ToPASeq: an R package for topology-based pathway analysis of microarray and RNAseq data

Ivana Ihnatova, Eva Budinska

November 13, 2014

Contents

Chapter 1

Introduction

This package de-novo implements or adjust the existing implementations of several different methods for topology-based pathway analysis of gene expression data from microarray and RNA-Seq technologies.

These high-throughput technologies are used for measuring of expression levels of thousands genes in one experiment often with the aim to find pathways and biological processes affected between two conditions. The information which biological processes are affected helps investigators to set-up biologically relevant hypotheses for further research.

To this end, a differential gene expression between conditions is assessed - by the means of specific methods, such as limma for instance, which produce lists of differentially expressed genes with specific statistics and p-values for each gene, as well as fold change of mean expression between compared groups.

Pathway analysis is the next step, where these differentially expressed genes are mapped to reference pathways derived from databases and relative enrichment is assessed. Methods of topology-based pathway analysis are the last generation of pathway analysis methods that take into account the topological structure of a pathway, which helps to increase specificity and sensitivity of the results.

This package implements seven topology-based pathway analysis methods that focus on identification of the pathways that are differentially affected between two conditions (Table 1.1). Each method is implemented as a single wrapper function which allows the user to call a method in a single command. In addition, this package offers a visualization of the results. The visualization is based on the `Rgraphviz` package and displays distribution of differential expression and topological significance of the nodes from one pathway. The user can simplify the pathway topology by merging selected sets of nodes into one (individual gene names is the only information that is lost in it).

Table 1.1: Methods included in the package.

Method	Ref.	Type	Implementation
TopologyGSA	[Massa *et al.*(2010)]	M	imported
DEGraph	[Jacob *et al.*(2010)]	M	imported
clipper	[Martini *et al.*(2012)]	M	imported
SPIA	[Tarca *et al.*(2009)], [Draghici *et al.*(2007)]	U	imported
TBS	[Al-Haj Ibrahim *et al.*(2012)]	U	de novo
PWEA	[Hung *et al.*(2010)]	U	de novo
TAPPA	[Gao and Wang(2007)]	U	de novo

M - multivariable, U - univariable

1.1 Input, output and general functionalities

The input data are either normalized (count) data or gene expression data as well as pathway topological structure.

For the sake of simplicity, our package offers in each wrapper function a preprocessing step for RNA-seq normalization - TMM [Robinson and Oshlack(2010] and DESeq [Anders and Huber(2010)]. If necessary, the functions also performs differential gene expression analysis through calling limma and DESeq2 packages.

To summarize, the wrapper functions give options to: 1) normalize the count data (for RNAseq) 2) apply differential expression analysis on gene-level, if applicable, and finally 3) perform topologifal pathway analysis. The functions provides output in a uniform format defined as a new S3 class topResult with basic methods (print, plot, summary) and methods for obtaining the individual parts of the output.

1.2 Pathway topological structure

Pathways and their topological structures are an important input for the analysis. They are represented as graphs $G = (V, E)$, where V denotes a set of vertices or nodes represented by genes and $E \subseteq V \times V$ is a set of edges between nodes (oriented or not, depending on the method) representing the interaction between genes. These structures are can be downloaded from public databases such as KEGG or Biocarta or are available through other packages such as `graphite`.

ToPASeq is build upon `graphite` R-package where pathways from seven public databases: KEGG, Biocarta, Reactome, NCI, SPIKE, HumanCyc, Panther were downloaded and parsed into a new S4 class `pathway`. The parsing process deals also with a special type of nodes that can be found in biological pathways. Protein complexes are expanded into cliques since it is assumed that all units from one complex interact with each other. A clique, from graph theory, is a subset of vertices such that every two vertices in the subset are connected by

an edge. On the other hand, gene families are expanded into separate nodes with same incoming and/or outgoing edges, because they are believed to be interchangable. The most important modification is the propagation of signal through the so called compound-mediated interactions. By compound-mediated interaction we mean an interaction that engages not only genes or their product but also other chemical compounds e.g. calcium ions. `graphite` is the first package that propagates signal through such interactions. For example, if gene A interacts with compound c and compound c with gene B then in a pathway topology gene A should interact with gene B. Please see [Sales *et al.*(2012)] for more details.

1.3 Preparing and manipulating pathways

The easiest way is to use pathway available through graphite. However, you might need to use your own pathway - the easiest way is to download it from some database (do not forget this pathway needs to contain topological information!) and convert it to the correct format using our specific functions for pathway conversion and manipulation.

Functions `AdjacencyMatrix2Pathway` and `graphNEL2Pathway` coerce either an adjacency matrix (binary matrix, where 1 means an edge between two genes) or `graphNEL` into `pathway`. For a reduction of a specified set of nodes (e.g. genes from the same class with similar function), which helps to simply the graphical graph representation, you can use function `reduceGraph`.

Any other topological manipulations can be achieved through `graphNEL` and conversion from and to `pathway`.

The normalized gene expression data or count data can be in two formats. One is an simple matrix were rows refer to genes and the other one is an `ExpressionSet`. There are four acceptable formats for the clincal data: the name or number of `phenoData` of `ExpressionSet` or a character or numeric vector that is coerced to factor. We will demonstrate the features of the package on the example of analysis of two datasets. For microarray data we will use the log2-transformed normalized expression data from the `DEGraph` package and for RNA-Seq data we will use the count data from `gageData` package. The pathway topologies are available as objects named according to the database they come from: `kegg`, `biocarta`, `reactome`, `nci` etc.

Chapter 2

Analysis of microarray data

In our example we will use the dataset `Loi2008_DEGraphVignette` from `DEGraph` package. It conatains the expression profiles of 255 patients with hormone-dependent breast cancer stored as a matrix. The aim of the study was to determine which genes are differentially expressed between tamoxifen-resistant and tamoxifen-sensitive samples. Gene expression data matrix and vector of class labels is stored as separate objects `exprLoi2008` and `classLoi2008`, respectively. In `classLoi2008`, `0` refers to a tamoxifen-resistant sample and `1` to a tamoxifen-sensitive one. We will not need the annotation data (`annLoi2008`) or KEGG pathways `grListKEGG` in our example. On the other hand, we will use a few first pathways from `KEGG`. The pathways were selected only in order to reduce the computational complexity of the analysis. Also, the outputs are displayed as comments following the command applying a method with high time requirements.

