# *Bioinfo_Project User Manual*

## 1) creazione_manifest.py

This script is responsible for processing the json and manifest files related to mRNA and miRNA, in order to produce the final manifest files that are used to download the data from the server.

The following global variables are used: pay attention to the names of the files used in the script and present in the folder.

```
mrnaJson = '2_mRNA.json'
mrnaManifest = '2_mRNA_manifest.txt'
mirnaJson = '1_miRNA.json'
mirnaManifest = '1_miRNA_manifest.txt'
```

Output:

```
nuovo_manifest_m.txt
nuovo_manifest_mi.txt
new_label_mi.txt
new_label_m.txt
```

## 2) Data download using the client gdc_client.exe

With the following commands (run as administrator) the files contained in **new_manifest_m.txt** and **new_manifest_mi.txt** are downloaded.

```
gdc-client.exe download -m
<absolute_path_to_project_folder>\n1_unione_maninfest\nuovo_manifest_m.txt --dir
.\mRNA\
```

```
gdc-client.exe download -m
<absolute_path_to_project_folder>\n1_unione_maninfest\nuovo_manifest_mi.txt --dir
.\miRNA\
```

N.B.: Check the confirmation message of "Successfully downloaded: 1160".

## 3) Dataset creation

- Move the files ”new_label_mi.txt” and ”new_label_m.txt” (which contain the labels) created in step(1) to the folder ./n2_creazione_dataset.
- Run scripts ”move_files_miRNA.py” and ”move_files_mRNA.py”.
- Run scripts ”create_dataset_mi.py” and ”create_dataset_m.py”.

Output:

```
dataset_miRNA.csv
dataset_mRNA.csv
```

## 4) Dataset preprocessing

Run the script dataset_Preprocessing.py twice, changing from time to time the parameter (the dataset filename) given to the script.
Input commands:

- dataset_Preprocessing.py dataset_miRNA.csv
- dataset_Preprocessing.py dataset_mRNA.csv

Output:

```
scaled_dataset_miRNA.csv
scaled_dataset_mRNA.csv
```

## 5) Feature Selection Methods

Run the script related to the desired FS method and answer the questions asked on the command line: the answers given by the user perform the method with different parameters.

```
a) Complete or reduced list of labels? (R = reduced/C = complete)
b) Which dataset do you want to process?(mi= miRNA/m = mRNA)
c) Do you want to save the dataset containing only the most relevant features? (Y/N)
d) How many features do you want to select?
```

a) Reduced = implements the FS on the dataset containing a smaller number of labels without considering all subcategories of tumors.
   Complete = implements the FS on the dataset containing all types of labels.
b) miRNA = use the microRNA dataset.
   mRNA = use the mRNA dataset.
c) Y = saves a dataset that contains only the most relevant features obtained from the FS method used.
   N = executes the FS method without saving the dataset.
d) Number of features to select (1-1000).

Output:

```
<FS>_dataset.csv
Result_FS.txt
```