

# How to test association of a group of genes with a clinical outcome?

Jelle Goeman

March 24, 2003

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Starting out</b>	<b>2</b>
<b>3</b>	<b>Testing</b>	<b>2</b>
<b>4</b>	<b>Variants</b>	<b>3</b>
<b>5</b>	<b>Diagnostic plots</b>	<b>5</b>
5.1	Permutations plot . . . . .	5
5.2	Gene Plot . . . . .	6
5.3	Checkerboard plot . . . . .	7
5.4	Regression Plot . . . . .	8

## 1 Introduction

This HowTo document explains how to test whether a given group of genes is significantly associated with a clinical outcome using the R-package **globaltest**. This explanation uses simulated data available with the package.

The Global Test tests whether the expression pattern of a group of genes (e.g. a pathway) is significantly related to a clinical outcome. It does this by testing whether the clinical outcome can be predicted on the basis of the expressions of the genes. The test rejects when the test statistic  $Q$  is high, which happens when those pairs of samples which have a similar gene expression pattern over the genes of interest also tend to have a similar clinical outcome.

For real data examples, for a more extensive explanation of the ideas behind the test and for the mathematical details we refer to the Technical Report:

J. J. Goeman, S.A. van de Geer, F. de Kort and J. C. van Houwelingen, *A global test for association of a group of genes with a clinical outcome*, Technical Report MI 2003-03, Mathematical Institute, Leiden University.

This is available from [www.math.leidenuniv.nl/~jgoeman](http://www.math.leidenuniv.nl/~jgoeman).

## 2 Starting out

First install the package and load the example data

```
> library(globaltest)
```

Loading required package: methods

```
> data(exampleX)
> data(exampleY)
```

Now we have **exampleX**, normalized microarray data for 1,000 genes and 40 samples and **exampleY**, a clinical outcome for each of the samples. This clinical outcome has a 1 for the samples in group 1 and a 0 for the samples in group 2.

To apply the Global Test to your own data, make sure the expression values are arranged in a matrix with the genes as rows and the samples as columns. Gene and samples names should be added as row- and column names to this matrix. The clinical outcome should be organized as a vector.

## 3 Testing

Suppose we are interested in a specific pathway containing 25 genes. We want to know whether this group of genes is associated with the clinical outcome. In this case these are the genes corresponding to rows 1 to 25 in the data matrix **exampleX**. If we have a vector of names of the genes we can look up their row numbers by applying the function **gene2ix** from the package **globaltest** (See **help(gene2ix)**).

Always first test all genes to see if the overall gene expression pattern is different for different clinical outcomes.

```
> gt.all <- globaltest(exampleX, exampleY)
> gt.all
```

Global Test result:  
1000 out of 1000 genes used; 40 samples

p value = 0.1304  
based on theoretical distribution

Test statistic Q = 78.74  
with expectation EQ = 75  
and standard deviation sdQ = 3.312 under the null hypothesis

We conclude that there is no evidence that the overall gene expression pattern for all 1,000 genes is associated with the clinical outcome.

Next we test the pathway of interest:

```
> pathway <- 1:25  
> gt.pw <- globaltest(exampleX, exampleY, test.genes = pathway)  
> gt.pw
```

Global Test result:  
25 out of 1000 genes used; 40 samples

p value = 0.000337  
based on theoretical distribution

Test statistic Q = 160.3  
with expectation EQ = 71.88  
and standard deviation sdQ = 20.08 under the null hypothesis

This means that the expression pattern of the pathway of interest is different between the samples in group 1 and group 2. Samples with similar clinical outcomes tend to have similar expression patterns for this pathway.

## 4 Variants

The procedure outlined in the preceding section is for a two-valued clinical outcome. The Global Test can also be applied to find association of gene expression patterns with a continuous clinical outcome using the option `model = 'linear'`.

