Tutorial: Microarray-based expression analysis part II: Quality control

December 15, 2009

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: Microarray-based expression analysis part II: Quality control

This tutorial is the second part of a series of tutorials about expression analysis. We continue working with the data set introduced in the first tutorial.

In this tutorial we will examine various methods to perform quality control of the data.

Transformation

First we inspect to what extent the variance in expression values depends on the mean. For this we create an **MA Plot**:

Toolbox | **Expression Analysis ( )** | **General Plots** | **Create MA Plot ( )**

Since the MA plot compares two samples, select two of the 12 arrays and click **Finish**. This will show a plot similar to the one shown in figure 1.

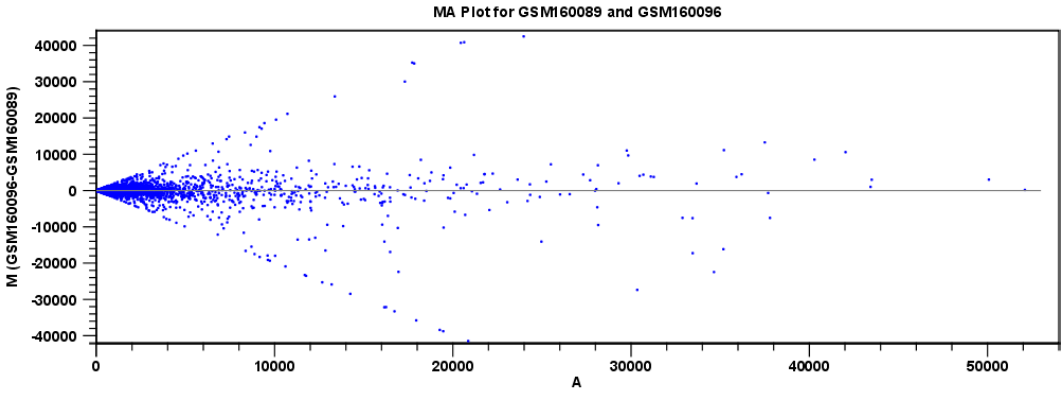


Figure 1: *MA plot before transformation.*

The X axis shows the mean expression level of a feature on the two arrays and the Y axis shows the difference in expression levels for a feature on the two arrays. From the plot shown in figure **??** it is clear that the variance increases with the mean. To remove some of the dependency, we want to **transform** the data:

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

Select the same arrays used for the plot, click **Next**, choose **Log 2** transformation and click **Finish**. Now create an MA plot again as described above, but when you click **Next** you can see that you now also have the option to choose **Transformed expression values** (see figure 2).

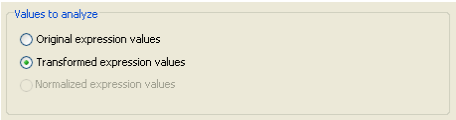


Figure 2: *Select the transformed expression values.*

Select the transformed values. You will see that these three selection boxes; Original, Transformed and Normalized expression values are used several places when expression values are used in a calculation.

Click **Finish**.

This will result in a quite different plot as shown in figure 3.

