

How To Use ESEA

Junwei Han

January 22, 2015

Contents

1 Overview	1
2 Get the edge background set data and pathways edge sets data	1
3 Discovering the dysregulated pathways based on edge set enrichment analysis	2
3.1 Calculate the differential correlation score for edges	2
3.2 Discovering the dysregulated pathways	3
4 Plot running result diagram	4
4.1 Plot global edge correlation profile	4
4.2 Plot running Edge enrichment score	5
4.3 Plot the pathway-result network diagram	5
4.4 Save a pathway-result network to a file which can be input to the Cytoscape software	8
5 Session Info	10

1 Overview

This vignette demonstrates how to easily use the ESEA package. The package can identify dysregulated pathways associated with a specific phenotype by investigating the changes of biological relationships of pathways in the context of gene expression data. Our system constructs a background set of edges by extracting pathway structure (e.g. interaction, regulation, modification, and binding etc.) from the seven public databases (KEGG; Reactome; Biocarta; NCI; SPIKE; HumanCyc; Panther) (see the section 2). Our system also constructs the edge sets of pathways for each of the above databases (see the section 2). The ESEA can quantify the change of correlation between genes for each edge based on gene expression data with cases and controls (see the section 3.1). The ESEA uses the weighted Kolmogorov-Smirnov statistic to calculate an edge enrichment score (EES), which reflects the degree to which a given pathway is associated the specific phenotype (see the section 3.2).

2 Get the edge background set data and pathways edge sets data

The section introduces how to obtain the edge background set data and pathways edge sets data. The human pathways data are collected from the seven popular public databases (KEGG; Reactome; Biocarta; NCI; SPIKE; HumanCyc; Panther). Each pathway in the above databases is converted into an undirected graph. Each node in the graph represents a gene, and each edge represents a relationship such as interaction, regulation or modification etc. between genes. The edge set for each pathway can

be extracted from the corresponding pathway graph. These pathway graphs are merged into a global gene interaction network. All the edges in the global network are used as the background set of edges.

For each pathway in the seven pathway database (KEGG; Biocarta; Reactome; NCI; SPIKE; HumanCyc; Panther), the edge set are extracted from the corresponding pathway graph. The edge sets of pathways are therefore created for each of the above databases.

The following commands can obtain the edge background set data and pathways edge sets data.

```
> #obtain the data for background set of edges.
> edgesbackground<-GetEdgesBackgrandData()
> edgesbackground[1:10,]
```

```
      V1      V2
[1,] "A1CF" "APOBEC1"
[2,] "A2M"  "APOA1"
[3,] "A2M"  "CDC42"
[4,] "A2M"  "F2"
[5,] "A2M"  "FOXO1"
[6,] "A2M"  "GAB1"
[7,] "A2M"  "GAB2"
[8,] "A2M"  "GRB2"
[9,] "A2M"  "KLKB1"
[10,] "A2M" "MMP1"
```

```
> #obtain the edge sets of pathways.
> pathwayEdge.db<-GetPathwayEdgeData()
```

3 Discovering the dysregulated pathways based on edge set enrichment analysis

The section introduces the Edge Set Enrichment Analysis (ESEA) method for identifying canonical biological pathways associated with a specific phenotype. ESEA identifies dysregulated pathways by investigating the changes of biological relationships of pathways in the context of gene expression data. ESEA integrates pathway structure (e.g. interaction, regulation, modification, and binding etc.) and differential correlation among genes. The biological pathways were collected from the seven public databases (KEGG; Reactome; Biocarta; NCI; SPIKE; HumanCyc; Panther). A background set of edges is constructed by extracting pathway structure from each pathway in the seven databases. The information-theoretic measure, mutual information (MI), is applied to quantify the change of correlation between genes for each edge based on gene expression data with cases and controls (see the section 3.1). An edge list was formed by ranking the edges according to their changes of correlation. Finally, the weighted Kolmogorov-Smirnov statistic is used to prioritize the pathways by mapping the edges in the pathway to the edge list (see the section 3.2).

3.1 Calculate the differential correlation score for edges

For each edge, we estimated the mutual information (MI) between two genes in the expression data of all samples and control samples respectively. The difference of MI between all samples and control samples is used as the differential correlation score of the edge.

