Highlighting the Power of Gene Set Enrichment Analysis using Simulation

Jean Fan

December 23, 2015

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states. GSEA can be particularly powerful when genes individually do not exhibit a statistically significant difference between two biological states, but when grouped together, show statistically significant concordant differences.

Here, we explore such a scenario through simulation using the Lightweight Iterative Geneset Enrichment in R (`liger`) package. To install `liger`, you can use `devtools`:

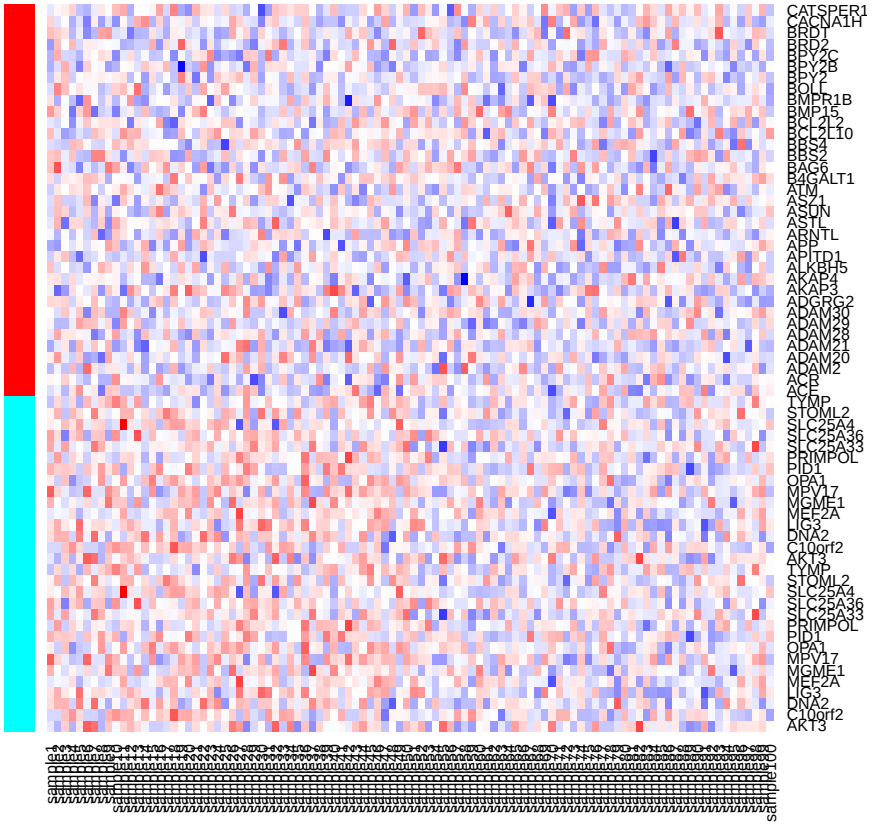
```
require(devtools)
devtools::install_github("JEFworks/liger")
```

To begin, we will simulate a weak differential expression within a known geneset between two biological samples.

```
set.seed(0)

library(liger)
# load gene set
data("org.Hs.GO2Symbol.list")
# get universe
universe <- unique(unlist(org.Hs.GO2Symbol.list))
# get genes in a geneset
gs <- org.Hs.GO2Symbol.list[[1]]
# geneset name
names(org.Hs.GO2Symbol.list)[1]
```

