Package 'BgeeDB'

May 27, 2017

Type Package

Title Annotation and gene expression data retrieval from Bgee database

Version 2.2.0

Date 2016-10-06

Author Andrea Komljenovic [aut, cre], Julien Roux [aut, cre]

Maintainer
Andrea Komljenovic <andreakomljenovic@gmail.com>, Frederic Bastian <bgee@sib.swiss>

Description A package for the annotation and gene expression data download from Bgee database, and TopAnat analysis: GO-like enrichment of anatomical terms, mapped to genes by expression patterns.

Depends R (>= 3.3.0), topGO, tidyr

Imports data.table, RCurl, digest, methods, stats, utils, dplyr, graph, Biobase

License GPL-2

VignetteBuilder knitr

biocViews Software, DataImport, Sequencing, GeneExpression, Microarray, GO, GeneSetEnrichment

Suggests knitr, BiocStyle, testthat, rmarkdown

LazyLoad yes

RoxygenNote 6.0.1

NeedsCompilation no

R topics documented:

Bgee-class 2
formatData 3
geneList 4
getAnnotation 5
getData 5
listBgeeRelease 6
listBgeeSpecies 7
loadTopAnatData 8
makeTable 9
topAnat 11

Index 13

`Bgee-class` *Bgee Reference Class*

Description

This is used to specify information at the beginning of a BgeeDB working session, for example, the targeted species and data type. An object of this class is then passed as argument to other functions of the package to provide these informations. See examples in vignette.

Details

Bgee (`http://bgee.org`) integrates different expression data types (RNA-seq, Affymetrix microarray, ESTs, and in-situ hybridizations) from multiple animal species. Expression patterns are based exclusively on curated "normal", healthy, expression data (e.g., no gene knock-out, no treatment, no disease), to provide a reference atlas of normal gene expression.

Fields

`species` A character indicating the species to be used, in the form "Genus_species", or a numeric indicating the species NCBI taxonomic id. Only species with data in Bgee will work. See the listBgeeSpecies() function to get the list of species available in the Bgee release used.

`dataType` A vector of characters indicating data type(s) to be used. To be chosen among:

- "rna_seq"
- "affymetrix"
- "est"
- "in_situ"

By default all data type are included: `c("rna_seq","affymetrix","est","in_situ")`. For download of quantitative expression data, a single data type should be chosen among "rna_seq" or 'affymetrix".

`pathToData` Path to the directory where the data files are stored. By default the working directory is used. If many analyses are launched in parallel, please consider re-using the cached data files instead of redownlaoding them for each analysis.

`release` Bgee release number to download data from, in the form "Release.subrelease" or "Release_subrelease", e.g., "13.2" or 13_2". Will work for release >=13.2. By default, the latest relase of Bgee is used.

`sendStats` A field specifying whether monitoring of users is performed for our internal usage statistics. This is useful to improve the settings of our servers and to get reliable usage statistics (e.g., when asking for funding for Bgee). No identification of the users is attempted, nor possible. Default to TRUE. This option can be set to FALSE, notably if all data files are in cache and that users want to be able to work offline.

`quantitativeData` A field specifying if a single type of quantitative expression data ("rna_seq" or "affymetrix") was specified and if it is available for targeted species, helping the package to know if it should proceed with the execution of getAnnotation() and getData() functions.

Examples

```
{
 bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
 bgee <- Bgee$new(species = "Mus_musculus")
}
```

`formatData` *Format RNA-seq or Affymetrix data downloaded from Bgee.*

Description

This function formats the data downloaded with the getData() function into an object of the Bioconductor "expressionSet" Class.

Usage

```
formatData(myBgeeObject, data, stats = NULL, callType = "all")
```

Arguments

`myBgeeObject`	A Reference Class Bgee object, notably specifying the targeted species and data type.
`data`	A list of data frames including data from multiple experiments, or a data frame including data from a single experiment.
`stats`	A character indicating what expression values should be used in the formatted data expressionSet object matrix. • "rpkm" for RNA-seq • "counts" for RNA-seq • "tpm" for RNA-seq (Bgee release 14 and above) • "intensities" for Affymetrix microarrays
`callType`	A character indicating whether intensities should be displayed only for present (i.e., expressed) genes, present high quality genes, or all genes (default). • "present" • "present high quality" • "all"

Value

If data was a list of data frames from multiple experiments, returns a list of ExpressionSet objects. If data was a data frame from a single experiment, returns an ExpressionSet object.

Author(s)

Andrea Komljenovic and Julien Roux.

