Network Enrichment Analysis using neaGUI Package

Setia Pramana, Woojoo Lee, Andrey Alexeyenko, Yudi Pawitan

April 2, 2013

1 Introduction

After discovering list of differentially expressed genes, the next challenge is to interpret them on a way that is consistent with biological hypotheses. It is known that genes work on networks. In humans there are around 22,258 to whom correspondence should be addressed protein-coding genes connected in 650,000 predicted interactions (Minguez and Dopazo, 2010). The knowledge of gene-protein network is now increasingly used as a based for characterizing gene sets. Alexeyenko et al., 2012 proposed a network enrichment analysis (NEA) method which systematically implements the network approach to describe novel gene sets with biologically meaningful functional categories were proposed.

The NEA method integrates functional information and network connectivity of nearly all protein-coding genes and quantifies the over/under-representation of the functional group members among the neighbors in the gene network rather than in the AGS itself. In the NEA methods a fast network randomization algorithm to obtain the distribution of any network statistics under the null hypothesis of no association between an AGS and FGS is used.

2 Implementation

The `neaGUI` was developed using the R-tcl/tk interface implemented in the `R-tcl/tk` package (Dalgaard, 2001) and is freely available for Windows and Linux from R forge site: `https://r-forge.r-project.org/projects/neagui/`. The `neaGUI` requires the following R packages: `tcltk`, `KEGG.db`, `GO.db`, `reactome.db`, `org.Hs.eg.db`, `AnnotationDbi`, and `hwriter`.

2.1 Installation

The following command can be used to install and load the package:

```
> install.packages("neaGUI", repos="http://R-Forge.R-project.org")
> library(neaGUI)
> neaGUI()
```

Note that when loading `neaGUI` will check if all required packages are already installed. If there are packages not yet installed, the required packages will be downloaded automatically (internet connection is required).

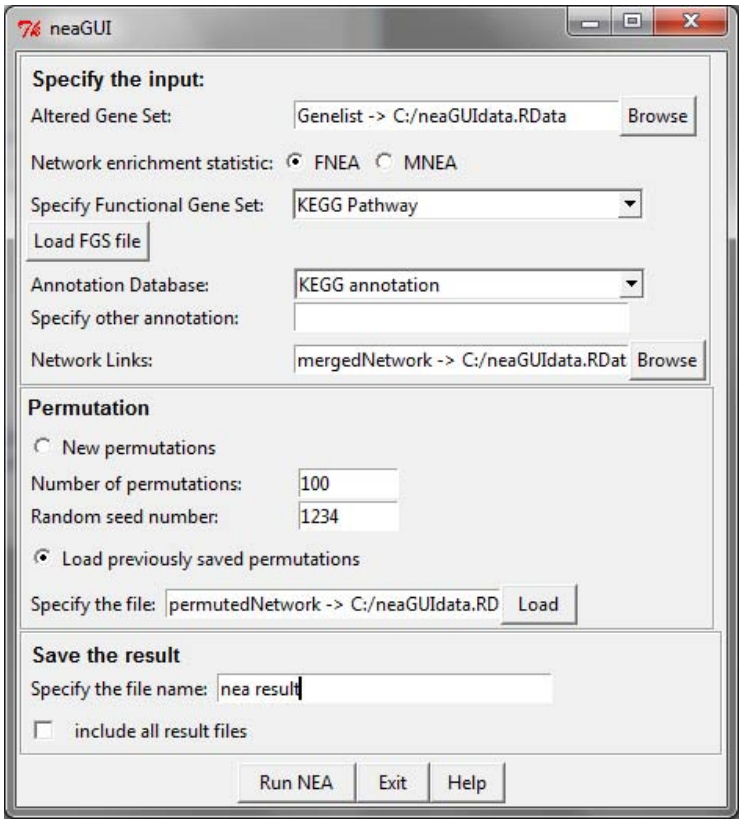

Figure 1: The neaGUI package main dialog box.

3 neaGUI Description

The main dialog box of `neaGUI` package is shown in Figure 1. The dialog box is divided into three parts: input specification, permutation, and output.

3.1 Input

The main input of neaGUI is a list of altered gene sets (AGSs) which is usually a list of differentially expressed genes and gene network links. The gene network input can be a vector of gene pairs or a list representing the network link. For the vector case, each element has a combined name of two gene symbols with space separation. The input can be an R object, csv file, or txt file.

