Ensemble of Gene Set Enrichment Analyses

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

July 20, 2017

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 twelve 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`, `fry` [8], *PADOG* [9], `camera` [10] and *GSVA* [11]. The `ora`, `gage`, `camera` and `gsva` methods depend on a competitive null hypothesis while the remaining eight methods are based on a self-contained null hypothesis. Conveniently, *EGSEA* is not limited to these twelve 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`, `fry` 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 and Bioconductor 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 the over-representation analysis on the EGSEA gene set collections that were adopted from MSigDB, KEGG and GeneSetDB databases.

2 Citation

- Alhamdoosh, M., Ng, M., Wilson, N. J., Sheridan, J. M., Huynh, H., Wilson, M. J., Ritchie, M. E. (2016). Combining multiple tools outperforms individual methods in gene set enrichment analyses. bioRxiv.

3 Installation instructions

The *EGSEA* package was developed so that it harmonizes with the existing R packages in the CRAN repository and 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")`
- *metap* provides a number of methods for meta-analysis of significance values. To install this package, type
  `install.packages("metap")`
- *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.
- *devtools* is needed to install packages from Bitbucket. It 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 in R console
  `install.packages("devtools")`

3.2.1 Bioconductor packages

The Bioconductor packages that need to be installed in order for *EGSEA* to work properly are: *PADOG*, *GSVA*, *AnnotationDbi*, *topGO*, *pathview*, *gage*, *globaltest*, *limma*, *edgeR*, *safe*, *org.Hs.eg.db*, *org.Mm.eg.db*, *org.Rn.eg.db*. They 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 EGSEAdata: 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 needed by *EGSEA* to perform gene set testing. *EGSEAdata* is available at Bioconductor and can be also installed from Bitbucket.

EGSEAdata can be installed from Bioconductor by running in R console the following commands

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

It can be also installed from Bitbucket from inside R console using the *devtools* package as follows

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

3.3 Installation

EGSEA can be installed from the Bioconductor project or the Bitbucket repository. We aim to only push the successfully tested versions to Bioconductor. Therefore, the Bitbucket version can have additional features that are not yet available in Bioconductor.

3.3.1 Bioconductor

To install the release version of *EGSEA* from Bioconductor, type in R console

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

To install the developmental version of *EGSEA* from Bioconductor, run the following commands in R console

```
library(BiocInstaller)
useDevel()
biocLite("EGSEA")
```

3.3.2 Bitbucket

To install the developmental version of *EGSEA* from Bitbucket, type in the R console

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

The stable release version of *EGSEA* can be obtained from Bitbucket by setting `ref = "Stabel_Release"` in the previous commands.

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 25,000 gene sets have been collated along with annotation information for each set (where available).

The *EGSEA* package has four indexing functions that utilize the gene set collections of *EGSEAdata*. They map the Entrez gene IDs of the input dataset into the gene sets of each collection and create an index for each collection. These fucntions also extract annotation information from *EGSEAdata* for each gene set to be displayed within the EGSEA HTML report. These functions are as follow

- `buildKEGGIdx` builds an index for the KEGG pathways collection and loads gene set annotation. Type ?buildKEGGIdx in the console to see how to use this function.
- `buildMSigDBIdx` builds indexes for the MSigDB gene set collections and loads gene set annotation. Type ?buildMSigDBIdx in the console to see how to use this function.
- `buildGeneSetDBIdx` builds indexes for the GeneSetDB collections and loads gene set annotation. Type ?buildGeneSetDBIdx in the console to see how to use this function.
- `buildIdx` is one-step method to build indexes for collections selected from the KEGG, MSigDB and GeneSetDB databases. Type ?buildIdx in the console to see how to use this function.

These four functions take a vector of Entrez Gene IDs and the species name and return an object (or list of objects) of class *GSCollectionIndex*.

Note: To use the *GSCollectionIndex* objects with the other *EGSEA* functions, the order of input ids vector should match that of the row names of the count matrix or the `voom` object.

4.2 EGSEA on a human dataset

The *EGSEA* package basically performs gene set enrichment analysis on a `voom` object generated by the `voom` function from the *limma* package. Actually, 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. Data are available from the GEO database www.ncbi.nlm.nih.gov/geo/ as series GSE79027.

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

```
library(EGSEA)
```

Then, the voom data object of this experiment is loaded from *EGSEAdata* as follows

```
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
```

A detailed explanation on how a voom can be created from a raw RNA-seq count matrix can be found in this workflow article [17].

Before the EGSEA pipeline is invoked, gene set collections need to be pre-processed and indexed using the EGSEA indexing functions as it was mentioned earlier. Here, indexes for the KEGG pathways and the c5 collection from the MSigDB are created as follows

```
gs.annots = buildIdx(entrezIDs = rownames(v$E), species = "human",
    msigdb.gsets = "c5", kegg.exclude = c("Metabolism"))
```

```
## [1] "Loading MSigDB Gene Sets ... "
## [1] "Loaded gene sets for the collection c5 ..."
## [1] "Indexed the collection c5 ..."
## [1] "Created annotation for the collection c5 ..."
## [1] "Building KEGG pathways annotation object ... "
```

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

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

The `gs.annots` is a list of two objects of class *GSCollectionIndex* that are labelled with "kegg" and "c5".

A quick summary of the collection indexes can be displayed using the `summary` function as follows

```
summary(gs.annots$kegg)
```

```
## KEGG Pathways (kegg): 203 gene sets - Version: NA, Update date: 07 March 2017
```

```
summary(gs.annots$c5)
```

```
## c5 GO Gene Sets (c5): 6166 gene sets - Version: 5.2, Update date: 07 March 2017
```

This shows the name, label and number of gene sets in the KEGG collection. Next, we select the base methods of the EGSEA analysis

```
baseMethods = egsea.base()[-c(2, 12)]
baseMethods
```

```
##  [1] "camera"     "safe"       "gage"       "padog"      "plage"      "zscore"
##  [7] "gsva"       "ssgsea"     "globaltest" "ora"
```

Another important parameter for the EGSEA analysis is the `sort.by` argument which determines how the gene sets are ordered. The possible values of this argument can be seen as follows

```
egsea.sort()
```

```
##  [1] "p.value"       "p.adj"         "avg.rank"      "med.rank"      "min.rank"
##  [6] "min.pvalue"    "vote.rank"     "avg.logfc.dir" "avg.logfc"     "direction"
## [11] "significance"  "camera"        "roast"         "safe"          "gage"
## [16] "padog"         "plage"         "zscore"        "gsva"          "ssgsea"
## [21] "globaltest"    "ora"           "fry"
```

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

```
# perform the EGSEA analysis set report = TRUE to generate
# HTML report.  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 = baseMethods, egsea.dir = "./il13-egsea-report",
    sort.by = "avg.rank", num.threads = 4, report = FALSE)
```

```
## [1] "EGSEA analysis has started"
```

```
## ##------ Thu Jul 20 21:23:13 2017 ------##
## [1] "Log fold changes are estimated using limma package ... "
## [1] "limma DE analysis is carried out ... "
## [1] "EGSEA is running on the provided data and c5 collection"
## ..camera*..safe*..gage*..padog*...plage*...zscore*...gsva*...ssgsea*..globaltest*..ora*
## [1] "EGSEA is running on the provided data and kegg collection"
## ..camera*..safe*..gage*..padog*...plage*...zscore*...gsva*...ssgsea*..globaltest*..ora*
## ##------ Thu Jul 20 21:39:37 2017 ------##
## [1] "EGSEA analysis took 983.78 seconds."
## [1] "EGSEA analysis has completed"
```

