Tutorial: Microarray-based expression analysis part III: Differentially expressed genes

August 23, 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

Tutorial: Microarray-based expression analysis part III: Differentially expressed genes

This tutorial is the third 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 identify and investigate the genes that are differentially expressed.

Statistical analysis

First we will carry out some statistical tests that we will use to identify the genes that are differentially expressed between the two groups:

Toolbox | **Expression Analysis** ( ) | **Statistical Analysis** | **On Gaussian Data** ( )

Select the experiment created in part I of the tutorials and click **Next**. Leave the parameters at the default and click **Next** again. You will now see a dialog as shown in figure 1.

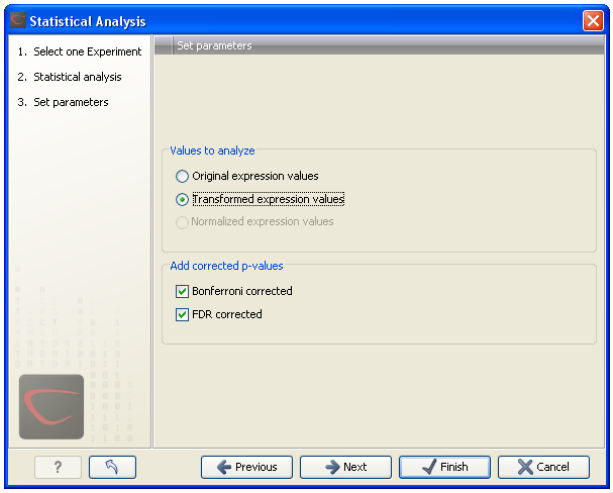

Figure 1: *Statistical analysis.*

As shown in figure 1 select the transformed expression values and check the two corrected p-values as well. You can read more about what they mean by clicking the **Help** (?) button in the dialog.

When you press **Finish**, a number of extra columns will be added to your experiment. For this analysis we will use the FDR p-value which is a measure that allows us to control how big a proportion of false positives (genes that we think are differentially expressed but really are not) we are willing to accept.

Click the **FDR p-value correction** column to sort it with the lowest values at the top. If you scroll down to values around 5E-4 you can clearly see the difference between using the FDR p-value and the Bonferroni-corrected p-value which is much stricter (p-values approaching 1 - see figure 2).

Filtering p-values

To do a more refined selection of the genes that we believe to be differentially expressed, we use the advanced filter located at the top of the experiment table. Click the **Advanced Filter** ( ) button and you will see that the simple text-based filter is now replaced with a more advanced

1373383_at	12	487,90	314,20	294,15	1,41	294,15	1,41	6,80	4,76E-5	5,31E-4	0,76
1384164_at	9	250,40	137,10	-137,50	-2,07	-137,50	-2,07	-6,80	4,76E-5	5,31E-4	0,76
1376058_at	12	2.835,50	1.787,30	1.740,20	2,60	1.740,20	2,60	6,79	4,77E-5	5,32E-4	0,76
1376813_at	12	553,60	350,80	335,75	1,60	335,75	1,60	6,79	4,78E-5	5,33E-4	0,76
1371592_at	12	220,00	177,70	146,88	1,65	146,88	1,65	6,79	4,83E-5	5,38E-4	0,77
1373792_at	12	1.155,20	673,30	671,77	1,48	671,77	1,48	6,78	4,84E-5	5,38E-4	0,77
1388810_at	12	325,10	217,80	-197,68	-1,25	-197,68	-1,25	-6,78	4,88E-5	5,43E-4	0,78
1367486_at	12	555,20	376,30	-328,52	-1,37	-328,52	-1,37	-6,77	4,91E-5	5,45E-4	0,78
1375312_at	0	155,50	95,00	97,78	6,22	97,78	6,22	6,77	4,91E-5	5,46E-4	0,78
1374135_at	9	352,30	185,40	-199,85	-1,41	-199,85	-1,41	-6,77	4,94E-5	5,48E-4	0,79
1372713_at	12	478,40	268,40	-291,45	-1,83	-291,45	-1,83	-6,76	4,96E-5	5,5E-4	0,79
1372930_at	12	553,20	418,30	355,72	2,28	355,72	2,28	6,76	4,97E-5	5,5E-4	0,79
1370285_at	12	591,80	296,10	323,78	1,69	323,78	1,69	6,76	4,98E-5	5,52E-4	0,79
1388538_at	12	276.10	164.70	-169.38	-1.47	-169.38	-1.47	-6.76	5.01E-5	5.54E-4	0.80

