



# Tutorial

## Microarray-based Expression Analysis

June 27, 2019

— Sample to Insight —

## Microarray-based Expression Analysis

Expression analysis often requires advanced skills in statistics, but this tutorial is intended to show a straight-forward example of how to identify and interpret the differentially expressed genes in samples from two different tissues using CLC Main Workbench 7.8. If you are familiar with the statistical concepts and issues within expression analysis, you may find this tutorial too simplistic, but we have favored a simple and quick introduction over an exhaustive explanation. The analysis of this tutorial is based on microarray data, but could also be applied to RNA-Seq data.

The data comes from a study of gene expression in tissues from cardiac left ventricle and diaphragm muscle of rats [van Lunteren et al., 2008]. During this series of tutorials, you will see:

- How to import and set up the data in an experiment with two groups
- How to perform quality checks on the data
- How to perform statistics and clustering to identify and visualize differentially expressed genes
- How to use annotations to categorize and interpret patterns among the differentially expressed genes in a biological context

### Importing array data and setting up and experiment

First, import the data set which can be downloaded from the Gene Expression Omnibus (GEO) database at NCBI: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6943&targ=gsm&form=text&view=data>.

After download, click **Import** (📁) in the Tool bar and choose the "Standard import" option. Select the file called "GSE6943" and choose where you want to save it in the Navigation Area of your workbench. You will now have 12 arrays in your **Navigation Area** as shown in figure 1.

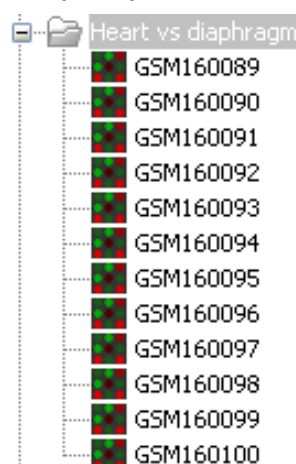


Figure 1: 12 microarrays have been imported.

The next step is to tell the workbench how the 12 samples are related. This is done by setting up an **Experiment** (📊), i.e., a set of samples and information about how the samples are related (which groups they belong to). The **Experiment** is also used to accumulate calculations like t-tests and clustering.

1. To set up the experiment, use the Launch button (🚀) to start the **Set Up Experiment** (🛠️) tool.
2. Select the 12 arrays that you have imported (see figure 2).

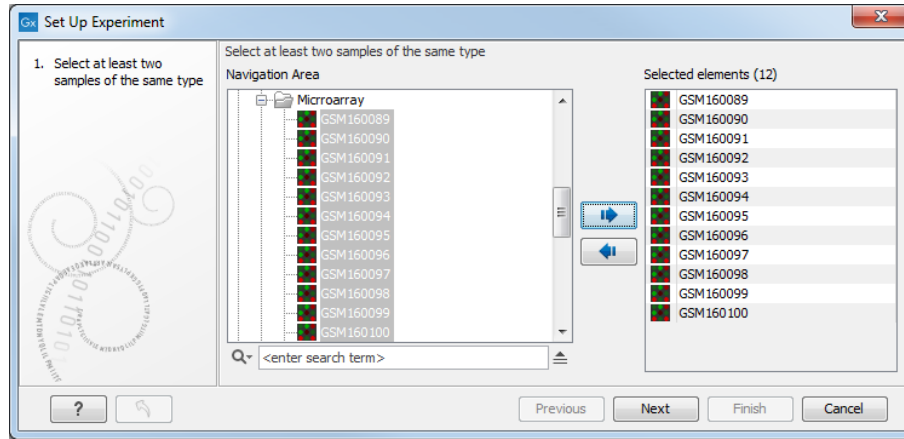


Figure 2: Select the 12 microarrays that have been imported.

3. In the next wizard window, you can define the number of groups in the experiment. Since we compare heart tissue with diaphragm tissue, we use a two-group comparison. Leave the option as **Unpaired** as in figure 3. Click on the button labeled **Next**.

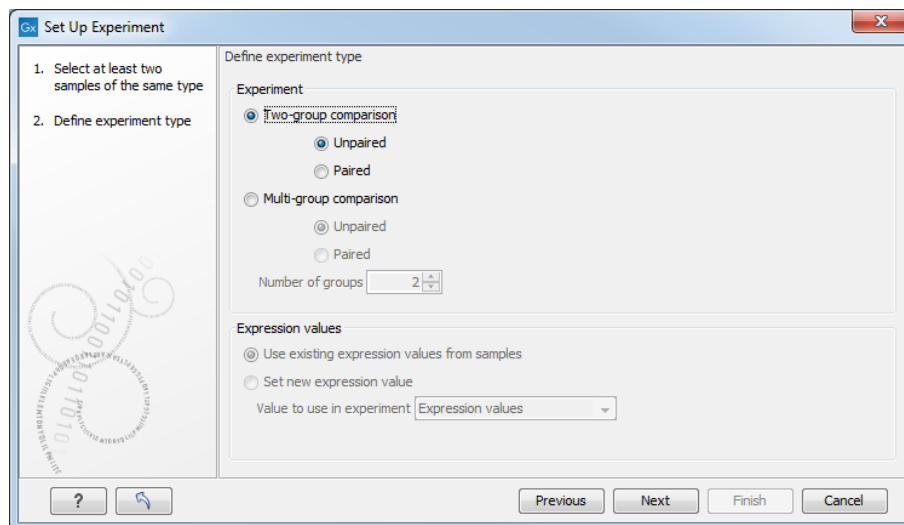


Figure 3: Defining the number of groups.

4. Name the first group **Heart** and the second group **Diaphragm** (figure 4) and click **Next**.
5. Here you see a list of all the samples preselected (figure 5). Select the first 6 samples (by clicking in the group column of the first sample and while holding down the mouse button you drag and select the other five samples), right-click and select **Heart**. Select the last 6 samples, right-click and select **Diaphragm**. In this way you define which group each sample belongs to. Click **Next**.
6. Chose to **Save** the experiment, indicate where in your Navigation Area you would like to save it and click **Finish**.

## Tutorial

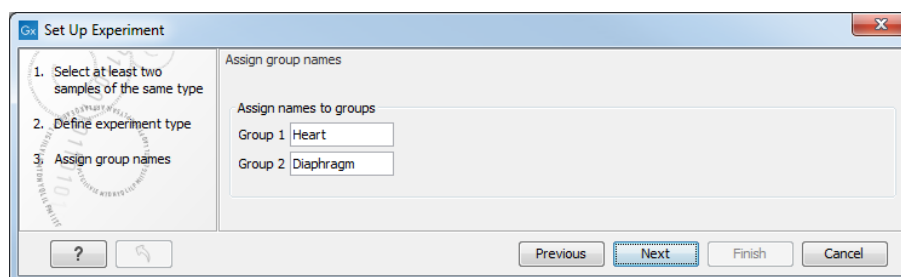


Figure 4: Naming the groups.

