Parametric Gene Set Enrichement Analysis

Karl Dykema

April 25, 2007

Laboratory of Computational Biology, Van Andel Research Institute

1 PGSEA basics

`PGSEA` analyzes gene expression data to determine miss-regulation of defined gene signatures or "molecular concepts". The gene signatures can be from a number of different sources but we have found the most informative ones to be experimentally derived lists of genes. (As opposed to canonical pathways...) There are many such lists that already exist in the literature as well as in publicly accessible repositories. To make use of these valuable resources, we have implemented a class to hold these gene lists called "smc", for "Simple Molecular Concepts." The class contains slots for holding many different types of information about the molecular concept, but all that is necessary is the "ids" slot. This is where the genes are stored. To begin we will show how to create your own basic concept:

```
> library(PGSEA)
> basic <- new("smc", ids = c("gene a", "gene b"), reference = "simple smc")
> str(basic)

Formal class 'smc' [package "PGSEA"] with 10 slots
  ..@ reference : chr "simple smc"
  ..@ desc      : chr(0)
  ..@ source    : chr(0)
  ..@ design    : chr(0)
  ..@ identifier: chr(0)
  ..@ species   : chr(0)
  ..@ data      : chr(0)
  ..@ private   : chr(0)
  ..@ creator   : chr(0)
  ..@ ids       : chr [1:2] "gene a" "gene b"
```

2 Concepts from ".gmt" files

A previous published method of gene set analysis termed "Gene Set Enrichment Analysis" http://www.broad.mit.edu/gsea/ also used lists of related genes and to hold their lists they use a file format called ".gmt". We have found it to be a simple and useful way to store our lists as well. To facilitate using this format in conjunction with our "smc" class, we have included the functions readGmt and writeGmt. An example of how to read in a ".gmt" file is shown below. The file provided contains some example molecular concepts. Genes induced and inhibited by Ras and Myc, as well as all genes on chromosome arms 5p and 5q have been included.

```
> datadir <- system.file("data", package = "PGSEA")
> sample <- readGmt(file.path(datadir, "sample.gmt"))
> str(sample[1])

List of 1
 $ ras UP - pmid: 16273092   NA :Formal class 'smc' [package "PGSEA"] with 10 slots
  .. ..@ reference : chr "ras UP - pmid: 16273092  "
  .. ..@ desc      : chr "NA "
  .. ..@ source    : chr(0)
  .. ..@ design    : chr(0)
  .. ..@ identifier: chr(0)
  .. ..@ species   : chr(0)
  .. ..@ data      : chr(0)
  .. ..@ private   : chr(0)
  .. ..@ creator   : chr(0)
  .. ..@ ids       : chr [1:181] "101" "154" "384" "490" ...
```

Below some example gene expression data from primary neuroblastoma tumors is analyzed by PGSEA using the provided example concepts. PGSEA is run with an index given to the appropriate reference samples, and the data for the [non-reference] samples is displayed with smcPlot.

```
> data(nbEset)
> pg <- PGSEA(nbEset@exprs, cl = sample, ref = 1:5)

[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6

> sub <- factor(c(rep(NA, 5), rep("NeuroB", 5), rep("NeuroB_MYC+",
+     5)))
> smcPlot(pg[, ], sub, scale = c(-12, 12), show.grid = T, margins = c(1,
+     1, 7, 13), col = .rwb)
```
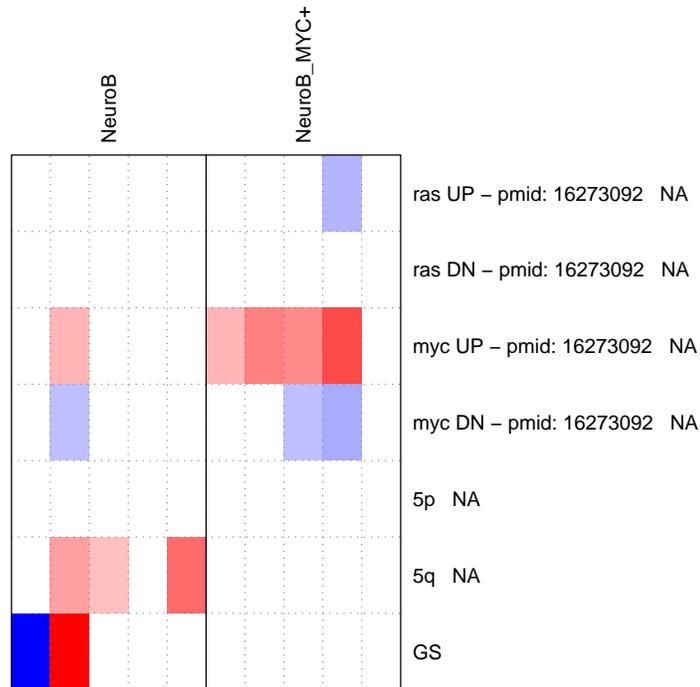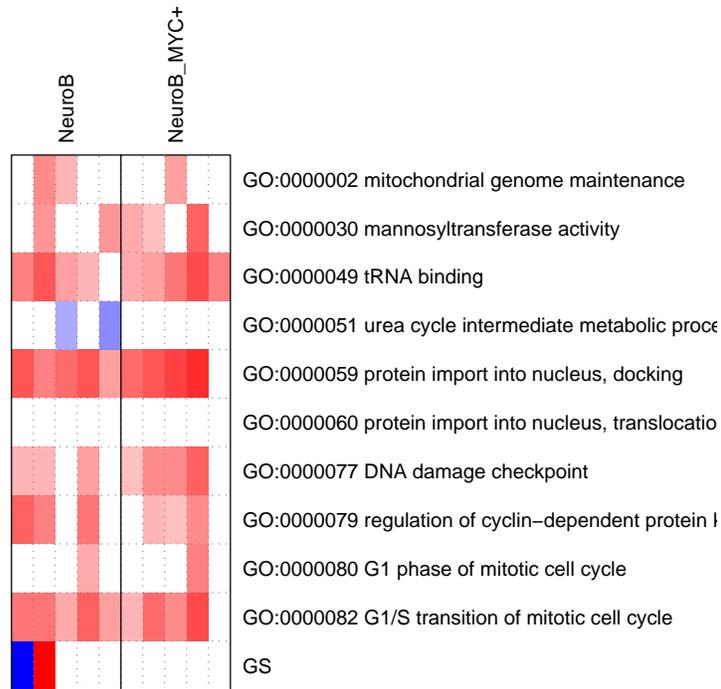
3 Using Gene Ontologies as concepts

