

EE 542 – Laboratory Assignment
Instructor: Young H. Cho
T.A.: Yue Shi
Due date: October 10 at 11:59pm

The git repo for all the codes provided in this lab: <https://github.com/yuesOctober/GDCproject/tree/yue>

Download the repo:

git clone <https://github.com/yuesOctober/GDCproject.git>

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.

What to turn in:

Go through the entire tutorial and do the Part 1 and Part 2 with the disease type: **Lung Squamous Cell Carcinoma**. In Part2, try a different model other than the one provided in the sample code and plot the ROC curve for the models.

Extra Credit: Explore the Gene Expression Quantification Data.

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
Simple Nucleotide Variation	Annotated Somatic Mutation
	Raw Simple Somatic Mutation
	Aggregated Somatic Mutation
	Masked Somatic Mutation
Transcriptome Profile	Gene Expression Quantification

	Isoform Expression Quantification
	miRNA expression Quantification

2. 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**
3. Click **Files** and select

Data Category: Transcriptome Profiling

Data type : miRNA Expression Quantification

Experimental Strategy: miRNA-Seq

The screenshot shows the GDC Data Portal interface. On the left, there are filters for 'Files' and 'Cases'. The 'Files' filter is active, showing a search bar with 'e.g. 142682.bam, 4f6e2e7a-b-...' and a list of filters: 'Data Category' (Transcriptome Profiling), 'Data Type' (miRNA Expression Quantification), 'Experimental Strategy' (miRNA-Seq), 'Workflow Type' (BCGSC miRNA Profiling), and 'Data Format' (TXT). On the right, there are buttons for 'Clear', 'Disease Type' (Liver Hepatocellular Carcinoma), 'IS', 'Access', 'IS', 'open', 'AND', 'Data Category' (Transcriptome Profiling), 'IS', 'Data Type' (miRNA Expression Quantification), 'AND', 'Experimental Strategy' (miRNA-Seq), 'Advanced Search', 'Add All Files to Cart', 'Manifest', 'View 373 Cases in Exploration', 'View Images', and 'Browse Annotations'. Below these, there are five circular charts labeled 'Primary Site', 'Project', 'Data Category', 'Data Type', and 'Data Format'. A 'Show More' button is next to the 'Data Format' chart. At the bottom, there is a table with columns: 'Access File Name', 'Cases Project', 'Data Category', 'Data Format', 'File Size', and 'Annotat'. The table shows one file: '0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt' with 1 case, TCGA-UHQ, Transcriptome Profiling, TXT, and 50.27 KB.

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.

4. 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.

