GEO RNASeq CountMatrix Extraction

Mohammad Reza Mohajeri

2023-10-04

==========

GSE135251

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135251 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE135251)

Supplementary file Size Download File type/resource

GSE135251_RAW.tar 43.7 Mb (http) TAR (of TXT)

“GSE135251_RAW.tar” file was downloaded from GEO from GSE135251, then the tarball was unpacked using “tar xf GSE135251_RAW.tar” in bash terminal

==========

Exploratory View of a Representative Sample File

```
Description for the explore-sample-file chunk:
  - Exploratory Inspection of a Sample File
  - Description: This chunk is an optional step to facilitate the exploratory inspection of individual sam
ple files. It provides an opportunity to understand the structure and content of the files before proceedi
ng to subsequent analyses.
```

```
# Load a gzipped RNA-Seq count data sample for inspection
p <- read.table(gzfile("/home/omidmoh1980/Downloads/GSM3998274_245-Ann-Daly_S13.counts.txt.gz"), sep="\t",
header = TRUE, row.names = 1)
```

Listing GSM Files

```
Description for the GSM-file-listing chunk:
    - This chunk identifies and lists all the GSM files (presumably from GEO) in the specified directory t
hat end with "txt.gz".
    -This is typically useful when one has downloaded raw data files from the GEO database and needs to pr
ocess or analyze them in R.
    - Make sure to unpack the tarball first (using "tar xf GSE135251_RAW.tar") before executing this chun
k.
```

```
# List all 'GSM' prefixed files in the specified directory
allGSMFiles <- list.files("/home/omidmoh1980/Downloads/", pattern="^GSM", full.names=TRUE)
allGSMFiles <- grep("txt.gz$", allGSMFiles, value=TRUE)
```

Loading and Processing Count Data

```
Description for the process-GSM-files chunk:
- This chunk reads in each GSM file and processes it to extract count data.
Specifically, it:
    1. Reads the delimited data from the file (assuming no header and using the first column as row name
s).
    2. Extracts a clean column name from the filename, removing any ".counts" and subsequent characters.
    3. Assigns the cleaned name as the column name for the count data.
    4. Returns the processed data.
      The end result is a list of matrices (countMatrix), where each matrix corresponds to a GSM file's co
unt data.
```

```
# Process each GSM file to extract count data:
countMatrix <- lapply(allGSMFiles, function(x){
    # Read the delimited data from the file
    r <- read.delim(x, header=FALSE, row.names=1, stringsAsFactors = FALSE)
    # Extract a clean column name from the filename
    colname <- gsub("\\.counts.*", "", basename(x))
    # Assign the modified name as the column name for the count data
    colnames(r) <- colname
    r
})
```

Combining Individual Count Matrices to Form a Comprehensive Count Matrix for Downstream Gene Expression Analyses (e.g. DESeq2, edgeR & limma)

```
  Description for the assemble-count-matrix chunk:
      - The 'do.call' function is employed here to bind together individual data frames from the 'countMat
rix' list.
      - Specifically:'do.call' takes a function (in this case, 'cbind') as its first argument.
      - The second argument to 'do.call' is our list of data frames, 'countMatrix'.
      - Within this context, 'do.call' effectively "spreads out" each data frame in 'countMatrix' and uses
'cbind' to bind them all together column-wise.
      - As a result, the single, combined data frame we get has each of its columns representing the data
extracted from a specific file in the 'allGSMFiles' list.
      - Each column is named according to the file from which its data was sourced.
```

```
# Combine multiple matrices/dataframes column-wise to create a single count matrix
  countMatrix <- do.call(cbind, countMatrix)
```

Exploring the Assembled Count Matrix

```
Description for the explore-count-matrix chunk:
    - This chunk provides an initial look at the assembled count matrix.
    - First, it displays the top rows of the matrix to give a glimpse of its structure.
    - Next, it outputs the matrix's dimensions to provide a sense of its size (number of genes vs number o
f samples).
```

```
# Display the first few rows of the combined count matrix
head(countMatrix)

# Display the dimensions of the count matrix
dim(countMatrix)
```