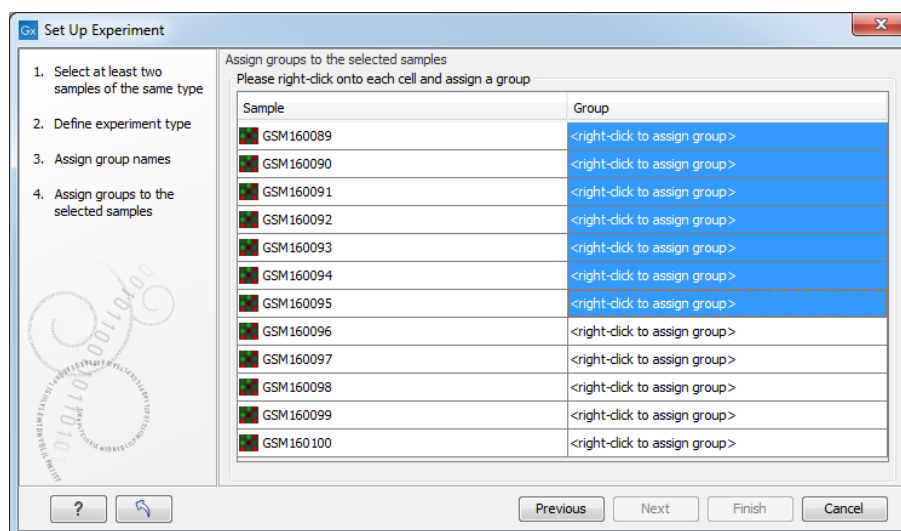
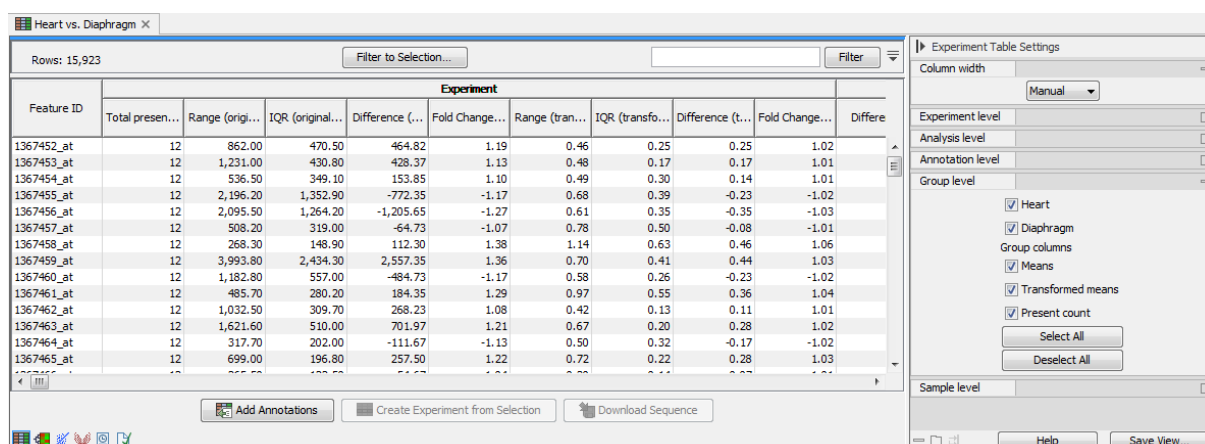


Figure 5: Assigning the samples to groups.

The experiment is created. Note that the information from samples located in the **Navigation Area** is copied into the experiment, so they now exist independently of each other.

Once it is created, the experiment can be opened in a table as shown in figure 6.



Feature ID	Total present	Range (original)	IQR (original)	Difference (original)	Fold Change (original)	Range (transformed)	IQR (transformed)	Difference (transformed)	Fold Change (transformed)	Difference (transformed)
1367452_at	12	862.00	470.50	464.82	1.19	0.46	0.25	0.25	1.02	
1367453_at	12	1,231.00	430.80	428.37	1.13	0.48	0.17	0.17	1.01	
1367454_at	12	536.50	349.10	153.85	1.10	0.49	0.30	0.14	1.01	
1367455_at	12	2,196.20	1,352.90	-772.35	-1.17	0.68	0.39	-0.23	-1.02	
1367456_at	12	2,095.50	1,264.20	-1,205.65	-1.27	0.61	0.35	-0.35	-1.03	
1367457_at	12	508.20	319.00	-64.73	-1.07	0.78	0.50	-0.08	-1.01	
1367458_at	12	268.30	148.90	112.30	1.38	1.14	0.63	0.46	1.06	
1367459_at	12	3,993.80	2,434.30	2,557.35	1.36	0.70	0.41	0.44	1.03	
1367460_at	12	1,182.80	557.00	-484.73	-1.17	0.58	0.26	-0.23	-1.02	
1367461_at	12	485.70	280.20	184.35	1.29	0.97	0.55	0.36	1.04	
1367462_at	12	1,032.50	309.70	268.23	1.08	0.42	0.13	0.11	1.01	
1367463_at	12	1,621.60	510.00	701.97	1.21	0.67	0.20	0.28	1.02	
1367464_at	12	317.70	202.00	-111.67	-1.13	0.50	0.32	-0.17	-1.02	
1367465_at	12	699.00	196.80	257.50	1.22	0.72	0.22	0.28	1.03	

Figure 6: The experiment table.

The table includes the expression values for each sample and in addition a few extra values have been calculated such as the range, the IQR (Interquartile Range), fold change and difference values and the present counts for the whole experiment and the individual groups (note that absent/present calls are not available on all kind of data).

## Quality checks

First we inspect to what extent the variance in expression values depends on the mean with an **MA Plot**.

1. Use the Launch button (🚀) to start the **Create MA Plot** (📊) tool.
2. Since the MA plot compares two samples, we will start out selecting one of the six first arrays (the one belonging to the group Heart). Click **Next**.
3. Select as control one of the six last arrays (the one belonging to the group Diaphragm) and click **Next**.
4. Leave the parameters as default (set to "Original expression values") and click **Next**.
5. Choose to open the result and click **Finish**.

This will show a plot similar to the one shown in figure 7.

