geneExpressionFromGEO: an R package to facilitate data reading from Gene Expression Omnibus (GEO)

by Davide Chicco
davidechicco@davidechicco.it
www.DavideChicco.it



Institute of Health Policy, Management and Evaluation
UNIVERSITY OF TORONTO

2023-06-03



Springer Link



Microarray Data Analysis pp 187–194 | Cite as

geneExpressionFromGEO: An R Package to Facilitate Data Reading from Gene Expression Omnibus (GEO)

Davide Chicco

Protocol | First Online: 14 December 2021

931 Accesses | **5** Citations

Part of the Methods in Molecular Biology book series (MIMB,volume 2401)

https://doi.org/10.1007/978-1-0716-1839-4_12

Gene Expression Omnibus (GEO) and its datasets

- GEO is an online repository of thousands of datasets of microarray gene expression and RNA methylation developed and maintained by the National Center for Biotechnology Information (NCBI)

- Datasets available on GEO can be used for new scientific analyses and for meta-analyses, by selecting datasets derived from the same platform. For example, one could download 10 datasets of patients with cancer all derived through the Affymetrix HG-U133A Human Genome U133A platform, do the batch correction, and run the same scientific analysis on all of them.

- This aspect is extremely useful for generalizability of scientific discoveries: a list of relevant genes found on one dataset can be confirmed by an analysis done on other datasets

R and Gene Expression Omnibus

- Data on GEO refer to probesets of the platform rather than to the gene symbols. Probesets are sets of fragments of DNA known as hybridization probes, and represent a "precise point" in the genome for that platform.

- Each probeset correspond to a single gene, but a single gene can correspond to multiple probesets

- For example, the 203072_at probeset of the GPL96 Affymetrix platform refer to the *MYO1E* gene symbol (Myosin IE protein coding gene)

- Associating probesets to gene symbols can be done in R through packages such as `jetset` and `annotate`, but it can be difficult for beginners

geneExpressionFromGEO

- CRAN and Bioconductor provide some software libraries to automatically download data from GEO (`GEOquery`) and to annotate gene symbols to probesets (`annotate, jetset`), but they're a bit difficult to use for beginners.

- So I thought about developing a new R package that would do both the jobs in a easier way. geneExpressionFromGEO has two main assets:

  - It allows users to download a dataset from GEO in a easy way, specifying just the GEO dataset code parameter

  - It associates gene symbols with the probesets of the downloaded dataset (feature available for 19 different platforms)

geneExpressionFromGEO

- Install it from CRAN and load it:

```
install.packages("geneExpressionFromGEO")
library("geneExpressionFromGEO")
```

geneExpressionFromGEO: Retrieves Gene Expression Dataset and Gene Symbols from GEO Code

A function that reads in the GEO code of a gene expression dataset, retrieves its data from GEO, (optional) retrieves the gene symbols of the dataset, and returns a simple dataframe table containing all the data. Platforms available: GPL11532, GPL23126, GPL6244, GPL80, GPL8300, GPL80, GPL96, GPL570, GPL571, GPL20115, GPL1293, GPL6102, GPL6104, GPL6883, GPL6884, GPL13497, GPL14550, GPL17077, GPL6480. GEO: Gene Expression Omnibus. ID: identifier code. The GEO datasets are downloaded from the URL <https://ftp.ncbi.nlm.nih.gov/geo/series/>.

Version:	0.4
Depends:	R (≥ 3.1.0)
Imports:	xml2, Biobase, annotate, GEOquery
Suggests:	knitr, rmarkdown
Published:	2021-01-25
Author:	Davide Chicco [aut, cre]
Maintainer:	Davide Chicco <davidechicco at davidechicco.it>
BugReports:	https://stackoverflow.com/questions/tagged/geneExpressionFromGEO
License:	GPL-3
URL:	https://github.com/davidechicco/geneExpressionFromGEO
NeedsCompilation:	no
CRAN checks:	geneExpressionFromGEO results

geneExpressionFromGEO

- Example on the GSE3268 dataset: gene expression profile of squamous lung cancer cells are used to identify genes that are differentially regulated



geneExpressionFromGEO

- Example on the GSE3268 dataset:

```
associateSymbolsToGenes <- TRUE
verbose <- TRUE
datasetCode <- "GSE3268"

geneExpressionDF <- getGeneExpressionFromGEO(datasetCode,
associateSymbolsToGenes, verbose)
```

