pcaGoPromoter version 1.10.0

Morten Hansen

October 13, 2014

1 Introduction

This R package provides functions to ease the analysis of Affymetrix DNA micro arrays by principal component analysis with annotation by GO terms and possible transcription factors.

2 Requirements

R version 2.14.0 or higher

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("pcaGoPromoter",dependencies=TRUE)
```

Rgraphviz from Bioconductor is needed to draw Gene Ontology tree. Note: Graphviz needs to be installed on the computer for Rgraphviz to work. See Rgraphviz README for installation.

3 Example

3.1 Load the library

```
> library("pcaGoPromoter")
```

3.2 Read in data set serumStimulation

```
> library("serumStimulation")
> data(serumStimulation)
```

The serumStimulation data set has been created from 13 CEL files - 5 controls, 5 serum stimulated with inhibitor and 3 serum stimulated without inhibitor. They are read with ReadAffy(), normalized with rma() and the expression data extracted with exprs(). All of these function are part of the affy package.

The arrays are most likely grouped in some sort of way. Create a factor vector to indicate the groups:

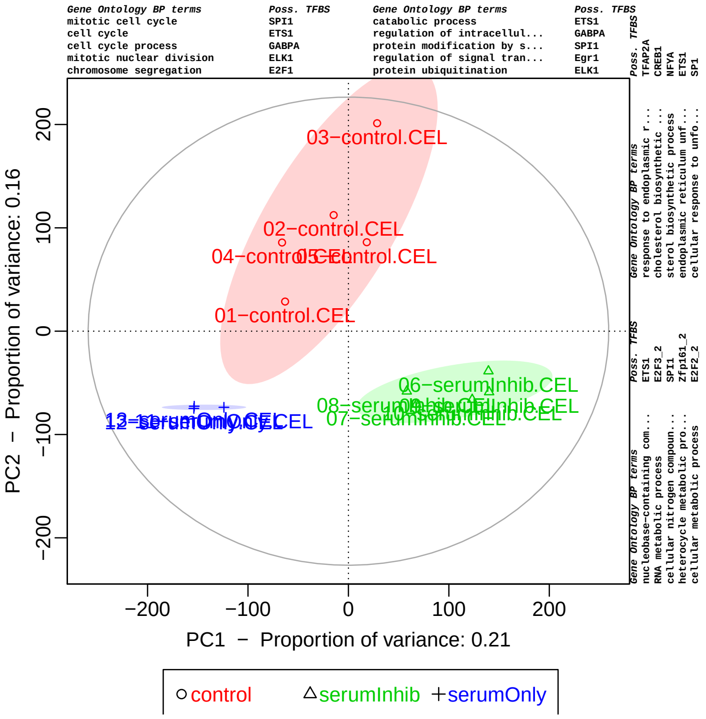
```
> groups <- as.factor( c( rep("control",5) , rep("serumInhib",5) ,
+                         rep("serumOnly",3) ) )
> groups
```

```
 [1] control    control    control    control    control    serumInhib
 [7] serumInhib serumInhib serumInhib serumInhib serumOnly  serumOnly 
[13] serumOnly 
Levels: control serumInhib serumOnly
```

3.3 Make PCA informative plot

This function "does-it-all". It will make a PCA plot and annotate the axis will GO terms and possible commmon transcription factors.

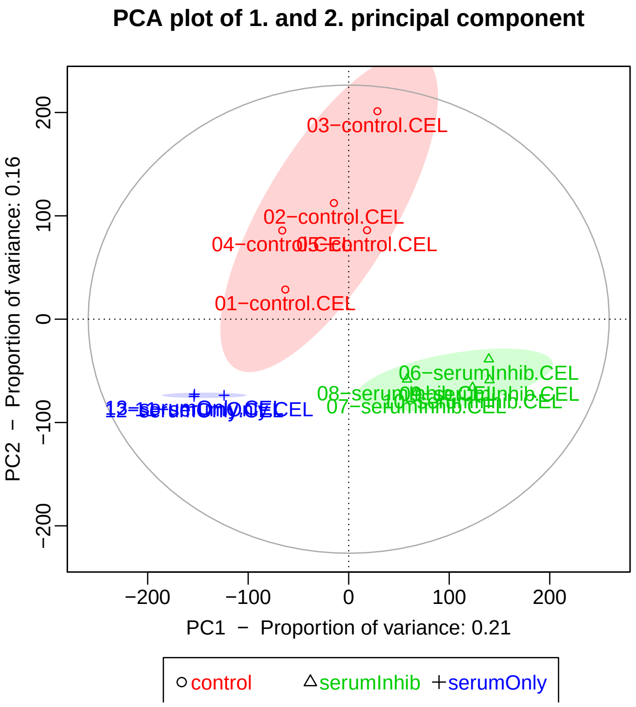
```
> pcaInfoPlot(serumStimulation,groups=groups)
```



3.4 Principal component analysis (PCA)

