Ensemble of Gene Set Enrichment Analyses

Monther Alhamdoosh[*], Milica Ng and Matthew Ritchie[†]

May 15, 2016

Contents

[*] m.hamdoosh@gmail.com
[†] mritchie@wehi.edu.au

1 Introduction

The *EGSEA* package implements the Ensemble of Gene Set Enrichment Analysis (EGSEA) algorithm that utilizes the analysis results of eleven prominent GSE algorithms in the literature to calculate collective significance scores for each gene set. These methods include: ora [1], *globaltest* [2], plage [3], *safe* [4], zscore [5], *gage* [6], ssgsea [7], roast [8], *PADOG* [9], camera [10] and *GSVA* [11]. The ora, gage, camera and gsva methods depend on a competitive null hypothesis while the remaining seven methods are based on a self-contained hypothesis. Conveniently, *EGSEA* is not limited to these eleven GSE methods and new GSE tests can be easily integrated into the framework. The plage, zscore and ssgsea algorithms are implemented in the *GSVA* package and camera and roast are implemented in the *limma* package. *EGSEA* was implemented with parallel computation enabled using the *parallel* package. There are two levels of parallelism in EGSEA:(i) parallelism at the method-level and (ii) parallelism at the experimental contrast level. A wrapper function was written for each individual GSE method to utilize existing R packages and create a universal interface for all methods. The ora method was implemented using the phyper function from the *stats* package, which estimates the hypergeometric distribution for a 2×2 contingency table.

RNA-seq reads are first aligned to the reference genome and mapped reads are assigned to annotated genomic features to obtain a summarized *count matrix*. The *EGSEA* package was developed so that it can accept a count matrix or a voom object. Most of the GSE methods were intrinsically designed to work with microarray expression values and not with RNA-seq counts, hence the voom transformation is applied to the count matrix to generate an expression matrix applicable for use with these methods [12] . Since gene set tests are most commonly applied when two experimental conditions are compared, a design matrix and a contrast matrix are used to construct the experimental comparisons of interest. The target collection of gene sets is indexed so that the gene identifiers can be substituted with the indices of genes in the rows of the count matrix. The GSE analysis is then carried out by each of the selected methods independently and an FDR value is assigned to each gene set. Lastly, the ensemble functions are invoked to calculate collective significance scores for each gene set.

The *EGSEA* package also allows for performing over-representation analysis on the EGSEA gene set collections that were adopted from MSigDB, KEGG and GeneSetDB databases.

2 Citation

- Monther Alhamdoosh, Milica Ng, Nicholas J. Wilson, Julie M. Sheridan, Huy Huynh, Michael J. Wilson and Matthew E. Ritchie. Combining multiple tools outperforms individual methods in gene set enrichment analyses.

3 Installation instructions

The *EGSEA* package was developed so that it harmonizes with the existing R packages in the CRAN repository or the Bioconductor project.

3.1 System prerequistes

EGSEA does not require any software package or library to be installed before it can be installed regardless of the operating system.

3.2 R package dependencies

The *EGSEA* package depends on several R packages that are not in the Bioconductor project. These packages are listed below:

- *HTMLUtils* facilitates automated HTML report creation, in particular framed HTML pages and dynamically sortable tables. It is used in *EGSEA* to generate the stats tables. To install it, type in the R console
  install.packages(”HTMLUtils”)
- *hwriter* has easy-to-use and versatile functions to output R objects in HTML format. It is used in this package to create the HTML pages of the EGSEA report. To install it,
  install.packages(”hwriter”)
- *ggplot2* is an implementation of the grammar of graphics in R. It is used in this package to create the summary plots. To install it, type
  install.packages(”ggplot2”)
- *gplots* has various R programming tools for plotting data. It is used in *EGSEA* to create heatmaps. To install it, run
  install.packages(”gplots”)
- *stringi* allows for fast, correct, consistent, portable, as well as convenient character string/text processing in every locale and any native encoding. It is used in generating the HTML pages. To install this package, type
  install.packages(”stringi”)
- *parallel* handles running much larger chunks of computations in parallel. It is used to carry out gene set tests on parallel. It is usually installed with R.

3.2.1 Bioconductor packages

The Bioconductor packages that need to be installed in order for *EGSEA* to function properly are: *PADOG*, *GSVA*, *AnnotationDbi*, *topGO*, *pathview*, *gage*, *globaltest*, *limma*, *edgeR*, *safe*, *org.Hs.eg.db*, *org.Mm.eg.db*, *org.Rn.eg.db*. Thesea packages can be installed from Biocondcutor using the following commands in R console

```
source("http://www.bioconductor.org/biocLite.R")
biocLite(c("PADOG", "GSVA", "AnnotationDbi", "topGO", "pathview",
    "gage", "globaltest", "limma", "edgeR", "safe", "org.Hs.eg.db",
    "org.Mm.eg.db", "org.Rn.eg.db"))
```

3.2.2 Essential data package

The gene set collections that are used by *EGSEA* were preprocessed and converted into R data objects to be used by the EGSEA functions. The data objects are stored in an R package, named *EGSEAdata*. It contains the gene set collections that are used by *EGSEA* to perform gene set testing. *EGSEAdata* can be installed from Bitbucket.

To install packages from Bitbucket, *devtools* should be installed. *devtools* Package devtools is available at CRAN. For Windows this seems to depend on having Rtools for Windows installed. You can download and install this from: http://cran.r-project.org/bin/windows/Rtools/

To install *devtools*, run

```
install.packages("devtools")
```

To install *EGSEAdata*, run in R console the following commands

```
library(devtools)
install_bitbucket("malhamdoosh/egseadata", ref = "Stable_Release")
```

3.3 Installation

3.3.1 Bioconductor

In R console, type

```
source("http://bioconductor.org/biocLite.R")
biocLite("EGSEA")
```

3.3.2 Bitbucket

To install EGSEA from bitbucket, type in the R console

```
library(devtools)
install_bitbucket("malhamdoosh/egsea", ref = "Stable_Release")
```

4 Quick start

4.1 EGSEA gene set collections

The Molecular Signatures Database (MSigDB) [13] v5.0 was downloaded from http://www.broadinstitute.org/gsea/msigdb (05 July 2015, date last accessed) and the human gene sets were extracted for each collection (h, c1, c2, c3, c4, c5, c6, c7). Mouse orthologous gene sets of these MSigDB collections were adopted from http://bioinf.wehi.edu.au/software/MSigDB/index.html [10]. EGSEA uses Entrez Gene identifiers [14] and alternate gene identifiers must be first converted into Entrez IDs. KEGG pathways [15] for mouse and human were downloaded using the *gage* package. To extend the capabilities of EGSEA, a third database of gene sets was downloaded from the GeneSetDB [16] http://genesetdb.auckland.ac.nz/sourcedb.html project. In total, more than 20,000 gene sets have been collated along with annotation information for each set (where available).

The *EGSEA* package has four main functions that utilizes the gene set collections of *EGSEAdata*. They map the dataset Entrez gene IDs into the available gene sets of each collection and create indexes for each gene set collection. They also compile annotation information for each gene set to be integrated with the final EGSEA report. These functions are:

- `buildKEGGIdxEZID` indexes the KEGG pathway gene sets and loads gene set annotation. Type ?buildKEGGIdxEZID in the console to see how to use this function.
- `buildMSigDBIdxEZID` indexes the MSigDB gene sets and loads gene set annotation. Type ?buildMSigDBIdxEZID in the console to see how to use this function.
- `buildGeneSetDBIdxEZID` indexes the GeneSetDB gene sets and loads gene set annotation. Type ?buildGeneSetDBIdxEZID in the console to see how to use this function.
- `buildIdxEZID` indexes the MSigDB and KEGG gene sets and loads gene set annotation. Type ?buildIdxEZID in the console to see how to use this function.