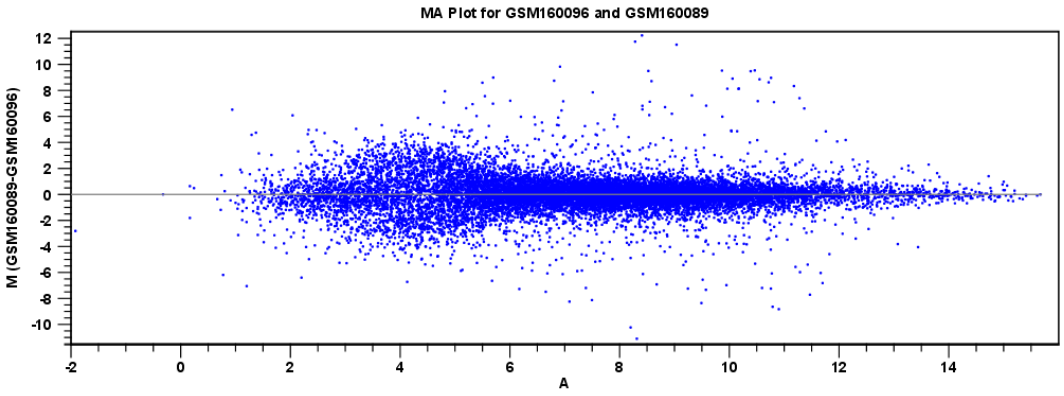
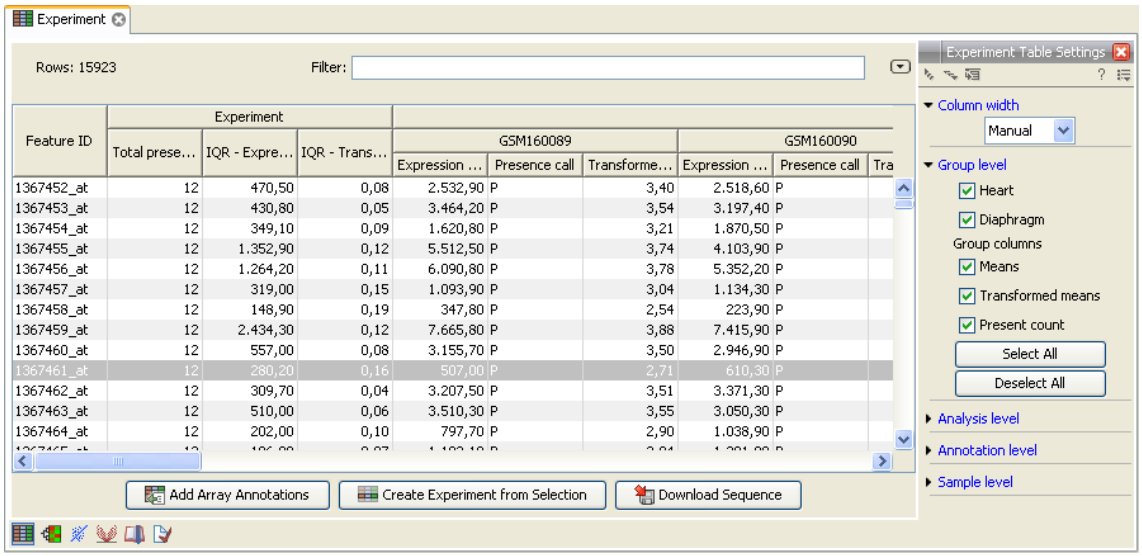


Figure 3: *MA plot after transformation.*

The much more symmetric and even spread indicates that the dependance of the variance on the mean is not as strong as it was before transformation.

We have now only transformed the values of the two samples used for the MA plot. The next step is to transform the expression values within the experiment, since this is the data we are going to use in the further analysis. The procedure is similar to before - this time you just select the experiment created in the first part of this tutorial series instead of the two arrays.

If you open the table, you will see that all the samples have an extra column with transformed expression values (see figure 4).



Feature ID	Experiment			GSM160089			GSM160090		
	Total prese...	IQR - Expre...	IQR - Trans...	Expression ...	Presence call	Transforme...	Expression ...	Presence call	Tra
1367452_at	12	470,50	0,08	2.532,90	P	3,40	2.518,60	P	
1367453_at	12	430,80	0,05	3.464,20	P	3,54	3.197,40	P	
1367454_at	12	349,10	0,09	1.620,80	P	3,21	1.870,50	P	
1367455_at	12	1.352,90	0,12	5.512,50	P	3,74	4.103,90	P	
1367456_at	12	1.264,20	0,11	6.090,80	P	3,78	5.352,20	P	
1367457_at	12	319,00	0,15	1.093,90	P	3,04	1.134,30	P	
1367458_at	12	148,90	0,19	347,80	P	2,54	223,90	P	
1367459_at	12	2.434,30	0,12	7.665,80	P	3,88	7.415,90	P	
1367460_at	12	557,00	0,08	3.155,70	P	3,50	2.946,90	P	
1367461_at	12	280,20	0,16	507,00	P	2,71	610,30	P	
1367462_at	12	309,70	0,04	3.207,50	P	3,51	3.371,30	P	
1367463_at	12	510,00	0,06	3.510,30	P	3,55	3.050,30	P	
1367464_at	12	202,00	0,10	797,70	P	2,90	1.038,90	P	

Figure 4: *Transformed expression values have been added to the table.*

There is also an extra column for transformed group means and transformed IQR.

