EMA - A R package for Easy Microarray data Analysis

Nicolas Servant[1,2,3,*], **Eleonore Gravier**[1,2,3,4], **Pierre Gestraud**[1,2,3], **Cecile Laurent**[1,2,3,6,7,8], **Caroline Paccard**[1,2,3], **Anne Biton**[1,2,3,5], **Isabel Brito**[1,2,3], **Jonas Mandel**[1,2,3], **Bernard Asselain**[1,2,3], **Emmanuel Barillot**[1,2,3], **Philippe Hupé**[1,2,3,5]

1. Institut Curie, Paris F-75248, France
2. INSERM, U900, Paris F-75248, France
3. Ecole des Mines ParisTech, Fontainebleau, F-77300 France
4. Institut Curie, Departement de Transfert, Paris F-75248, France
5. CNRS, UMR144, Paris F-75248, France
6. CNRS, UMR3347, Orsay F-91405, France
7. INSERM, U1021, Orsay F-91405, France
8. Université Paris-Sud 11, Orsay F-91405, France
*Contact author: ema-support@curie.fr

Keywords: Microarray analysis

The increasing number of methodologies and tools currently available to analyse gene expression microarray data can be confusing for non specialist users. Based on the experience of biostatisticians of Institut Curie, we propose both a clear analysis strategy and a selection of tools to investigate microarray gene expression data. The most usual and relevant existing *R* functions were discussed, validated and gathered in an easy-to-use *R* package (**EMA**) devoted to gene expression microarray analysis.

Removing noise and systematic biases is performed using the most famous techniques for Affymetrix GeneChip normalisation. The data are then filtered to both reduce the noise and increase the statistical power of the subsequent analysis. Exploratory approaches based on R packages such as **FactoMineR**, or **mostclust** and classically used to find clusters of genes (or samples) with similar profiles are also offered. Supervised approaches, as Significance Analysis of Microarrays (**siggenes** package) approach or ANOVA functions, are proposed to identify differentially expressed genes (DEG) and functional enrichment of the DEG list is assessed based on **GOstat** package.

The package includes a vignette which describes the detailed biological/clinical analysis strategy used at Institut Curie. Most of the functions were improved for ease of use (fewer command lines, default parameters tested and chosen to be optimal). Relevant, enhanced and easy-to-interpret text and graphic outputs are offered. The package is available on The Comprehensive R Archive Network repository.

References

Bertoni, A. and G. Valentini (2007). Model order selection for bio-molecular data clustering. *BMC Bioinformatics 8 Suppl 2*, S7.

Falcon, S. and R. Gentleman (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics 23*, 257–258.

Le, S., J. Josse, and F. Husson (2008). Factominer: an R package for multivariate analysis. *Journal of statistical software 25*, 1–18.

Servant, N., G. Eleonore, P. Gestraud, C. Laurent, C. Paccard, A. Biton, I. Brito, J. Mandel, B. Asselain, E. Barillot, and P. Hupé (2010). Ema - a R package for easy microarray data analysis. *BMC Research Notes 3*, 277.

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A 98*, 5116–21.