GSCA: Gene Set Correlation Analysis

YounJeong Choi

June 8, 2009

1 Introduction

`LungCancer3` in the package includes matched data matrices from the three lung cancer microarray experiments, along with the gene set definitions (in Entrez Gene ID) and their descriptions.

Gene matching by Entrez Gene ID provides 3,924 common genes across the 3 studies of interest. Expression matrix rows are named in Entrez Gene IDs. Multiple transcripts representing one ID have been averaged at the log level before matching.

```
> library(GSCA)
> data(LungCancer3)
> str(LungCancer3$data)
```

```
List of 3
 $ Harvard : num [1:3924, 1:156] 8.6 5.68 8.11 4.34 9.22 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3924] "5595" "7075" "1843" "4319" ...
  .. ..$ : chr [1:156] "Tumor1" "Tumor2" "Tumor3" "Tumor4" ...
 $ Stanford: num [1:3924, 1:46] -0.108 2.474 0.615 -1.567 -0.338 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3924] "5595" "7075" "1843" "4319" ...
  .. ..$ : chr [1:46] "Tumor1" "Tumor2" "Tumor3" "Tumor4" ...
 $ Michigan: num [1:3924, 1:96] 10.27 10.07 10.62 7.61 10.82 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3924] "5595" "7075" "1843" "4319" ...
  .. ..$ : chr [1:96] "Tumor1" "Tumor2" "Tumor3" "Tumor4" ...
```

```
> str(LungCancer3$info$GSdef[1:3])
```

```
List of 3
 $ GO:0008150: Named chr [1:3631] "5595" "8850" "4137" "644" ...
  ..- attr(*, "names")= chr [1:3631] "Row1" "Row10" "Row100" "Row1000" ...
 $ GO:0008152: Named chr [1:2348] "5595" "8850" "4137" "644" ...
  ..- attr(*, "names")= chr [1:2348] "Row1" "Row10" "Row100" "Row1000" ...
 $ GO:0002526: Named chr [1:46] "2683" "4153" "717" "12" ...
  ..- attr(*, "names")= chr [1:46] "Row1052" "Row1097" "Row1193" "Row1291" ...
```

```
> str(LungCancer3$info$Name)

 Named chr [1:3649] "biological_process" "metabolic process" ...
 - attr(*, "names")= chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
```

The `info` portion in `LungCancer3` contains gene set definitions `GSdef` as a list of 3,649 gene sets. Each entry in the list is a character vector of Entrez Gene IDs that belong to the corresponding gene set. A particular set can be pulled out, for example, GO:0008033 as shown here.

```
> GS <- LungCancer3$info$GSdef
> GSdesc <- LungCancer3$info$Name
> setid <- "GO:0008033"
> GS[[setid]]

Row1413 Row1554 Row1891 Row2842 Row2872 Row3217
 "3735"  "6301"    "16"  "6741" "10940" "10556"

> GSdesc[setid]

       GO:0008033
"tRNA processing"
```

2 Single-study differential correlation

The function `singleDC` calculates a dispersion index between the tumor and normal groups for every gene set within a study. It also calculates permutation-based p-values based on the given number of sample permutations. For example, if the Michigan study is of interest, one can do the following. We try with 3 permutations to save time.

```
> dc.M <- singleDC(data = LungCancer3$data$Michigan, group = c(86,
+     10), GSdefList = GS, nperm = 3, permDI = TRUE)
> str(dc.M)

List of 3
 $ DI    : Named num [1:3649] 0.431 0.434 0.404 0.427 0.422 ...
  ..- attr(*, "names")= chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
 $ pvalue: Named num [1:3649] 0.667 0.667 0.333 0.667 0.667 ...
  ..- attr(*, "names")= chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
 $ permv : num [1:3649, 1:3] 0.431 0.434 0.404 0.427 0.422 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
  .. ..$ : chr [1:3] "P1" "P2" "P3"
```

3 Meta-analysis of multiple studies

The function `metaDI` calculates a dispersion index between two studies for every gene set. We choose to use the sign of correlation difference between tumor and normal groups from each study here, but the function itself can be used with the raw correlations as well. It also calculates permutation-based p-values based on the given number of permutations. The permutation here is annotated gene pair permutation. For example, if the Harvard and Michigan studies are of interest, one can do the following. We try with 3 permutations to save time.

```
> data.grouped <- list(Tumor = list(Harvard = LungCancer3$data$Harvard[,
+     1:139], Stanford = LungCancer3$data$Stanford[, 1:41], Michigan = LungCancer3$data$Mich
+     1:86]), Normal = list(Harvard = LungCancer3$data$Harvard[,
+     140:156], Stanford = LungCancer3$data$Stanford[, 42:46],
+     Michigan = LungCancer3$data$Michigan[, 87:96]))
> corr.t <- lapply(lapply(data.grouped$Tumor, t), cor, use = "pairwise.complete.obs")
> corr.n <- lapply(lapply(data.grouped$Normal, t), cor, use = "pairwise.complete.obs")
> cor.diff <- list(Harvard = corr.t$Harvard - corr.n$Harvard, Michigan = corr.t$Michigan -
+     corr.n$Michigan)
> cor.diff.sign <- list(Harvard = apply((cor.diff$Harvard > 0),
+     2, as.numeric) - apply((cor.diff$Harvard < 0), 2, as.numeric),
+     Michigan = apply((cor.diff$Michigan > 0), 2, as.numeric) -
+         apply((cor.diff$Michigan < 0), 2, as.numeric))
> for (i in 1:length(cor.diff.sign)) {
+     rownames(cor.diff.sign[[i]]) <- colnames(cor.diff.sign[[i]])
+ }
> dist.HM <- metaDI(cor.diff.sign$Harvard, cor.diff.sign$Michigan,
+     GS, 3, permDI = TRUE)
> str(dist.HM)

List of 3
 $ DI    : Named num [1:3649] 1.39 1.39 1.33 1.39 1.39 ...
  ..- attr(*, "names")= chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
 $ pvalue: Named num [1:3649] 0.333 1 0.333 0.333 0.333 ...
  ..- attr(*, "names")= chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
 $ permv : num [1:3649, 1:3] 1.39 1.39 1.33 1.39 1.39 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:3649] "GO:0008150" "GO:0008152" "GO:0002526" "GO:0006629" ...
  .. ..$ : chr [1:3] "P1" "P2" "P3"
```

