— Global Testing —

Ulrich Mansmann, Reinhard Meister, Manuela Hummel

Practical DNA Microarray Analysis, March 2008, Heidelberg
http://compdiag.molgen.mpg.de/ngfn/pma2008mar.shtml

Abstract. *This is the tutorial for the exercises about global testing on day 4 of the course on Practical DNA Microarray Analysis. The global test introduced by Goeman et al. (2004) and the global ANCOVA approach of Mansmann and Meister (2005) and Hummel et al. (2008) are practiced.*

1 Preliminaries

Load data and packages. Have a look on the data. It stems from a study on breast cancer from van't Veer et al. (2002). There is an expression matrix (`vantVeer`), a data frame giving phenotype information for all samples (`phenodata`) and a list of nine cancer related pathways (`pathways`) each consisting of corresponding probe set names. In order to provide a feasible data set we reduced the available information of gene expression of 1113 genes. These are exactly the genes of the nine cancer related pathways. We take one gene from the orignial data additionally to the expression set, namely 'AL137718'. This gene is part of the original van't Veer prognosis signature. We will use it to demonstrate how signature genes can be related to pathways.

```
> library(GlobalAncova)
> library(globaltest)
> data(vantVeer)
> data(phenodata)
> data(pathways)
> dim(vantVeer)
> vantVeer[1:10, 1:10]
> str(phenodata)
> str(pathways)
```

Assume we are interested in differential expression between relevant prognostic groups, defined by the development of distant metastases within five years (`metastases`). Further we have covariate information about the tumor grade (`grade`) and the Estrogen receptor status (`ERstatus`). If we attach the phenotype data frame to the R search path all its variables can be accessed by simply giving their names.

```
> attach(phenodata)
> table(metastases)
```

2 Global Testing of a Single Pathway

We start by applying global tests to all genes in the dataset so that differences in the overall gene-expression pattern can be demonstrated. We would like to study if the genes in the cancer related pathways

contain prognostic power with repect to future metastases (globaltest). (Therefore we leave out gene 'AL137718', which does not belong to any of the pathways.) We are also interested in the differences in mean expression between both prognostic groups (GlobalAncova). The use of GlobalAncova in this situation is quite artificial because GlobalAncova is not deviced for prognostic problems. However, it is of interest to compare both procedures. Both, globaltest and GlobalAncova provide permutation as well as approximative p-values. Here we just calculate the latter ones.

```
> index <- rownames(vantVeer) != "AL137718"
> gt.all <- globaltest(X = vantVeer[index, ], Y = metastases)
> gt.all
> ga.all <- GlobalAncova(xx = vantVeer[index, ], group = metastases,
+     method = "approx")
> ga.all
```

The globaltest checks if gene expression allows for predicting future metastases while GlobalAncova assesses difference in mean gene expression between both groups of patients.

GlobalAncova may also be called in a more general way by definition of two linear models that shall be compared. Hence model formulas for the full model containing all parameters and the reduced model, where the terms of interest are omitted, have to be given. An alternative is to provide the formula for the full model and a character vector naming the terms of interest. Consequently we could run the same analysis as above with two possible further function calls shown below. In both cases a data frame with information about all variables for each sample is required. (In the case of microarray data this can usually be the corresponding pData object.) Such model definitions will be useful for more complex analysis tasks (see later).

```
> GlobalAncova(xx = vantVeer[index, ], formula.full = ~metastases,
+     formula.red = ~1, model.dat = phenodata, method = "approx")
> GlobalAncova(xx = vantVeer[index, ], formula.full = ~metastases,
+     test.terms = "metastases", model.dat = phenodata,
+     method = "approx")
```

From the result we conclude that the overall gene expression profile for all genes is associated with the clinical outcome.

Now we consider a special group of genes, e.g. the p53-signalling pathway. We apply the global test to this pathway using the options genesets and test.genes, respectively.

```
> p53 <- pathways$p53_signalling
> gt.p53 <- globaltest(X = vantVeer, Y = metastases, genesets = p53)
> gt.p53
> ga.p53 <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = p53, method = "approx")
> ga.p53
```

3 Adjusting for Covariates

The adjustment for covariate information is possible with both methods. For example we can adjust for the Estrogen receptor status.

```
> rownames(phenodata) <- Sample
> gt.adj <- globaltest(X = vantVeer, Y = metastases ~ ERstatus,
```

```
+     adjust = phenodata, genesets = p53)
> gt.adj
> ga.adj <- GlobalAncova(xx = vantVeer, group = metastases,
+     covars = ERstatus, test.genes = p53, method = "approx")
> ga.adj
```

With the more general `GlobalAncova` function call we would simply adjust the definitions of model formulas, namely `formula.full = ~ metastases + ERstatus` and `formula.red = ~ ERstatus`.

