GOGOgadget Package

Xiaoping Li

January 17, 2017

GOGOgadget Package Manual

Introduction

GOGOgadget contains 5 basic functions to assist RNA seq data **Ontology** analysis. At moment, it best suits the data from *Deinococcus radiodurans*. Four functions are: DBconstruct, GOGOpick, GOGOscavenger, GOGOsummary, GOGOsearch. The GO term databases include PANTHER sequence classfication and UniProt. Details please see PANTHER; UniProt

workflow

1. DBconstruct() : this function constructs the basic database for GO terms reference. It returns 1) a tidy database with various kind of identifier for convenient access. 2) a general annotation from NCBI database. 3) GO types and terms from PANTHER 4) provide other information for future manipulation and extraction.

DBconstruct(ftp, species, short, write = F, ...)

ftp: the ftp file name for the species in PANTHER. For Drad, it's "PTHR11.0_deinococcus". species: the whole latin name for the species. E.g. "Deinococcus radiodurans" short: short name for the organism. E.g. "dra" write: if false, not to produce a csv file for the database information in your current directory. ...: if write = T, specify the name for the csv file to be saved.

2. GOGOpick(): this function extracts GO information from the DB database according to your edgeR or DESeq2 processed data sets. It picks genes that are significantly differetially expressed(FDR or padj < 0.05, logFC > 1 or $<$ -1). It also adds GO terms, GOtypes, identifiers and other information to those genes. It generate 3 csv files with 1) up regulated genes, 2) down regulated genes, 3) combining Up and Down genes into your current directory([dra]upReg.csv, [dra]downReg.csv, [dra]UpDown.csv). These 3 files do not contain GO terms. It also generates a csv file containing GO terms for all the up and down genes if write = T, and specify the file name in "...". If **not match** were found in the database, it returns NA or "No Match" for corresponding information columns.

GOGOpick(dataset, database, species, short, write = TRUE, ...)

dataset: edgeR or DESeq2 processed dataset database: DBconstruct() processed species: the whole latin name of your organism short: the short name for your organism write: if true: write the GOterm added data.frame into a csv file, file name is defined by "..."; if FALSE, just create a data.frame for downstream process.

3. GOGOscavenger(): This function compares the list of up and down genes(UpDown.csv) to the data set(e.g. pick.csv) produced by GOGOpick. It looks for the genes in UpDown.csv but not in the pick.csv. The reason probably those genes do not have GO terms in PANTHER. GOGOscavenger() will reference UniProt for GO terms for those genes and add those genes to the pick.csv list.

GOGOscavenger(whole, pick, taxiID, write = F, ...)

whole: UpDown.csv pick: pick.csv taxiID: e.g. 243230 (dra) write: if true, it produces a csv file with up and down genes annotated by GO terms in PANTHER and UniProt; if FALSE, it returns a data frame.

4. GOGOsummary(): This funciton summarize the dataset GOGOscavenger process. It present the GO term proportions in GO types(MF, BP, CC, PC) and in up/down genes as a form of pie chart.

GOGOsummary(scavenger, gotype)

scavenger: GOGOscavenger() processed data set gotype: MF, BP, CC, PC

5. GOGOsearch(): This function helps you narrow down and returns the genes you are interested.

GOGOsearch(scav, help = T, view = c("full", "search"), goterm, gotype = "PC", ...)

scav: GOGOscavenger() processed data set help: if TRUE, provides all the GO terms in the input dataset, it is convenient if you don't know what terms to look for. view: "full" shows you in each GO type, how many GO terms are up regulated or down regulated; if "search", won't show the all previously mentioned information. goterm: specify what GO term you look for. gotype: MF, BP, CC, PC

Example

Load package

```
library(GOGOgadget)
```

```
## Warning: replacing previous import 'KEGGREST::listDatabases' by
## 'biomartr::listDatabases' when loading 'GOGOgadget'
```

```
## Warning: replacing previous import 'UniProt.ws::select' by 'dplyr::select'
## when loading 'GOGOgadget'
```

1. Step 1: Load database using DBconstruct()

```
database <- DBconstruct("PTHR11.0_deinococcus", "Deinococcus radiodurans", "dra", write = F)
```

```
## [1] "The proteome of 'Deinococcus_radiodurans' has been downloaded to '_ncbi_downloads/proteome' and
```

```
head(database)
```

```
##      refID  geneID proteinID            Family_name
## 1 DR_B0043 1799808 NP_051583       FAMILY NOT NAMED
## 2 DR_B0043 1799808 NP_051583       FAMILY NOT NAMED
## 3 DR_B0043 1799808 NP_051583       FAMILY NOT NAMED
## 4 DR_B0043 1799808 NP_051583       FAMILY NOT NAMED
## 5  DR_0502 1800258 NP_294225 ALANYL-TRNA SYNTHETASE
## 6  DR_0502 1800258 NP_294225 ALANYL-TRNA SYNTHETASE
##                     Sub_family_name      goIDs
## 1              UPF0226 PROTEIN YHHS          0
## 2              UPF0226 PROTEIN YHHS          0
## 3              UPF0226 PROTEIN YHHS          0
## 4              UPF0226 PROTEIN YHHS          0
## 5 ALANINE--TRNA LIGASE, CYTOPLASMIC GO:0016874
## 6 ALANINE--TRNA LIGASE, CYTOPLASMIC GO:0003676
##                                                                annotation
## 1 integral membrane protein LmrP (plasmid) [Deinococcus radiodurans R1]
## 2 integral membrane protein LmrP (plasmid) [Deinococcus radiodurans R1]
## 3 integral membrane protein LmrP (plasmid) [Deinococcus radiodurans R1]
## 4 integral membrane protein LmrP (plasmid) [Deinococcus radiodurans R1]
```

```
## 5      alanyl-tRNA synthetase-like protein [Deinococcus radiodurans R1]
## 6      alanyl-tRNA synthetase-like protein [Deinococcus radiodurans R1]
##   goType              goTerms Pathway
## 1     MF                    0    <NA>
## 2     BP                    0    <NA>
## 3     CC                    0    <NA>
## 4     PC                    0    <NA>
## 5     MF      ligase activity    <NA>
## 6     MF nucleic acid binding    <NA>
```