The function egsea returns an object of class *EGSEAResults*, which is described next in Section 5. To generate an HTML report of the EGSEA analysis results, you need to set `report=TRUE`. Then, the EGSEA report can be launched by opening ./il13-egsea-report/index.html. A quick summary of the top ten significant gene sets from each collection and for each contrast including the comparative analysis, if there are more than one contrast, can be displayed using the `summary` function as follows

```
summary(gsa)
```

```
## **** Top 10 gene sets in the c5 GO Gene Sets collection ****
## ** Contrast X24IL13-X24 **
## GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE_MEMBRANE | GO_CLATHRIN_COATED_VESICLE_MEMBRANE
## GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE | GO_ICOSANOID_BIOSYNTHETIC_PROCESS
## GO_FATTY_ACID_DERIVATIVE_BIOSYNTHETIC_PROCESS | GO_UNSATURATED_FATTY_ACID_BIOSYNTHETIC_PROCESS
## GO_POSITIVE_REGULATION_OF_CYTOKINE_SECRETION | GO_MHC_CLASS_II_PROTEIN_COMPLEX
## GO_MHC_CLASS_II_RECEPTOR_ACTIVITY | GO_LEUKOTRIENE_METABOLIC_PROCESS
##
## ** Contrast X24IL13Ant-X24IL13 **
## GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE_MEMBRANE | GO_POSITIVE_REGULATION_OF_NF_KAPPAB_IMPORT_INT
## GO_MHC_CLASS_II_PROTEIN_COMPLEX | GO_CLATHRIN_COATED_VESICLE_MEMBRANE
## GO_IGG_BINDING | GO_POSITIVE_REGULATION_OF_ACUTE_INFLAMMATORY_RESPONSE
## GO_CXCR_CHEMOKINE_RECEPTOR_BINDING | GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE
## GO_POSITIVE_REGULATION_OF_INTERLEUKIN_1_SECRETION | GO_MHC_CLASS_II_RECEPTOR_ACTIVITY
##
## ** Comparison analysis **
## GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE_MEMBRANE | GO_CLATHRIN_COATED_VESICLE_MEMBRANE
## GO_CLATHRIN_COATED_ENDOCYTIC_VESICLE | GO_MHC_CLASS_II_PROTEIN_COMPLEX
## GO_POSITIVE_REGULATION_OF_NF_KAPPAB_IMPORT_INTO_NUCLEUS | GO_MHC_CLASS_II_RECEPTOR_ACTIVITY
## GO_POSITIVE_REGULATION_OF_CYTOKINE_SECRETION | GO_POSITIVE_REGULATION_OF_ACUTE_INFLAMMATORY_RE
## GO_IGG_BINDING | GO_ANTIGEN_BINDING
##
## **** Top 10 gene sets in the KEGG Pathways collection ****
## ** Contrast X24IL13-X24 **
## Amoebiasis | Asthma
## Intestinal immune network for IgA production | Endocrine and other factor-regulated calcium re
## Viral myocarditis | HTLV-I infection
## Prion diseases | Proteoglycans in cancer
## Hematopoietic cell lineage | Legionellosis
```

```
##
## ** Contrast X24IL13Ant-X24IL13 **
## Malaria | Viral myocarditis
## NOD-like receptor signaling pathway | Toll-like receptor signaling pathway
## Asthma | Legionellosis
## Hematopoietic cell lineage | Rheumatoid arthritis
## HTLV-I infection | Melanoma
##
## ** Comparison analysis **
## Asthma | Viral myocarditis
## Intestinal immune network for IgA production | Amoebiasis
## HTLV-I infection | NOD-like receptor signaling pathway
## Hematopoietic cell lineage | Malaria
## Legionellosis | Toll-like receptor signaling pathway
```

To run the EGSEA analysis with all the gene set collections that are avilable in the *EGSEAdata* package, use the `buildIdx` function to create the gene set indexes as follows

```
gs.annots = buildIdx(entrezIDs = rownames(v$E), species = "human",
    gsdb.gsets = "all")
```

```
## [1] "Loading MSigDB Gene Sets ... "
## [1] "Loaded gene sets for the collection h ..."
## [1] "Indexed the collection h ..."
## [1] "Created annotation for the collection h ..."
## [1] "Loaded gene sets for the collection c1 ..."
## [1] "Indexed the collection c1 ..."
## [1] "Created annotation for the collection c1 ..."
## [1] "Loaded gene sets for the collection c2 ..."
## [1] "Indexed the collection c2 ..."
## [1] "Created annotation for the collection c2 ..."
## [1] "Loaded gene sets for the collection c3 ..."
## [1] "Indexed the collection c3 ..."
## [1] "Created annotation for the collection c3 ..."
## [1] "Loaded gene sets for the collection c4 ..."
## [1] "Indexed the collection c4 ..."
## [1] "Created annotation for the collection c4 ..."
## [1] "Loaded gene sets for the collection c5 ..."
## [1] "Indexed the collection c5 ..."
## [1] "Created annotation for the collection c5 ..."
## [1] "Loaded gene sets for the collection c6 ..."
## [1] "Indexed the collection c6 ..."
## [1] "Created annotation for the collection c6 ..."
## [1] "Loaded gene sets for the collection c7 ..."
## [1] "Indexed the collection c7 ..."
## [1] "Created annotation for the collection c7 ..."
## [1] "Loading GeneSetDB Gene Sets ... "
## [1] "Created the GeneSetDB Gene Sets collection ... "
```

```
## 56 gene sets from the GeneSetDB gsdbgo collection do not have valid GO ID.
## They will be removed.
## [1] "Building KEGG pathways annotation object ... "
```

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

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

5 S4 classes and methods

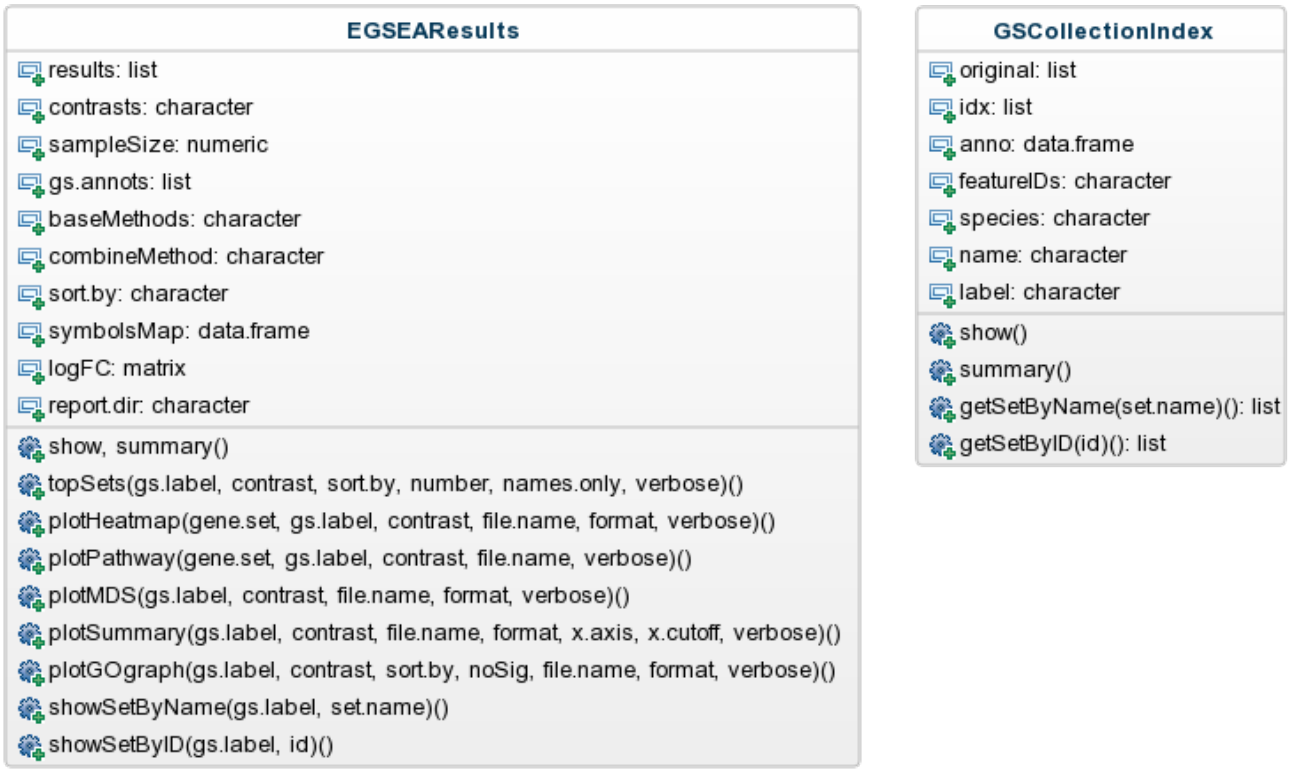


Figure 1: The S4 classes and methods of *EGSEA*.

