Package 'gageData'

April 4, 2014

Type Package

Title Auxillary data for gage package

Version 2.0.3

Date 2013-9-24

Author Weijun Luo

Maintainer Weijun Luo <luo_weijun@yahoo.com>

Description This is a supportive data package for the software package, gage. However, the data supplied here are also useful for gene set or pathway analysis or microarray data analysis in general. In this package, we provide two demo microarray dataset: GSE16873 (a breast cancer dataset from GEO) and BMP6 (originally published as an demo dataset for GAGE, also registered as GSE13604 in GEO). This package also includes commonly used gene set data based on KEGG pathways and GO terms for major research species, including human, mouse, rat and budding yeast. Mapping data between common gene IDs for budding yeast are also included.

biocViews ExperimentData, Cancer

Suggests gage, pathview, genefilter

License GPL (>=2.0)

LazyLoad yes

Depends R (>= 2.10)

R topics documented:

bmp6 *A microarray dataset on BMP6 treated mesenchymal stem cells*

Description

This dataset describes mesenchymal stem cell response to BMP6 treatment. This is a typical small dataset with as few as two samples per condition like in most experimental studies. BMP6 treated samples and controls are one-on-one matched. This data has been extensively analyzed in GAGE paper, and was used as the primary demo data in earlier versions of gage package.

Usage

```
data(bmp6)
```

Details

This dataset is also available through Gene Expression Omnibus (GEO) with accession number GSE13604. Notice that bmp6 dataset is processed differently than GSE13604. bmp6 dataset used a updated probe set definition (CDF) file based on Entrez Gene mapping, while GSE13604 used the original CDF based on UniGene mapping.

Source

GEO Dataset GSE13604: <URL: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13604>

References

Luo, W., Friedman, M., Shedden K., Hankenson, K. and Woolf, P GAGE: Generally Applicable Gene Set Enrichment for Pathways Analysis. BMC Bioinformatics 2009, 10:161

Examples

```
data(bmp6)
colnames(bmp6)

#kegg analysis
if(require(gage)){
data(kegg.gs)
lapply(kegg.gs[1:3],head)
head(rownames(bmp6))
bmp6.kegg.p <- gage(bmp6, gsets = kegg.gs,
    ref =c(1,3), samp = c(2,4))
}
```

genesets *Common gene set data collections*

Description

The gene set data collections derived from KEGG, GO and BioCarta databases.

Usage

```
data(kegg.sets.hs)
data(go.sets.hs)
data(carta.hs)
data(kegg.sets.mm)
data(go.sets.mm)
data(kegg.sets.rn)
data(go.sets.rn)
data(kegg.sets.sc)
data(go.sets.sc)

data(sigmet.idx.hs)
data(go.subs.hs)
data(sigmet.idx.mm)
data(go.subs.mm)
data(sigmet.idx.rn)
data(go.subs.rn)
data(sigmet.idx.sc)
data(go.subs.sc)
```

Format

kegg.sets.hs is a named list of 229 elements. Each element is a character vector of member gene Entrez IDs for a single KEGG pathway. Type `head(kegg.sets.hs, 3)` for the first 3 gene sets or pathways.

go.sets.hs is a named list of 17202 elements. Each element is a character vector of member gene Entrez IDs for a single Gene Ontology term. Type `head(go.sets.hs, 3)` for the first 3 gene sets or GO terms.

sigmet.idx.hs is a index numbers of sinaling and metabolic pathways in kegg.set.gs. In other words, KEGG pathway include other types of pathway definitions, like "Global Map" and "Human Diseases", which are frequently undesirable in pathway analysis. Therefore, `kegg.sets.hs[sigmet.idx.hs]` gives you the "cleaner" gene sets of sinaling and metabolic pathways only.

go.subs.hs is a named list of three elements: "BP", "CC" and "MF", corresponding to biological process, cellular component and molecular function subtrees. It may be more desirable to conduct separated GO enrichment test on each of these 3 subtrees as shown in the example code.

carta.hs is a named list of 259 elements. Each element is a character vector of member gene Entrez IDs for a single BioCarta pathway. Type `head(carta.hs, 3)` for the first 3 gene sets or pathways.

These are just KEGG, GO and BioCarta gene sets for the default species, i.e. human. KEGG or GO geen sets for other species including mouse (.mm), rat (.rn) and yeast (.sc) have similar structure as their counterparts for human. In addition to the individual species, KEGG gene sets for KEGG Orthology (.ko) is also provided. This is useful for metagenomics, microbiome and non-KEGG species data analysis.

Details

The human gene set data were compiled using Entrez Gene IDs, gene set names and mapping information from multiple Bioconductor packages, including: org.Hs.eg.db, KEGG.db, GO.db and cMAP. Please check the corresponding packages for more information.