2. Step2: Load edgeR or DESeq2 processed data set and using GOGOpick() to pick significantly differentially expressed genes and add in GO terms, types and other information.

```
dataset <- "[V2]drad_24_diff_edgeR.csv"
pick <- GOGOpick(dataset, database, "Deinococcus radiodurans", "dra", write = T, "pick.csv")
```

```
## [1] "The proteome of 'Deinococcus_radiodurans' has been downloaded to '_ncbi_downloads/proteome' and
```

```
head(pick)
```

```
##         refID  geneID proteinID      Family_name     Sub_family_name
## 1392 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1393 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1394 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1395 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 3038  DR_2281 1797858 NP_296002 RIBONUCLEASE III      RIBONUCLEASE 3
## 3039  DR_2281 1797858 NP_296002 RIBONUCLEASE III      RIBONUCLEASE 3
##           goIDs     logFC
## 1392          0 -5.855846
## 1393          0 -5.855846
## 1394          0 -5.855846
## 1395          0 -5.855846
## 3038 GO:0016788 -1.385570
## 3039 GO:0003676 -1.385570
##                                                      annotation goType
## 1392 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     MF
## 1393 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     BP
## 1394 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     CC
## 1395 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     PC
## 3038  hypothetical protein DR_2281 [Deinococcus radiodurans R1]     MF
## 3039  hypothetical protein DR_2281 [Deinococcus radiodurans R1]     MF
##                                        goTerms Pathway DE
## 1392                                         0    <NA> -1
## 1393                                         0    <NA> -1
## 1394                                         0    <NA> -1
## 1395                                         0    <NA> -1
## 3038 hydrolase activity, acting on ester bonds    <NA> -1
## 3039                      nucleic acid binding    <NA> -1
```

3.Step3: Use GOGOscavenger() to find genes that do not have GO hits in PANTHER and reroute to UniProt to look for GO terms.243230 is the taxonomy ID for *Deinococcus radiodurans*.

```
UpDown <- read.csv("[dra]UpDown.csv", stringsAsFactors = F)

head(UpDown)
```

```
##   X    refID  geneID proteinID     logFC
```

```
## 1 1  DR_0972 1799034 NP_294696  1.154176
## 2 2  DR_1500 1800389 NP_295223  1.729625
## 3 3   DR_t13 1799677  No match -2.133357
## 4 4 DR_A0209 1797985 NP_285532 -1.702373
## 5 5 DR_A0212 1797972 NP_296362 -2.425909
## 6 6 DR_A0210 1797981 NP_296361 -1.382044
##                                                                                annotation
## 1                                 hypothetical protein DR_0972 [Deinococcus radiodurans R1]
## 2                               NADH dehydrogenase I subunit F [Deinococcus radiodurans R1]
## 3                                                                                    <NA>
## 4                               peptide ABC transporter permease [Deinococcus radiodurans R1]
## 5                                 hypothetical protein DR_A0212 [Deinococcus radiodurans R1]
## 6 peptide ABC transporter, periplasmic peptide-binding protein [Deinococcus radiodurans R1]
##   DE
## 1  1
## 2  1
## 3 -1
## 4 -1
## 5 -1
## 6 -1
```

```
scav <- GOGOscavenger(UpDown, pick, 243230, write = T, "[dra]hr24_ontology.csv")

head(scav)
```

```
##         refID  geneID proteinID      Family_name     Sub_family_name
## 1392 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1393 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1394 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 1395 DR_A0267 1798063 NP_285590 FAMILY NOT NAMED SUBFAMILY NOT NAMED
## 3038  DR_2281 1797858 NP_296002 RIBONUCLEASE III      RIBONUCLEASE 3
## 3039  DR_2281 1797858 NP_296002 RIBONUCLEASE III      RIBONUCLEASE 3
##           goIDs     logFC
## 1392          0 -5.855846
## 1393          0 -5.855846
## 1394          0 -5.855846
## 1395          0 -5.855846
## 3038 GO:0016788 -1.385570
## 3039 GO:0003676 -1.385570
##                                                    annotation goType
## 1392 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     MF
## 1393 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     BP
## 1394 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     CC
## 1395 hypothetical protein DR_A0267 [Deinococcus radiodurans R1]     PC
## 3038  hypothetical protein DR_2281 [Deinococcus radiodurans R1]     MF
## 3039  hypothetical protein DR_2281 [Deinococcus radiodurans R1]     MF
##                                        goTerms Pathway DE
## 1392                                         0    <NA> -1
## 1393                                         0    <NA> -1
## 1394                                         0    <NA> -1
## 1395                                         0    <NA> -1
## 3038 hydrolase activity, acting on ester bonds    <NA> -1
## 3039                      nucleic acid binding    <NA> -1
```

