**EE 542 – Laboratory Assignment**

Instructor: Young H. Cho

T.A.: Yue Shi

*Due date:*

*Report due: October 22, 11:59pm*

*Demo due: Oct 24 ,11:59 pm*

**The git repo for all the codes provided in this lab:** [**https://github.com/yuesOctober/GDCproject/tree/yue**](https://github.com/yuesOctober/GDCproject/tree/yue)

*Download the repo:*

*git clone* [*https://github.com/yuesOctober/GDCproject.git*](https://github.com/yuesOctober/GDCproject.git)

*Basic git command:* *https://confluence.atlassian.com/bitbucketserver/basic-git-commands-776639767.html*

**GDC Data Lab:**

**In this lab, you will learn:**

1. **How to download, integrate, and preprocess files related to a particular disease type, and how to use the data obtained.**
2. **As an example, you will go through the entire process to get the miRNA files, and the related file metadata, case metadata to the disease Liver Hepatocellular Carcinoma**
3. **You will apply the machine learning package to the miRNA matrix extracted to detect normal/ cancer samples.**

**Steps you need to do with this lab:**

1. **Go through the entire tutorial and do the Part 1 and Part 2 with miRNA expression data for all cancer types and do a multiclass classification. You can treat all the samples from normal tissues from different cancer types as one single class.**
2. **In Part2, try a different model other than the one provided in the sample code.**
3. **Visualize the features after feature selection with t-SNE and PCA method [3]**
4. **Plot the evaluation metrics and ROC curve for the model.**

***Submission guideline:***

***Each team should create a github repo and provide the link to your repo for in your slide. You need have a readme file explaining how to run your source codes. For video demo submission, you need show the steps to run your code and explain. Only one submission per team is needed.***

***Notes:***

***In this lab, the miRNA expression quantification data is used, however in your final project, you could use different types or combination of different types of biomarkers, e.g. somatic mutation, copy number variation , DNA methylation etc for cancer type idenfication.***

***Part 1: Data download, integration and preprocess.***

**1. Introduction to GDC data:**

**Read the document below to get a sense of GDC data.**

<https://gdc.cancer.gov/about-data>

Biomarker Data:

|  |  |
| --- | --- |
| Data Category | Data Type |
| DNA Methylation | [Methylation Beta Value](https://portal.gdc.cancer.gov/repository?facetTab=files&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.primary_site%22%2C%22value%22%3A%5B%22Liver%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.access%22%2C%22value%22%3A%5B%22open%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.data_category%22%2C%22value%22%3A%5B%22DNA%20Methylation%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.data_type%22%2C%22value%22%3A%5B%22Methylation%20Beta%20Value%22%5D%7D%7D%5D%7D&searchTableTab=cases) |
| Simple Nucleotide Variation | Annotated Somatic Mutation |
| Raw Simple Somatic Mutation |
| Aggregated Somatic Mutation |
| Masked Somatic Mutation |
| Transcriptome Profile | Gene Expression Quantification |
| Isoform Expression Quanfitication |
| miRNA expression Quantification |

1. **Example: Downloading miRNA files of Disease: Liver Hepatocellular Carcinoma**

miRNA Expression Quantification is a table that associates miRNA IDs with read count and a normalized count in reads-per-million-miRNA-mapped. <https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/miRNA_Pipeline/>