The other options need to be specified are the following:

1. The statistics of the NEA method , fixed NEA (FNEA, default) and maximum NEA (MNEA). FNEA depends on the size of AGS while MNEA avoids the dependence by taking the largest statistic across AGS. More detail please see Alexeyenko et al. (2012).

2. Functional gene sets (FGS). Options to specify known gene sets, databases of group of genes describing their biological activities are necessary. The

options provided are GO ontologies (cellular component, biological process, moelcular function) or KEGG pathway. Users can provide their own functional gene sets as a list. Please see the format of AGS input.

3. Annotation database. The annotation database options which will provide detailed information about the pathways. Besides two provided annotation database options, KEGG pathways(based on KEGG.db package) and Gene Ontologies (based on GO.db), user defined or other annotation database can be specified.

4. Gene network links. The network input is a vector of gene pairs (two gene symbols with separation) or a list representing the network link. Users are recommended to use a comprehensive merged gene networks discussed by Alexeyenko et al. (2012) which contains 1,445,027 links of 16,299 distinct Human Proteome Organisation (HUPO) genes (Alexeyenko and Sonnhammer, 2009). The merged network can be downloaded from `http://www.meb.ki.se/~yudpaw/papers/neaGUIdata.RData`. Example of network input can be found here.

5. Number of permutation. The default number of permutation is 100, however user can increase the number. Note that this permutation procedure only needs to be performed once for each network list. Then the result of permuted network can be used for other analysis for the same network list using options "load previously saved permutations". A permuted network for the merged network using 100 permutations can be obtained from `http://www.meb.ki.se/~yudpaw/papers/neaGUIdata.RData`.

6. Name of the output. Note that the result files will be saved in the same directory of the AGS input directory.

3.2 Output

There are three output files, which the name is defined by users, that will be produced by the neaGUI:

1. An html file
2. A csv file containing summary of the result
3. An R workspace containing complete results (if the option "include all result files" is checked).

Figure 2 shows the resulting html file which contain the summary of the results. Here the the imputed FGS (pathways) are shown (sorted by on te p-values) with corresponding: number of observed and expected network links, number of genes, number of AGS genes, z-score, p-values based on network permutations, and the false discovery rate (FDR). The same information is also given in the csv file. Note that eash pathway IDs provide a hyper link to the pathways/ontologies description in the KEGG or GO web site. The same information is also given in the csv file.

The R workspace output contains more detail results from the neaGUI which includes: Nea Result (the same as the csv file), the specified AGS, FGS, Network, AnnotationDB, and the permuted Network (if it is specified before, it will be the same as the input, if not, a new permuted network will be provided).

The result of neaGUI

AGS: length: 10
FGS: KEGG
Annotation: KEGG.db
Statistics: FNEA
Number of network links: 30
Number of Permutations: 100

PATH_ID	PATH_NAME	Number_links	Expected_links	Number_of_Genes	Number_of_AGS_genes	Z_score	P_value	FDR
hsa03450	Non-homologous end-joining	1	0.1089	14	0	2.846192	0.217822	1
hsa01100	Metabolic pathways	3	1.4455	1131	2	1.425263	0.396040	1
hsa00010	Glycolysis / Gluconeogenesis	1	0.2079	65	0	1.942114	0.415842	1
hsa00020	Citrate cycle (TCA cycle)	1	0.2079	30	0	1.942114	0.415842	1
hsa00620	Pyruvate metabolism	1	0.2079	40	0	1.942114	0.415842	1
hsa04146	Peroxisome	1	0.2079	79	0	1.942114	0.415842	1
hsa05140	Leishmaniasis	1	0.2079	73	0	1.942114	0.415842	1
hsa05142	Chagas disease (American trypanosomiasis)	1	0.2079	104	0	1.942114	0.415842	1
hsa05146	Amoebiasis	1	0.2079	106	0	1.942114	0.415842	1
hsa05200	Pathways in cancer	1	0.2079	327	0	1.942114	0.415842	1
hsa05222	Small cell lung cancer	1	0.2079	85	0	1.942114	0.415842	1
hsa04110	Cell cycle	1	0.2475	128	0	1.734907	0.495050	1
hsa04144	Endocytosis	1	0.2970	203	0	1.403442	0.554455	1
hsa00240	Pyrimidine metabolism	1	0.3663	99	0	1.130002	0.653465	1
hsa04114	Oocyte meiosis	1	0.3861	114	0	1.053685	0.673267	1
hsa03040	Spliceosome	0	0.9604	128	0	-1.036577	0.693069	1
hsa05145	Toxoplasmosis	1	0.3663	133	0	1.209547	0.693069	1
hsa00330	Arginine and proline metabolism	1	0.3762	54	1	1.185103	0.712871	1

Figure 2: An html output of the neaGUI package.

