

# netGO User's Manual

Contact: Jinhwan Kim [kjh0530@unist.ac.kr](mailto:kjh0530@unist.ac.kr)

2019-08-06

## 1. Introduction

**netGO** is an R-Shiny package for network-integrated pathway enrichment analysis. It also provides the conventional Fisher's exact test. Specifically, it provides user-interactive visualization of enrichment analysis results and related networks. The netGO package is available at Github (<https://github.com/unistbig/netGO>). Currently, netGO provides network and annotation gene-set data for four species including human, mouse, yeast, and Arabidopsis thaliana. These data are all available from another repository (<https://github.com/unistbig/netGO-Data/>)

## 2. Installation and Example codes

1) Prerequisite R packages can be installed by simply executing the following R codes:

```
install.packages(c('devtools', 'Rcpp', 'shinyjs', 'DT', 'doParallel', 'foreach', 'parallel', 'htmlwidgets', 'googleVis', 'V8'))
library(devtools)
install_github('unistbig/shinyCyJS')
```

2) netGO can be installed and executed as follows:

```
library(devtools) # load devtools to use 'install_github' function
install_github('unistbig/netGO') # install netGO
library(netGO) # load netGO
DownloadExampleData() # Download and load example datasets
obj = netGO(genes = brca[1:20], genesets = genesets, network = network, genesetV = genesetV) # Executing netGO
netGOVis(obj = obj, genes = brca[1:20], genesets = genesets, R = 50, network = network) # Visualization of the result
```

## 3. Data

Human, mouse, yeast and Arabidopsis data are available at <https://github.com/unistbig/netGO-Data>.

\* Tip: Human STRING and MSigDB C2 data are directly loaded using 'DownloadExampleData()' function.

Species	Data Type	File name	Object name
Human	Network	1) Human/networkString.RData 2) Human/networkHumannet.RData	network
	Gene-set (mSigDB C2)	Human/c2gs.RData	genesets
	Pre-calculated interaction data  <div>NOTE: After downloading *1.RData and *2.RData, <u>genesetV1</u> and <u>genesetV2</u> must be <u>joined</u> by a row! <u>genesetV=rbind(genesetV1,genesetV2)</u></div>	1) For <i>STRING</i> , Human/genesetVString1.RData Human/genesetVString2.RData	genesetV1 genesetV2
		2) For <i>HumanNet</i> , Human/genesetVHumannet1.RData Human/genesetVHumannet2.RData	genesetV1 genesetV2
Mouse	Network	networkMousenet.RData	network
	Gene-set (KEGG)	KEGGmouse.RData	genesets
	Pre-calculated interaction data	genesetVMousenet.RData	genesetV
Yeast	Network	networkYeastnet.RData	network
	Gene-set (KEGG)	KEGGyeast.RData	genesets

Arabidopsis	Network	networkAranet.RData	network
	Gene-set (KEGG)	KEGGara.RData	genesets

Table M1. Network and gene-set data provided in netGO

## 4. Functions

The two main functions of netGO package are '**netGO**' ([for enrichment test](#)) and '**netGOVis**' ([for visualization of the test results](#)).

### 1) netGO

netGO function takes six arguments: **1) genes**, **2) genesets**, **3) network**, **4) genesetV**, **5) alpha (optional)** and **6) nperm (optional)**. It returns the *p*-values of gene-sets (dataframe) derived from netGO and Fisher's exact test. Note that the members in **genes** (denoted by A, B, C here) should be given in gene symbols when using the default STRING and MSigDB data. Other types of gene names are also acceptable if the corresponding customized data (network and gene-set data) are uploaded. Descriptions of each argument are as follows:

- ① **genes**: A character vector of input genes (e.g., differentially expressed genes).
- ② **genesets**: A list of gene-sets consisting of groups of genes (e.g., C2 category of MSigDB).
- ③ **network**: A numeric matrix of network data. The range of network score is [0,1].

	A	B	C
A	0	0.1	0.76
B	0.1	0	0.324
C	0.76	0.324	0

Figure M1. Example for network data.

- ④ **genesetV**: A numeric matrix of pre-calculated interaction data between genes and gene-sets. The matrix dimension becomes [ {# of genes} X {# of gene-sets} ]. This matrix can be obtained using **BuildGenesetV** function with the network and gene-set objects as input arguments. Calculating this matrix corresponds to the preprocessing step of netGO.

```
genesetV = BuildGenesetV(network, genesets)
```

	GS1	GS2	GS3
A	0.837	1.647	0.074
B	0	1.750	0.113
C	0.464	0.486	0.442

Figure M2. Example for genesetV.

- ⑤ **alpha** (optional): A numeric parameter balancing the weights between the relative and absolute network scores (See (1) in the main text). The value is between 0 and 1 and the default is 0.5.
- ⑥ **nperm** (optional): The number of resampling. Default is 10,000.

### 2) netGOVis

netGOVis also takes six input arguments: **1) obj**, **2) genes**, **3) genesets**, **4) network**, **5) R (optional)**, **6) Q (optional)**. This function visualizes the test results on web browser (google chrome is recommended). The resulting network graphs and table are downloadable.

- ① **obj**: data frame of the test results obtained from 'netGO' function. It consists of three columns including 1) gene-set name and *p*-values calculated from 2) netGO (netGOP) and 3) Fisher's exact test (FisherP).

- ② R (optional): Gene-set rank threshold, default is 50 (Top 50 gene-sets in either method will be displayed).
- ③ Q (optional): Gene-set Q-value threshold, default is 1. (Gene-sets with Q-values below this threshold will be displayed)
- ④ genes, genesets, network: the same arguments in the 'netGO' function.