id	filename	md5	size	state				
baa65cc1-acb7-46c0-b68b-ce11600b476d	0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt	0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt	50268	live	1a94e462dd73d4bbb8a2866e89d568e6	50261	live	
59334a08-a05e-42c8-9440-4176fbf177fe	455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt	455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt	50137	live	66e27e0fbbef4f0de482c0e1b7194d8	50137	live	
9a7a5f6e-5b59-48d4-8a06-608ddd5e739e	8c8e48d1-f62f-4d9d-9435-ef0bc4fcd646.mirbase21.mirnas.quantification.txt	8c8e48d1-f62f-4d9d-9435-ef0bc4fcd646.mirbase21.mirnas.quantification.txt	50046	live	86650cd63b0d18211f37d4275f974877	50046	live	
1e78c9a5-aed7-4c17-8c99-ba2220053948	7bb700da-edf4-4831-90ef-8d252f59025f.mirbase21.mirnas.quantification.txt	7bb700da-edf4-4831-90ef-8d252f59025f.mirbase21.mirnas.quantification.txt	50206	live	4ce6c76f6a16f19084d4698346ea8895	50206	live	
517895c1-a30d-4582-9f41-fcb9ddd251e1	f7f8c1ac-96cb-49bf-a485-dc8404105191.mirbase21.mirnas.quantification.txt	f7f8c1ac-96cb-49bf-a485-dc8404105191.mirbase21.mirnas.quantification.txt	50133	live	ab7e824cf406d4a7257c95535c456b50	50133	live	
9b23f8cb-6b59-4040-b7c8-ba4fa08eba55	466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt	466776cb-6906-4da2-b788-a05a154decf3.mirbase21.mirnas.quantification.txt	50402	live	aec98f0de51a1ae776f88218f9c6676d	50402	live	
0b74f41b-1771-4f42-8181-ca6fc7686b5c	6bee6719-9ee9-4561-8c59-1667f2632d52.mirbase21.mirnas.quantification.txt	6bee6719-9ee9-4561-8c59-1667f2632d52.mirbase21.mirnas.quantification.txt	50294	live	e3eca02afce43633544f433c0db4d7d3	50294	live	
963ea266-2577-4251-b47d-168d6c807c72	820f4803-0895-4741-865c-fdd98fbc4f7b.mirbase21.mirnas.quantification.txt	820f4803-0895-4741-865c-fdd98fbc4f7b.mirbase21.mirnas.quantification.txt	50361	live	08f2c8199f2f1690f6191d14e89bcbac	50361	live	
9a1850a4-7050-4c64-8d3d-50d040cb7c89	bb405500-016a-4353-8c87-987be99855d5.mirbase21.mirnas.quantification.txt	bb405500-016a-4353-8c87-987be99855d5.mirbase21.mirnas.quantification.txt			f4d0688c8c31ef9a0c98c0db2b68d51de			
74026969-eab8-44f1-8746-d89d5a458ba1	45da1c01-0316-4dbf-939b-4e75df75e7.mirbase21.mirnas.quantification.txt	45da1c01-0316-4dbf-939b-4e75df75e7.mirbase21.mirnas.quantification.txt			f2d03488568b2446f558f05e08f9e7c5			
772f0a50-c019-4085-a1c9-6534ff57f482	f7332a1d-ba16-44cd-b6c8-2639fdd568bf.mirbase21.mirnas.quantification.txt	f7332a1d-ba16-44cd-b6c8-2639fdd568bf.mirbase21.mirnas.quantification.txt			a09f95360861e0ae81ff46ec3597b2ed			

5. 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

Unzip:

Unzip gdc-client_v1.3.0_OSX_x64.zip

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

Binary Distributions

Links to the binary distributions for supported platforms are provided below.

- [gdc-client_v1.3.0_Windows_x64.zip](#)
- [gdc-client_v1.3.0_Ubuntu14.04_x64.zip](#)
- [gdc-client_v1.3.0_OSX_x64.zip](#)
- [gdc-client_v1.3.0_CentOS7_x64_Beta.zip](#)
- Try out the new Beta GDC Data Transfer Tool User Interface!

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.

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

```
mkdir live_miRNA
cd live_miRNA
```
 - b. 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

```
100% [#####] Time: 0:00:00 69.62 kB/s
100% [#####] Time: 0:00:00 74.38 kB/s
100% [#####] Time: 0:00:00 66.29 kB/s
100% [#####] Time: 0:00:00 73.63 kB/s
100% [#####] Time: 0:00:00 67.27 kB/s
100% [#####] Time: 0:00:00 69.77 kB/s
Successfully downloaded: 425
```

7. 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.