- The dataset is stored into `geneExpressionDF`, which is an R dataframe of 22,283 rows (probesets) and 11 columns (10 samples and the gene symbol variable)

geneExpressionFromGEO

```
> geneExpressionDF <- getGeneExpressionFromGEO("GSE3268",  associateSymbolsToGenes, verb
ose)
Processed URL: https://ftp.ncbi.nlm.nih.gov/geo/series/GSE3nnn/GSE3268
Found 1 file(s)
GSE3268_series_matrix.txt.gz
Using locally cached version: /tmp/RtmpzJdbrl/GSE3268_series_matrix.txt.gz
[1] "sort(names(platform_ann_df))"
 [1] "Annotation Date"                  "ENTREZ_GENE_ID"
 [3] "GB_ACC"                           "Gene Ontology Biological Process"
 [5] "Gene Ontology Cellular Component" "Gene Ontology Molecular Function"
 [7] "Gene Symbol"                      "Gene Title"
 [9] "ID"                               "RefSeq Transcript ID"
[11] "Representative Public ID"         "Sequence Source"
[13] "Sequence Type"                    "Species Scientific Name"
[15] "SPOT_ID"                          "Target Description"

[start] loop for the association of the gene symbols to the probeset ID's: completed
4.49% 8.98% 13.46% 17.95% 22.44% 26.93% 31.41% 35.90% 40.39% 44.88% 49.36% 53.85% 58.34%
 62.83% 67.32% 71.80% 76.29% 80.78% 85.27% 89.75% 94.24% 98.73%
 [end] loop for the association of the gene symbols to the probeset ID's
```

geneExpressionFromGEO

```
 > geneExpressionDF %>% head()
          GSM73386 GSM73387 GSM73388 GSM73389 GSM73390 GSM73391 GSM73392
1007_s_at    9.185   10.529    9.618   10.343    9.369   10.484    9.610
1053_at      6.282    6.803    6.389    6.717    6.402    6.616    6.395
117_at       6.508    6.514    6.480    6.427    6.587    6.557    6.658
121_at       8.945    8.898    9.004    9.017    9.145    9.017    9.032
1255_g_at    3.974    4.007    4.142    4.157    4.296    4.294    4.043
1294_at      8.272    7.508    8.542    7.994    8.342    7.674    8.213
          GSM73393 GSM73394 GSM73395        GeneSymbol
1007_s_at   10.934    9.372   11.332
1053_at      6.600    6.340    7.067              RFC2
117_at       6.486    6.799    6.436             HSPA6
121_at       8.791    8.719    8.725              PAX8
1255_g_at    4.068    4.043    4.082             GUCA1A
1294_at      7.267    8.649    7.667 MIR5193 /// UBA7
```

probesets of the platform

geneExpressionFromGEO

```
 > geneExpressionDF %>% head()
           GSM73386 GSM73387 GSM73388 GSM73389 GSM73390 GSM73391 GSM73392
1007_s_at     9.185   10.529    9.618   10.343    9.369   10.484    9.610
1053_at       6.282    6.803    6.389    6.717    6.402    6.616    6.395
117_at        6.508    6.514    6.480    6.427    6.587    6.557    6.658
121_at        8.945    8.898    9.004    9.017    9.145    9.017    9.032
1255_g_at     3.974    4.007    4.142    4.157    4.296    4.294    4.043
1294_at       8.272    7.508    8.542    7.994    8.342    7.674    8.213
           GSM73393 GSM73394 GSM73395        GeneSymbol
1007_s_at    10.934    9.372   11.332
1053_at       6.600    6.340    7.067              RFC2
117_at        6.486    6.799    6.436             HSPA6
121_at        8.791    8.719    8.725              PAX8
1255_g_at     4.068    4.043    4.082             GUCA1A
1294_at       7.267    8.649    7.667 MIR5193 /// UBA7
```

probesets of the platform

associated gene symbols from EntrezGene

geneExpressionFromGEO

- Future developments: add more platforms for which it is possible to associate probesets with gene symbols

Thank for your attention!

Questions and collaboration proposals: davidechicco@davidechicco.it

More information can be found on my website: www.DavideChicco.it