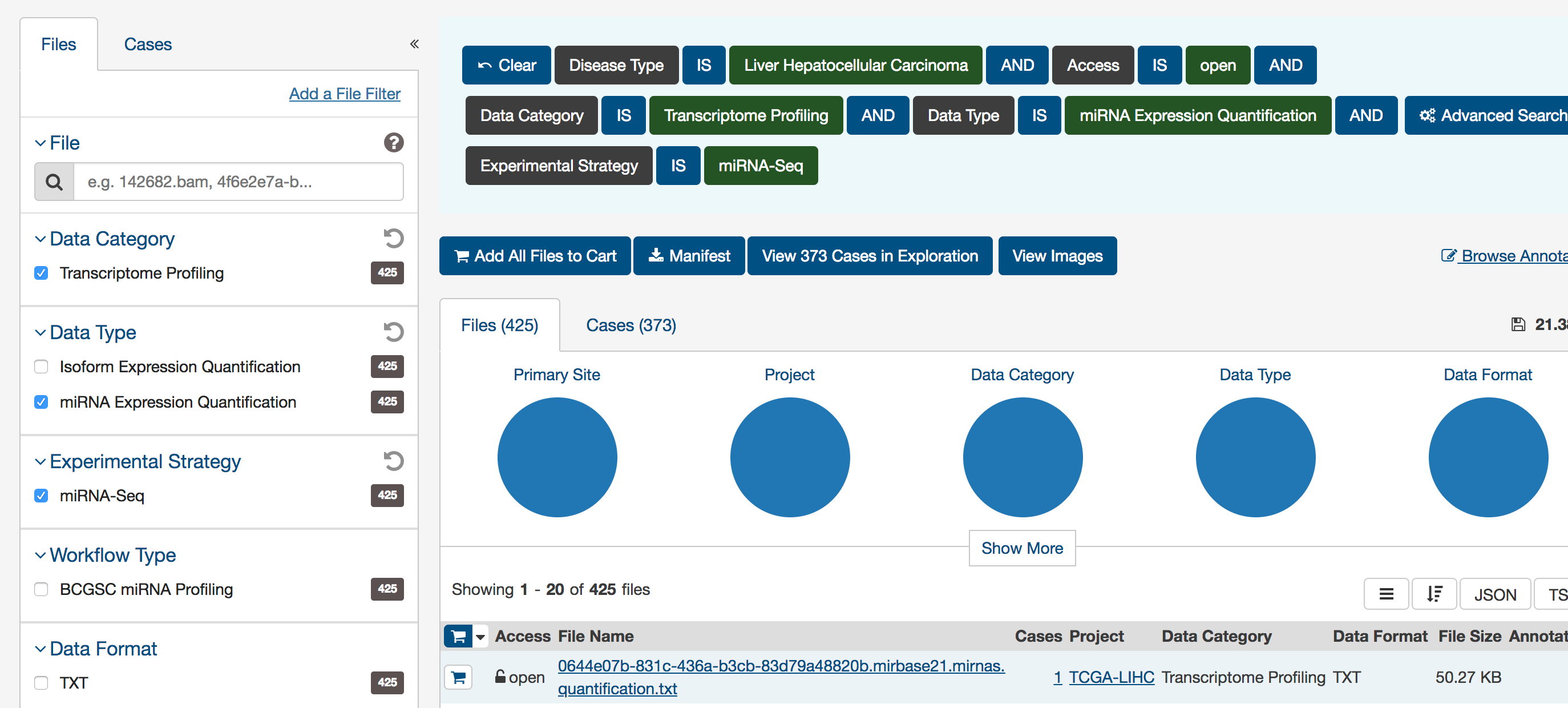
Download **Expression Quantification data:** miRNA sequence data

1. Go to the data portal *https://portal.gdc.cancer.gov/repository*, on the left side there are two tabs : ***Files*** and ***Cases***
2. Click ***Cases*** and select a disease type: [**Liver Hepatocellular Carcinoma**](https://portal.gdc.cancer.gov/repository?facetTab=cases)
3. Click **Files** and select

*Data Category*: **Transcriptome Profiling**

*Data type* : **miRNA Expression Quantification**

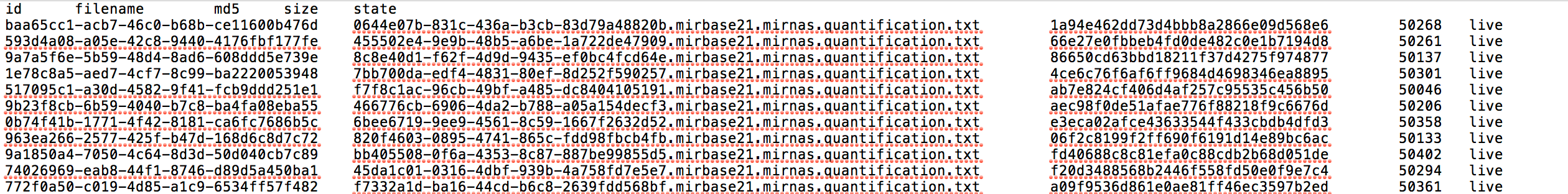
*Experimental Strategy:* **miRNA-Seq**

****

You will see 373 cases and 425 files. That means there are duplicates for some cases. Also in those cases, there are some normal cases without cancer.