8. 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
```

9. 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.

Showing 181 - 200 of 425 files

| File UUID                            | Access | File Name                                                                | Cases | Project   | Filter Columns                                    | File Size | Annotation |
|--------------------------------------|--------|--------------------------------------------------------------------------|-------|-----------|---------------------------------------------------|-----------|------------|
| cb08a98b-5a79-4fb1-937a-67b1b83a66a9 | open   | c1c3e084-2f21-4e22-978d-67d4d119898c.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | Restore Defaults                                  | 50.4 KB   |            |
| 8f75ba5d-4bf1-4f06-b64e-f692f24b2161 | open   | 66aff8a7-d401-44e9-99be-ef0e8d3d211c.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> File UUID     | 50.42 KB  |            |
| fb0b1cdc-3000-4e54-af93-d1324c12e8df | open   | 6cbec671-f495-42de-be76-51c92313e4a3.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> Access        | 50.25 KB  |            |
| 118ec81b-2bab-40cb-9b57-2ae8ad52f192 | open   | 43529af3-1a79-4746-beb0-f7e6f6e4494b.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> File Name     | 50.32 KB  |            |
| 5fa6477b-1a72-4d2d-b889-e98480a456c1 | open   | 342aff96-4185-4216-ba76-672a53535719.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> Cases         | 50.5 KB   |            |
| a77aabc0-cbc7-42d9-bf19-ecfdd20de09e | open   | 17b55c6c-0cb8-4ac9-98a0-a11e132c2ef9.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> Project       | 50.43 KB  |            |
| e08956b0-a3fb-4591-aff1-113c41c4b4f1 | open   | 359c393d-f522-462f-b476-d8f158566038.mirbase21.mirnas.quantification.txt | 1     | TCGA-LIHC | <input checked="" type="checkbox"/> Data Category | 50.39 KB  |            |
|                                      |        |                                                                          |       |           | <input checked="" type="checkbox"/> Data Format   |           |            |
|                                      |        |                                                                          |       |           | <input checked="" type="checkbox"/> Size          |           |            |
|                                      |        |                                                                          |       |           | <input checked="" type="checkbox"/> Annotations   |           |            |
|                                      |        |                                                                          |       |           | <input type="checkbox"/> Data Type                |           |            |
|                                      |        |                                                                          |       |           | <input type="checkbox"/> Experimental Strategy    |           |            |

Screenshot of a downloaded file:

```
{
 "file_name": "0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt",
 "data_format": "TXT",
 "access": "open",
 "file_id": "baa65cc1-acb7-46c0-b68b-ce11600b476d",
 "data_category": "Transcriptome Profiling",
 "file_size": 50268,
 "cases": [
 {
 "project": {
 "project_id": "TCGA-LIHC"
 },
 "case_id": "7bdc5f86-4d7d-4f1f-bc23-ab51fa9fb947"
 }
],
 "file_name": "455502e4-9e9b-48b5-a6be-1a722de47909.mirbase21.mirnas.quantification.txt",
 "data_format": "TXT",
 "access": "open",
 "file_id": "593d4a08-a05e-42c8-9440-4176fbf177fe",
 "data_category": "Transcriptome Profiling",
 "file_size": 50261,
 "cases": [
 {
 "project": {
 "project_id": "TCGA-LIHC"
 },
 "case_id": "801b1d2c-eb6f-4eef-a00b-83da939d755a"
 }
]
}
```

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](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](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.

| cases.0.samples.0.tissue_type        | cases.0.samples.0.tumor_descriptor                  | file_id                                                      | data_category               | cases.0.samples.0.sample_type | file_name                                                                | cases.0.samples.0.submitter_id |
|--------------------------------------|-----------------------------------------------------|--------------------------------------------------------------|-----------------------------|-------------------------------|--------------------------------------------------------------------------|--------------------------------|
| 0.case_id                            | id                                                  | cases.0.samples.0.ports.0.analytes.0.aliquots.0.submitter_id | cases.0.samples.0.sample_id | cases.0.submitter_id          | cases.0.samples.0.sample_id                                              | cases.0.samples.0.sample_id    |
| 76d02c30-f4b0-4d7b-85cf-f022c14cf0ae | miRNA Expression Quantification Primary Tumor       | 0644e07b-831c-436a-b3cb-83d79a48820b                         | Transcriptome Profiling     | TCGA-ZS-A9CE-01A-11R-A376-13  | 0644e07b-831c-436a-b3cb-83d79a48820b.mirbase21.mirnas.quantification.txt | TCGA-ZS-A9CE-01A-11R-A376-13   |
| d690-48ac-9630-5105acc239111         | miRNA Expression Quantification Primary Tumor       | baa65cc1-acb7-46c0-b68b-ce11600b476d                         | Transcriptome Profiling     | TCGA-ZS-A9CE-01A-11R-A376-13  | baa65cc1-acb7-46c0-b68b-ce11600b476d                                     | TCGA-ZS-A9CE-01A-11R-A376-13   |
| d5c5ae8e-9f61-4977-ba2e-df4d2486edc4 | miRNA Expression Quantification Primary Tumor       | 593d4a08-a05e-42c8-9440-4176fbf177fe                         | Transcriptome Profiling     | TCGA-2Y-A9H4-01A-11R-A30M-13  | 593d4a08-a05e-42c8-9440-4176fbf177fe                                     | TCGA-2Y-A9H4-01A-11R-A30M-13   |
| d1d0dfc2802a                         | miRNA Expression Quantification Primary Tumor       | 801b1d2c-eb6f-4eef-a00b-83da939d755a                         | Transcriptome Profiling     | TCGA-2Y-A9H4-01A-11R-A30M-13  | 801b1d2c-eb6f-4eef-a00b-83da939d755a                                     | TCGA-2Y-A9H4-01A-11R-A30M-13   |
| d62a0ef1-2a67-47e7-b468-2a00c25ebab4 | miRNA Expression Quantification Primary Tumor       | 9a7a5f6e-5b59-48d4-8ad6-608dd5e739e                          | Transcriptome Profiling     | TCGA-BC-A69I-01A-11R-A310-13  | 9a7a5f6e-5b59-48d4-8ad6-608dd5e739e                                      | TCGA-BC-A69I-01A-11R-A310-13   |
| e23cb453c279                         | miRNA Expression Quantification Solid Tissue Normal | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                         | Transcriptome Profiling     | TCGA-BC-A69I-01A-11R-A310-13  | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                                     | TCGA-BC-A69I-01A-11R-A310-13   |