The above functions require a vector of Entrez Gene IDs and the species name. To use the output of these functions with the *EGSEA* functions, the order of gene ids in the *entrezIDs* parameter should match that of the row names of the count matrix or the `voom` object.

4.2 A simple example

The *EGSEA* package basically performs gene set enrichment analysis on a `voom` object generated by the `voom` function from the *limma* package. It was primarily developed to extend the limma-voom RNA-seq analysis pipeline.

To quickly start with *EGSEA* analysis, an example on analyzing a human IL-13 dataset is presented here. This experiment aims to identify the biological pathways and diseases associated with the cytokine Interleukin 13 (IL-13) using gene expression measured in peripheral blood mononuclear cells (PBMCs) obtained from 3 healthy donors. The expression profiles of *in vitro* IL-13 stimulation were generated using RNA-seq technology for 3 PBMC samples at 24 hours. The transcriptional profiles of PBMCs without IL-13 stimulation were also generated to be used as controls. Finally, an IL-13Rα1 antagonist was introduced into IL-13 stimulated PBMCs and the gene expression levels after 24h were profiled to examine the neutralization of IL-13 signaling by the antagonist. Only two samples were available for the last condition. Single-end 100bp reads were obtained via RNA-seq from total RNA using a HiSeq 2000 Illumina sequencer. TopHat was used to map the reads to the human reference genome (GRCh37.p10). HTSeq was then used to summarize reads into a gene-level count matrix. The TMM method from the *edgeR* package was used to normalize the RNA-seq counts.

To perform EGSEA analysis on this dataset, the *EGSEA* package is first loaded:

```
library(EGSEA)

## Loading required package:  Biobase

## Loading required package:  BiocGenerics

## Loading required package:  parallel

##
## Attaching package:  'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ, clusterExport,
##    clusterMap, parApply, parCapply, parLapply, parLapplyLB, parRapply,
##    parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs

## The following objects are masked from 'package:base':
##
##     Filter, Find, Map, Position, Reduce, anyDuplicated, append, as.data.frame,
##     cbind, colnames, do.call, duplicated, eval, evalq, get, grep, grepl,
##     intersect, is.unsorted, lapply, lengths, mapply, match, mget, order, paste,
##     pmax, pmax.int, pmin, pmin.int, rank, rbind, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with 'browseVignettes()'.  To
##     cite Bioconductor, see 'citation("Biobase")', and for packages
##     'citation("pkgname")'.

## Loading required package:  gage

## Loading required package:  AnnotationDbi

## Loading required package:  stats4

## Loading required package:  IRanges

## Loading required package:  S4Vectors

##
## Attaching package:  'S4Vectors'

## The following objects are masked from 'package:base':
##
##     colMeans, colSums, expand.grid, rowMeans, rowSums

## Loading required package:  topGO

## Loading required package:  graph

## Loading required package:  GO.db

##

## Loading required package:  SparseM

##
## Attaching package:  'SparseM'

## The following object is masked from 'package:base':
##
##     backsolve

##
## groupGOTerms:  GOBPTerm, GOMFTerm, GOCCTerm environments built.
```

```
##
## Attaching package:  ’topGO’

## The following object is masked from ’package:IRanges’:
##
##     members

## The following object is masked from ’package:gage’:
##
##     geneData

## Loading required package:  pathview

## Loading required package:  org.Hs.eg.db

##

## ##############################################################################
## Pathview is an open source software package distributed under GNU General
## Public License version 3 (GPLv3).  Details of GPLv3 is available at
## http://www.gnu.org/licenses/gpl-3.0.html.  Particullary, users are required to
## formally cite the original Pathview paper (not just mention it) in publications
## or products.  For details, do citation("pathview") within R.
##
## The pathview downloads and uses KEGG data.  Non-academic uses may require a KEGG
## license agreement (details at http://www.kegg.jp/kegg/legal.html).
## ##############################################################################

##
## KEGG.db contains mappings based on older data because the original resource was
##  removed from the the public domain before the most recent update was produced.
##  This package should now be considered deprecated and future versions of
##  Bioconductor may not have it available.  Users who want more current data are
##  encouraged to look at the KEGGREST or reactome.db packages

##

##

##

##
```

Then, the voom data object of this experiment from *EGSEAdata* is loaded to perform the EGSEA analysis:

```
library(EGSEAdata)
data(il13.data)
v = il13.data$voom
names(v)

## [1] "genes"   "targets" "E"       "weights" "design"

v$design

##   X24 X24IL13 X24IL13Ant X40513 X40913
## 1   0       1          0      0      0
## 2   0       0          1      0      0
## 3   1       0          0      1      0
## 4   0       1          0      1      0
## 5   1       0          0      0      1
## 6   0       1          0      0      1
## 7   0       0          1      0      1
## 8   1       0          0      0      0
## attr(,"assign")
## [1] 1 1 1 2 2
## attr(,"contrasts")
## attr(,"contrasts")$`d1$samples$group`
## [1] "contr.treatment"
##
## attr(,"contrasts")$`d1$samples$Date`
## [1] "contr.treatment"
```

```
contrasts = il13.data$contra
contrasts
```

```
##             Contrasts
## Levels       X24IL13 - X24 X24IL13Ant - X24IL13
##   X24                   -1                    0
##   X24IL13                1                   -1
##   X24IL13Ant             0                    1
##   X40513                 0                    0
##   X40913                 0                    0
```

Before the EGSEA function is called gene set collection(s) needs to be preprocessed and indexed using EGSEA indexing functions that were presented earlier. For example, to use the KEGG pathway collections without the Metabolism pathways, type

```
# prepare gene set collections
gs.annots = buildIdxEZID(entrezIDs = rownames(v$E), species = "human",
    msigdb.gsets = "c5", kegg.exclude = c("Metabolism"))
```

```
## [1] "Reading Broad Gene Sets ... "
## [1] "Created the gs.annot$original for c5 \n..."
## [1] "Created the gs.annot$idx for c5 ..."
## [1] "Created the gs.annot$anno for c5 ..."
## [1] "Building KEGG pathways annotation object ... "
```

```
names(gs.annots)
```

```
## [1] "c5"   "kegg"
```

