Tutorial

Microarray Expression Analysis

April 8, 2025

Sample to Insight

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Tutorial

Microarray Expression Analysis

Introduction

The purpose of this tutorial is to show how to identify differentially expressed genes in microarray data from two different tissues using *CLC Main Workbench* or *CLC Genomics Workbench*. We focus on the following:

- Import microarray data.
- Set up an experiment with two groups.
- Perform quality control.
- Perform statistical analyses to identify and visualize differentially expressed genes.
- Optionally perform annotation tests to categorize and interpret patterns among the differentially expressed genes in a biological context.

Data used in this tutorial

This tutorial uses the data from "Contrast between cardiac left ventricle and diaphragm muscle in expression of genes involved in carbohydrate and lipid metabolism" by van Lunteren et al 2008 (GSE6943, science/article/abs/pii/S1569904807003229).

The authors investigated transcriptional differences between **6 heart** and **6 diaphragm** rat tissue samples using Affymetrix microarray analyses.

Prerequisites

For this tutorial, you must be working with *CLC Main Workbench* or *CLC Genomics Workbench* 25.0 or higher. Note that higher versions may produce slightly different results than those shown here.

General tips

- Throughout this tutorial, we provide links to relevant manual pages, which we recommend exploring for additional details.
- Tools can be found in the Toolbox, but it is often easier to launch them using Quick Launch (), found in the top toolbar (shortcut Ctrl+Shift+T or ⌘ +Shift+T on Mac). Quick Launch displays the full Toolbox path, making it easy to identify the location of the tool if needed.
- The in-built manual can be accessed by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.
- Within wizards, the **Reset** button can be used to change settings to their default values.
- Colors and gradients in plots can be changed by clicking on them in the Side Panel.
- Columns in tables can be hidden by unchecking their name in the Side Panel.
- Columns in tables can be used to sort the rows, by successively clicking on the column name until the desired order (indicated by an arrow next to the column name) is achieved.
- Most of the tools of *CLC Workbench* require multiple inputs. When many data elements need to be selected, all elements located under a folder can be added by using the options **Add folder contents** or **Add folder contents (recursively)** found in the right-click menu.
- Many data elements produced by *CLC Workbench* tools have multiple views, indicated as icons in the lower left corner of elements opened in the View Area. Clicking on one of the view icons while pressing the Ctrl (⌘ on Mac) key will open in split view such that both views are visible at the same time. Often, if viewing a table and a graphical representation in split view, selecting entries in the table will highlight them in the graphical representation. The order of the views can be changed using drag and drop, see Arrange views in View Area.

Import the data

We start by downloading and importing the tutorial data.

1. Download the tutorial data.
2. Start the *CLC Main Workbench* or *CLC Genomics Workbench*.
3. Import the data using Standard Import:
   (a) Launch **Standard Import** ( ) using Quick Launch ( ).
   (b) Locate the tutorial data called "GSE6943.txt" using the **Add files** button and select **Automatic import** (figure 1).

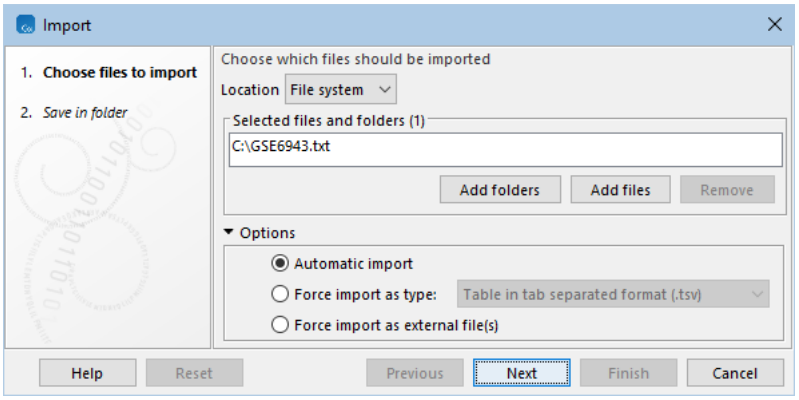

Figure 1: *Standard Import configured to import the tutorial data.*

   (c) In the next step, select a suitable location in the Navigation Area, create a new folder called "Microarray Tutorial Data" and choose to save the imported data there. Click on **Finish**.

