Comparison of `liger` with `fgsea`

Jean Fan

June 26, 2018

`liger` is just one of many methods for gene set enrichment analysis. `fgsea` is another similar method that uses a faster cumulative statistic calculation on preranked values. Here, we compare `liger` to `fgsea`.

Comparison

We will use the example data and example gene sets that come with the `fgsea` package. Note that `exampleRanks` is a numeric vector where each value is fold-change or differential expression z-score between two biological conditions (some type of metric used for ranking genes) and `examplePathways` is a list of lists where each entry is a gene name corresponding to `exampleRanks`.

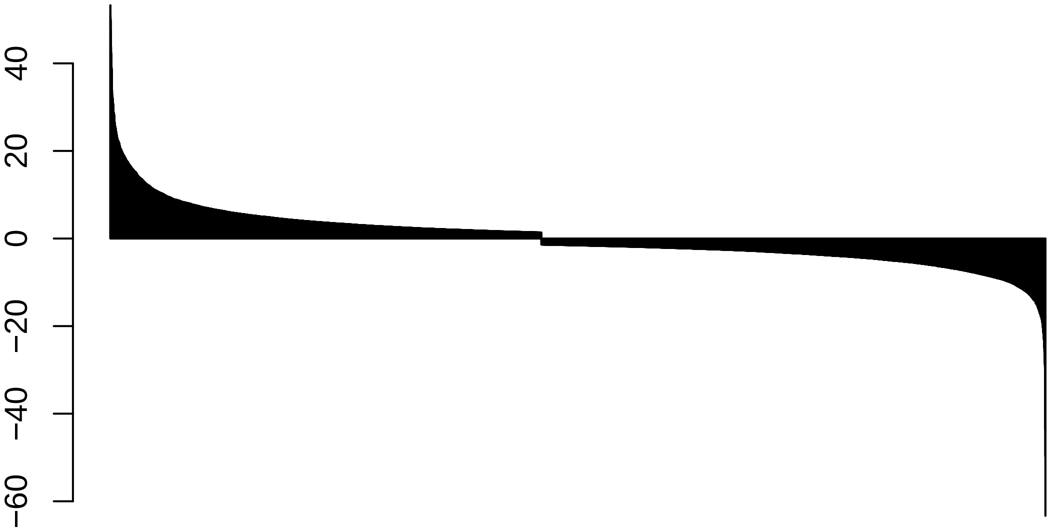
```r
# example pathways from fgsea
library(fgsea)
data(examplePathways)
data(exampleRanks)

head(examplePathways)
```

```
## $`1221633_Meiotic_Synapsis`
##  [1] "12189"     "13006"     "15077"     "15078"     "15270"    
##  [6] "15512"     "16905"     "16906"     "19357"     "20842"    
## [11] "20843"     "20957"     "20962"     "21749"     "21750"    
## [16] "22196"     "23856"     "24061"     "28113"     "50878"    
## [21] "56739"     "57321"     "64009"     "66654"     "69386"    
## [26] "71846"     "74075"     "77053"     "94244"     "97114"    
## [31] "97122"     "97908"     "101185"    "140557"    "223697"   
## [36] "260423"    "319148"    "319149"    "319150"    "319151"   
## [41] "319152"    "319153"    "319154"    "319155"    "319156"   
## [46] "319157"    "319158"    "319159"    "319160"    "319161"   
## [51] "319565"    "320332"    "320558"    "326619"    "326620"   
## [56] "360198"    "497652"    "544973"    "625328"    "667250"   
## [61] "100041230" "102641229" "102641751" "102642045"
## 
## $`1368092_Rora_activates_gene_expression`
## [1] "11865"  "12753"  "12894"  "18143"  "19017"  "19883"  "20787"  "217166"
## [9] "328572"
## 
## $`1368110_Bmal1:Clock,Npas2_activates_circadian_gene_expression`
##  [1] "11865"  "11998"  "12753"  "12952"  "12953"  "13170"  "14068" 
##  [8] "18143"  "18626"  "18627"  "19013"  "19883"  "20893"  "59027" 
## [15] "79362"  "217166"
## 
## $`1445146_Translocation_of_Glut4_to_the_Plasma_Membrane`
##  [1] "11461"     "11465"     "11651"     "11652"     "12313"    
##  [6] "12314"     "12315"     "16568"     "16569"     "16579"    
## [11] "17274"     "17884"     "17886"     "17913"     "17918"    
## [16] "19079"     "19082"     "19325"     "19341"     "20336"    
## [21] "20528"     "20619"     "20909"     "20912"     "22318"
## [26] "22627"     "22628"     "22629"     "22630"     "22631"
## [31] "53413"     "54401"     "55948"     "56044"     "57915"
## [36] "66482"     "68328"     "68365"     "68938"     "69940"
## [41] "102058"    "105504"    "107371"    "108079"    "108097"
## [46] "108099"    "210789"    "211446"    "240028"    "241113"
## [51] "241694"    "100039786" "102634437" "102641200" "102641764"
##
## $`186574_Endocrine-committed_Ngn3+_progenitor_cells`
## [1] "18012" "18088" "18506" "53626"
##
## $`186589_Late_stage_branching_morphogenesis_pancreatic_bud_precursor_cells`
## [1] "11925"  "15205"  "21410"  "246086"
```

```r
head(exampleRanks)
```

```
##    170942    109711     18124     12775     72148     16010
## -63.33703 -49.74779 -43.63878 -41.51889 -33.26039 -32.77626
```

```r
barplot(sort(exampleRanks, decreasing=TRUE), names.arg='')
```



