Package 'pathprintGEOData'

October 18, 2017

Version 1.2.0

Title Pathway fingerprint vectors representing a subsection of arrays from the GEO data repository

Description Pathway Fingerprint vectors that have been pre-calculated for ~160,000 publicly available arrays from the GEO corpus, spanning 6 species and 31 platforms, together with their associated metadata.

Author Gabriel Altschuler

Maintainer Sokratis Kariotis <s.kariotis@sheffield.ac.uk>

License GPL

biocViews ExperimentData, ExpressionData, MicroarrayData, GEO, Homo_sapiens_Data, Mus_musculus_Data, Drosophila_melanogaster_Data, Rattus_norvegicus_Data, Caenorhabditis_elegans_Data, Danio_rerio_Data,Genome, SequencingData, OneChannelData, TwoChannelData,PathwayInteractionDatabase

NeedsCompilation no

Suggests pathprint, SummarizedExperiment

Depends R (>= 3.4)

RoxygenNote 6.0.1

R topics documented:

`pathprintSummarizedGEOData`
GEO pathway fingerprint matrix and their metadata data frame

Description

The package GEOMetaDB (https://bioconductor.org/packages/release/bioc/html/GEOmetadb.html) was used to obtain a list of the most highly represented one-channel gene expression platforms (GPLs) in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) that profiled Homo sapiens, Mus musculus, Rattus norvegicus, Danio rerio, Drosophila melanogaster, and Caenorhabditis elegans and their associated arrays (GSMs). The package GEOquery (https://bioconductor.org/packages/release/bioc/html/GEOquery.html) was used to retrieve normalized expression tables for all of the experiments, all normalization methods were accepted. After discarding incomplete records, the expression data was mapped to Entrez Gene identifications using systematically updated annotations from AILUN (Array Information Library Universal Navigator, http://ailun.ucsf.edu/). Multiple probes were merged to unique Entrez Gene IDs by taking the mean probe set intensity.

H. sapiens canonical pathway gene sets were compiled from Reactome, Wikipathways, and KEGG (Kyoto Encyclopedia of Genes and Genomes). Static modules were constructed independently by decomposing a network that extended curated pathways with non-curated sources of information, including protein-protein interactions, gene co-expression, protein domain interaction, GO annotations and text-mined protein interactions (see Altschuler et al). M. musculus, R. norvegicus, D. rerio, D. melanogaster, and C. elegans gene sets were inferred using homology based on the HomoloGene database (https://www.ncbi.nlm.nih.gov/homologene).

Pathway expression scores were calculated for each pathway in each array based on the mean squared ranked expression of the member genes. The full set of GEO experiments was used to calculate a static pathway expression background distribution for each pathway across each platforms. A signed probability of expression (POE) was calculated based on a two-component uniform-normal mixture model, representing the probability that a pathway expression score has significant low (negative) or high (positive) expression. POE values were converted to a ternary score (-1,0,1) by application of a symmetric threshold to produce the final pathprint matrix.

For complete details of method and references please see Altschuler et al. (2013, PMID: 23890051).

The fingerprint matrix contains ternary scores for 633 pathways that have been pre-calculated for 188,390 publicly available arrays from the GEO corpus, spanning 6 species and 31 platforms, using the method described in Altschuler et al. (2013, PMID: 23890051).

The metadata data frame includes experiment IDs, platform, species and a selection of the record description provided by the GEO database.

Usage

```
data(result)
```

Format

The metadata contain the following 7 variables for 188390 samples.

`GSM` GEO sample ID

`GSE` GEO series ID

`GPL` GEO platform ID

`Species` GEO description - Species

`Title` GEO description - Title

`Source` GEO description - Source

`Characteristics` GEO description - Characteristic

Source

Primary data was retrieved from http://www.ncbi.nlm.nih.gov/geo/

References

Altschuler et al. Pathprinting: An integrative approach to understand the functional basis of disease. *Genome Med* (2013) vol.5 pp. 68

Barrett et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic acids research* (2007) vol. 35 (Database issue) pp. D760-5

Examples

```
library(SummarizedExperiment)
library(pathprintGEOData)
# load  the data
data(compressed_result)

# objects in environment
objects()

# object type
class(result)

# package attribute of resulted object
attr(result,"package")

# SummarizedExperiment objectc specifications
head(result)

# extract GEO.fingerprint.matrix
GEO.fingerprint.matrix = assays(result)$fingerprint

# extract GEO.metadata.matrix
GEO.metadata.matrix = colData(result)

# get dimensions
dim(GEO.fingerprint.matrix)
dim(GEO.metadata.matrix)
```

Index