Tutorial: RNA-Seq analysis part IV: Spikes and quality control

February 9, 2010

CLC bio
Finlandsgade 10-12 · 8200 Aarhus N · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

Tutorial: RNA-Seq analysis part IV: Spikes and quality control

This tutorial is the fourth part of a series of tutorials about RNA-Seq analysis. We continue working with the data set introduced in the first tutorial, although in this tutorial we are no longer considering only a subset of the data.

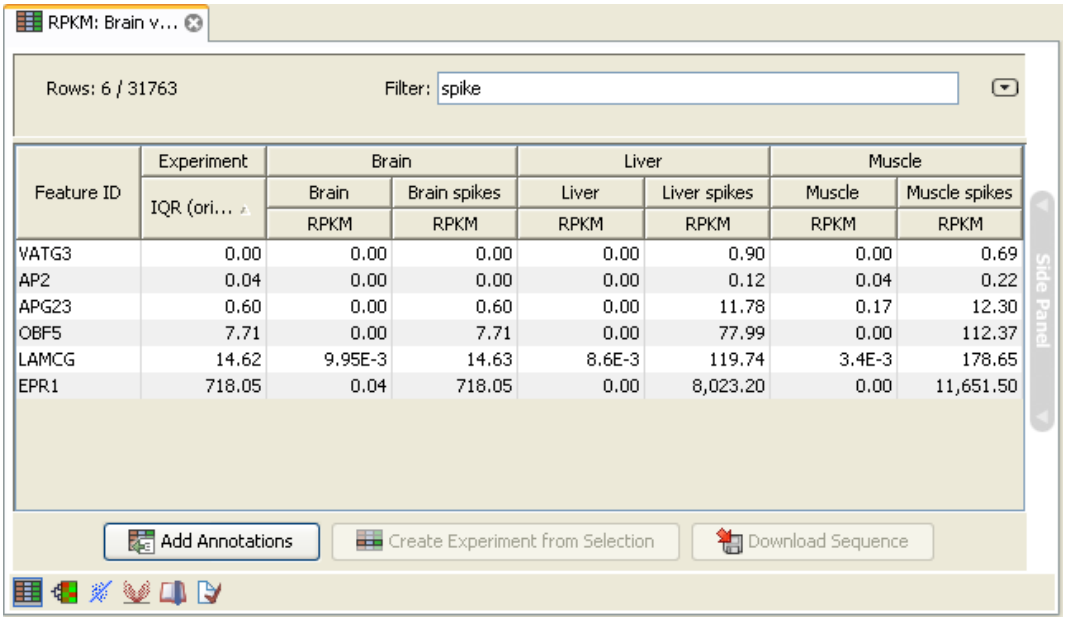
In this tutorial we will focus on quality control. First, we will examine spike-ins in the data set, and second you will learn about general quality control tools.

Inspecting the spike reads

The data set includes six samples that come from three groups corresponding to the three different types of tissue. Within each group of samples, one sample has six spike-in genes, where controlled amounts of mRNA from *Arabidopsis* and phage lambda templates have been introduced in the sample prior to sequencing. We will now inspect the six spike-in genes and check if they are expressed in the spiked samples only as we would expect.

For this and the rest of the tutorial, we will work on the full data set. The data set is so large that on a powerful eight-core workstation computer with 32 GB RAM, it takes more than an hour to run one sample. Therefore, we have performed the analysis beforehand and included the final result with the files you downloaded in the first tutorial. The full RNA-Seq results with contigs of all the mouse genes take up too much space, so they are not included. Instead, we have set up experiments with all six samples.

Open the experiment in the *RPKM* folder and type "spike" in the filter to only include the six spike genes (when we ran the full analysis, we made an artificial sequence called "Spike chromosome" containing the six spike genes). Next, click the **IQR** column header to sort the genes according to the interquartile range of expression. You should now have a view similar to figure 1.