We will test all gene sets of a particular size.

```r
# filter pathways to certain size
size <- lapply(examplePathways, length)
vi <- size > 15 & size < 500
table(vi)
```

```
## vi
## FALSE  TRUE
##   702   755
```

```r
examplePathways <- examplePathways[vi]
```

Now, we run both methods.

```r
# run fgsea
start_time <- Sys.time()
fgseaRes <- fgsea(pathways = examplePathways,
                  stats = exampleRanks,
                   nperm=10000)
end_time <- Sys.time()
print(end_time - start_time)
```

```
## Time difference of 1.549495 secs
```

```
# run liger
library(liger)
start_time <- Sys.time()
ligerRes <- iterative.bulk.gsea(exampleRanks,
                     set.list=examplePathways,
                     n.rand = c(100, 1000, 10000)
                     )
```

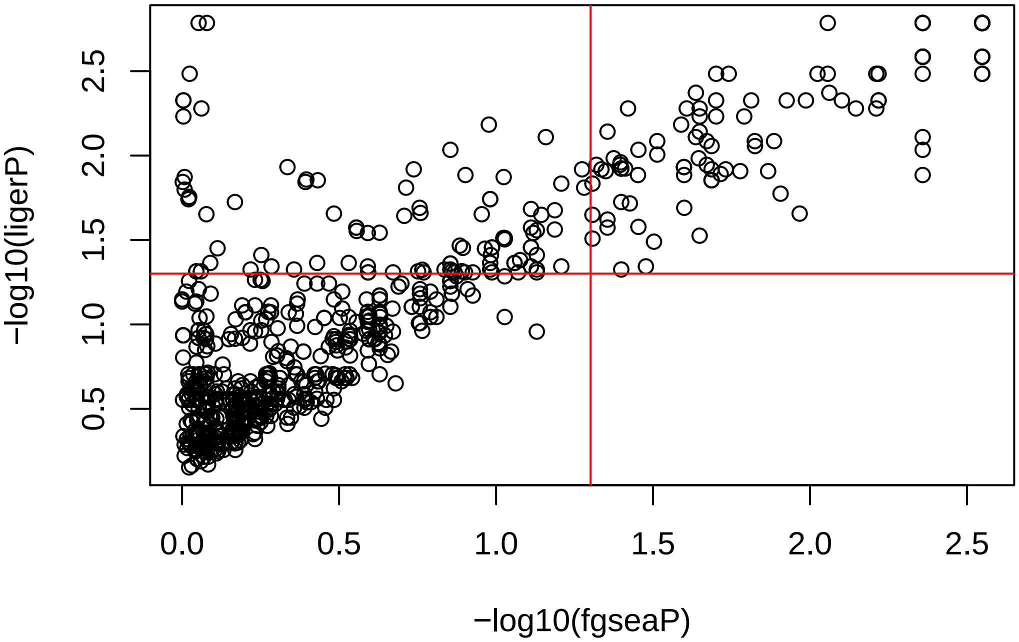
```
## initial: [1e+02 - 448] [1e+03 - 212] [1e+04 - 101] done
```

```
end_time <- Sys.time()
print(end_time - start_time)
```

```
## Time difference of 3.838567 mins
```