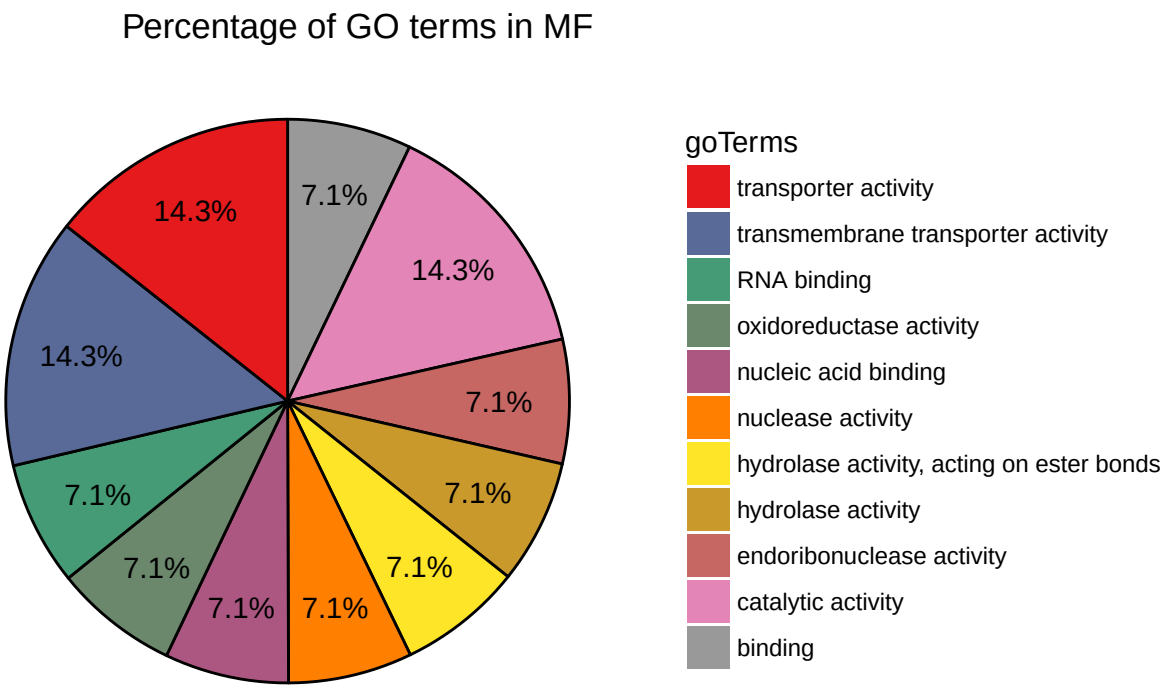
```
tail(scav)
```

```
##          refID  geneID proteinID Family_name Sub_family_name      goIDs
## 19164   DR_t13    <NA>      <NA>        <NA>            <NA>       <NA>
## 1      DR_0972 1799034 NP_294696        <NA>            <NA>       <NA>
## 2     DR_A0212 1797972 NP_296362        <NA>            <NA>       <NA>
## 3     DR_A0210 1797981 NP_296361        <NA>            <NA>       <NA>
## 4      DR_2054 1799983 NP_295777        <NA>            <NA>       <NA>
## 5      DR_1299 1798875 NP_295023        <NA>            <NA> GO:0016021
##           logFC
## 19164 -2.133357
## 1      1.154176
## 2     -2.425909
## 3     -1.382044
## 4     -2.089734
## 5     -1.292373
##                                                                                annotation
## 19164                                                                           tRNA-Ala
## 1                                    hypothetical protein DR_0972 [Deinococcus radiodurans R1]
## 2                                   hypothetical protein DR_A0212 [Deinococcus radiodurans R1]
## 3     peptide ABC transporter, periplasmic peptide-binding protein [Deinococcus radiodurans R1]
## 4                                    hypothetical protein DR_2054 [Deinococcus radiodurans R1]
## 5                                    hypothetical protein DR_1299 [Deinococcus radiodurans R1]
##       goType                         goTerms Pathway DE
## 19164     PC                            tRNA    <NA> -1
## 1       <NA>                            <NA> Uniprot  1
## 2       <NA>                            <NA> Uniprot -1
## 3       <NA>                            <NA> Uniprot -1
## 4       <NA>                            <NA> Uniprot -1
## 5         CC integral component of membrane Uniprot -1
```

4.Step4: This step is to summarize the data in scav. Use GOGOsummary.