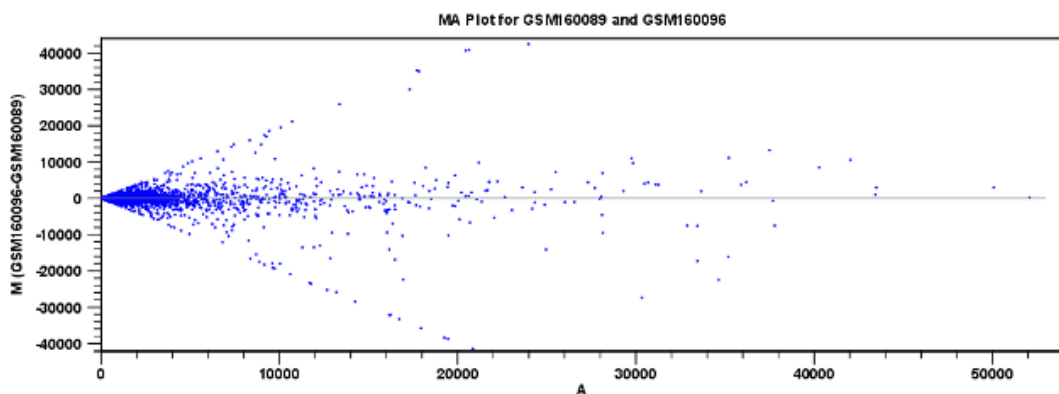


Figure 7: MA plot before transformation.

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 7 it is clear that the variance increases with the mean.

To remove some of the dependency, we want to **transform** the data.

1. Use the Launch button (🚀) to start the **Transform** (📊) tool.
2. Select in the first dialog the same two arrays as used for the plot and click **Next**.
3. Leave the "Values to analyze" as "Original expression values" but change the "Logarithm transformation" to **Log 2**.
4. Choose to Save the data and click **Finish**.
5. Now create an MA plot again as described above and using the samples you just transformed. Click **Next**.

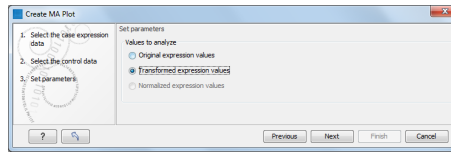


Figure 8: Select the transformed expression values.

6. In the set parameters window, you can see that you now also have the option to choose **Transformed expression values** as parameters (see figure 8).

Original, Transformed and Normalized expression values are used several places when expression values are used in a calculation. Select the "transformed expression values" option and click **Next**.

7. Choose to Save or Open the plot and click **Finish**.

This will result in a quite different plot as shown in figure 9.

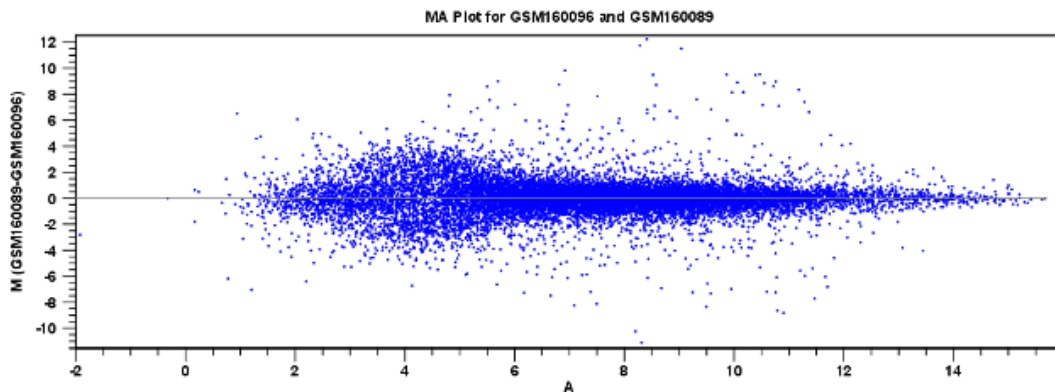


Figure 9: MA plot after transformation.

The much more symmetric and even spread indicates that the dependence 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. Hence, we need to transform the expression values within the entire experiment, as we will use the transformed data in the further analysis.

1. To transform the entire experiment, use again the Launch button (🔧) to start the **Transform** (📊) tool.
2. Select the experiment Heart vs. Diaphragm and click **Next**.
3. Choose **Log 2** transformation.
4. Save the experiment before clicking **Finish**.

If you open the table, you will see that all the samples have an extra column with transformed expression values ( figure 10).

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

Rows: 15,923											
Feature ID	Experiment			GSM160089			GSM160090			GSM160091	
	Total presen...	IQR (original...	IQR (transfo...	Expression values	Presence call	Transformed...	Expression values	Presence call	Transformed...	Expression values	Prese
1367452_at	12	470.50	0.25	2,532.90 P		11.31	2,518.60 P		11.30	2,384.60 P	
1367453_at	12	430.80	0.17	3,464.20 P		11.76	3,197.40 P		11.64	3,487.10 P	
1367454_at	12	349.10	0.30	1,620.80 P		10.66	1,870.50 P		10.87	1,538.60 P	
1367455_at	12	1,352.90	0.39	5,512.50 P		12.43	4,103.90 P		12.00	5,746.50 P	
1367456_at	12	1,264.20	0.35	6,090.80 P		12.57	5,352.20 P		12.39	5,614.90 P	
1367457_at	12	319.00	0.50	1,093.90 P		10.10	1,134.30 P		10.15	736.40 P	
1367458_at	12	148.90	0.63	347.80 P		8.44	223.90 P		7.81	261.40 P	
1367459_at	12	2,434.30	0.41	7,665.80 P		12.90	7,415.90 P		12.86	7,075.90 P	
1367460_at	12	557.00	0.26	3,155.70 P		11.62	2,946.90 P		11.52	3,589.70 P	

Figure 10: Transformed expression values have been added to the table.

**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 (such as 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** (📊).

1. Launch **Create Box Plot** (📊).
2. Select the experiment and click **Next**.
3. Choose the option **Transformed expression values** and click on the button labeled **Next**.
4. Chose to open your results before clicking on **Finish**.

The box plot is shown in figure 11.

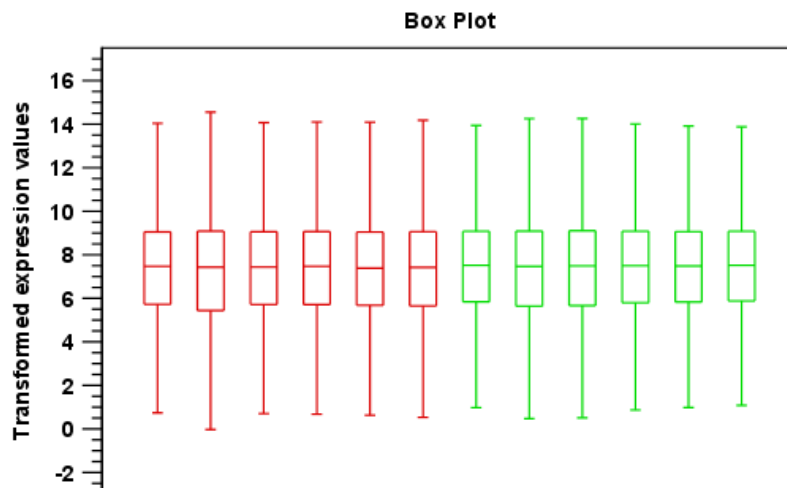


Figure 11: A box plot of the 12 samples in the experiment, colored by group.