| d7fa37f0-47d0-4f67-a36b-f4dbbd1a63cb | miRNA Expression Quantification Solid Tissue Normal | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                         | Transcriptome Profiling     | TCGA-BC-A69I-01A-11R-A310-13  | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                                     | TCGA-BC-A69I-01A-11R-A310-13   |
| edf4-4831-80ef-8d252f590257          | miRNA Expression Quantification Primary Tumor       | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                         | Transcriptome Profiling     | TCGA-BC-A69I-01A-11R-A310-13  | 70a6cc90-17f0-4064-8ebb-be81f9b1d7fd                                     | TCGA-BC-A69I-01A-11R-A310-13   |
| Profiling                            | TCGA-BC-A10R                                        | 1ea00561-3d1d-49d0-a3f7-d34eb8fec235                         | Transcriptome Profiling     | TCGA-BC-A10R                  | 1ea00561-3d1d-49d0-a3f7-d34eb8fec235                                     | TCGA-BC-A10R                   |
| A10R-11A-11R-A130-13                 | TCGA-BC-A10R                                        | 1ea00561-3d1d-49d0-a3f7-d34eb8fec235                         | Transcriptome Profiling     | TCGA-BC-A10R                  | 1ea00561-3d1d-49d0-a3f7-d34eb8fec235                                     | TCGA-BC-A10R                   |
| 6fa420dd-c227-4153-8c3e-f85a55340bd1 | miRNA Expression Quantification Primary Tumor       | 517095c1-a30d-4582-9f41-fcb9dd251e1                          | Transcriptome Profiling     | TCGA-CC-A71L-01A-11R-A343-13  | 517095c1-a30d-4582-9f41-fcb9dd251e1                                      | TCGA-CC-A71L-01A-11R-A343-13   |
| TCGA-CC-A71L-01A                     | TCGA-CC-A71L-01A                                    | 517095c1-a30d-4582-9f41-fcb9dd251e1                          | Transcriptome Profiling     | TCGA-CC-A71L-01A-11R-A343-13  | 517095c1-a30d-4582-9f41-fcb9dd251e1                                      | TCGA-CC-A71L-01A-11R-A343-13   |
| 1445b31d-812d-4dd0-8965-0a970b9ea562 | miRNA Expression Quantification Primary Tumor       | 466776cb-6906-4da2-b788-a05a154decf3                         | Transcriptome Profiling     | TCGA-DD-A118-01A-11R-A130-13  | 466776cb-6906-4da2-b788-a05a154decf3                                     | TCGA-DD-A118-01A-11R-A130-13   |
| 6cb46502-3901-4568-ac07-ab03e7615b03 | miRNA Expression Quantification Primary Tumor       | 9b23f8cb-6b59-4040-b7c8-ba4fa08eba55                         | Transcriptome Profiling     | TCGA-DD-A118-01A-11R-A130-13  | 9b23f8cb-6b59-4040-b7c8-ba4fa08eba55                                     | TCGA-DD-A118-01A-11R-A130-13   |
| TCGA-DD-A118-01A                     | TCGA-DD-A118-01A                                    | 9b23f8cb-6b59-4040-b7c8-ba4fa08eba55                         | Transcriptome Profiling     | TCGA-DD-A118-01A-11R-A130-13  | 9b23f8cb-6b59-4040-b7c8-ba4fa08eba55                                     | TCGA-DD-A118-01A-11R-A130-13   |
| b1da5474-3d42-433c-918e-22cb5c353a13 | miRNA Expression Quantification Primary Tumor       | 7d914d07-f250-4c7a-8fa2-df14e708936e                         | Transcriptome Profiling     | TCGA-DD-A118-01A-11R-A130-13  | 7d914d07-f250-4c7a-8fa2-df14e708936e                                     | TCGA-DD-A118-01A-11R-A130-13   |
| 8ca1ceb5-1122-4bd8-bf4f-615658f0ba1b | miRNA Expression Quantification Primary Tumor       | 6bee6719-9ee9-4561-8c59-1667f2632d52                         | Transcriptome Profiling     | TCGA-DD-A118-01A-11R-A130-13  | 6bee6719-9ee9-4561-8c59-1667f2632d52                                     | TCGA-DD-A118-01A-11R-A130-13   |