EGSEA implements two S4 classes to perform its functionalities efficiently. The *GSCollectionIndex* stores an indexed gene set collection, which can be used to perform an EGSEA analysis, and the *EGSEAResults* stores the results of an EGSEA analysis. Each class has several slots and S4 methods to enable the user explore EGSEA results efficiently and effectively (Figure 1).

The *GSCollectionIndex* class has seven slots and four S4 methods that are defined as follows

- *GSCollectionIndex* slots:
  - `original` is a list of character vectors, each stores the Entrez Gene IDs of a gene set.
  - `idx` is a list of character vectors, each stores only the indexes of the mapped genes of a set.
  - `anno` is a data frame that stores additional annotation for each gene set.

  - `featureIDs` is a character vector of the Entrez Gene IDs that were used to index the gene sets.
  - `species` is a character that stores the species name. It accepts
  - `name` is a character that stores a short description of the gene set collection.
  - `label` is a character that stores a label for the collection to identify it from other collections when multiple collections are used for an EGSEA analysis.
- *GSCollectionIndex* S4 methods:
  - `show` displays the content of the gene set collection.
  - `summary` displays a brief summary of the gene set collection.
  - `getSetByName` returns a list of the details of gene sets given their names.
  - `getSetByID` returns a list of the details of gene sets given their IDs.

The *EGSEAResults* class has eleven slots and ten S4 methods that are defined as follows

- *EGSEAResults* slots:
  - `results` is a list that stores the EGSEA analysis results in a hierarchical format (Figure 2). The *comparison* element only exists when more than one contrast are analyzed. The *ind.results* only exists if the EGSEA function argument `print.base = TRUE`.
  - `limmaResults` is a limma linear fit model. This is only defined when `keep.limma = TRUE`.
  - `contrasts` is a character vector of contrast names.
  - `sampleSize` is a numeric value of the number of samples.
  - `gs.annots` is a list of objects of class *GSCollectionIndex* that stores the indexed gene set collections that were used in the EGSEA analysis.
  - `baseMethods` is a character vector of the base GSE methods that were used in the EGSEA analysis.
  - `baseInfo` is a list that stores additional information on the base methods (e.g., version).
  - `combineMethod` is a character value of the name of p-value combinging method.
  - `sort.by` is a character value of the sorting EGSEA score.
  - `symbolsMap` is a data frame of two columns that stores the mapping of the Entrez Gene IDs to their gene symbols.
  - `logFC` is a matrix of the calculated (or provided) logFC values where columns correspond to contrasts and rows correspond to genes.
  - `report` is a logical value indicates whether an HTML report was generated or not.
  - `report.dir` is a character value of the HTML report directory (if it was generated).
- *EGSEAResults* S4 methods:
  - `show` displays the parameters of the EGSEAResults object.
  - `summary` displays a brief summary of the EGSEA analysis results.
  - `topSets` extracts a table of the top-ranked gene sets from an EGSEA analysis.
  - `plotSummary` generates a Summary plot of an EGSEA analysis for a given gene set collection and a selected contrast.
  - `plotMethods` generates a multi-dimensional scaling (MDS) plot for the gene set rankings of the base methods of an EGSEA analysis.
  - `plotHeatmap` generates a heatmap of fold changes for a selected gene set.
  - `plotPathway` generates a visual map for a selected KEGG pathway.
  - `showSetByName` displays the details of gene sets given their names and their collection.
  - `showSetByID` displays the details of gene sets given their IDs and their collection.
  - `limmaTopTable` returns a dataframe of the top table of the limma analysis for a given contrast. This is only defined when `keep.limma = TRUE`.
  - `getlimmaResults` returns the linear model fit produced by `limma::eBayes`. This is only defined when `keep.limma = TRUE`.
  - `getSetScores` returns a dataframe of the gene set enrichment scores per sample. This can

be only calculated using specific base methods, namely, "ssgsea". This is only defined when `keep.set.scores = TRUE`.

– `plotSummaryHeatmap` generates a summary heatmap for the top n gene sets of the comparative analysis of multiple contrasts.

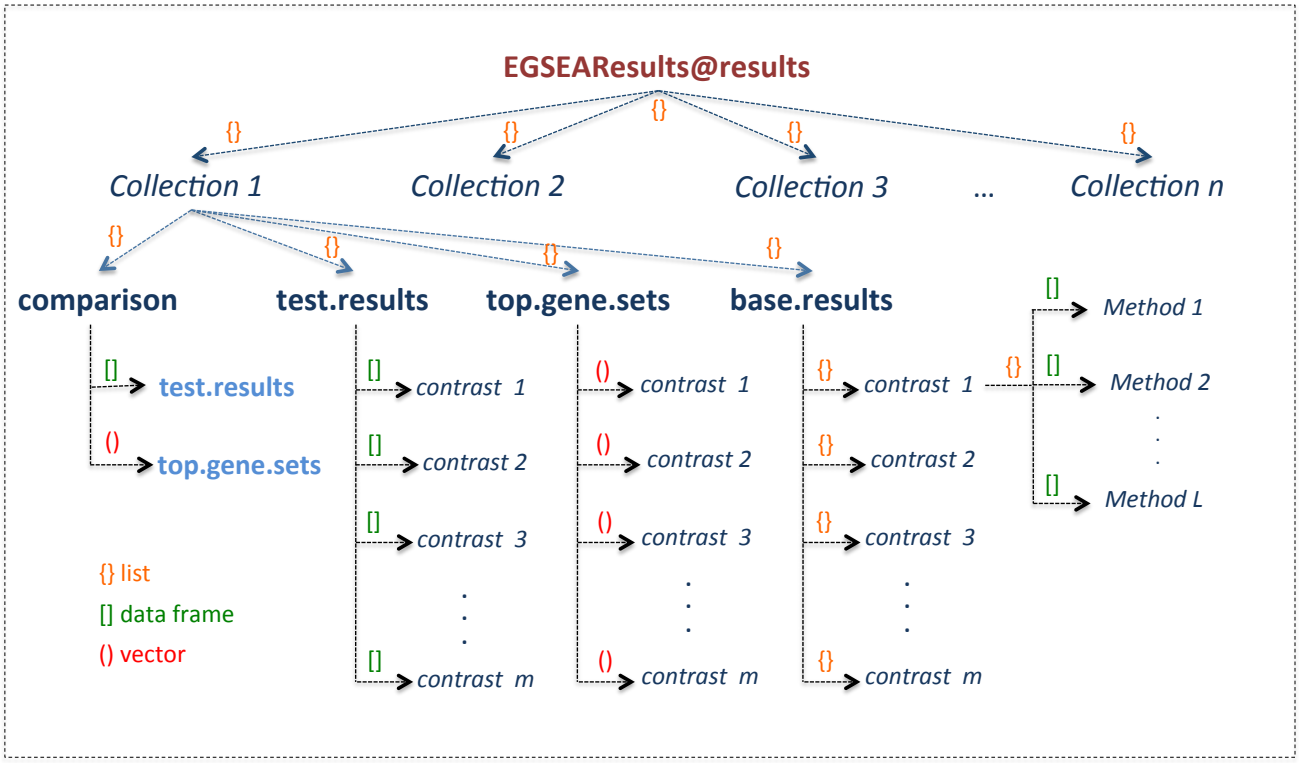


Figure 2: The structure of the slot `results` of the class *EGSEAResults*.