Once the import is completed, 12 microarray samples are visible in the **Navigation Area** (figure 2).

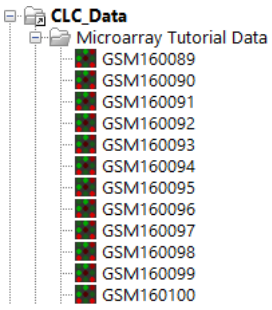

Figure 2: *The 12 imported microarray samples.*

Set up an experiment

The next step is to set up a microarray experiment, to define how the 12 samples are related. An **Experiment** is a set of samples and information about how those samples are related (which groups they belong to). An **Experiment** can also be used to accumulate results from, e.g., t-tests and clustering.

1. Launch **Set Up Microarray Experiment** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select at least two samples of the same type", select the 12 microarray samples (figure 3).

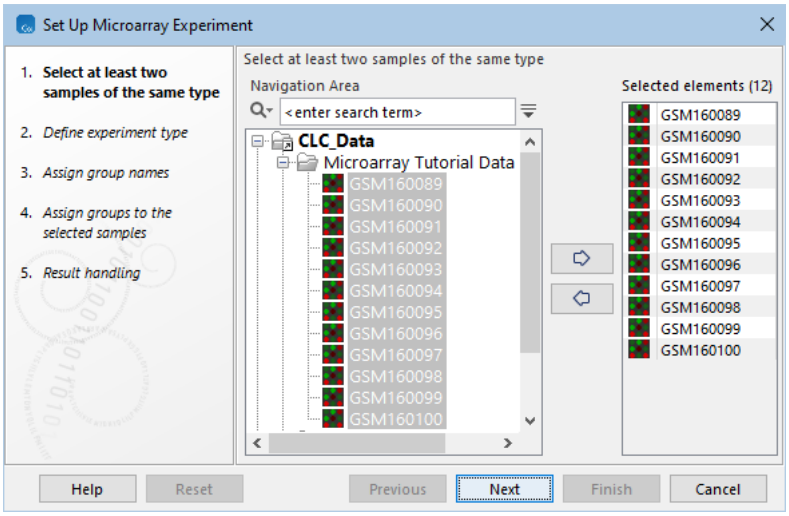

Figure 3: *All 12 samples are used as input.*

3. In the next step, "Define experiment type", define the number of groups in the experiment. Since we want to compare heart and diaphragm tissue, select **Two-group comparison** and **Unpaired** (figure 4).

