**GenomeCruzer\_preprocessing\_instructions**

The starting point is the gathering of RNA and CNA starting data from [www.cbioportal.org/datasets](http://www.cbioportal.org/datasets). The databases are named and referenced as follows:

A screenshot of a computer

Description automatically generated

Once chosen the dataset of interest, we should look for the data\_mrna\_seq and data\_cna documents, as well as for the data\_clinical\_sample.txt file. It should be noted that the formats used in the following process are the same as the ones in the cBioportal platform.

The next steps for the creation of the database are listed below:

* Definition of the samples’ dataset
* Definition of the genic dataset
* Definition of the molecular data
* Database generation
* Sample cluster annotation
* Gene cluster annotation

**1. Definition of the samples’ dataset**

From the data\_clinical\_sample.txt file, the following columns are retained and the samples list is created:

#Patient Identifier Sample Identifier

#Identifier to uniquely specify a patient. A unique sample identifier.

#STRING STRING

#1 1

PATIENT\_ID SAMPLE\_ID

**2. Definition of the genic dataset**

The genic database is structured as a tab delimited csv file. It should contain a numeric UNIQUE\_ID and an associated GENE\_ID, together with the chromosome location of the genes and their start position, end position, ARM\_ID, BAND\_ID and SUB BAND ID. The formatting of the table is shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **UNIQUE\_ID** | **GENE\_ID** | **chromosome\_name** | **start\_position** | **end\_position** | **ARM\_ID** | **BAND\_ID** | **SUB BAND ID** |

A default genic database of 19603 genes is used.

**3. Definition of the molecular data**

**3.1 RNA matrix**

Duplicates removal is performed and the numerical entrez id should be the first column of the file. In the same folder, the meta-data for this dataset is created as follows:

cancer\_study\_identifier: Name\_of\_the\_study

genetic\_alteration\_type: MRNA\_EXPRESSION

datatype: CONTINUOUS

stable\_id: rna\_seq\_mrna\_capture

show\_profile\_in\_analysis\_tab: true

profile\_name: mRNA expression (ILMN-Linear)

profile\_description: Expression levels (Log2, RNAseq)

data\_filename: data\_mrna.tsv

**3.2 CNA matrix**

The matrix is generated starting from segmentation data using the *code.R* script. This code takes as an input the genic dataset and the segmentation data and after extracting unique genes and samples, CNA values are collected for the segments, even in the case of overlapping ones.

The meta-data is created for this dataset as well, as follows:

cancer\_study\_identifier: Name\_of\_the\_study

genetic\_alteration\_type: COPY\_NUMBER\_ALTERATION

datatype: LOG2-VALUE

stable\_id: log2CNA

show\_profile\_in\_analysis\_tab: true

profile\_name: Log2 copy-number values.

profile\_description: Copy-number values for each gene (from shallow seq).

data\_filename: cna\_matrix.tsv

**3.3 Methylation**

**4. Database Generation**

The database is generated using a json file, containing the instructions and the paths to the files needed. A template is available for the creation of this file.

Through the use of a prompt, the database is generated. Once in the folder where the database should be created, the following instructions are given to *cBioImporter.exe* program:

...\Genome Cruzer\1.9.0\bin>cBioImporter.exe -j nuovo\_db\code\database\_cBioImport.json --out nuovo\_db\Outputs\database.db

**5. Sample cluster annotation**

After the database has been created, we could provide a sample clustering, given as column clustering onto *DbManager.exe* program. This clustering takes the form of a TAB delimited text file, containing all the samples in the database (samples not present in the database should not be listed). An example for the heading of ths file is shown below:

COLUMNS 2

UNIQUE\_ID CLUS\_ID\_1 CLUS\_ID\_2

**6. Gene cluster annotation**

Gene clustering is provided as row clustering onto *DbManager.exe* program. Some pre-defined files could be found in the clusterings folder, as an example.