```
> pcaOutput <- pca(serumStimulation)
```

```
> plot(pcaOutput, groups=groups)
```



Proportion of variance is noted along the axis. In this case there are 3 groups in the data set - control, serumInhib and serumOnly. There is a clear separation of the groups along the 1. principal component (X-axis). The 2. principal component shown a difference between the controls and the serum stimulated.

3.5 Get loadings from PCA

We would like to have the first 1365 probe ids (2,5 %) from 2. principal component in the negative (serum stimulated) direction.

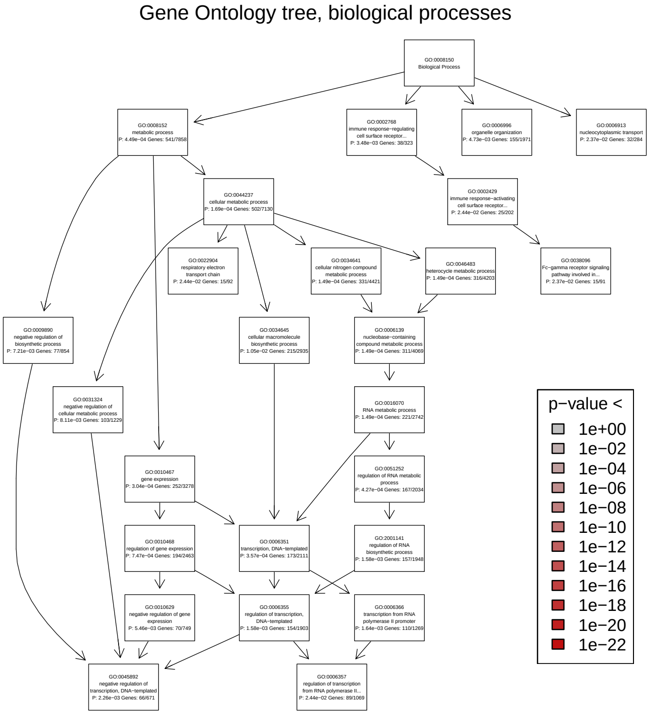
```
> loadsNegPC2 <- getRankedProbeIds( pcaOutput, pc=2, decreasing=FALSE )[1:1365]
```

3.6 Create Gene Ontology tree from loadings

Note: In this step you will be asked to install the necessary data packages.