Figure 2: *FDR p-values compared to Bonferroni-corrected p-values.*

filter. Select **Diaphragm vs Heart transformed - FDR p-value correction** in the first drop-down box, select **<** in the next, and enter 0.0005 (or 0,0005 depending on your locale settings). Click **Apply** or press **Enter**.

This will filter the table so that only values below 0.0005 are shown (see figure 3).

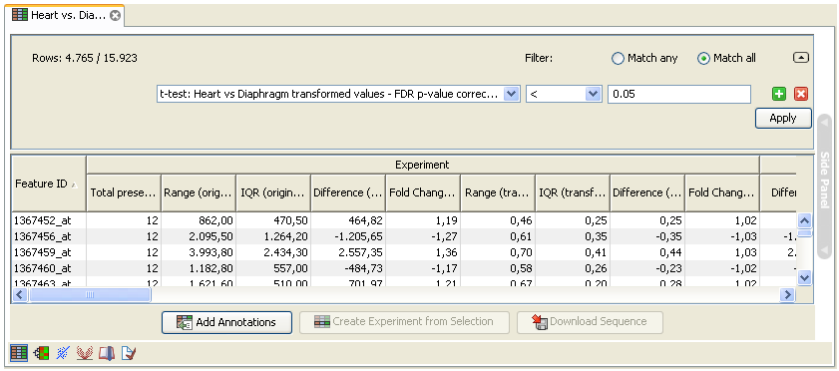

Figure 3: *Filtering on FDR p-values.*

You can see that 1471 genes fulfilled this criterion (marked with a red circle).

Inspecting the volcano plot

Another way of looking at this data is to click the **Volcano Plot** ( ) at the bottom of the view. Press and hold the Ctrl key while you click (⌘ on Mac).

This will make a split view of the experiment table and the volcano plot as shown in figure 4.

The volcano plot shows the difference between the means of the two groups on the X axis and the $-\log_{10}$ p-values on the Y axis.

If you now select the genes in the table (click in the table and press Ctrl + A /⌘ + A on Mac), you can see that the corresponding dots are selected in the volcano plot (see figure 5).

Filtering absent/present calls and fold change

Besides filtering on low p-values, we may also take the absent/present status of the features into consideration. The absent/present status is assigned by the Affymetrix software. There can be a number of reasons why a gene is called *absent*, and sometimes it is simply because the signal is very weak. When a gene is called absent, we may not wish to include it in the list of differentially expressed genes, so we want to filter these out as well.