Feature ID	Experiment	Brain		Liver		Muscle	
	IQR (ori...	Brain	Brain spikes	Liver	Liver spikes	Muscle	Muscle spikes
		RPKM	RPKM	RPKM	RPKM	RPKM	RPKM
VATG3	0.00	0.00	0.00	0.00	0.90	0.00	0.69
AP2	0.04	0.00	0.00	0.00	0.12	0.04	0.22
APG23	0.60	0.00	0.60	0.00	11.78	0.17	12.30
OBF5	7.71	0.00	7.71	0.00	77.99	0.00	112.37
LAMCG	14.62	9.95E-3	14.63	8.6E-3	119.74	3.4E-3	178.65
EPR1	718.05	0.04	718.05	0.00	8,023.20	0.00	11,651.50

Figure 1: *Comparing the expression of the spikes.*

The important thing to check here is that we see expression of the spike genes in the spike samples but not in the rest. The six spike transcripts were added in different concentrations to the spike samples, and this is why we see the varying levels of expression.

Select all six rows, and switch to the scatter plot ( ) to visualize this trend. Choose to compare expression value of *Brain spikes* and *Brain*, and you will be able to see some red points in the scatter plot. Change the **Dot type** to **Dot** in the **Side Panel** to see the red dots more clearly. The scatter plot is shown in figure 2.

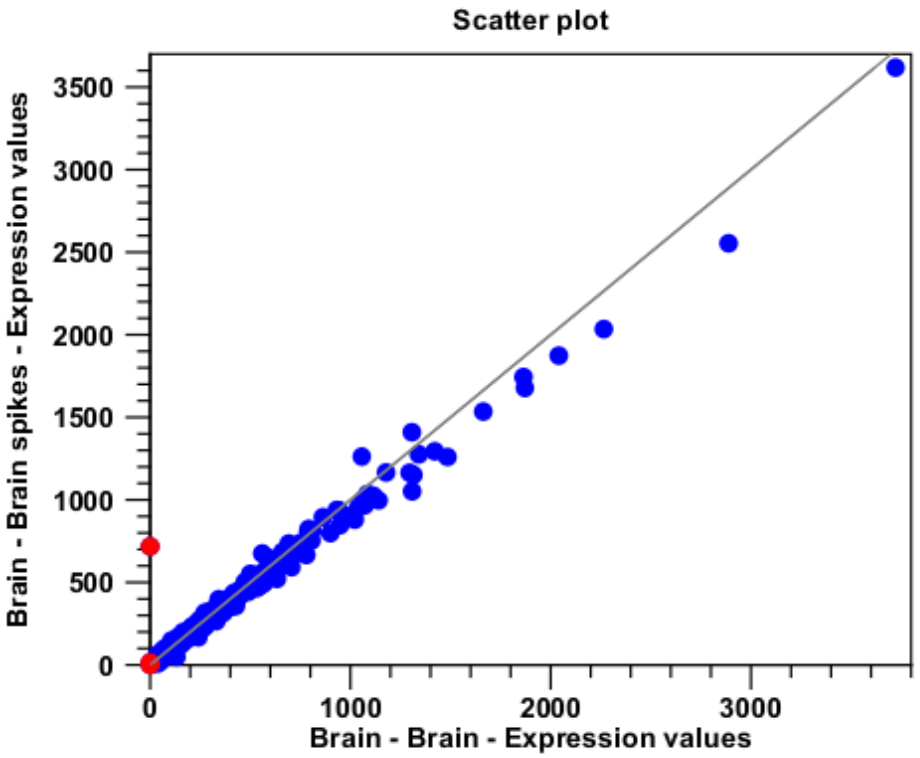


Figure 2: *Comparing the expression of the spikes in a scatter plot.*

At first glance, you only see one clear outlier at the bottom, whereas the other red dot (or rather, 'dots') seem to have more similar expression in the two samples (close to zero in both). But if you **Zoom in** ( ) on these dots, you can see that 3 of them are outliers, although to a smaller extent (see figure 3).

Checking within and between group variability

We have now confirmed that the spike controls look fine (you can check the other two groups in the scatter plot if you like). The next step in our quality control efforts is to check whether the overall variability of the samples reflect their grouping. In other words we want the samples from the same group to be relatively homogenous and distinguishable from the samples from the other groups.

