



# Cancer Genomics

data mining through The Cancer Genome  
Atlas (TCGA)

Xianglong Tan  
(Biological Chemistry department, UCLA DGSOM)  
[xianglongtan@mednet.ucla.edu](mailto:xianglongtan@mednet.ucla.edu)

# Syllabus

## 1. Introduction and Data Downloading Guide

GDC Portal, barcode, website analysis tool

## 2. Utilizing the R Package TCGAbiolinks

Query, retrieve, prepare data, maftools, DESeq

## 3. Integrated Data Analysis

K-M Curve, UMAP

R > 4.1 and Rstudio is necessary from Day2

## Goals

The ability to retrieve and analyze specific cohorts of data from TCGA and additional sources, tailored to your individual research interests.

## Grades

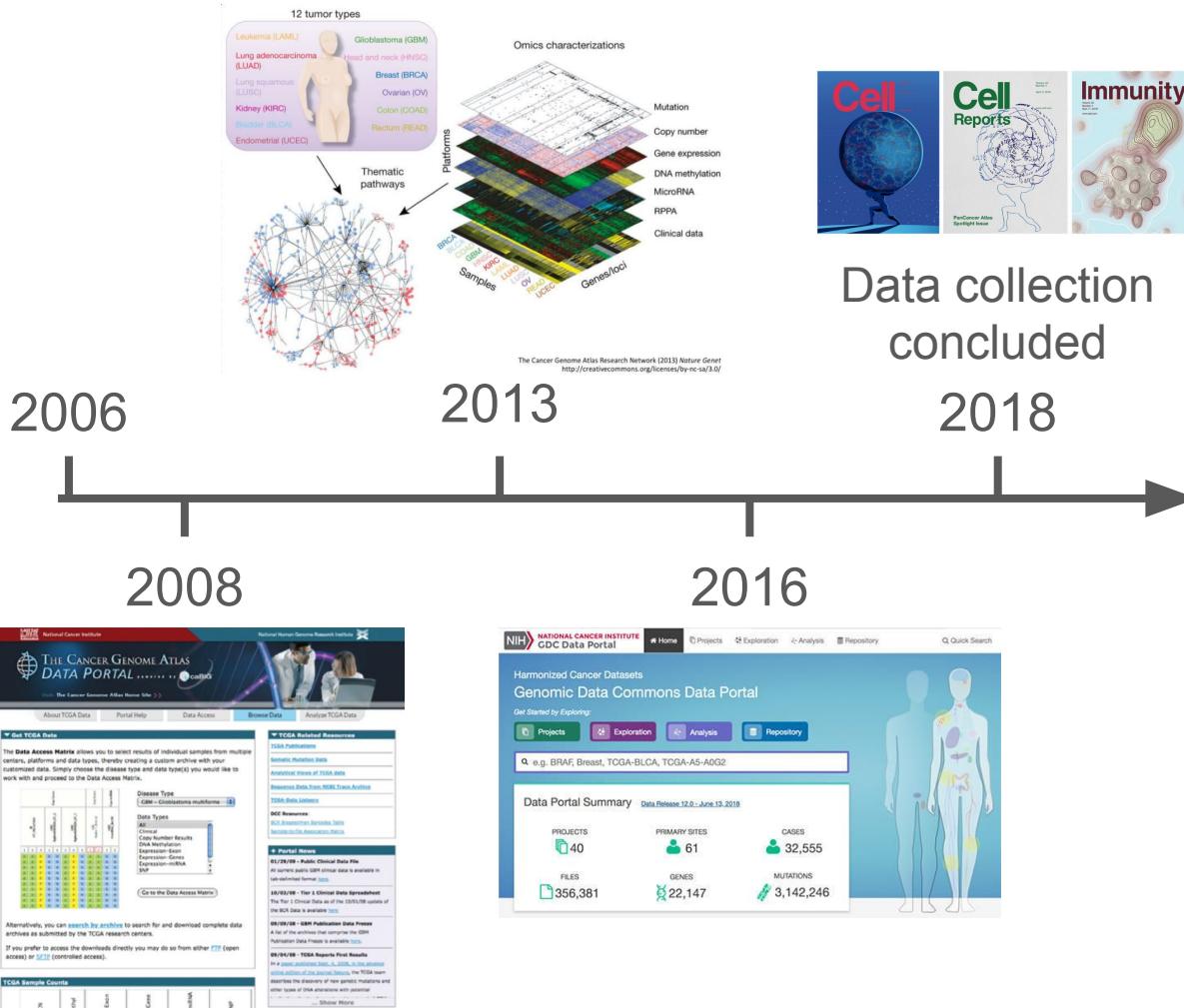
for students who enroll for academic credit only