Gene set for other 3 species included here, was built similarly. The users are encourage to build their own gene set collections for more species in a similar way or to use the Bioconductor package GSEABase.

Source

Human data come from multiple Bioconductor packages, including: org.Hs.eg.db, KEGG.db, GO.db and cMAP.

References

Entrez Gene <URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene> KEGG pathways <URL: ftp://ftp.genome.ad.jp/pub/kegg/pathways> Gene Ontology <URL: http://www.geneontology.org/> cMAP <URL: http://cmap.nci.nih.gov/PW>

Examples

```
if(require(gage)){
#load expression and gene set data
data(gse16873)
cn=colnames(gse16873)
hn=grep(HN,cn, ignore.case =TRUE)
dcis=grep(DCIS,cn, ignore.case =TRUE)

data(kegg.sets.hs)
data(sigmet.idx.hs)
kegg.sigmet.hs=kegg.sets.hs[sigmet.idx.hs]

#make sure the gene IDs are the same for expression data and gene set
#data
lapply(kegg.sets.hs[1:3],head)
head(rownames(gse16873))

#GAGE analysis with signaling and metabolic pathways
gse16873.kegg.sets.p <- gage(gse16873, gsets = kegg.sigmet.hs,
    ref = hn[1:3], samp = dcis[1:3])

data(go.sets.hs)
data(go.subs.hs)
names(go.subs.hs)
```

```
go.mf.hs=go.sets.hs[go.subs.hs$MF]

#GAGE analysis with GO Molecular Function gene sets
gse16873.kegg.sets.p <- gage(gse16873, gsets = go.mf.hs,
    ref = hn[1:3], samp = dcis[1:3])
}
```

`gse16873.full` *GSE16873: a breast cancer microarray dataset*

Description

GSE16873 is a breast cancer study (Emery et al, 2009) downloaded from Gene Expression Omnibus (GEO). GSE16873 covers twelve patient cases, each with HN (histologically normal), ADH (ductal hyperplasia), and DCIS (ductal carcinoma in situ) RMA samples. Hence, there are 12*3=36 microarray hybridizations or samples interesting to us plus 4 others less interesting in the full dataset, gse16873.full. Dataset gse16873 in gage and gse16873.2 in this package are half datasets each with only HN and DCIS samples of 6 patients. Dataset gse16873.affyid is similar to gse16873 in gage package, except that row IDs are Affymetrix probe set IDs instead of Entrez Gene IDs. This is becuase Affymetrix original CDF (hgu133a) instead of Entrez Gene based on CDF was used when processing the raw data.

Details section below gives more information on the datasets.

Usage

```
data(gse16873.full)
data(gse16873.2)
data(gse16873.affyid)
```

Details

Due to the size limit of the software package gage, we split GSE16873 into two halves, each including 6 patients with their HN and DCIS but not ADH tissue types. The gage package only includes the first half dataset for 6 patients as the example dataset gse16873. Most of our demo analyses are done on the first half dataset, except for the advanced analysis where we use both halves datasets with all 12 patients.

Raw data for these two half datasets were processed separately using two different methods, FARMS and RMA, respectively to generate the non-biological data heterogeneity. The first half dataset is named as gse16873, the second half dataset named gse16873.2. We also have this full dataset, gse16873.full, which includes all HN, ADH and DCIS samples of all 12 patients, processed together using FARMS.

Source

GEO Dataset GSE16873: <URL: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16873>

References

Emery LA, Tripathi A, King C, Kavanah M, Mendez J, Stone MD, de las Morenas A, Sebastiani P, Rosenberg CL: Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. Am J Pathol 2009, 175:1292-302.