We will load the package, the data and subset of the pathways with

```
> library(ToPASeq)
> library(DEGraph)
> data(Loi2008_DEGraphVignette)
> pathways<-kegg[1:5]
> ls()

[1] "annLoi2008"   "classLoi2008" "exprLoi2008"
[4] "grListKEGG"   "pathways"
```

2.1 TopologyGSA

TopologyGSA represents a multivariable method in which the expression of genes is modelled with Gausian Graphical Models with covariance matrix reflecting the pathway topology. It uses the the Iterative Proportional Scaling algorithm to estimate the covariance matrices. The testing procedure is a two-step process. First the equality of covariance matrices is testes via a likelihood

ratio test. Then, when the null hypothesis of equality of covariance matrices is not rejected, the differential expression is testes via multivariate analysis of variance. On the other hand, when the convariance matrices are not equal, then Behrens-Fisher method for testing the equality of means in a two sample problem with unequal covariance matrices is employed.

The method can be used with a single command

```
> top<-TopologyGSA(exprLoi2008, classLoi2008, pathways, type="MA", nperm=200)
> #99 node labels mapped to the expression data
> #Average coverage 31.47657 %
> #0 (out of 5) pathways without a mapped node
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> res(top)
> #                                      t.value df.mean1 df.mean2 p.value
> #Acute myeloid leukemia                3080.663       30      224   0.000
> #Adherens junction                     1102.830       10      244   0.040
> #Adipocytokine signaling pathway       3196.432       25      229   0.000
> #Adrenergic signaling in cardiomyocytes 2178.476       26      228   0.055
> #African trypanosomiasis               1400.088        8      246   0.000
> #                                      lambda.value df.var  p.value.var
> #Acute myeloid leukemia                   217.92044    165 3.622794e-03
> #Adherens junction                         39.92094     10 1.749659e-05
> #Adipocytokine signaling pathway          192.81336    121 3.595452e-05
> #Adrenergic signaling in cardiomyocytes   169.47418     80 2.211953e-08
> #African trypanosomiasis                   13.77192     15 5.428926e-01
> #                                      qchisq.value var.equal
> #Acute myeloid leukemia                   195.97336         1
> #Adherens junction                         18.30704         1
> #Adipocytokine signaling pathway          147.67353         1
> #Adrenergic signaling in cardiomyocytes   101.87947         1
> #African trypanosomiasis                   24.99579         0
>
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. The `nperm` argument sets the number of permutations to be used in the statistical tests. By default both mean and variance tests are run, this can be changed to only variance test by setting `test="var"`. Also the node labels of pathway topologies are converted into entrezIDs. This is controlled with arguments `convert`, and `IDs`. A converstion into the gene symbols is available too. Please note, that

the node labels should be the same as the rownames of gene expression data matrix. The threshold for variance test is specified with `alpha` argument. The implementation allows also testing of all the cliques present in the graph by setting `testCliques=TRUE`. Please note that these tests may take quite a long time.

2.2 DEGraph

Another multivariable method implemented in the package is DEGraph. This method assumes the same direction in the differential expresion of genes belonging to a pathway. It performs the regular Hotelling's T2 test in the graph-Fourier space restricted to its first k components which is more powerful than test in the full graph-Fourier space or in the original space.

We apply the method with

```
> deg<-DEGraph(exprLoi2008, classLoi2008, pathways, type="MA")

99 node labels mapped to the expression data
Average coverage 31.47657 %
0 (out of 5) pathways without a mapped node

> res(deg)

                                    Overall.p
Acute myeloid leukemia              0.03521929
Adherens junction                   NA
Adipocytokine signaling pathway     0.008440407
Adrenergic signaling in cardiomyocytes 0.05739118
African trypanosomiasis             0.2342124
                                    Comp1.p
Acute myeloid leukemia              0.1532096
Adherens junction                   NA
Adipocytokine signaling pathway     0.03920983
Adrenergic signaling in cardiomyocytes 0.1538293
African trypanosomiasis             0.0472761
                                    Comp1.pFourier
Acute myeloid leukemia              0.03521929
Adherens junction                   NA
Adipocytokine signaling pathway     0.008440407
Adrenergic signaling in cardiomyocytes 0.05739118
African trypanosomiasis             0.2342124
                                    Comp1.graph Comp1.k
Acute myeloid leukemia              ?           4
Adherens junction                   NA          NA
Adipocytokine signaling pathway     ?           1
Adrenergic signaling in cardiomyocytes ?        3
```

```
African trypanosomiasis                ?           1
                                       Comp2.p
Acute myeloid leukemia                 0.006982534
Adherens junction                      NA
Adipocytokine signaling pathway        NA
Adrenergic signaling in cardiomyocytes 0.492055
African trypanosomiasis                0.02562905
                                       Comp2.pFourier
Acute myeloid leukemia                 0.0004994694
Adherens junction                      NA
Adipocytokine signaling pathway        NA
Adrenergic signaling in cardiomyocytes 0.7744589
African trypanosomiasis                0.1517751
                                       Comp2.graph Comp2.k
Acute myeloid leukemia                 ?           1
Adherens junction                      NA          NA
Adipocytokine signaling pathway        NA          NA
Adrenergic signaling in cardiomyocytes ?           1
African trypanosomiasis                ?           1
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with arguments `convert`, and `IDs`. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. Since, the DEGraph method runs a statistical test for each connected component of a pathway, a method for assigning a global p-value for whole pathway is needed. The user can select from three approaches: the minimum, the mean and the p-value of the biggest component. This is specified via `overall` argument. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or modified t-statistic from `limma` (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway.

2.3 clipper

The last multivariable method available within this package is called clipper. This method is similar to the topologyGSA as it uses the same two-step approach. However, the Iterative Proportional Scaling algorithm was subsituted with a shrinkage procedure of James-Stein-type which additionally allows proper estimates also in the situation when number of samples is smaller than the number of genes in a pathway. The tests on a pathway-level are follwed with a search for the most affected path in the graph.

The method can be applied with

```
> cli<-Clipper( exprLoi2008, classLoi2008, pathways,type="MA", test="mean")
> #99 node labels mapped to the expression data
> #Average coverage 31.47657 %
> #0 (out of 5) pathways without a mapped node
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> res(cli)
> #                                    alphaVar alphaMean  maxScore activation
> #Acute myeloid leukemia                0.788     0.008  4.336307  0.1255490
> #Adherens junction                     0.087     0.027        NA         NA
> #Adipocytokine signaling pathway       0.675     0.000 33.209403  0.8012589
> #Adrenergic signaling in cardiomyocytes 0.108     0.042        NA         NA
> #African trypanosomiasis               0.966     0.005        NA         NA
> #                                      impact
> #Acute myeloid leukemia             0.3846154
> #Adherens junction                         NA
> #Adipocytokine signaling pathway    0.5000000
> #Adrenergic signaling in cardiomyocytes    NA
> #African trypanosomiasis                   NA
> #
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway    32;51422;53632;5562;5563;5564;5565;5571;2538;51422
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> #
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway    32;51422;53632;5562;5563;5564;5565;5571,2538;51422
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with arguments `convert`, and `IDs`. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. Also, both mean and variance tests are run, this can be changed to only variance test by setting

`test="var"`. The `nperm` controls the number of permutations in the statistical tests. Similarly as in topologyGSA, the implementation allows testing of all the cliques present in the graph by setting `testCliques=TRUE`. Please note that these tests may take quite a long time.

2.4 SPIA

The most well-known topology-based pathway analysis method is SPIA. In there, two evidences of differential expression of a pathway are combined. The first evidence is a regular so called overrepresentation analysis in which the statistical significance of the number of differentially expressed genes belonging to a pathway is assessed. The second evidence reflects the pathway topology and it is called the pertubation factor. The authors assume that a differentially expressed gene at the begining of a pathway topology (e.g. a receptor in a signaling pathway) has a stronger effect on the functionality of a pathway than a differentially expressed gene at the end of a pathway (e.g. a transcription factor in a signaling pathway). The pertubation factors of all genes are calculated from a system of linear equations and then combined within a pathway. The two evidences in a form of p-values are finally combined into a global p-value, which is used to rank the pathways.