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 12 from another data set, you can see the difference.

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 removing the bad quality sample.

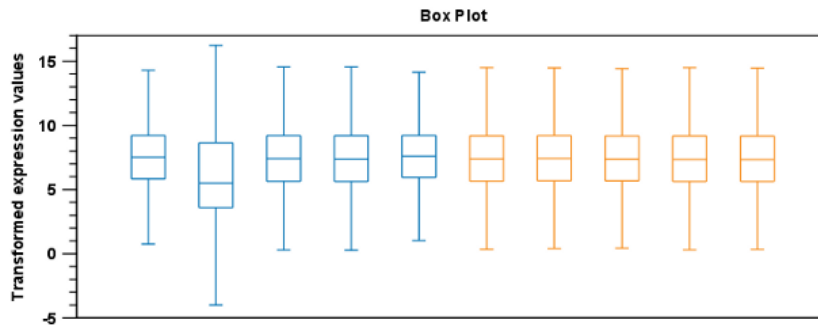



Figure 12: A box plot showing one sample that stands out from the rest.

## 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 of the other group.

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

1. : Launch **Principal Component Analysis** ().
2. Select the experiment and click **Next**.
3. Keep the parameters as set by default to "Original expression values" and click **Next**.
4. **Save** the data before clicking on **Finish**.

This will create a PCA plot as shown in figure 13.

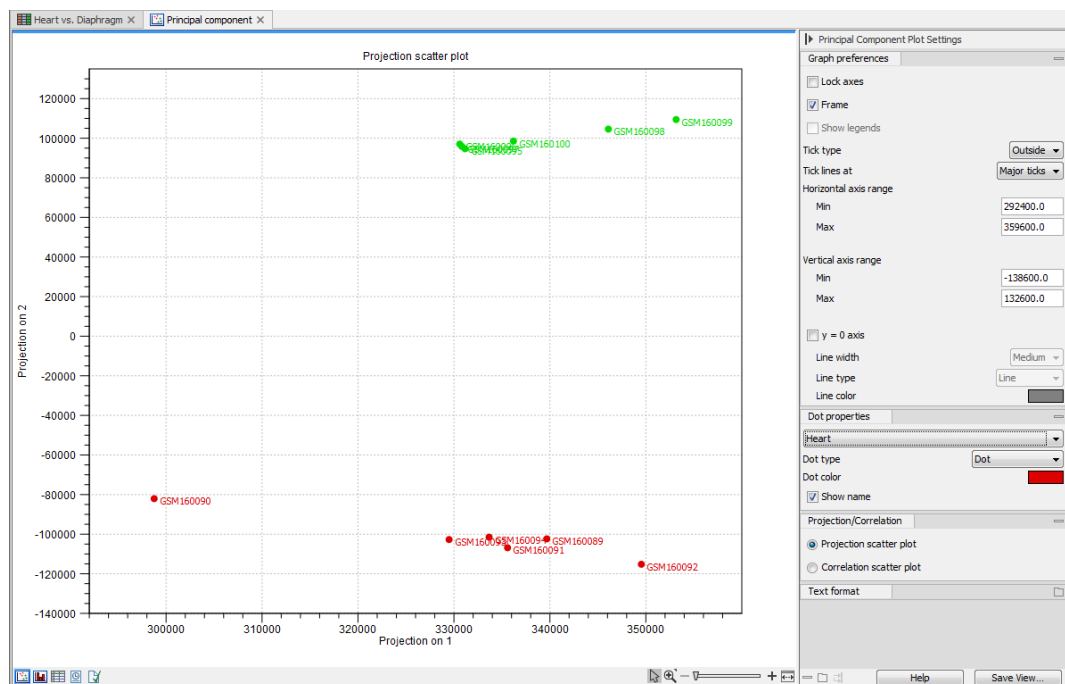



Figure 13: A *principal component analysis* colored by group.



The plot shows the projection of the samples onto the two-dimensional space spanned by the first and second principal components (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 the plot shows that they cluster together. 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. In the "Dot properties" section, select *GSM160090* in the drop-down box and choose to "Show names". In this way you can control the coloring and dot types of the different samples and groups.

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:

1. Launch **Hierarchical Clustering of Samples** (  ).
2. Select the experiment and click **Next**.
3. Leave the parameters at their default as seen in figure 14). You can use the **Reset** button if you are not sure whether you have previously changed the parameters for the tool. Click **Next**.

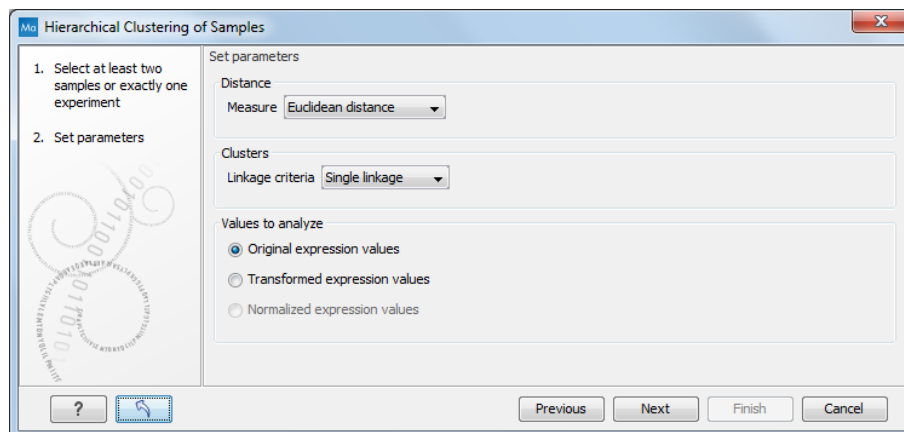


Figure 14: Set parameters for the Hierarchical Clustering of Samples tool.

4. Choose to Save the output and click **Finish**.

This will display a heat map showing the clustering of samples at the bottom (see figure 15).

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. You can use the buttons at the bottom of the editor to switch between different views in figure 16.