We can plot the -log10(corrected p-values) for both approaches and assess their correspondence.

```
# compare
fgseaP <- fgseaRes$padj; names(fgseaP) <- fgseaRes$pathway
fgseaP <- fgseaP[names(examplePathways)]
ligerP <- ligerRes$q.val; names(ligerP) <- rownames(ligerRes)
ligerP <- ligerP[names(examplePathways)]

# plot
par(mfrow=c(1,1), mar=rep(5,4))
plot(-log10(fgseaP), -log10(ligerP))
abline(v = -log10(0.05), col='red')
abline(h = -log10(0.05), col='red')
```



Each dot here is a gene set. The x position is the -log10(p-value) of the gene set from `fgsea` while the y-axis is from `liger`. While there does appear to be a good general correspondence (strong diagonal), notice a set of gene sets that are very significant in `liger` but not in `fgsea`. Let's take a closer look at what are these gene sets.

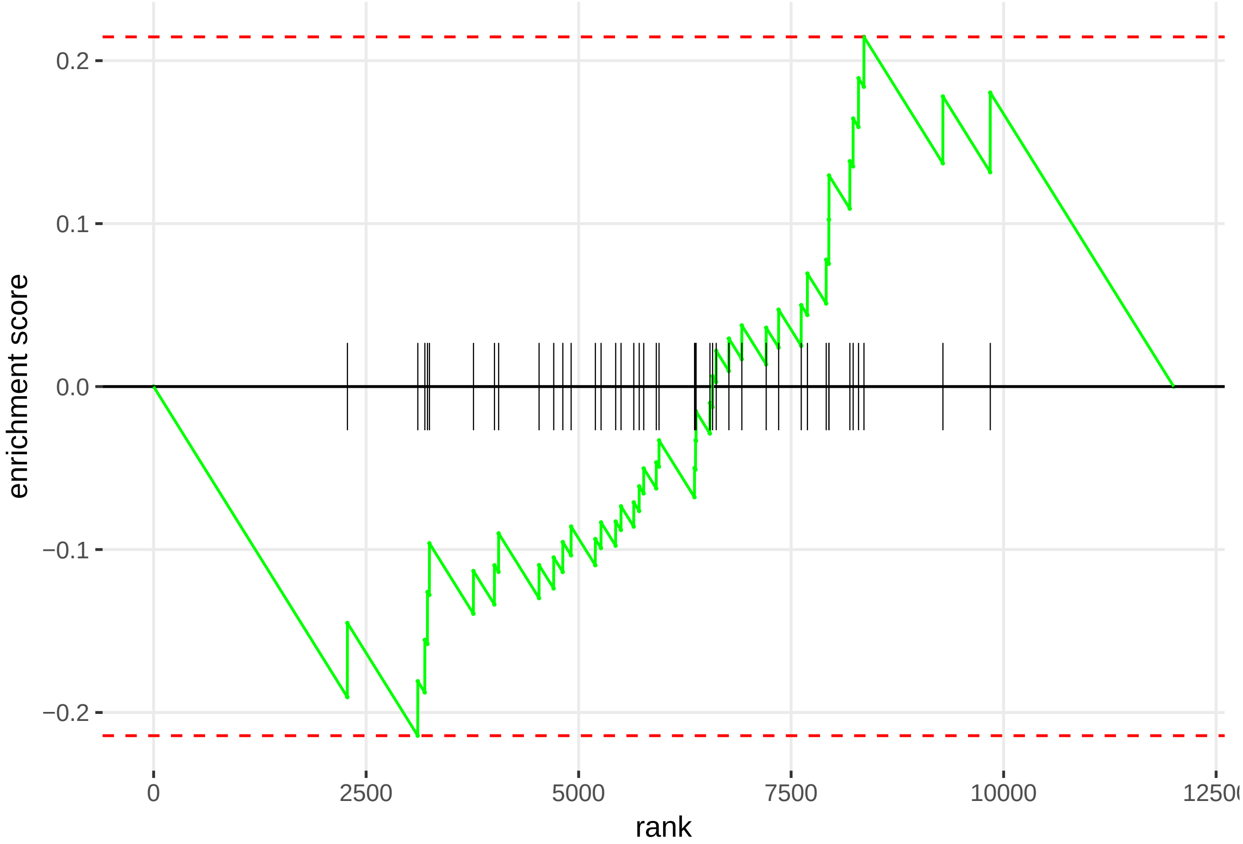
```
# maximal difference (most inconsistent) between methods
diff <- abs(-log10(fgseaP) - -log10(ligerP))
diff <- sort(diff, decreasing=TRUE)
# pick most inconsistent
gs <- names(diff)[1]
print(fgseaP[gs])
```

```
## 5991461_Peptide_chain_elongation
##                         0.8859868
```