Next, we show how these different methods can be used to query the EGSEA results. To obtain a quick overview of the parameters of IL-13 EGSEA analysis

```
show(gsa)
```

```
## An object of class "EGSEAResults"
##  Total number of genes: 17343
##  Total number of samples: 8
##  Contrasts: X24IL13-X24, X24IL13Ant-X24IL13
##  Base GSE methods: camera (limma:3.32.3), safe (safe:3.16.0), gage (gage:2.26.1), padog (PADOG
##  P-values combining method: wilkinson
##  Sorting statistic: avg.rank
##  Organism: Homo sapiens
##  HTML report generated: No
##  Tested gene set collections:
##  c5 GO Gene Sets (c5): 6166 gene sets - Version: 5.2, Update date: 07 March 2017
##  KEGG Pathways (kegg): 203 gene sets - Version: NA, Update date: 07 March 2017
##  EGSEA version: 1.4.1
##  EGSEAdata version: 1.4.0
## Use summary(object) and topSets(object, ...) to explore this object.
```

The EGSEA analysis results can be queried in different ways. For example, the top 10 gene sets of the KEGG collection for the contrast X24IL13-X24 can be retrieved as follows

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

```
## Extracting the top gene sets of the collection
## KEGG Pathways for the contrast X24IL13-X24
##  Sorted by avg.rank
##  [1] "Amoebiasis"
##  [2] "Asthma"
##  [3] "Intestinal immune network for IgA production"
##  [4] "Endocrine and other factor-regulated calcium reabsorption"
##  [5] "Viral myocarditis"
##  [6] "HTLV-I infection"
##  [7] "Prion diseases"
##  [8] "Proteoglycans in cancer"
##  [9] "Hematopoietic cell lineage"
## [10] "Legionellosis"
```

Here the gene sets are ordered based on the value of the argument `sort.by` when the EGSEA analysis was invoked, i.e., the *avg.rank* in this example. However, the top gene sets based on a selected EGSEA score, e.g. ORA ranking, can be retreieved as follows

```
t = topSets(gsa, contrast = 1, gs.label = "c5", sort.by = "ora",
    number = 10, names.only = FALSE)
```

```
## Extracting the top gene sets of the collection
## c5 GO Gene Sets for the contrast X24IL13-X24
##  Sorted by ora
```

```
t
```

```
##                                                 Rank      p.value        p.adj vote.rank
## GO_IMMUNE_RESPONSE                                 1 4.223452e-30 2.602913e-26         5
## GO_IMMUNE_SYSTEM_PROCESS                           2 6.084844e-26 1.875045e-22      6050
## GO_DEFENSE_RESPONSE                                3 1.786982e-25 3.671056e-22      1555
## GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS             4 4.501501e-21 6.935687e-18      3925
## GO_EXTRACELLULAR_SPACE                             5 1.407154e-20 1.445381e-17      5555
## GO_INNATE_IMMUNE_RESPONSE                          6 1.016920e-19 8.953256e-17      1590
## GO_RESPONSE_TO_CYTOKINE                            7 2.449741e-19 1.887219e-16      5075
## GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS          8 1.327858e-18 8.183590e-16        10
## GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS     9 3.646334e-18 2.042942e-15        10
## GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS   10 3.146056e-17 1.491473e-14      3825
##                                                 avg.rank med.rank   min.pvalue min.rank
## GO_IMMUNE_RESPONSE                                2498.1   2516.5 4.692724e-31        1
## GO_IMMUNE_SYSTEM_PROCESS                          3284.3   3411.5 6.760938e-27        2
## GO_DEFENSE_RESPONSE                               2596.1   2536.5 1.985535e-26        3
## GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS            3188.1   3081.5 5.001667e-22        4
## GO_EXTRACELLULAR_SPACE                            3658.1   4299.5 1.563504e-21        5
## GO_INNATE_IMMUNE_RESPONSE                         2004.8   1986.5 1.016920e-20        4
## GO_RESPONSE_TO_CYTOKINE                           3015.3   3340.5 2.721934e-20        4
```

```
## GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS          2266.1   1478.0 1.327858e-19        3
## GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS     3151.2   3631.5 4.051483e-19        9
## GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS    3166.3   2804.0 3.495618e-18       10
##                                                  avg.logfc avg.logfc.dir direction
## GO_IMMUNE_RESPONSE                                1.656540      1.775304        Up
## GO_IMMUNE_SYSTEM_PROCESS                          1.558363      1.632048        Up
## GO_DEFENSE_RESPONSE                               1.666075      1.786638        Up
## GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS            1.493355      1.485645        Up
## GO_EXTRACELLULAR_SPACE                            1.851236      1.942962        Up
## GO_INNATE_IMMUNE_RESPONSE                         1.806955      2.010952        Up
## GO_RESPONSE_TO_CYTOKINE                           1.946023      2.173693        Up
## GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS         1.991152      2.204445        Up
## GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS    1.600137      1.657213        Up
## GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS   1.545729      1.463488        Up
##                                                  significance camera safe gage padog plage
## GO_IMMUNE_RESPONSE                                  100.00000   1962 1174 5644     1  3071
## GO_IMMUNE_SYSTEM_PROCESS                             79.88931   6049 1127 5221   917  2602
## GO_DEFENSE_RESPONSE                                  84.26414   1553 2353 5367  1136  2720
## GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS               60.46068   3923 1150 5545   380  2240
## GO_EXTRACELLULAR_SPACE                               73.55714   5555 1185 5973  3165  4906
## GO_INNATE_IMMUNE_RESPONSE                            68.42096   1589 2384    4   918  2601
## GO_RESPONSE_TO_CYTOKINE                              72.19984   5074 1228 5791     4  2031
## GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS            70.88090   6077 1252    9     3  1704
## GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS       55.46152   5832 1129 5231     9  2605
## GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS      50.42693   3823 1192 5469   385  1785
##                                                  zscore gsva ssgsea globaltest ora
## GO_IMMUNE_RESPONSE                                 3293 3995   4636       1204   1
## GO_IMMUNE_SYSTEM_PROCESS                           4221 5900   5488       1316   2
## GO_DEFENSE_RESPONSE                                3308 4014   4341       1166   3
## GO_REGULATION_OF_IMMUNE_SYSTEM_PROCESS             5790 5731   5902       1216   4
## GO_EXTRACELLULAR_SPACE                             4066 5787   4533       1406   5
## GO_INNATE_IMMUNE_RESPONSE                          3565 3835   4216        930   6
## GO_RESPONSE_TO_CYTOKINE                            4650 5084   5568        716   7
## GO_CELLULAR_RESPONSE_TO_CYTOKINE_STIMULUS          3888 4513   4554        653   8
## GO_POSITIVE_REGULATION_OF_RESPONSE_TO_STIMULUS     4658 5747   5289       1003   9
## GO_POSITIVE_REGULATION_OF_IMMUNE_SYSTEM_PROCESS    6158 5681   5924       1236  10
```

This can be useful to identify over-represented GO terms since GO gene sets in the c5 collection are based on ontologies which do not necessarily comprise co-regulated genes. More information on the first gene set can be retrieved as follows

```
showSetByName(gsa, "c5", rownames(t)[1])
```

```
## ID: M14329
## GeneSet: GO_IMMUNE_RESPONSE
## BroadUrl: http://www.broadinstitute.org/gsea/msigdb/cards/GO_IMMUNE_RESPONSE.html
## Description: Any immune system process that functions in the calibrated response of an organis
## PubMedID:
## NumGenes: 829/1100
```

```
## Contributor: Gene Ontology
## Ontology: BP
## GOID: GO:0006955
```

The *NumGenes* shows the number of your dataset genes that were mapped to this gene set out of the total number of genes in the set. This ratio mainly depends on the filtering criteria that are used for constructing the count matrix.

Similarly, the top gene sets of the comparative analysis can be retrieved as follows

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

```
## Extracting the top gene sets of the collection
## KEGG Pathways for the contrast comparison
##  Sorted by avg.rank
```

```
t
```

```
##  [1] "Asthma"
##  [2] "Viral myocarditis"
##  [3] "Intestinal immune network for IgA production"
##  [4] "Amoebiasis"
##  [5] "HTLV-I infection"
##  [6] "NOD-like receptor signaling pathway"
##  [7] "Hematopoietic cell lineage"
##  [8] "Malaria"
##  [9] "Legionellosis"
## [10] "Toll-like receptor signaling pathway"
```

More information on the first gene set of the comparative analysis can be retrieved as follows

```
showSetByName(gsa, "kegg", rownames(t)[1])
```

Next, the visualization capabilities of EGSEA are explored. The results can be visualized at the experiment-level using the MDS plot, Summary or GO Graph plots, or at the set-level using heatmaps and pathway maps.