Box plot

To examine and compare the overall distribution of the expression values in the samples you may use a **Box plot** ( ). Box plots can be used to get an overall impression of the locations of the distributions, and to some extend the spread of the distributions.

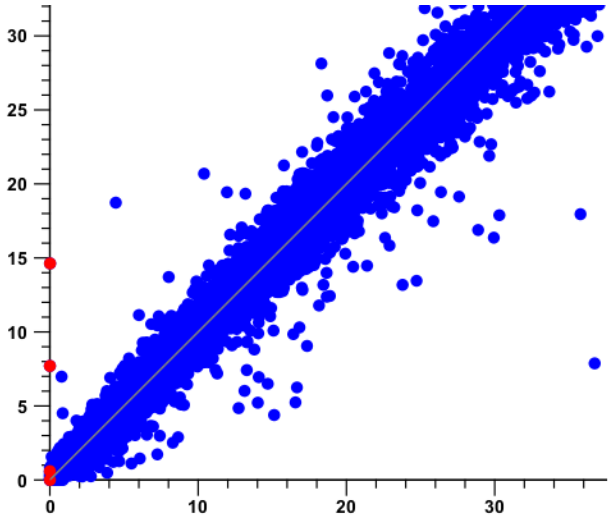


Figure 3: *Comparing the expression of the spikes in a scatter plot, zoomed in.*

First, create a box plot based on the RPKM-based experiment:

Toolbox | **Expression Analysis (**) | **Quality Control** | **Create Box Plot (**)

Select the experiment that is based on RPKM expression values and click **Next** and **Finish**.

The box plot is shown in figure 4.

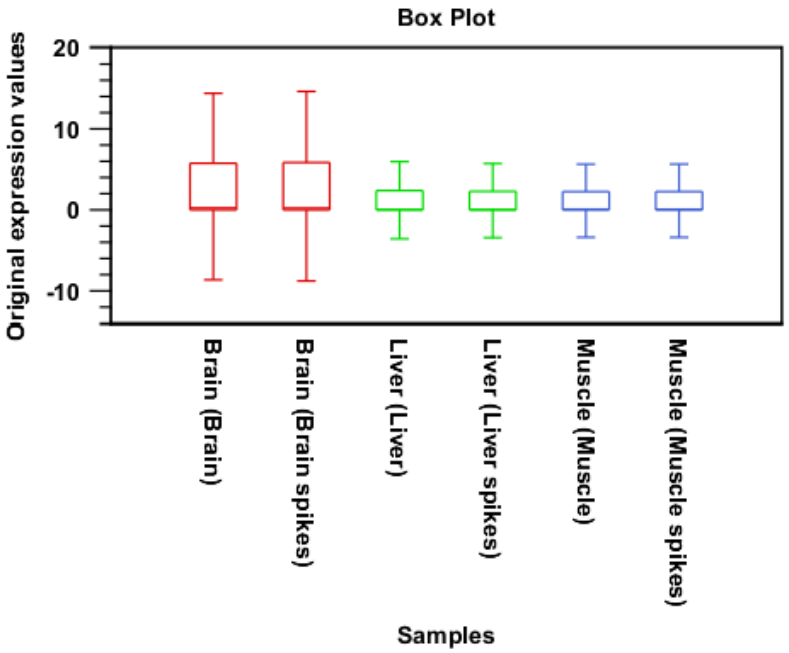


Figure 4: *A box plot of the 6 samples in the experiment, colored by group.*

The median and mean lines (you can display the mean in the **Side Panel**) show the median and mean expression values in the samples and the boxes extend from the p'th to the 100-p'th percentile of the sets of expression values in the samples. Thus, the box plot view is rather

sensitive to the choice of the percentile value, and you may get a better impression of how the distributions compare by trying different percentiles.

The distributions of the expression values are dominated by a lot of really small values and much fewer but much larger values. To diminish the effect on the box plots of the few very high values, you can square root transform the values and create box plots for the transformed values. First, transform the values in the experiment:

> **Toolbox** | **Expression Analysis** ( ) | **Transformation and Normalization** | **Transform** ( )

Select the experiment, click **Next** and select **Square root transformation**. Click **Finish**. Now, create a new box plot, but this time make sure to select **Transformed expression values** in the second step. Figure 5 shows the box plot before and after square root transformation.