1. Click on the **Manifest download.** This will download the manifest file for use with GDC data transfer tool.

The Manifest file contains the id, filename, md5, size and patient state.



1. Data transfer tool Download:

<https://gdc.cancer.gov/access-data/gdc-data-transfer-tool>

Download the version according to your OS type.

Command line to **download** and **unzip** a **OSX** version:

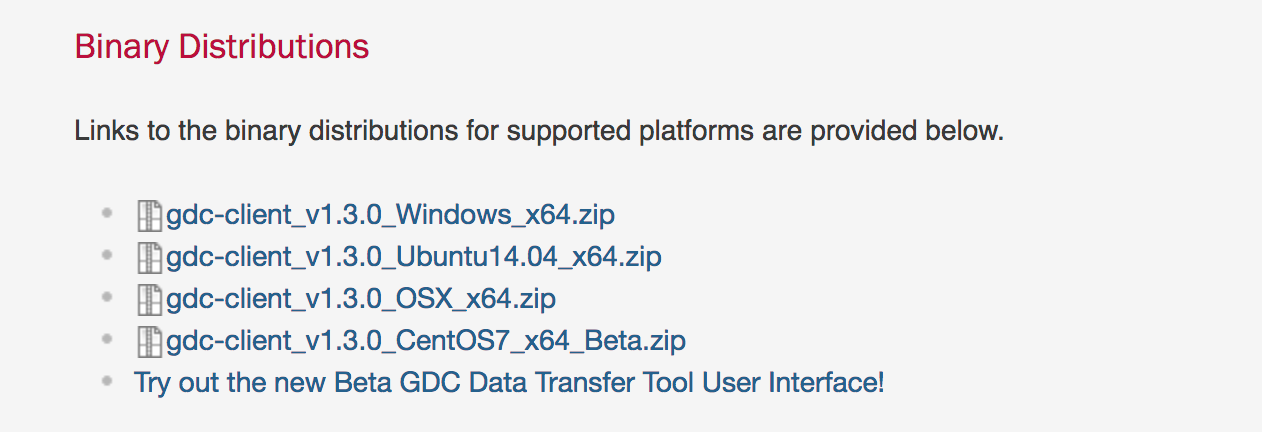
***Download:***

***wget -c -t 0*** [***https://gdc.cancer.gov/files/public/file/gdc-client\_v1.3.0\_OSX\_x64.zip***](https://gdc.cancer.gov/files/public/file/gdc-client_v1.3.0_OSX_x64.zip)

***Unzip:***

***Unzip gdc-client\_v1.3.0\_OSX\_x64.zip***

***Note: For other versions, just replace with the corresponding OS version file name.***



### [System Recommendations](https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Getting_Started/#system-recommendations)

**The system recommendations for using the GDC Data Transfer Tool are as follows:**

* **OS: Linux (Ubuntu 14.x or later), OS X (10.9 Mavericks or later), or Windows (7 or later)**
* **CPU: At least eight 64-bit cores, Intel or AMD**
* **RAM: At least 8 GiB**
* **Storage: Enterprise-class storage system capable of at least 1 Gb/s (gigabit per second) write throughput and sufficient free space for BAM files.**

**Please use a AWS machine instance if your own laptop does not meet the requirement.**

1. Download the files with gdc-client tool :
2. make a directory for the data:

mkdir live\_miRNA

cd live\_miRNA

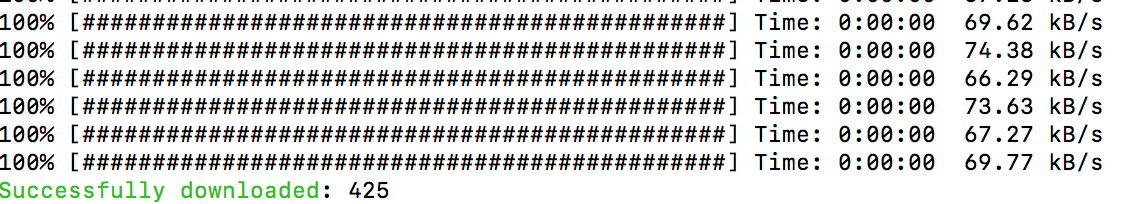
1. Download with gdc-client.

./<path-to-gdc-client>/gdc-client download –m <path-to-manifest-file>

e.g.

