



EpiMed Open Course – Session 3

Manage public omics data from NCBI GEO

Ekaterina Flin

27/03/2020

Public repositories for omics data

Nowadays, many repositories with public omics data are available online.

Below are listed a few popular repositories:

- NCBI GEO: <https://www.ncbi.nlm.nih.gov/geo>
- Array Express: <https://www.ebi.ac.uk/arrayexpress>
- The Cancer Genome Atlas (TCGA): <https://portal.gdc.cancer.gov>
- The Clinical Proteomic Tumor Analysis Consortium (CPTAC): <https://cptac-data-portal.georgetown.edu/cptacPublic/>
- Genotype-Tissue Expression (GTEx): <https://www.gtexportal.org/home/>

How to access, download, manage and process the data?

Let's consider an example:

- **NCBI GEO:** <https://www.ncbi.nlm.nih.gov/geo>

We are interested in a particular dataset:

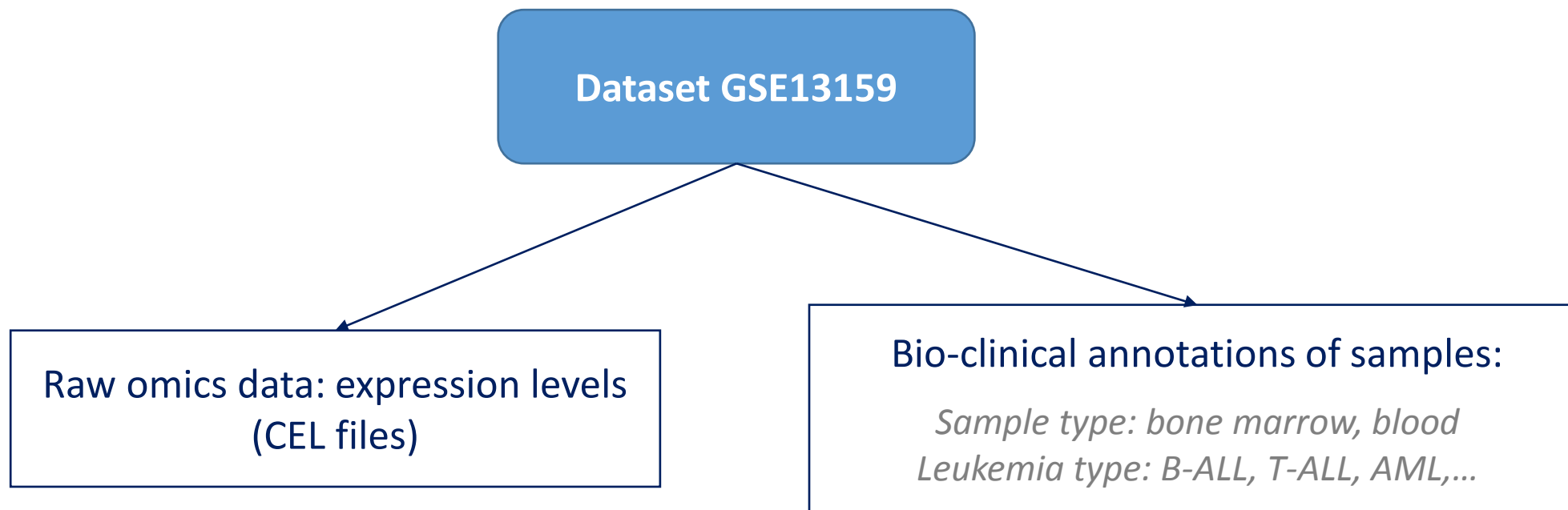
GSE13159 - Microarray Innovations in LEukemia (MILE) study

This dataset contains **2 096** samples of **transcriptomic** data obtained with **microarray technology**

Example: GSE13159 dataset from NCBI GEO

Usually, a dataset contains raw omics data and corresponding bio-clinical annotations. Sometimes, processed data are also available in addition to raw data (or instead of raw data).

We need to manage both raw omics data and bio-clinical annotations.