The performance of the EGSEA base methods on a selected contrast can be visualized usign an MDS plot that shows how different methods rank a gene set collection (Figure 3). For example, the performance of the various methods on the contrast X24IL13-X24 and the KEGG collection can be plotted as follows

```
plotMethods(gsa, gs.label = "kegg", contrast = 1, file.name = "X24IL13-X24-kegg-methods")
```

```
## Generating methods plot for the collection
## KEGG Pathways and for the contrast X24IL13-X24
## character(0)
```

The overall EGSEA significance of all gene sets in a given collection and a selected contrast can be visualized using the summary plots (Figure 4) as follows

```
plotSummary(gsa, gs.label = "kegg", contrast = 1, file.name = "X24IL13-X24-kegg-summary")
```

```
## Generating Summary plots for the collection
## KEGG Pathways and for the contrast X24IL13-X24
```

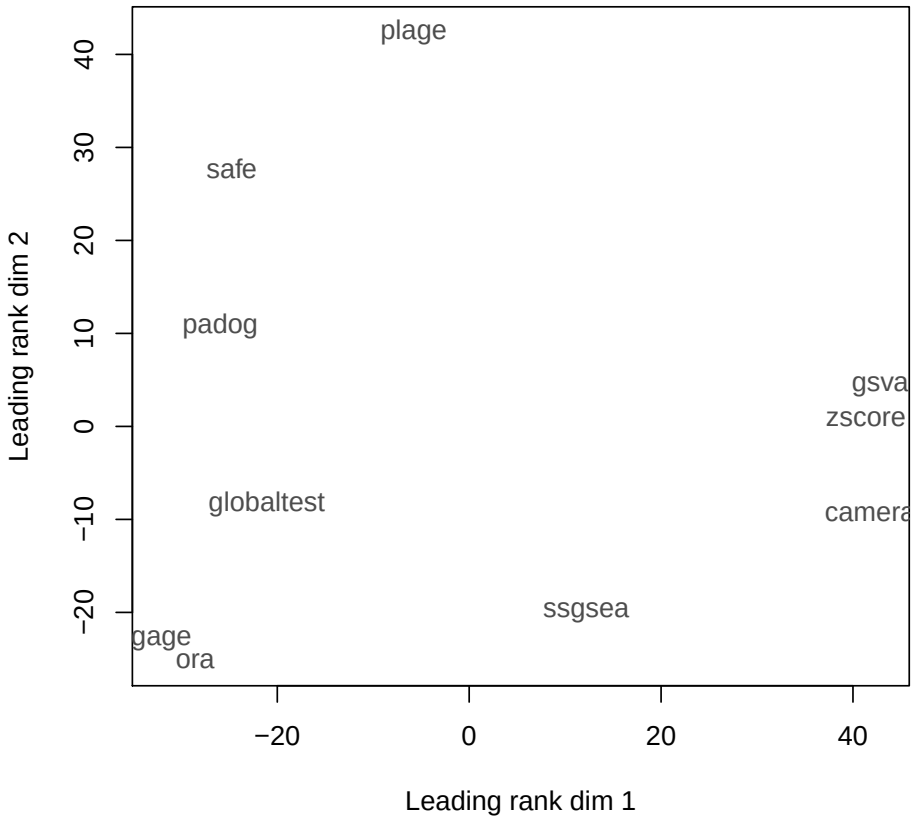


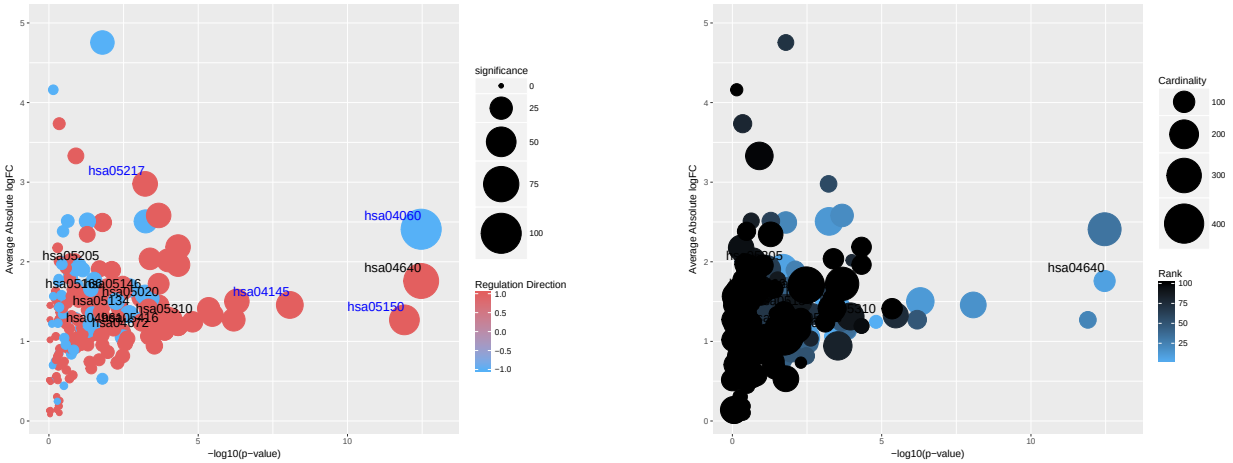
Figure 3: The performance of multiple GSE methods on the contrast X24IL13-X24.

Gene set IDs are used to highlight significant sets on the Summary plot. To obtain additional information on these gene sets, the function `showSetByID` can be used as follows

```
showSetByID(gsa, gs.label = "kegg", c("hsa04060", "hsa04640"))
```

```
## ID: hsa04060
## GeneSet: Cytokine-cytokine receptor interaction
## NumGenes: 194/270
## Type: Signaling
##
## ID: hsa04640
## GeneSet: Hematopoietic cell lineage
## NumGenes: 85/97
## Type: Signaling
```

Gene Ontology (GO) graphs can be generated for the three categories of GO terms: Biological Process (BP), Molecular Function (MF) and Cellular Componenet (CC). There are two GO term collections in the package *EGSEAdata*: c5 from MSigDB and gsdbgo from GeneSetDB. To generate the GO graphs for c5 collection on the contrast X24IL13-X24



(a) Directional summary plot

(b) Ranking summary plot

Figure 4: Summary plots for the contrast X24IL13-X24 on the KEGG pathways collection.

```
plotGOGraph(gsa, gs.label = "c5", file.name = "X24IL13-X24-c5-top-",
    sort.by = "avg.rank")
```

```
## Generating GO Graphs for the collection c5 GO Gene Sets
##  and for the contrast X24IL13-X24 based on the avg.rank
```

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

```
## ( 10280 GO terms found.  )
```

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

```
## ( 14293 GO terms and 34029 relations.  )
```

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

```
## ( 11862 genes annotated to the GO terms.  )
```

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

```
## ( 3572 GO terms found.  )
```

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

```
## ( 4079 GO terms and 5169 relations.  )
```

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

```
## ( 11930 genes annotated to the GO terms.  )
```

```
##
## Building most specific GOs .....
## ( 1492 GO terms found.  )
##
## Build GO DAG topology ..........
## ( 1780 GO terms and 3552 relations.  )
##
## Annotating nodes ...............
## ( 12536 genes annotated to the GO terms.  )
## 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
```

This command generates three graphs, one for each GO category and, by default, displays the top 5 significant terms in each category. For example, Figure 5 shows the BP graph.

Heatmaps of the gene fold changes can be gereated for a selected gene set as follows

```
plotHeatmap(gsa, "Asthma", gs.label = "kegg", contrast = 1, file.name = "asthma-hm")
```

```
## Generating heatmap for Asthma from the collection
## KEGG Pathways and for the contrast X24IL13-X24
```

Figure 6 shows the Asthma gene set heatmap. For the KEGG collection, a pathway map that shows the gene interactiosn can be generated as follows

```
plotPathway(gsa, "Asthma", gs.label = "kegg", file.name = "asthma-pathway")
```

```
## Generating pathway map for Asthma from the collection
## KEGG Pathways and for the contrast X24IL13-X24
```

Figure 7 shows the Asthma pathway map with nodes coloured based on the gene fold changes in the contrast X24IL13-X24.