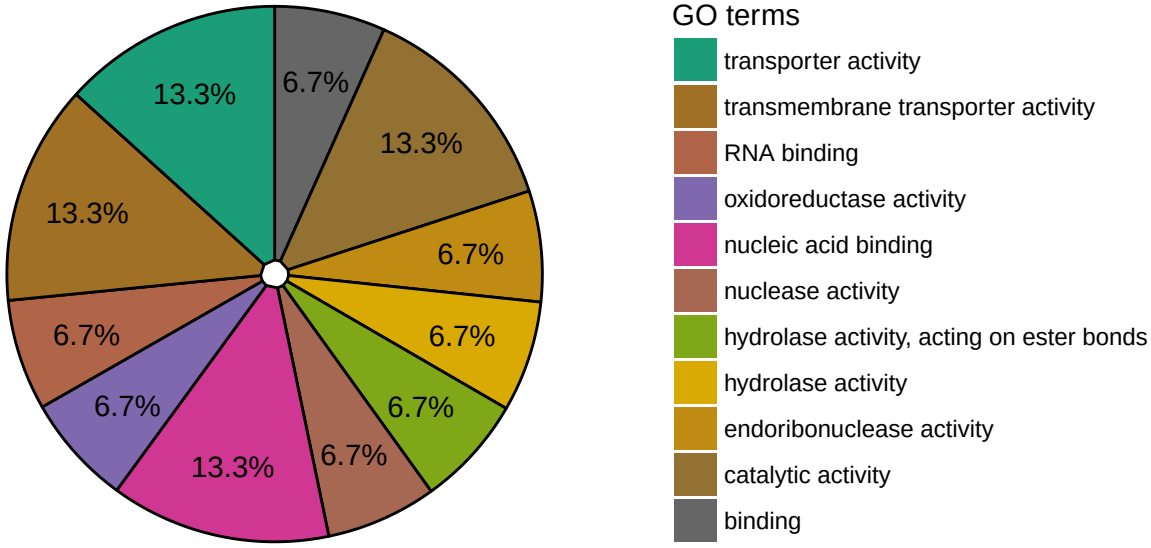
```
GOGOsummary(scav, "MF")
```

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


Percentage of GO terms in MF
14.3%
7.1%
14.3%
14.3%
7.1%
7.1%
7.1%
7.1%
7.1%
7.1%
7.1%
goTerms
transporter activity
transmembrane transporter activity
RNA binding
oxidoreductase activity
nucleic acid binding
nuclease activity
hydrolase activity, acting on ester bonds
hydrolase activity
endoribonuclease activity
catalytic activity
binding