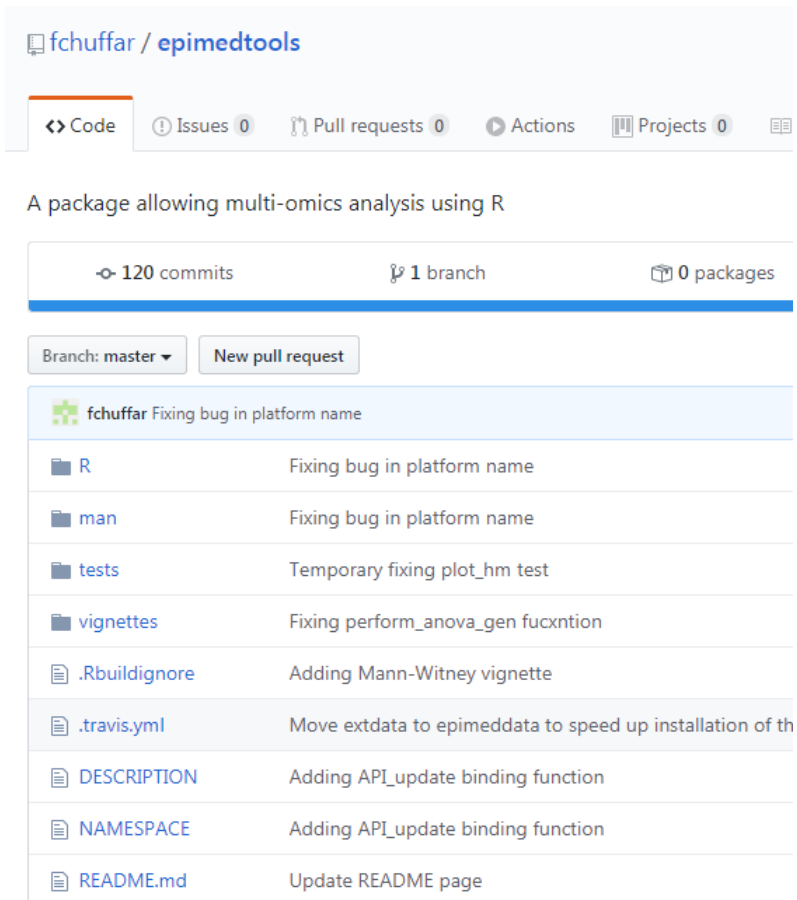


<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13159>

Download and process omics data with R package “epimedtools”

epimedtools on Github: <https://github.com/fchuffar/epimedtools>

Package allowing multi-omics analysis using R, developed by **Florent Chuffart**



fchuffar / epimedtools

Code Issues 0 Pull requests 0 Actions Projects 0

A package allowing multi-omics analysis using R

120 commits 1 branch 0 packages

Branch: master New pull request

fchuffar Fixing bug in platform name

R	Fixing bug in platform name
man	Fixing bug in platform name
tests	Temporary fixing plot_hm test
vignettes	Fixing perform_anova_gen fucxntion
.Rbuildignore	Adding Mann-Witney vignette
.travis.yml	Move extdata to epimeddata to speed up installation of th
DESCRIPTION	Adding API_update binding function
NAMESPACE	Adding API_update binding function
README.md	Update README page

Installation

To get the current development version from github, you need first to install following packages from bioconductor:

- Biobase
- affy
- GEOquery

```
# install.packages("BiocManager")
BiocManager::install(c("Biobase", "affy", "GEOquery"))
```

Then, install `epimedtool` :

```
# install.packages("devtools")
devtools::install_github("fchuffar/epimedtools")
```

Vignettes

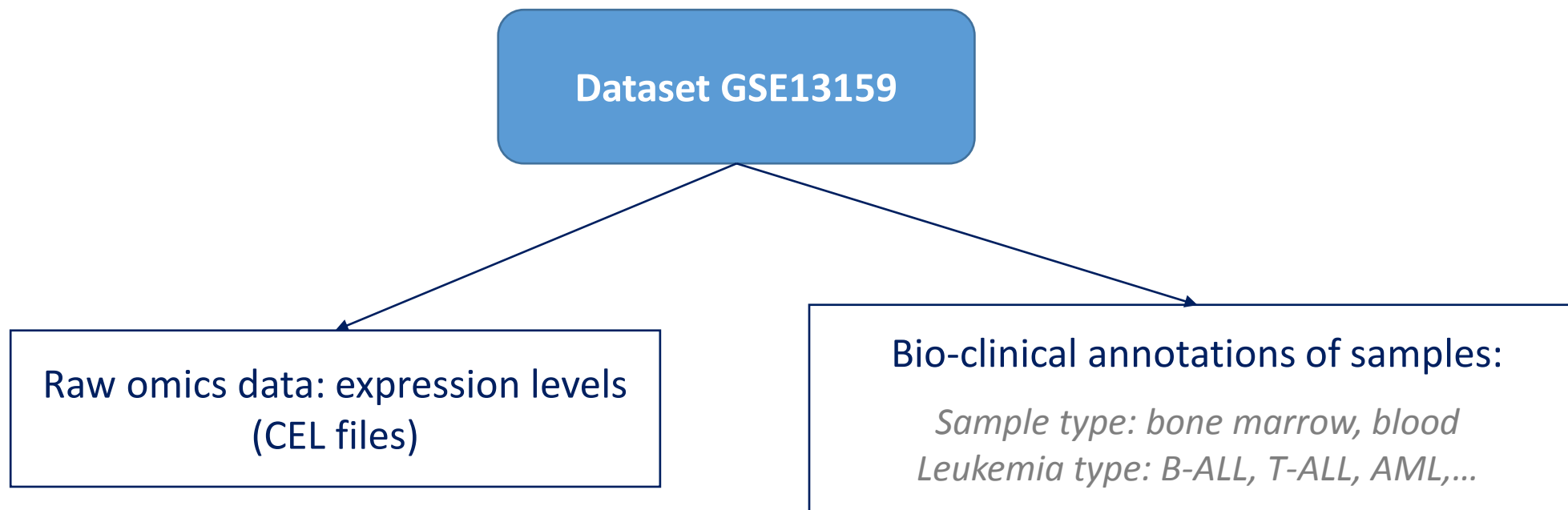
To browse available vignettes:

```
browseVignettes(package = 'epimedtools')
```