4 Testing Several Pathways Simultaneously

The user can apply `globaltest` and `GlobalAncova` to compute p–values for a couple of pathways with one call.

```
> gt.pw <- globaltest(X = vantVeer, Y = metastases, genesets = pathways)
> gt.pw
> ga.pw <- GlobalAncova(xx = vantVeer, group = metastases,
+     test.genes = pathways, method = "approx")
> ga.pw
```

Afterwards a suitable correction for multiple testing has to be applied. Note however that due to the extremely high correlations between these tests, many procedures that correct for multiple testing here are inappropriate. An appropriate method would be for example the (rather conservative) Holm correction.

```
> gt.pw.raw <- p.value(gt.pw)
> gt.pw.adj <- p.adjust(gt.pw.raw, "holm")
> gt.result <- cbind(raw = gt.pw.raw, Holm = gt.pw.adj)
> gt.result
> ga.pw.raw <- ga.pw[, "p.approx"]
> ga.pw.adj <- p.adjust(ga.pw.raw, "holm")
> ga.result <- cbind(raw = ga.pw.raw, Holm = ga.pw.adj)
> ga.result
```

5 Analysis of Arbitrary Clinical Variables

With `GlobalAncova` also clinical variables with more than two groups or even continuous ones can be considered. For demonstration we investigate differential expression for the three ordered levels of tumor grade and again only the p53-signalling pathway.

```
> ga.grade <- GlobalAncova(xx = vantVeer, formula.full = ~grade,
+     formula.red = ~1, model.dat = phenodata, test.genes = p53,
+     method = "approx")
> ga.grade
```

6 Gene–Gene Interaction

Now we want to go into the matter of other interesting biological questions. For example one might ask if there exists interaction between the expression of special genes (e.g. genes from a prognosis signature)

and the expression of genes in a certain pathway. This question can be answered by viewing the expression values of the signature genes as linear regressors and by testing their effects on the expression pattern of the pathway genes. For demonstration we pick the gene 'AL137718' which is not part of any of the pathways. Assume that we also want to adjust for the Estrogen receptor status.

```
> signature.gene <- "AL137718"
> model <- data.frame(phenodata, signature.gene = vantVeer[signature.gene,
+     ])
> ga.signature <- GlobalAncova(xx = vantVeer, formula.full = ~signature.gene +
+     ERstatus, formula.red = ~ERstatus, model.dat = model,
+     test.genes = p53, method = "approx")
> ga.signature
```

7 Co–Expression

Next we want to analyse co–expression regarding the clinical outcome of building distant metastases within five years. This can be done by simply adding the variable `metastases` to the full and reduced model, respectively. Such layout corresponds to testing the linear effect of the signature gene stratified not only by Estrogen receptor status but also by `metastases`.

```
> ga.coexpr <- GlobalAncova(xx = vantVeer, formula.full = ~metastases +
+     signature.gene + ERstatus, formula.red = ~metastases +
+     ERstatus, model.dat = model, test.genes = p53, method = "approx")
> ga.coexpr
```

Supposably the most interesting question in this case concerns differential co–expression. Differential co–expression is on hand if the effect of the signature gene behaves different in both `metastases` groups. In a one dimensional context this would become manifest by different slopes of the regression lines. Hence what we have to test is the interaction between `metastases` and `signature.gene`.

```
> ga.diffcoexpr <- GlobalAncova(xx = vantVeer, formula.full = ~metastases *
+     signature.gene + ERstatus, formula.red = ~metastases +
+     signature.gene + ERstatus, model.dat = model, test.genes = p53,
+     method = "approx")
> ga.diffcoexpr
```

With `globaltest` we can also test gene-gene interaction, also adjusted for phenotype groups. But it is not possible to test for differential co-expression or the influence of more than one signature gene on a pathway. On the other hand it is able to deal with survival times as clinical outcome.

