GSE_tutorial

Luke Torre-Healy

7/3/2019

Getting started

Run install.dependencies.R, then build the package before you do anything.

Look for Build in the top right tab, click Install and Restart.

Query Geo using SOFT files

Here is the most rigorous version of querying, I don't recommend using this approach. However, it might be nice to run once to see how a SOFT file is formatted and how long it takes. It's certainly less efficient than the second example. For now, I've set eval to FALSE to not waste time.

```
getGEOfile("GSE71729",
           destdir = ".")

gfile.s <- system.file("extdata", "GSE71729.soft.gz", package = "rnaGinesis")

gset.s <- getGEO(filename = gfile.s,
                 GSEMatrix =FALSE,
                 AnnotGPL=FALSE,
                 getGPL=FALSE)
```

The option below is much cleaner and quicker.

```
gset <- getGEO(GEO = "GSE71729",
                 GSEMatrix =TRUE,
                 AnnotGPL=FALSE,
                 getGPL=FALSE)
## Found 1 file(s)
## GSE71729_series_matrix.txt.gz
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_REF = col_character()
## )
## See spec(...) for full column specifications.
# ---------------------------------------------
class(gset)
## [1] "list"
#its currently a list, we want the expression set. The line below reduces it to what we want
gset <- gset$GSE71729_series_matrix.txt.gz
# ----------------
class(gset)
## [1] "ExpressionSet"
## attr(,"package")
## [1] "Biobase"
```

^ you can run getGEO with $GSE.... at the end once you know what it is. It's more compact for future people. I just don't know exactly which GSE you'll be using

experimentData has info about where/who did the work, assayData should have counts, phenoData should have sample info. sometimes these are poorly/incorrectly formatted, requires some searching. use GEOaccession (google) to find more specifics