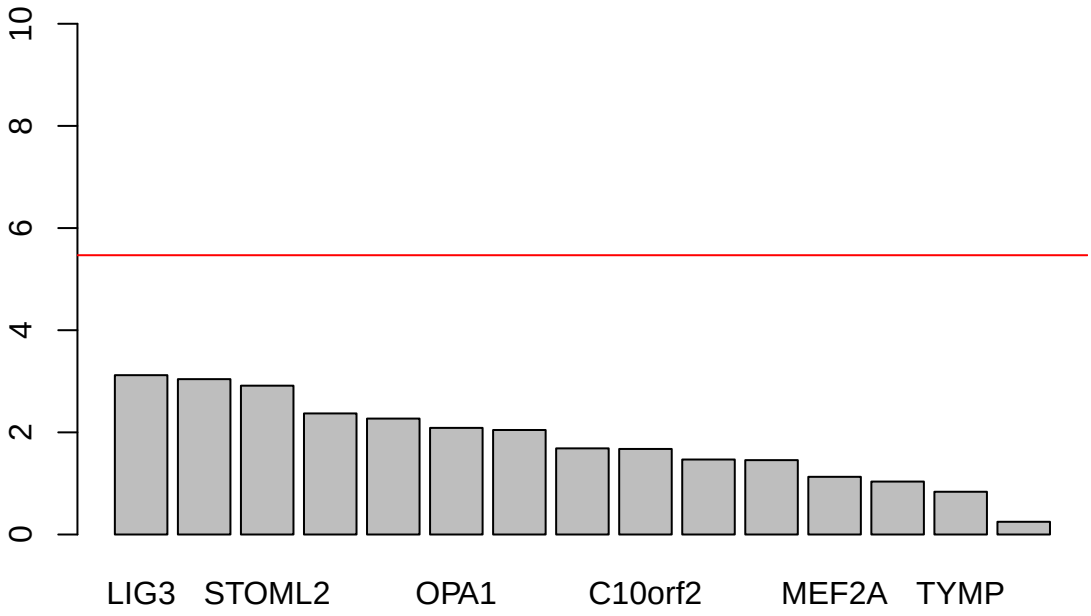
```
## [1] "GO:0000002"
```

```
# make random data
Nsamples <- 100
Mgenes <- length(universe)
mat <- matrix(rnorm(Nsamples * Mgenes, 5, 10), Mgenes, Nsamples)
colnames(mat) <- paste0('sample', 1:Nsamples)
rownames(mat) <- universe

# let half the samples be in one biological state and the other half a different one
# simulate differential expression for genes in the geneset
mat[gs, 1:round(Nsamples/2)] <- rnorm(length(gs)*round(Nsamples/2), 10, 10)

# we can visualize this weak differential expression visually in a heatmap
# visualize weakly differentially expressed genes and another 50 genes
vi <- c(gs, universe[1:50])
# label supposedly differentially expressed genes
heatmap(mat[vi,], Rowv=NA, Colv=NA, scale="none",
        col=colorRampPalette(c("blue", "white", "red"))(100),
        RowSideColors = rainbow(2)[as.factor(vi %in% gs)])
```



Even visually, it's somewhat difficult to tell which genes are supposedly differentially expressed. We can also quantify the extent of the differential expression between our two biological states using a T-test or other metrics for assessing the significance of differential expression.

```
# run differential expression analysis
vals <- sapply(1:nrow(mat), function(i) {
  pv <- t.test(mat[i, 1:round(Nsamples/2)], mat[i, round(Nsamples/2+1):Nsamples])$p.val
  pv
})
names(vals) <- rownames(mat)
vals <- -log10(vals)
# look at -log10(p values) for genes that are supposedly differentially expressed
barplot(sort(vals[gs], decreasing=TRUE), ylim=c(0, 10))
# multiple testing correction line
bonf <- function(a, n) { 1 - (1-a) ** (1/n) }
abline(h = -log10(bonf(0.05, nrow(mat))), col="red")
```



After multiple testing correction, none of the genes, including those we simulated to be differentially expressed, were actually picked up as significantly differentially expressed. In a real world situation, we may be tempted to end our analysis here and conclude that nothing is significantly differentially expressed between the two biological states and thus there is no significant difference.

However, we can perform GSEA, for example, on 100 different a priori defined gene sets to look for statistically significant concordant differences.

