

# Reducing the Non-Discovery Rate by use of auxiliary variables in microarray experiments

Maëla Kloareg<sup>(1)</sup>, David Causeur<sup>(1,2)</sup> and Marie Damon<sup>(3)</sup>

<sup>(1)</sup>: IRMAR, UMR 6625 CNRS, Agrocampus Rennes, 65 rue de St-Brieuc, CS 84215, 35042 Rennes Cedex, France

<sup>(2)</sup>: CREST-ENSAI, France

<sup>(3)</sup>: INRA, UMR SENAH, St-Gilles, France

Although multiple testing issues have been widely discussed in the statistical literature for a long time, novel approaches have emerged in recent years for the analysis of gene expression data. In this situation, the main goal is to identify the genes that show good evidence of being differentially expressed under two conditions (eg. treatments, genotypes or times in studies of kinetics).

In the recent discussions about a type I error rate that would yield to less conservative decision rules than the traditional Bonferroni procedure, a major innovation has come from Benjamini and Hochberg (1995), who define the false discovery rate (FDR) as the proportion of true  $H_0$  among the tests for which  $H_0$  is rejected. Benjamini and Hochberg (1995) also provide a decision rule, shown by Benjamini and Yekutieli (2001) to control the FDR under a large class of positive dependency between the test statistics.

Attempts to improve the existing methods always involve better knowledge of the responses' dependency structure. However the high dimensionality of the data usually prohibits the modelling of the whole set of gene expressions' joint distribution. As mentioned by Kendzioriski *et al.* (2003), treating variables as independent tend to be less efficient than some Bayesian approaches, which take advantage of the shared information between genes. In many situations, relating the gene expressions to phenotypic variables can also give insight into the correlation structure of sets of genes. But integrating biological relevant knowledge and gene expressions in differential analysis is still much less usual than in multivariate exploratory data analysis for instance.

The aim of our paper is to propose a testing method based on moderated t-statistics that integrates external information to improve the power of the usual testing strategies. This external information is supposed to be available in the sample for which the microarray data are observed and also on additional items for which microarrays are not available. Improving inference by use of auxiliary variables in such a double-sampling framework is not novel in some areas of statistics. Indeed, multiple sampling strategies are usually dedicated to improve estimation procedures, but more rarely to testing issues.

---

1. Tel.: +33-2-23-48-58-81; fax: +33-2-23-48-58-71.

*E-mail address*: maëla.kloareg@agrocampus-rennes.fr (M. Kloareg)

In a multivariate regression framework, many papers have dealt with the optimal allocation of the measurements of the outcome and the auxiliary variable (see Causeur, 2005). The starting point of the present paper comes from Causeur and Husson (2007), who adapted the methodology to optimal testing procedures.

Figure 1 gives an idea of the gain made possible by the use of a relevant auxiliary variable, which intra-group correlation with the gene expressions is  $\rho$ . The simulated double-sampling scheme is based on  $n = 20$  microarrays with  $p = 100$  genes (50 under  $H_1$ ) and  $N = 150$  measurements of the covariate. The Non-Discovery Rate (NDR) is derived for a usual (single-sampling) Benjamini-Hochberg procedure and for a modified (double-sampling) procedure.

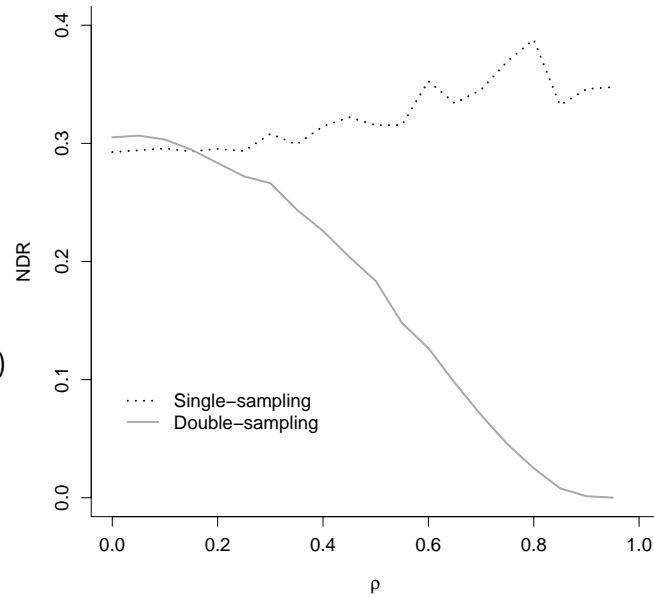


Fig. 1: NDR for various values of  $\rho$ .

In the talk, the double-sampling method is implemented in a situation where microarray data are used to select the genes that affect the degree of muscle destructure in pigs.

## References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**, 289–300.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependence. *Ann. Statist.*, **29**, pp. 1165–1188.
- Causeur, D. (2005). Optimal sampling from concomitant variables for regression problems. *Journal of Statistical planning and Inference*. **128**, 289–301.
- Causeur, D. and Husson, F. (2007). Asymptotic Distribution of Double-Sampling Tests for General Linear Hypotheses. *Statistics*, to appear.
- Kendzioriski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric Empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, **22** (24), 3899–3914.