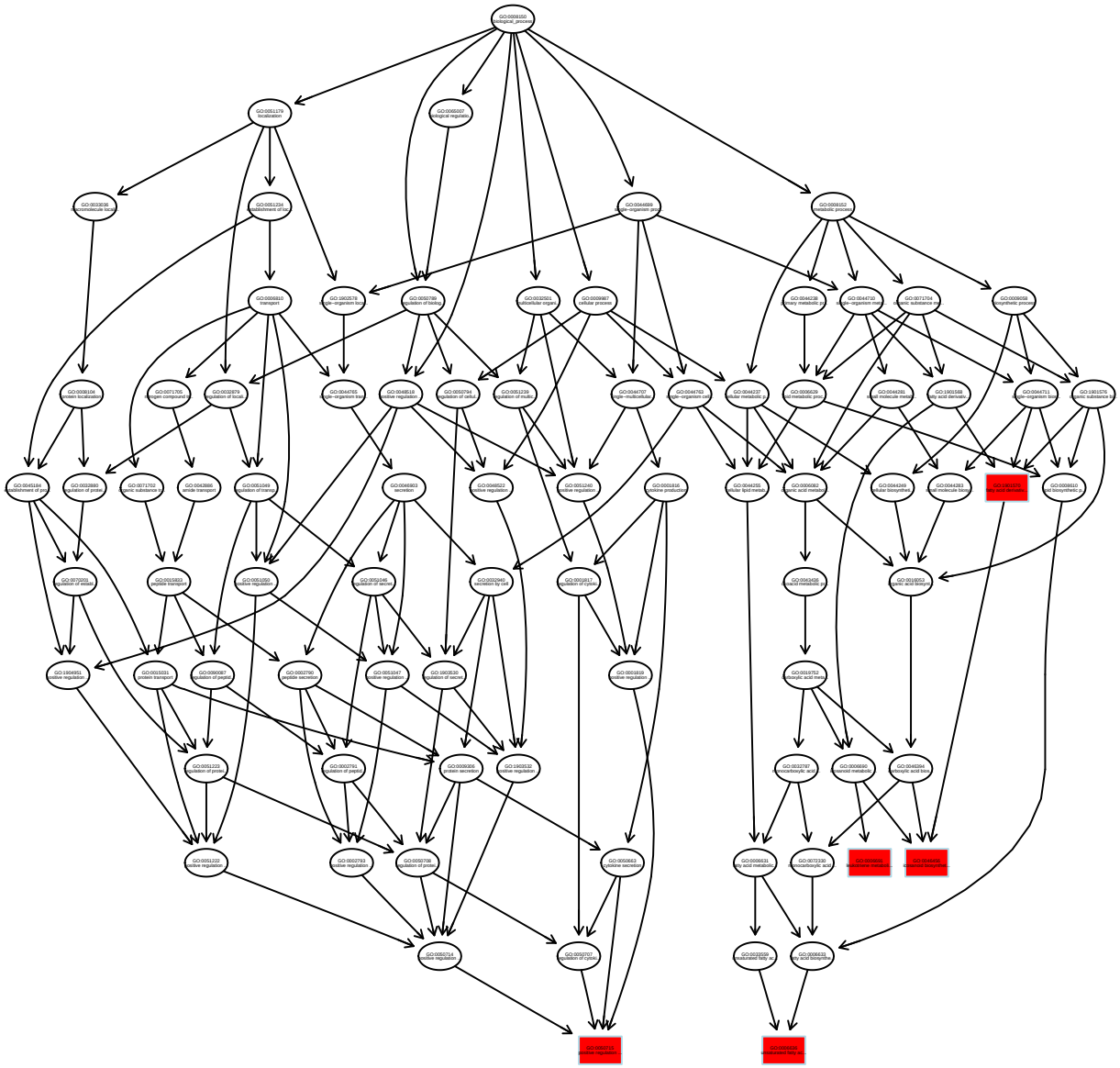


Figure 5: The top significant Biological Processes (BP) from the c5 collection.

Similar reporting capabilities are also provided for the comparative analysis results of EGSEA. A summary plot that compares two contrasts can be generated as follows

```
plotSummary(gsa, gs.label = "kegg", contrast = c(1, 2), file.name = "kegg-summary-cmp")
```

```
## Generating Summary plots for the collection
## KEGG Pathways and for the comparison X24IL13-X24 vs X24IL13Ant-X24IL13
```

Figure 8 shows a comparative summary plot between the two contrasts of the IL-13 dataset. It clearly highlights the effectiveness of the IL-13 antagonist in neuteralizing most of the IL-13 stimulated pathways.

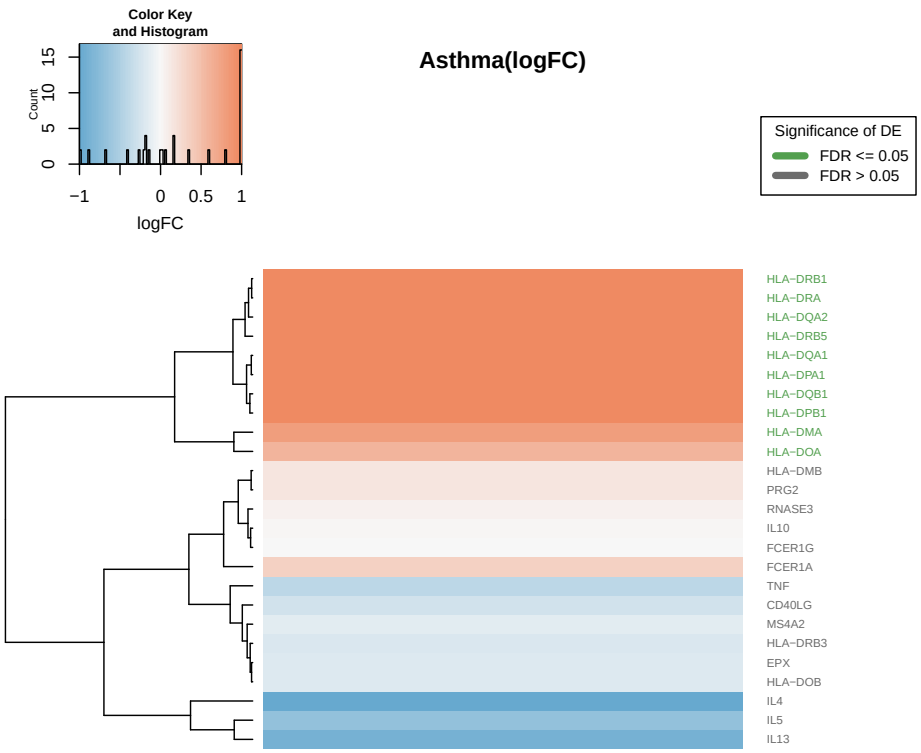


Figure 6: Asthma heatmap for the contrast X24IL13-X24

Alternatively, a summary heatmap for all the contrasts at the gene set level can generated as follows

```
plotSummaryHeatmap(gsa, gs.label = "kegg", show.vals = "p.adj",
    file.name = "il13-sum-heatmap")
```

```
## Generating summary heatmap for the collection KEGG Pathways
## sort.by: avg.rank, hm.vals: avg.rank, show.vals: p.adj
```

Figure 9 shows a summary heatmap for the rankings of top 20 gene sets of the comparative analysis across all the contrasts. The EGSEA adjusted p-values are displayed on the heatmap for each gene set. This can help to identify gene sets that are highly ranked/sgnificant in multiple contrasts.