Furthermore, if the sample size is small, the asymptotic formula's used to calculate the p-value might be inaccurate. In that case a permutation version of the test might be used.

```
> gt.pw <- globaltest(exampleX, exampleY, test.genes = pathway,  
+   permutation = TRUE)  
> gt.pw
```

Global Test result:  
25 out of 1000 genes used; 40 samples

p value = 6e-04  
based on 10000 permutations

Test statistic  $Q = 160.3$   
with expectation  $EQ = 73.75$   
and standard deviation  $sdQ = 20.57$  under the null hypothesis

The option `permutation = TRUE` generates 10,000 permutations. If desired, with the option `nperm` the user can change the number of permutations. Choosing fewer permutations will speed up the program, choosing more will generate a more accurate p-value.

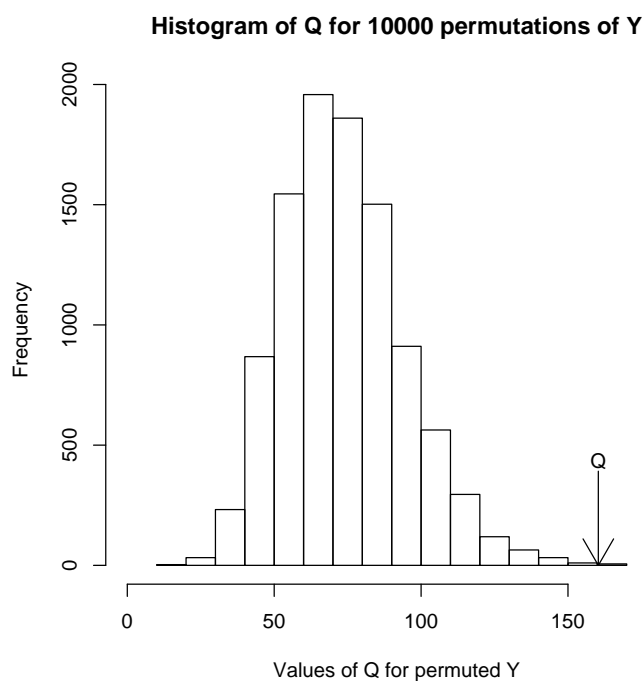
## 5 Diagnostic plots

There are various types of diagnostic plots available in the package `globaltest` to help the user interpret the test result.

### 5.1 Permutations plot

The first is the permutations plot. It can only be used if the permutation version of the Global Test was used. It plots the values of the test statistic  $Q$  calculated for permutations of the clinical outcome in a histogram. The observed value of  $Q$  for the true values of the clinical outcome is marked with an arrow.

```
> permutations(gt.pw)
```

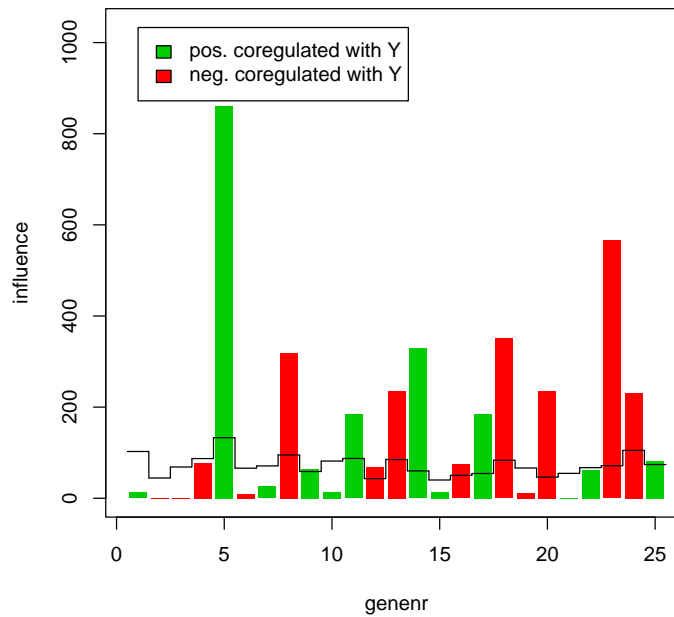


## 5.2 Gene Plot

The second diagnostic plot is the Gene Plot which can be used to assess the influence of each gene on the outcome of the test. The Gene Plot has a bar and a reference line for each gene tested. The bar indicates the influence of each gene on the test statistic (the test statistic for the group is the average of the bars for the genes). The reference line gives the expected height of the bar under the null hypothesis that the gene is not associated with the clinical outcome. Finally the bars are colored to indicate a positive or a negative association of the gene with the clinical outcome.