```
> spi<-SPIA(exprLoi2008, classLoi2008,pathways , type="MA", logFC.th=-1)

99 node labels mapped to the expression data
Average coverage 31.47657 %
0 (out of 5) pathways without a mapped node
Acute myeloid leukemia
Adherens junction
Adipocytokine signaling pathway
Adrenergic signaling in cardiomyocytes
African trypanosomiasis
0 denoted as 0
 1 denoted as 1
 Contrasts:  1 - 0
Found 40 differentially expressed genes

Done pathway 1 : Acute myeloid leukemia..
Done pathway 2 : Adherens junction..
Done pathway 3 : Adipocytokine signaling pathwa..
Done pathway 4 : Adrenergic signaling in cardio..
Done pathway 5 : African trypanosomiasis..

> res(spi)

                                     pSize NDE       pNDE
Adipocytokine signaling pathway         25   8 0.04877082
```

```
African trypanosomiasis                    8   3 0.14980853
Adherens junction                         10   3 0.24909633
Acute myeloid leukemia                    30   5 0.64330485
Adrenergic signaling in cardiomyocytes    26   4 0.71166426
                                                  tA pPERT
Adipocytokine signaling pathway          0.175400722 0.631
African trypanosomiasis                 -0.009264582 0.985
Adherens junction                       -0.268304125 0.603
Acute myeloid leukemia                  -0.497198579 0.274
Adrenergic signaling in cardiomyocytes  -0.334295837 0.554
                                                     p      pFdr
Adipocytokine signaling pathway         0.1379023 0.6027761
African trypanosomiasis                 0.4299218 0.6027761
Adherens junction                       0.4349569 0.6027761
Acute myeloid leukemia                  0.4822208 0.6027761
Adrenergic signaling in cardiomyocytes  0.7612173 0.7612173
                                                 pFWER    Status
Adipocytokine signaling pathway         0.6895113 Activated
African trypanosomiasis                 1.0000000 Inhibited
Adherens junction                       1.0000000 Inhibited
Acute myeloid leukemia                  1.0000000 Inhibited
Adrenergic signaling in cardiomyocytes  1.0000000 Inhibited
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. The default thresholds for the differential expression analysis of genes (the moderated t-test from `limma` is used) are set with arguments `logFC.th` and `p.val.th`. The user can omit one of these criteria by setting the agrument negative value, as is shown also in the example. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or modified t-statistic from `limma` (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway.

2.5 TAPPA

TAPPA was among the first topology-based pathway analysis methods. It was inspired in chemointformatics and their models for predicting the structure of molecules. In TAPPA, the gene expression values are standardized and sigma-transformed within a samples. Then, a pathway is seen a molecule, individual

genes as atoms and the energy of a molecule is a score defined for one sample. This score is called Pathway Connectivity Index. The difference of expression is assessed via a common univariable two sample test - Mann-Whitney in our implemetation.

```
> tap<-TAPPA(exprLoi2008, classLoi2008, pathways, type="MA")

99 node labels mapped to the expression data
Average coverage 31.47657 %
0 (out of 5) pathways without a mapped node
0 denoted as 1
 1 denoted as 2

> res(tap)

                                     valid1      median1
Acute myeloid leukemia                   68 -0.006965550
Adherens junction                        68 -0.021800852
Adipocytokine signaling pathway          68 -0.002324606
Adrenergic signaling in cardiomyocytes   68  0.004798401
African trypanosomiasis                  68 -0.045194060
                                        min1      max1
Acute myeloid leukemia           -0.5952676 0.5004099
Adherens junction                -0.1520576 0.1333737
Adipocytokine signaling pathway  -0.3464005 0.3572598
Adrenergic signaling in cardiomyocytes -0.1776318 0.1407472
African trypanosomiasis          -0.2920941 0.3175060
                                     valid2       median2
Acute myeloid leukemia                  187 -0.0492809545
Adherens junction                       187 -0.0027597436
Adipocytokine signaling pathway         187  0.0009354723
Adrenergic signaling in cardiomyocytes  187 -0.0120154851
African trypanosomiasis                 187  0.0001076088
                                        min2      max2
Acute myeloid leukemia           -0.8393242 0.6727377
Adherens junction                -0.1535961 0.1239743
Adipocytokine signaling pathway  -0.4469216 0.5097148
Adrenergic signaling in cardiomyocytes -0.2228221 0.1994799
African trypanosomiasis          -0.3490664 0.5230598
                                          p     p.adj
Acute myeloid leukemia           0.19006652 0.3167775
Adherens junction                0.11696076 0.2924019
Adipocytokine signaling pathway  0.42613127 0.4261313
Adrenergic signaling in cardiomyocytes 0.27420712 0.3427589
African trypanosomiasis          0.05354071 0.2677035
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type`

argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. The user can also specified whether the normalization step (standardization and sigma-transformation) should be perfomed (`normalize=TRUE`). If `verbose=TRUE`, function prints out the titles of pathways as their are analysed. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or modified t-statistic from `limma` (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway.

2.6 TBS

TBS is another method that works with gene-level statistics and a list of differentially expresed genes. The pathway topology is incorporated as the number of downstream differentially expressed genes. The gene-level log fold-changes are weigted by this numeber and sumed up into a pathway-level score. A statistical significance is assessed by a permutations of genes.