***./~/Downloads/gdc-client –m ~/Downloads/gdc\_manifest.2018-08-23.txt***

After successful downloads, you will see



1. Check the successful download:

Since large volumes of data are downloaded, it is important to check the file integrity. You could use the md5 checksum to check the integrity of downloaded files.

Run the code: python3 check.py

A sample python 3 code **check.py** is provided.

1. If some files fail download, use the following command:

***./<path-to-gdc-client>/gdc-client download <id>***

e.g.

../gdc-client download fa63ce14-b9b5-4041-9df7-3b86ba9ede16

1. Once we get the biomarker files. We also need get the case ids related to the files .

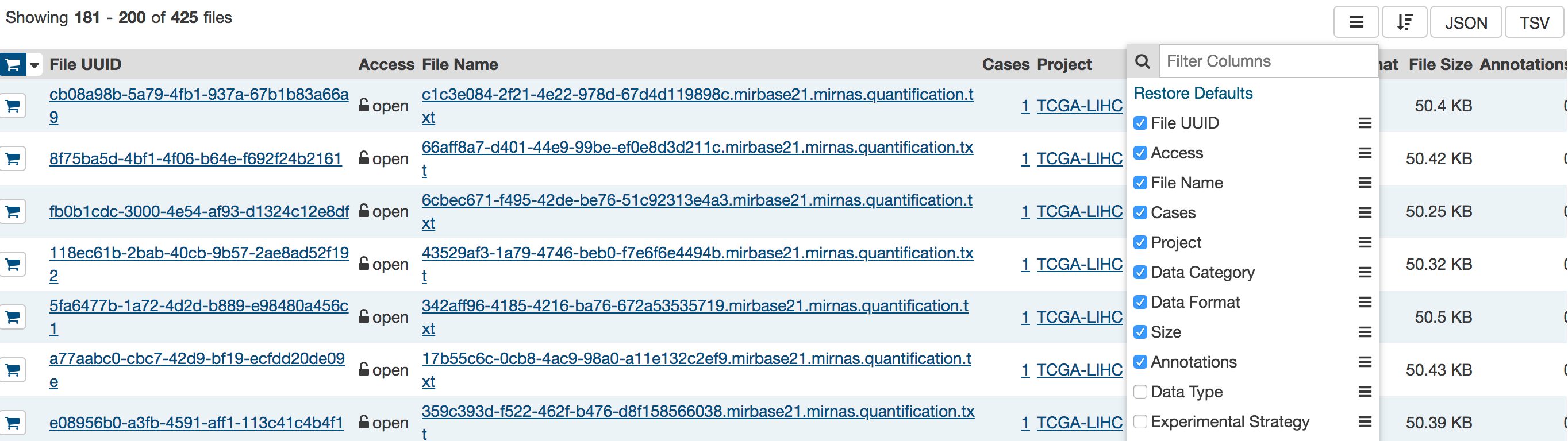
This is because we need correlate the biomarker files with the corresponding case clinical/ biospecimen files.

Here we need to write some python codes to extract all the file\_ids and the corresponding case\_ids for future use.

