# Tutorial: Microarray-based expression analysis part IV: Annotation test

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

# Tutorial: Microarray-based expression analysis part IV: Annotation test

This tutorial is the fourth and final part of a series of tutorials about expression analysis. We continue working with the data set introduced in the first tutorial and analyzed in part two and three.

In this tutorial we will annotate the gene list and use the annotations to see if there is a pattern in the biological annotations of the genes in the list of candidate differentially expressed genes.

We use two different methods for annotation testing: Hypergeometric Tests on Annotations and Gene Set Enrichment Analysis (GSEA).

## Importing and adding the annotations

First step is to import an annotation file used to annotate the arrays. In this case, the data were produced using an Affymetrix chip, and the annotation file can be downloaded from the web site http://www.affymetrix.com. You can access the file by search for **RAE230A**. Note that you have to sign up in order to download the file (this is a free service).

To import the annotation file, click **Import** ( ) in the Tool bar and select the file.

Next, annotate the experiment with the annotation file:

**Toolbox** | **Expression Analysis ( )** | **Annotation Test** | **Add Annotations ( )**

Select the experiment created in the previous tutorial and the annotation file ( ) and click **Next** and **Finish**.

## Inspecting the annotations

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.

An important annotation is the **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.

## Processes that are over-represented in the small list

The first annotation test will show whether any of the GO biological processes are over-represented in our small list of 142 differentially expressed genes relate to the full set of genes measured:

**Toolbox** | **Expression Analysis ( )** | **Annotation Test** | **Hypergeometric Tests on Annotations ( )**

Select the two experiments (the original full experiment and the small subset of 142 genes) and click **Next**. Select **GO biological process** and **Transformed expression values** (see figure 1).

Click **Next** and **Finish** to perform the test. The result is shown in figure 2.

This table lists the GO categories according to p-values for this test. If you take number 2,

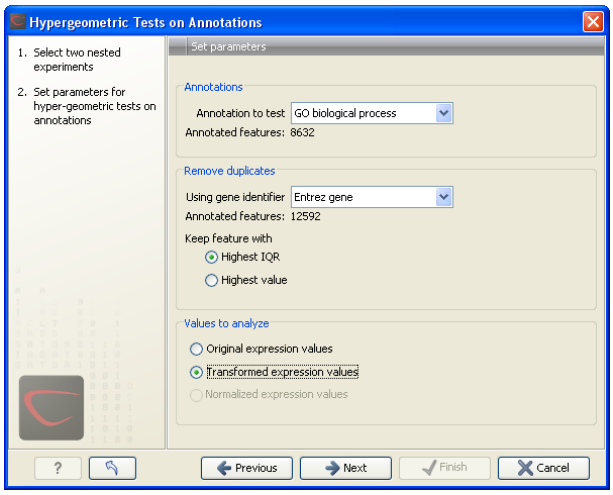
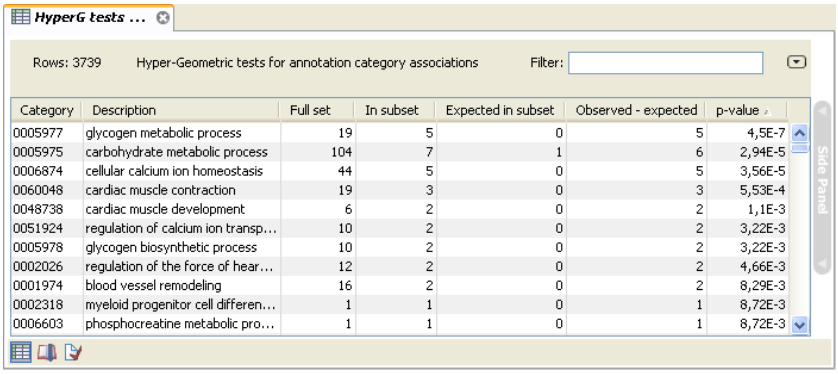



Figure 1: *Testing on GO biological process.*




| Category | Description | Full set | In subset | Expected in subset | Observed - expected | p-value |
|---|---|---|---|---|---|---|
| 0005977 | glycogen metabolic process | 19 | 5 | 0 | 5 | 4,5E-7 |
| 0005975 | carbohydrate metabolic process | 104 | 7 | 1 | 6 | 2,94E-5 |
| 0006874 | cellular calcium ion homeostasis | 44 | 5 | 0 | 5 | 3,56E-5 |
| 0060048 | cardiac muscle contraction | 19 | 3 | 0 | 3 | 5,53E-4 |
| 0048738 | cardiac muscle development | 6 | 2 | 0 | 2 | 1,1E-3 |
| 0051924 | regulation of calcium ion transp... | 10 | 2 | 0 | 2 | 3,22E-3 |
| 0005978 | glycogen biosynthetic process | 10 | 2 | 0 | 2 | 3,22E-3 |
| 0002026 | regulation of the force of hear... | 12 | 2 | 0 | 2 | 4,66E-3 |
| 0001974 | blood vessel remodeling | 16 | 2 | 0 | 2 | 8,29E-3 |
| 0002318 | myeloid progenitor cell differen... | 1 | 1 | 0 | 1 | 8,72E-3 |
| 0006603 | phosphocreatine metabolic pro... | 1 | 1 | 0 | 1 | 8,72E-3 |

Figure 2: *The result of testing on GO biological process.*

carbohydrate metabolic process, there are 104 genes in this category in the full set, if the subset was randomly chosen you would have expected 1 gene to be in the subset. But because there are 7 genes in this subset, this process is over-represented and given a p-value of 2.63E-5.

