About pathprintGEOData

May 7, 2020

1 Description

This package contains the data used by the pathprint package, including fingerprint and metadata data frames and chipframe. The fingerprint matrices contain pathway Fingerprint vectors that have been pre-calculated for 390,000 publicly available arrays from the GEO corpus, spanning 6 species and 31 platforms. All data are accompanied by their associated metadata.

The data in this package were obtained using the method described by Altschuler et al. (2013, PMID: 23890051). The package GEOquery was used to retrieve normalized expression tables for all of the experiments of each platform, all normalization methods were accepted. The expression data was mapped to Entrez Gene identifications using systematically updated annotations from AILUN(Array Information Library Universal Navigator). 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. M. musculus, R. norvegicus, D. rerio, D. melanogaster, and C. elegans gene sets were inferred using homology based on the HomoloGene database. 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 platform. 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.

2 Using pathprintGEOData with pathprint package

The data in this package are primarily used by the pathprint package. For the following examples to work, the pathprint package needs to be installed. For further explanations of some of the functions mentioned in the examples please refer to pathprint. Furthermore, the SummarizedExperiment package is required to extract the two matrices from the SummarizedExperiment object.

```
> # use the pathprint library
> library(pathprint)
> library(SummarizedExperiment)
> library(pathprintGEOData)
> # load  the data
> data(SummarizedExperimentGEO)
> ds = c("chipframe", "genesets","pathprint.Hs.gs","platform.thresholds",
+     "pluripotents.frame")
> data(list = ds)
> # see available platforms
> names(chipframe)
 [1] "GPL570"   "GPL1261"  "GPL339"   "GPL96"    "GPL81"    "GPL8321"
 [7] "GPL8300"  "GPL571"   "GPL2986"  "GPL6947"  "GPL6883"  "GPL6104"
[13] "GPL6102"  "GPL6884"  "GPL6887"  "GPL6885"  "GPL6103"  "GPL6105"
[19] "GPL3921"  "GPL4685"  "GPL1319"  "GPL200"   "GPL72"    "GPL1322"
[25] "GPL341"   "GPL85"    "GPL1355"  "GPL2700"  "GPL2995"  "GPL6333"
[31] "GPL91"    "GPL6244"  "GPL6246"  "GPL16570" "GPL16686"
> # extract GEO.fingerprint.matrix and GEO.metadata.matrix
> GEO.fingerprint.matrix = assays(geo_sum_data)$fingerprint
> GEO.metadata.matrix = colData(geo_sum_data)
> # create consensus fingerprint for pluripotent samples
> pluripotent.consensus<-consensusFingerprint(
+     GEO.fingerprint.matrix[,pluripotents.frame$GSM],
+     threshold=0.9)
> # calculate distance from the pluripotent consensus
> geo.pluripotentDistance<-consensusDistance(
+     pluripotent.consensus, GEO.fingerprint.matrix)
[1] "Scaling against max length, 158"
> # plot histograms
> par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
> geo.pluripotentDistance.hist<-hist(
+     geo.pluripotentDistance[,"distance"],
+     nclass = 50, xlim = c(0,1),
+     main = "Distance from pluripotent consensus")
> par(mar = c(7, 4, 4, 2))
> hist(geo.pluripotentDistance[
```

```
+     pluripotents.frame$GSM, "distance"],
+     breaks = geo.pluripotentDistance.hist$breaks,
+     xlim = c(0,1),
+     main = "",
+     xlab = "above: all GEO, below: pluripotent samples")
> # annotate top 100 matches not in original seed with metadata
> geo.pluripotentDistance.noSeed<-geo.pluripotentDistance[
+     !(rownames(geo.pluripotentDistance)
+     %in%
+     pluripotents.frame$GSM),
+     ]
> top.noSeed.meta<-GEO.metadata.matrix[
+     match(
+     head(rownames(geo.pluripotentDistance.noSeed), 100),
+                               rownames(GEO.metadata.matrix)),
+     ]
> print(top.noSeed.meta[, c(1:4)])
DataFrame with 100 rows and 4 columns
                  GSM         GSE         GPL      Species
           <character> <character> <character>  <character>
GSM1017386  GSM1017386    GSE41439      GPL570 Homo sapiens
GSM1017387  GSM1017387    GSE41439      GPL570 Homo sapiens
GSM1017392  GSM1017392    GSE41439      GPL570 Homo sapiens
GSM1032052  GSM1032052    GSE42073      GPL570 Homo sapiens
GSM1032053  GSM1032053    GSE42073      GPL570 Homo sapiens
...               ...         ...         ...          ...
GSM172875    GSM172875     GSE7179      GPL570 Homo sapiens
GSM172877    GSM172877     GSE7179      GPL570 Homo sapiens
GSM172878    GSM172878     GSE7179      GPL570 Homo sapiens
GSM172879    GSM172879     GSE7179      GPL570 Homo sapiens
GSM181756    GSM181756     GSE7500      GPL570 Homo sapiens
```