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  dataMouseGSE30617 <- getData(bgee, experimentId = "GSE30617")
  dataMouseGSE30617.rpkm <- formatData(bgee,
                                       dataMouseGSE30617,
                                       callType = "present",
                                       stats = "rpkm")
}
```

`geneList` *Example of gene list object used to run a topAnat enrichment test, created on April 18th, 2017. The format of the gene list is the same as the gene list required to build a "'topGOdata"' object in the "'topGO"' package: a vector with background genes as names, and 0 or 1 values depending if a gene is in the foreground or not. In this example the foreground genes are zebrafish genes annotated with "spermatogenesis" in the Gene Ontology (or annotated to children terms of "spermatogenesis"), and the background is composed of all zebrafish genes with at least one Gene Ontology annotation. The gene list was built using the biomaRt package, and the code used can be found in the vignette of the package.*

Description

Example of gene list object used to run a topAnat enrichment test, created on April 18th, 2017. The format of the gene list is the same as the gene list required to build a "'topGOdata"' object in the "'topGO"' package: a vector with background genes as names, and 0 or 1 values depending if a gene is in the foreground or not. In this example the foreground genes are zebrafish genes annotated with "spermatogenesis" in the Gene Ontology (or annotated to children terms of "spermatogenesis"), and the background is composed of all zebrafish genes with at least one Gene Ontology annotation. The gene list was built using the biomaRt package, and the code used can be found in the vignette of the package.

Usage

```
geneList
```

Format

A named vector of factor values with 22141 elements. The factor levels are "0" for the 22103 genes in the background and "1" for the 38 genes in the foreground. Vector names are the Ensembl IDs of the zebrafish genes.

`getAnnotation` *Retrieve Bgee experiments annotation for targeted species and data type.*

Description

This function loads the annotation of experiments and samples of quantitative expression datasets (rna_seq, affymetrix) that are available from Bgee.

Usage

```
getAnnotation(myBgeeObject)
```

Arguments

`myBgeeObject`	A Reference Class Bgee object, notably specifying the targeted species and data type.

Value

A list of two elements, including a data frame of the annotation of experiments for chosen species (field "experiment.annotation") and a data frame of the annotation of chips/libraries from these experiments (field "sample.annotation").

Author(s)

Andrea Komljenovic and Julien Roux.

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  myAnnotation <- getAnnotation(bgee)
}
```

`getData` *Retrieve Bgee RNA-seq or Affymetrix data.*

Description

This function loads the quantitative expression data and presence calls for samples available from Bgee (rna_seq, affymetrix).

Usage

```
getData(myBgeeObject, experimentId = NULL)
```

Arguments

myBgeeObject	A Reference Class Bgee object, notably specifying the targeted species and data type.
experimentId	An ArrayExpress or GEO accession, e.g., GSE30617. Default is NULL: takes all available experiments for targeted species and data type.

Value

If experimentId is not specified, returns a list of data frames with data from all experiments for targeted species and data type. If experimentId is specified, returns a data frame with data from this experiment.

Author(s)

Andrea Komljenovic and Julien Roux.

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  dataMouse <- getData(bgee)
  dataMouseGSE30617 <- getData(bgee, experimentId = "GSE30617")
}
```

---

listBgeeRelease *List Bgee releases available to use with BgeeDB package*

Description

Returns information on available Bgee releases, the access URL for FTP and webservice, and the date of release

Usage

```
listBgeeRelease(release = NULL)
```

Arguments

release	A character specifying a targeted release number. In the form "Release.subrelease" or "Release_subrelease", e.g., "13.2" or 13_2". If not specified, all available releases are shown.

Value

A data frame with information on Bgee releases.

Author(s)

Julien Roux

Examples

```
{
 listBgeeRelease()
}
```

`listBgeeSpecies` *List species in the Bgee database and the available data types for each of them*

Description

Returns information on available species in Bgee

Usage

```
listBgeeSpecies(release = NULL, ordering = NULL, allReleases = NULL,
  removeFile = TRUE)
```

Arguments

`release`	A character specifying a targeted release number. In the form "Release.subrelease" or "Release_subrelease", e.g., "13.2" or 13_2". If not specified, the latest release is used.
`ordering`	A numeric indicating which column should be used to sort the data frame. Default NULL, returning unsorted data frame.
`allReleases`	A data frame with information on all releases. Avoid redownloading this information if .getBgeeRelease() already called.
`removeFile`	Boolean indicating whether the downloaded file should be deleted. Default to TRUE.

Value

A data frame with species Id, genus name, species name, common name and data type availability for targeted Bgee release

Author(s)

Julien Roux

Examples

