Automated Validation of Homo sapiens Presence in GEO Datasets

"Mohammad Reza Mohajeri"

"2023-07-15"

Abstract

The "Automated Validation of Homo sapiens Presence in GEO Datasets" script aids researchers in efficiently identifying and validating GEO datasets with Homo sapiens samples. Key functionalities include:

1. Displaying the GEO accession number for dataset correlation.
2. Reporting total samples and Homo sapiens count.
3. Highlighting the relevant column for potential manual verification.
4. Confirming when all samples match the Homo sapiens criterion.

This tool streamlines dataset screening, ensuring researchers work with relevant and accurate data.

Example output:
"In the dataset GSE161420 , there are 60 samples in total, and 60 of them are Homo sapiens."
"Homo sapiens found in the following columns: organism ch1"
"All samples are human."

The differential expression analysis of multiple datasets offers a unique opportunity to leverage larger amounts of data for more robust conclusions. However, combining datasets presents challenges, particularly when ensuring their compatibility. To guide this process, we utilize the following approach:

1. Criteria Definition:

- Organism:

Identify the target organism of interest (e.g., Homo sapiens, Mus musculus).

- Sample Size:

Determine an acceptable range for the total number of samples and specific conditions.

- Sample Origin:

Specify the origin of the samples, be it from tissues (like blood or biopsies), cell lines, or isolated cells.

- Platform and Technology:

Establish the preferred technology or platform, such as Microarray, RNA-Seq, or Single-cell RNA-Seq.

2. Data Exploration:

- Using platforms like the GEO database, identify potential studies that might match the criteria. This exploration can be done through PubMed, GEO, or other relevant databases.

- Catalog the relevant studies, noting their GEO accession numbers, in an organized manner (e.g., an Excel spreadsheet).

3.Automated Screening:

- With the list of GEO accession numbers, utilize the script to automatically check each dataset's phenotypic data.

- The script scans for specific criteria, such as the organism, and provides a detailed summary for each dataset. For instance, if the criterion is "Homo sapiens", the script will output:

"In the dataset GSE161420 , there are 60 samples in total, and 60 of them are Homo sapiens."
"Homo sapiens found in the following columns: organism ch1"
"All samples are human."

This output serves multiple purposes:

- Dataset Identification:

Clearly identifies the dataset being analyzed.

- Sample Count:

Provides a total count of samples in the dataset and how many meet the criterion.

- Column Location:

Offers a direct pointer to the column where the criterion was found, facilitating manual verification if needed.

- Confirmation:

If all samples match the criterion, the script confirms it, making it easier to decide on dataset inclusion.

By following this approach, researchers can more efficiently sift through multiple datasets, ensuring that only the most relevant and compatible datasets are selected for combined differential expression analysis.

```
# Library imports
library(GEOquery)
```

```
# Dataset Identifier Initialization
dataset_id <- "GSE161420"
```

```r
# Data Retrieval
dataset = GEOquery::getGEO(GEO = dataset_id)
```

```
## Found 1 file(s)

## GSE161420_series_matrix.txt.gz
```

```r
# Specifying the Platform
dataset_idx <- 1 # Default set to 1

# Extracting Phenotype Data
if (length(dataset) >= dataset_idx) {
  pheno_data <- Biobase::phenoData(dataset[[dataset_idx]])@data
} else {
  stop("The specified index does not exist in the dataset.")
}

# Initialize a counter for human samples
human_samples <- 0

# Initialize a list to store column names where 'Homo sapiens' is found
columns_with_humans <- list()

# Loop through each column to search for 'Homo sapiens'
for(colname in names(pheno_data)) {
  if(is.factor(pheno_data[[colname]]) || is.character(pheno_data[[colname]])) {
    num_human_in_col <- sum(pheno_data[[colname]] == 'Homo sapiens', na.rm = TRUE)

    if(num_human_in_col > 0) {
      human_samples <- human_samples + num_human_in_col
      columns_with_humans <- c(columns_with_humans, list(colname))
    }
  }
}

# Find out the total number of samples
total_samples <- nrow(pheno_data)

# Create and print the final message
if(human_samples > 0) {
  final_message <- paste(
    "In the dataset", dataset_id,
    ", there are", total_samples, "samples in total, and",
    human_samples, "of them are Homo sapiens.")

  print(final_message)

  columns_message <- paste(
    "Homo sapiens found in the following columns:",
    paste(columns_with_humans, collapse=", "))

  print(columns_message)
```

```
  if(human_samples == total_samples) {
    print("All samples are human.")
  } else {
    print("Not all samples are human.")
  }
} else {
  print("No Homo sapiens samples were found.")
}
```

```
## [1] "In the dataset GSE161420 , there are 60 samples in total, and 60 of them are Homo sapiens."
## [1] "Homo sapiens found in the following columns: organism_ch1"
## [1] "All samples are human."
```