Comparing spread and distribution

In order to perform meaningful statistical analysis and inferences from the data, you need to ensure that the samples are comparable. Systematic differences between the samples that are likely to be due to noise (e.g. differences in sample preparation and processing) rather than true biological variability should be removed. To examine and compare the overall distribution of the transformed expression values in the samples you may use a **Box plot** ( ):

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

Select the experiment and click **Next**. Choose the **Transformed expression values** and **Finish**.

The box plot is shown in figure 5.

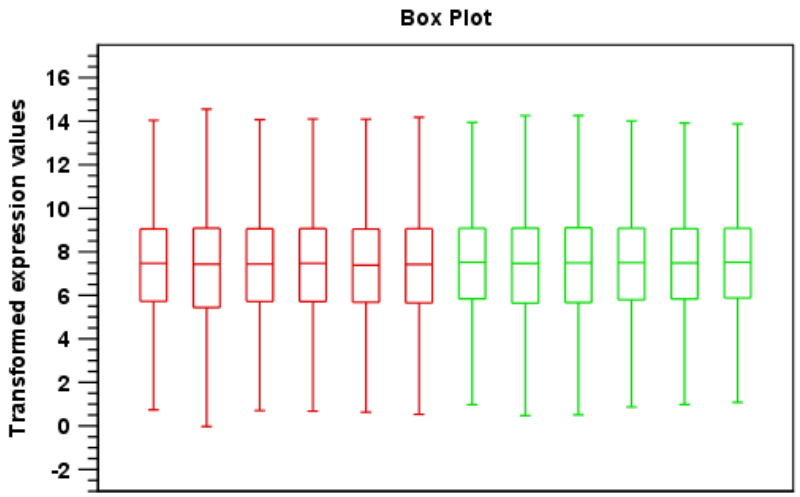


Figure 5: *A box plot of the 12 samples in the experiment, colored by group.*

This plot looks very good because none of the samples stands out from the rest. If you compare this plot to the one shown in figure 6 from another data set, you can see the difference.

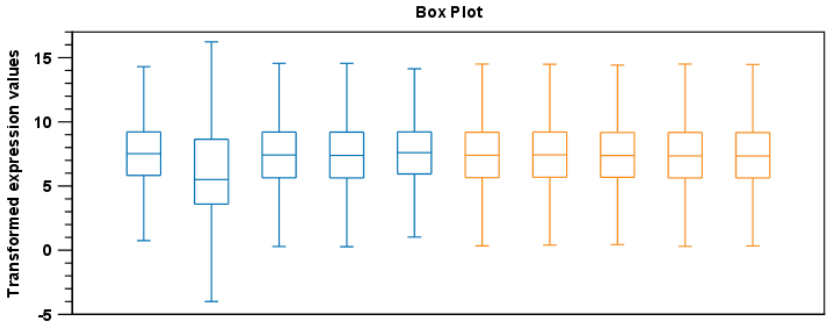


Figure 6: *A box plot showing one sample that stands out from the rest.*

The second sample from the left has a distribution that is quite different from the others. If you have a data set like this, then you should consider either to remove the bad quality sample.

Group differentiation

The next step in the quality control is to check whether the overall variability of the samples reflect their grouping. In other words we want the replicates to be relatively homogenous and distinguishable from the samples from the other group.

First, we perform a **Principal Component Analysis (PCA)**:

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

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

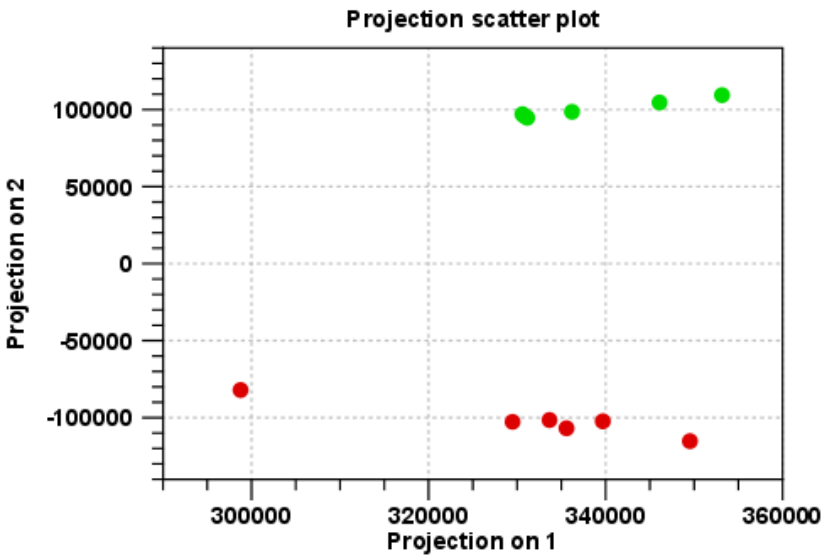


Figure 7: *A principal component analysis colored by group.*

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 also group very nicely in the plot. There is only one outlier - to see which sample it is, place the mouse cursor on the dot for a second, and you will see that it is the *GSM160090* from the *Heart* group.