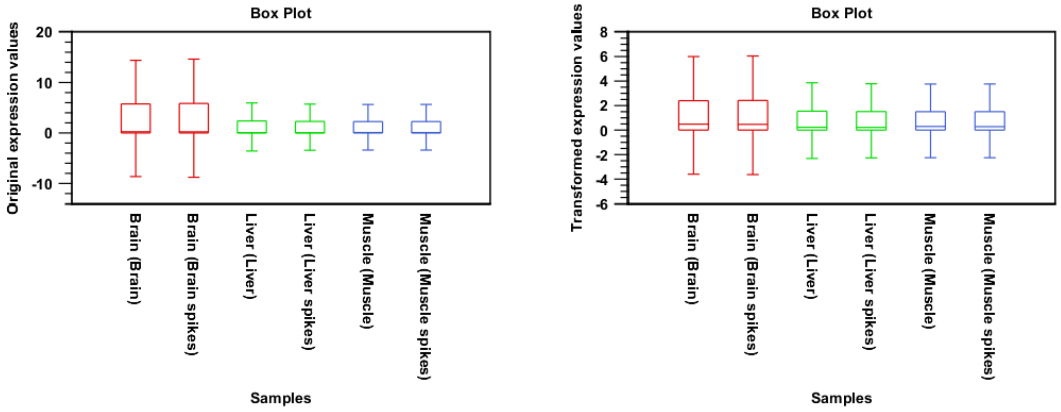


Figure 5: *To the left: original box plot based on RPKM values. To the right: box plot based on square root transformed expression values.*

After square root transformation, the distributions appear more a bit more similar.

Now, create similar box plots for the experiment based on total exon reads. You can see the result in figure 6 that again shows the box plot before and after square root transformation, this time based on total exon reads rather than RPKM.

Overall, the box plots indicate that the *locations* of the distributions of the expression values in the samples are similar both for RPKM and for Total Exon Reads, but there is considerable difference in the *spread* of the values. The distributions of RPKM values for samples for the same tissue are highly similar. This is expected, as the samples are technical replicates and the sample size is factored out in the RPKM (see part III of the tutorials). The high variability of the 'Total exon reads' counts is obviously related to the numbers of (mapped) reads in the samples.

Principal component analysis (PCA)

We continue to work with the experiment based on RPKM expression values. Next, we perform a **Principal Component Analysis (PCA)**:

> **Toolbox** | **Expression Analysis** ( ) | **Quality Control** | **Principal Component Analysis** ( )

Select the experiment, click **Next** and **Finish**. This will create a PCA plot as shown in figure 7.

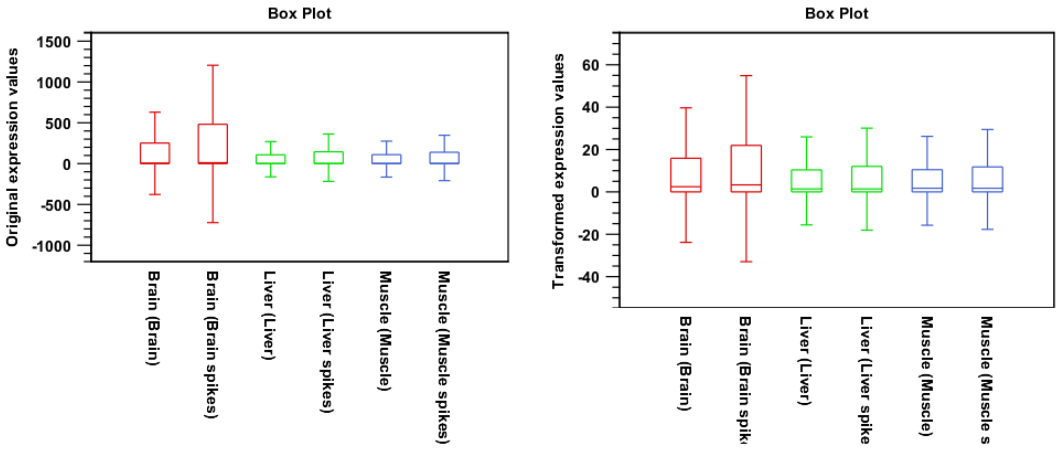


Figure 6: *To the left: original box plot based on total exon reads. To the right: box plot based on square root transformed expression values.*