4 Command Line nea

For the users who like to use command line nea, the package also provide a function called to run the NEA analysis.

```
nea(ags, fgs, fgslib = NULL, network, pnet = NULL, nperm = 50,
stat="F", seed = NULL)
```

The `ags` is a vector of altered genes. Gene symbols (upper case) are used as a default. The `fgs` is a list defined by user or a character to specify GO ontologies or KEGG pathway. User can provide their own functional gene sets as a list. Options to specify GO ontologies or KEGG pathway are "CC","BP","MF", "KEGG" and "Reactome" (cellular component, biological process, moelcular function and KEGG pathway, Reactome). The `fgslib` is to specify the name of annotation data. To use GO terms or KEGG pathways, a specific annotation data should be specified.

The option `network` is used to defined the network link. `pnet` is for specifying a list of randomly permuted networks. If we do not have it, the default is null. Last option is `stat` with two types of network enrichment statistic: FNEA and MNEA. FNEA (stat="F") depends on the size of ags while MNEA (Stat="M") avoids the dependence by taking the largest statistic across AGS. Default is FNEA.

```
> library(neaGUI)
> AGS<-c("AIFM3","DIMT1L","ADNP","AHCYL1","EIF4H","RGL1",
+                  "SEC23IP","EIF4A1","CSNK2B","NOS3")
> NETWORK<-c("DNAJC6 RGL1","C1ORF156 NCBP2","AHCYL1 RTN3",
+                  "PLK4 SKIV2L2","C22ORF28 MESDC2","TINP1 UTP23",
+                  "HEATR3 MVD","WBP11 XAB2","CSNK2B PA2G4","GCN1L1 RRM2",
+                  "DIMT1L SMC1A","GPN3 THOC3", "DLG3 GPHN",
+                  "C19ORF29 EXOSC4","AIFM3 SFXN5","HSPA1L RUVBL2",
```

```
+                 "DLAT EIF4A1","ADNP XRCC5", "NOS2 NOS3","CIZ1 TLK2",
+                 "MRPL49 RPS7","GSPT1 SLK","LUC7L2 SEC23IP","DHX8 IGF2BP3",
+                 "CNTROB SASS6","MRPS12 RPLP2","DHODH EIF4H","GINS3 KIF23",
+                 "ANXA5 TGFBI","CDK5 PMM1")
> res <- nea(ags=AGS, fgs = "KEGG", fgslib = "KEGG.db", network=NETWORK,
+               pnet = NULL, nperm = 10, stat="F", seed = 1234)
> head(res$MainResult)

   PATH_ID                                PATH_NAME Number_links Expected_links
1 hsa00010             Glycolysis / Gluconeogenesis            1     0.36363636
2 hsa00020                Citrate cycle (TCA cycle)            1     0.36363636
3 hsa00030                Pentose phosphate pathway            0     0.00000000
4 hsa00040 Pentose and glucuronate interconversions            0     0.00000000
5 hsa00051          Fructose and mannose metabolism            0     0.09090909
6 hsa00052                     Galactose metabolism            0     0.00000000
  Number_of_Genes Number_of_AGS_genes    Z_score   P_value FDR
1              65                   0  1.2613124 0.7272727   1
2              30                   0  1.2613124 0.7272727   1
3              27                   0  0.0000000 1.0000000   1
4              32                   0  0.0000000 1.0000000   1
5              36                   0 -0.3015113 1.0000000   1
6              27                   0  0.0000000 1.0000000   1
```

The result matrix is similar with the html output discussed in the previous section.

References

1. A. Alexeyenko and E. L. Sonnhammer. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Research*, 19:1107-1116, 2009.

2. A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehti, and Y. Pawitan. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, 13:226, 2012.

3. P. Dalgaard. The R-Tcl/Tk interface. *DSC 2001 Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, Vienna, Austria, 2001.

4. P. Minguez and J. Dopazo. Functional genomics and networks: new approaches in the extraction of complex gene modules expert. *Rev. Proteomics*, 7(1):55-63, 2010.