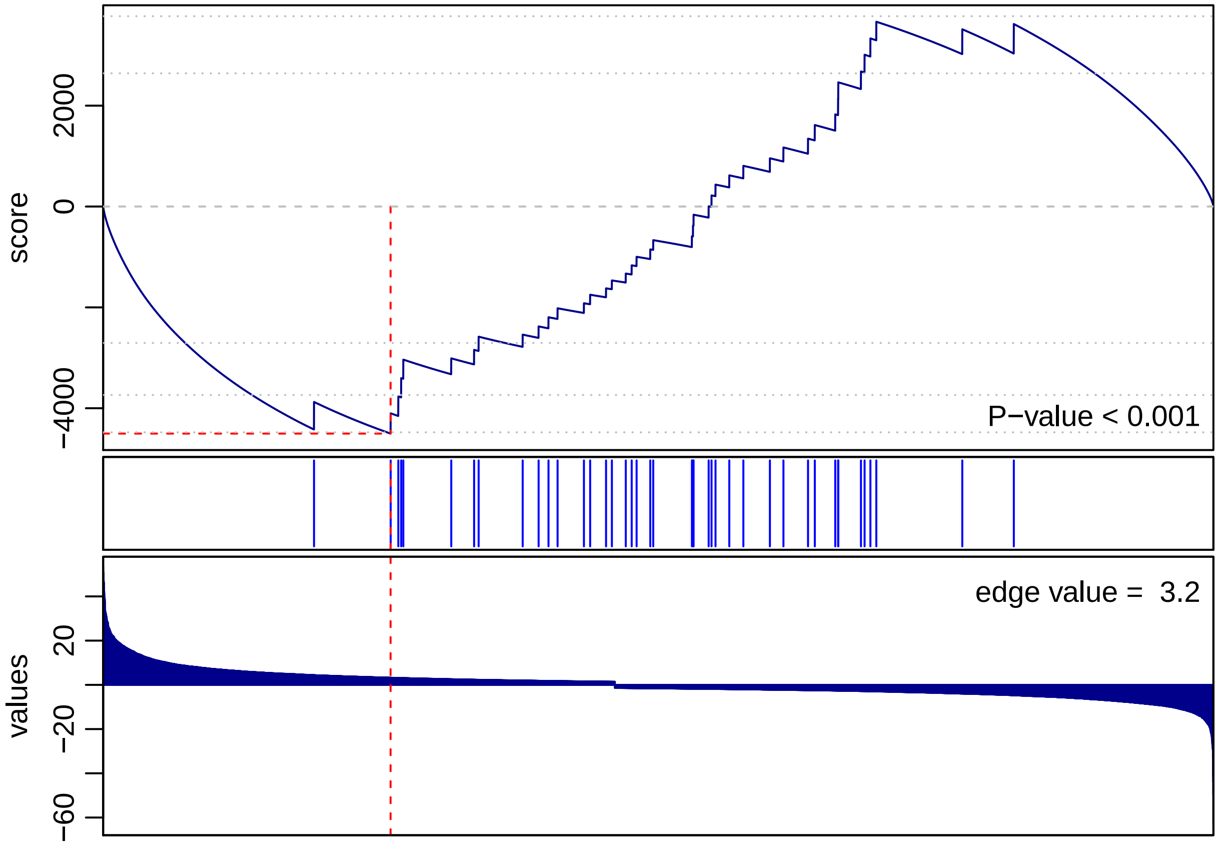
```
print(ligerP[gs])
```

```
## 5991461_Peptide_chain_elongation
##                       0.001638967
```

```
# fgsea
plotEnrichment(examplePathways[[gs]], exampleRanks)
```



```
# liger
gsea(exampleRanks, examplePathways[[gs]])
```