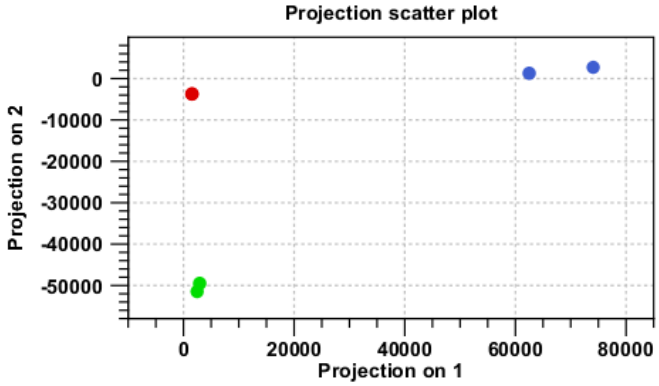


Figure 7: *A principal component analysis colored by group.*

The plot shows the projection of the samples onto the two-dimensional space spanned by the first and second principal component. (These are the orthogonal directions in which the data exhibits the largest and second-largest variability).

The dots are colored according to the groups, and they group very nicely in the plot (the two red red dots are on top of each other).

You can display the names of the samples in the plot using the settings under **Dot properties** in the **Side Panel** to the right of the view (see the result in figure 8).

This PCA was done on RPKM-based expression values. As you saw in the previous tutorial, the RPKM normalizes for the sample size. In order to show the importance of doing this for this kind of analysis, perform a new PCA on the experiment located in the *total exon reads* folder.

The result is shown in figure 9.

Although the replicates still group together, the grouping is not near as clear as in figure 8 which uses the RPKM-based expression values.