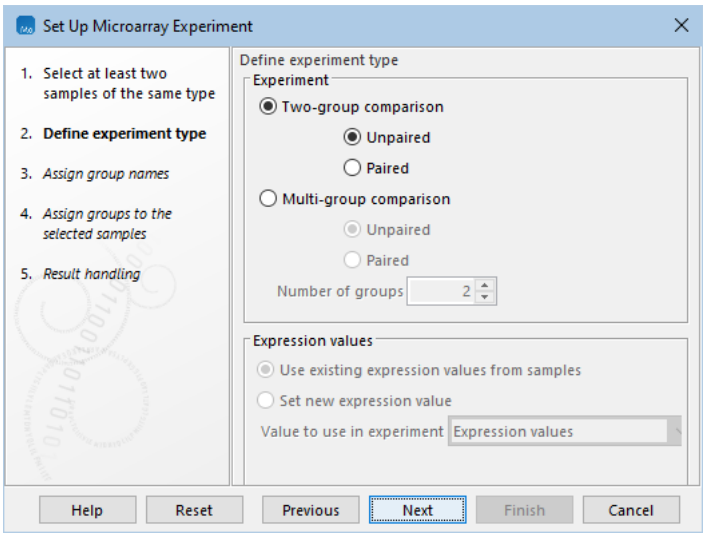

Figure 4: *Two-group comparison is selected.*

4. In the next step, "Assign group names", name Group 1 "Heart" and Group 2 "Diaphragm" (figure 5).

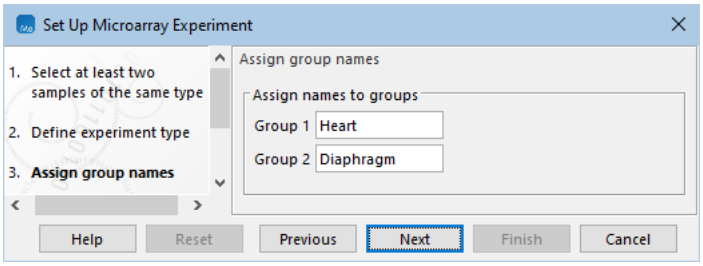

Figure 5: *The groups are named after the two tissue types.*

5. In the next step, "Assign groups to the selected samples", define which group each sample belongs to. Select the first six samples by left-clicking in the "Group" column of the first sample, holding down the mouse button, and dragging to select the other five samples. Right-click and select **Heart**. Select the last six samples in the same way, right-click and select **Diaphragm** (figure 6).

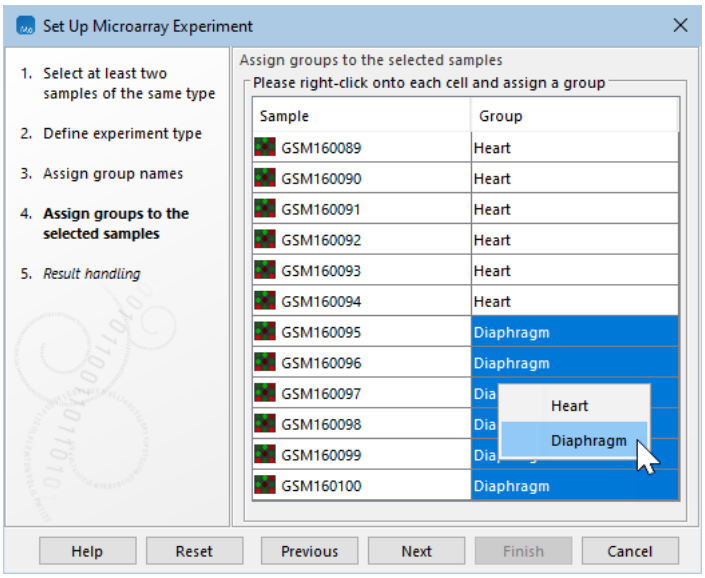

Figure 6: *Six samples are assigned to each group.*

6. In the next step, "Result handling", choose **Save**.
7. In the last step, choose to save the results in the "Microarray Tutorial Data" folder.

   Click on **Finish**.

The resulting `experiment table` includes the feature (i.e., gene) expression values for each sample and different summary values across samples.

Perform quality control

To perform meaningful statistical analysis and inferences, it is essential to first ensure high data quality. We will therefore now perform various quality control analyses.

MA plot and transformation

First we will create an MA plot to assess whether any data normalization and/or transformation is needed. Note that an MA plot can only be used to compare two samples at a time.

1. Launch **Create MA Plot** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select the case expression data", select one of the Heart samples (GSM160089).
3. In the next step, "Select the control data", select one of the Diaphragm samples (GSM160095).
4. In the next step, "Set parameters", select **Original expression values**.
5. In the last step, "Result handling", choose **Open**.

   Click on **Finish**.

The resulting MA plot ( ) (figure 7, left) shows the mean gene expression levels on the X-axis and difference in gene expression levels on the Y-axis for GSM160089 vs. GSM160095. Here, we can see that the variance increases with mean gene expression levels, suggesting that the data should be transformed.