8 Plots

The functions `geneplot` (`globaltest`) and `Plot.genes` (`GlobalAncova`) visualize the influence of individual genes on the test result. The `sampleplot` (`globaltest`) shows the influence of samples on the test result. If a sample has a positive bar, its expression profile is relatively similar to that of samples which have the same value of the clinical variable. Similarly, `Plot.subjects` (`GlobalAncova`) shows how well the model fits to the individual samples. If an individual does not fit into the expression pattern of its clinical group, negative values can occur. A small p-value will therefore generally coincide with many positive bars. (Figure 1)

```
> geneplot(gt.p53)
> Plot.genes(xx = vantVeer, group = metastases, test.genes = p53)
> sampleplot(gt.p53)
> Plot.subjects(xx = vantVeer, group = metastases, test.genes = p53,
+     legendpos = "bottomright")
```

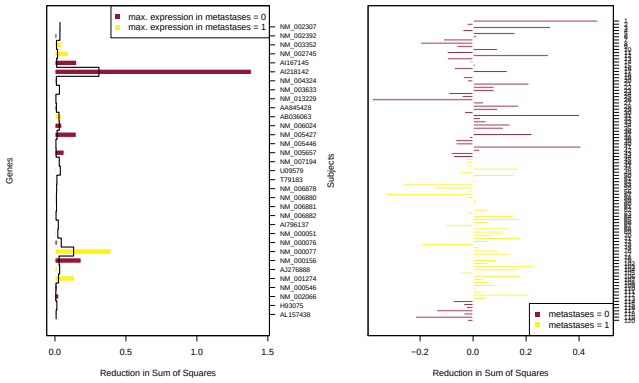

Figure 1: Gene and subjects plot (with `GlobalAncova`).

With `GlobalAncova` also plots for more general models are available. For example we could ask whether there exists differential expression between the three different tumour stages. In both plots a variable for defining the coloring can be chosen by the user. In the subjects plot samples can be sorted for a better overview. In figure 2 e.g. we see that patients with tumour grade 1 have quite homogeneous expression patterns whereas patients with grade 3 are more heterogeneous.

```
> Plot.subjects(xx = vantVeer, formula.full = ~grade, formula.red = ~1,
+     model.dat = phenodata, test.genes = p53, Colorgroup = "grade",
+     sort = TRUE, legendpos = "topleft")
```

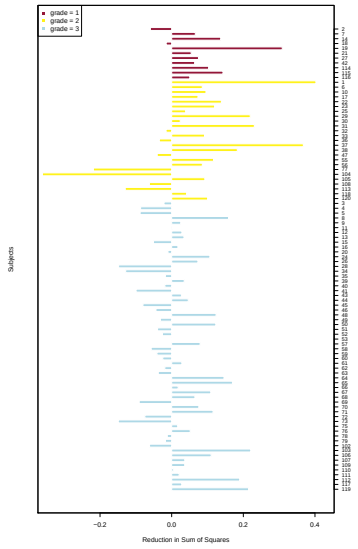

Figure 2: Subjects plot.

The package `globaltest` provides the checkerboard as another diagnostic plot. It shows similarities between samples. For high similarities the respective squares are colored white, for relatively different samples they are colored black. (Figure 3)

```
> checkerboard(gt.p53)
```

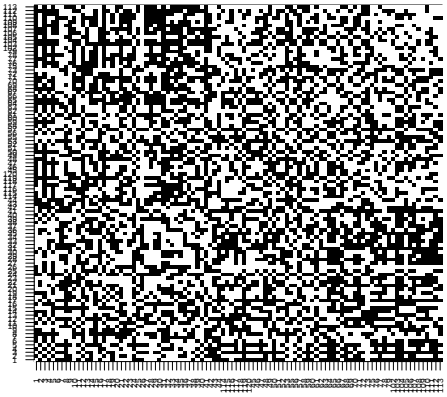

Figure 3: Checkerboard plot.

9 References

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; 20(1): 93-9.

Holm S. A simple sequentially rejective multiple test procedure. Scand. J. Statist 1979; 6: 65-70.

Hummel M, Meister R and Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. Bioinformatics 2008; 24 (1): 78-85.

Mansmann U and Meister R. Testing differential gene expression in functional groups. Methods Inf Med 2005; 44 (3): 449-53.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-536.