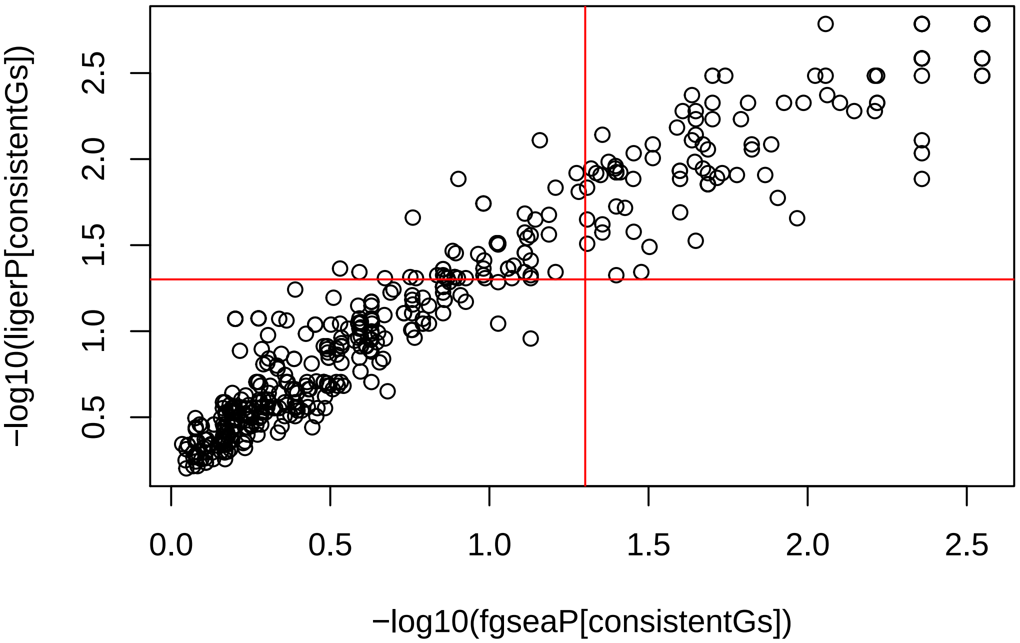


```
## [1] 0.001
```

What we can see is that `liger` detected a significant lack of genes in this gene set among the most highly ranked genes as noted by the positive `edge` but negative `sscore`. This particular type of enrichment testing may be important if we want to make claims about certain gene sets never being highly differentially expressed (depleted in representation) but are not necessarily down-regulated.

As `fgsea` does not detect such patterns, to make our comparison between the two methods more appropriate, we will restrict to gene sets for which `liger` detects a consistent `sscore` and `edge` (ie. both positive suggesting upregulation or both negative suggesting downregulation).

```
# make comparable
vi <- ligerRes$sscore * ligerRes$edge > 0
consistentGs <- rownames(ligerRes)[vi]
par(mfrow=c(1,1), mar=rep(5,4))
plot(-log10(fgseaP[consistentGs]), -log10(ligerP[consistentGs]))
abline(v = -log10(0.05), col='red')
abline(h = -log10(0.05), col='red')
```



Now, results are highly consistent between the two approaches.