- 25% Participation
- 50% Quiz
- 25% Homework

# Day 1 – Introduction and Guide to Data Download

# What is TCGA

# The Cancer Genome Atlas (TCGA)



Data collection concluded

NATIONAL CANCER INSTITUTE  
THE CANCER GENOME ATLAS

## TCGA BY THE NUMBERS

TCGA produced over **2.5 PETABYTES** of data

To put this into perspective, **1 petabyte** of data is equal to

**212,000 DVDs**



TCGA data describes **33 DIFFERENT TUMOR TYPES** and **10 RARE CANCERS** ...including

...based on paired tumor and normal tissue sets collected from

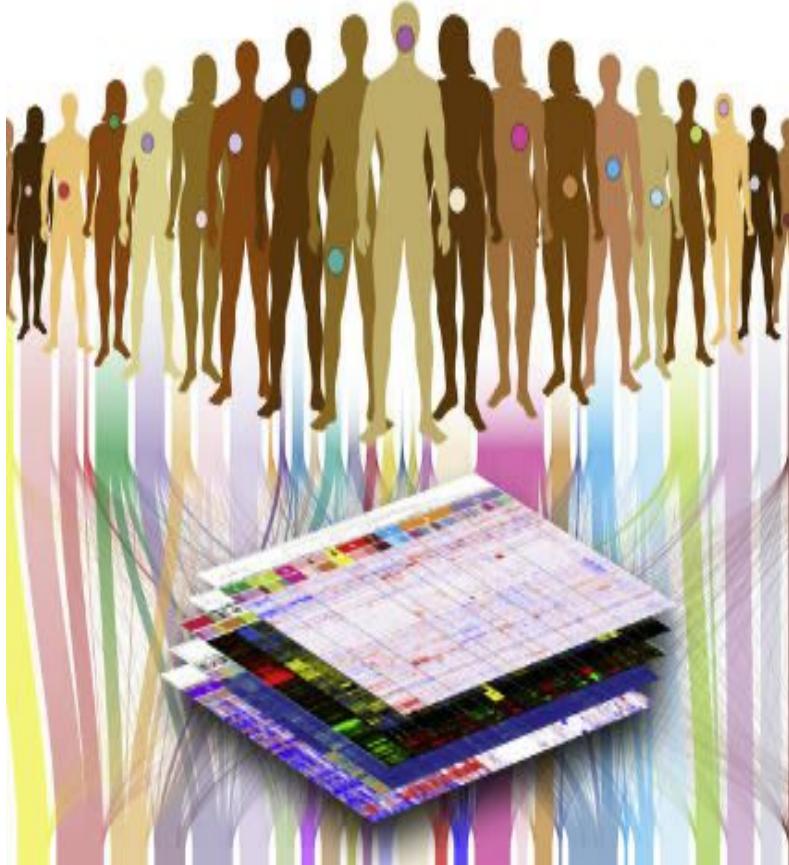
**11,000 PATIENTS**

...using **7 DIFFERENT DATA TYPES**



# The Cancer Genome Atlas (TCGA)

- Comprehensive Data Collection
- Integration of Multiple Data Types
- Open Access
- High-Quality Standards
- Standardized Computational Pipelines



# Exploring the diverse data types of TCGA

- **Clinical Data**
- **Biospecimen Data**
- **Pathology Reports**
- **Imaging Data**
- **Copy Number Variation (CNV) Data** – W13: Genetic Analysis
- **DNA Sequencing Data** – W8: Variant Calling with GATK
- **Methylation Data** – W6: BS-Seq DNA Methylation analysis with Hoffman2 and R
- **Microsatellite Instability Data**
- **miRNA Sequencing Data**
- **mRNA Expression Data** – W5: RNA-Seq analysis
- **Protein Expression Data**

# Data availability

## Summary of Data Types Collected

The data collected for a specific case in TCGA may have differed according to sample quality and quantity, cancer type, or technology available at the time of analysis. Below is a general summary of the types of clinical, molecular characterization, and other types of data that may have been generated for the different cancer types studied.