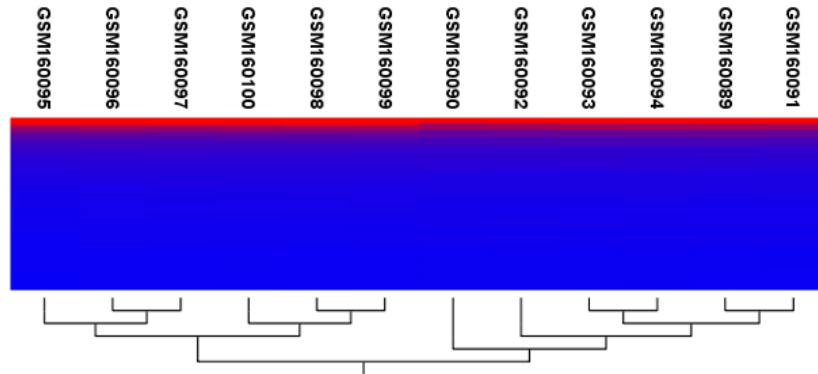


Figure 15: Sample clustering.




Figure 16: Different views on an experiment.

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.

## Statistical analyses

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

1. Launch the **Gaussian Statistical Analysis**  tool.
2. Select the experiment Heart vs. Diaphragm and click **Next**.
3. Leave the parameters as default as shown in figure 17 and click **Next** again.

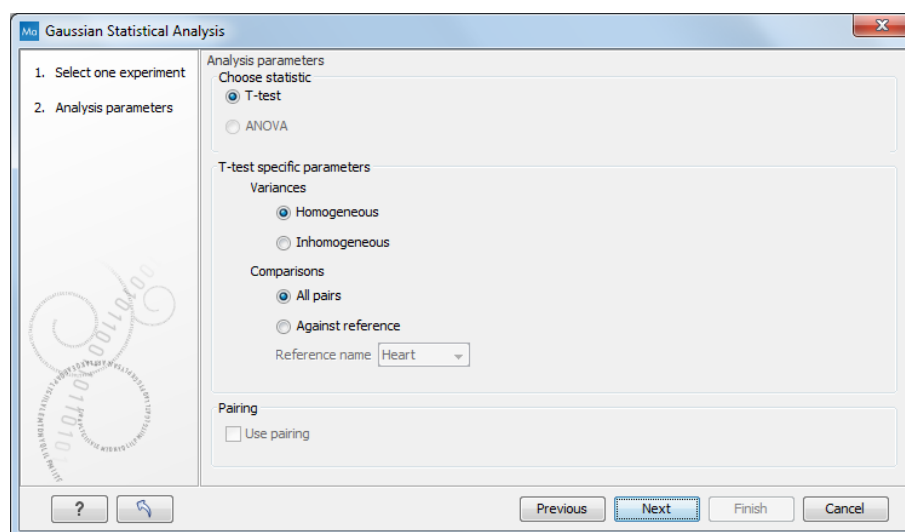


Figure 17: Set parameters for the Gaussian Statistical Analysis tool.

4. Select the "transformed expression values" option and check the two corrected p-values as well (figure 18). You can read more about what they mean by clicking the **Help** button in the dialog.

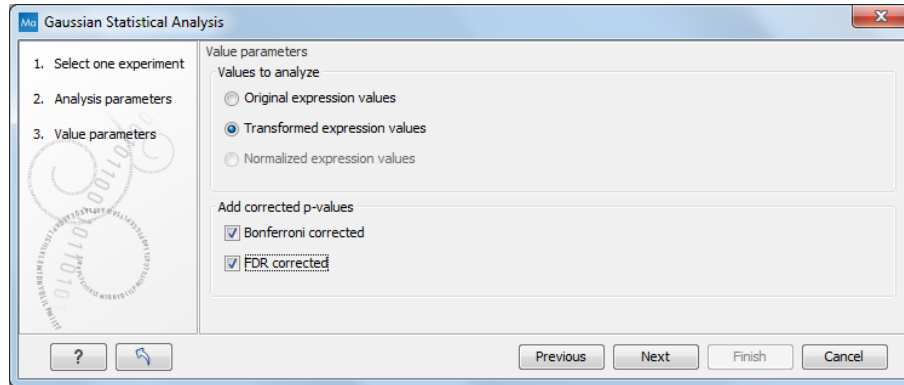


Figure 18: Statistical analysis.

5. Save the results (it will be saved in the existing table) and click on **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.

In the "t-test" section of the table, 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 which of the FDR p-value and the Bonferroni-corrected p-value is much stricter (Bonferroni p-values approaching 1).

**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 filter. Select **t-test: Heart vs Diaphragm transformed values - 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 **Filter**.

This will filter the table so that only values below 0.0005 are shown (see figure 19).

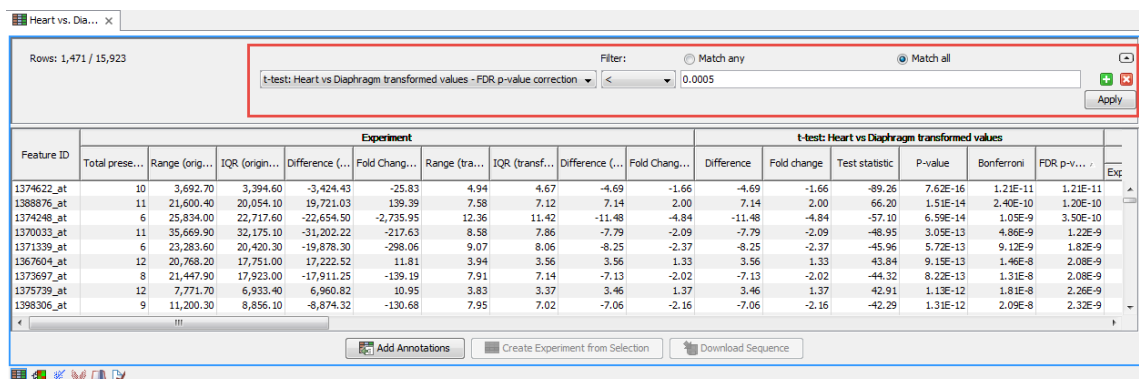


Figure 19: Filtering on FDR p-values.

**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) to open the plot

in split view.

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 in the table all the genes that were left after applying the filter (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 20).

