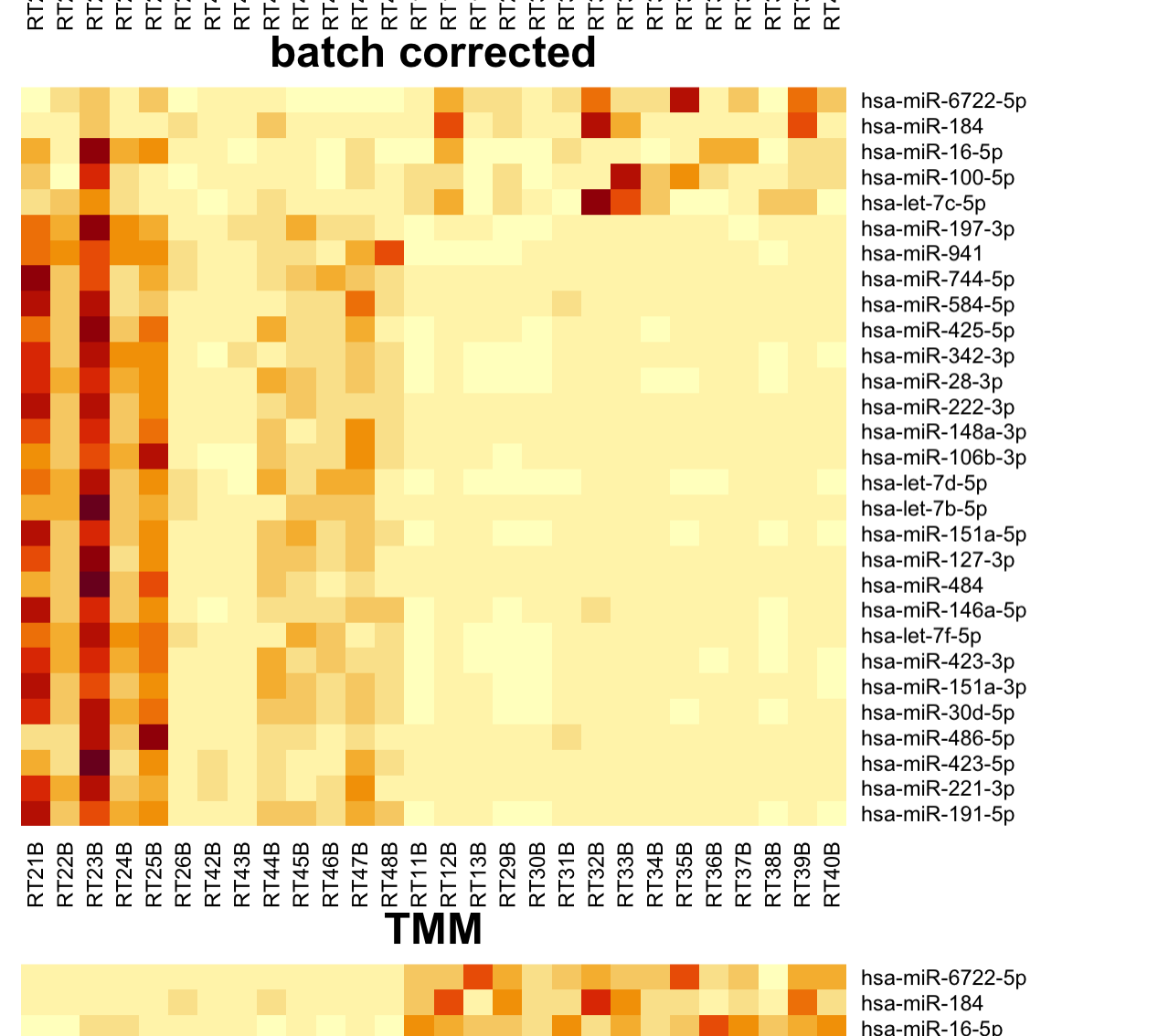
A heatmap is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions.

