Using functions *samrocNboot* and *pava.fdr*

Per Broberg

November 24, 2005

Analysis of the data from Golub *et al.*

Consider the microarray experiment in Golub et al. (1999) where ALL and AML types are compared. The data are available within package *multtest*. We can analyse those data in SAGx with the function *samrocNboot*. The ideas behind it are presented in Broberg (2003). Briefly, the method relies on a penalised t-test statistica $d = (\bar{x}_1 - \bar{x}_2)/(S + a)$ Efron et al. (2001). Example code now follows

```
> library(multtest)
> data(golub)
> set.seed(849867)
> samroc.res <- samrocNboot(data = golub, formula = ~as.factor(golub.cl))
> samroc.res$p0
```

```
[1] 0.3730628
```

The function *samrocNboot* is used to perform a penalised t-test. The estimated proportion unchanged genes equals 0.37. The distribution of p-values is shown in Figure 1, which confirms that many genes are changed. Furthermore, using the function *pava.fdr* we obtain estimates of the FDR and of the local FDR, see Figure 2. This function is presented in Broberg (2005) and combines the local FDR estimator of Aubert et al. (2004) with Poisson regression (see Efron (2004)) and isotonic regression.

```
> par(bg = "cornsilk")
> hist(samroc.res$pvalue, xlab = "p-value", main = "", col = "orange",
+     freq = F)
> print(abline(samroc.res$p0, 0, col = "red"))
```
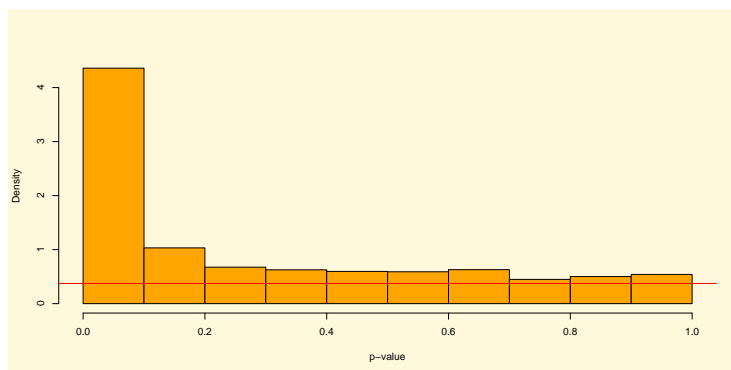
NULL

Figure 1: Histogram of the *p*-values generated by function *samrocNboot*

```
> par(bg = "cornsilk")
> fdrs <- pava.fdr(ps = samroc.res$pvalues)
> plot(samroc.res$pvalues, fdrs$pava.local.fdr, type = "n", xlab = "p-value",
+     ylab = "False Discovery Rate (FDR)")
> lines(lowess(samroc.res$pvalues, fdrs$pava.local.fdr), col = "red")
> lines(lowess(samroc.res$pvalues, fdrs$pava.fdr), col = "blue")
> legend(0.1, 0.9, pch = NULL, col = c("red", "blue"), c("pava local FDR",
+     "pava FDR"), lty = 1)
```
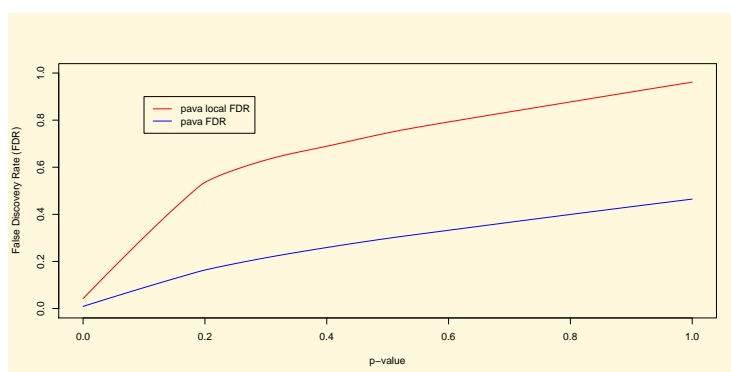

Figure 2: Scatter plot of the local false discovery rate and the false discovery rate as estimated by function *pava.fdr*

References

J Aubert, A Bar-Hen, JJ Daudin, and S Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. *BMC Bioinformatics*, 5:125, 2004. doi: http://dx.doi.org/10.1186/1471-2105-5-125.

P Broberg. Statistical methods for ranking differentially expressed genes. *Genome Biology*, 4:R41, 2003. doi: http://dx.doi.org/10.1186/gb-2003-4-6-r41. URL http://genomebiology.com/2003/4/6/R41.

Per Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6(1):199, 2005. ISSN 1471-2105. doi: http://dx.doi.org/10.1186/1471-2105-6-199. URL http://www.biomedcentral.com/1471-2105/6/199.

Bradley Efron. Selection and estimation for large-scale simultaneous inference. preprint available at the http://www-stat.stanford.edu/ brad/papers/Selection.pdf, 2004.

Bradley Efron, Robert Tibshirani, John Storey, and Victoria Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 2001.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.