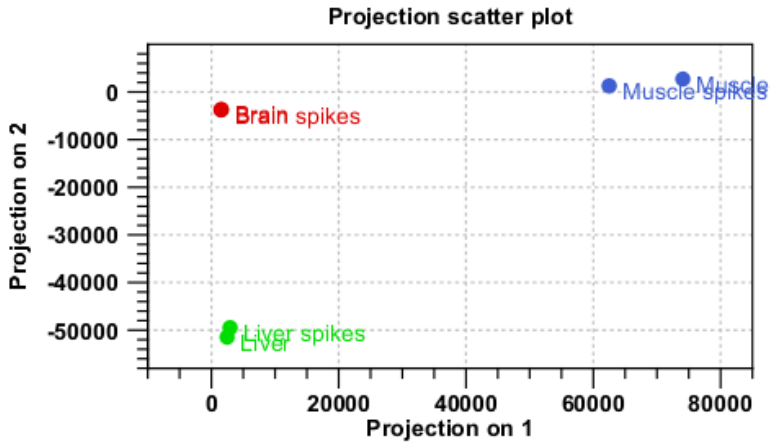


Figure 8: *Displaying the sample names.*

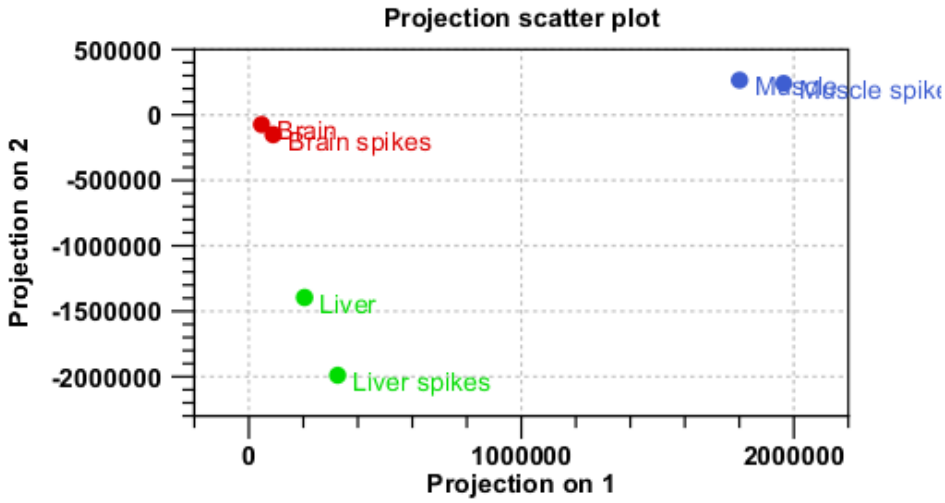


Figure 9: *PCA plot using total exon reads.*

Hierarchical clustering

In order to complement the principal component analysis, we will also do a hierarchical clustering of the samples to see if the samples cluster in the groups we expect:

> **Toolbox** | **Expression Analysis** ( ) | **Quality Control** | **Hierarchical Clustering of Samples** ( )

Select the experiment from the *RPKM* folder and click **Next**. Leave the parameters at their default and click **Finish**.

This will display a heat map showing the clustering of samples at the bottom (see figure 10).

The replicates cluster nicely together as expected, and it looks like the pattern of expression is more similar between brain and liver than between muscle and the other two groups.

Note that the heat map is not a new element to be stored in the **Navigation Area** - it is just

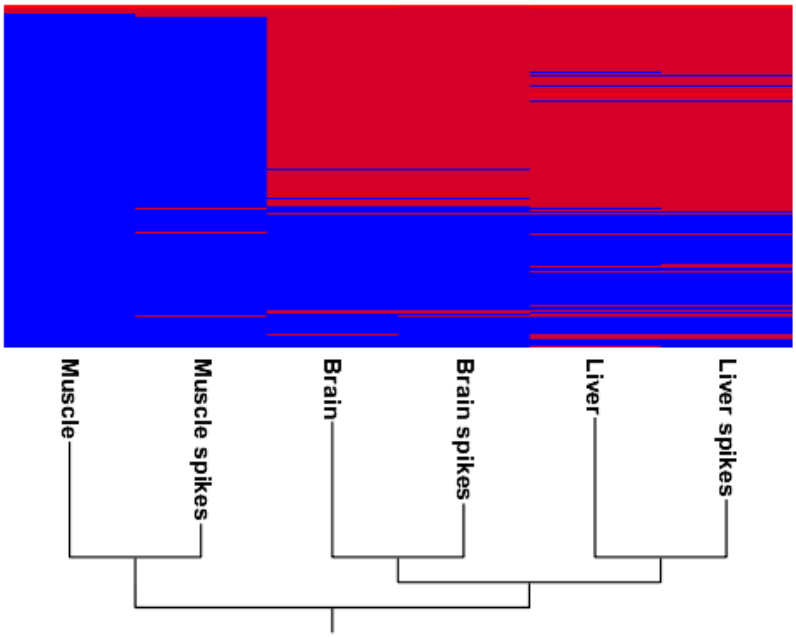


Figure 10: *Sample clustering. Adjusting the settings in the Side Panel to put the sample names at the bottom.*

another way of looking at the experiment (note the buttons to switch between different views in figure 11.



Figure 11: *Different views on an experiment.*

As a conclusion to this fourth tutorial on RNA-Seq, we can confirm that the spike signals are clear and unambiguous, and we can conclude that both the PCA and hierarchical clustering shows that the replicates are fairly homogenous. Just as in the previous tutorial, we can see the effect of using RPKM over total exon reads as expression measure. When doing this kind of quality control, it may be advisory to use the RPKM expression value as this is implicitly standardized between samples (see tutorial part II). Alternatively, you can choose to use the total exon reads counts and then normalize them before performing quality control (you find the **Normalize** ( ) tool here: Toolbox -> Expression Analysis -> Transformation and Normalization).

One of the reasons why the samples cluster so nicely is that there is no biological variation in the samples. The two samples in each group are technical replicates, so what we are really measuring is the quality of the method. In experiments with biological replicates, you would expect to see much more intra-group variability.