The function `calEdgeCorScore` can calculate the differential correlation scores of two genes in the edge. The following commands can calculate the differential correlation scores of edges in a given gene expression dataset.

```

> #get example data
> dataset<-GetExampleData("dataset")
> class.labels<-GetExampleData("class.labels")
> controlcharacter<-GetExampleData("controlcharacter")
> #get the data for background set of edges
> edgesbackground<-GetEdgesBackgroundData()
> #Calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharacter, edgesbackground)
> #print the top ten results to screen
> EdgeCorScore[1:10]

      AANAT|ACLY      AANAT|ACSL1      AANAT|ACSL3      AANAT|ACSL4      AANAT|ACSL6
-0.05406403    -0.28411395    -0.07583330    -0.07442451      0.14375235
      AANAT|ASMT      AANAT|DDC      AANAT|MAOA      AANAT|MAOB      AANAT|SLC25A16
-0.24033758    -0.06331058    -0.10133133    -0.01922738      0.20428877

>
> #Each element is the differential correlation score of an edge and whose name
> # correspond to the edge in the background set of edges.

```

3.2 Discovering the dysregulated pathways

ESEA identify dysregulated pathways by investigating the changes of biological relationships of pathways in the context of gene expression data. The weighted Kolmogorov-Smirnov statistic is used to evaluate each pathway and the permutation is used to calculate the statistical significance of pathways.

The function `ESEA.Main` can identify the dysregulated pathways. The following commands can identify the dysregulated pathways in a given gene expression dataset with default parameters.

```

> #get example data
> dataset<-GetExampleData("dataset")
> class.labels<-GetExampleData("class.labels")
> controlcharacter<-GetExampleData("controlcharacter")
> #get the data for background set of edges
> edgesbackground<-GetEdgesBackgroundData()
> #get the edge sets of pathways
> pathwayEdge.db<-GetPathwayEdgeData()
> #calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharacter, edgesbackground)
> #identify dysregulated pathways by using the function ESEA.Main
> Results<-ESEA.Main(
+ EdgeCorScore,
+ pathwayEdge.db,
+ weighted.score.type = 1,
+ pathway = "kegg",
+ gs.size.threshold.min = 15,
+ gs.size.threshold.max = 1000,
+ reshuffling.type = "edge.labels",
+ nperm =10,
+ p.val.threshold=-1,
+ FDR.threshold = 0.05,
+ topgs =1
+ )

```

```
[1] "Running ESEA Analysis..."
```

```
> #print the summary results of pathways to screen  
> Results[[1]][[1]][1:5,]
```

		GS	SOURCE	SIZE	ES	NES	NOM	p-val
1	Cysteine and methionine metabolism	kegg		66	0.42012	1.7047		0
2	Phototransduction	kegg		38	0.38751	1.4631		0
3	Colorectal cancer	kegg		80	0.36905	1.3853		0
4	Inflammatory bowel disease (IBD)	kegg		58	0.36125	1.3765		0
5	Vibrio cholerae infection	kegg		44	0.36971	1.3732		0

	FDR	q-val	Tag	Edge	Signal
1	0	0.288		0.147	0.246
2	0	0.421		0.257	0.313
3	0	0.338		0.194	0.272
4	0	0.241		0.123	0.212
5	0	0.364		0.21	0.287

```
> #print the detail results of pathways to screen  
> Results[[2]][[1]][1:5,]
```

#	EdgeID	List	Loc	EdgeCorScore	RES	CORE_ENRICHMENT
1	1	PLCB2 PLCB3	73744	-0.347	0.0154	YES
2	2	PLCB1 PRKCA	70182	-0.231	-0.0699	YES
3	3	PLCB3 PRKCG	68833	-0.209	-0.141	YES
4	4	PLCB4 PRKCG	67811	-0.194	-0.207	YES
5	5	TNF VCAM1	66075	-0.174	-0.258	YES

```
> #write the summary results of pathways to tab delimited file.  
> write.table(Results[[1]][[1]], file = "kegg-SUMMARY RESULTS Gain-of-correlation.txt",  
+ quote=F, row.names=F, sep = "\t")  
> write.table(Results[[1]][[2]], file = "kegg-SUMMARY RESULTS Loss-of-correlation.txt",  
+ quote=F, row.names=F, sep = "\t")  
> #write the detail results of genes for each pathway with FDR.threshold< 0.05 to tab delimited file.  
> for(i in 1:length(Results[[2]])){  
+ PathwayList<-Results[[2]][[i]]  
+ filename <- paste(names(Results[[2]][i]),".txt", sep="", collapse="")  
+ write.table(PathwayList, file = filename, quote=F, row.names=F, sep = "\t")  
+ }
```