```
{
 listBgeeSpecies()
 # species present in a specific Bgee release
 listBgeeSpecies(release = "13.2")
 # in order to order species according to theit taxonomical IDs
 listBgeeSpecies(ordering = 1)
}
```

`loadTopAnatData` *Retrieve data from Bgee to perform GO-like enrichment of anatomical terms, mapped to genes by expression patterns.*

Description

This function loads a mapping from genes to anatomical structures based on calls of expression in anatomical structures. It also loads the structure of the anatomical ontology.

Usage

```
loadTopAnatData(myBgeeObject, callType = "presence", confidence = "all",
  stage = NULL)
```

Arguments

`myBgeeObject`	An output object from Bgee$new().
`callType`	A character of indicating the type of expression calls to be used for enrichment. Only calls for significant detection of expression are implemented so far ("presence"). Differential expression calls, based on differential expression analysis, might be implemented in the future.
`confidence`	A character indicating if only high quality present calls should be retrieved. Options are "all" or "high_quality". Default is "all".
`stage`	A character indicating the targeted developmental stages for the analysis. Developmental stages can be chosen from the developmental stage ontology used in Bgee (available at https://github.com/obophenotype/developmental-stage-ontologies). If a stage is specified, the expression pattern mapped to this stage and all children developmental stages (substages) will be retrieved. Default is NULL, meaning that expression patterns of genes are retrieved regardless of the developmental stage displaying expression; this is equivalent to specifying stage="UBERON:0000104" (life cycle, the root of the stage ontology). For information, the most useful stages (going no deeper than level 3 of the ontology) include:

- UBERON:0000068 (embryo stage)
  - UBERON:0000106 (zygote stage)
  - UBERON:0000107 (cleavage stage)
  - UBERON:0000108 (blastula stage)
  - UBERON:0000109 (gastrula stage)
  - UBERON:0000110 (neurula stage)
  - UBERON:0000111 (organogenesis stage)
  - UBERON:0007220 (late embryonic stage)
  - UBERON:0004707 (pharyngula stage)
- UBERON:0000092 (post-embryonic stage)
  - UBERON:0000069 (larval stage)
  - UBERON:0000070 (pupal stage)
  - UBERON:0000066 (fully formed stage)

Details

The expression calls come from Bgee (http://bgee.org), that integrates different expression data types (RNA-seq, Affymetrix microarray, ESTs, or in-situ hybridizations) from multiple animal species. Expression patterns are based exclusively on curated "normal", healthy, expression data (e.g., no gene knock-out, no treatment, no disease), to provide a reference atlas of normal gene expression. Anatomical structures are identified using IDs from the Uberon ontology (browsable at http://www.ontobee.org/ontology/UBERON). The mapping from genes to anatomical structures includes only the evidence of expression in these specific structures, and not the expression in their substructures (i.e., expression data are not propagated). The retrieval of propagated expression data might be implemented in the future, but meanwhile, it can be obtained using specialized packages such as topGO, see the `topAnat.R` function.

Value

A list of 4 elements:

- A `gene2anatomy` list, mapping genes to anatomical structures based on expression calls.
- A `organ.names` data frame, with the name corresponding to UBERON IDs.
- A `organ.relationships` list, giving the relationships between anatomical structures in the UBERON ontology (based on parent-child "is_a" and "part_of" relationships).
- The Bgee class object thta was used to retrieve the data.

Author(s)

Julien Roux

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  myTopAnatData <- loadTopAnatData(bgee)
}
```

---

makeTable *Formats results of the enrichment test on anatomical structures.*

Description

This function loads the results from the topGO test and creates an output table with organ names, fold enrichment and FDR. Data are sorted by p-value and only terms below the specified FDR cutoff are included.

Usage

```
makeTable(topAnatData, topAnatObject, results, cutoff = 1, ordering = 7)
```

Arguments

topAnatData	A list produced by the function loadTopAnatData().
topAnatObject	An object produced by the function topAnat().
results	A result object, produced by the runtest() function of topGO.
cutoff	An FDR cutoff between 0 and 1. Only terms with FDR lower than this cutoff are included. Default is 1, meaning that all terms are included.
ordering	A numeric indicating which column should be used to sort the data frame. If the column number is preceded by a \"-\" sign, results are displayed in decreasing ordering. Default is "7", returning data frame sorted by p-values in increasing order.

Value

A data frame with significantly enriched anatomical structures, sorted by p-value.