To closely see how the antagonist works for a given pathway, a comparative heatmap can be generated as follows

```
plotHeatmap(gsa, "Asthma", gs.label = "kegg", contrast = "comparison",
    file.name = "asthma_hm_cmp")
```

```
## Generating heatmap for Asthma from the collection
## KEGG Pathways and for the contrast comparison
```

File not found
asthma-pathway.png
Please report this to package maintainers at GitHub or via email m.hamdoosh@gmail.com

Figure 7: Asthma pathway map for the contrast X24IL13-X24

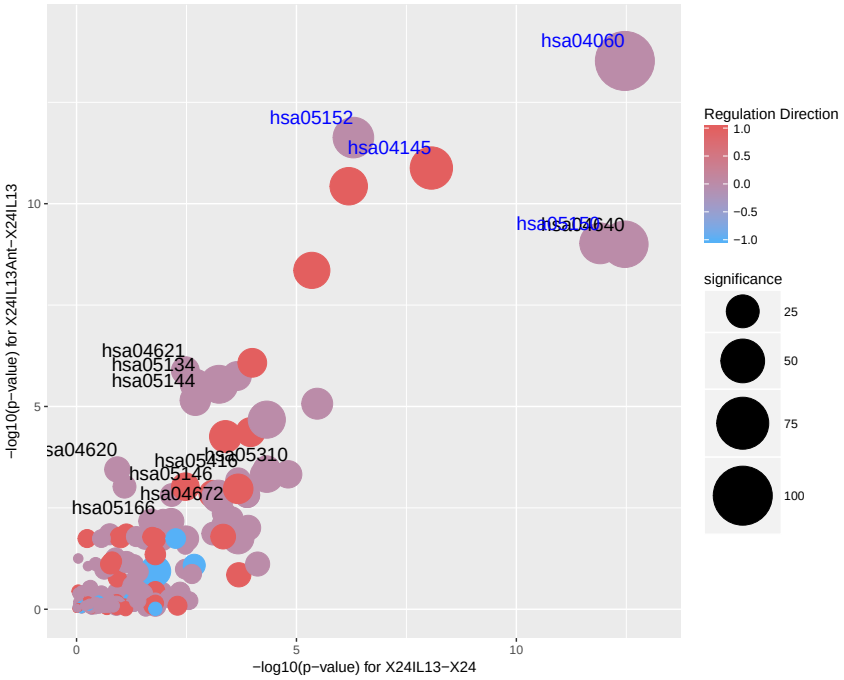


Figure 8: Summary plot for the comparative analysis

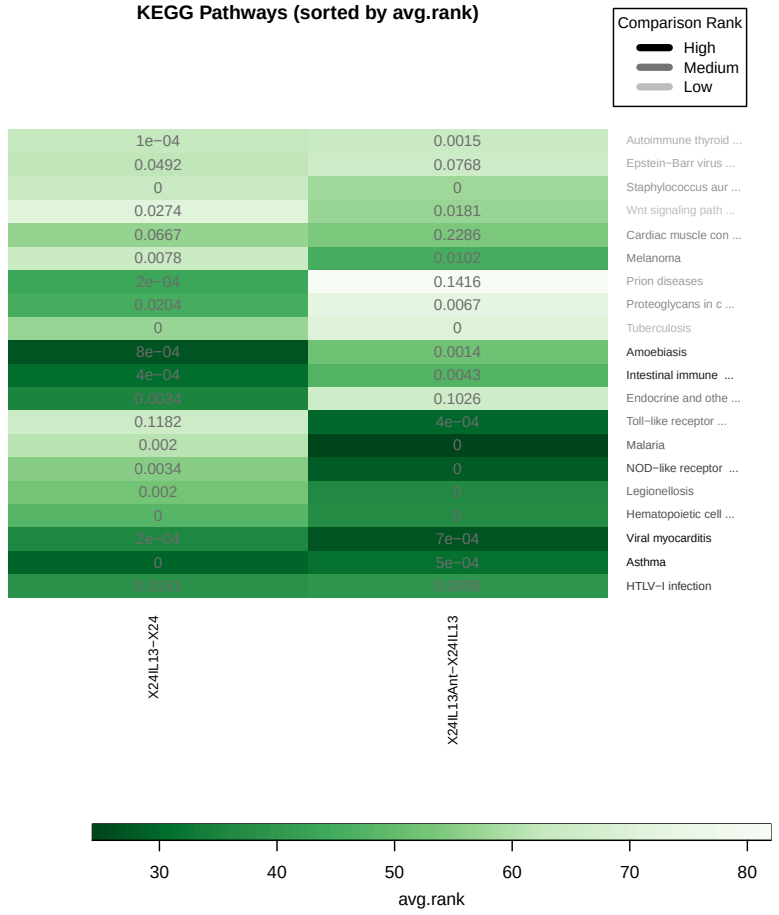


Figure 9: Summary heatmap for the comparative analysis

The heatmap clearly shows that all the genes that were stimulated by IL-13 were reveresed when the antagonist was introduced (Figure 10). Finally, a comparative pathway map can be used to quickly see which genes of the Asthma pathway are affected by IL-13 stimulation (Figure 11) and can be generated as follows

```
plotPathway(gsa, "Asthma", gs.label = "kegg", contrast = 0, file.name = "asthma-pathway-cmp")
```

```
## Generating pathway map for Asthma from the collection
## KEGG Pathways and for the contrast comparison
```

6 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

Figure 10: Asthma heatmap for the comparative analysis

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 combined p-value assigned to γ_i and is calculated using six different methods.
- 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

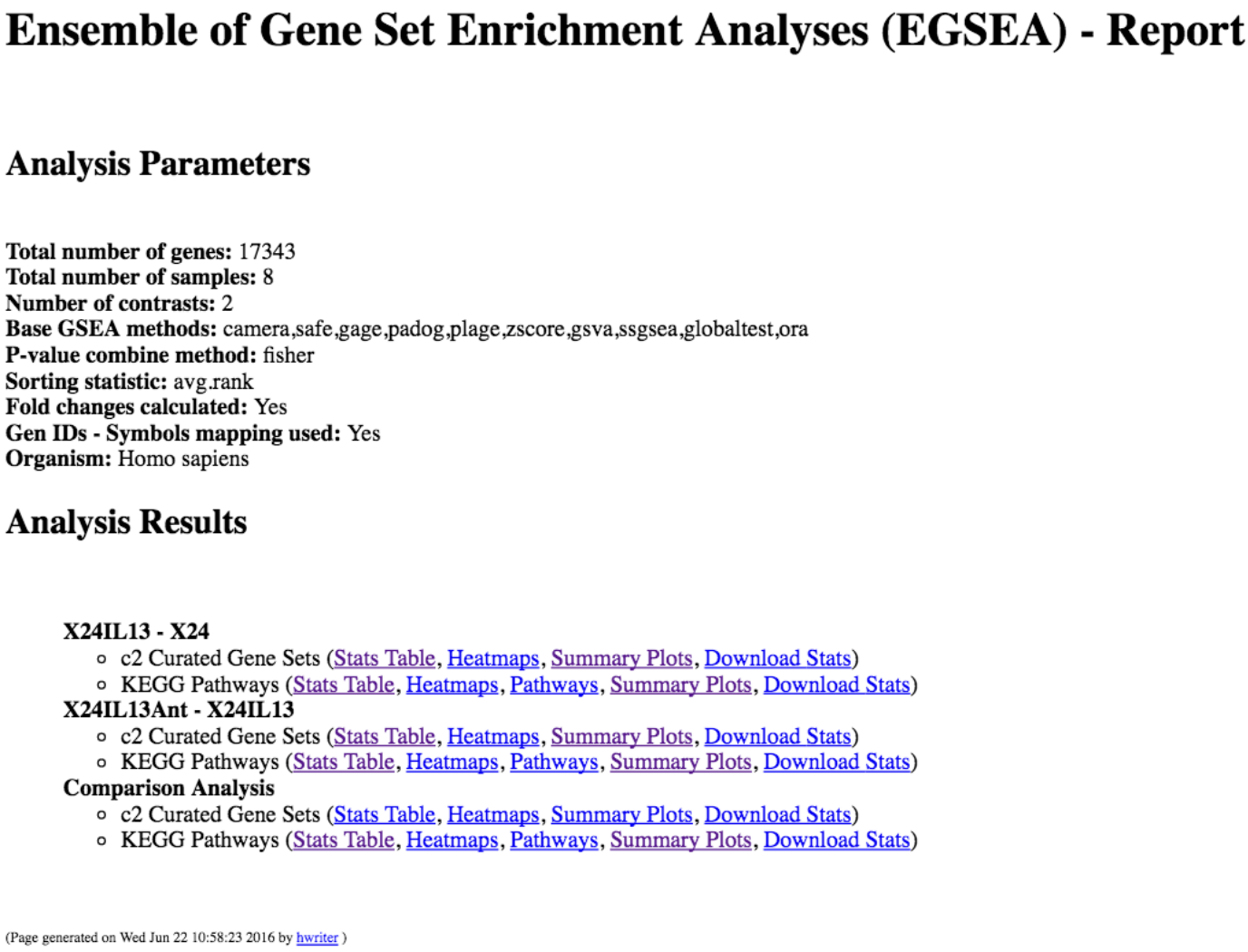
Figure 11: Asthma pathway map for the comparative analysis

2.