Finally, the EGSEA analysis can be invoked using the egsea function as follows

```
# perform the EGSEA analysis set display.top = 20 to display
# more gene sets. It takes longer time to run.
gsa = egsea(voom.results = v, contrasts = contrasts, gs.annots = gs.annots,
    symbolsMap = v$genes, baseGSEAs = egsea.base()[-2], display.top = 3,
    sort.by = "avg.rank", egsea.dir = "./il13-egsea-report",
    num.threads = 4)
```

```
## [1] "Log fold changes are estimated using limma package ... "
## [1] "EGSEA is running on the provided data and c5 gene sets"
## [1] "Writing out the top-ranked gene sets for each contrast .. \nC5 gene sets"
## [1] "The top gene sets for contrast X24IL13 - X24 are:"
##                                                   ID        p.adj
## INNATE_IMMUNE_RESPONSE                        M3064 1.849187e-12
## POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS  M4046 2.743478e-06
## REGULATION_OF_IMMUNE_RESPONSE                M12336 2.127039e-06
## [1] "The top gene sets for contrast X24IL13Ant - X24IL13 are:"
##                                      ID        p.adj
## INFLAMMATORY_RESPONSE           M10617 0.000000e+00
## DEFENSE_RESPONSE                 M3458 0.000000e+00
## FATTY_ACID_BIOSYNTHETIC_PROCESS M11481 8.080358e-07
## [1] "GO graphs are being generated for top-ranked GO terms based on \np-values ... "

##
## Building most specific GOs .....

## ( 9828 GO terms found.  )

##
## Build GO DAG topology ..........

## ( 13656 GO terms and 32783 relations.  )

##
## Annotating nodes ...............

## ( 11749 genes annotated to the GO terms.  )

##
## Building most specific GOs .....

## ( 3395 GO terms found.  )

##
## Build GO DAG topology ..........

## ( 3908 GO terms and 4929 relations.  )

##
## Annotating nodes ...............

## ( 11797 genes annotated to the GO terms.  )

##
## Building most specific GOs .....

## ( 1403 GO terms found.  )

##
## Build GO DAG topology ..........
## ( 1664 GO terms and 3283 relations.  )

##
## Annotating nodes ...............

## ( 12389 genes annotated to the GO terms.  )

## [1] "X24IL13 - X24"

## Loading required package:  Rgraphviz

## Loading required package:  grid

##
## Attaching package:  'grid'

## The following object is masked from 'package:topGO':
##
##    depth

##
## Attaching package:  'Rgraphviz'

## The following objects are masked from 'package:IRanges':
##
##    from, to

## The following objects are masked from 'package:S4Vectors':
##
##    from, to

## [1] "X24IL13Ant - X24IL13"
## [1] "Heat maps are being generated for top-ranked gene sets based on \nlogFC ... "
## [1] "Summary plots are being generated ... "
## /tmp/RtmpjfoAyN/Rbuild29502520932/EGSEA/vignettes/il13-egsea-report/ranked-gene-sets-fisher/js
## [1] "EGSEA is running on the provided data and kegg gene sets"
## [1] "Writing out the top-ranked gene sets for each contrast .. \nKEGG gene sets"
## [1] "The top gene sets for contrast X24IL13 - X24 are:"
##                                                   Type        p.adj
## Asthma                                         Disease 2.816639e-09
## Intestinal immune network for IgA production Signaling 7.624705e-08
## Amoebiasis                                     Disease 1.110637e-07
## [1] "The top gene sets for contrast X24IL13Ant - X24IL13 are:"
##                                          Type        p.adj
## NOD-like receptor signaling pathway  Signaling 3.635776e-06
## Asthma                                 Disease 7.619107e-07
## Toll-like receptor signaling pathway Signaling 3.155081e-08
## Pathway maps are being generated for top-ranked
##  pathways based
## on logFC ...
## [1] "    X24IL13 - X24"
## [1] "    X24IL13Ant - X24IL13"
## [1] "Heat maps are being generated for top-ranked gene sets based on \nlogFC ... "
## [1] "Summary plots are being generated ... "
## [1] "Comparison summary plots are being generated  ... "
## Pathway maps are being generated for top-ranked
##  comparative
## pathways based on logFC ...
```

```
topSets(gsa, contrast = 1, gs.label = "kegg", number = 10)
```

```
##  [1] "Asthma"
##  [2] "Intestinal immune network for IgA production"
##  [3] "Amoebiasis"
##  [4] "Viral myocarditis"
##  [5] "Endocrine and other factor-regulated calcium reabsorption"
##  [6] "Legionellosis"
##  [7] "HTLV-I infection"
##  [8] "Prion diseases"
##  [9] "Toxoplasmosis"
## [10] "Proteoglycans in cancer"
```

```
topSets(gsa, contrast = 1, gs.label = "kegg", sort.by = "ora",
    number = 10, names.only = FALSE)
```

```
##                                             p.value        p.adj vote.rank avg.rank
## Cytokine-cytokine receptor interaction 0.000000e+00 0.000000e+00         5     56.0
## Staphylococcus aureus infection        9.992007e-15 6.761258e-13        10     69.2
## Hematopoietic cell lineage             0.000000e+00 0.000000e+00         5     50.9
## Phagosome                              2.950185e-10 1.197775e-08       130     68.0
## Tuberculosis                           1.567893e-08 2.652353e-07         5     59.2
## Leishmaniasis                          7.178571e-09 1.619166e-07        20     80.0
## Cell adhesion molecules (CAMs)         4.376893e-09 1.110637e-07        10     79.0
## Rheumatoid arthritis                   5.168753e-08 7.494692e-07        10     96.0
## Asthma                                 5.550027e-11 2.816639e-09        10     26.0
## Type I diabetes mellitus               9.536817e-06 8.417278e-05       100     98.2
##                                        med.rank   min.pvalue min.rank avg.logFC Direction
## Cytokine-cytokine receptor interaction     47.0 2.687045e-16        1 0.7563135        -1
## Staphylococcus aureus infection            74.5 1.836804e-15        2 0.6701979         1
## Hematopoietic cell lineage                 14.0 8.441405e-12        2 0.7557242        -1
## Phagosome                                  64.0 1.685628e-11        4 0.5389219         1
## Tuberculosis                               58.0 1.103390e-09        4 0.4968563         1
## Leishmaniasis                              40.0 1.915394e-09        6 0.5805793         1
## Cell adhesion molecules (CAMs)             86.0 1.007610e-08        7 0.5597125         1
## Rheumatoid arthritis                       90.0 1.760956e-08        8 0.5947603        -1
## Asthma                                     13.0 1.490995e-07        9 0.7024946         1
## Type I diabetes mellitus                  110.0 2.629903e-07       10 0.6005818        -1
##                                        Significance camera safe gage padog plage zscore
## Cytokine-cytokine receptor interaction    100.00000     30   76    1     1   170     64
## Staphylococcus aureus infection            81.88529     87  114    9    62     6     93
## Hematopoietic cell lineage                 99.92208     12   16    2     2   128    203
## Phagosome                                  42.86011    129   87   26    13    41     96
## Tuberculosis                               32.80426    111    83   33    12    42     74
## Leishmaniasis                              39.58128    181    26   18    18    31    146
## Cell adhesion molecules (CAMs)             39.07866     14    98   10    81   152    113
## Rheumatoid arthritis                       36.57453    202    14    8    71   159    161
## Asthma                                     60.30259     10    70   11    39     9     21
## Type I diabetes mellitus                   24.56941    128   121   23    99    51    143
##                                          gsva ssgsea globaltest ora
## Cytokine-cytokine receptor interaction   78    118         21   1
## Staphylococcus aureus infection         134    182          3   2
## Hematopoietic cell lineage              103      6         34   3
## Phagosome                               124    137         23   4
## Tuberculosis                            119    109          4   5
## Leishmaniasis                           180    145         49   6
## Cell adhesion molecules (CAMs)          193     31         91   7
## Rheumatoid arthritis                     98    157         82   8
## Asthma                                   15     11         65   9
## Type I diabetes mellitus                146    165         96  10
```

```
topSets(gsa, contrast = "comparison", gs.label = "kegg", number = 10)
```

```
##  [1] "Asthma"
##  [2] "Viral myocarditis"
##  [3] "Amoebiasis"
##  [4] "HTLV-I infection"
##  [5] "Legionellosis"
##  [6] "NOD-like receptor signaling pathway"
##  [7] "Intestinal immune network for IgA production"
##  [8] "Cytokine-cytokine receptor interaction"
##  [9] "Malaria"
## [10] "Endocrine and other factor-regulated calcium reabsorption"
```

The egsea returns a list of elements, one for each gene set collection and one for the comparative analysis. Each element is also a list of two elements: the top.gene.sets, which stores the top *display.top* gene sets for each contrast and the test.results, which stores the stores the EGSEA test results for all gene sets along with the ensemble and individual rankings.