```
> GOtreeOutput <- GOtree( input = loadsNegPC2)
```

```
> plot(GOtreeOutput,legendPosition = "bottomright")
```


Gene Ontology tree, biological processes
GO:0008150 Biological Process
GO:0008152 metabolic process P: 4.49e−04 Genes: 541/7858
GO:0002768 immune response−regulating cell surface receptor... P: 3.48e−03 Genes: 38/323
GO:0006996 organelle organization P: 4.73e−03 Genes: 155/1971
GO:0006913 nucleocytoplasmic transport P: 2.37e−02 Genes: 32/284
GO:0044237 cellular metabolic process P: 1.69e−04 Genes: 502/7130
GO:0002429 immune response−activating cell surface receptor... P: 2.44e−02 Genes: 25/202
GO:0022904 respiratory electron transport chain P: 2.44e−02 Genes: 15/92
GO:0034641 cellular nitrogen compound metabolic process P: 1.49e−04 Genes: 331/4421
GO:0046483 heterocycle metabolic process P: 1.49e−04 Genes: 316/4203
GO:0038096 Fc−gamma receptor signaling pathway involved in... P: 2.37e−02 Genes: 15/91
GO:0009890 negative regulation of biosynthetic process P: 7.21e−03 Genes: 77/854
GO:0034645 cellular macromolecule biosynthetic process P: 1.05e−02 Genes: 215/2935
GO:0006139 nucleobase−containing compound metabolic process P: 1.49e−04 Genes: 311/4069
GO:0031324 negative regulation of cellular metabolic process P: 8.11e−03 Genes: 103/1229
GO:0016070 RNA metabolic process P: 1.49e−04 Genes: 221/2742
GO:0010467 gene expression P: 3.04e−04 Genes: 252/3278
GO:0051252 regulation of RNA metabolic process P: 4.27e−04 Genes: 167/2034
GO:0010468 regulation of gene expression P: 7.47e−04 Genes: 194/2463
GO:0006351 transcription, DNA−templated P: 3.57e−04 Genes: 173/2111
GO:2001141 regulation of RNA biosynthetic process P: 1.58e−03 Genes: 157/1948
GO:0010629 negative regulation of gene expression P: 5.46e−03 Genes: 70/749
GO:0006355 regulation of transcription, DNA−templated P: 1.58e−03 Genes: 154/1903
GO:0006366 transcription from RNA polymerase II promoter P: 1.64e−03 Genes: 110/1269
GO:0045892 negative regulation of transcription, DNA−templated P: 2.26e−03 Genes: 66/671
GO:0006357 regulation of transcription from RNA polymerase II... P: 2.44e−02 Genes: 89/1069
p−value <
1e+00
1e−02
1e−04
1e−06
1e−08
1e−10
1e−12
1e−14
1e−16
1e−18
1e−20
1e−22

Output to PDF file is advised. This can be done by coping output to a PDF file:

```
> dev.copy2pdf(file="GOtree.pdf")
```

Function 'GOtree()' also outputs a list of GO terms order by p-value.

```
> head(GOtreeOutput$sigGOs,n=10)
```

```
           GOid genesInTerm totalGenesInTerm      pValue
287  GO:0006139         311             4069 0.000148921
1012 GO:0016070         221             2742 0.000148921
1597 GO:0034641         331             4421 0.000148921
2142 GO:0046483         316             4203 0.000148921
1955 GO:0044237         502             7130 0.000169008
862  GO:0010467         252             3278 0.000303639
333  GO:0006351         173             2111 0.000356736
2434 GO:0051252         167             2034 0.000426658
723  GO:0008152         541             7858 0.000449290
863  GO:0010468         194             2463 0.000747396
                                        GOterm
287  nucleobase-containing compound metabolic process
1012                         RNA metabolic process
1597     cellular nitrogen compound metabolic process
2142                 heterocycle metabolic process
1955                    cellular metabolic process
862                            gene expression
```

```
333                       transcription, DNA-templated
2434              regulation of RNA metabolic process
723                                  metabolic process
863                     regulation of gene expression
```

3.7 Get list of possible transcription factors

To get possible transcription factors, use function primo() function.

```
> TFtable <- primo( loadsNegPC2 )
> head(TFtable$overRepresented)
```

```
     id baseId pwmLength     gene      pValue
1  9326 MA0098         6     ETS1 2.30355e-08
2 10235 PB0113        17   E2F3_2 1.08742e-07
3  9308 MA0080         6     SPI1 3.92539e-05
4 10321 PB0199        14 Zfp161_2 7.41396e-05
5 10234 PB0112        17   E2F2_2 9.72520e-05
6 10132 PB0010        14   Egr1_1 1.08150e-04
```

The output shows you which possible transcription factors (genes) the supplied probes have in common.

3.8 Get a list of probe ids for a specific transcription factor

```
> probeIds <- primoHits( loadsNegPC2 , id = 9343 )
> head(probeIds)
```

```
[1] "NM_001121"    "NM_016824"    "NM_001114380" "NM_002209"    "NM_003342"
[6] "NM_006403"
```