## A different approach: 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 e.g. 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.

**Toolbox** | **Expression Analysis (** ) | **Annotation Test** | **Gene Set Enrichment Analysis (GSEA) (** )

Select the original full experiment and click **Next**. In this step, make sure the **GO biological process** is chosen (see figure 3.

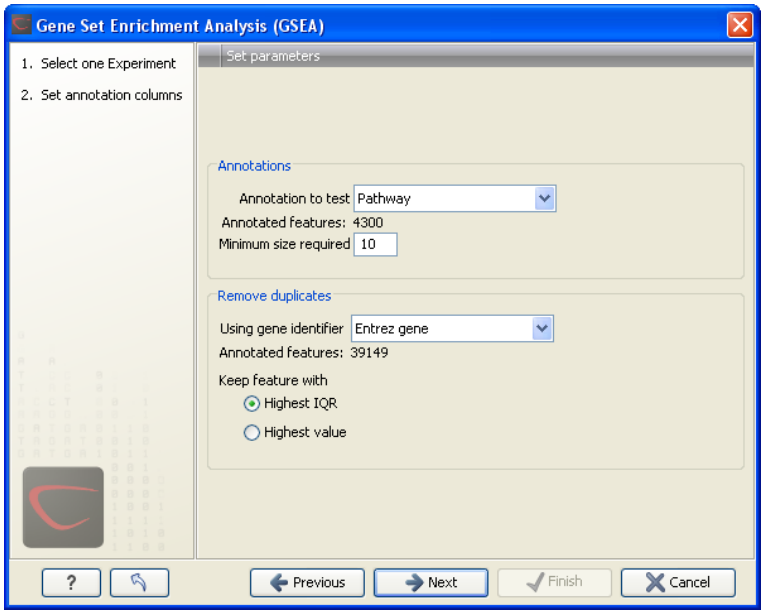
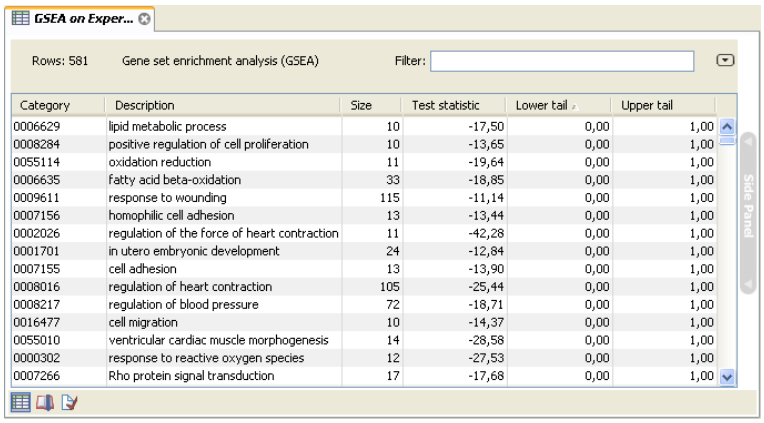



Figure 3: *Gene set enrichment analsysis based on GO biological process.*

Click **Next** and select the **Transformed expression values**. Click **Finish**. The result is shown in figure 4.




| Category | Description | Size | Test statistic | Lower tail | Upper tail |
|---|---|---|---|---|---|
| 0006629 | lipid metabolic process | 10 | -17,50 | 0,00 | 1,00 |
| 0008284 | positive regulation of cell proliferation | 10 | -13,65 | 0,00 | 1,00 |
| 0055114 | oxidation reduction | 11 | -19,64 | 0,00 | 1,00 |
| 0006635 | fatty acid beta-oxidation | 33 | -18,85 | 0,00 | 1,00 |
| 0009611 | response to wounding | 115 | -11,14 | 0,00 | 1,00 |
| 0007156 | homophilic cell adhesion | 13 | -13,44 | 0,00 | 1,00 |
| 0002026 | regulation of the force of heart contraction | 11 | -42,28 | 0,00 | 1,00 |
| 0001701 | in utero embryonic development | 24 | -12,84 | 0,00 | 1,00 |
| 0007155 | cell adhesion | 13 | -13,90 | 0,00 | 1,00 |
| 0008016 | regulation of heart contraction | 105 | -25,44 | 0,00 | 1,00 |
| 0008217 | regulation of blood pressure | 72 | -18,71 | 0,00 | 1,00 |
| 0016477 | cell migration | 10 | -14,37 | 0,00 | 1,00 |
| 0055010 | ventricular cardiac muscle morphogenesis | 14 | -28,58 | 0,00 | 1,00 |
| 0000302 | response to reactive oxygen species | 12 | -27,53 | 0,00 | 1,00 |
| 0007266 | Rho protein signal transduction | 17 | -17,68 | 0,00 | 1,00 |

Figure 4: *The result of a gene set enrichment analsysis based on GO biological process.*

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.

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