Here we show the expression of the 5 down-regulated and 24 up-regulated miRNAs across the samples. Samples from *RT21B* to *RT48B* are **pre-radiotherapy** samples, while samples from *RT11B* to *RT40B* are **Healthy** samples.

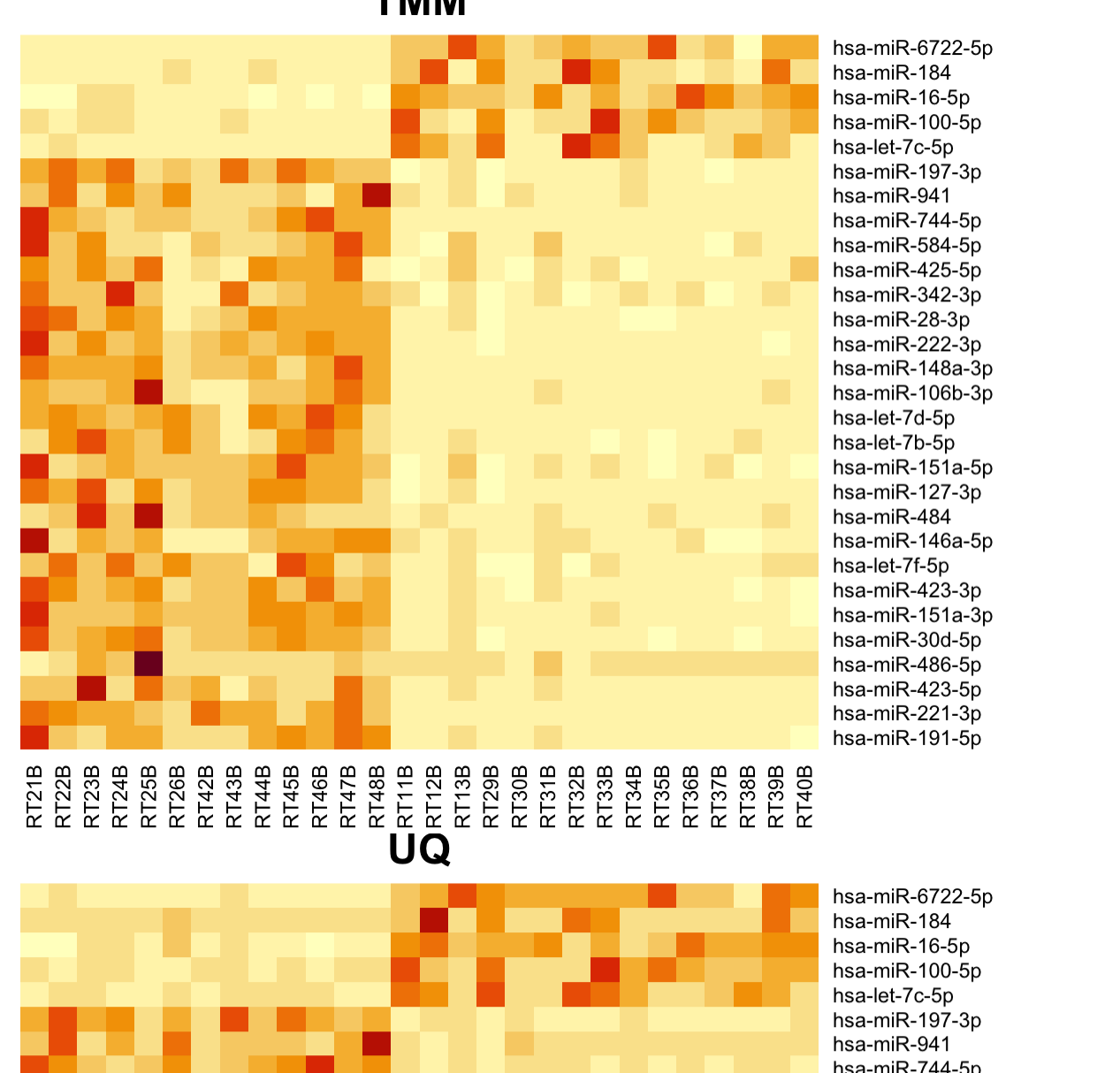
unnormalised



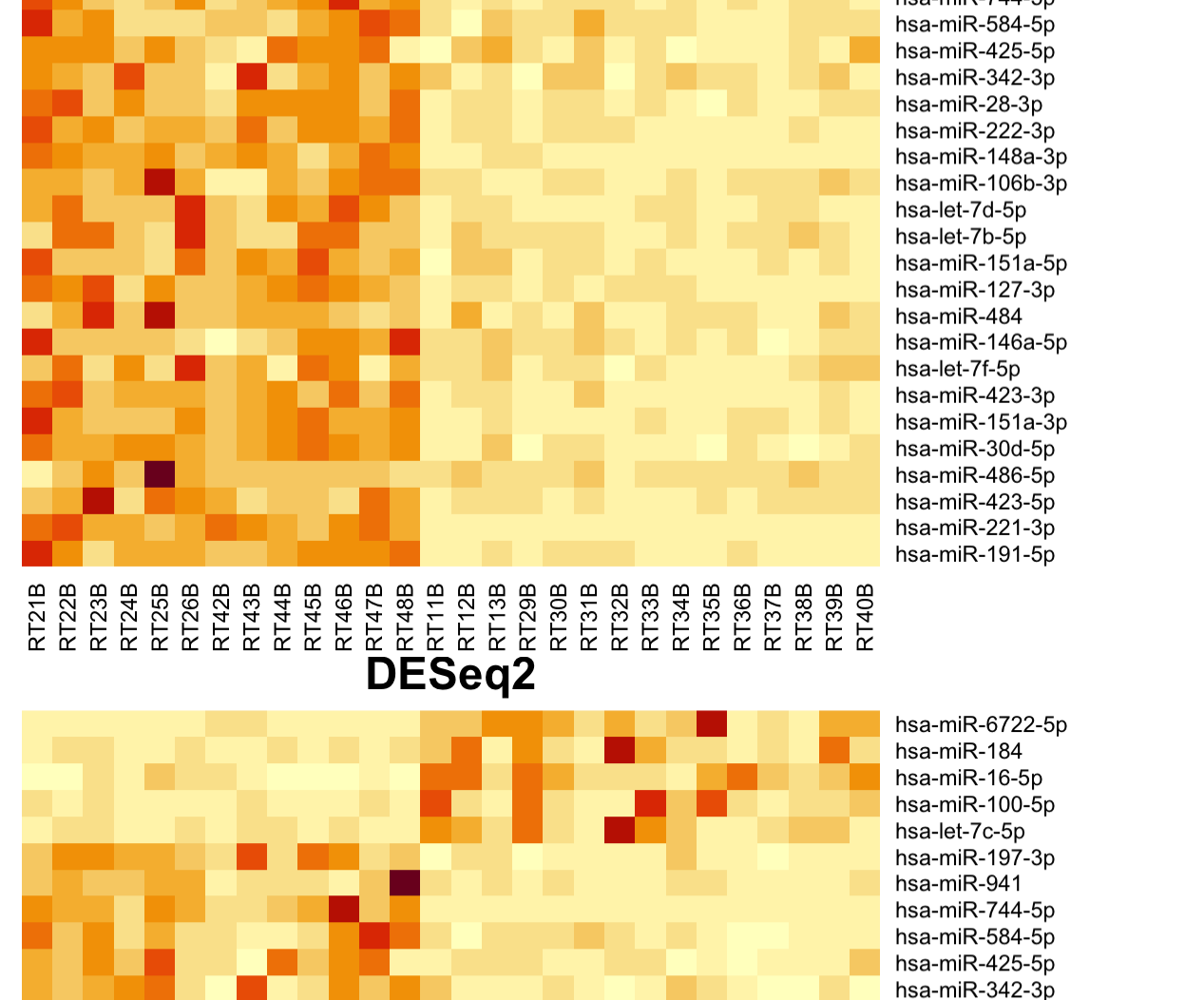
batch corrected