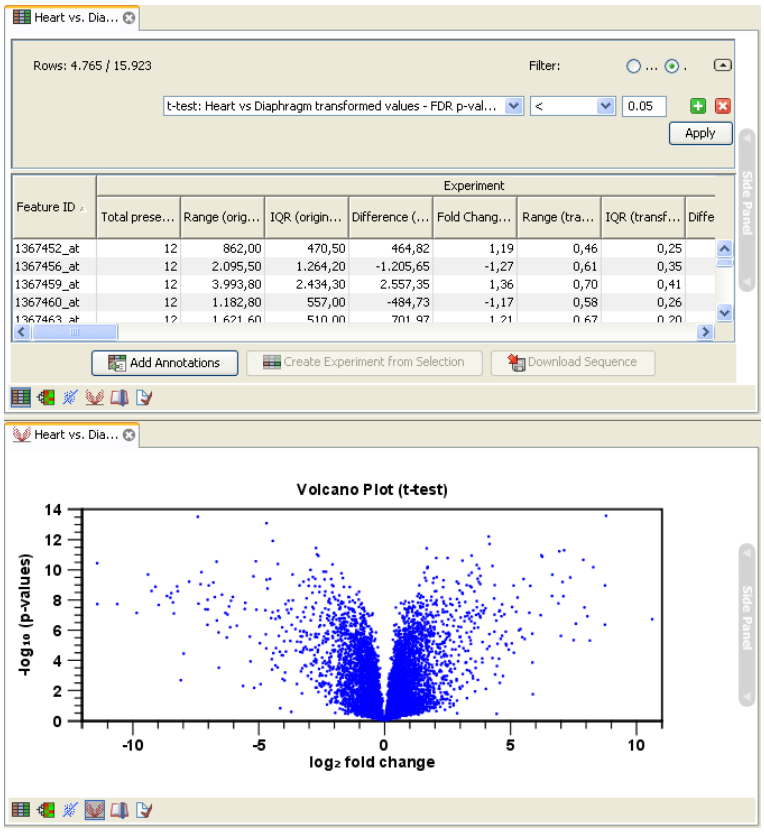

Figure 4: *Split view of volcano plot and experiment table.*

This can be done in several ways - in our approach we say that for any gene there must not be more than one absent call in each group. Thus, we add more criteria to the filter by clicking the **Add search criterion** (+) button twice and enter the limit for present calls as shown in figure 6.

Before applying this filter, 1471 genes were selected, and this list is now reduced to 1093.

Often the results of microarray experiments are verified using other methods such as QPCR, and then we may want to filter out genes that exhibit differences in expression that are so small that we will not be able to verify them with another method. This is done by adding one last criterion to the filter: Difference should have an absolute value higher than 2 (as we are working with log transformed data, the group mean difference is really the *fold change*, so this filter means that we require a fold change above 2).

This final filtering is shown in figure 7.

Note that the **abs value >** is important because the difference could be negative as well as positive.

The result is that we end up with a list of 142 genes that are likely candidates to exhibit differential expression in the two groups.

Click one of the rows and press Ctrl + A (⌘ +A on Mac) to select the 142 genes. You can now inspect the selection in the volcano plot below as shown in figure 8.