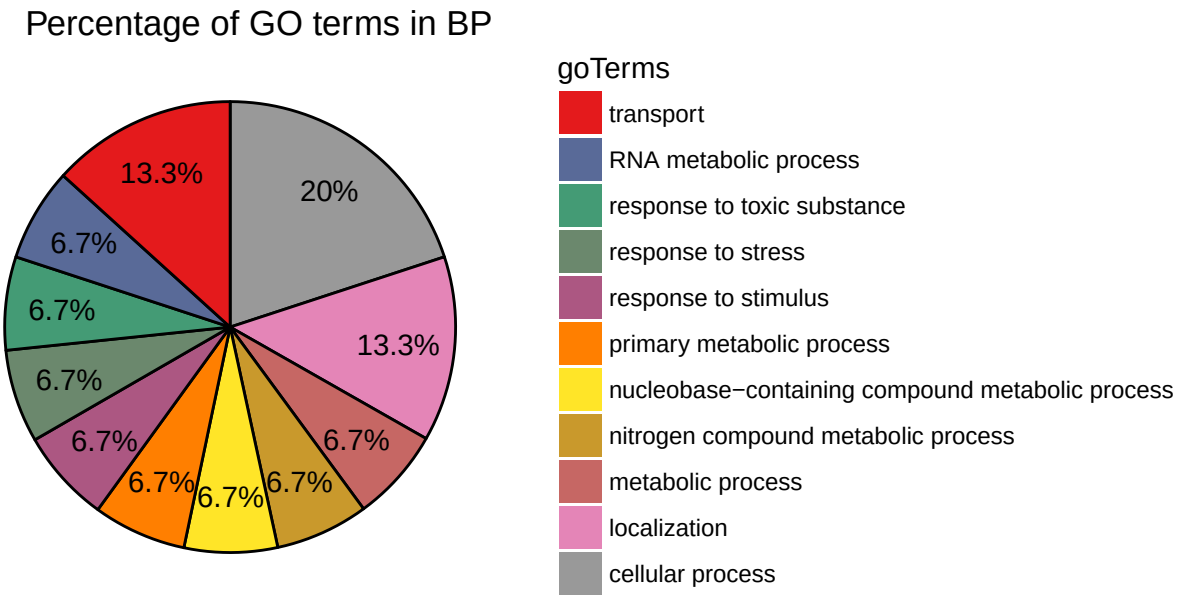
```
##
## [[2]]
```

GO terms percentage in DOWN genes in MF

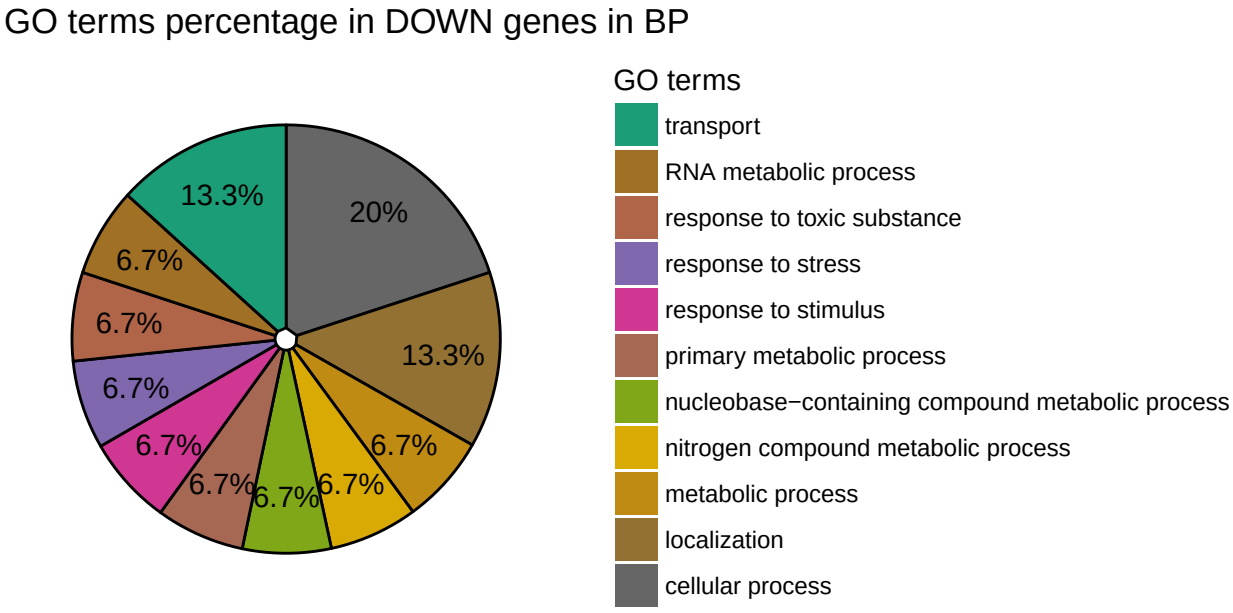



```
GOGOsummary(scav, "BP")
```

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

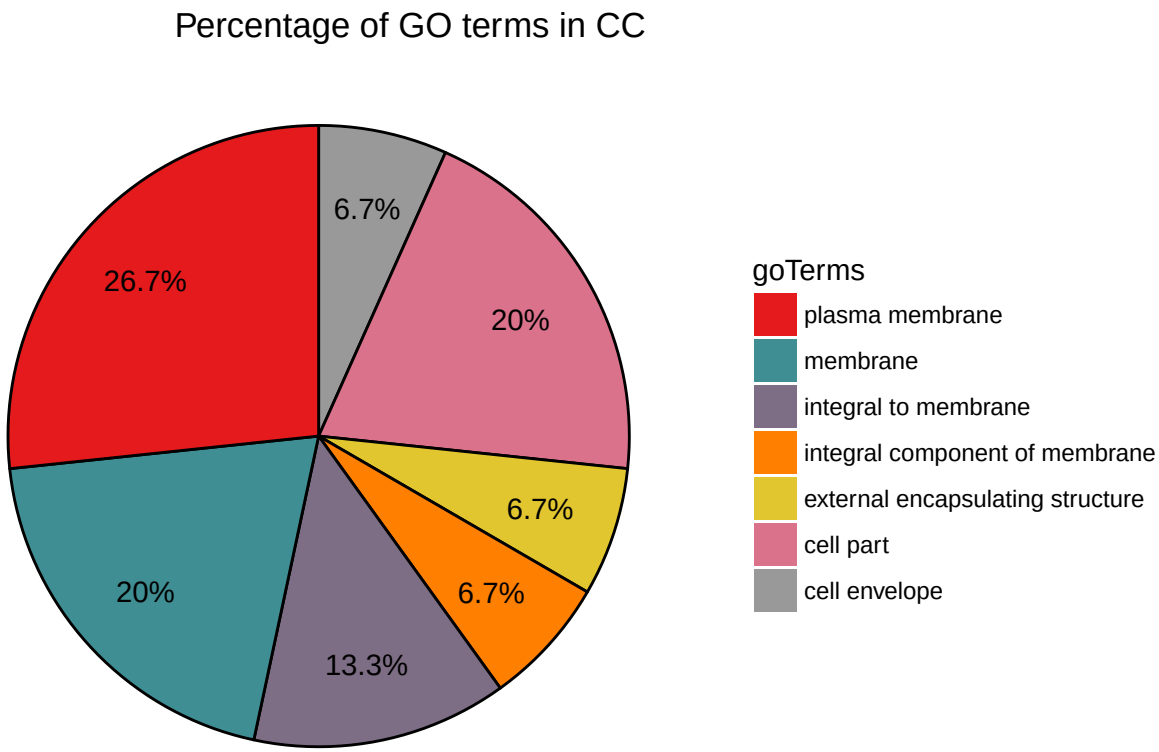

Percentage of GO terms in BP
13.3%
20%
6.7%
6.7%
13.3%
6.7%
6.7%
6.7%
6.7%
6.7%
6.7%
goTerms
transport
RNA metabolic process
response to toxic substance
response to stress
response to stimulus
primary metabolic process
nucleobase−containing compound metabolic process
nitrogen compound metabolic process
metabolic process
localization
cellular process

```
##
## [[2]]
```

