GO Enrichment

Run the `render()` function below and everything will be run with report at end.

```
library(rmarkdown)
render("skeleton_GO.Rmd", "pdf_document", output_file = paste(sample1,"_",sample2,"_","GO.pdf",sep=""))
```

Read in YAML guide

```
library(yaml)
yamls <- yaml.load_file("de.yml")
sample1 <- yamls$sample1
sample2 <- yamls$sample2

sample1
sample2
```

```
library(goseq)
library(GO.db)
```

Setting up the DE table for GO analysis

File Input

Input the output from DE analysis. This is made for a list that includes only the significant genes.

```
sigOnly <- read.table(paste(sample1,"_",sample2,"_DE_sig.txt", sep=""), header = TRUE, fill = TRUE)
head(sigOnly)
```

```
##                  ITAG  logFC logCPM    PValue       FDR
## 1 Solyc00g013180.1.1 -2.175  7.902 3.617e-05 1.439e-03
## 2 Solyc00g014830.2.1 -1.629  7.932 1.889e-03 3.594e-02
## 3 Solyc00g187050.2.1 -4.840  6.410 3.044e-17 1.493e-14
## 4 Solyc00g277510.1.1 -2.270  5.616 3.667e-05 1.448e-03
## 5 Solyc00g282510.1.1 -2.378  2.655 2.092e-03 3.883e-02
## 6 Solyc01g005270.2.1 -1.653  5.860 2.755e-03 4.840e-02
##
## 1              NADH-ubiquinone oxidoreductase chain 4 (AHRD V1 ***- Q5M9Z8_TOBAC); contains Interpro
## 2                            NADH-quinone oxidoreductase subunit D (AHRD V1 ***- C4YU74_9RICK); conta
## 3                                       Leucyl aminopeptidase (AHRD V1 ***- D7DWG6_NOSA0); contains I
## 4                                     Photosystem Q(B) protein (AHRD V1 ***- Q95B48_PRUPE); contains
## 5                                          Phenylalanine ammonia-lyase (AHRD V1 ***- B5LAW0_CAPAN); c
## 6 SEC14 cytosolic factor family protein (AHRD V1 **-- D7M8H6_ARALY); contains Interpro domain(s)  IPR
##         AGI symbol
## 1 ATMG00580   NAD4
## 2 ATMG00510   NAD7
## 3 AT4G30920   <NA>
## 4 ATCG00020   PSBA
## 5 AT2G37040 ATPAL1
## 6 AT1G14820   <NA>
##
## 1
## 2
## 3 cytosol aminopeptidase family protein; Identical to Leucine aminopeptidase 3, chloroplast precurso
## 4
## 5
## 6
##   X..identity alignment.length e.value bit.score percent.query.align
## 1       92.91              141   7e-70       259               79.10
## 2       94.67              394   0e+00       778               99.49
## 3       66.39              488  4e-177       617               95.09
## 4       91.04               67   4e-23       102               97.01
## 5       80.55              581   0e+00      1013               97.97
## 6       66.24              237   1e-93       339               97.51
```

```r
dim(sigOnly)
```

```
## [1] 412  14
```

```r
colnames(sigOnly)
```

```
##  [1] "ITAG"                "logFC"               "logCPM"
##  [4] "PValue"              "FDR"                 "SGN_annotation"
##  [7] "AGI"                 "symbol"              "gene_name"
## [10] "X..identity"         "alignment.length"    "e.value"
## [13] "bit.score"           "percent.query.align"
```

```r
colnames(sigOnly)[1] <- "itag"
```

Subset

First I need to subset the list to up or down regulated, then add a new colum that specififys 1. This column is need to for merging.

```r
upITAG <- subset(sigOnly, logFC > 0, select = c(itag))
upITAG$up <- 1

downITAG <- subset(sigOnly, logFC < 0, select = c(itag))
downITAG$down <- 1

allITAG <- subset(sigOnly, select = c(itag))
allITAG$all <- 1
```

Merge I - with normalized ITAG length gene list

read in guide.

```r
geneLength <- read.csv("../normalized_genes_length.csv")
head(geneLength)
```

```
##                 itag length
## 1 Solyc00g005040.2.1    357
## 2 Solyc00g005050.2.1    588
## 3 Solyc00g005060.1.1    273
## 4 Solyc00g005070.1.1     81
## 5 Solyc00g005080.1.1    297
## 6 Solyc00g005150.1.1   1143
```

```r
#isolate just the gene list
genes <- subset(geneLength, select = c(itag))
```