The EGSEA report of this experiment can be launched from ./il13-egsea-report/index.html.

Finally, the EGSEA analysis can be run with all the gene set collections that are avilable in the *EGSEAdata* package as follows

```
gs.annots = buildIdxEZID(entrezIDs = rownames(v$E), species = "human")
```

```
## [1] "Reading Broad Gene Sets ... "
## [1] "Created the gs.annot$original for h \n..."
## [1] "Created the gs.annot$idx for h ..."
## [1] "Created the gs.annot$anno for h ..."
## [1] "Created the gs.annot$original for c1 \n..."
## [1] "Created the gs.annot$idx for c1 ..."
## [1] "Created the gs.annot$anno for c1 ..."
## [1] "Created the gs.annot$original for c2 \n..."
## [1] "Created the gs.annot$idx for c2 ..."
## [1] "Created the gs.annot$anno for c2 ..."
## [1] "Created the gs.annot$original for c3 \n..."
## [1] "Created the gs.annot$idx for c3 ..."
## [1] "Created the gs.annot$anno for c3 ..."
## [1] "Created the gs.annot$original for c4 \n..."
## [1] "Created the gs.annot$idx for c4 ..."
## [1] "Created the gs.annot$anno for c4 ..."
## [1] "Created the gs.annot$original for c5 \n..."
## [1] "Created the gs.annot$idx for c5 ..."
## [1] "Created the gs.annot$anno for c5 ..."
## [1] "Created the gs.annot$original for c6 \n..."
## [1] "Created the gs.annot$idx for c6 ..."
## [1] "Created the gs.annot$anno for c6 ..."
## [1] "Created the gs.annot$original for c7 \n..."
## [1] "Created the gs.annot$idx for c7 ..."
## [1] "Created the gs.annot$anno for c7 ..."
## [1] "Building KEGG pathways annotation object ... "

gsetdb.annots = buildGeneSetDBIdxEZID(entrezIDs = rownames(v$E),
    species = "human")

## [1] "Building custom pathways annotation object ... "

gs.annots = c(gs.annots, gsetdb.annots)
names(gs.annots)

##  [1] "h"         "c1"        "c2"        "c3"        "c4"        "c5"        "c6"
##  [8] "c7"        "kegg"      "gsdbdrug" "gsdbdis"   "gsdbgo"    "gsdbpath" "gsdbreg"
```

5 Ensemble of Gene Set Enrichment Analysis

Given an RNA-seq dataset D of samples from N experimental conditions, K annotated genes $g_k(k = 1, \cdots, K)$, L experimental comparisons of interest $C_l(l = 1, \cdots, L)$, a collection of gene sets Γ and M methods for gene set enrichment analysis, the objective of a GSE analysis is to find the most relevant gene sets in Γ which explain the biological processes and/or pathways that are perturbed in expression in individual comparisons and/or across multiple contrasts simultaneously. Numerous statistical gene set enrichment analysis methods have been proposed in the literature over the past decade. Each method has its own characteristics and assumptions on the analyzed dataset and gene sets tested. In principle, gene set tests calculate a statistic for each gene individually $f(g_k)$ and then integrate these significance scores in a framework to estimate a set significance score $h(\gamma_i)$.

We propose seven statistics to combine the individual gene set statistics across multiple methods, and to rank and hence identify biologically relevant gene sets. Assume a collection of gene sets Γ, a given gene set $\gamma_i \in \Gamma$, and that the GSE analysis results of M methods on γ_i for a specific comparison (represented by ranks R_i^m and statistical significance scores p_i^m, where $m = 1, \cdots, M$ and $i = 1, \cdots, |\Gamma|$) are given. The EGSEA scores can then be devised, for each experimental comparison, as follows:

- The p-value score is the average p-value assigned to γ_i
- The minimum p-value score is the smallest p-value calculated for γ_i
- The minimum rank score of γ_i is the smallest rank assigned to γ_i
- The average ranking score is the mean rank across the M ranks
- The median ranking score is the median rank across the M ranks
- The majority voting score is the most commonly assigned bin ranking
- The significance score assigns high scores to the gene sets with strong fold changes and high statistical significance

It is worth noting that the p-value score can only be calculated under the independence assumption of individual gene set tests, and thus it is not an accurate estimate of the ensemble gene set significance, but can still be useful for ranking results. The significance score is scaled into $[0, 100]$ range for each gene set collection. To learn more about the calculation of each EGSEA score, the original paper of this work is available at Section 2.

6 EGSEA report

Since the number of annotated gene set collections in public databases continuously increases and there is a growing trend towards generating dynamic analytical tools, our software tool was developed to enable users to interactively navigate through the analysis results by generating an HTML *EGSEA Report*. The report presents the results in different ways. For example, the *Stats table* displays the top n gene sets (where n is selected by the user) for each experimental comparison and includes all calculated statistics. Hyperlinks are enabled wherever possible, to access additional information on the gene sets such as annotation information. The gene expression fold changes can be visualized using heat maps for individual gene sets (Fig. 1) or projected onto pathway maps where available (e.g. KEGG gene sets) (Fig. 2). The most significant Gene Ontology (GO) terms for each comparison can be viewed in a GO graph that shows their relationships (Fig. 3).

Additionally, EGSEA creates summary plots for each gene set collection to visualize the overall statistical significance of gene sets (Fig. 4). Two types of summary plots are generated: (i) a plot that emphasizes the gene regulation direction and the significance score of a gene set and (ii) a plot that emphasizes the set cardinality and its rank. EGSEA also generates a multidimensional scaling (MDS) plot that shows how various GSE methods rank a collection of gene sets (Fig. 5). This plot gives insights into the similarity of different methods on a given dataset. Finally, the reporting capabilities of EGSEA can be used to extend any existing or newly developed GSE method by simply using only that method.

Similar reporting capabilities are also provided for the comparative analysis results of EGSEA (Fig. 6 and Fig. 7).

6.1 Comparative analysis

Unlike most GSE methods that calculate a gene set enrichment score for a given gene set under a single experimental contrast (e.g. disease vs. control), the comparative analysis proposed here allows researchers to estimate the significance of a gene set across multiple experimental contrasts. This analysis helps in the identification of biological processes that are perturbed by multiple experimental conditions simultaneously. Comparative significance scores are calculated for a gene set.