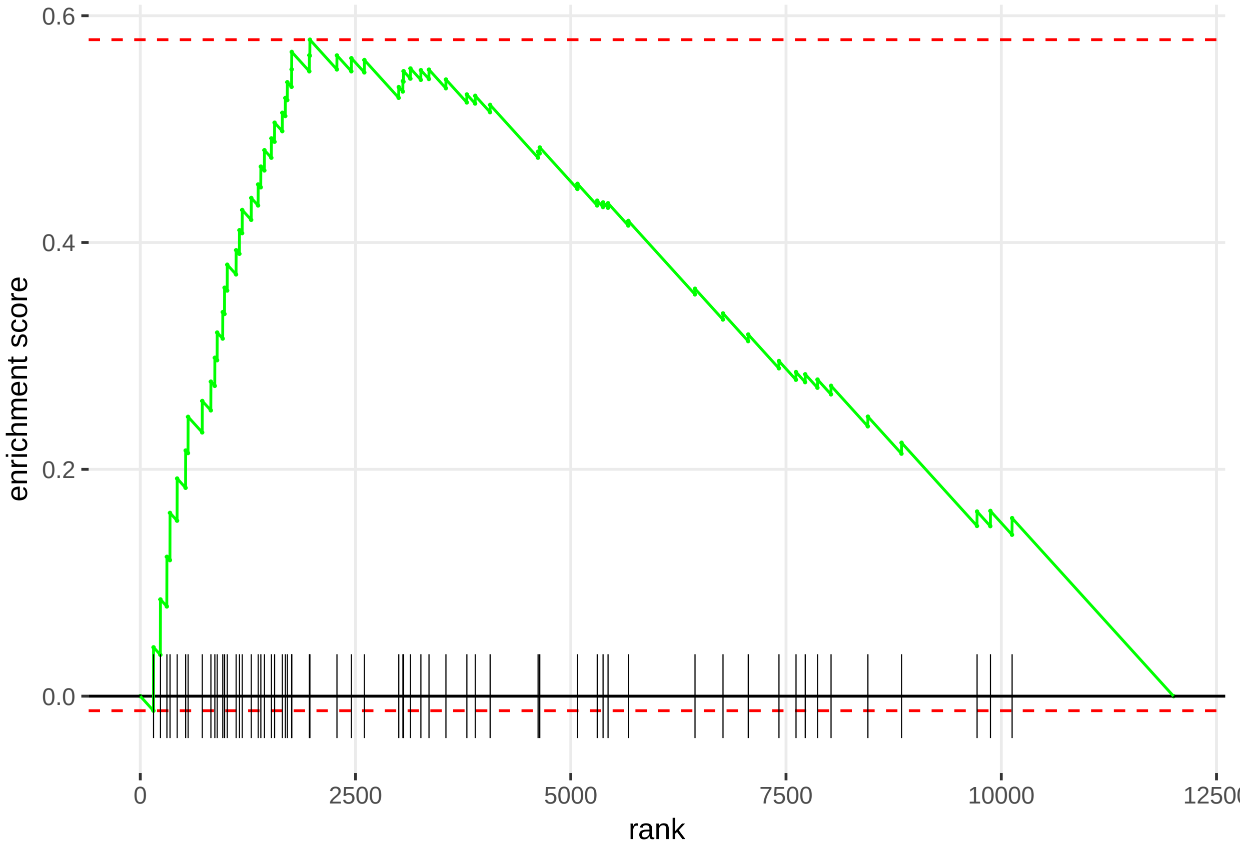
```
# pick a significant gene set
gs <- names(which(fgseaP[consistentGs]==min(fgseaP[consistentGs])))[1]
print(fgseaP[gs])
```

```
## 5990977_DNA_Replication_Pre-Initiation
##                            0.002829135
```