4 Plot running result diagram

4.1 Plot global edge correlation profile

The function `PlotGlobEdgeCorProfile` can plot global edge correlation profile for differential correlation scores of edges.

```
> #get example data  
> dataset<-GetExampleData("dataset")  
> class.labels<-GetExampleData("class.labels")  
> controlcharactor<-GetExampleData("controlcharactor")  
> #get the data for background set of edges
```

```

> edgesbackgrand<-GetEdgesBackgrandData()
> #calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor, edgesbackgrand)
>
> #plot global edge correlation profile
> PlotGlobEdgeCorProfile(EdgeCorScore)

```

Figure 1 shows the global edge correlation profile for differential correlation scores of edges.

4.2 Plot running Edge enrichment score

The function `PlotRunEnrichment` can plot running edge enrichment score for the pathway result.

```

> #get example data
> dataset<-GetExampleData("dataset")
> class.labels<-GetExampleData("class.labels")
> controlcharactor<-GetExampleData("controlcharactor")
> #get the data for background set of edges
> edgesbackgrand<-GetEdgesBackgrandData()
> #get the edge sets of pathways
> pathwayEdge.db<-GetPathwayEdgeData()
> #calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)
> #identify dysregulated pathways by using the function ESEA.Main
> #Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)
> Results<-GetExampleData("PathwayResult")
> #obtain the detail results of genes for a significant pathway
> PathwayResult<-Results[[2]][1]
>
> #Plot running edge enrichment score for the pathway result
> PlotRunEnrichment(EdgeCorScore,PathwayResult,weighted.score.type = 1)

```

Figure 2 shows the running edge enrichment score for the pathway result

4.3 Plot the pathway-result network diagram

The function `PlotPathwayGraph` can plot the pathway-result network diagram, and the edges which contribute to the pathway enrichment score are marked with red.

```

> #get example data
> dataset<-GetExampleData("dataset")
> class.labels<-GetExampleData("class.labels")
> controlcharactor<-GetExampleData("controlcharactor")
> #get the data for background set of edges
> edgesbackgrand<-GetEdgesBackgrandData()
> #get the edge sets of pathways
> pathwayEdge.db<-GetPathwayEdgeData()
> #calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)
> #identify dysregulated pathways by using the function ESEA.Main
> #Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)