Examples

```
##usage of the full dataset
data(gse16873.full)
#column/smaple names
cn=colnames(gse16873.full)
hn=grep(HN,cn, ignore.case =TRUE)
sh=grep(SH,cn, ignore.case =TRUE)
adh=grep(ADH,cn, ignore.case =TRUE)
dcis=grep(DCIS,cn, ignore.case =TRUE)

#multi-state comparison based on f-test
fac=rep(NA, 40)
fac[hn]=hn
fac[sh]=sh
fac[adh]=adh
fac[dcis]=dcis
if (require(genefilter)){
    fstats=rowFtests(gse16873.full[, -sh],
    as.factor(fac[-sh]))[,1,drop=FALSE]
    fstats=cbind(fstat=fstats)

    ## Not run:
    if(require(gage)){
        data(kegg.gs)
        lapply(kegg.gs[1:6],head)
        head(rownames(fstats))
        #feed fstats as single-column matrix into gage
        gse16873.fstats.kegg.p <- gage(fstats, gsets = kegg.gs,
                  ref = NULL, samp = NULL)
        head(gse16873.fstats.kegg.p$greater)
    }

## End(Not run)
}

##for usage of the half datasets, check the help information for
##heter.gage function in the gage package.

#use of gse16873.affyid
## Not run:
if(require(hgu133a.db) & require(gage)){
data(gse16873.affyid)
affyid=rownames(gse16873.affyid)
egids2 <- hgu133aENTREZID[affyid]
annots=toTable(egids2)
str(annots)
```

```
gse16873.affyid=gse16873.affyid[annots$probe_id,]
#if multiple probe sets map to a gene, select the one with maximal IQR
iqrs=apply(gse16873.affyid, 1, IQR)
sel.rn=tapply(1:nrow(annots), annots$gene_id, function(x){
x[which.max(iqrs[x])]
})
gse16873.egid=gse16873.affyid[sel.rn,]
rownames(gse16873.egid)=names(sel.rn)

cn=colnames(gse16873.egid)
hn=grep(HN,cn, ignore.case =T)
dcis=grep(DCIS,cn, ignore.case =T)
data(kegg.gs)
gse16873.kegg.p.affy <- gage(gse16873.egid, gsets = kegg.gs,
    ref = hn, samp = dcis)
#result should be similar to that of using gse16873
}

## End(Not run)
```

`hnrnp.cnts` *RNA-seq dataset on HNRNPC knockdown and control HeLa cells*

Description

This dataset describes HeLa cell response to RNA-binding protein hnRNP C (HNRNPC) knock down. There are two replicate samples from each HNRNPC knockdown condition/siRNA (KD1 and KD2). In addition, there are four control HeLa cell samples. This is a typical RNA-seq dataset with two experimental groups. Experiment and control samples in this study should be treated as unpaired. This data is used as the primary demo data in the RNA-seq pathway analysis workflow of gage package.

Usage

```
data(hnrnp.cnts)
```

Details

This dataset is also available through EBI ArrayExpress with accession number E-MTAB-1147. The raw reads data in zipped FASTQ format (fastq.gz) were downloaded and mapped to human reference genome (hg19)using tophat2. Then reads mapped to annotated gene regions were counted using `summarizeOverlaps` function from `GenomicRanges` package.

Source

EBI ArrayExpress Experiment E-MTAB-1147: <URL: http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1147/>

References

Luo, W., Friedman, M., Shedden K., Hankenson, K. and Woolf, P GAGE: Generally Applicable Gene Set Enrichment for Pathways Analysis. BMC Bioinformatics 2009, 10:161

Examples

```
#data preparation
data(hnrnp.cnts)
cnts=hnrnp.cnts
libsizes=colSums(cnts)
size.factor=libsizes/exp(mean(log(libsizes)))
cnts.norm=t(t(cnts)/size.factor)
sel.rn=rowSums(cnts.norm) != 0
cnts.norm=cnts.norm[sel.rn,]
cnts.norm=log2(cnts.norm+8)

#kegg analysis
if(require(gage)){
data(kegg.gs)
lapply(kegg.gs[1:3],head)
head(rownames(cnts))
ref.idx=5:8
samp.idx=1:4
cnts.kegg.p <- gage(cnts.norm, gsets = kegg.gs,
    ref = ref.idx, samp = samp.idx, compare ="unpaired")
}

#GO analysis
if(require(gage)){
data(go.sets.hs)
data(go.subs.hs)
#molecular function (MF) terms only here as a quick example,
#biological process (BP) and cellular component (CC) term analysis
#could be even more informative.
cnts.mf.p <- gage(cnts.norm, gsets = go.sets.hs[go.subs.hs$MF],
    ref = ref.idx, samp = samp.idx, compare ="unpaired")
}
```

---

sc.gene *Common IDs used for budding yeast (Saccharomyces cerevisa) genes*

Description

These two data provide mapping between Entrez IDs, official symbols and ORF (open reading frame) IDs for budding yeast genes. These data are useful for yeast microarray data analysis. sc.gene is a 3-column matrix listing the Entrez IDs, official symbols and ORF (open reading frame) IDs for all known genes. orf2eg is a named vector mapping ORF IDs to Entrez IDs.

Usage

```
data(sc.gene)
data(orf2eg)
```

Details

These mapping data is may be used together with functions `eg2sym` and `sym2eg` in the gage package or similar functions. Check the examples for these functions in gage package.

Source

These mapping data were compiled using the gene data from NCBI Entrez Gene database.

Similar information can also be derived from Bioconductor package org.Sc.sgd.db. Please check the package for more information.

References

Entrez Gene <URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>

Examples

```
data(sc.gene)
head(sc.gene)
data(orf2eg)
head(orf2eg)

## for more example, check \code{eg2sym} and \code{sym2eg} funtions in
## the gage package.
```

Index