Example: GSE13159 dataset from NCBI GEO

Usually, a dataset contains raw omics data and corresponding bio-clinical annotations. Sometimes, processed data are also available in addition to raw data (or instead of raw data).

We need to manage both raw omics data and bio-clinical annotations.



<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13159>

The clinical data are presented in NCBI GEO in a semi-structured or a plain text format. Unstructured text format is **difficult to use directly for analysis**. We usually **need to convert the text into a structured format**.

Sample GSM329407

Query DataSets for GSM329407

Status	Public on Sep 30, 2009
Title	MILES stage 1 data N1_0001
Sample type	RNA

Source name	Leukemia patient sample
-------------	-------------------------

Organism	Homo sapiens
----------	------------------------------

Characteristics	sample type: bone marrow leukemia class: mature B-ALL with t(8;14)
-----------------	---

Treatment protocol	Samples are from untreated patients.
--------------------	--------------------------------------

Growth protocol	not applicable
-----------------	----------------

Extracted molecule	total RNA
--------------------	-----------

GSE13159: Bio-clinical annotations

The clinical data are presented in NCBI GEO in a semi-structured or a plain text format. Unstructured text format is **difficult to use directly for analysis**. We usually **need to convert the text into a structured format**.

Sample GSM329407

Query DataSets for GSM329407

Status	Public on Sep 30, 2009
Title	MILES stage 1 data N1_0001
Sample type	RNA

Source name	Leukemia patient sample
-------------	-------------------------

Organism	Homo sapiens
----------	------------------------------

Characteristics	sample type: bone marrow leukemia class: mature B-ALL with t(8;14)
-----------------	---

Treatment protocol	Samples are from untreated patients.
--------------------	--------------------------------------

Growth protocol	not applicable
-----------------	----------------

Extracted molecule	total RNA
--------------------	-----------



Structured JSON format

```
{
  "tissue": "bone marrow",
  "icdo": "C42.1",
  "pathology": "cancer",
  "subtype": "mature B-ALL with t(8;14)"
}
```


GSE13159: Bio-clinical annotations

The clinical data are presented in NCBI GEO in a semi-structured or a plain text format. Unstructured text format is **difficult to use directly for analysis**. We usually **need to convert the text into a structured format**.

Sample GSM329407

Query DataSets for GSM329407

Status Public on Sep 30, 2009
Title MILES stage 1 data N1_0001
Sample type RNA

Source name Leukemia patient sample

Organism [Homo sapiens](#)

Characteristics sample type: bone marrow
leukemia class: mature B-ALL with t(8;14)

Treatment protocol Samples are from untreated patients.

