GeneGroupAnalysisBetaCS: a package to perform gene-set significance analysis for data sets with cross-sectional or time series designs.

Alejandro Quiroz-Zárate[1] and John Quackenbush[1,2,3]

[1]Department of Biostatistics, Harvard University
[2]Computational Biology and Functional Genomics Laboratory, Dana-Farber Cancer Institute, Harvard School of Public Health
[3]Center for Cancer Computational Biology, Dana-Farber Cancer Institute

May 16, 2012

Contents

1 Introduction

The *GeneGroupAnalysis* package provides functions to perform statistical identification of gene functional classes that behave in a distinct manner between the phenotypes of interest for data sets under cross-sectional or time series designs. This package includes (i) functions to perform gene set comparison (ii) examples to visualize the results of such comparisons.

1.1 Installation

GeneGroupAnalysisBetaCS requires that *MCMCpack*, *AnnotationDbi*, *annotate*, *tcltk* and *R* ($>=$ 2.10.0) are installed. These should be installed automatically when you install *GeneGroupAnalysisBetaCS*. To install *GeneGroupAnalysis*, source biocLite from bioconductor:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("GeneGroupAnalysisBetaCS")
```

Load the *GeneGroupAnalysis*, into your current workspace:

```
> library(GeneGroupAnalysisBetaCS)
```

1.2 Further help

To view the *GeneGroupAnalysisBetaCS* description and a summary of all the functions within *GeneGroupAnalysisBetaCS*, type the following:

```
> library(help=GeneGroupAnalysisBetaCS)
```

1.3 Citing

We are delighted if you use this package. Please do email us if you find a bug or have a suggestion. We would be very grateful if you could cite:

Quiroz-Zarate A and Quackenbush J (2012). *Manuscript in preparation* .

2 Two simple cases: from gene expression to functional class selection.

We will very briefly demonstrate the use of some functions in *GeneGroupAnalysisBetaCS* by providing its application on two data sets with different experimental design.

We use the *breastCancerVDX* data library from Bioconductor for demonstration purposes under a cross-sectional design. This data set corresponds to the data set from [1].

Minn, AJ and colleagues used Affymetrix U133A Gene Chips to profile gene expression in 286 fresh-frozen tumor samples from patients with lymph-node-negative breast cancer who were treated during $1980 - 95$, but who did not receive systemic neoadjuvant or adjuvant therapy. These samples correspond from the data set used in [3] with GEO reference accession number GSE2034, from the tumor bank at the Erasmus Medical Center in Rotterdam, Netherlands. An additional 58 estrogen receptor-negative samples were added from [1] GEO (GSE5327). In total 209 tumor samples are classified as ER+ and 135 as ER-. Even though this data set comes from a 5-year follow-up design, the way the data is conceived for this analysis is cross-sectional. For purposes of providing an example on the use of the functions on the package the data set with cross-sectional design has a 70% of the samples with ER+ and with ER- where selected randomly selected.

2.1 Example: Data analysis under a cross-sectional setting.

This is an example on how to perform an analysis with the proposed method in [2] for a data set with cross-sectional design. This example is divided in two parts. The data preparation and the execution of the Gibb's sampler function.

2.1.1 Data preprocessing stage

The original gene expression data set Minn AJ and colleagues [1] has a U133A Affymetrix platform. The normalized data set was saved to the variable `vdx` in the *breastCancerVDX* data library from Bioconductor.

```
> library(breastCancerVDX)
> library(GeneGroupAnalysisBetaCS)
> library(hgu133a.db)
> library(annotate)
> data(vdx,package="breastCancerVDX")
> #-Normalized expression data set
> minn.data.expr=exprs(vdx)
> #--- Checking that the columns correspond to their respective phenotype data id
> #all(colnames(minn.data.expr)==rownames(pData(vdx)))
```

The data on `minn.data.expr` has the following appearance (maybe with a different column order):

	ER +			ER −		
	Sample 1	…	Sample 209	Sample 1	…	Sample 135
Probe ID 1	7.84	…	7.31	7.11	…	6.98
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Probe ID q	6.14	…	6.81	6.52	…	6.37

The rows correspond the the measurements for the Affymetrix identifiers and the columns correspond to the gene expression measurements on the samples.

3 Session Info

- R version 2.15.0 (2012-03-30), `x86_64-apple-darwin9.8.0`
- Locale: `C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: AnnotationDbi 1.18.0, Biobase 2.16.0, BiocGenerics 0.2.0, DBI 0.2-5, GeneGroupAnalysisBetaCS 1.0, MASS 7.3-18, MCMCpack 1.2-3, RSQLite 0.11.1, annotate 1.34.0, breastCancerVDX 1.0.3, coda 0.14-7, hgu133a.db 2.7.1, lattice 0.20-6, org.Hs.eg.db 2.7.1
- Loaded via a namespace (and not attached): IRanges 1.14.2, grid 2.15.0, stats4 2.15.0, tcltk 2.15.0, tools 2.15.0, xtable 1.7-0

References

[1] Minn AJ, Gupta GP, Padua D, Bos P, Nguyen DX, Nuyten D, Kreike B, Zhang Y, Wang Y, Ishwaran H, Foekens JA, Van de Vijver M and Massagué J: Lung Metastasis Genes Couple Breast Tumor Size and Metastatic Spread. *PNAS*, **104(16)**, 6740-6745. 2007.

[2] Quiroz-Zarate A and Quackenbush J XXXX: Genes as repeated measures of gene-set significance *Journal* **Vol(Num):Page 1-Page N**. 2011.

[3] Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yanh F, Talantov D, Timmermans M, Gelder, MEMG, Yu J, Jatkoe T, Berns EMJJ, Atkins D and Foekens JA: Gene-expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer. *Lancet*, **365**, 671-679. 2005.

[4] Wollbold J, Huber R, Pohlers D, Koczan D, Guthke R, Kinne RW and Gausmann U: Adapted Boolean Network Models for Extracellular Matrix Formation. *BMC Systems Biology*, **3(77)**, 2009.