An interesting application of the comparative analysis would be finding pathways or biological processes that are activated by a stimulation with a particular cytokine yet are completely inhibited when the cytokine's

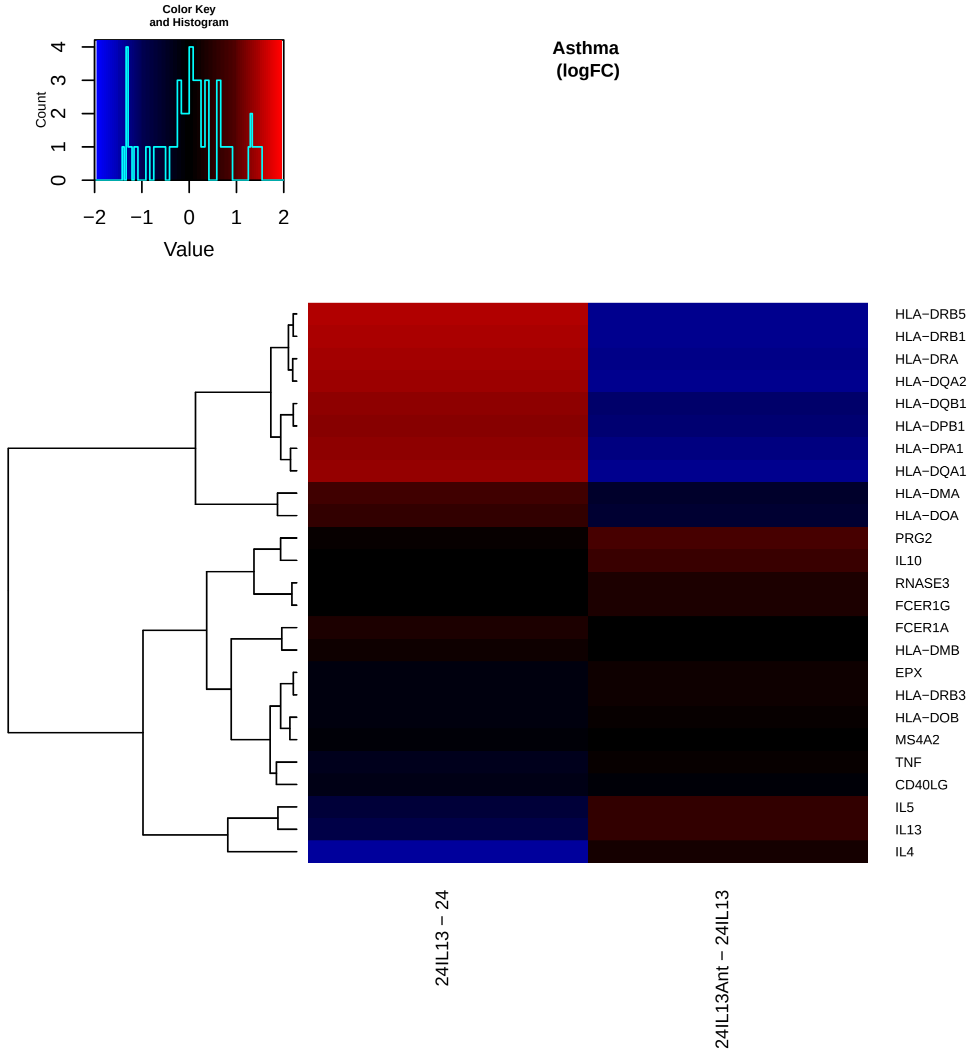

Figure 1: Asthma heatmap for the comparative analysis

receptor is blocked by an antagonist, revealing the functions uniquely associated with the signaling of that particular receptor as in the experiment below.

7 EGSEA on a non-human dataset

Epithelial cells from the mammary glands of female virgin 8-10 week-old mice were sorted into three populations of basal, luminal progenitor (LP) and mature luminal (ML) cells. Three independent samples from each population were profiled via RNA-seq on total RNA using an Illumina HiSeq 2000 to generate 100bp single-end read libraries. The *Rsubread* aligner was used to align these reads to the mouse reference genome (*mm10*) and mapped reads were summarized into gene-level counts using `featureCounts` with default settings. The raw counts are also normalized using the TMM method. Data are available from the GEO database as series

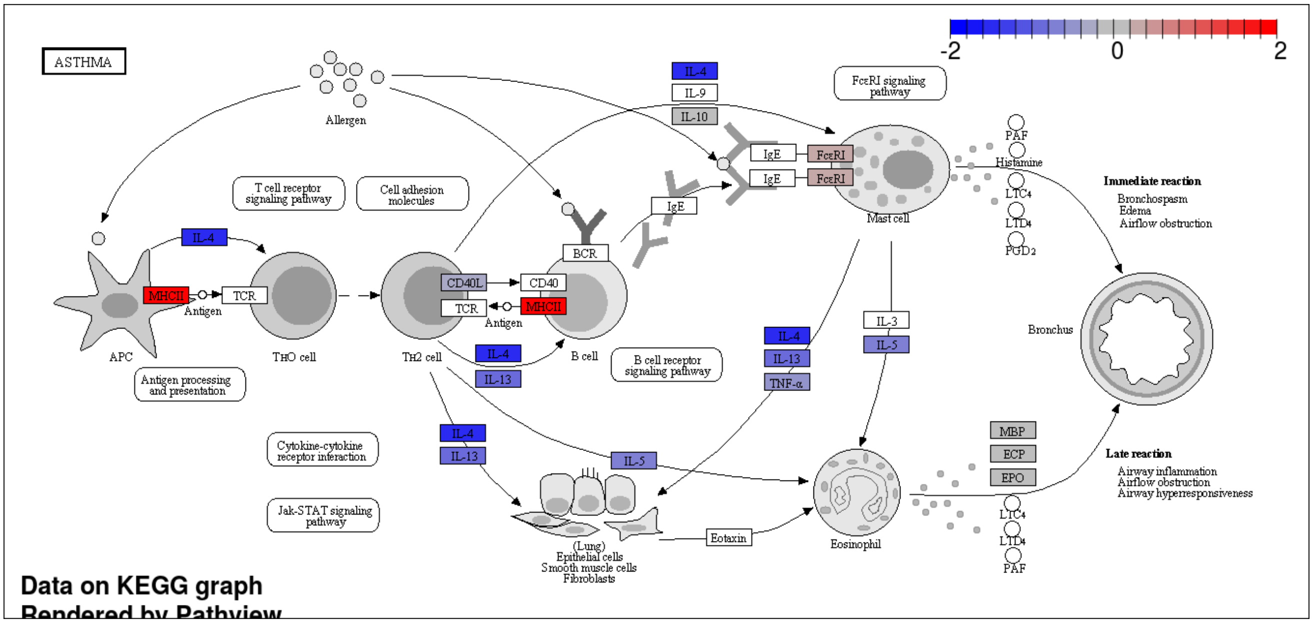

Figure 2: Asthma pathway map for the contrast X24IL13-X24

GSE63310.

To perform EGSEA analysis on this dataset, the following commands are invoked in the R console

```
# load the mammary dataset
library(EGSEA)
library(EGSEAdata)
data(mam.data)
v = mam.data$voom
names(v)
v$design
contrasts = mam.data$contra
contrasts
# build the gene set collections
gs.annots = buildIdxEZID(entrezIDs = rownames(v$E), species = "mouse",
    msigdb.gsets = "c2", kegg.exclude = "all")
names(gs.annots)
# create Entrez IDs - Symbols map
symbolsMap = v$genes[, c(1, 3)]
colnames(symbolsMap) = c("FeatureID", "Symbols")
symbolsMap[, "Symbols"] = as.character(symbolsMap[, "Symbols"])
# replace NA Symbols with IDs
na.sym = is.na(symbolsMap[, "Symbols"])
symbolsMap[na.sym, "Symbols"] = symbolsMap[na.sym, "FeatureID"]
# perform the EGSEA analysis set report = TRUE to generate
# the EGSEA interactive report
gsa = egsea(voom.results = v, contrasts = contrasts, gs.annots = gs.annots,
    symbolsMap = symbolsMap, baseGSEAs = egsea.base()[-c(2, 5,
```