```
> tbs<-TBS( exprLoi2008, classLoi2008, pathways, type="MA",  logFC.th=-1, nperm=100)
> #99 node labels mapped to the expression data
> #Average coverage 31.47657 %
> #0 (out of 5) pathways without a mapped node
> #0 denoted as 0
> # 1 denoted as 1
> # Contrasts:  0 - 1
> #Found 40 differentially expressed genes
> #Preparing permutation table and downstream list
> #Observed scores..
> #Random scores..
> #100
> #Normalization and p-values...
> res(tbs)
> #                                      TBS.obs.norm    p     p.adj
> #Acute myeloid leukemia                  -0.8012546 0.90 0.9000000
> #Adherens junction                        2.9052652 0.03 0.1250000
> #Adipocytokine signaling pathway          0.8461749 0.10 0.1666667
> #Adrenergic signaling in cardiomyocytes  -0.5548923 0.80 0.9000000
> #African trypanosomiasis                  1.5028307 0.05 0.1250000
>
```

Arguments of this functions are almost the same as in `SPIA`. Apart from the expected arguments: a gene expression data matrix, a vector of class labels and

a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. The default thresholds for the differential expression analysis of genes (the moderated t-test from `limma` is used) are set with arguments `logFC.th` and `p.val.th`. The user can omit one of these criteria by setting the agrument negative value, as is shown also in the example. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or modified t-statistic from `limma` (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway. There is one extra argument `nperm` which controls the number of permutations.

2.7 PWEA

The last method available in this package is called PathWay Enrichment Analysis (PWEA). This is actually a weigthed form of common Gene Set Enrichment Analysis (GSEA). The weights are called Topological Influence Factor (TIF) and are defined as a geometic mean of ratios of Pearson's correlation coefficient and the distance of two genes in a pathway. The weights of genes outside a pathway are assigned randomly from normal distribution with parameters estimated from the weights of genes in all pathways. A statistical significance of a pathway is assessed via Kolmogorov-Simirnov-like test statistic comparing two cumulative distribution functions with class label permutations.

```
> pwe<-PWEA(exprLoi2008, classLoi2008, pathways,  type="MA", nperm=100)
> #99 node labels mapped to the expression data
> #Average coverage 31.47657 %
> #0 (out of 5) pathways without a mapped node
> #0 denoted as 0
> # 1 denoted as 1
> # Contrasts:  0 - 1
> #Preparing data..
> #100
> #Processing gene set:
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> res(pwe)
> #                                                ES    p p.adj
```

```
> #Acute myeloid leukemia                  0.1995347 0.81  0.81
> #Adherens junction                       0.5757274 0.67  0.81
> #Adipocytokine signaling pathway         0.3272288 0.32  0.81
> #Adrenergic signaling in cardiomyocytes 0.3888446 0.68  0.81
> #African trypanosomiasis                 0.3544996 0.46  0.81
```

Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the node labels of pathway topologies are converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of gene expression data matrix. The `alpha` parameter sets a threshold for gene weights. The purpose of this filtering is to reduce the possiblity that a weight of a gene that is tighly correlated with a few genes are lowered by the weak correlation with other genes in a pathway.The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or modified t-statistic from `limma` (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway. The `nperm` argument controls the number of permutations.

Chapter 3

Analysis of RNA-Seq data

All of the methods metioned in the previus chapter were designed for the microarray data. However, the RNA-Seq technology is gaining popularity and becomes widely used. Unfortunatelly, the topology-based pathway analysis methods are not available for this type of the data. Therefore, we adapted the selected methods for RNA-Seq count matrices. Two types of adaptations were used. If a method works directly with the expression profiles (multivariable methods and TAPPA), then the count matrix is normalized and transformed either by TMM or DESeq2 method. The remaining methods use also or only the gene-level statistics like log fold-change. The differential expression analysis of genes with either `DESeq2` or `limma` package is a part of their implementation.

We will use the data from `gageData` for an example analysis.

```
> library(gageData)
> data(hnrnp.cnts)
> hnrnp.cnts<-hnrnp.cnts[rowSums(hnrnp.cnts)>0,]
> group<-c(rep("sample",4), rep("control",4))
> pathways<-kegg[1:10]
```

3.1 TopologyGSA

TopologyGSA represents a multivariable method in which the expression of genes is modelled with Gausian Graphical Models with covariance matrix reflecting the pathway topology. It uses the the Iterative Proportional Scaling algorithm to estimate the covariance matrices. The testing procedure is a two-step process. First the equality of covariance matrices is testes via a likelihood ratio test. Then, when the null hypothesis of equality of covariance matrices is not rejected, the differential expression is testes via multivariate analysis of variance. On the other hand, when the convariance matrices are not equal, then Behrens-Fisher method for testing the equality of means in a two sample problem with unequal covariance matrices is employed.

The method can be used with a single command

```
> top<-TopologyGSA(hnrnp.cnts, group, pathways[1:3], type="RNASeq", nperm=1000)
> #528 node labels mapped to the expression data
> #Average coverage 83.16538
> #0 (out of 10) pathways without a mapped node
> #Normalization method was not specified. TMM used as default
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> #Alanine, aspartate and glutamate metabolism
> #Aldosterone-regulated sodium reabsorption
> #Allograft rejection
> #alpha-Linolenic acid metabolism
>
> res(top)
> #data frame with 0 columns and 1 rows
>
>
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"TMM"` method is used for the normalization. The user can select `DESeq2` by setting argument `norm.method` to `"DESeq2"`. The `nperm` argument sets the number of permutations to be used in the statistical tests. Other default settings are: both mean and variance tests are calculated, this can be changed to only variance test by setting `test="var"`. Also the node labels of pathway topologies are converted into entrezIDs. This is controlled with arguments `convert`, and `IDs`. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. The threshold for variance test is specified with `alpha` argument. The implementation allows also testing of all the cliques present in the graph by setting `testCliques=TRUE`. Please note that these tests may take quite a long time.

Unfortunatelly, this method requires more samples than nodes in a pathway. Therefore there is an empty output in the example above.

3.2 DEGraph

Another multivariable method implemented in the package is DEGraph. This method assumes the same direction in the differential expresion of genes belonging to a pathway. It performs the regular Hotelling's T2 test in the graph-Fourier space restricted to its first k components which is more powerful than test in the full graph-Fourier space or in the original space.