The function returns a legend for the plot, connecting the gene numbers appearing in the plot to the gene names.

```
> legend <- geneplot(gt.pw)
```



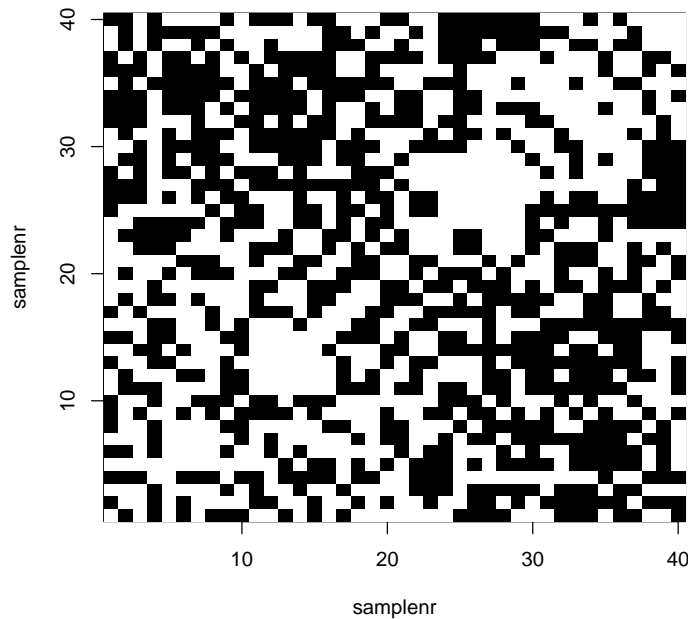
### 5.3 Checkerboard plot

The third and fourth diagnostic plots can both be used to assess the influence of each of the samples on the test result. The checkerboard plot visualizes the similarity between samples. It makes a square figure with the samples both on the X and on the Y-axis, so that it shows all possible comparisons between the samples. Samples which are relatively similar are coded white and samples which are relatively dissimilar are coded black.

For easier interpretation the samples are sorted by their clinical outcome. If the test was (very) significant and the clinical outcome has two values, a typical block-like structure will appear. If the clinical outcome was continuous and the test is significant, the black squares will tend to be away from the diagonal. By looking at these patterns some things can be learned about the structure of the data. For example, by looking at samples which deviate from the main pattern, outlying samples can be detected.

The function `checkerboard` also returns a legend to link the numbers appearing in the plot to the sample names.

```
> legend <- checkerboard(gt.pw)
```



## 5.4 Regression Plot

Using the regression plot a more precise assessment can be made of the influence of each sample on the result of the test. This graph plots the same pairs of samples, showing the covariance between their clinical outcomes on the X-axis and the covariance between their gene expression patterns on the Y-axis. The comparisons of each sample with itself have been excluded.

The test statistic of the Global Test can be seen as a regression-coefficient for this plot, so it is visualized by drawing a least squares regression line. If this regression line is steep, the test statistic has a large value (and is possibly significant).

The influence of specific samples can be assessed by drawing a second regression line through only those points in the plot, which are comparisons involving the sample of interest. For example if we are interested in sample nr. 40, we take the points corresponding to the pairs (1,40) up to (39,40). If the regression line drawn through only these points deviates much from the general line, the sample deviates from the general pattern. This is especially the case if this line has a negative slope, which means that the sample is more similar (in its gene expression pattern) to the samples with a different clinical outcome than to samples with a similar clinical outcome.

If we want to test sample nr. 40, we say

```
> regressionplot(gt.pw, 40)
```

Samples investigated:

sample40