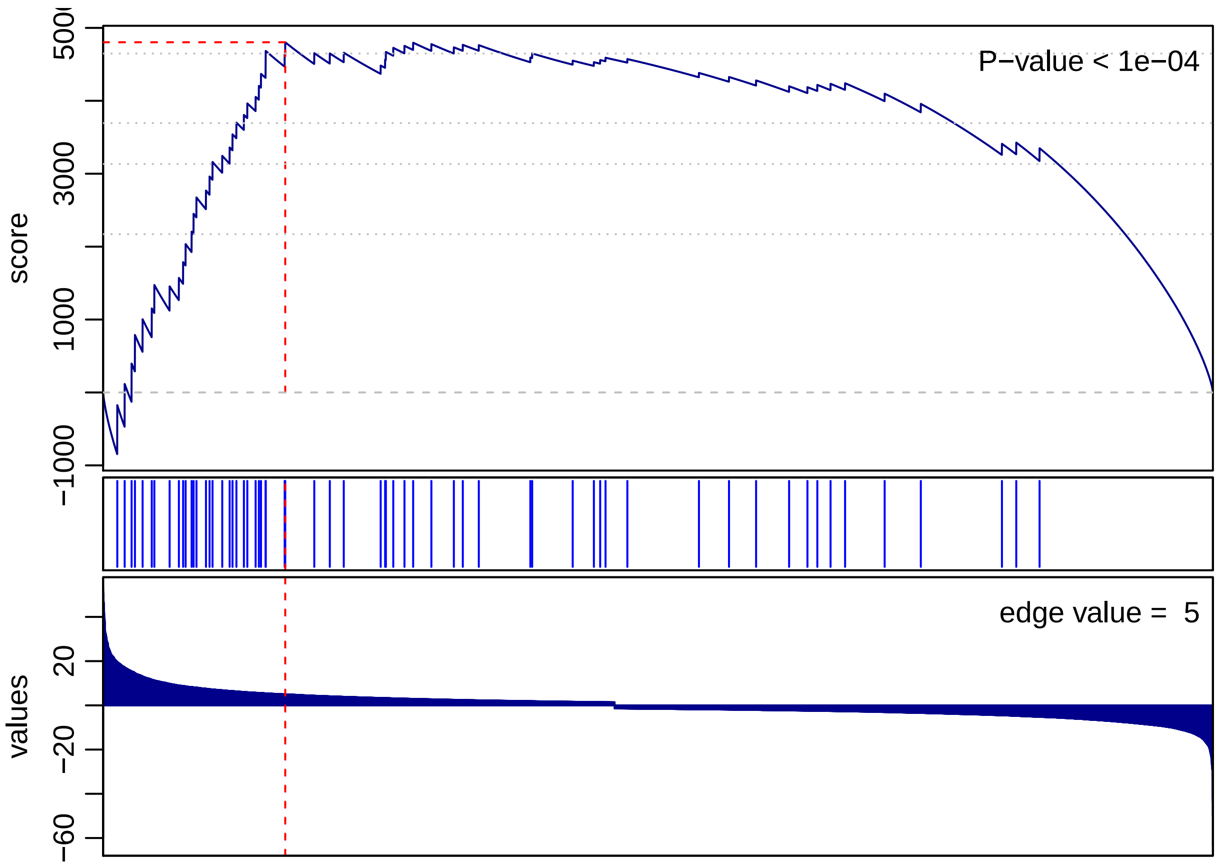
```
print(ligerP[gs])
```

```
## 5990977_DNA_Replication_Pre-Initiation
##                            0.001638967
```

```
# fgsea
plotEnrichment(examplePathways[[gs]], exampleRanks)
```



```
# liger
gsea(exampleRanks, examplePathways[[gs]])
```



```
## [1] 1e-04
```

Conclusion

In conclusion, `fgsea` provides a very fast test for gene sets where ranked gene values are appropriate. Both `fgsea` and `liger` offer very comparable results when looking for significantly upregulated or downregulated gene sets.

Ultimately, the appropriateness of gene set enrichment analysis approaches will depend on your question of interest. If you are only looking to test simply for over-representation of a set of genes, perhaps a hypergeometric test will be sufficient. If you care about the magnitude of the gene expression fold-change used in your gene ranking, a purely rank-based approach may be less optimal. If you are only interested in consistent upregulation and downregulation, significant results pointing to a depletion in representation among highly ranked genes may not be useful and should be ignored.

What ever gene set enrichment analysis you choose and whatever hypotheses they may help you generate, given the multitude of issues associated with gene sets, their accuracy, particularly as they pertain to your biological system of study, additional biological validation is always encouraged.