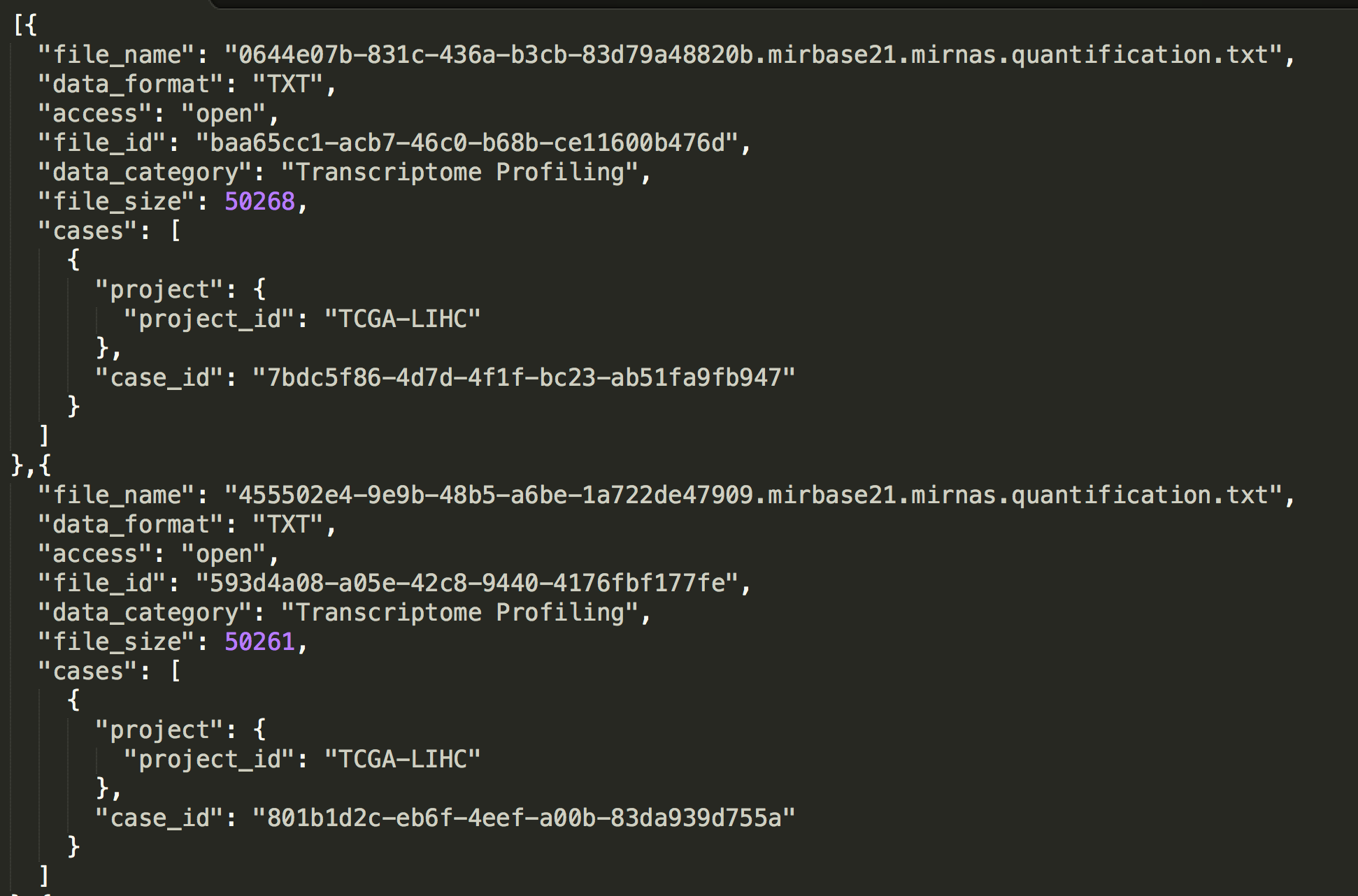
Get the cases related to the files:

The code ***parse\_file\_case\_id.py***is provided.

Click on the tab , and check all the following items, then click on the **JSON** tab. It will download the case ids for the files.



Screenshot of a downloaded file:



11． Get the meta data for the files and corresponding cases:

The source code: ***request\_meta.py***

The fields for the files and cases:

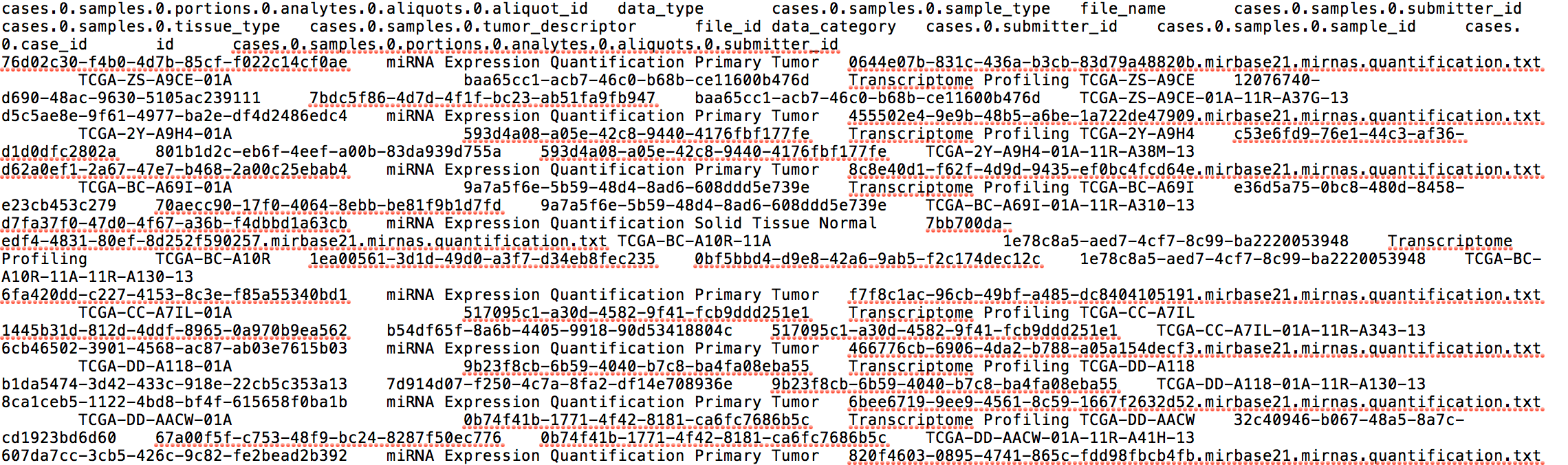
File fields:

<https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/#file-fields>

case fields:

<https://docs.gdc.cancer.gov/API/Users_Guide/Appendix_A_Available_Fields/#case-fields>

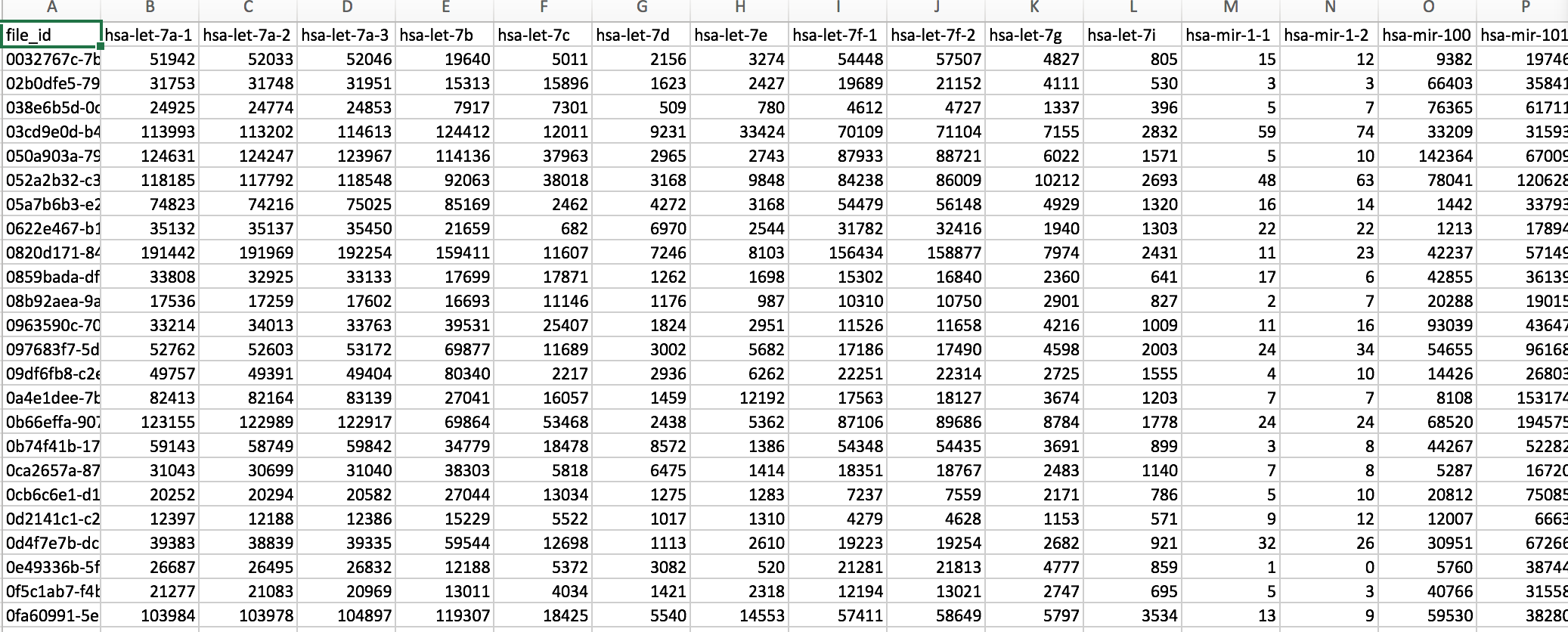
Once we get the meta data for the miRNA files, we can see that some samples come from a normal solid tissue and some others come from tumor.