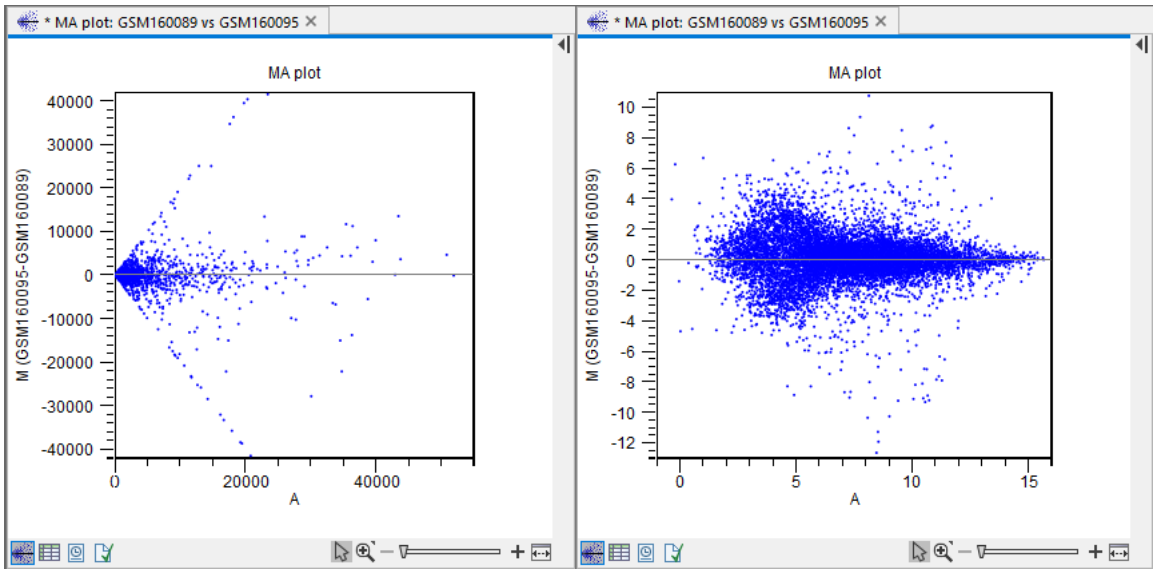

Figure 7: *MA plot before transformation (left) and after log2-transformation (right).*

We will now transform the data and then make a new MA plot using the transformed data:

1. Launch **Transform** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select either samples or an experiment", select GSM160089 and GSM160095.

3. In the next step, "Set parameters", select **Original expression values** under "Values to analyze".

   Under "Transformation method", select **Logarithm transformation** and select "Log2" from the drop-down menu.

4. In the last step, "Result handling", choose **Save**.

   Click on **Finish**.

5. Create an MA plot using the transformed data for GSM160089 and GSM160095, as described in steps 1 to 5.

   This time, in the "Set parameters" step, choose **Transformed expression values**.

The resulting MA plot shows a much more symmetric and even spread (figure 7, right). We will therefore now transform the entire experiment:

1. Repeat steps 1 to 4.

   This time, in the "Select either samples or an experiment" step, select the "Heart vs. Diaphragm" experiment ( ).

Box plot

Next, we will assess whether 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 we will `create a box plot`:

1. Launch **Create Box Plot** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select samples or an experiment", select the "Heart vs. Diaphragm" experiment ( ).
3. In the next step, "Set parameters", choose **Transformed expression values**.
4. In the last step, "Result handling", choose **Open**.

   Click on **Finish**.

The resulting `box plot` shows that all samples have similar distributions (figure 8). This indicates that the samples are comparable. If any of the samples showed a different distribution, they should have been considered for removal.