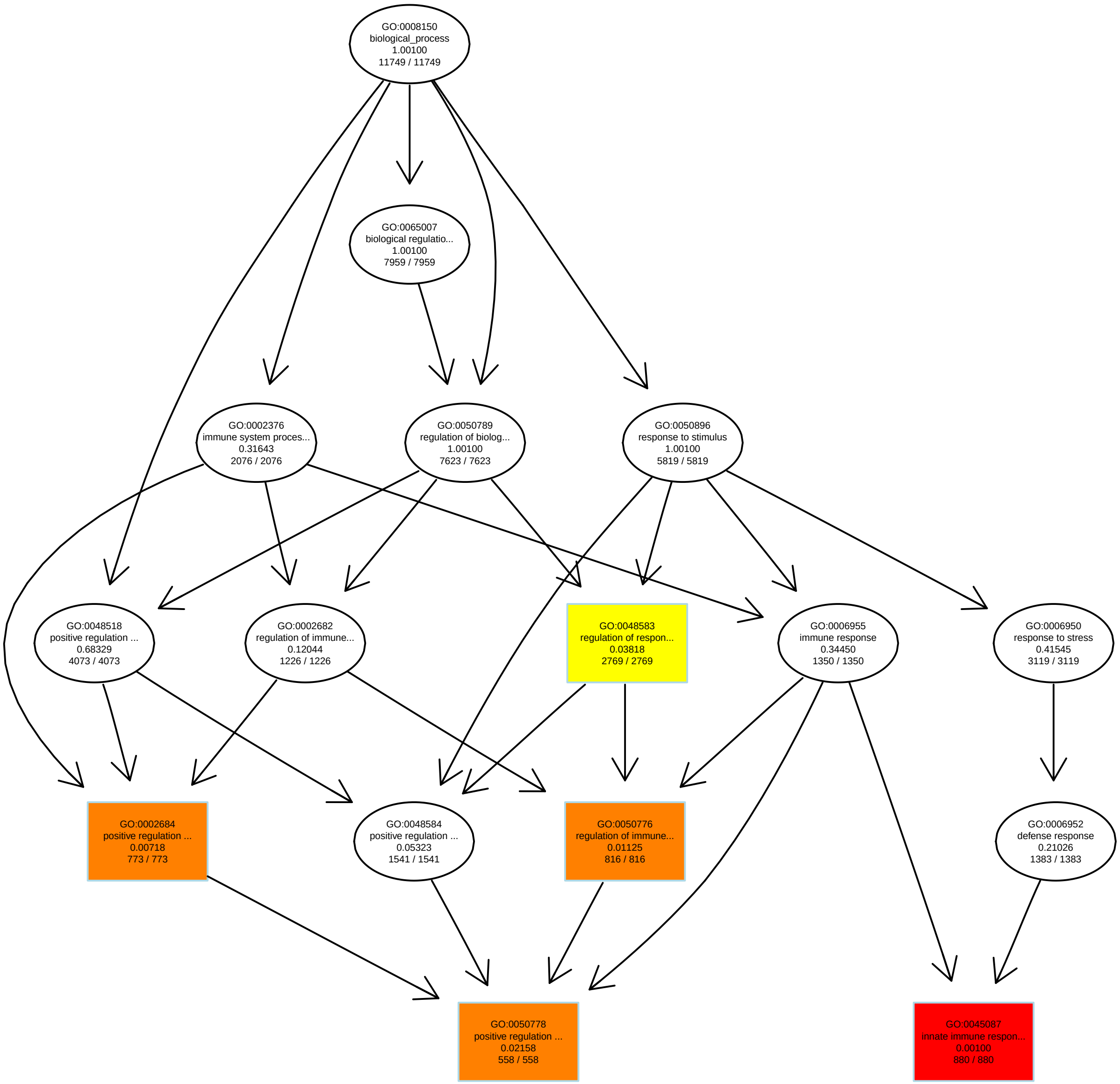

Figure 3: The top significant Biological Processes (BP) from GO terms.

```
        6, 9)], display.top = 20, sort.by = "med.rank", egsea.dir = "./mam-egsea-report",
    num.threads = 4, report = FALSE)
# show top 20 comparative gene sets in C2 collection
topSets(gsa, contrast = "comparison", gs.label = "c2", number = 20)
```

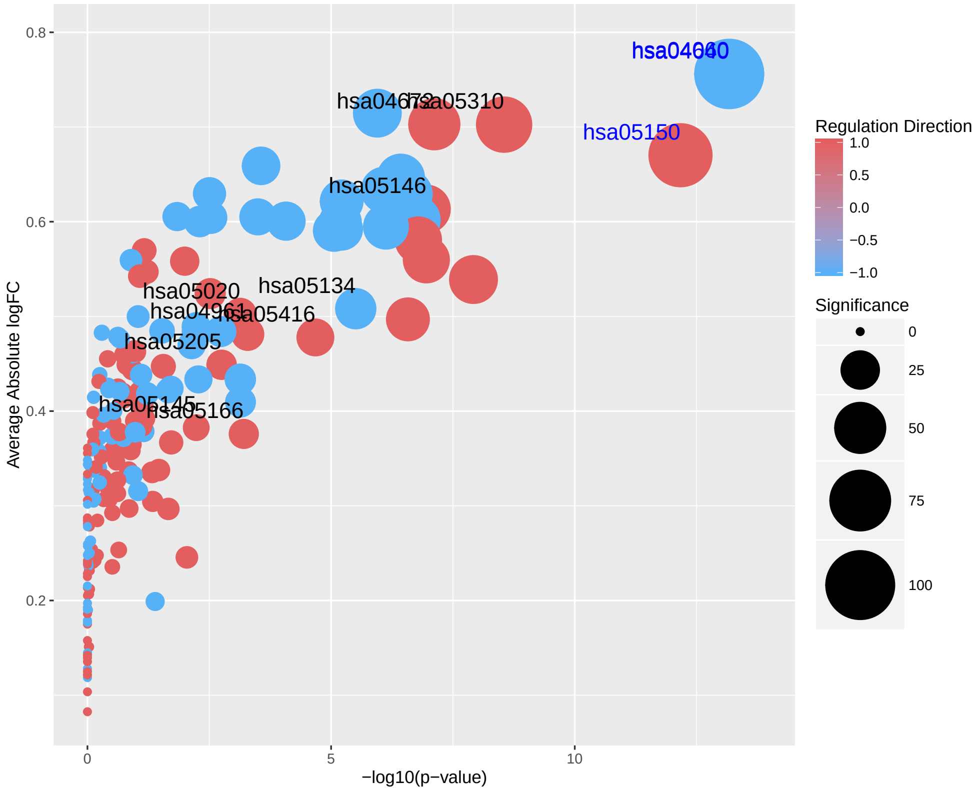

Figure 4: Summary plot for the contrast X24IL13-X24

8 EGSEA on a count matrix

The EGSEA analysis can be also performed on the count matrix directly without the need of having a `voom` object in advance. The `egsea.cnt` can be invoked on a count matrix given the group of each sample is provided with design and contrast matrices as it is illustrated in this example. This function uses the `voom` function from the *limma* pakcage to convert the RNA-seq counts into expression values.