*Raw data (e.g. BAMs), germline and non-validated mutations, and genotypes are under controlled access (indicated in red). Derived data is available open access (exceptions are noted in table below).*

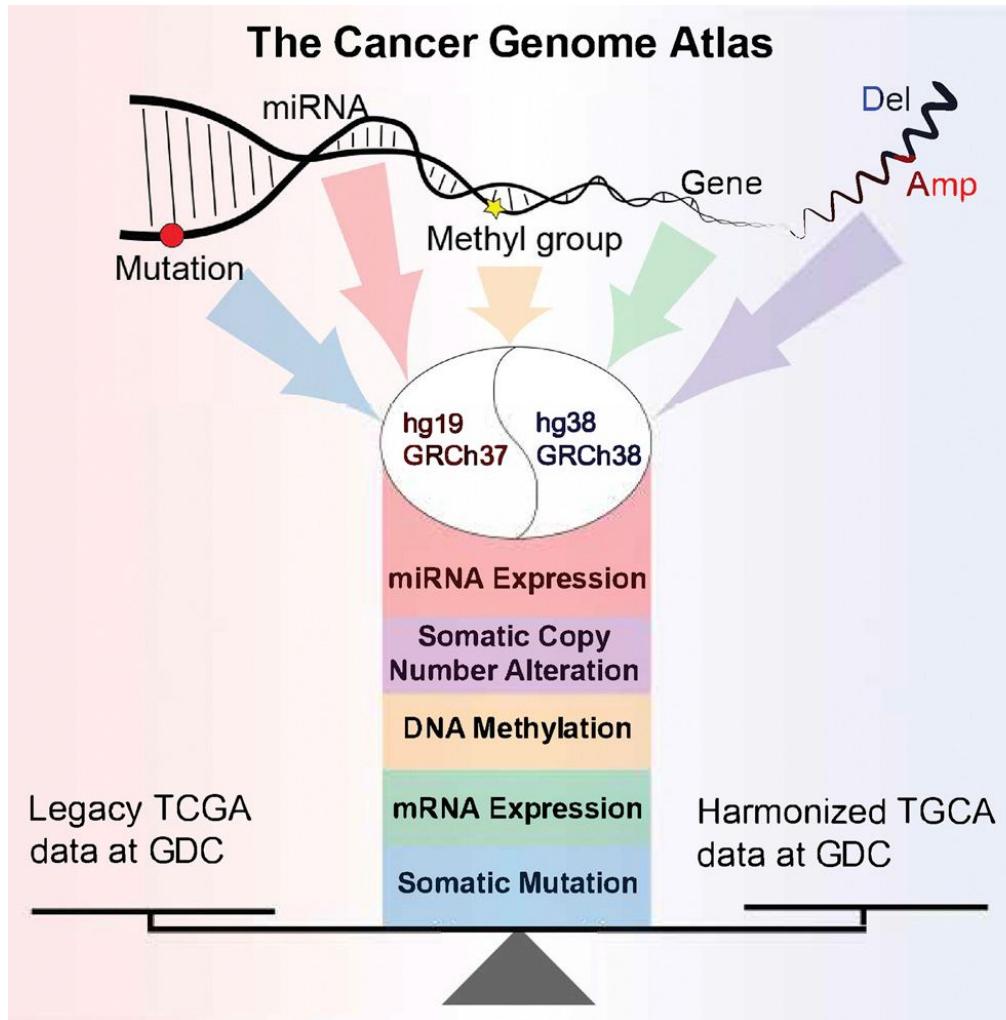
All data collected and processed by the program is available at the [Genomic Data Commons](#) (GDC), including TCGA publication supplemental and associated data files. Questions about locating or accessing data should be directed to the [GDC support team](#). Resources for TCGA users and [TCGA FAQs](#) are available.

Experimental protocols for each platform can be found in [individual publications](#).

Obtaining access for controlled data:

<https://gdc.cancer.gov/access-data/obtaining-access-controlled-data>

# Harmonized Data



Encyclopedia:

<https://docs.gdc.cancer.gov/Encyclopedia/>

# Methods for Mining TCGA Data

- Direct via **GDC Portal**
- **GDC Client Software** (Linux, MacOs, Windows)
- R-packages(**TCGAbiolinks**, RTCGA.....)

# The GDC Portal

https://portal.gdc.cancer.gov

NATIONAL CANCER INSTITUTE  
GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary Data Release 37.0 - March 29, 2023

PROJECTS 78

PRIMARY SITES 68

CASES 86,962

FILES 931,947

GENES 22,501

MUTATIONS 2,885,293

Cases by Major Primary Site

Cancer Type	Cases (in thousands)
Adrenal Gland	0.1
Bile Duct	0.2
Bladder	0.5
Bone	0.3
Bone Marrow	8.5
Brain	0.5
Breast	8.5
Cervix	0.2
Colorectal	7.5
Esophagus	0.2
Eye	0.1
Head and Neck	2.5
Kidney	3.5
Liver	0.5
Lung	12.0
Lymph Nodes	0.2
Nervous System	4.5
Ovary	3.0
Pancreas	2.5
Pleura	0.5
Prostate	1.5
Skin	2.5
Soft Tissue	1.5
Stomach	1.5
Testis	0.5
Thymus	0.5
Thyroid	1.5
Uterus	1.5