PCA plot and hierarchical clustering

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

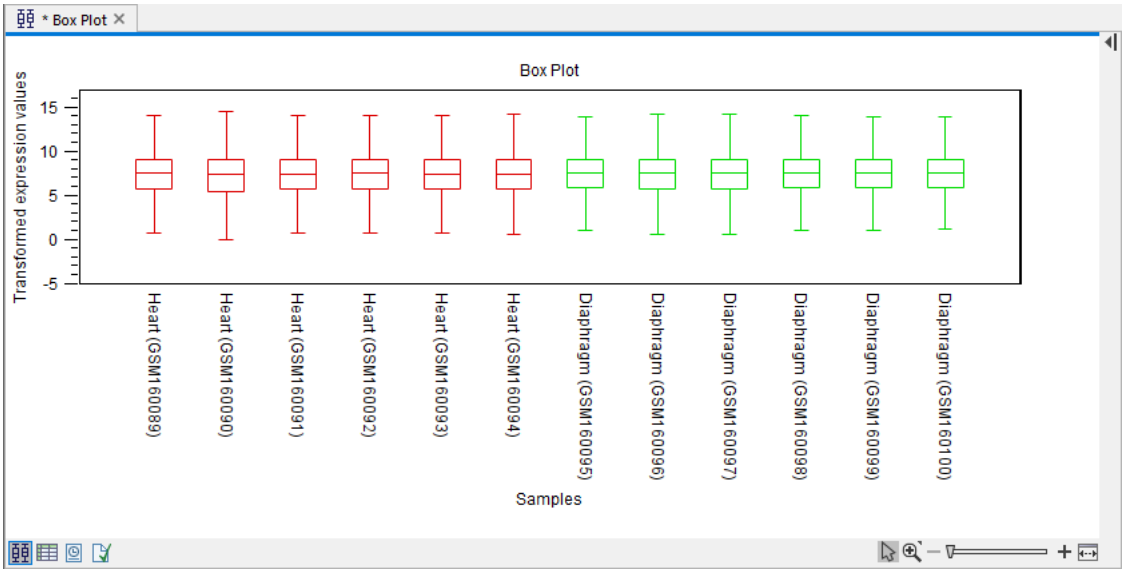

Figure 8: *Box plot of the 12 samples in the experiment.*

First, we will perform a `principal component analysis (PCA)`:

1. Launch **Principal Component Analysis** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select an experiment or samples", select the "Heart vs. Diaphragm" experiment ( ).
3. In the next step, "Set parameters", choose **Original expression values**.
4. In the last step, "Result handling", choose **Open**.

   Click on **Finish**.

In the resulting `PCA plot`, the Heart samples are colored red and the Diaphragm samples are colored green. Check "Show name" in the "Dot properties" Side Panel palette.

The samples clearly cluster by group, although one sample from the Heart group (GSM160090) is an outlier (figure 9).

To complement the principal component analysis, we will now also perform `hierarchical clustering` to see if the samples cluster according to group:

1. Launch **Hierarchical Clustering of Samples** ( ) using Quick Launch ( ).
2. In the first wizard step, "Select at least two samples or exactly one experiment", select the "Heart vs. Diaphragm" experiment ( ).
3. In the next step, "Set parameters", keep the default settings.
4. In the last step, "Result handling", choose **Open**.

   Click on **Finish**.