Author(s)

Julien Roux

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  myTopAnatData <- loadTopAnatData(bgee)
  geneList <- as.factor(c(rep(0, times=90), rep(1, times=10)))
  names(geneList) <- c("ENSMUSG00000064370", "ENSMUSG00000064368", "ENSMUSG00000064367",
                       "ENSMUSG00000064363", "ENSMUSG00000065947", "ENSMUSG00000064360",
                       "ENSMUSG00000064358", "ENSMUSG00000064357", "ENSMUSG00000064356",
                       "ENSMUSG00000064354", "ENSMUSG00000064351", "ENSMUSG00000064345",
                       "ENSMUSG00000064341", "ENSMUSG00000029757", "ENSMUSG00000079941",
                       "ENSMUSG00000053367", "ENSMUSG00000016626", "ENSMUSG00000037816",
                       "ENSMUSG00000036781", "ENSMUSG00000022519", "ENSMUSG00000079606",
                       "ENSMUSG00000068966", "ENSMUSG00000038608", "ENSMUSG00000047473",
                       "ENSMUSG00000038542", "ENSMUSG00000025386", "ENSMUSG00000028145",
                       "ENSMUSG00000024816", "ENSMUSG00000020978", "ENSMUSG00000055373",
                       "ENSMUSG00000038155", "ENSMUSG00000046408", "ENSMUSG00000030032",
                       "ENSMUSG00000042249", "ENSMUSG00000071909", "ENSMUSG00000039670",
                       "ENSMUSG00000032501", "ENSMUSG00000054252", "ENSMUSG00000068071",
                       "ENSMUSG00000067578", "ENSMUSG00000074892", "ENSMUSG00000027905",
                       "ENSMUSG00000058216", "ENSMUSG00000078754", "ENSMUSG00000062101",
                       "ENSMUSG00000043633", "ENSMUSG00000071350", "ENSMUSG00000021639",
                       "ENSMUSG00000059113", "ENSMUSG00000049115", "ENSMUSG00000053310",
                       "ENSMUSG00000043832", "ENSMUSG00000063767", "ENSMUSG00000026775",
                       "ENSMUSG00000038537", "ENSMUSG00000078716", "ENSMUSG00000096820",
                       "ENSMUSG00000075089", "ENSMUSG00000049971", "ENSMUSG00000014303",
                       "ENSMUSG00000056054", "ENSMUSG00000033082", "ENSMUSG00000020801",
                       "ENSMUSG00000030590", "ENSMUSG00000026188", "ENSMUSG00000014301",
                       "ENSMUSG00000073491", "ENSMUSG00000014529", "ENSMUSG00000036960",
                       "ENSMUSG00000058748", "ENSMUSG00000047388", "ENSMUSG00000002204",
                       "ENSMUSG00000034285", "ENSMUSG00000109129", "ENSMUSG00000035275",
                       "ENSMUSG00000051184", "ENSMUSG00000034424", "ENSMUSG00000041828",
                       "ENSMUSG00000029416", "ENSMUSG00000030468", "ENSMUSG00000029911",
                       "ENSMUSG00000055633", "ENSMUSG00000027495", "ENSMUSG00000029624",
                       "ENSMUSG00000045518", "ENSMUSG00000074259", "ENSMUSG00000035228",
```

```
                    "ENSMUSG00000038533", "ENSMUSG00000030401", "ENSMUSG00000014602",
                    "ENSMUSG00000041827", "ENSMUSG00000042345", "ENSMUSG00000028530",
                    "ENSMUSG00000038722", "ENSMUSG00000075088", "ENSMUSG00000039629",
                    "ENSMUSG00000067567", "ENSMUSG00000057594", "ENSMUSG00000005907",
                    "ENSMUSG00000027496")
  myTopAnatObject <- topAnat(myTopAnatData, geneList)
  resFis <- runTest(myTopAnatObject, algorithm = 'elim', statistic = 'fisher')
  ## Format results
  tableOver <- makeTable(myTopAnatData, myTopAnatObject, resFis, 0.1)
}
```

`topAnat` *Produces an object allowing to perform GO-like enrichment of anatomical terms using the topGO package*

Description

This function produces a topAnatObject, ready to use for gene set enrichment testing using functions from the topGO package. This object uses the Uberon ontology instead of the GO ontology.

Usage

```
topAnat(topAnatData, geneList, nodeSize = 10, ...)
```

Arguments

`topAnatData`	a list including a gene2anatomy list, an organ.relationships list and an organ.names data.frame, produced by the function loadTopAnatData().
`geneList`	Vector indicating foreground and background genes. Names of the vector indicate the background genes. Values are 1 (gene in foreground) or 0 (gene not in foreground).
`nodeSize`	Minimum number of genes mapped to a node for it to be tested. Default is 10.
`...`	Additional parameters as passed to build topGOdata object in topGO package.

Details

To perform the enrichment test for expression in anatomical structures for each term of Uberon ontology (browsable at http://www.ontobee.org/ontology/UBERON), the data are formatted to use the topGO package for testing. This package is interesting because it propagates the mapping of gene to terms to parent terms, and it possesses a pannel of enrichment tests and decorrelation methods. Expert users should be able to use information from the topAnatObject to test enrichment with other packages than topGO.

Value

topAnatObject, a topAnatData class object, ready for gene set enrichment testing with topGO.

Author(s)

Julien Roux

Examples