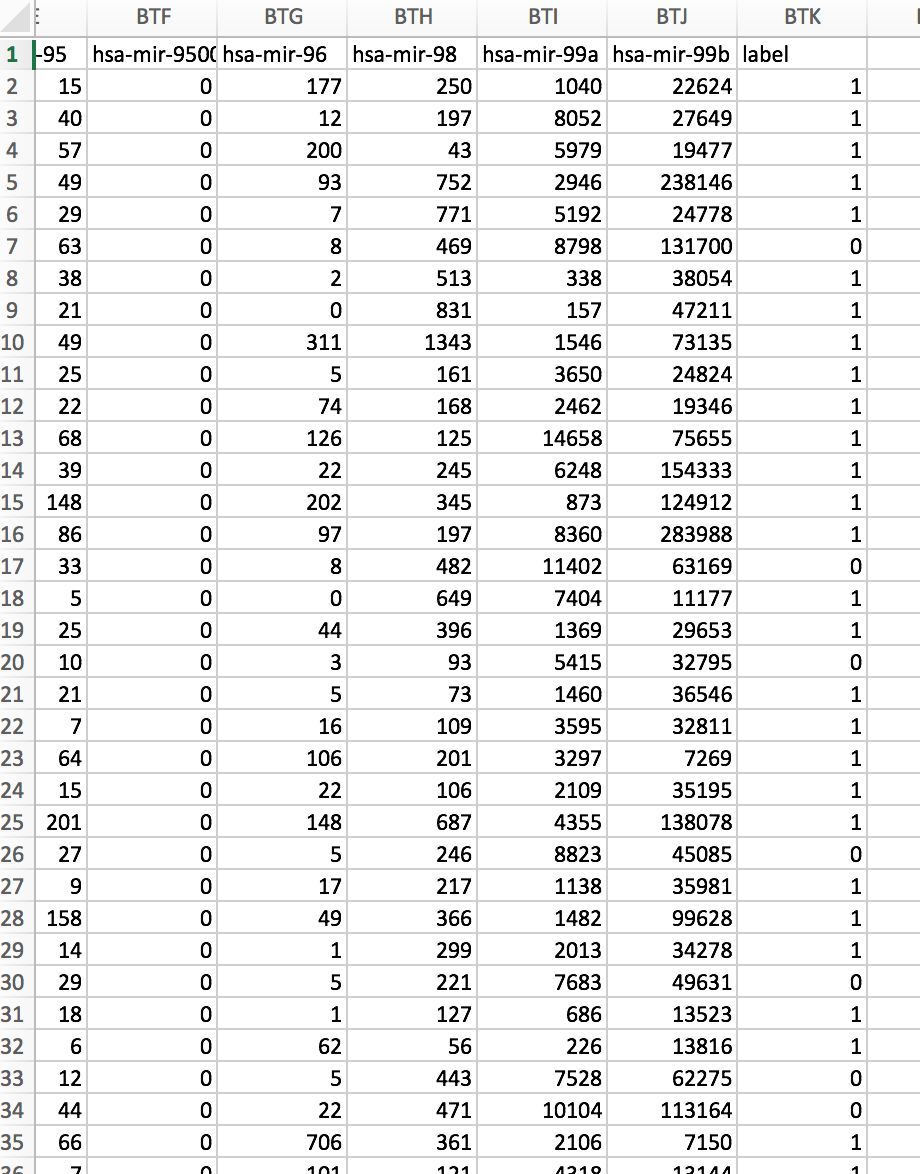


12. Now we could generate the miRNA matrix for all the files with labeled normal or tumor.

The miRNA seq that comes from tumor is labeled with 1, and normal tissue is labeled with 0.