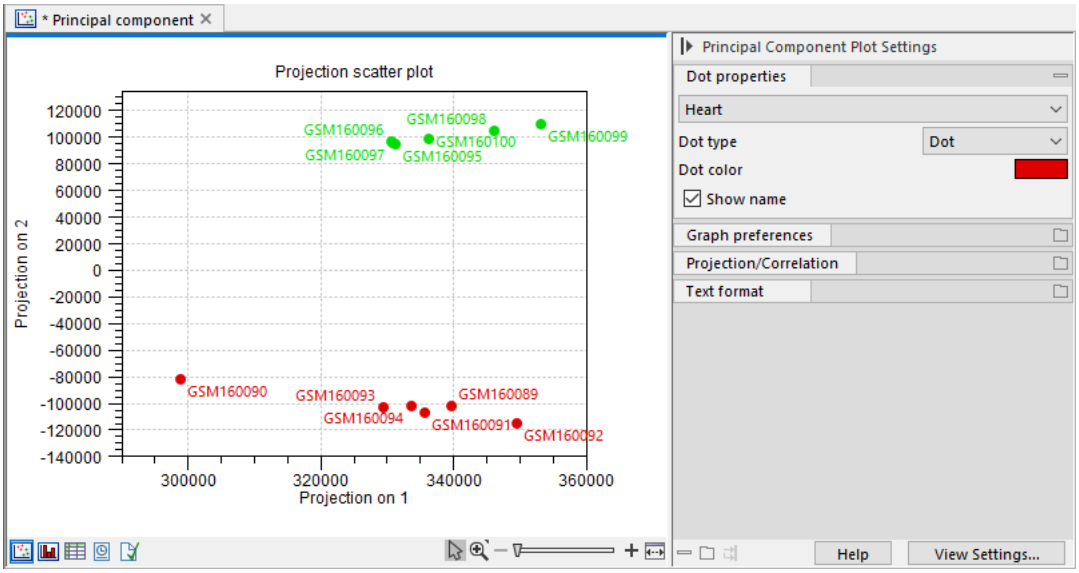

Figure 9: *Principal component analysis showing samples colored by group.*

The resulting `heat map` is added as a view to the experiment and shows how the samples cluster (figure 10).

The two overall clusters 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.

As both the principal component analysis and hierarchical clustering confirm the grouping of the samples, we can confidently conclude that the quality of the samples is satisfactory.

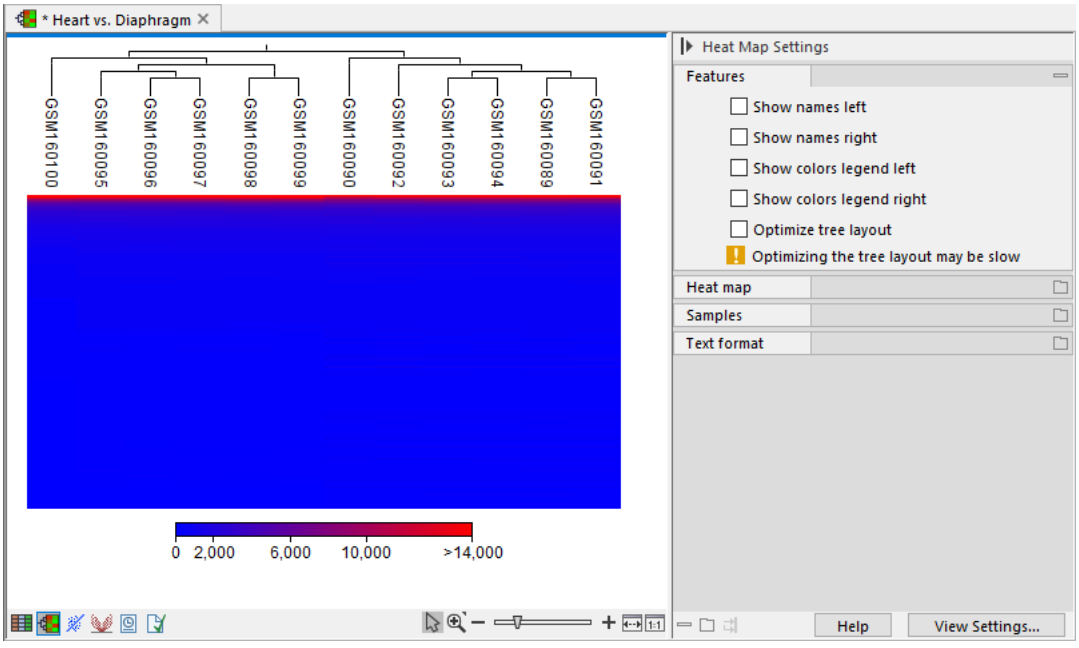

Figure 10: *Heat map view of the experiment showing how the 12 samples cluster.*

Perform statistical analysis

We will now carry out statistical tests for identifying genes that are differentially expressed between the Heart and Diaphragm groups.

T-test

First, we will perform a `t-test` to compare expression values between the two groups:

1. Launch `Gaussian Statistical Analysis` ( ) using Quick Launch ( ).
2. In the first wizard step, "Select one experiment", select the "Heart vs. Diaphragm" experiment ( ).
3. In the next step, "Analysis parameters", keep the default settings.
4. In the next step, "Value parameters", select **Transformed expression values** under "Values to Analyze".
5. In the next step, "Result handling", choose **Save**.
6. In the last step, choose to save the results in the "Microarray Tutorial Data" folder in the Navigation Area. This will save the results to the existing "Heart vs. Diaphragm" experiment.