Growth protocol not applicable

Extracted molecule total RNA

Structured JSON format

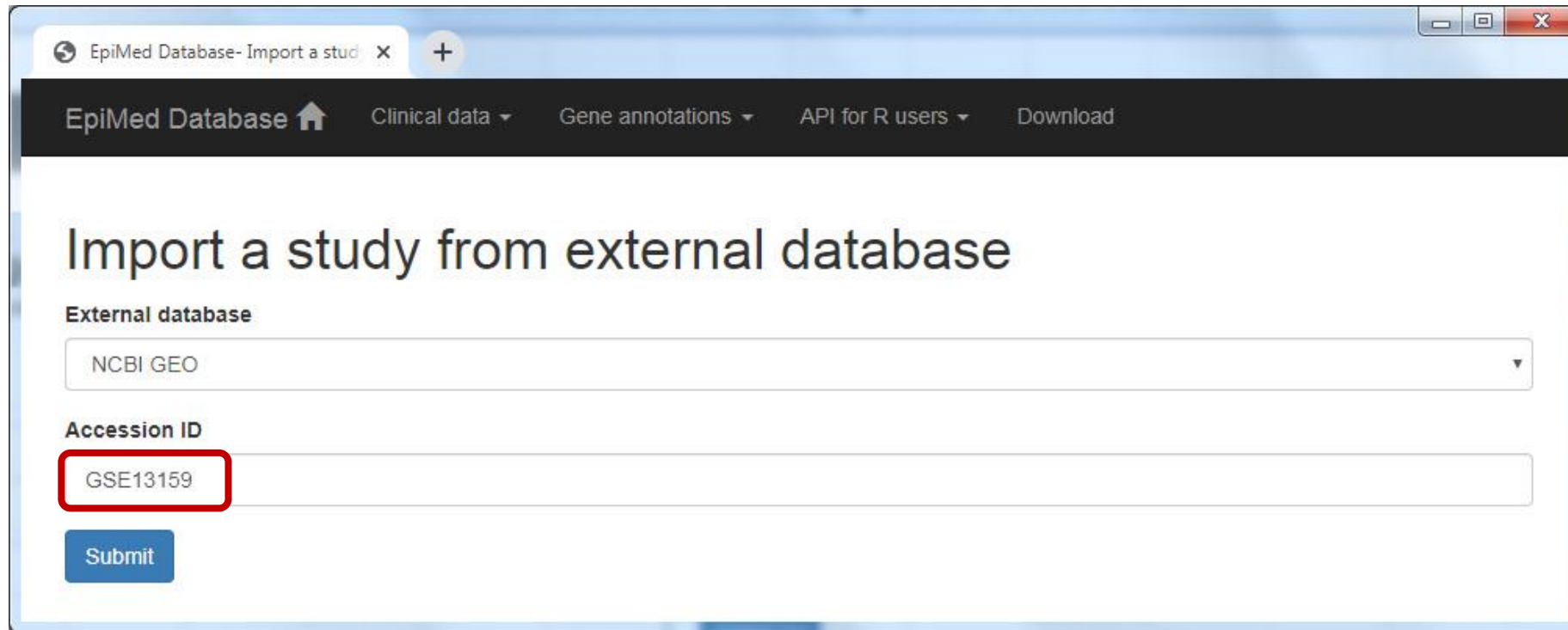
```
{  
  "tissue": "bone marrow",  
  "icdo": "C42.1",  
  "pathology": "cancer",  
  "subtype": "mature B-ALL with t(8;14)"  
}
```

Structured Excel-like format

tissue	icdo	pathology	subtype
bone marrow	C42.1	cancer	mature B-ALL with t(8;14)

Download and process clinical annotations with EpiMed Database

EpiMed Database can help you to organize clinical annotations from NCBI GEO in a structured format (Excel, CSV).



The screenshot shows a web browser window with the EpiMed Database interface. The page title is "Import a study from external database". The "External database" dropdown menu is set to "NCBI GEO". The "Accession ID" text input field contains "GSE13159" and is highlighted with a red rectangular border. A blue "Submit" button is located below the input fields. The browser's address bar shows "EpiMed Database- Import a stud" and the page has a dark navigation bar with links for "Clinical data", "Gene annotations", "API for R users", and "Download".

Demonstration: <http://epimed.univ-grenoble-alpes.fr/database/series>

Download and process clinical annotations with EpiMed Database

EpiMed Database can help you to organize clinical annotations from NCBI GEO in a structured format (Excel, CSV).

The image shows two browser windows from the EpiMed Database. The left window, titled 'Import a study from external database', has a navigation bar with 'EpiMed Database', 'Clinical data', 'Gene annotations', 'API for R users', and 'Download'. Below the navigation bar, the page title is 'Import a study from external database'. Under the heading 'External database', there is a text input field containing 'NCBI GEO'. Under the heading 'Accession ID', there is a text input field containing 'GSE13159', which is highlighted with a red rectangle. A blue 'Submit' button is located below the input fields. A red arrow points from the 'Submit' button to the right window.

The right window, titled 'Download experimental grouping', has the same navigation bar. The page title is 'Download experimental grouping'. There is a green button labeled 'Import a study from external database' and a search bar containing 'GSE13159' with a search icon. Below the search bar, it says 'Found 1 studies. Please select one or several studies and click on download button.' There is a checked checkbox next to the text 'GSE13159 (2096) - Microarray Innovations in LEukemia (MILE) study: Stage 1 data'. Below this, there is a blue button labeled 'Download Excel'.

Demonstration: <http://epimed.univ-grenoble-alpes.fr/database/series>

Link to **R package “epimedtools”** (developed by Florent Chuffart)

<https://github.com/fchuffar/epimedtools>

Link to **EpiMed Database** (developed by Ekaterina Flin)

<http://epimed.univ-grenoble-alpes.fr/database/>

Questions? Please, send an e-mail to

epimed-open-course@univ-grenoble-alpes.fr