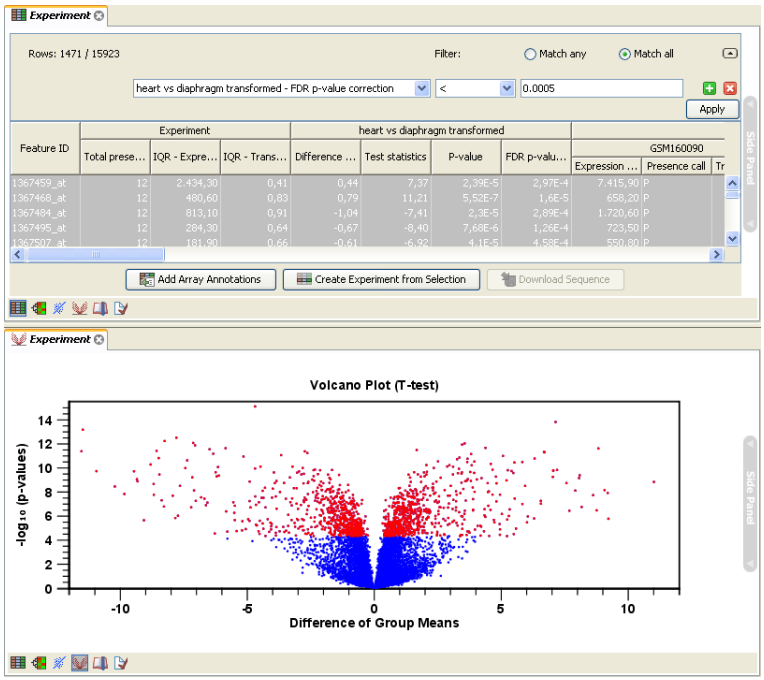

Figure 5: *Volcano plot where selected dots are colored red.*

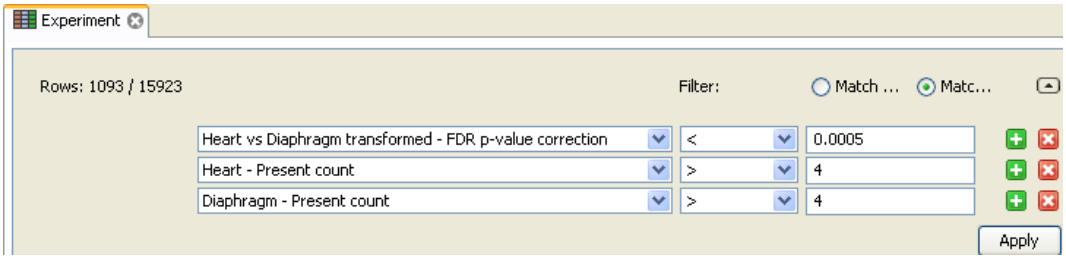

Figure 6: *Filtering genes where at least 5 out of 6 calls in each group are present.*

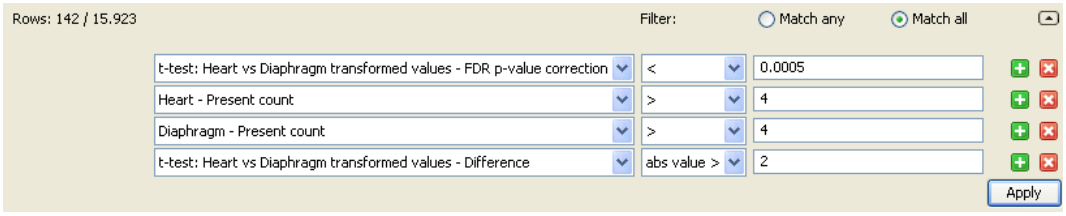

Figure 7: *The absolute value of group mean difference should be larger than 2.*

Saving the gene list

Before we proceed to the final part of the tutorials, we save the list of genes; click **Create Experiment from Selection (**  **)**

This will create a new experiment based on the selection. **Save** () the new experiment next to the old one.

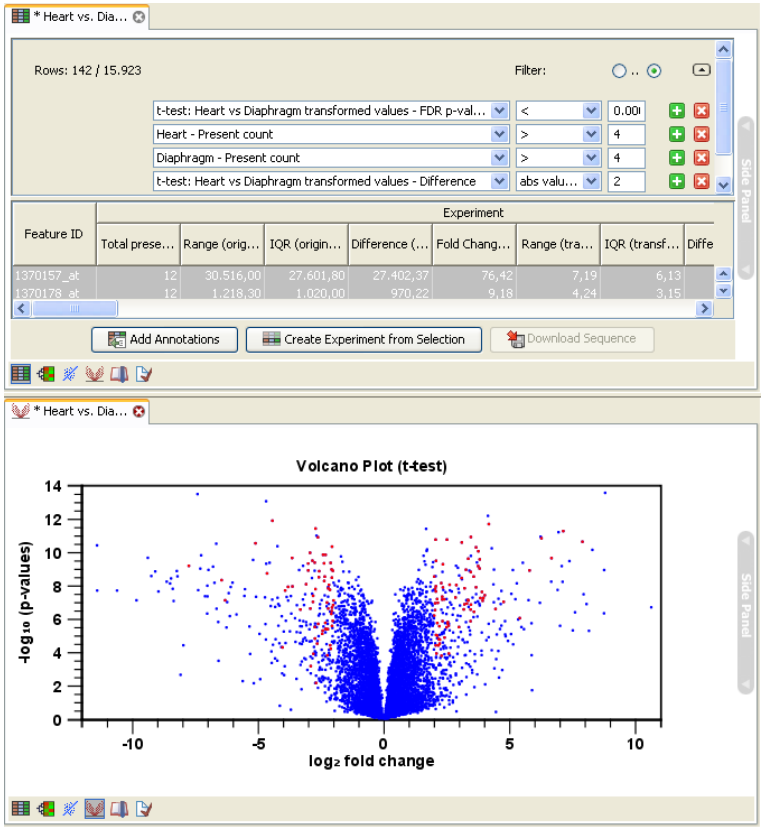

Figure 8: *142 genes out of 15923 selected.*