We apply the method with

```
> deg<-DEGraph(hnrnp.cnts, group, pathways, type="RNASeq")

530 node labels mapped to the expression data
Average coverage 82.98681 %
0 (out of 10) pathways without a mapped node
Normalization method was not specified. TMM used as default

> res(deg)[,1:4]

                                          Overall.p
Acute myeloid leukemia                    0.0283905
Adherens junction                         0.1343409
African trypanosomiasis                   0.6626785
Alanine, aspartate and glutamate metabolism 0.1417877
Aldosterone-regulated sodium reabsorption   0.1738535
Allograft rejection                       0.8546771
alpha-Linolenic acid metabolism           0.07924667
                                          Comp1.p
Acute myeloid leukemia                    NA
Adherens junction                         NA
African trypanosomiasis                   NA
Alanine, aspartate and glutamate metabolism NA
Aldosterone-regulated sodium reabsorption   NA
Allograft rejection                       0.733728
alpha-Linolenic acid metabolism           NA
                                          Comp1.pFourier
Acute myeloid leukemia                    0.0283905
Adherens junction                         0.1343409
African trypanosomiasis                   0.6626785
Alanine, aspartate and glutamate metabolism 0.1417877
Aldosterone-regulated sodium reabsorption   0.1738535
Allograft rejection                       0.8546771
alpha-Linolenic acid metabolism           0.07924667
                                          Comp1.graph
Acute myeloid leukemia                    ?
Adherens junction                         ?
African trypanosomiasis                   ?
Alanine, aspartate and glutamate metabolism ?
Aldosterone-regulated sodium reabsorption   ?
Allograft rejection                       ?
alpha-Linolenic acid metabolism           ?
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"TMM"` method is used for the normalization. The user can select

DESeq2 by setting argument norm.method to "DESeq2". The node labels of pathway topologies are automatically converted into entrezIDs. This is controlled with arguments convert, and IDs. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. Since, the DEGraph method runs a statistical test for each connected component of a pathway, a method for assigning a global p-value for whole pathway is needed. The user can select from three approaches: the minimum, the mean and the p-value of the biggest component. This is specified via overall argument. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (gene.stat="logFC") or modified t-statistic from limma (gene.stat="stats"). These statistics are later used in the visualization of a selected pathway.

3.3 clipper

The last multivariable method available within this package is called clipper. This method is similar to the topologyGSA as it uses the same two-step approach. However, the Iterative Proportional Scaling algorithm was subsituted with a shrinkage procedure of James-Stein-type which additionally allows proper estimates also in the situation when number of samples is smaller than the number of genes in a pathway. The tests on a pathway-level are follwed with a search for the most affected path in the graph.

The method can be applied with

```
> cli<-Clipper(hnrnp.cnts, group, pathways, type="RNASeq", test="mean")
> #528 node labels mapped to the expression data
> #Average coverage 83.16538
> #0 (out of 10) pathways without a mapped node
> #Normalization method was not specified. TMM used as default
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> #Alanine, aspartate and glutamate metabolism
> #Alcoholism
> #Aldosterone-regulated sodium reabsorption
> #Allograft rejection
> #alpha-Linolenic acid metabolism
> res(cli)[1:2,]
> #                       alphaVar alphaMean  maxScore activation    impact
> #Acute myeloid leukemia    0.022     0.008 18.290959  0.3782696 0.2592593 3728;4609;3728;5
> #Adherens junction         0.035     0.018  1.956012  0.1415808 0.1052632
> #
```

```
> #Acute myeloid leukemia 3728;5371;5914;861,3728;4609,3728;5467,3728;595,3728;6932,3728;693
> #Adherens junction                                          2260;6615;7046;7048,4087;4088;40
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"TMM"` method is used for the normalization. The user can select `DESeq2` by setting argument `norm.method` to `"DESeq2"`. The node labels of pathway topologies are automatically converted into entrezIDs. This is controlled with arguments `convert`, and `IDs`. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. Also, both mean and variance tests are run, this can be changed to only variance test by setting `test="var"`. The `nperm` controls the number of permutations in the statistical tests. Similarly as in topologyGSA, the implementation allows testing of all the cliques present in the graph by setting `testCliques=TRUE`. Please note that these tests may take quite a long time.

3.4 SPIA

The most well-known topology-based pathway analysis method is SPIA. In there, two evidences of differential expression of a pathway are combined. The first evidence is a regular so called overrepresentation analysis in which the statistical significance of the number of differentially expressed genes belonging to a pathway is assessed. The second evidence reflects the pathway topology and it is called the pertubation factor. The authors assume that a differentially expressed gene at the begining of a pathway topology (e.g. a receptor in a signaling pathway) has a stronger effect on the functionality of a pathway than a differentially expressed gene at the end of a pathway (e.g. a transcription factor in a signaling pathway). The pertubation factors of all genes are calculated from a system of linear equations and then combined within a pathway. The two evidences in a form of p-values are finally combined into a global p-value, which is used to rank the pathways.

```
> spi<-SPIA(hnrnp.cnts, group, pathways, type="RNASeq", logFC.th=-1)

530 node labels mapped to the expression data
Average coverage 82.98681 %
0 (out of 10) pathways without a mapped node
test was not specified. 'vstlimma' used as default
control denoted as 0
 sample denoted as 1
 Contrasts:  sample - control
Found 7415 differentially expressed genes
```

```
Done pathway 1 : Acute myeloid leukemia..
Done pathway 2 : Adherens junction..
Done pathway 3 : Adipocytokine signaling pathwa..
Done pathway 4 : Adrenergic signaling in cardio..
Done pathway 5 : African trypanosomiasis..
Done pathway 6 : Alcoholism..
Done pathway 7 : Aldosterone-regulated sodium r..
```

> res(spi)

```
                                       pSize NDE
African trypanosomiasis                   20   4
Adherens junction                         65  34
Adipocytokine signaling pathway           57  21
Acute myeloid leukemia                    50  25
Alcoholism                               137  48
Adrenergic signaling in cardiomyocytes   125  54
Aldosterone-regulated sodium reabsorption  25   9
                                           pNDE
African trypanosomiasis                0.99223153
Adherens junction                      0.08567161
Adipocytokine signaling pathway        0.86339795
Acute myeloid leukemia                 0.20023351
Alcoholism                             0.97847131
Adrenergic signaling in cardiomyocytes 0.52820365
Aldosterone-regulated sodium reabsorption 0.82132126
                                              tA pPERT
African trypanosomiasis                -2.6242010 0.069
Adherens junction                      -0.0429354 0.988
Adipocytokine signaling pathway        -4.1884018 0.150
Acute myeloid leukemia                  0.5419265 0.855
Alcoholism                              9.3106872 0.219
Adrenergic signaling in cardiomyocytes -3.0993658 0.561
Aldosterone-regulated sodium reabsorption  0.3402195 0.792
                                              p
African trypanosomiasis                0.2520465
Adherens junction                      0.2936544
Adipocytokine signaling pathway        0.3942274
Acute myeloid leukemia                 0.4733542
Alcoholism                             0.5443803
Adrenergic signaling in cardiomyocytes 0.6567413
Aldosterone-regulated sodium reabsorption 0.9302183
                                           pFdr pFWER
African trypanosomiasis                0.7621324     1
Adherens junction                      0.7621324     1
Adipocytokine signaling pathway        0.7621324     1
```

```
Acute myeloid leukemia                    0.7621324     1
Alcoholism                                0.7621324     1
Adrenergic signaling in cardiomyocytes    0.7661982     1
Aldosterone-regulated sodium reabsorption 0.9302183     1
                                             Status
African trypanosomiasis                   Inhibited
Adherens junction                         Inhibited
Adipocytokine signaling pathway           Inhibited
Acute myeloid leukemia                    Activated
Alcoholism                                Activated
Adrenergic signaling in cardiomyocytes    Inhibited
Aldosterone-regulated sodium reabsorption Activated
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"limma"` method is used for the differential expression analysis on gene-level. The user can select `DESeq2` by setting argument `test` to `"DESeq2"`. The node labels of pathway topologies are automaticaly converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. The default thresholds for the differential expression analysis of genes are set with arguments `logFC.th` and `p.val.th`. The user can omit one of these criteria by setting the agrument negative value, as is shown also in the example. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or test statistic (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway.

3.5 TAPPA

TAPPA was among the first topology-based pathway analysis methods. It was inspired in chemointformatics and their models for predicting the structure of molecules. In TAPPA, the gene expression values are standardized and sigma-transformed within a samples. Then, a pathway is seen a molecule, individual genes as atoms and the energy of a molecule is a score defined for one sample. This score is called Pathway Connectivity Index. The difference of expression is assessed via a common univariable two sample test - Mann-Whitney in our implemetation.