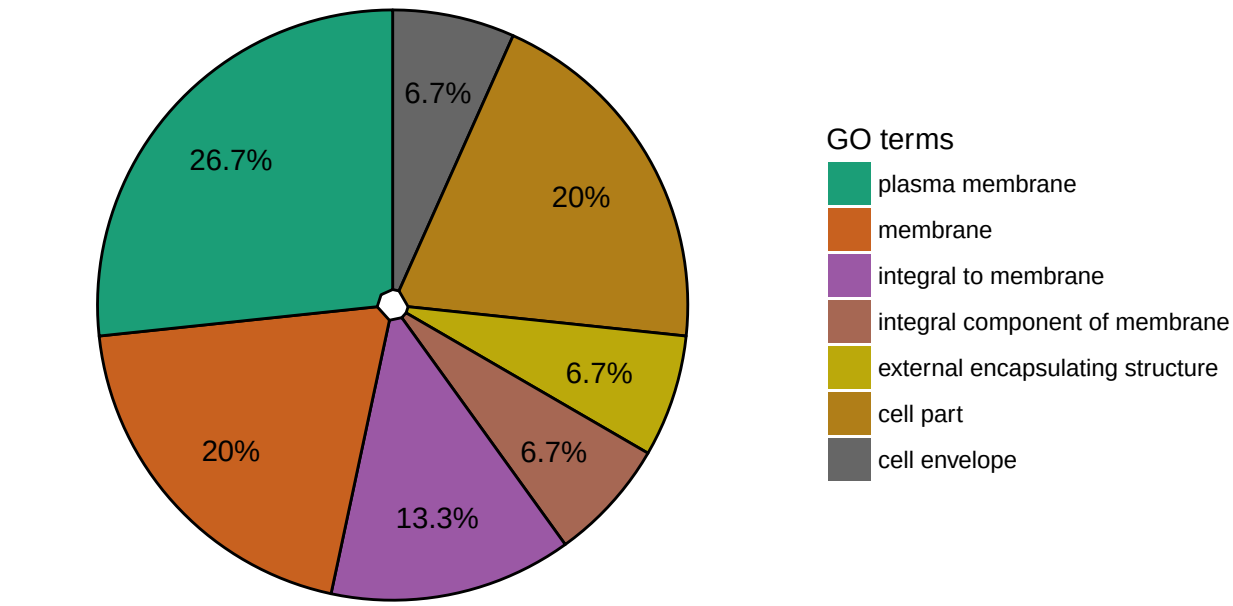



```
GOGOsummary(scav, "CC")
```

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

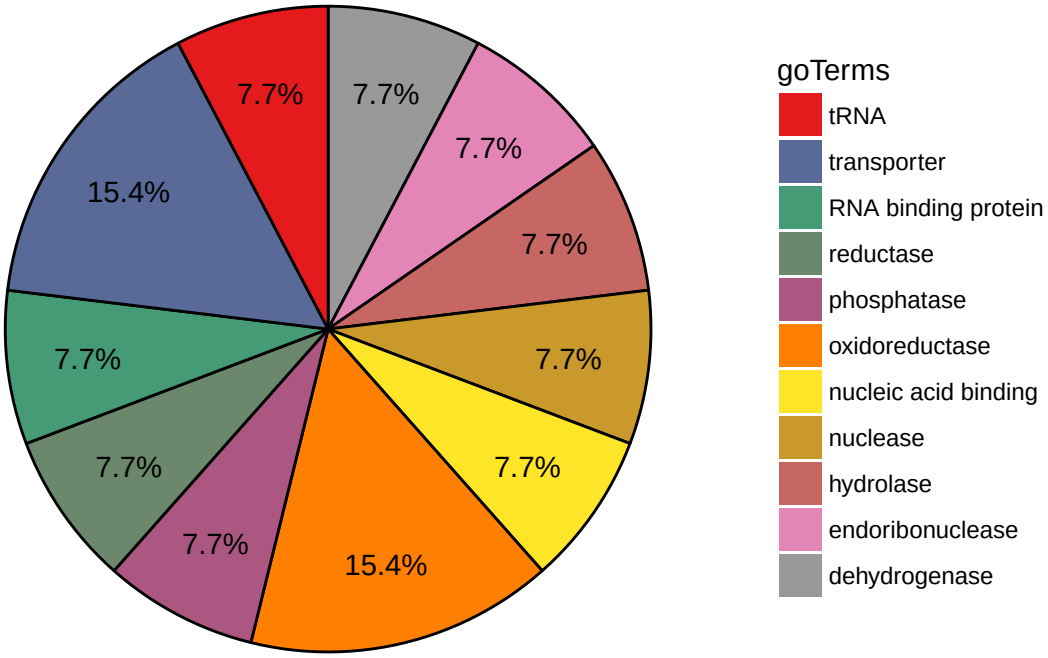

Percentage of GO terms in CC
6.7%
26.7%
20%
6.7%
20%
6.7%
13.3%
goTerms
plasma membrane
membrane
integral to membrane
integral component of membrane
external encapsulating structure
cell part
cell envelope

```
##
## [[2]]
```