Here, the IL-13 human dataset is reanalyzed using the count matrix.

```
# load the count matrix and other relevant data
library(EGSEAdata)
data(il13.data.cnt)
cnt = il13.data.cnt$counts
group = il13.data.cnt$group
group
design = il13.data.cnt$design
contrasts = il13.data.cnt$contra
genes = il13.data.cnt$genes
# build the gene set collections
gs.annots = buildIdxEZID(entrezIDs = rownames(cnt), species = "human",
    msigdb.gsets = "none", kegg.exclude = c("Metabolism"))
# perform the EGSEA analysis set report = TRUE to generate
# the EGSEA interactive report
```

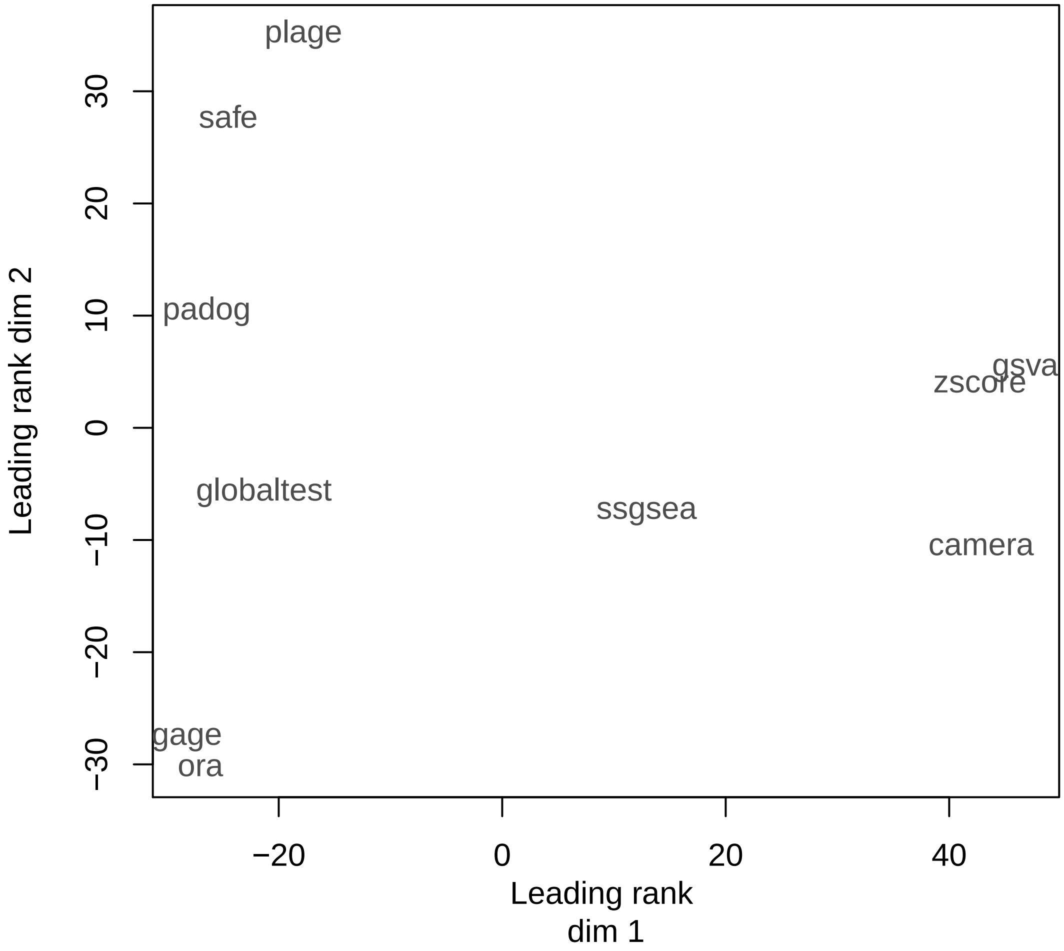

Figure 5: The performance of multiple GSE methods on the contrast X24IL13-X24.

```
gsa = egsea.cnt(counts = cnt, group = group, design = design,
    contrasts = contrasts, gs.annots = gs.annots, symbolsMap = genes,
    baseGSEAs = egsea.base()[-2], display.top = 5, sort.by = "avg.rank",
    egsea.dir = "./il13-egsea-cnt-report", num.threads = 4, report = FALSE)
```

9 EGSEA on a list of genes

Since performing simple over-representation analysis on large collections of gene sets is not readily available in Bioconductor, an ORA analysis was augmented to the *EGSEA* package so that all the reporting capabilities of EGSEA are enabled.