   Click on **Finish**.

Open the "Heart vs. Diaphragm" experiment. A number of extra columns have been added as a t-test section of the table (figure 11, top).

To select the genes indicative of differential expression, we will use the `advanced filter` located at the top of the table:

1. Click on the arrow at the top-right corner ( ).
2. Select "t-test: Heart vs Diaphragm transformed values - `FDR p-value correction`" and "<" in the drop-down menus.
3. Enter "0.0005" (or "0,0005" depending on your locale settings) in the search field.

   Click on **Filter**. This will filter the table so that only values below 0.0005 are shown (figure 11, top).

Another way of looking at the results from the t-test is to click on the `volcano plot` icon ( ) at the lower left corner of the view (figure 11, bottom).

Now select all the genes in the table that were left after applying the filter above (figure 11, top) by clicking in the table and pressing Ctrl+A (⌘ +A on Mac). The corresponding genes are then shown in red in the volcano plot (figure 11, bottom).

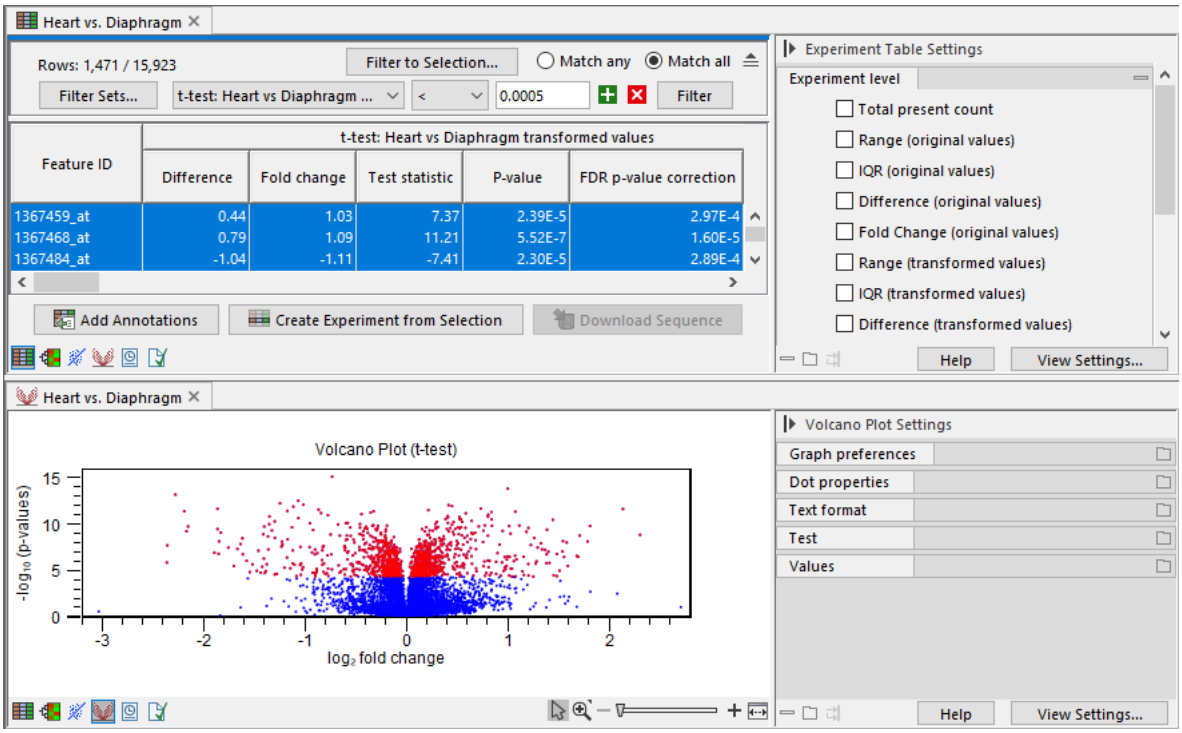

Figure 11: *Top: Filtering t-test results in the experiment table on FDR p-values. Note that not all columns are shown. Bottom: Volcano plot showing genes with FDR <0.0005 in red.*

Filtering on absent/present status and fold change

We will now filter the t-test results on additional criteria:

- Filter genes with an absent status.