```
> tap<-TAPPA(hnrnp.cnts, group, pathways, type="RNASeq")

530 node labels mapped to the expression data
Average coverage 82.98681 %
0 (out of 10) pathways without a mapped node
```

```
Normalization method was not specified. TMM used as default
control denoted as 1
 sample denoted as 2
```

> res(tap)

```
                                          valid1
Acute myeloid leukemia                         4
Adherens junction                              4
Adipocytokine signaling pathway                4
Adrenergic signaling in cardiomyocytes         4
African trypanosomiasis                        4
Alanine, aspartate and glutamate metabolism    4
Alcoholism                                     4
Aldosterone-regulated sodium reabsorption      4
Allograft rejection                            4
alpha-Linolenic acid metabolism                4
                                         median1
Acute myeloid leukemia                 0.5803186
Adherens junction                      0.5856506
Adipocytokine signaling pathway        0.2270070
Adrenergic signaling in cardiomyocytes -0.2832096
African trypanosomiasis                -0.3735032
Alanine, aspartate and glutamate metabolism  0.9500718
Alcoholism                             -1.5346674
Aldosterone-regulated sodium reabsorption    0.1640505
Allograft rejection                    -0.4317980
alpha-Linolenic acid metabolism        -0.8223371
                                             min1
Acute myeloid leukemia                 0.5203395
Adherens junction                      0.5790818
Adipocytokine signaling pathway        0.2010982
Adrenergic signaling in cardiomyocytes -0.2918570
African trypanosomiasis                -0.4445059
Alanine, aspartate and glutamate metabolism  0.9294232
Alcoholism                             -1.5872543
Aldosterone-regulated sodium reabsorption    0.1474352
Allograft rejection                    -0.4394848
alpha-Linolenic acid metabolism        -0.8846071
                                             max1
Acute myeloid leukemia                 0.6058688
Adherens junction                      0.5934445
Adipocytokine signaling pathway        0.2795654
Adrenergic signaling in cardiomyocytes -0.2525965
African trypanosomiasis                -0.2288833
Alanine, aspartate and glutamate metabolism  0.9903976
```

```
Alcoholism                                -1.3471744
Aldosterone-regulated sodium reabsorption  0.1773254
Allograft rejection                       -0.4080421
alpha-Linolenic acid metabolism           -0.5102183
                                          valid2
Acute myeloid leukemia                         4
Adherens junction                              4
Adipocytokine signaling pathway                4
Adrenergic signaling in cardiomyocytes         4
African trypanosomiasis                        4
Alanine, aspartate and glutamate metabolism    4
Alcoholism                                     4
Aldosterone-regulated sodium reabsorption      4
Allograft rejection                            4
alpha-Linolenic acid metabolism                4
                                            median2
Acute myeloid leukemia                     0.5401001
Adherens junction                          0.6091719
Adipocytokine signaling pathway            0.1991280
Adrenergic signaling in cardiomyocytes    -0.2436956
African trypanosomiasis                   -0.3344318
Alanine, aspartate and glutamate metabolism 1.0269876
Alcoholism                                -1.6259737
Aldosterone-regulated sodium reabsorption  0.1527353
Allograft rejection                       -0.4102112
alpha-Linolenic acid metabolism           -0.7195674
                                               min2
Acute myeloid leukemia                     0.4781089
Adherens junction                          0.5808310
Adipocytokine signaling pathway            0.1529440
Adrenergic signaling in cardiomyocytes    -0.3807526
African trypanosomiasis                   -0.4381323
Alanine, aspartate and glutamate metabolism 0.9485746
Alcoholism                                -2.0779406
Aldosterone-regulated sodium reabsorption  0.1201292
Allograft rejection                       -0.4205114
alpha-Linolenic acid metabolism           -0.7635583
                                               max2
Acute myeloid leukemia                     0.5889745
Adherens junction                          0.6157619
Adipocytokine signaling pathway            0.2638716
Adrenergic signaling in cardiomyocytes    -0.1837874
African trypanosomiasis                   -0.2425912
Alanine, aspartate and glutamate metabolism 1.0769587
Alcoholism                                -1.2792329
Aldosterone-regulated sodium reabsorption  0.1833585
```

```
Allograft rejection                             -0.3532748
alpha-Linolenic acid metabolism                 -0.5599943
                                                         p
Acute myeloid leukemia                          0.3428571
Adherens junction                               0.1142857
Adipocytokine signaling pathway                 0.3428571
Adrenergic signaling in cardiomyocytes          0.6857143
African trypanosomiasis                         0.8857143
Alanine, aspartate and glutamate metabolism     0.2000000
Alcoholism                                      0.8857143
Aldosterone-regulated sodium reabsorption       1.0000000
Allograft rejection                             0.2000000
alpha-Linolenic acid metabolism                 0.3428571
                                                     p.adj
Acute myeloid leukemia                          0.5714286
Adherens junction                               0.5714286
Adipocytokine signaling pathway                 0.5714286
Adrenergic signaling in cardiomyocytes          0.9795918
African trypanosomiasis                         0.9841270
Alanine, aspartate and glutamate metabolism     0.5714286
Alcoholism                                      0.9841270
Aldosterone-regulated sodium reabsorption       1.0000000
Allograft rejection                             0.5714286
alpha-Linolenic acid metabolism                 0.5714286
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"TMM"` method is used for the normalization. The user can select `DESeq2` by setting argument `norm.method` to `"DESeq2"`. The node labels of pathway topologies are automatically converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. The user can also specified whether the normalization step (standardization and sigma-transformation) should be perfomed (`normalize=TRUE`). If `verbose=TRUE`, function prints out the titles of pathways as their are analysed. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or test statistic (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway.