| TCGA-DD-AACW-01A                     | TCGA-DD-AACW-01A                                    | 6bee6719-9ee9-4561-8c59-1667f2632d52                         | Transcriptome Profiling     | TCGA-DD-AACW-01A-11R-A41H-13  | 6bee6719-9ee9-4561-8c59-1667f2632d52                                     | TCGA-DD-AACW-01A-11R-A41H-13   |
| cd1923bd6d60                         | miRNA Expression Quantification Primary Tumor       | 0b74f41b-1771-4f42-8181-ca6fc7686b5c                         | Transcriptome Profiling     | TCGA-DD-AACW-01A-11R-A41H-13  | 0b74f41b-1771-4f42-8181-ca6fc7686b5c                                     | TCGA-DD-AACW-01A-11R-A41H-13   |
| 67a00f5f-c753-48f9-bc24-8287f50ec776 | miRNA Expression Quantification Primary Tumor       | 820f4603-0895-4741-865c-fdd98fbcb4fb                         | Transcriptome Profiling     | TCGA-DD-AACW-01A-11R-A41H-13  | 820f4603-0895-4741-865c-fdd98fbcb4fb                                     | TCGA-DD-AACW-01A-11R-A41H-13   |
| 607da7cc-3cb5-426c-9c82-fe2bead2b392 | miRNA Expression Quantification Primary Tumor       | 820f4603-0895-4741-865c-fdd98fbcb4fb                         | Transcriptome Profiling     | TCGA-DD-AACW-01A-11R-A41H-13  | 820f4603-0895-4741-865c-fdd98fbcb4fb                                     | TCGA-DD-AACW-01A-11R-A41H-13   |

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*

| A            | B            | C            | D            | E          | F          | G          | H          | I            | J            | K          | L          | M           | N           | O           | P           |
|--------------|--------------|--------------|--------------|------------|------------|------------|------------|--------------|--------------|------------|------------|-------------|-------------|-------------|-------------|
| file_id      | hsa-let-7a-1 | hsa-let-7a-2 | hsa-let-7a-3 | hsa-let-7b | hsa-let-7c | hsa-let-7d | hsa-let-7e | hsa-let-7f-1 | hsa-let-7f-2 | hsa-let-7g | hsa-let-7i | hsa-mir-1-1 | hsa-mir-1-2 | hsa-mir-100 | hsa-mir-101 |
| 0032767c-7b  | 51942        | 52033        | 52046        | 19640      | 5011       | 2156       | 3274       | 54448        | 57507        | 4827       | 805        | 15          | 12          | 9382        | 19746       |
| 02b0dfe5-79  | 31753        | 31748        | 31951        | 15313      | 15896      | 1623       | 2427       | 19689        | 21152        | 4111       | 530        | 3           | 3           | 66403       | 35841       |
| 038e6b5d-0c  | 24925        | 24774        | 24853        | 7917       | 7301       | 509        | 780        | 4612         | 4727         | 1337       | 396        | 5           | 7           | 76365       | 61711       |
| 03cd9e0d-b4  | 113993       | 113202       | 114613       | 124412     | 12011      | 9231       | 33424      | 70109        | 71104        | 7155       | 2832       | 59          | 74          | 33209       | 31593       |
| 050a903a-75  | 124631       | 124247       | 123967       | 114136     | 37963      | 2965       | 2743       | 87933        | 88721        | 6022       | 1571       | 5           | 10          | 142364      | 67005       |
| 052a2b32-c3  | 118185       | 117792       | 118548       | 92063      | 38018      | 3168       | 9848       | 84238        | 86009        | 10212      | 2693       | 48          | 63          | 78041       | 120628      |
| 05a7b6b3-e2  | 74823        | 74216        | 75025        | 85169      | 2462       | 4272       | 3168       | 54479        | 56148        | 4929       | 1320       | 16          | 14          | 1442        | 33793       |
| 0622e467-b1  | 35132        | 35137        | 35450        | 21659      | 682        | 6970       | 2544       | 31782        | 32416        | 1940       | 1303       | 22          | 22          | 1213        | 17894       |