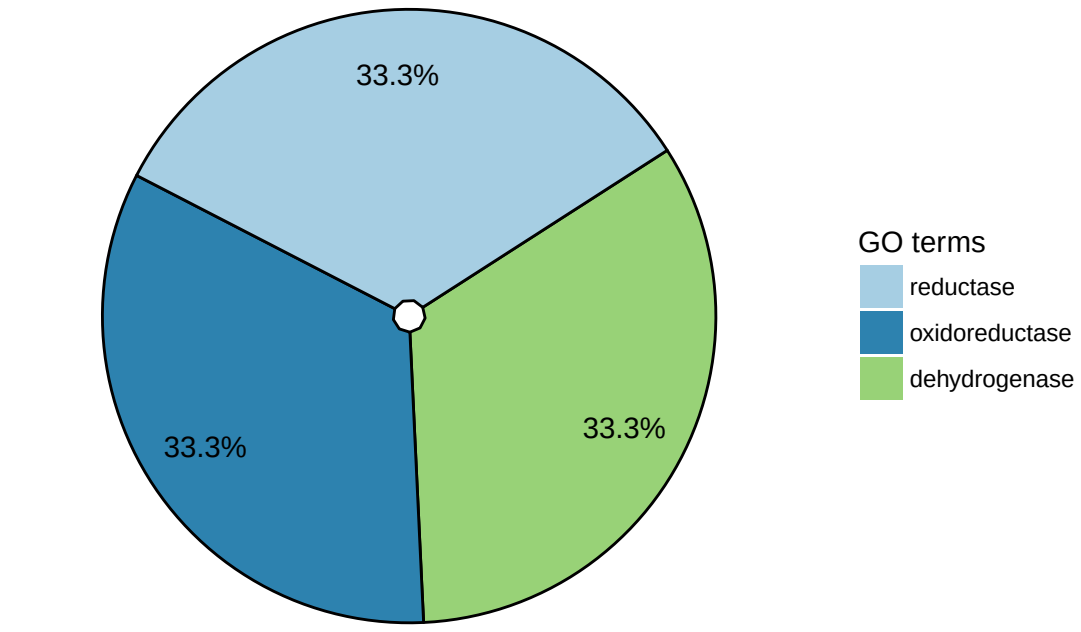



```
GOGOsummary(scav, "PC")
```

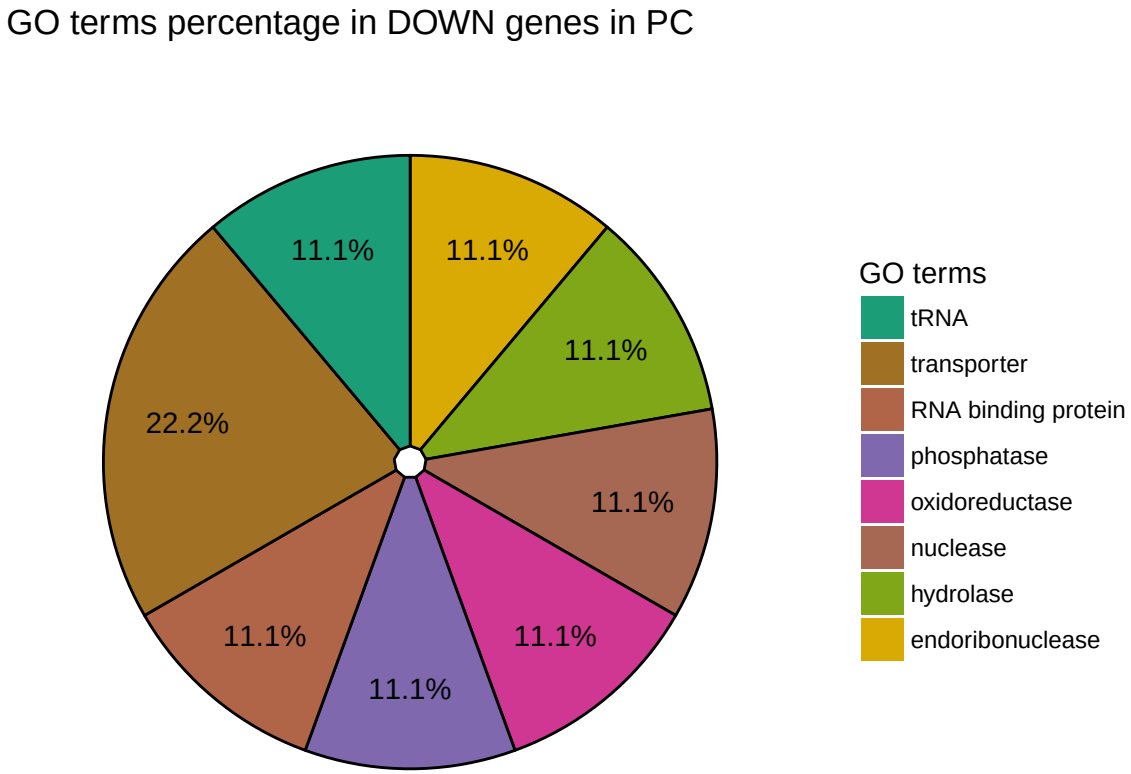
```
## [[1]]
```


Percentage of GO terms in PC
7.7%
7.7%
7.7%
15.4%
7.7%
7.7%
7.7%
7.7%
7.7%
15.4%
goTerms
tRNA
transporter
RNA binding protein
reductase
phosphatase
oxidoreductase
nucleic acid binding
nuclease
hydrolase
endoribonuclease
dehydrogenase

```
##
## [[2]]
```


GO terms Percentage in UP genes in PC
33.3%
33.3%
33.3%
GO terms
reductase
oxidoreductase
dehydrogenase

```
##
## [[3]]
```