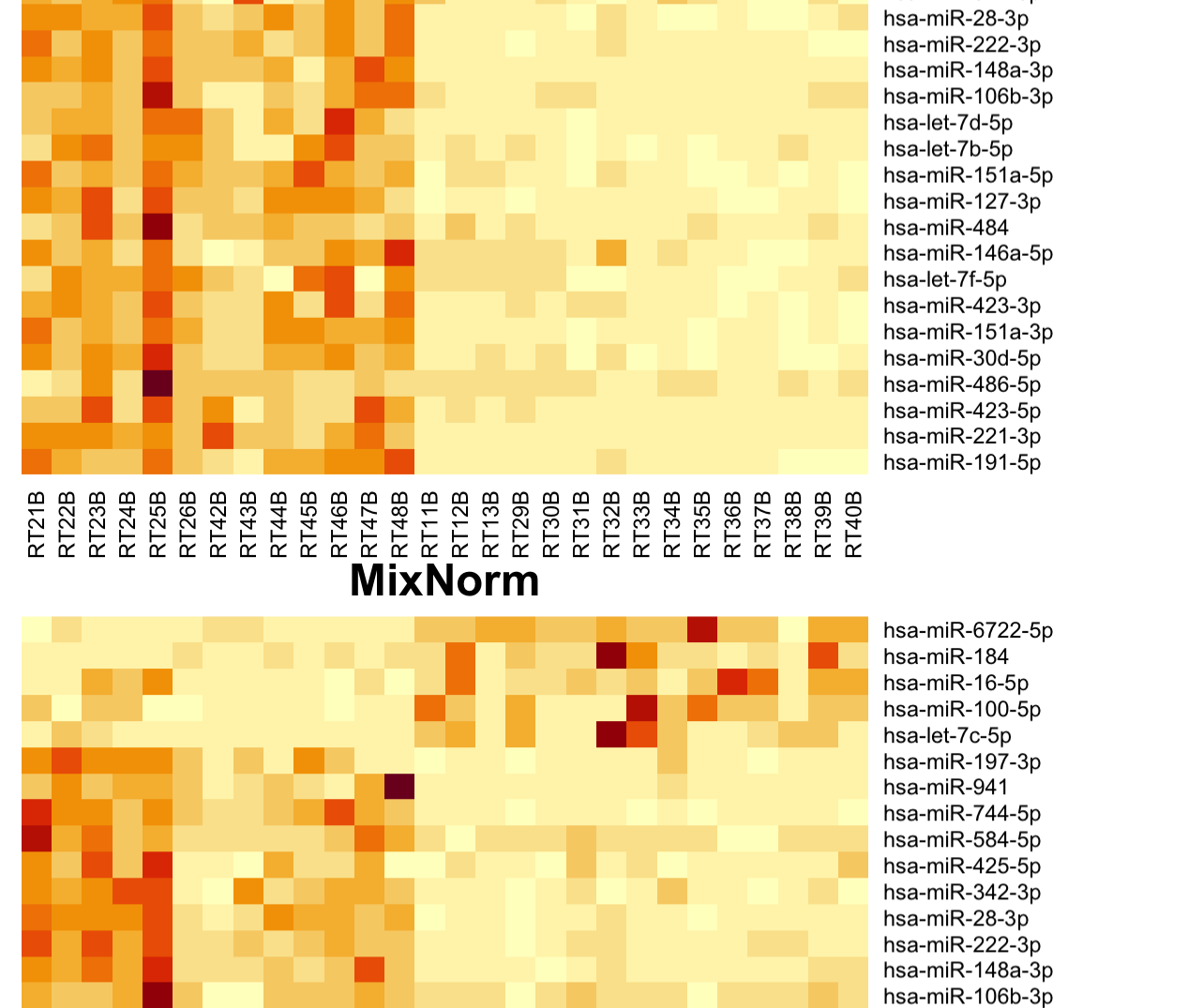
TMM



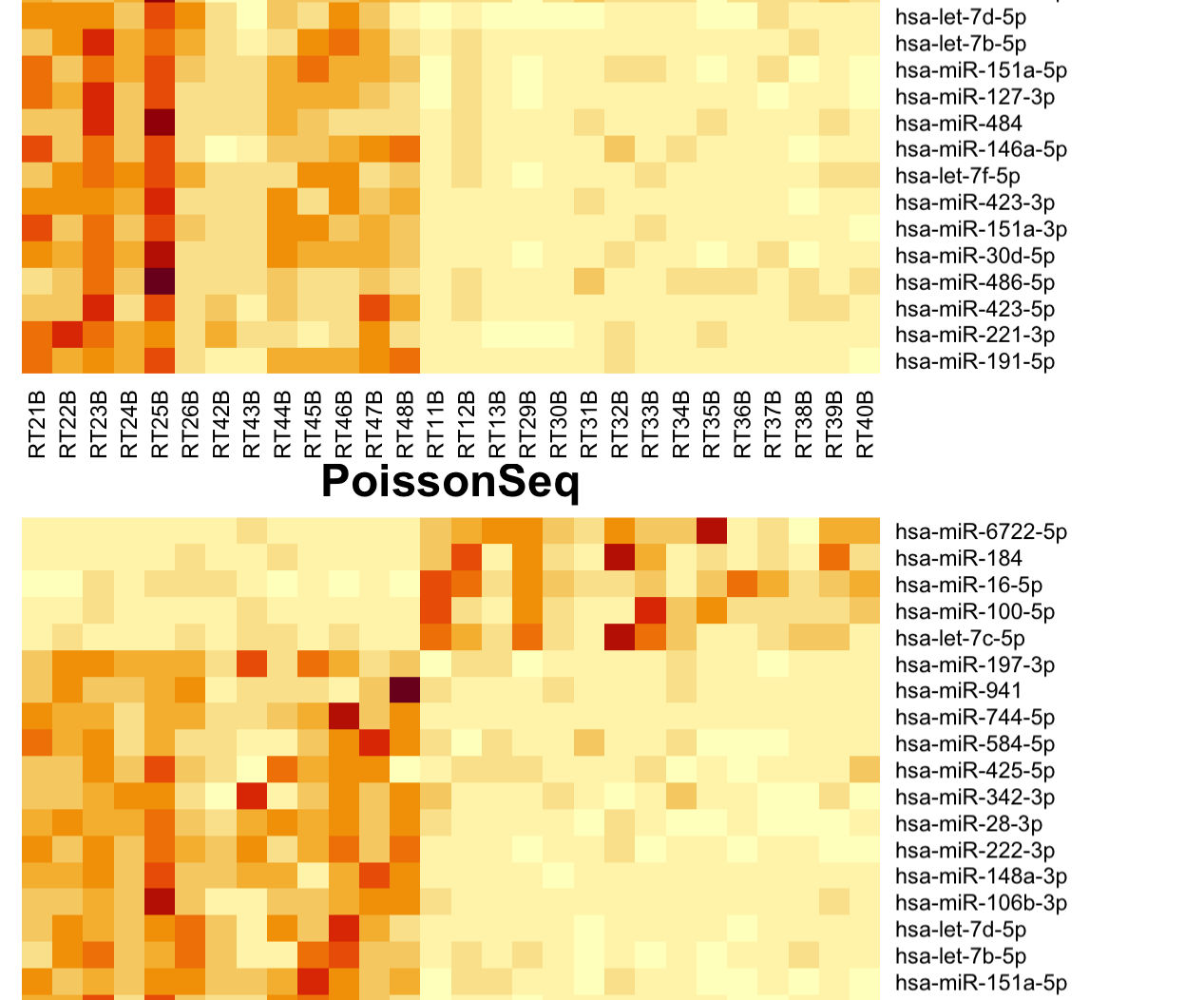
UQ



DESeq2



MixNorm



PoissonSeq