| 0820d171-8c  | 191442       | 191969       | 192254       | 159411     | 11607      | 7246       | 8103       | 156434       | 158877       | 7974       | 2431       | 11          | 23          | 42237       | 57145       |
| 0859bada-df  | 33808        | 32925        | 33133        | 17699      | 17871      | 1262       | 1698       | 15302        | 16840        | 2360       | 641        | 17          | 6           | 42855       | 36135       |
| 08b92aea-9a  | 17536        | 17259        | 17602        | 16693      | 11146      | 1176       | 987        | 10310        | 10750        | 2901       | 827        | 2           | 7           | 20288       | 19015       |
| 0963590c-7c  | 33214        | 34013        | 33763        | 39531      | 25407      | 1824       | 2951       | 11526        | 11658        | 4216       | 1009       | 11          | 16          | 93039       | 43647       |
| 097683f7-5d  | 52762        | 52603        | 53172        | 69877      | 11689      | 3002       | 5682       | 17186        | 17490        | 4598       | 2003       | 24          | 34          | 54655       | 96168       |
| 09df6fb8-c2c | 49757        | 49391        | 49404        | 80340      | 2217       | 2936       | 6262       | 22251        | 22314        | 2725       | 1555       | 4           | 10          | 14426       | 26803       |
| 0a4e1dee-7b  | 82413        | 82164        | 83139        | 27041      | 16057      | 1459       | 12192      | 17563        | 18127        | 3674       | 1203       | 7           | 7           | 8108        | 153174      |
| 0b66effa-90f | 123155       | 122989       | 122917       | 69864      | 53468      | 2438       | 5362       | 87106        | 89686        | 8784       | 1778       | 24          | 24          | 68520       | 194575      |
| 0b74f41b-17  | 59143        | 58749        | 59842        | 34779      | 18478      | 8572       | 1386       | 54348        | 54435        | 3691       | 899        | 3           | 8           | 44267       | 52282       |
| 0ca2657a-87  | 31043        | 30699        | 31040        | 38303      | 5818       | 6475       | 1414       | 18351        | 18767        | 2483       | 1140       | 7           | 8           | 5287        | 16720       |
| 0cb6c6e1-d1  | 20252        | 20294        | 20582        | 27044      | 13034      | 1275       | 1283       | 7237         | 7559         | 2171       | 786        | 5           | 10          | 20812       | 75085       |
| 0d2141c1-c2  | 12397        | 12188        | 12386        | 15229      | 5522       | 1017       | 1310       | 4279         | 4628         | 1153       | 571        | 9           | 12          | 12007       | 6663        |
| 0d4f7e7b-dc  | 39383        | 38839        | 39335        | 59544      | 12698      | 1113       | 2610       | 19223        | 19254        | 2682       | 921        | 32          | 26          | 30951       | 67266       |
| 0e49336b-5f  | 26687        | 26495        | 26832        | 12188      | 5372       | 3082       | 520        | 21281        | 21813        | 4777       | 859        | 1           | 0           | 5760        | 38744       |
| 0f5c1ab7-f4f | 21277        | 21083        | 20969        | 13011      | 4034       | 1421       | 2318       | 12194        | 13021        | 2747       | 695        | 5           | 3           | 40766       | 31558       |
| 0fa60991-5e  | 103984       | 103978       | 104897       | 119307     | 18425      | 5540       | 14553      | 57411        | 58649        | 5797       | 3534       | 13          | 9           | 59530       | 38280       |

|    |     | BTF         | BTG        | BTH        | BTI         | BTJ         | BTK   |  |
|----|-----|-------------|------------|------------|-------------|-------------|-------|--|
| 1  | 95  | hsa-mir-950 | hsa-mir-96 | hsa-mir-98 | hsa-mir-99a | hsa-mir-99b | label |  |