3.6 TBS

TBS is another method that works with gene-level statistics and a list of differentially expresed genes. The pathway topology is incorporated as the number of

downstream differentially expressed genes. The gene-level log fold-changes are weigted by this numeber and sumed up into a pathway-level score. A statistical significance is assessed by a permutations of genes.

```
> tbs<-TBS(hnrnp.cnts, group, pathways, type="RNASeq",  logFC.th=-1, nperm=100)
> #528 node labels mapped to the expression data
> #Average coverage 83.16538
> #0 (out of 10) pathways without a mapped node
> #test was not specified. 'vstlimma' used as default
> #Found 5702 differentially expressed genes
> #Preparing permutation table and downstream list
> #Observed scores..
> #Random scores..
> #100
> #Normalization and p-values...
> res(tbs)
> #                                          TBS.obs.norm    p      p.adj
> #Acute myeloid leukemia                      -1.6325413 0.05 0.06250000
> #Adherens junction                           -3.9416308 0.01 0.01666667
> #Adipocytokine signaling pathway             -3.1989858 0.00 0.00000000
> #Adrenergic signaling in cardiomyocytes     -16.1777366 0.00 0.00000000
> #African trypanosomiasis                     -4.0834773 0.00 0.00000000
> #Alanine, aspartate and glutamate metabolism  0.0137086 0.44 0.48888889
> #Alcoholism                                  -4.1997338 0.00 0.00000000
> #Aldosterone-regulated sodium reabsorption    1.9996012 1.00 1.00000000
> #Allograft rejection                         -3.4004380 0.01 0.01666667
> #alpha-Linolenic acid metabolism             -2.6720346 0.02 0.02857143
```

Arguments of this functions are almost the same as in `SPIA`. Apart from the expected arguments: a gene expression data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"limma"` method is used for the differential expression analysis on gene-level. The user can select `DESeq2` by setting argument `test` to `"DESeq2"`. The node labels of pathway topologies are automatically converted into entrezIDs. This is controlled with `IDs` argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. The default thresholds for the differential expression analysis of genes are set with arguments `logFC.th` and `p.val.th`. The user can omit one of these criteria by setting the agrument negative value, as is shown also in the example. The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (`gene.stat="logFC"`) or test statistic (`gene.stat="stats"`). These statistics are later used in the visualization of a selected pathway. The last argument `nperm` controls the number of permutations.

3.7 PWEA

The last method available in this package is called PathWay Enrichment Analysis (PWEA). This is actually a weigthed form of common Gene Set Enrichment Analysis (GSEA). The weights are called Topological Influence Factor (TIF) and are defined as a geometic mean of ratios of Pearson's correlation coefficient and the distance of two genes in a pathway. The weights of genes outside a pathway are assigned randomly from normal distribution with parameters estimated from the weights of genes in all pathways. A statistical significance of a pathway is assessed via Kolmogorov-Simirnov-like test statistic comparing two cumulative distribution functions with class label permutations.

```
> pwe<-PWEA(hnrnp.cnts, group, pathways, type="RNASeq", nperm=100)
> #528 node labels mapped to the expression data
> #Average coverage 83.16538
> #0 (out of 10) pathways without a mapped node
> #test was not specified. 'vstlimma' used as default
> #Preparing data..
> #1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25
> #Acute myeloid leukemia
> #Adherens junction
> #Adipocytokine signaling pathway
> #Adrenergic signaling in cardiomyocytes
> #African trypanosomiasis
> #Alanine, aspartate and glutamate metabolism
> #Alcoholism
> #Aldosterone-regulated sodium reabsorption
> #Allograft rejection
> #alpha-Linolenic acid metabolism
> res(pwe)
> #                                                ES    p     p.adj
> #Acute myeloid leukemia                      0.3526104 0.29 0.4142857
> #Adherens junction                           0.3829831 1.00 1.0000000
> #Adipocytokine signaling pathway             0.3102945 1.00 1.0000000
> #Adrenergic signaling in cardiomyocytes      0.3611207 0.20 0.3333333
> #African trypanosomiasis                     0.3272899 0.20 0.3333333
> #Alanine, aspartate and glutamate metabolism 0.2720946 0.20 0.3333333
> #Alcoholism                                  0.4708293 0.86 1.0000000
> #Aldosterone-regulated sodium reabsorption   0.3951037 0.20 0.3333333
> #Allograft rejection                         0.9421248 0.03 0.3000000
> #alpha-Linolenic acid metabolism             0.6587026 0.20 0.3333333
```

Apart from the expected arguments: a count data matrix, a vector of class labels and a list of pathways, the user needs to specify the `type` argument which decides on the type of the data (`"MA"` is used for expression microarray and `"RNA-Seq"` for RNA-Seq data). The others arguments are optional. By default, the `"limma"` method is used for the differential expression analysis on gene-level and

TMM for data normalization prior to calculating the TIFs. The user can select DESeq2 by setting argument test to "DESeq2". The node labels of pathway topologies are automaticaly converted into entrezIDs. This is controlled with IDs argument. A converstion into the gene symbols is available too. Please note, that the node labels should be the same as the rownames of count data matrix. The alpha parameter sets a threshold for gene weights. The purpose of this filtering is to reduce the possiblity that a weight of a gene that is tighly correlated with a few genes are lowered by the weak correlation with other genes in a pathway.The implementation returns also a gene-level statistics of the differential expression of genes and the user can select between log fold-change (gene.stat="logFC") or test statistic (gene.stat="stats"). These statistics are later used in the visualization of a selected pathway. The nperm argument controls the number of permutations.

Chapter 4

Outputs and visualization of the results for one pathway

All the functions mentioned in this vignette return an object of class `topResult`. It is a list with three slots. The first one is called `res` and contains a data frame of the results for all the pathways. The actual informations there differ among the methods and are described in the manual. The second slot is called `topo.sig` and it is a list of topological significances of genes in pathways. The term topologial significance means scores used to measure the importance of a gene in a pathway. The higher the score the more important gene. It is `NULL` for TAPPA and DEGraph method, because they do not provide any measure of this kind. The last slot contains the log fold-changes or test statistics of differential expression at gene level. They are necessary in the `plot` function for all the methods except TopologyGSA and Clipper.

The `plot()` function has three necessarry arguments when it is to be applied on `topResult` object. The first one is an output from any of the methods. The second one is either a name of a pathway or its number in a list of pathways. And the last one is a list of pathways used in the analysis.