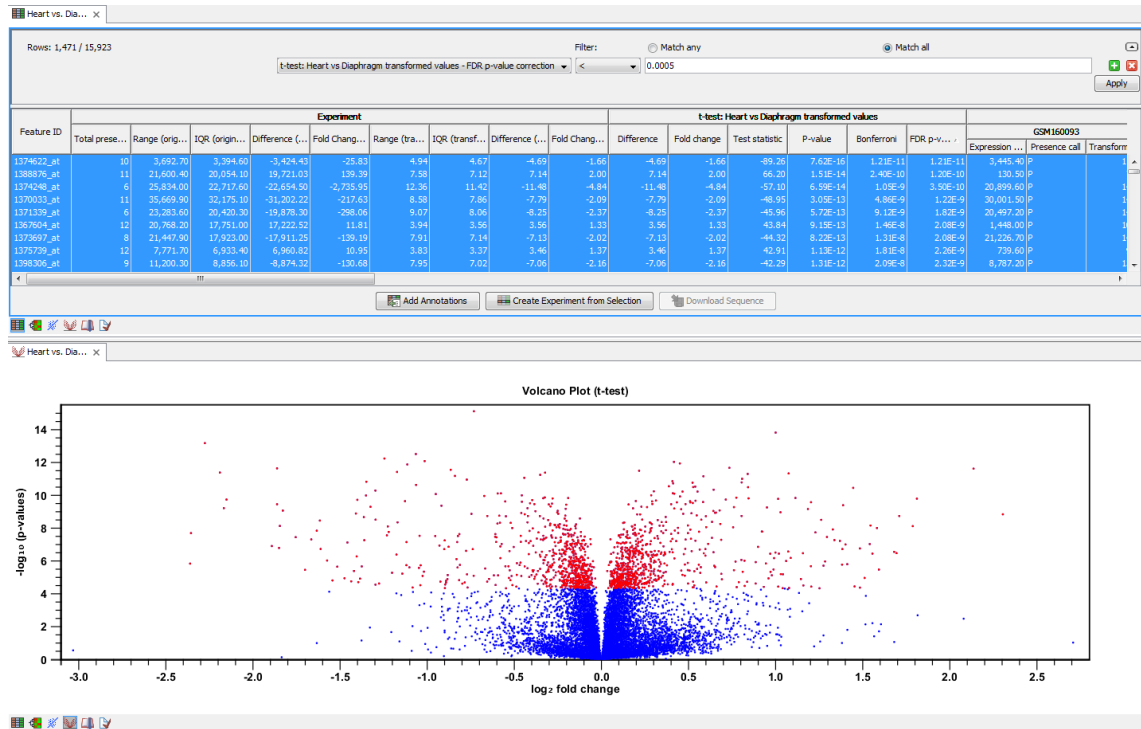


Figure 20: Volcano plot where selected dots are colored red.

## 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.

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 21. Click **Filter**.

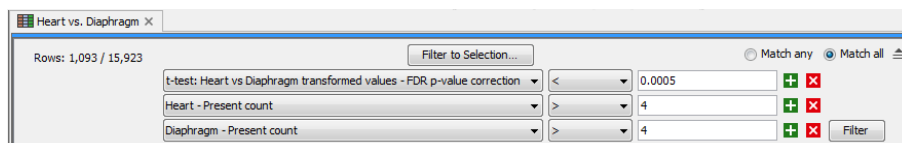


Figure 21: Filtering genes where at least 5 out of 6 calls in each group are present.

## Tutorial

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). Click **Filter**.

This final filtering is shown in figure 22.

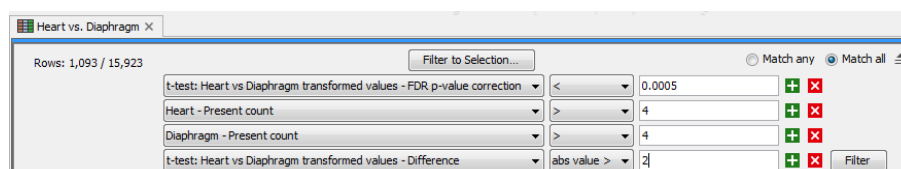


Figure 22: The absolute value of group mean difference should be larger than 2.

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 genes that are likely candidates to exhibit differential expression in the two groups.

Select all remaining rows and inspect the selection in the volcano plot below as shown in figure 23.

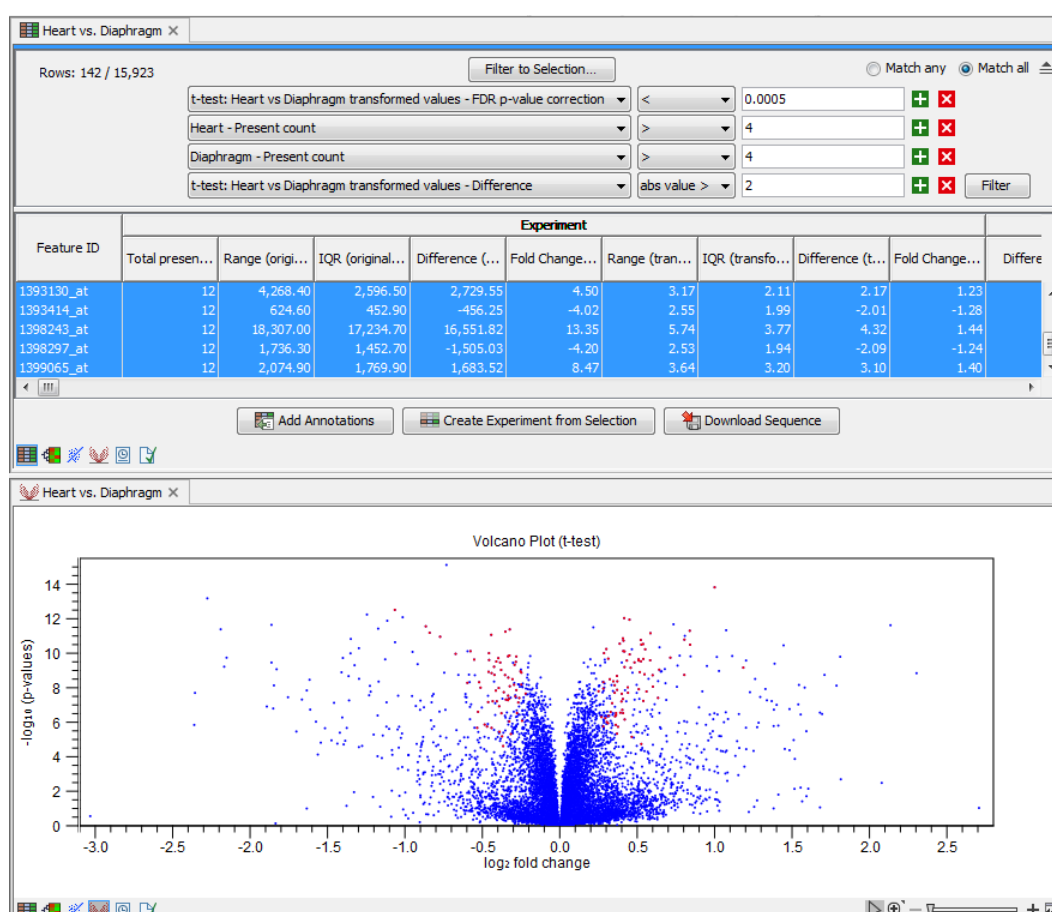


Figure 23: 142 genes out of 15923 selected.