| 2  | 15  | 0           | 177        | 250        | 1040        | 22624       | 1     |  |
| 3  | 40  | 0           | 12         | 197        | 8052        | 27649       | 1     |  |
| 4  | 57  | 0           | 200        | 43         | 5979        | 19477       | 1     |  |
| 5  | 49  | 0           | 93         | 752        | 2946        | 238146      | 1     |  |
| 6  | 29  | 0           | 7          | 771        | 5192        | 24778       | 1     |  |
| 7  | 63  | 0           | 8          | 469        | 8798        | 131700      | 0     |  |
| 8  | 38  | 0           | 2          | 513        | 338         | 38054       | 1     |  |
| 9  | 21  | 0           | 0          | 831        | 157         | 47211       | 1     |  |
| 10 | 49  | 0           | 311        | 1343       | 1546        | 73135       | 1     |  |
| 11 | 25  | 0           | 5          | 161        | 3650        | 24824       | 1     |  |
| 12 | 22  | 0           | 74         | 168        | 2462        | 19346       | 1     |  |
| 13 | 68  | 0           | 126        | 125        | 14658       | 75655       | 1     |  |
| 14 | 39  | 0           | 22         | 245        | 6248        | 154333      | 1     |  |
| 15 | 148 | 0           | 202        | 345        | 873         | 124912      | 1     |  |
| 16 | 86  | 0           | 97         | 197        | 8360        | 283988      | 1     |  |
| 17 | 33  | 0           | 8          | 482        | 11402       | 63169       | 0     |  |
| 18 | 5   | 0           | 0          | 649        | 7404        | 11177       | 1     |  |
| 19 | 25  | 0           | 44         | 396        | 1369        | 29653       | 1     |  |
| 20 | 10  | 0           | 3          | 93         | 5415        | 32795       | 0     |  |
| 21 | 21  | 0           | 5          | 73         | 1460        | 36546       | 1     |  |
| 22 | 7   | 0           | 16         | 109        | 3595        | 32811       | 1     |  |
| 23 | 64  | 0           | 106        | 201        | 3297        | 7269        | 1     |  |
| 24 | 15  | 0           | 22         | 106        | 2109        | 35195       | 1     |  |
| 25 | 201 | 0           | 148        | 687        | 4355        | 138078      | 1     |  |
| 26 | 27  | 0           | 5          | 246        | 8823        | 45085       | 0     |  |
| 27 | 9   | 0           | 17         | 217        | 1138        | 35981       | 1     |  |
| 28 | 158 | 0           | 49         | 366        | 1482        | 99628       | 1     |  |
| 29 | 14  | 0           | 1          | 299        | 2013        | 34278       | 1     |  |
| 30 | 29  | 0           | 5          | 221        | 7683        | 49631       | 0     |  |
| 31 | 18  | 0           | 1          | 127        | 686         | 13523       | 1     |  |
| 32 | 6   | 0           | 62         | 56         | 226         | 13816       | 1     |  |
| 33 | 12  | 0           | 5          | 443        | 7528        | 62275       | 0     |  |
| 34 | 44  | 0           | 22         | 471        | 10104       | 113164      | 0     |  |
| 35 | 66  | 0           | 706        | 361        | 2106        | 7150        | 1     |  |
| 36 | 7   | 0           | 101        | 121        | 4210        | 12144       | 1     |  |

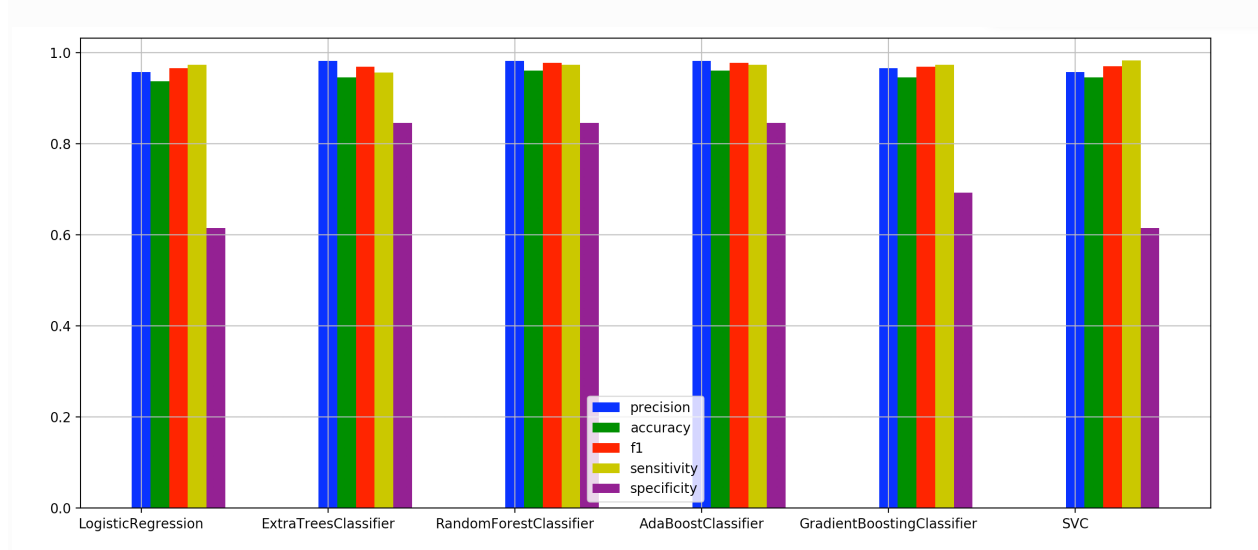
**Part 2: Apply Machine Learning Package (sklearn) to the above data.**

Sample code provided: *predict.py*

The steps:

1. Data standardization.
2. Feature selection.
3. Model fit with gridsearch cross-validation.
4. Evaluation: Precision, Sensitivity, Accuracy, F1-score, Specificity

Please explain the evaluation metrics.



Please try a different model other than the models used in the sample code. Also plot the ROC curve for the model applied.

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 Cancer* **volume 118**, pages1492–1501 (2018).