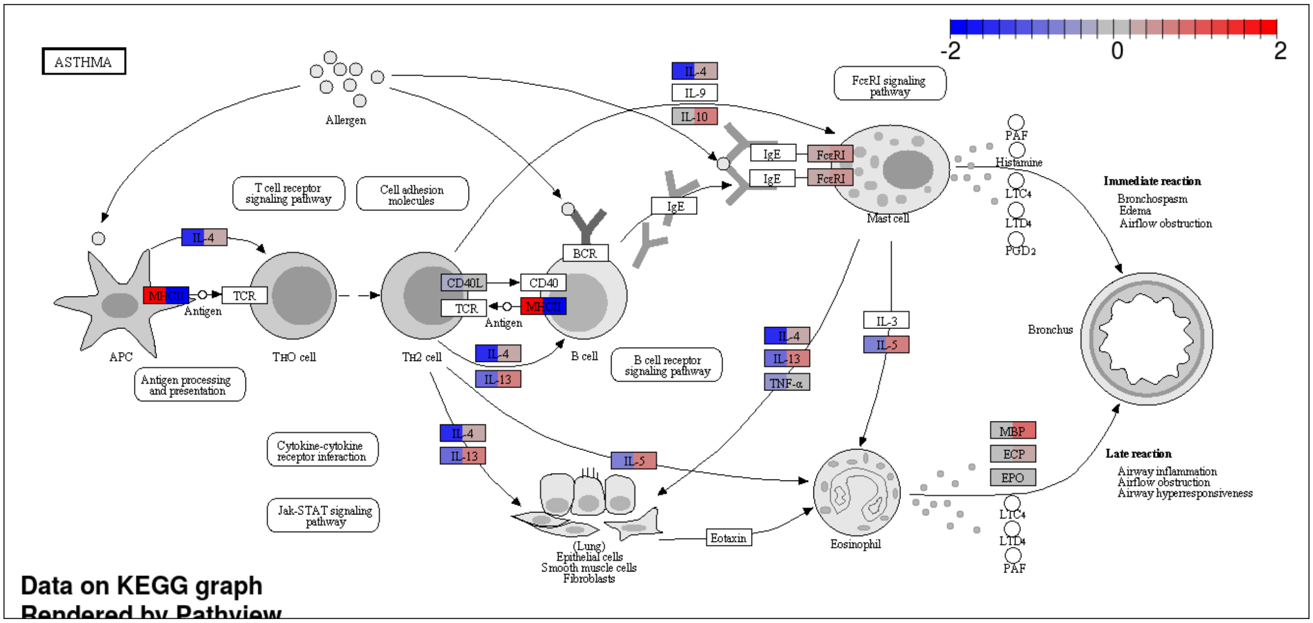

Figure 6: Asthma pathway map for the comparative analysis

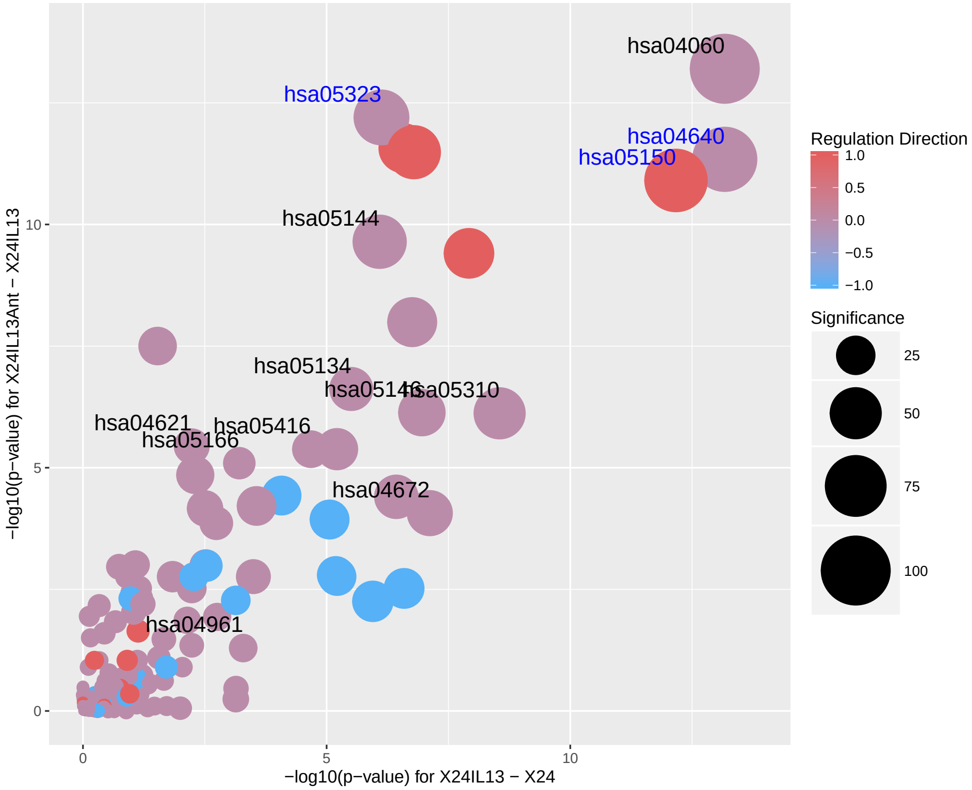

Figure 7: Summary plot for the comparative analysis

To perform ORA using the DE genes of the *X24IL13-X24* contrast from the IL-13 dataset, cut-off thresholds of p-value=0.05 and logFC = 1 are used to select a subset of DE genes. Then, the `egsea.ora` function is invoked as it is illulstrated in the following example

```
# load IL-13 dataset
library(EGSEAdata)
data(il13.data)
voom.results = il13.data$voom
contrast = il13.data$contra
# find Differentially Expressed genes
library(limma)

##
## Attaching package:  ’limma’

## The following object is masked from ’package:BiocGenerics’:
##
##    plotMA

vfit = lmFit(voom.results, voom.results$design)
vfit = contrasts.fit(vfit, contrast)
vfit = eBayes(vfit)
# select DE genes (Entrez IDs and logFC) at p-value <= 0.05
# and |logFC| >= 1
top.Table = topTable(vfit, coef = 1, number = Inf, p.value = 0.05,
    lfc = 1)
deGenes = as.character(top.Table$FeatureID)
logFC = top.Table$logFC
names(logFC) = deGenes
# build the gene set collection index
gs.annots = buildIdxEZID(entrezIDs = deGenes, species = "human",
    msigdb.gsets = "none", kegg.exclude = c("Metabolism"))

## [1] "Building KEGG pathways annotation object ... "

# perform the ORA analysis set report = TRUE to generate the
# EGSEA interactive report
gsa = egsea.ora(entrezIDs = deGenes, universe = as.character(voom.results$genes[,
    1]), logFC = logFC, title = "X24IL13-X24", gs.annots = gs.annots,
    symbolsMap = top.Table[, c(1, 2)], display.top = 5, egsea.dir = "./il13-egsea-ora-report",
    num.threads = 4, report = FALSE)

## [1] "EGSEA is running on the provided data and kegg gene sets"
## [1] "   Running ORA for X24IL13-X24"
## [1] "Running ORA on all \ncontrasts ... COMPLETED "
## [1] "Writing out the top-ranked gene sets for each contrast .. \nKEGG gene sets"
## [1] "The top gene sets for contrast X24IL13-X24 are:"
##                                              Type        p.adj
## Cytokine-cytokine receptor interaction Signaling 1.830767e-13
## Staphylococcus aureus infection          Disease 1.830767e-13
## Hematopoietic cell lineage             Signaling 3.714218e-10
## Phagosome                              Signaling 5.562572e-10
## Tuberculosis                             Disease 2.912950e-08
```

10 Non-standard gene set collections

Scientists usually have their own lists of gene sets and are interested in finding which sets are significant in the investigated dataset. Additional collections of gene sets can be easily added and tested using the EGSEA algorithm. The `buildCustomIdxEZID` function indexes newly created gene sets and attach gene set annotation if provided. To illustrate the use of this function, assume a list of gene sets is available where each gene set is represented by a character vector of Entrez Gene IDs. In this example, 50 gene sets were selected from the KEGG collection and then they were used to build a custom gene set collection index.

```
library(EGSEAdata)
data(il13.data)
v = il13.data$voom
# load KEGG pathways
data(kegg.pathways)
# select 50 pathways
gsets = kegg.pathways$human$kg.sets[1:50]
gsets[1]

## $`hsa00010 Glycolysis / Gluconeogenesis`
##  [1] "10327"  "124"    "125"    "126"    "127"    "128"    "130"    "130589" "131"
## [10] "160287" "1737"   "1738"   "2023"   "2026"   "2027"   "217"    "218"    "219"
## [19] "220"    "2203"   "221"    "222"    "223"    "224"    "226"    "229"    "230"
## [28] "2538"   "2597"   "26330"  "2645"   "2821"   "3098"   "3099"   "3101"   "3939"
## [37] "3945"   "3948"   "441531" "501"    "5105"   "5106"   "5160"   "5161"   "5162"
## [46] "5211"   "5213"   "5214"   "5223"   "5224"   "5230"   "5232"   "5236"   "5313"
## [55] "5315"   "55276"  "55902"  "57818"  "669"    "7167"   "80201"  "83440"  "84532"
## [64] "8789"   "92483"  "92579"  "9562"

# build custom gene set collection using these 50 pathways
gs.annots = buildCustomIdxEZID(entrezIDs = rownames(v$E), gsets = gsets,
    species = "human")

## [1] "Building custom pathways annotation object ... "

names(gs.annots)

## [1] "original"   "idx"          "anno"         "label"        "featureIDs" "species"
## [7] "name"

colnames(gs.annots$anno)

## [1] "ID"       "GeneSet"   "NumGenes"
```

The `buildCustomIdxEZID` creates an annotation data frame for the gene set collection if the *anno* parameter is not provided. Once the gene set collection is indexed, it can be used with any of the *EGSEA* functions: `egsea`, `egsea.cnt` or `egsea.ora`.

11 Add new GSE method

If you have an interesting gene set test method that you would like to add to the EGSEA framework, please contact us and we will be happy to add your method to the next release of *EGSEA*.

References

[1] S Tavazoie et al. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–5, 1999.

[2] Jelle J Goeman et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–9, 2004.

[3] John Tomfohr et al. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, 2005.

[4] William T Barry et al. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, 21(9):1943–9, 2005.

[5] Eunjung Lee et al. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217, 2008.

[6] Weijun Luo et al. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10:161, 2009.

[7] David A Barbie et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–12, 2009.

[8] Di Wu et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–82, 2010.

[9] Adi Laurentiu Tarca et al. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13:136, 2012.

[10] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.

[11] Sonja Hänzelmann et al. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, 2013.

[12] Charity W Law et al. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(2):R29, 2014.

[13] Aravind Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

[14] Donna Maglott et al. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database issue):D54–8, 2005.

[15] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[16] Hiromitsu Araki et al. GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, 2:76–82, 2012.