You can display this information in the plot using the settings in the **Side Panel** to the right of the view:

> **Dot properties** | **select GSM160090 in the drop-down box** | **Show names**

In this way you can control the coloring and dot types of the different samples and groups (see figure 8).

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** ( )

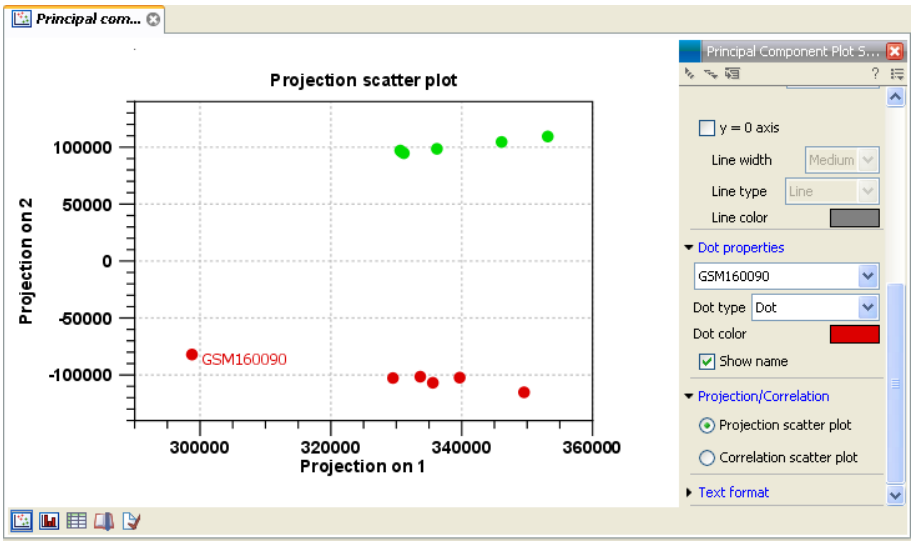


Figure 8: *Naming the outlier.*

Select the experiment 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 9).

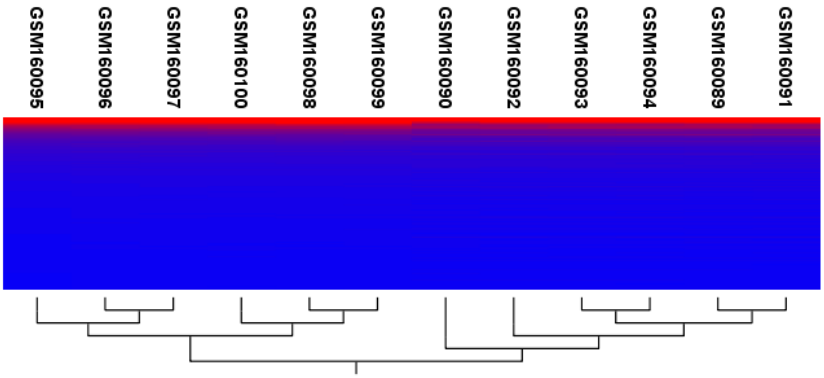


Figure 9: *Sample clustering.*

The two overall groups formed are identical to the grouping in the experiment. You can double-check by placing your mouse on the name of the sample - that will show which group it belongs to.

Since both the principal component analysis and the hierarchical clustering confirms the grouping of the samples, we have no reason to be sceptical about the quality of the samples and we conclude that the data is OK.

Note that the heat map is not a new element to be stored in the **Navigation Area** - it is just another way of looking at the experiment (note the buttons to switch between different views in figure 10.

In part III of the tutorial series we will be looking into the different views in more detail.

To summarize this part about quality control, it looks like the data have good quality, and we are now ready to proceed to the next step where we do some statistical analysis to see which genes are differentially expressed.



Figure 10: *Different views on an experiment.*