7 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* (Figure 12). 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 (Figure 6 and 10) or projected onto pathway maps where available (e.g. KEGG gene sets) (Figure 7 and 11). The most significant Gene Ontology (GO) terms for each comparison can be viewed in a GO graph that shows their relationships (Figure 5).

Additionally, EGSEA creates summary plots for each gene set collection to visualize the overall statistical significance of gene sets (Figure 4 and 8). 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 (Figure 3). This plot gives insights into the similarity of different



Ensemble of Gene Set Enrichment Analyses (EGSEA) - Report

Analysis Parameters

Total number of genes: 17343
Total number of samples: 8
Number of contrasts: 2
Base GSEA methods: camera,safe,gage,padog,plage,zscore,gsva,ssgsea,globaltest,ora
P-value combine method: fisher
Sorting statistic: avg.rank
Fold changes calculated: Yes
Gen IDs - Symbols mapping used: Yes
Organism: Homo sapiens

Analysis Results

X24IL13 - X24
- c2 Curated Gene Sets (Stats Table, Heatmaps, Summary Plots, Download Stats)
- KEGG Pathways (Stats Table, Heatmaps, Pathways, Summary Plots, Download Stats)

X24IL13Ant - X24IL13
- c2 Curated Gene Sets (Stats Table, Heatmaps, Summary Plots, Download Stats)
- KEGG Pathways (Stats Table, Heatmaps, Pathways, Summary Plots, Download Stats)

Comparison Analysis
- c2 Curated Gene Sets (Stats Table, Heatmaps, Summary Plots, Download Stats)
- KEGG Pathways (Stats Table, Heatmaps, Pathways, Summary Plots, Download Stats)

(Page generated on Wed Jun 22 10:58:23 2016 by hwriter)

Figure 12: The main page of the EGSEA HTML report.

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.

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

8 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 GSE63310.

To perform EGSEA analysis on this dataset, the following commands can be 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$contra
contrasts
# build the gene set collections
gs.annots = buildIdx(entrezIDs = rownames(v$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"])
na.sym
symbolsMap[na.sym, "Symbols"] = symbolsMap[na.sym, "FeatureID"]
# perform the EGSEA analysis set report = TRUE to generate
# the EGSEA interactive report
baseMethods = c("camera", "safe", "gage", "padog", "zscore",
    "gsva", "globaltest", "ora")
gsa = egsea(voom.results = v, contrasts = contrasts, gs.annots = gs.annots,
    symbolsMap = symbolsMap, baseGSEAs = baseMethods, sort.by = "med.rank",
    num.threads = 4, report = FALSE)
# show top 20 comparative gene sets in C2 collection
summary(gsa)
topSets(gsa, gs.label = "c2", contrast = "comparison", number = 20)
```

9 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 = buildIdx(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
gsa = egsea.cnt(counts = cnt, group = group, design = design,
    contrasts = contrasts, gs.annots = gs.annots, symbolsMap = genes,
    baseGSEAs = egsea.base()[-c(2, 12)], sort.by = "avg.rank",
    num.threads = 4, report = FALSE)
```

10 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.

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 = buildIdx(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 analysis has started"
## ##------ Thu Jul 20 21:41:47 2017 ------##
## [1] "The ensemble mode was disabled. No sufficient number of base methods."
## [1] "EGSEA is running on the provided data and kegg collection"
## .ora*
## ##------ Thu Jul 20 21:41:48 2017 ------##
## [1] "EGSEA analysis took 0.200000000000045 seconds."
## [1] "EGSEA analysis has completed"
```

11 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 `buildCustomIdx` 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.annot = buildCustomIdx(entrezIDs = rownames(v$E), gsets = gsets,
    species = "human")
```

```
## [1] "Created the User-Defined Gene Sets collection ... "
```

```
class(gs.annot)
```

```
## [1] "GSCollectionIndex"
## attr(,"package")
## [1] "EGSEA"
```

```
show(gs.annot)
```

```
## An object of class "GSCollectionIndex"
##  Number of gene sets: 49
##  Annotation columns: ID, GeneSet, NumGenes
##  Total number of indexing genes: 17343
##  Species: Homo sapiens
##  Collection name: User-Defined Gene Sets
##  Collection uniqe label: custom
##  Database version: NA
##  Database update date: Thu Jul 20 21:41:48 2017
```

The `buildCustomIdx` 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`.

12 Adding 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*. We do not allow users to add new methods by themselves because this procedure is a straightforward and is a method-dependent.

13 Packages used

```
sessionInfo()
```

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows Server 2012 R2 x64 (build 9600)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=C                           LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
##  [1] grid      stats4    parallel  stats     graphics  grDevices utils     datasets
##  [9] methods   base
##
## other attached packages:
##  [1] limma_3.32.3         Rgraphviz_2.20.0     EGSEAdata_1.4.0      EGSEA_1.4.1
##  [5] pathview_1.16.1      org.Hs.eg.db_3.4.1   topGO_2.28.0         SparseM_1.77
##  [9] GO.db_3.4.1          graph_1.54.0         AnnotationDbi_1.38.1 IRanges_2.10.2
## [13] S4Vectors_0.14.3     gage_2.26.1          Biobase_2.36.2       BiocGenerics_0.22.0
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-131             bitops_1.0-6             matrixStats_0.52.2
##  [4] GSVA_1.24.1              R2HTML_2.3.2             bit64_0.9-7
##  [7] RColorBrewer_1.1-2       httr_1.2.1               rprojroot_1.2
## [10] tools_3.4.1              backports_1.1.0          doRNG_1.6.6
## [13] R6_2.2.2                 KernSmooth_2.23-15       DBI_0.7
## [16] lazyeval_0.2.0           colorspace_1.3-2         bit_1.1-12
## [19] compiler_3.4.1           formatR_1.5              pkgmaker_0.22
## [22] labeling_0.3             caTools_1.17.1           KEGGgraph_1.38.0
## [25] scales_0.4.1             PADOG_1.18.0             stringr_1.2.0
## [28] digest_0.6.12            rmarkdown_1.6            XVector_0.16.0
## [31] pkgconfig_2.0.1          htmltools_0.3.6          HTMLUtils_0.1.7
## [34] highr_0.6                rlang_0.1.1              RSQLite_2.0
## [37] hwriter_1.3.2            gtools_3.5.0             RCurl_1.95-4.8
## [40] magrittr_1.5             KEGG.db_3.2.3            Matrix_1.2-10
## [43] Rcpp_0.12.12             munsell_0.4.3            stringi_1.1.5
## [46] yaml_2.1.14              edgeR_3.18.1             zlibbioc_1.22.0
## [49] globaltest_5.30.0        gplots_3.0.1             KEGGdzPathwaysGEO_1.14.0
## [52] plyr_1.8.4               blob_1.1.0               gdata_2.18.0
## [55] lattice_0.20-35          splines_3.4.1            Biostrings_2.44.2
## [58] annotate_1.54.0          KEGGREST_1.16.0          locfit_1.5-9.1
## [61] knitr_1.16               hgu133plus2.db_3.2.3     hgu133a.db_3.2.3
## [64] rngtools_1.2.4           codetools_0.2-15         XML_3.98-1.9
```

```
## [67] GSA_1.03                evaluate_0.10.1          metap_0.8
## [70] png_0.1-7               foreach_1.4.3            org.Mm.eg.db_3.4.1
## [73] gtable_0.2.0            org.Rn.eg.db_3.4.1       ggplot2_2.2.1
## [76] xtable_1.8-2            survival_2.41-3          tibble_1.3.3
## [79] iterators_1.0.8         safe_3.16.0              registry_0.3
## [82] memoise_1.1.0           BiocStyle_2.4.0          GSEABase_1.38.0
```

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.

[17] CW Law, M Alhamdoosh, S Su, GK Smyth, and ME Ritchie. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger [version 1; referees: 1 approved]. *F1000Research*, 5(1408), 2016. doi: 10.12688/f1000research.9005.1.