**Saving the gene list** Before we proceed to the final part of the tutorial, we save the list of genes; click on **Create Experiment from Selection** (📄). This will create a new experiment based on the selection. **Save** (💾) the new experiment next to the old one in your Navigation Area.

## Importing Annotations

The first step is to import an annotation file used to annotate the arrays.

1. In this case, the data were produced using an Affymetrix chip, and the annotation file can be downloaded from the web site <https://www.thermofisher.com>. You can access the file by searching for **RAE230A**. Choose the item called "Current NetAffx Annotation Files: RAE230A Annotations, CSV format, Release 36". Note that this is a free service but you still have to sign up in order to download the file.
2. To import the annotation file in the workbench, click **Import** (📄) | **Standard Import** in the Tool bar and select the "RAE230A.na36.annot.csv" zip file.
3. Next, Launch **Add Annotations** (📄).
4. Select the experiment created in the previous tutorial and the annotation file (📄) (figure 24) before clicking **Next**.

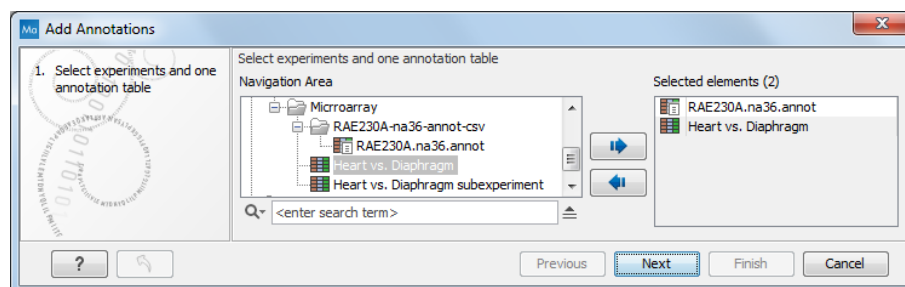


Figure 24: Annotating the experiment.

5. Leave the parameters as default (figure 25) and click **Next**.

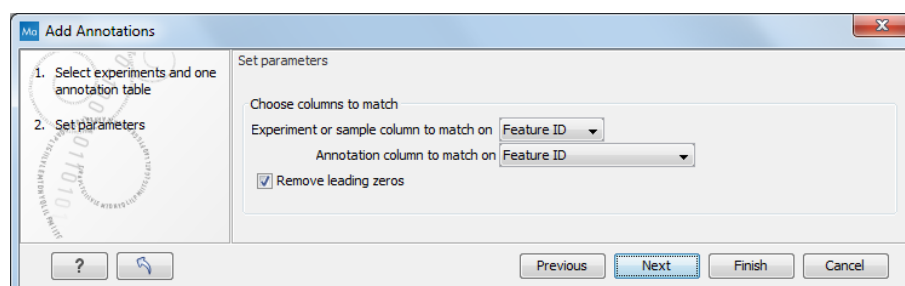


Figure 25: Parameters of the Add Annotations tool.

6. **Save** the results and click on **Finish**.

Re-open the experiment. When you look in the **Side Panel** of the experiment, there are a lot of options to show and hide columns in the table. This can be done on several levels. At the **Annotation level** you find a list of all the annotations. Some are shown per default, others you will have to click to show. At the bottom of the list, a button "Deselect all" allows you to deselect

all annotations at once. Then reselect **Gene title** which describes the gene and is much more informative than the Feature ID. Further down the list you find the annotation type **GO biological process**. We will use this annotation in the next two analyses.

## Hypergeometric Tests on Annotations

The first annotation test will show whether any of the GO biological processes are over-represented in our sub experiment, i.e., the list of differentially expressed genes in the full set of genes measured.

1. Launch the **Hypergeometric Tests on Annotations** (🎯) tool.
2. Select the two experiments (the original full experiment and the small subset of 142 genes) and click **Next** (figure 26).

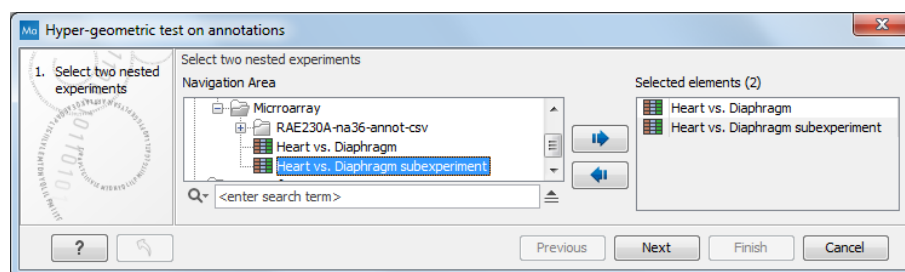


Figure 26: Selecting an experiment and a nested sub-experiment.

3. Select **GO biological process** and **Transformed expression values** (see figure 27). Click **Next**.

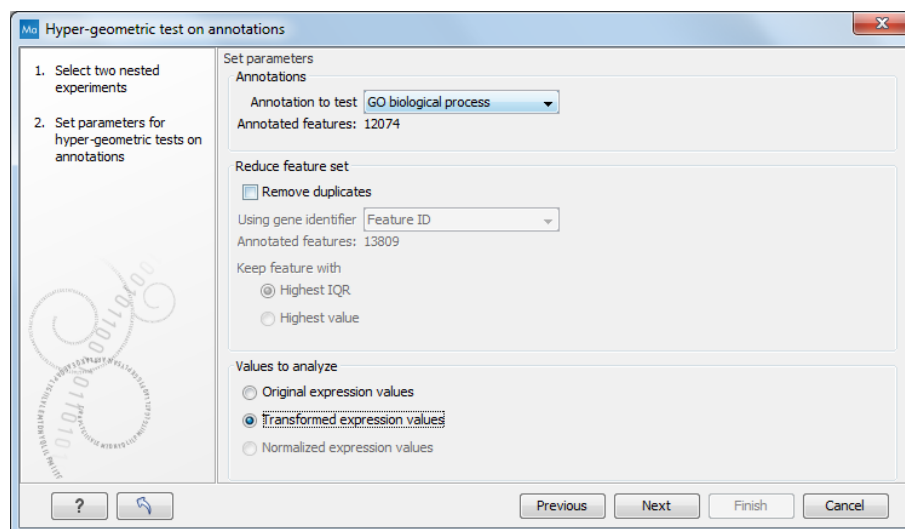


Figure 27: Testing on GO biological process.