```

Edge List Correlation Profile

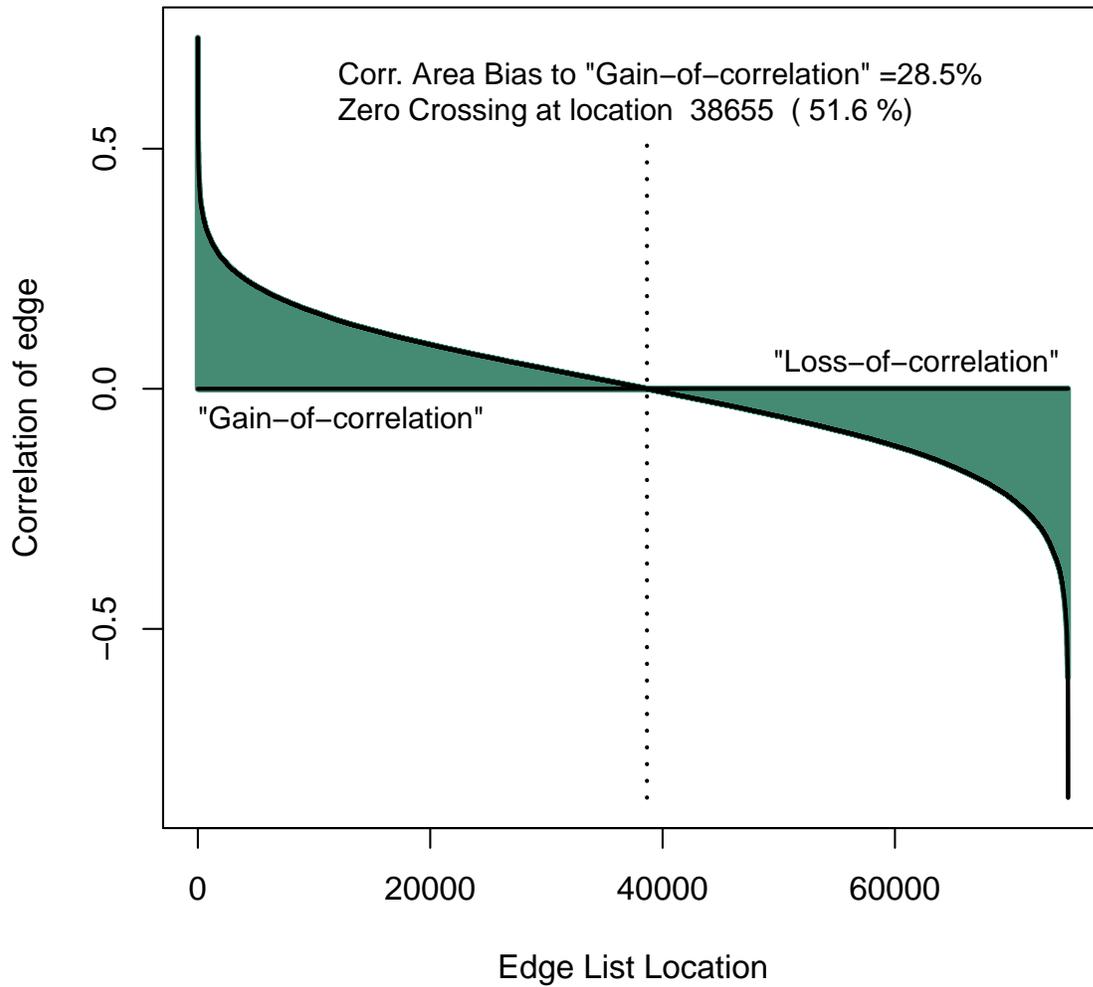


Figure 1: The visualization of global edge correlation profile for differential correlation scores of edges