5.Step5: Search for the genes you are interested in certain funcitons. Using GOGOsearch().

```
GOGOsearch(scav, F, "full", "transporter", "MF")
```

```
## Warning in GOGOsearch(scav, F, "full", "transporter", "MF"): Argument help
## = F: not showing all the GO therms in your scav data set. To see, set help
## = T.
```

```
##
## MF full view:
##                                              regulation
##                                               down
##   binding                                        1
##   catalytic activity                             2
##   endoribonuclease activity                      1
##   hydrolase activity                             1
##   hydrolase activity, acting on ester bonds      1
##   nuclease activity                              1
##   nucleic acid binding                           1
##   oxidoreductase activity                        1
##   RNA binding                                    1
##   transmembrane transporter activity             2
##   transporter activity                           2
##
## BP full view:
##                                                 regulation
##                                                  down
```

```
##   cellular process                                        3
##   localization                                            2
##   metabolic process                                       1
##   nitrogen compound metabolic process                     1
##   nucleobase-containing compound metabolic process        1
##   primary metabolic process                               1
##   response to stimulus                                    1
##   response to stress                                      1
##   response to toxic substance                             1
##   RNA metabolic process                                   1
##   transport                                               2
##
## CC full view:
##                                    regulation
##                                     down
##   cell envelope                        1
##   cell part                            3
##   external encapsulating structure     1
##   integral component of membrane       1
##   integral to membrane                 2
##   membrane                             3
##   plasma membrane                      4
##
## PC full veiw:
##                        regulation
##                         down up
##   dehydrogenase            0  1
##   endoribonuclease         1  0
##   hydrolase                1  0
##   nuclease                 1  0
##   nucleic acid binding     1  0
##   oxidoreductase           1  1
##   phosphatase              1  0
##   reductase                0  1
##   RNA binding protein      1  0
##   transporter              2  0
##   tRNA                     1  0
##
##
## >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>> Search Results >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
##
## [1] "The terms found in MF that are matched:"
##
## transmembrane transporter activity,transporter activity,transporter
##
##
## [1] "********** transporter up regulated in MF classification. ************"
## Up reguated genes not found
##
## [1] "********* transporter down regulated in MF classification. ************"
##      refID     logFC
## 1 DR_A0208 -1.373335
## 3 DR_A0209 -1.702373
##                                                    annotation
```

```
## 1 peptide ABC transporter permease [Deinococcus radiodurans R1]
## 3 peptide ABC transporter permease [Deinococcus radiodurans R1]
```

```
GOGOsearch(scav, T, "search", "reductase", "MF")
```

```
## MF GO terms:
##  [1] "hydrolase activity, acting on ester bonds"
##  [2] "nucleic acid binding"
##  [3] "RNA binding"
##  [4] "nuclease activity"
##  [5] "binding"
##  [6] "catalytic activity"
##  [7] "endoribonuclease activity"
##  [8] "hydrolase activity"
##  [9] "oxidoreductase activity"
## [10] "transmembrane transporter activity"
## [11] "transporter activity"
##
## BP GO terms:
##  [1] "cellular process"
##  [2] "nucleobase-containing compound metabolic process"
##  [3] "RNA metabolic process"
##  [4] "nitrogen compound metabolic process"
##  [5] "metabolic process"
##  [6] "primary metabolic process"
##  [7] "response to stimulus"
##  [8] "response to toxic substance"
##  [9] "response to stress"
## [10] "localization"
## [11] "transport"
##
## CC GO terms:
## [1] "plasma membrane"                  "cell part"
## [3] "membrane"                         "integral to membrane"
## [5] "cell envelope"                    "external encapsulating structure"
## [7] "integral component of membrane"
##
## PC GO terms:
##  [1] "cellular process"
##  [2] "nucleobase-containing compound metabolic process"
##  [3] "RNA metabolic process"
##  [4] "nitrogen compound metabolic process"
##  [5] "metabolic process"
##  [6] "primary metabolic process"
##  [7] "response to stimulus"
##  [8] "response to toxic substance"
##  [9] "response to stress"
## [10] "localization"
## [11] "transport"
```

```
## Warning in GOGOsearch(scav, T, "search", "reductase", "MF"): Not showing
## the results for all the GO types. To see, set argument view = 'full'
```

```
## >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>> Search Results <<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
##
## [1] "The terms found in MF that are matched:"
```

```
##
## oxidoreductase activity,oxidoreductase,reductase
##
##
## [1] "********** reductase up regulated in MF classification. ************"
## Up reguated genes not found
##
## [1] "********* reductase down regulated in MF classification. ***********"
##      refID     logFC
## 1 DR_A0214 -1.241899
##                                                      annotation
## 1 trp repressor binding protein WrbA [Deinococcus radiodurans R1]
```

```
GOGOsearch(scav, F, "search", "membrane", "CC")
```

```
## Warning in GOGOsearch(scav, F, "search", "membrane", "CC"): Argument help =
## F: not showing all the GO therms in your scav data set. To see, set help =
## T.

## Warning in GOGOsearch(scav, F, "search", "membrane", "CC"): Not showing the
## results for all the GO types. To see, set argument view = 'full'

## >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>> Search Results >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
##
## [1] "The terms found in CC that are matched:"
##
## plasma membrane,transmembrane transporter activity,membrane,integral to membrane,integral component
##
##
## [1] "********** membrane up regulated in CC classification. ************"
## Up reguated genes not found
##
## [1] "********* membrane down regulated in CC classification. ***********"
##       refID     logFC
## 1  DR_A0214 -1.241899
## 2  DR_A0208 -1.373335
## 5  DR_A0207 -1.204668
## 7  DR_A0209 -1.702373
## 10  DR_1299 -1.292373
##                                                       annotation
## 1  trp repressor binding protein WrbA [Deinococcus radiodurans R1]
## 2    peptide ABC transporter permease [Deinococcus radiodurans R1]
## 5       hypothetical protein DR_A0207 [Deinococcus radiodurans R1]
## 7    peptide ABC transporter permease [Deinococcus radiodurans R1]
## 10       hypothetical protein DR_1299 [Deinococcus radiodurans R1]
```