KEGGandMetacoreDzPathwaysGEO : Disease Datasets from GEO

Gaurav Bhatti

17 April 2014

1 Overview of KEGGandMetacoreDzPathwaysGEO data package

`KEGGandMetacoreDzPathwaysGEO` is a collection of 18 GEO datasets for which the phenotype is a disease with a corresponding pathway in either of the two popular gene to pathway annotation databases, KEGG and Metacore. These datasets were used as gold standard in comparing gene set analysis methods [1]. Details about the individual datasets including sample tissue, target disease pathway, etc may be obtained by typing:

```r
> ?KEGGandMetacoreDzPathwaysGEO
```

at the R prompt. In order to access all the datasets available in the package, type the following:

```r
> mysets=data(package="KEGGandMetacoreDzPathwaysGEO")$results[,"Item"]
> mysets
```

The microarray data from the GEO database along with the associated metadata is stored as ExpressionSet class. "The ExpressionSet class is designed to combine several different sources of information into a single convenient structure. An ExpressionSet can be manipulated (e.g., subsetted, copied) conveniently, and is the input or output from many Bioconductor functions." [2]. An example dataset is shown below:

```r
> library(KEGGandMetacoreDzPathwaysGEO)
> data(GSE1145)
> show(GSE1145)
```

ExpressionSet (storageMode: lockedEnvironment)
assayData: 54675 features, 26 samples
  element names: exprs
protocolData: none
A similar data package, *KEGGDzPathwaysGEO*, is already available for installation in Bioconductor. It contains additional 24 GEO datasets for which the phenotype is a disease with a corresponding pathway in the KEGG database. These datasets were used to test the performance of an in-house pathway analysis method which has also been implemented as a Bioconductor package, *PADOG* [3].

These datasets may be used to compare existing gene set pathway analysis methods or to test the performance of novel methods.

**References**