## 5. Network visualization

netGO visualization page consists of three parts such as **Network**, **Table**, and **Bubble**.

- 1) **Network** panel displays the input genes, selected gene-set, and the network connections between the two.
  - Sky blue nodes represent input genes (e.g., DE genes)
  - Yellow nodes represent genes in the selected gene-set
  - Green nodes represent the intersection of input genes and the gene-set.
  - The edge width represents the strength of interaction between two nodes.
  - \* Nodes without edges are discarded.
  - \* The gene-set can be selected by clicking on the gene-set name in the **Table** on the right side.
  - \* The users can download the graph image as SVG format.
- 2) **Table** contains the names of gene-sets and their p-values evaluated from netGO and Fisher's exact test, respectively. It is downloadable by clicking the 'Download Table' button in the upper right of the table (Figure M3).

Gene-set name	netGO q-value	Fisher's exact test q-value
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_DUCTAL_NORMAL_DN	0.0422	0.0035
TURASHVILI_BREAST_DUCTAL_CARCINOMA_VS_LOBULAR_NORMAL_DN	0.0422	1
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_DUCTAL_NORMAL_DN	0.0422	1
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_LOBULAR_NORMAL_UP	0.0422	0.5193
WILCOX_RESPONSE_TO_PROGESTERONE_DN	0.0422	1
WONG_ENDMETRIUM_CANCER_DN	0.0422	1

Showing 1 to 6 of 82 entries

Figure M3. Table panel that shows the list of significant gene-sets

- 3) **Bubble** module includes the bubble chart of significant gene-sets. For each gene-set, the overlap (x-axis) and network score (y-axis) are represented where sum of the two scores is the integrated score  $P(T \rightarrow A)$  as below.

$$P(T \rightarrow A) = \underbrace{\frac{|T \cap A|}{|T|}}_{\text{Overlap score}} + \underbrace{\frac{1}{|T|} \sum_{x \in T-A} AI(x, A)^a \cdot RI(x, A)^{1-a}}_{\text{Network score}}$$

Where

$$AI(x, A) = \frac{1}{|A|} \sum_{y \in A} I(x, y)$$

$$RI(x, A) = \frac{1}{|N(x)|} \sum_{y \in A} I(x, y)$$

$T$  and  $A$  are target and annotation gene-sets, respectively;  $|\cdot|$  denotes the size of a set;  $I(x, y)$  represents the interaction score between  $x$  and  $y$  genes normalized to unit interval  $[0, 1]$ ;  $N(x)$  is the set of all scores of interactions to  $x$  and  $|N(x)|$  is their summation.

Note that high network score means a dense interaction between input genes and gene-set.

- The size of bubble represents the significance of each gene-set ( $-\log_{10} p - value$ ).
- Colors of bubble indicate the method that detected the gene-set as significant.

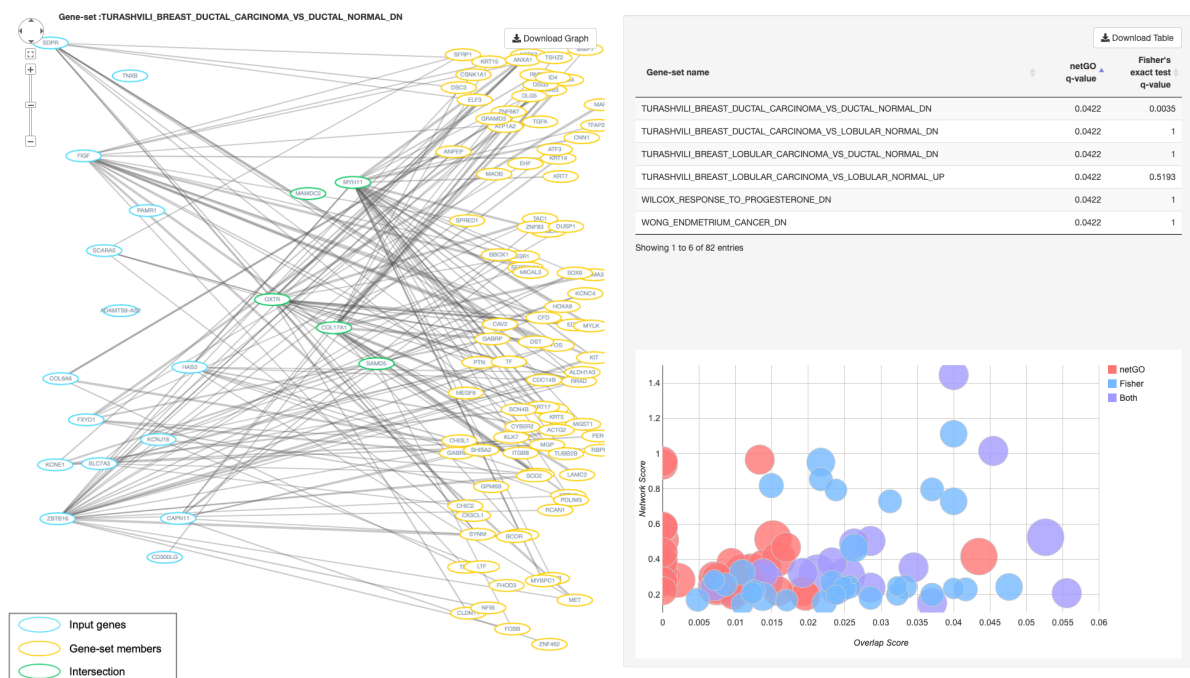


Figure M4. netGO interface