4. Click **Save** and **Finish** to perform the test.

The result is shown in figure 28.

This table lists the GO categories. Click on the header "In subset" twice to sort values according to how many genes in each categories were in the subset. If you take the first row, there are



Rows: 13,549 Hyper-Geometric tests for annotation category associations					
Category	Description	Full set	In subset	Expected in s...	Observed - e... p-value
0008150	biological process	8668	85	85	0
0044699	single-organism process	6456	73	57	16 5.10E-5
0009987	cellular process	7544	73	66	7 0.05
0044763	single-organism cellular process	5285	64	46	18 6.47E-5
0065007	biological regulation	6660	63	59	4 0.18
0050789	regulation of biological process	6256	57	55	2 0.37
0050794	regulation of cellular process	5916	51	52	-1 0.63
0008152	metabolic process	5522	38	49	-11 0.99
0071704	organic substance metabolic process	4871	34	43	-9 0.98
0044237	cellular metabolic process	4608	34	41	-7 0.94
0032502	developmental process	3323	33	29	4 0.22
0071840	cellular component organization or biogenesis	3033	32	27	5 0.13
0016043	cellular component organization	2984	32	26	6 0.11

Figure 28: The result of testing on GO biological process.


almost 10,000 genes in this category in the full set, and 85 were expected in the subset. The analysis found 85 in the subset, so this biological process does not seem to be over-represented. On the other hand, on the second row, 57 genes related to "single-organism process" were expected in the subset (based on having 6456 in the full set). 73 were found, so this process is over-represented and given a p-value of 5.10E-5. To find easily over and under represented processes, click on the header "Observed - expected" and check processes at both end of the list.

## Gene Set Enrichment Analysis (GSEA)

The hypergeometric tests on annotations uses a pre-defined subset of differentially expressed genes as a starting point and compares the annotations in this list to those of the genes in the full experiment. The exact limit for this subset is somewhat arbitrary - in our case we could have chosen a p-value less than 0.005 instead of 0.0005 and it would lead to a different result.

Furthermore, only the most apparently differentially expressed genes are used in the subset - one could easily imagine that other categories would be significant based on more genes with lower fold change or higher p-values.

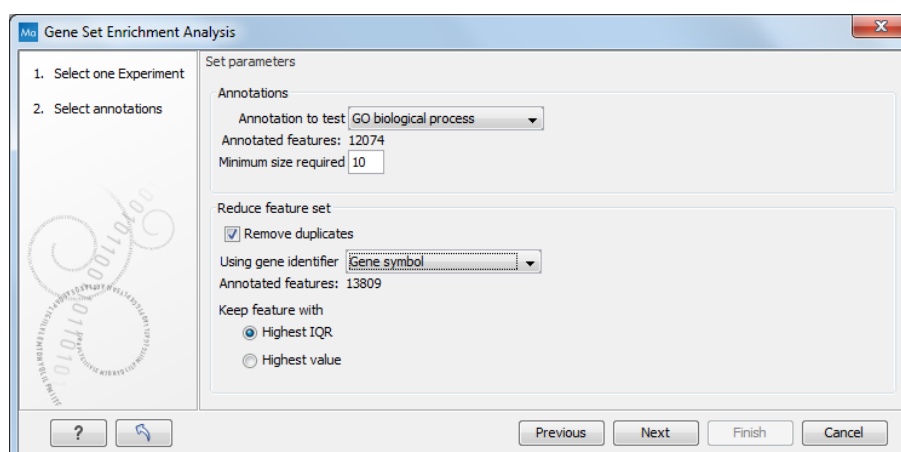
The Gene Set Enrichment Analysis (GSEA) does not take an *a priori* defined list of differentially expressed genes and compares it to the full list - it uses a single experiment. It ranks the genes on p-value and analyzes whether there are some categories that are over-represented in the top of the list.

1. Launch **Gene Set Enrichment Analysis (GSEA)** .
2. Select the original full experiment and click **Next**.
3. In the next step, make sure the **GO biological process** is chosen, and check the option to **remove duplicates** using "gene symbol" as identifier (see figure 29). Click **Next**.
4. Select the **Transformed expression values**. The number of permutations for p-value calculation is left at 10 000. Click **Next**.
5. Chose to **Open** the result and click on **Finish**.

The result is shown in figure 30.

The table is sorted on the lower tail so that the GO categories where up-regulated genes in the first group are over-represented are placed at the top, and the GO categories where up-regulated genes in the second group are over-represented are placed at the bottom.





Gene Set Enrichment Analysis

1. Select one Experiment  
2. Select annotations

Set parameters

Annotations

Annotation to test: GO biological process

Annotated features: 12074

Minimum size required: 10

Reduce feature set

☒ Remove duplicates

Using gene identifier: Gene symbol

Annotated features: 13809

Keep feature with

☒ Highest IQR

☐ Highest value

Previous Next Finish Cancel

Figure 29: Gene set enrichment analysis based on GO biological process.

Rows: 4,938 Gene set enrichment analysis (GSEA)					
Category	Description	Size	Test statistic	Lower tail	Upper tail
0044260	cellular macromolecule metabolic process	3325	-8.73	0.00	1.00
0044264	cellular polysaccharide metabolic process	50	-22.41	0.00	1.00
0048634	regulation of muscle organ development	103	-18.30	0.00	1.00
0048641	regulation of skeletal muscle tissue development	44	-34.32	0.00	1.00
0045661	regulation of myoblast differentiation	39	-25.04	0.00	1.00
0005976	polysaccharide metabolic process	55	-20.22	0.00	1.00
0005977	glycogen metabolic process	44	-25.10	0.00	1.00
0014888	striated muscle adaptation	22	-29.34	0.00	1.00
0048742	regulation of skeletal muscle fiber development	10	-31.05	0.00	1.00
0043170	macromolecule metabolic process	3741	-8.78	0.00	1.00
0014733	regulation of skeletal muscle adaptation	14	-31.30	0.00	1.00
0007519	skeletal muscle tissue development	52	-29.86	0.00	1.00
0006112	energy reserve metabolic process	53	-21.26	0.00	1.00
0003009	skeletal muscle contraction	33	-28.57	0.00	1.00

Figure 30: The result of a gene set enrichment analysis based on GO biological process.

Note that we could have chosen to filter away genes with less reliable measurements from the experiment (as shown previously) before subjecting it to the GSEA analysis in order to limit noise and aim for a more robust result.

## Bibliography

[van Lunteren et al., 2008] van Lunteren, E., Spiegler, S., and Moyer, M. (2008). Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism. *Respir Physiol Neurobiol*, 161(1):41–53.