```
{
  bgee <- Bgee$new(species = "Mus_musculus", dataType = "rna_seq")
  myTopAnatData <- loadTopAnatData(bgee)
  geneList <- as.factor(c(rep(0, times=90), rep(1, times=10)))
  names(geneList) <- c("ENSMUSG00000064370", "ENSMUSG00000064368", "ENSMUSG00000064367",
                       "ENSMUSG00000064363", "ENSMUSG00000065947", "ENSMUSG00000064360",
                       "ENSMUSG00000064358", "ENSMUSG00000064357", "ENSMUSG00000064356",
                       "ENSMUSG00000064354", "ENSMUSG00000064351", "ENSMUSG00000064345",
                       "ENSMUSG00000064341", "ENSMUSG00000029757", "ENSMUSG00000079941",
                       "ENSMUSG00000053367", "ENSMUSG00000016626", "ENSMUSG00000037816",
                       "ENSMUSG00000036781", "ENSMUSG00000022519", "ENSMUSG00000079606",
                       "ENSMUSG00000068966", "ENSMUSG00000038608", "ENSMUSG00000047473",
                       "ENSMUSG00000038542", "ENSMUSG00000025386", "ENSMUSG00000028145",
                       "ENSMUSG00000024816", "ENSMUSG00000020978", "ENSMUSG00000055373",
                       "ENSMUSG00000038155", "ENSMUSG00000046408", "ENSMUSG00000030032",
                       "ENSMUSG00000042249", "ENSMUSG00000071909", "ENSMUSG00000039670",
                       "ENSMUSG00000032501", "ENSMUSG00000054252", "ENSMUSG00000068071",
                       "ENSMUSG00000067578", "ENSMUSG00000074892", "ENSMUSG00000027905",
                       "ENSMUSG00000058216", "ENSMUSG00000078754", "ENSMUSG00000062101",
                       "ENSMUSG00000043633", "ENSMUSG00000071350", "ENSMUSG00000021639",
                       "ENSMUSG00000059113", "ENSMUSG00000049115", "ENSMUSG00000053310",
                       "ENSMUSG00000043832", "ENSMUSG00000063767", "ENSMUSG00000026775",
                       "ENSMUSG00000038537", "ENSMUSG00000078716", "ENSMUSG00000096820",
                       "ENSMUSG00000075089", "ENSMUSG00000049971", "ENSMUSG00000014303",
                       "ENSMUSG00000056054", "ENSMUSG00000033082", "ENSMUSG00000020801",
                       "ENSMUSG00000030590", "ENSMUSG00000026188", "ENSMUSG00000014301",
                       "ENSMUSG00000073491", "ENSMUSG00000014529", "ENSMUSG00000036960",
                       "ENSMUSG00000058748", "ENSMUSG00000047388", "ENSMUSG00000002204",
                       "ENSMUSG00000034285", "ENSMUSG00000109129", "ENSMUSG00000035275",
                       "ENSMUSG00000051184", "ENSMUSG00000034424", "ENSMUSG00000041828",
                       "ENSMUSG00000029416", "ENSMUSG00000030468", "ENSMUSG00000029911",
                       "ENSMUSG00000055633", "ENSMUSG00000027495", "ENSMUSG00000029624",
                       "ENSMUSG00000045518", "ENSMUSG00000074259", "ENSMUSG00000035228",
                       "ENSMUSG00000038533", "ENSMUSG00000030401", "ENSMUSG00000014602",
                       "ENSMUSG00000041827", "ENSMUSG00000042345", "ENSMUSG00000028530",
                       "ENSMUSG00000038722", "ENSMUSG00000075088", "ENSMUSG00000039629",
                       "ENSMUSG00000067567", "ENSMUSG00000057594", "ENSMUSG00000005907",
                       "ENSMUSG00000027496")
  myTopAnatObject <- topAnat(myTopAnatData, geneList, nodeSize=1)
}
```

Index

∗Topic **datasets**
geneList, 4

Bgee *(*Bgee-class*)*, 2
Bgee-class, 2

formatData, 3

geneList, 4
getAnnotation, 5
getData, 5

listBgeeRelease, 6
listBgeeSpecies, 7
loadTopAnatData, 8

makeTable, 9

topAnat, 11