  Each gene is assigned an absent/present status by the Affymetrix microarray software. There can be a number of reasons why a gene is absent, and sometimes it is simply because the signal is very weak. Here, we will filter out all genes that are absent more than once in either of the groups.

  - Click on the **Add** (+) icon and set the filtering options to "Heart - Present count", ">=", and "5" (figure 12).
  - Click on the **Add** (+) icon and set the filtering options to "Diaphragm - Present count", ">=", and "5" (figure 12).

- Filter genes with small fold changes.

  We will now filter out genes with negligible expression differences between the two groups, thus only keeping the genes with a pronounced biological impact on the Heart vs. Diaphragm phenotype.

  - Click on the **Add** (+) icon and set the filtering options to "t:test Heart vs Diaphragm transformed values - Difference", "abs value >", and "2" (figure 12).

    Since the data is log-transformed, the group mean difference is really the fold change, so this filter means that we require a fold change above 2.

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

- Click on **Filter**.

  This results in a list of genes that are likely differentially expressed between the two groups.

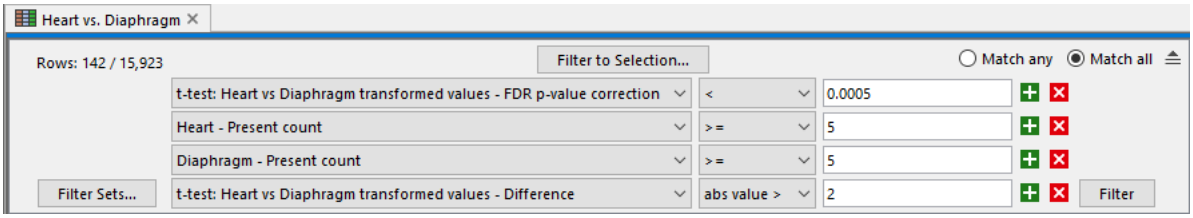

Figure 12: *Filtering genes based on present count and fold change.*

The filtered list of genes indicative of differential expression can be saved:

1. Select all the genes left after filtering by clicking in the table and pressing Ctrl+A (⌘ +A on Mac).
2. Click on **Create Experiment from Selection (**  **)**.
3. In the wizard that opens, choose to **Save** the new experiment in the "Microarray Tutorial Data" folder in the Navigation Area.

Perform annotation tests

Annotation information, such as Gene Ontology categories or KEGG pathways associated with a feature (e.g., a gene), can be used in `annotation tests` to detect significant patterns among experiment features. This may help interpret the analysis results from experiments involving a large number of features within a biological context.

Annotation testing tools require that relevant annotations be associated with the experiment prior to analysis:

1. Import annotations from a `supported format` using `Standard Import`.
2. Associate the annotations with the experiment using `Add Annotations` ().

Subsequently, the following analyses can be performed:

1. `Hypergeometric Test on Annotations` ().

   This test measures whether annotation categories of features in a smaller list (e.g., the filtered genes indicative of differential expression) are over- or under-represented relative to those in a larger list (e.g., all genes in the "Heart vs. Diaphragm" experiment).

   In other words, this test can be used to determine whether specific biological functions or pathways are more prominently associated with a filtered subset of features compared to the full dataset.

2. Gene Set Enrichment Analysis (GSEA) ().

   All features in an experiment are initially ranked based on e.g. the test statistic derived from a differential expression analysis. A GSEA test is then carried out for each annotation category to assess whether the ranks of features within that category are evenly distributed throughout the ranked list.

   In other words, this test can be used to identify biological functions or pathways relevant to comparisons between different groups, without the need to predefine a subset of genes.