4 Network display

4.1 Single-study network

`plotNW` is a wrapper function of `plot.graph` in `Rgraphviz` package with specific settings. Nodes represent genes and edges represent relationships between genes.

Edges are colored in from blue to red indicating correlation value of -1 to 1. This way we do not need to threshold correlation values for drawing edges. One can also choose to color nodes according to the corresponding genes' DE information. Arguments that are accepted by `plot.graph` can be passed. Figure 1 below is an example with the set GO:0019216 drawn for the Michigan normal group, with hypothetical DE (differential expression) information.

```
> setid <- "GO:0019216"
> gid <- GS[[setid]]
> ss <- c("SERPINA3", "SOD1", "SCAP", "NPC2", "ADIPOQ", "PRKAA1",
+     "AGT", "PPARA", "BMP6", "BRCA1")
> plotNW(genes = ss, cormatrix = corr.n$Michigan[gid, gid], node.de = c(rep(1,
+     5), rep(0, 5)), ncolor = "yellow", ecolor = bluered(201),
+     ewidth = 5, gtype = "circo")
```

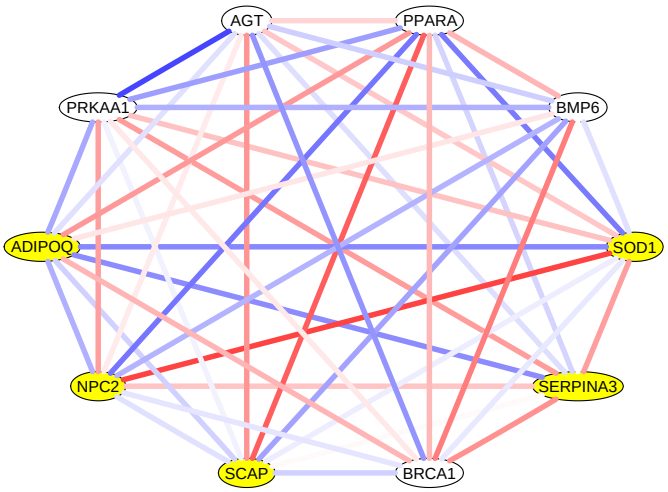

Figure 1: Network display of a gene set. Edge colors range from blue to red to indicate the correlation range of from -1 to 1. Differentially expressed gene nodes are colored in yellow.

4.2 Multi-edge network

The function `plotMNW` places nodes on a circle boundary and connects every pair of nodes with multiple edges corresponding to multiple studies of interest. Figure 2 below is an example of GO:0008033 with three studies.

```
> setid <- "GO:0008033"
> gid <- GS[[setid]]
> ss <- c("KARS", "SARS", "AARS", "SSB", "POP1", "RPP30")
> cormatrix.list <- list(corr.n$Harvard[gid, gid], corr.n$Stanford[gid,
+     gid], corr.n$Michigan[gid, gid])
> plotMNW(cormatrix.list, genes = ss, mycolors = bluered(201),
+     ncolor = "yellow", node.de = c(rep(1, 3), rep(0, 3)), lwd = 5,
+     jt = 0.3)
```

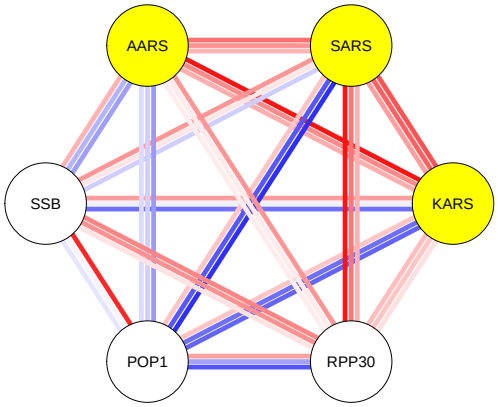

Figure 2: Multi-edge network display. Three lines per edge represent three study-specific correlations. Edges with similar color across studies indicate study agreement for the corresponding gene pair. Colors range from blue to red to indicate the correlation range of from -1 to 1.