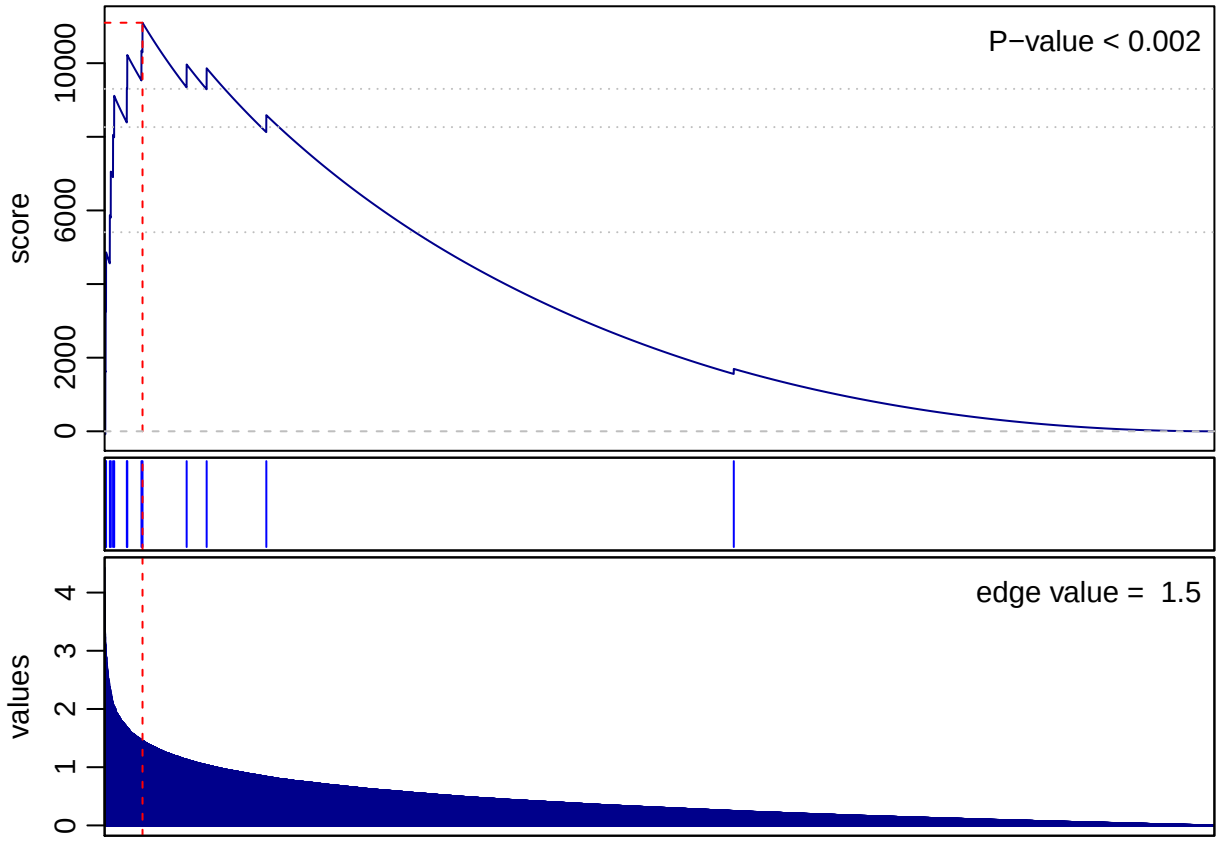
```
# run iterative bulk gsea
gseaVals <- iterative.bulk.gsea(values = vals, set.list = org.Hs.GO2Symbol.list[1:100], n.rand=500)
```

```
## initial: [5e+02 - 10] done
```

```
# identify significant genesets
gseaSig <- rownames(gseaVals[gseaVals$q.val < 0.05 & gseaVals$sscore > 0,])
gseaSig
```

```
## [1] "GO:0000002"
```

```
# look at plots
for(i in seq_along(gseaSig)) {
  gs <- org.Hs.GO2Symbol.list[[gseaSig[i]]]
  gsea(values=vals, geneset=gs, mc.cores=1, plot=TRUE, n.rand=500)
}
```



Despite no individual gene being statistically significantly differentially expressed between our two biological states, GSEA identifies a significantly enriched geneset, `GO:0000002`, which is exactly the geneset that we simulated to show concordant differences. Therefore, by pooling genes within these a priori defined genesets, we are able to increase our statistical power to identify differences between our two biological states.

R Session Info

```
sessionInfo()
```

```
## R version 3.3.3 (2017-03-06)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: macOS  10.14
##
## locale:
## [1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] liger_1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.14       matrixStats_0.52.2 digest_0.6.13
##  [4] rprojroot_1.2      backports_1.1.1    magrittr_1.5
##  [7] evaluate_0.10.1    stringi_1.1.6      rmarkdown_1.10
```

```
## [10] tools_3.3.3        stringr_1.2.0      tinytex_0.8
## [13] xfun_0.3           yaml_2.1.16        parallel_3.3.3
## [16] htmltools_0.3.6    knitr_1.20
```