Next, to create a list of concepts we will use the function `go2smc`. This function will convert the entire GO datbase into 'smc' objects. This is quite a large number to work with, so we will just use a few of them for illustration.

```
> mcs <- go2smc()[1:10]
> pg <- PGSEA(nbEset@exprs, cl = mcs, ref = 1:5)

[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
[1] 6
[1] 7
[1] 8
[1] 9
[1] 10

> smcPlot(pg[, ], sub, scale = c(-12, 12), show.grid = T, margins = c(1,
+     1, 7, 20), col = .rwb)
```

|	NeuroB	NeuroB_MYC+	
			GO:0000002 mitochondrial genome maintenance
			GO:0000030 mannosyltransferase activity
			GO:0000049 tRNA binding
			GO:0000051 urea cycle intermediate metabolic proce
			GO:0000059 protein import into nucleus, docking
			GO:0000060 protein import into nucleus, translocatio
			GO:0000077 DNA damage checkpoint
			GO:0000079 regulation of cyclin−dependent protein k
			GO:0000080 G1 phase of mitotic cell cycle
			GO:0000082 G1/S transition of mitotic cell cycle
			GS

4   Concepts compiled at VAI

Lastly, we have included a list of concepts that we have manually compiled from various sources. We have found these concepts to very informative in our analysis.

```
> data(VAImcs)
> pg <- PGSEA(nbEset@exprs, cl = VAImcs, ref = 1:5)

[1]  1
[1]  2
[1]  3
[1]  4
[1]  5
[1]  6
[1]  7
[1]  8
[1]  9
[1] 10
[1] 11
[1] 12
[1] 13
```

```
[1] 14
[1] 15
[1] 16
[1] 17
[1] 18
[1] 19
[1] 20
[1] 21
[1] 22
[1] 23
[1] 24
[1] 25
[1] 26
[1] 27
[1] 28
[1] 29
[1] 30
[1] 31
[1] 32
[1] 33
[1] 34
[1] 35
[1] 36
[1] 37
[1] 38
[1] 39

> smcPlot(pg[, ], sub, scale = c(-5, 5), show.grid = T, margins = c(1,
+     1, 8, 14), col = .rwb, r.cex = 0.7)
```
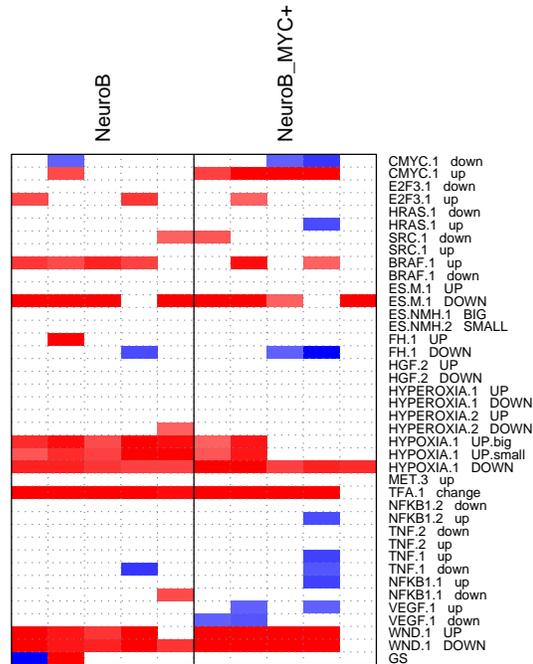
We included this particular example dataset because the tumors were tested for chromosomal aberrations which lead to amplification of important genes. The amplification we are attempting to illustrate is that of MYC. We have five samples that are confirmed to have the amplification and five samples that have standard copy numbers. Our concept of interest is "CMYC.1 up", which is the second row from the top in the above plot. It was provided by Bild et al. (PMID: 16273092). PGSEA reports four samples with confirmed MYC amplification have a positive score, while only one of five samples without the amplification has a positive score. These results may not be overly impressive, but in the interest of file size we could only include a portion of full the data set. Originally there were 101 samples, 81 of which had the amplification and 20 which did not. When PGSEA is run on the entire dataset, 18/20 (.90) of samples with confirmed amplification have show increased activity, and only 5/81 (.06) of samples confirmed to have no amplification also show increased activity.

These results are quite interesting and demonstrate the feasibility of using a technique such as PGSEA to quickly analyze a wide variety of aspects about a given data set. This type of analysis depends solely on informative concepts, so that is where we believe the efforts of the community should be directed.