First merge each table to geneLength

```r
upITAGmerge <- merge(genes, upITAG, by = "itag", all= TRUE)
downITAGmerge <- merge(genes, downITAG, by = "itag", all= TRUE)
allITAGmerge <- merge(genes, allITAG, by = "itag", all= TRUE)
```

Merge II - Merge them all together.

```r
matrixGOupdown <- merge(upITAGmerge, downITAGmerge, by = "itag", all = TRUE)
matrixGOupdownall <- merge(matrixGOupdown, allITAG, by = "itag", all = TRUE)
matrixGO <- merge(matrixGOupdownall, geneLength, by = "itag", all = TRUE)
```

Clean Up

```r
matrixGO[is.na(matrixGO)] <- 0
head(matrixGO)
```

```
##                 itag up down all length
## 1 Solyc00g005040.2.1  0    0   0    357
## 2 Solyc00g005050.2.1  0    0   0    588
## 3 Solyc00g005060.1.1  0    0   0    273
## 4 Solyc00g005070.1.1  0    0   0     81
## 5 Solyc00g005080.1.1  0    0   0    297
## 6 Solyc00g005150.1.1  0    0   0   1143
```

This is if you want to write out the table of the GO matrix. #write.table(matrixGO, “mydata.txt”, sep=“͡”, quote= FALSE)

GO enrichment

The is the input of the GOslim categories. There are only two columns 1. itag and 2. go

```r
pat <- matrixGO
head(pat)
```

```
##                  itag up down all length
## 1 Solyc00g005040.2.1  0    0   0    357
## 2 Solyc00g005050.2.1  0    0   0    588
## 3 Solyc00g005060.1.1  0    0   0    273
## 4 Solyc00g005070.1.1  0    0   0     81
## 5 Solyc00g005080.1.1  0    0   0    297
## 6 Solyc00g005150.1.1  0    0   0   1143
```

```r
cate <- read.table("../melted.GOTable.txt",header=TRUE)
head(cate)
```

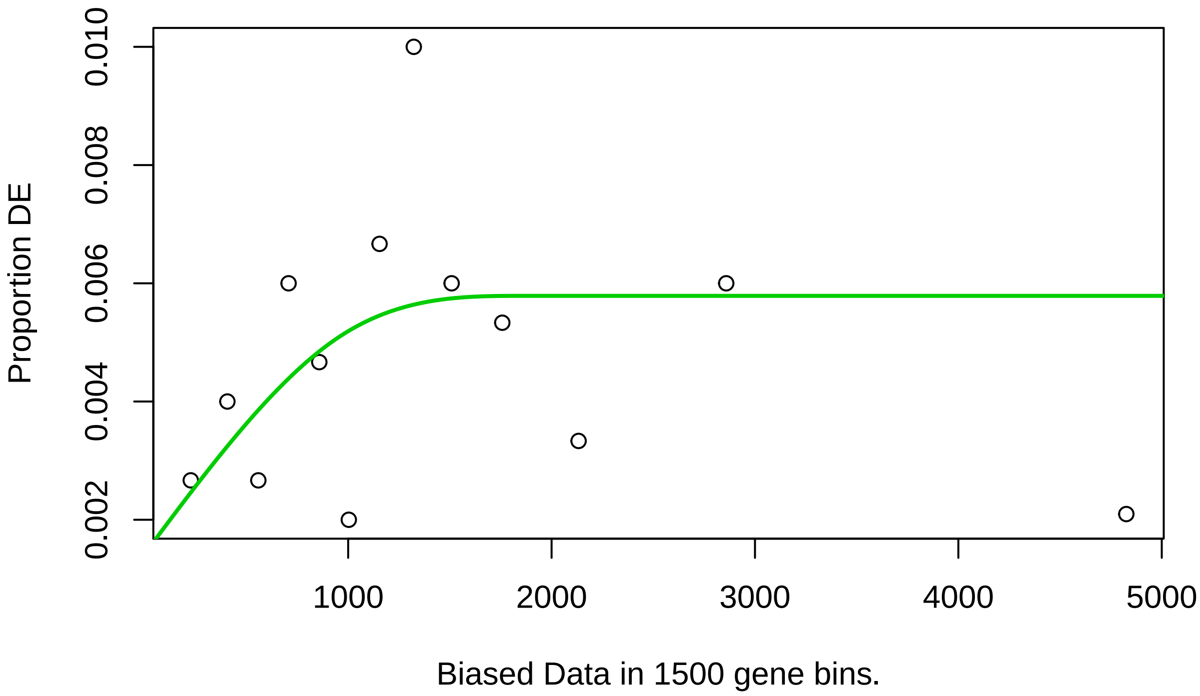
```
##                  itag         go
## 1 Solyc00g005000.2.1 GO:0006508
## 2 Solyc00g005040.2.1 GO:0005774
## 3 Solyc00g005050.2.1 GO:0005829
## 4 Solyc00g005080.1.1 GO:0005524
## 5 Solyc00g005130.1.1 GO:0006508
## 6 Solyc00g005150.1.1 GO:0003676
```