Extract relevant data

```
# ex has the counts, use exprs()
ex <- exprs(gset)
str(ex)
```

```
##  num [1:19749, 1:357] 1.685 1.373 0.926 2.95 3.09 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:19749] "A1BG" "A1CF" "A2BP1" "A2LD1" ...
##   ..$ : chr [1:357] "GSM1843893" "GSM1843894" "GSM1843895" "GSM1843896" ...
```

```
# make a featInfo object that just has the rownames of ex, usually gene names or ensembl ids, etc
featInfo <- data.frame(SYMBOL = rownames(ex))
str(featInfo)
```

```
## 'data.frame':    19749 obs. of  1 variable:
##  $ SYMBOL: Factor w/ 19749 levels "A1BG","A1CF",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
# phenoData has more info about the samples
samps <-gset@phenoData
str(samps)
## Formal class 'AnnotatedDataFrame' [package "Biobase"] with 4 slots
##   ..@ varMetadata      :'data.frame':    49 obs. of  1 variable:
##   .. ..$ labelDescription: chr [1:49] NA NA NA NA ...
##   ..@ data             :'data.frame':    357 obs. of  49 variables:
##   .. ..$ title                                          : Factor w/ 357 levels "46336-Normal-Pancreas
##   .. ..$ geo_accession                                  : chr [1:357] "GSM1843893" "GSM1843894" "GSM1
##   .. ..$ status                                         : Factor w/ 1 level "Public on Sep 07 2015":
##   .. ..$ submission_date                                : Factor w/ 1 level "Aug 04 2015": 1 1 1 1 1
##   .. ..$ last_update_date                               : Factor w/ 1 level "Sep 07 2015": 1 1 1 1 1
##   .. ..$ type                                           : Factor w/ 1 level "RNA": 1 1 1 1 1 1 1 1 1
##   .. ..$ channel_count                                  : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1
##   .. ..$ source_name_ch1                                : Factor w/ 1 level "Human Reference": 1 1 1
##   .. ..$ organism_ch1                                   : Factor w/ 1 level "Homo sapiens": 1 1 1 1 1
##   .. ..$ characteristics_ch1                            : Factor w/ 1 level "sample type: Stratagene
##   .. ..$ molecule_ch1                                   : Factor w/ 1 level "total RNA": 1 1 1 1 1 1
##   .. ..$ extract_protocol_ch1                           : Factor w/ 1 level "Qiagen Rneasy Kit": 1 1
##   .. ..$ label_ch1                                      : Factor w/ 1 level "Cy3": 1 1 1 1 1 1 1 1 1
##   .. ..$ label_protocol_ch1                             : Factor w/ 1 level "As described in Agilent
##   .. ..$ taxid_ch1                                      : Factor w/ 1 level "9606": 1 1 1 1 1 1 1 1 1
##   .. ..$ source_name_ch2                                : Factor w/ 21 levels "AbWall_Metastasis",..:
##   .. ..$ organism_ch2                                   : Factor w/ 1 level "Homo sapiens": 1 1 1 1 1
##   .. ..$ characteristics_ch2                            : Factor w/ 29 levels "cell line/tissue: AbWa
##   .. ..$ characteristics_ch2.1                          : Factor w/ 5 levels "tissue type: Metastasis
##   .. ..$ characteristics_ch2.2                          : Factor w/ 48 levels "stroma_subtype_0na_1lo
##   .. ..$ characteristics_ch2.3                          : Factor w/ 7 levels "","death_event_1death_0
##   .. ..$ characteristics_ch2.4                          : Factor w/ 3 levels "","tumor_subtype_0na_1c
##   .. ..$ characteristics_ch2.5                          : Factor w/ 4 levels "","stroma_subtype_0na_1
##   .. ..$ molecule_ch2                                   : Factor w/ 1 level "total RNA": 1 1 1 1 1 1
##   .. ..$ extract_protocol_ch2                           : Factor w/ 1 level "Qiagen Rneasy Kit": 1 1
##   .. ..$ label_ch2                                      : Factor w/ 1 level "Cy5": 1 1 1 1 1 1 1 1 1
##   .. ..$ label_protocol_ch2                             : Factor w/ 1 level "As described in Agilent
##   .. ..$ taxid_ch2                                      : Factor w/ 1 level "9606": 1 1 1 1 1 1 1 1 1
##   .. ..$ hyb_protocol                                   : Factor w/ 1 level "As described in Agilent
##   .. ..$ scan_protocol                                  : Factor w/ 1 level "Fluorescent array images
##   .. ..$ description                                    : Factor w/ 357 levels "46336","46337",..: 23
##   .. ..$ data_processing                                : Factor w/ 1 level "log2 Cy5 signal was anal
##   .. ..$ platform_id                                    : Factor w/ 1 level "GPL20769": 1 1 1 1 1 1 1
##   .. ..$ contact_name                                   : Factor w/ 1 level "Richard,,Moffitt": 1 1 1
##   .. ..$ contact_institute                              : Factor w/ 1 level "University of North Caro
##   .. ..$ contact_address                                : Factor w/ 1 level "450 West Drive": 1 1 1 1
##   .. ..$ contact_city                                   : Factor w/ 1 level "Chapel Hill": 1 1 1 1 1
##   .. ..$ contact_state                                  : Factor w/ 1 level "NC": 1 1 1 1 1 1 1 1 1 1
##   .. ..$ contact_zip/postal_code                        : Factor w/ 1 level "27599": 1 1 1 1 1 1 1 1
##   .. ..$ contact_country                                : Factor w/ 1 level "USA": 1 1 1 1 1 1 1 1 1
##   .. ..$ supplementary_file                             : Factor w/ 357 levels "ftp://ftp.ncbi.nlm.ni
##   .. ..$ data_row_count                                 : Factor w/ 1 level "19749": 1 1 1 1 1 1 1 1
##   .. ..$ cell line/tissue:ch2                           : chr [1:357] "BXPC3" "Capan1" "Capan2" "CFPA
##   .. ..$ death_event_1death_0censor:ch2                 : chr [1:357] NA NA NA NA ...
##   .. ..$ sample type:ch1                                : chr [1:357] "Stratagene Human reference RNA
##   .. ..$ stroma_subtype_0na_1low_2normal_3activated:ch2: chr [1:357] "1" "1" "1" "1" ...
```

```
##   .. ..$ survival_months:ch2                              : chr [1:357] NA NA NA NA ...
##   .. ..$ tissue type:ch2                                  : chr [1:357] NA NA NA NA ...
##   .. ..$ tumor_subtype_0na_1classical_2basal:ch2          : chr [1:357] "2" "2" "2" "2" ...
##   ..@ dimLabels         : chr [1:2] "sampleNames" "sampleColumns"
##   ..@ .__classVersion__:Formal class 'Versions' [package "Biobase"] with 1 slot
##   .. .. ..@ .Data:List of 1
##   .. .. .. ..$ : int [1:3] 1 1 0
```

What to do with our extracted data

We like to make lists with four headings, **$ex, $sampInfo, $metaData, and $featInfo**. Keeping them the same allows us to use prewritten codes for parsing and analysis

```
Moffitt_data <- list()

Moffitt_data$ex <- ex
Moffitt_data$sampInfo <- samps
Moffitt_data$metadata <- list(log.transformed = F)
# This just indicates we have raw counts^
Moffitt_data$featInfo <- featInfo
```

Now if you save the above list, its the cleaned, extracted data from the online portal ready to be manipulated.