

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1 1
A A G G A 0 0 0 0 0
G A T G A 0 1 1 1 1
T A G A T 0 0 1 1 1
G A T G A 1 0 1 1 1
T G T A G 1 0 0 1 1
T A G T A 0 0 0 1 1
G A T A T 1 0 0 1 1
G A G T G 1 1 1 1 1
A G A T T 1 1 1 1 1
G A G T A 1 1 1 1 1
T G A T G 1 1 1 1 1
A T T A G 1 1 1 1 1
T A G A T 1 1 1 1 1
T A G T G 1 1 1 1 1
G A G A G 1 1 1 1 1
G T A G T 1 1 1 1 1
G A T A G 1 1 1 1 1
T A G A G 1 1 1 1 1
A G A G A 1 1 1 1 1
G A G A G 1 1 1 1 1
A G A G A 1 1 1 1 1
G A G A G 1 1 1 1 1
A G A G A 1 1 1 1 1
T A G A G 1 1 1 1 1

Tutorial

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

August 23, 2010

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 or under 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,

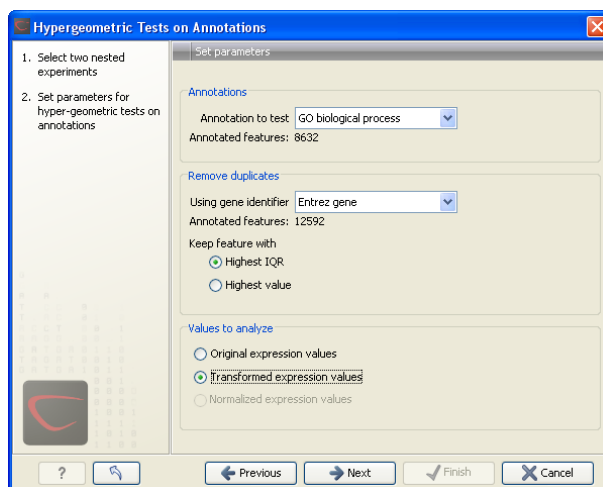


Figure 1: Testing on GO biological process.

Rows: 3,740 Hyper-Geometric ... Filter:

Category	Description	Full set	In subset	Expected i...	Observed ...	p-value
0005977	glycogen me...	19	5	0	5	3,32E-7
0005975	carbohydrat...	104	7	1	6	2,03E-5
0006874	cellular calciu...	44	5	0	5	2,64E-5
0060048	cardiac musc...	19	3	0	3	4,55E-4
0048738	cardiac musc...	6	2	0	2	9,56E-4
0006807	nitrogen com...	8	2	0	2	1,77E-3
0051924	regulation of...	10	2	0	2	2,81E-3
0005978	glycogen bio...	11	2	0	2	3,41E-3
0002026	regulation of...	12	2	0	2	4,07E-3
0001974	blood vessel ...	16	2	0	2	7,25E-3
0002318	myeloid prog...	1	1	0	1	8,12E-3
0006603	phosphocrea...	1	1	0	1	8,12E-3
0042424	catecholamin...	1	1	0	1	8,12E-3
0031275	regulation of...	1	1	0	1	8,12E-3
0046439	L-cysteine m...	1	1	0	1	8,12E-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).

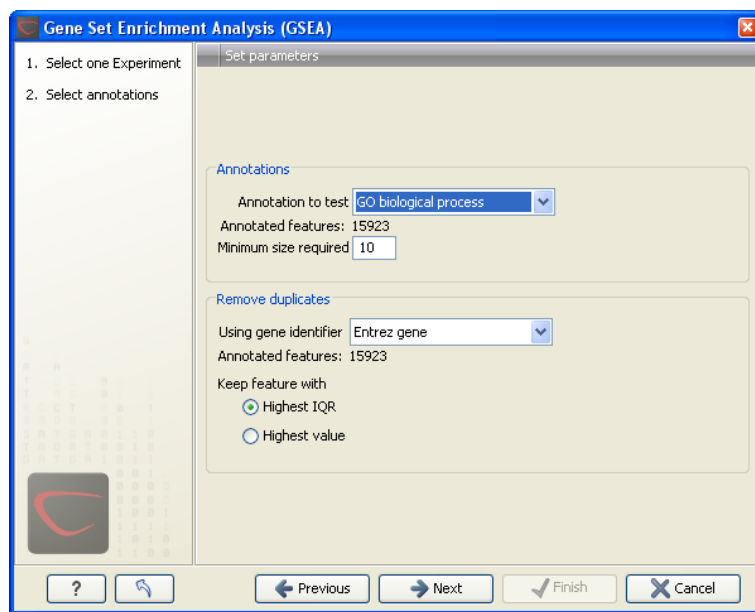


Figure 3: Gene set enrichment analysis based on GO biological process.

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
0006412	translation	204	-13,92	0,00	1,00
0006941	striated muscle ...	15	-28,52	0,00	1,00
0006936	muscle contract...	26	-37,56	0,00	1,00
0006937	regulation of m...	17	-30,63	0,00	1,00
0007519	skeletal muscle ...	30	-20,45	1E-4	1,00
0005977	glycogen meta...	19	-19,62	2E-4	1,00
0007517	muscle develop...	21	-15,18	1,7E-3	1,00
0001501	skeletal develo...	42	-14,10	2,1E-3	1,00
0006094	gluconeogenesis	16	-14,19	3,5E-3	1,00
0009749	response to glu...	37	-12,23	5,4E-3	0,99
0006414	translational el...	12	-13,48	6E-3	0,99
0001756	somitogenesis	13	-12,34	6,8E-3	0,99
0007528	neuromuscular ...	13	-12,81	7,8E-3	0,99
0005978	glycogen biosy...	11	-12,03	8,8E-3	0,99
0007200	...	25	-10,62	0,7E-2	0,00

Figure 4: The result of a gene set enrichment analysis 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.