Subseting for GO analysis

Specify the column you are interested in `pat$all` refers to all the DE gene regardless if they are up or down regulated. If you want to specify down regulated, specify `pat$down`. I am going to put this into a loop, where each time the loop goes thought it will perform GO enrichment on all three types of lists of significant genes and them write them to a table.

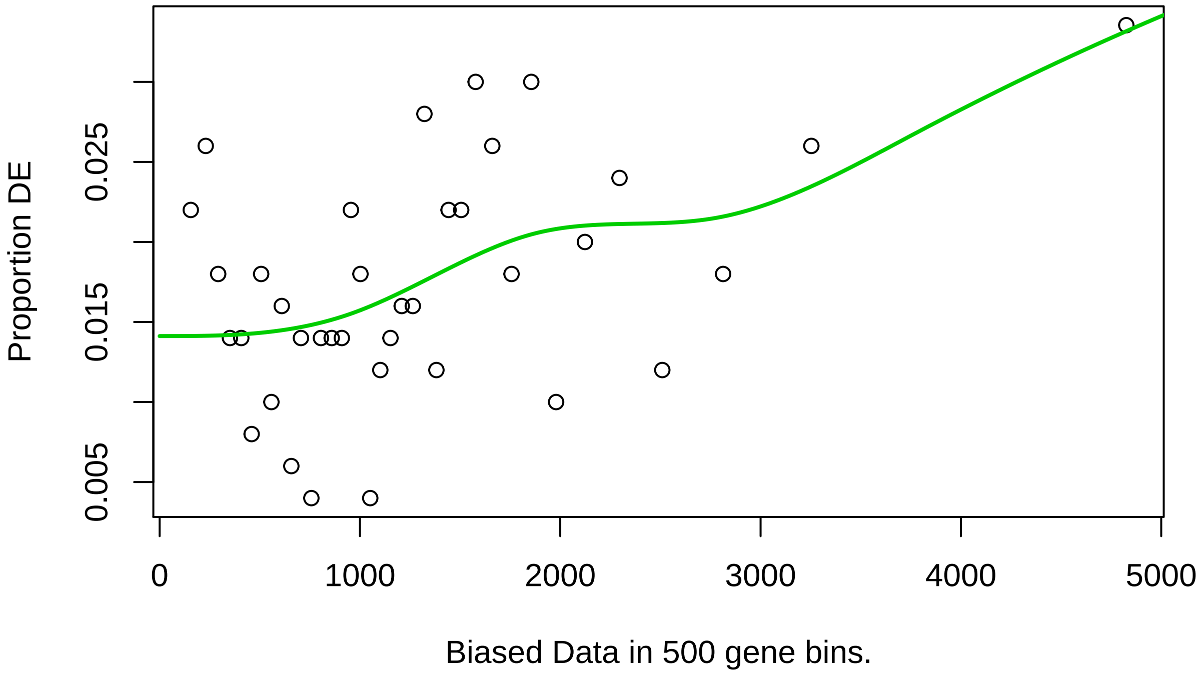
```r
sigType <- c("up", "down", "all")

for(type in sigType) {

genes = as.integer(pat[,type])
names(genes) = pat$itag
table(genes)
length(genes)

pwf = nullp(genes,bias.data=pat$length)

GO.wall = goseq(pwf,gene2cat = cate)
head(GO.wall)

#This is going to correct for multiple testing.  You can specify the p-value cut-off of GO categories y

enriched.GO = GO.wall$category[p.adjust(GO.wall$over_represented_pvalue, method = "BH") < 0.05]

enriched.GO

my.GO <- as.character(enriched.GO)
my.GO.table <- Term(my.GO)
my.GO.table
t <- as.matrix(my.GO.table)

print(type) #this is for the knitr document
print(t) #this is for the knitr document

write.table(t, file=paste(sample1,"_",sample2,"DE1_sigonly_",type,"_GO.txt", sep=""))
}
```

```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=T (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```



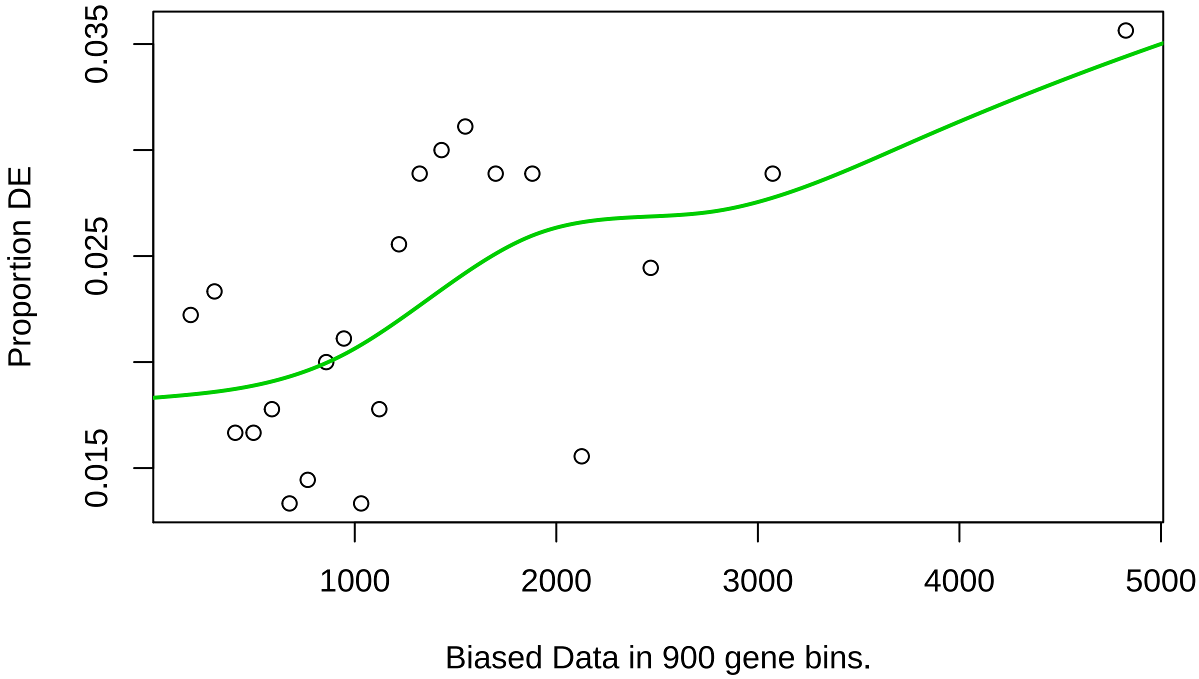
```
## [1] "up"
##      [,1]
```