Alanine, aspartate and glutamate metabolism.Gain-of-correlati

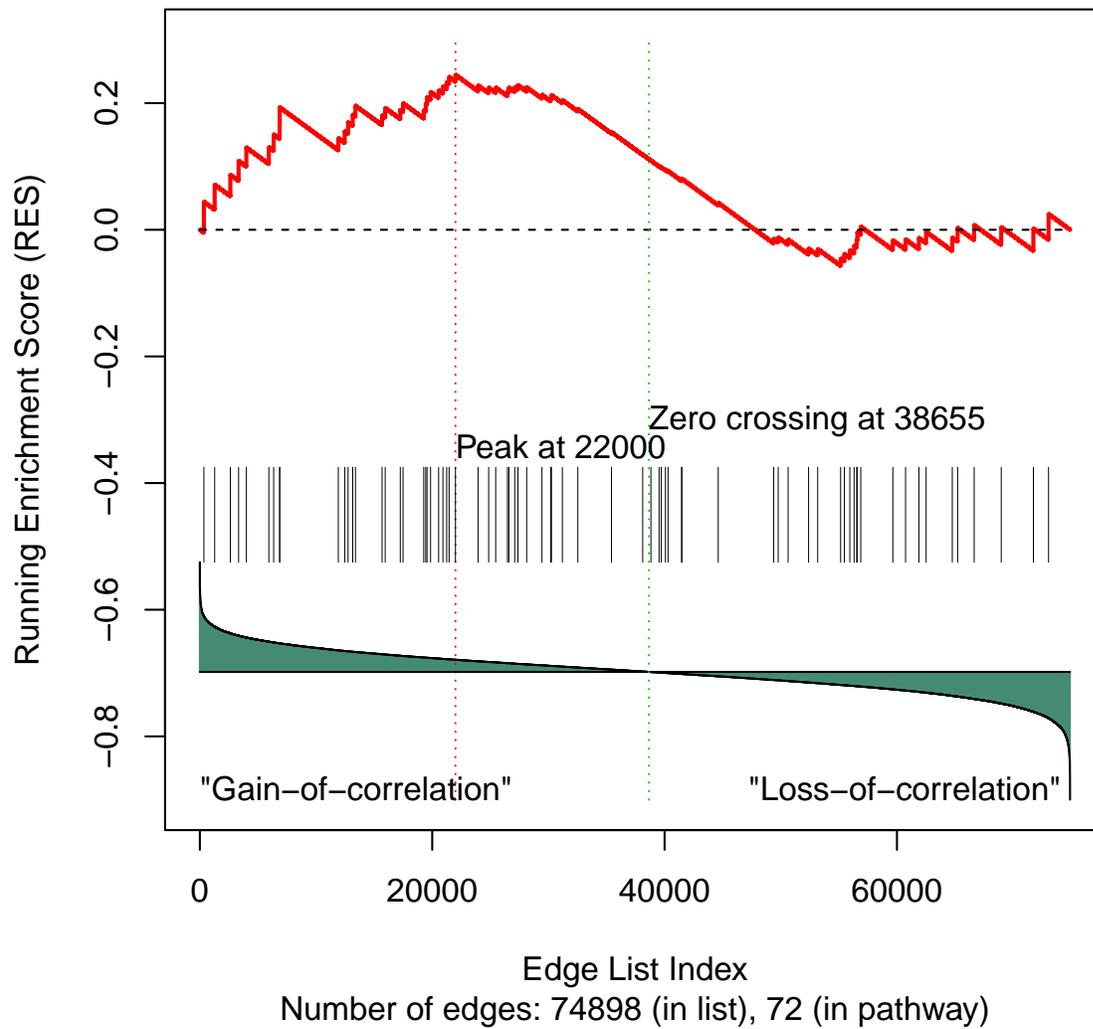


Figure 2: The visualization of the running edge enrichment score for the pathway result

```

> Results<-GetExampleData("PathwayResult")
> #obtain the detail results of genes for a significant pathway
> PathwayNetwork<-Results[[2]][[1]]
>
> #Plot the pathway-result network diagram, the edges which contribute to the ES are labeled with red.
> PlotPathwayGraph(PathwayNetwork,layout=layout.random)

```

Figure 3 shows the pathway-result network diagram, the edges which contribute to the ES are labeled with red

4.4 Save a pathway-result network to a file which can be input to the Cytoscape software

The function `SavePathway2File` can save a pathway-result network to a file which can be input to the Cytoscape software.

```

> #get example data
> dataset<-GetExampleData("dataset")
> class.labels<-GetExampleData("class.labels")
> controlcharactor<-GetExampleData("controlcharactor")
> #get the data for background set of edges
> edgesbackgrand<-GetEdgesBackgrandData()
> #get the edge sets of pathways
> pathwayEdge.db<-GetPathwayEdgeData()
> #calculate the differential correlation score for edges
> EdgeCorScore<-calEdgeCorScore(dataset, class.labels, controlcharactor,edgesbackgrand)
> #identify dysregulated pathways by using the function ESEA.Main
> #Results<-ESEA.Main(EdgeCorScore,pathwayEdge.db)
> Results<-GetExampleData("PathwayResult")
> #obtain the detail results of genes for a significant pathway
> PathwayNetwork<-Results[[2]][[1]]
> #save the pathway-result network to a file which can be input to the Cytoscape software.
> SavePathway2File(PathwayNetwork,layout=layout.circle,file="Graph")

[1] "...writing to file"

```


5 Session Info

The script runs within the following session:

R version 3.1.2 (2014-10-31)

Platform: i386-w64-mingw32/i386 (32-bit)

locale:

[1] LC_COLLATE=C

[2] LC_CTYPE=Chinese (Simplified)_People's Republic of China.936

[3] LC_MONETARY=Chinese (Simplified)_People's Republic of China.936

[4] LC_NUMERIC=C

[5] LC_TIME=Chinese (Simplified)_People's Republic of China.936

attached base packages:

[1] stats graphics grDevices utils datasets methods base

other attached packages:

[1] ESEA_1.0 parmigene_1.0.2 XML_3.98-1.1 igraph_0.7.1

loaded via a namespace (and not attached):

[1] BiocGenerics_0.12.1 graph_1.44.1 parallel_3.1.2

[4] stats4_3.1.2 tools_3.1.2

References

[Subramanian *et al.*, 2005] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. et al. (2005) Gene set enrichment analysis: a knowledgebased approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102, 15545-15550.

[Mani *et al.*, 2008] Mani, K.M., Lefebvre, C., Wang, K., Lim, W.K., Basso, K., Dalla-Favera, R. and Califano, A. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular systems biology*, 4, 169.

[de la Fuente *et al.*, 2008] de la Fuente, A. (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet*, 26, 326-333.