gageData

March 19, 2011

`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.gs)
data(go.gs)
data(carta.gs)
data(kegg.mm)
data(go.mm)
data(kegg.rn)
data(go.rn)
data(kegg.sc)
data(go.sc)
```

Format

kegg.gs is a named list of 205 elements. Each element is a character vector of member gene Entrez IDs for a single KEGG pathway. Type `head(kegg.gs, 3)` for the first 3 gene sets or pathways.

go.gs is a named list of ~10000 elements. Each element is a character vector of member gene Entrez IDs for a single Gene Ontology term. Type `head(go.gs, 3)` for the first 3 gene sets or GO terms.

carta.gs 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.gs, 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.

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.gs)

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

#GAGE analysis
gse16873.kegg.p <- gage(gse16873, gsets = kegg.gs,
    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. Details section below gives more information.

Usage

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

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)

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)
}
}

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

`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