The source code: ***gen\_miRNA\_matrix.py***





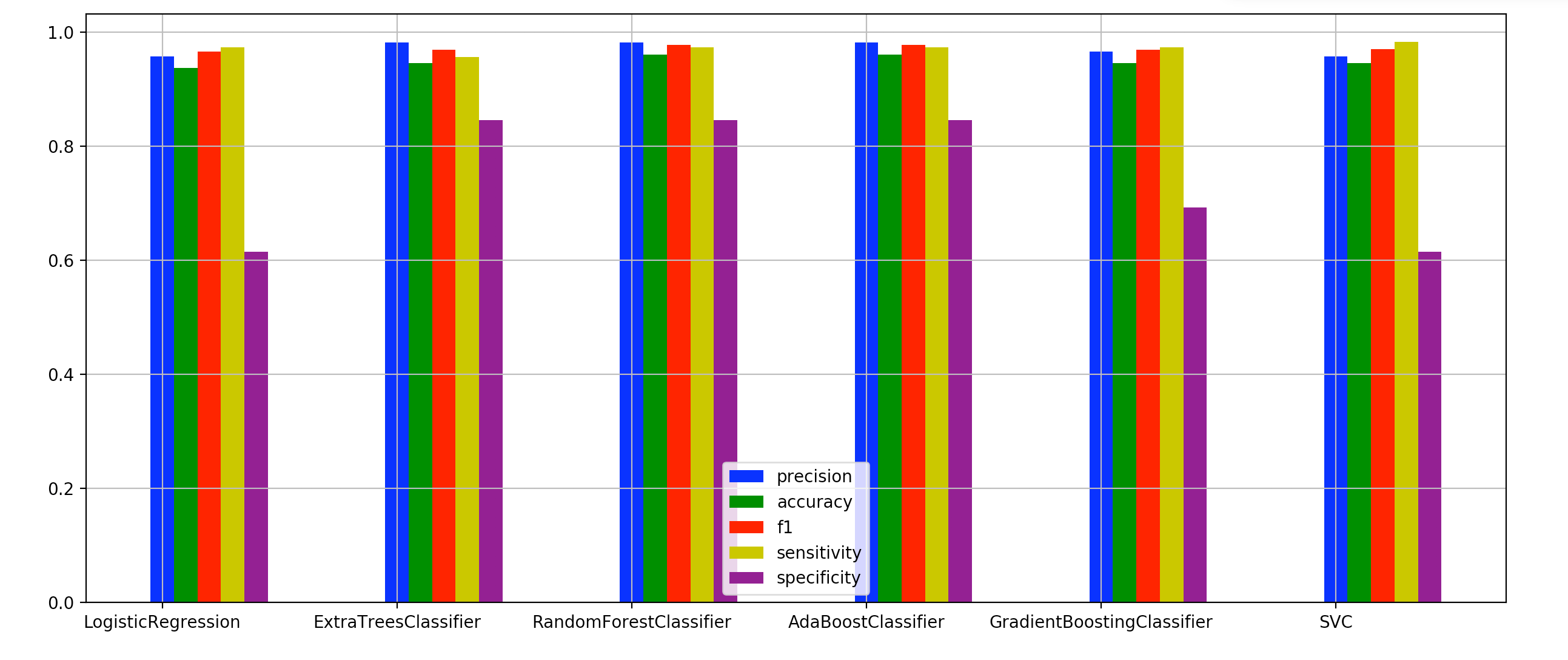
***Part 2: Apply Machine Learning Package (sklearn) to the above data.***

Sample code provided: ***predict.py***

*The steps are as following:*

1. Data standardization.
2. Train and test data split.
3. Feature selection.
4. Model hyper-parameters tuning with cross validation
5. Model prediction with the best hyper-parameters
6. Evaluation: Precision, Sensitivity, Accuracy, F1-score, Specificity

The result is shown in Figure 1.



***Fig.1 Performance Evaluation for Different ML models***

Below are some good reference papers for your project.

**Reference:**

[1] Hyeongmin Kim & Yong-Min Kim ,“Pan-cancer analysis of somatic mutations and transcriptomes reveals common functional gene clusters shared by multiple cancer types,” *Scientific Reports,***volume 8**, Article number: 6041 (2018) ,https://www.nature.com/articles/s41598-018-24379-y

[2] Marieke Lydia Kuijjer, Joseph Nathaniel Paulson, Peter Salzman, Wei Ding & John Quackenbush, “Cancer subtype identification using somatic mutation data,”, British Journal of Cancervolume 118, pages1492–1501 (2018).

[3] https://medium.com/@luckylwk/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b