The final visualization of the results for one pathway is method specific. Three arguments that are common to all methods are:

- `IDs` - the type of gene labels in the original data, `"entrez"` by default
- `graphIDS` - the type of gene labels to be used in plot, `"symbol"` by default
- `layout` - the layout of the graph from `Rgraphviz` package, `"dot"` by default, other possibilities are e.g. `"neato"` or `"twopi"`

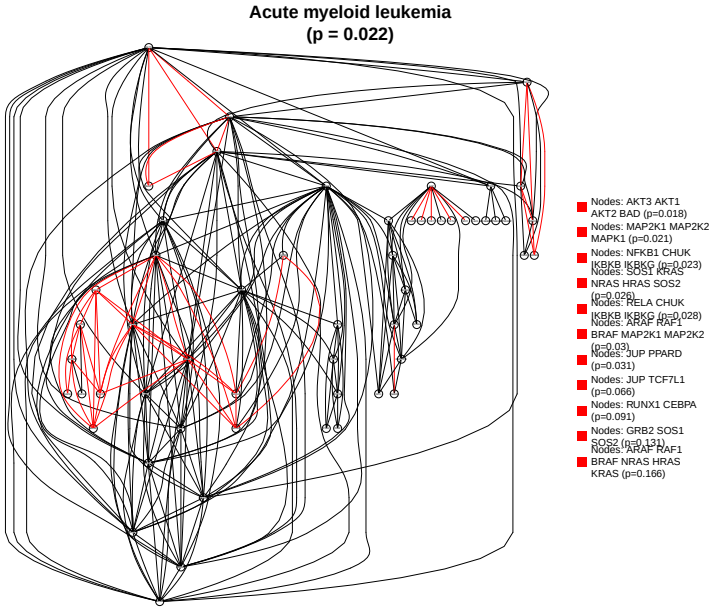
The significant cliques are enhanced in the results of TopologyGSA and Clipper. Since the whole analysis with these method is done on transformed topology (moralized then triangulated graphs), the transformed topology is also drawn in the visualization. The user can specify the color which used for edges between nodes from a significant clique (default value is `cli.color="red"` and

can be either a character or a function that returns a color pallette) and the color of nodes (default value is `cli.node.color="white"`. The `alpha` controls the significance threshold for the cliques. If `add.legend=TRUE` then a legend is drawn containing the colors of edges of individual cliques, their genes and p-value. The `intersp` can be used to adjust the space between items of legened.

```
> library(gageData)
> data(hnrnp.cnts)
> group<-c(rep("sample",4), rep("control",4))
> hnrnp.cnts<-hnrnp.cnts[rowSums(hnrnp.cnts)>0,]
> cli<-Clipper(hnrnp.cnts, group, kegg[1:2], type="RNASeq", testCliques=TRUE)

115 node labels mapped to the expression data
Average coverage 90.28822 %
0 (out of 2) pathways without a mapped node
Normalization method was not specified. TMM used as default
Acute myeloid leukemia
Adherens junction
Testing cliques...
------------------
Acute myeloid leukemia
Adherens junction

> plot(cli,1, kegg)
>
```



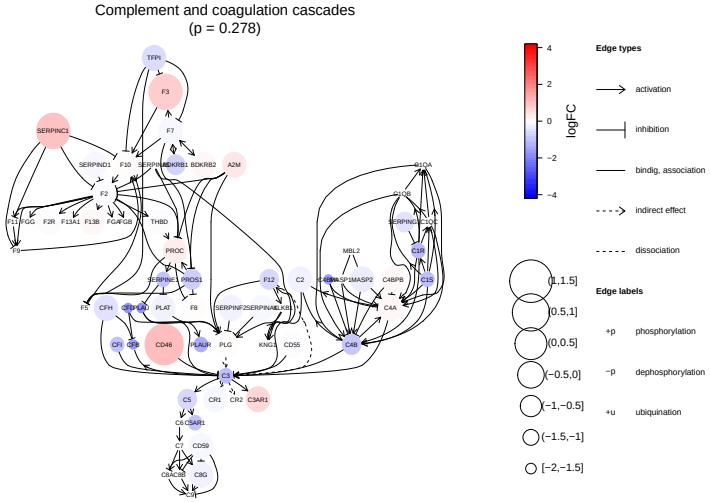


Figure 4.1:

In the visualization of the results from TBS, PWEA or SPIA method, the nodes are colored accoring to the selected gene-level statistic and the size of node reflects the topological significance of a node. Because TAPPA and DEGraph do not provide any specific topological or statistical measure at gene-level, only the coloring of the nodes according to gene-level statistics is used. The user can specify the number of breaks for gene statistics and topological significance of genes (default values are 100 and 5, `breaks=c(100,5)`), colors in the pallete for the gene statistics (default is `pallete.colors=c("blue","white", "red")` and a color for missing nodes `na.col="grey"`. The `stats` argument controls the label of the gene statistics and `title` controls whether the name of a pathway and its p-value should be written as a title. The user can also adjust the size of the nodes (`nodesize`) and font (`fontsize`)

```
> library(gageData)
> data(hnrnp.cnts)
> group<-c(rep("sample",4), rep("control",4))
> hnrnp.cnts<-hnrnp.cnts[rowSums(hnrnp.cnts)>0,]
> spi<-SPIA(hnrnp.cnts, group, kegg[45:50], type="RNASeq", logFC.th=-1)
> plot(spi,"Complement and coagulation cascades", kegg[45:50], fontsize=50)
>
```

Bibliography

[Al-Haj Ibrahim *et al.*(2012)] Al-Haj Ibrahim, M., Jassim, S., Cawthorne, M. A., and Langlands, K. (2012). A topology-based score for pathway enrichment. *J Comput Biol*.

[Anders and Huber(2010)] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

[Dillies *et al.*(2013)] Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloe, D., Le Gall, C., Schaeffer, B., Le Crom, S., Guedj, M., and Jaffrezic, F. (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, **14**(6), 671–683.

[Draghici *et al.*(2007)] Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, **17**(10), 000.

[Gao and Wang(2007)] Gao, S. and Wang, X. (2007). Tappa: topological analysis of pathway phenotype association. *Bioinformatics*, **23**(22), 3100–3102.

[Hung *et al.*(2010)] Hung, J.-H., Whitfield, T., Yang, T.-H., Hu, Z., Weng, Z., and DeLisi, C. (2010). Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biology*, **11**(2), R23.

[Jacob *et al.*(2010)] Jacob, L., Neuvial, P., and Dudoit, S. (2010). Gains in Power from Structured Two-Sample Tests of Means on Graphs. *ArXiv e-prints*.

[Martini *et al.*(2012)] Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2012). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*.

[Massa *et al.*(2010)] Massa, M., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology*, **4**(1), 121.

[R Core Team(2014)] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Robinson and Oshlack(2010] Robinson, M. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biology*, **11**(3), R25.

[Sales *et al.*(2012)] Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). graphite - a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**(1), 20.

[Tarca *et al.*(2009)] Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.