```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=T (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```



```
## [1] "down"
##            [,1]
## GO:0016841 "ammonia-lyase activity"
## GO:0016168 "chlorophyll binding"
## GO:0009772 "photosynthetic electron transport in photosystem II"
## GO:0004397 "histidine ammonia-lyase activity"
## GO:0045548 "phenylalanine ammonia-lyase activity"
## GO:0006559 "L-phenylalanine catabolic process"
## GO:0030076 "light-harvesting complex"
## GO:0018298 "protein-chromophore linkage"
## GO:0009523 "photosystem II"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0009698 "phenylpropanoid metabolic process"
## GO:0045156 "electron transporter, transferring electrons within the cyclic electron transport pathwa
## GO:0030077 "plasma membrane light-harvesting complex"
## GO:0015979 "photosynthesis"
```

```
## Using manually entered categories.
## For 2936 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=T (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
```



```
## [1] "all"
##            [,1]
## GO:0016841 "ammonia-lyase activity"
## GO:0030076 "light-harvesting complex"
## GO:0016168 "chlorophyll binding"
## GO:0004397 "histidine ammonia-lyase activity"
## GO:0045548 "phenylalanine ammonia-lyase activity"
## GO:0009523 "photosystem II"
## GO:0018298 "protein-chromophore linkage"
## GO:0009772 "photosynthetic electron transport in photosystem II"
## GO:0006559 "L-phenylalanine catabolic process"
## GO:0009698 "phenylpropanoid metabolic process"
## GO:0000287 "magnesium ion binding"
## GO:0009535 "chloroplast thylakoid membrane"
## GO:0006351 "transcription, DNA-templated"
```