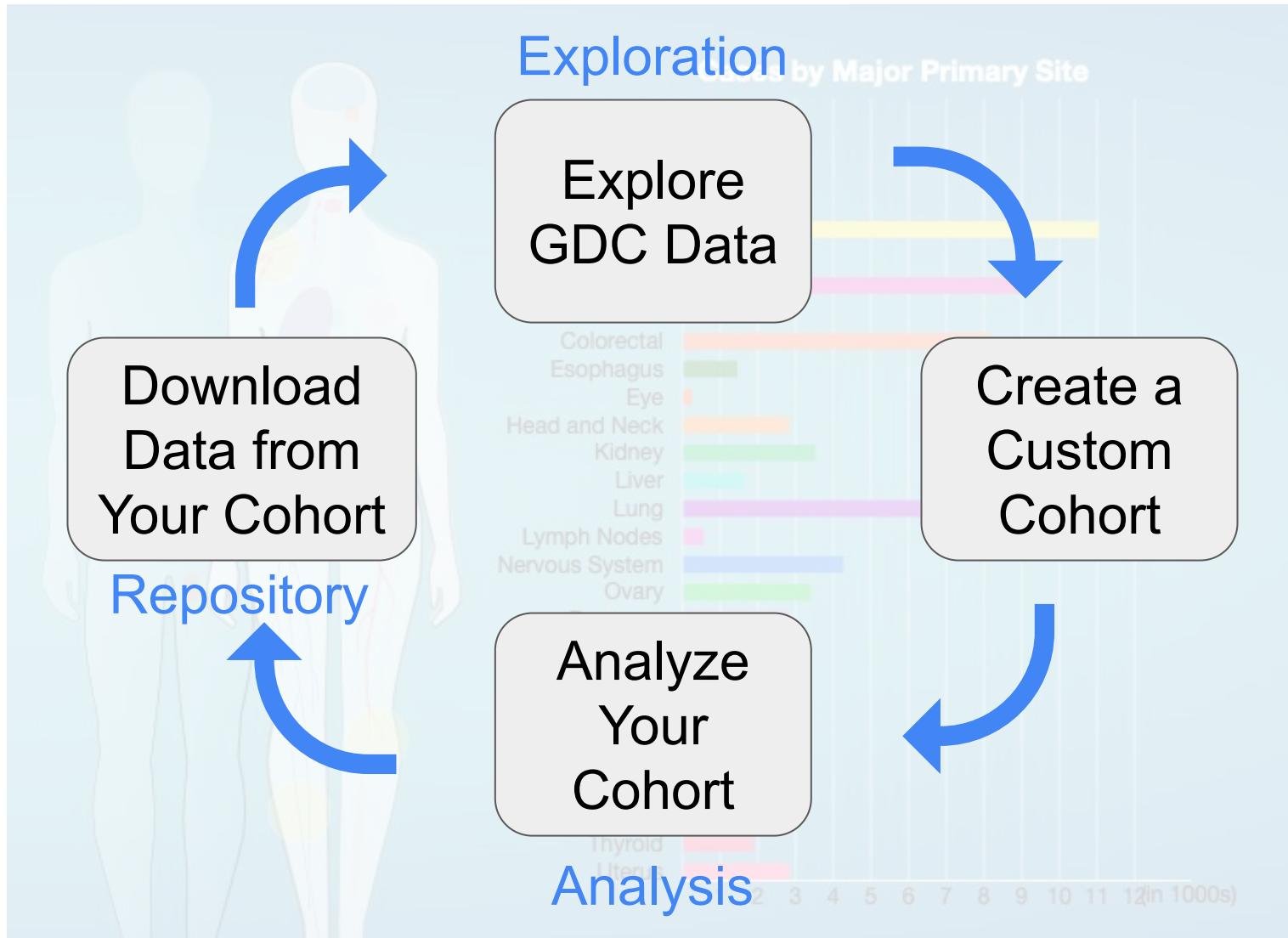
GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Data Portal Website API Data Transfer Tool Documentation Data Submission Portal Publications

<https://portal.gdc.cancer.gov/>

# Workflow by using GDC Portal



# An example of using GDC Portal

1. Build a cohort of female lung adenocarcinoma, >45 years.
2. Do a comparison of this cohort with a similar cohort except for males, and with tobacco smoking status.
3. Do an analysis on some major clinical properties (survival, gender, race...) for this cohort.
4. Download the top few txn files from this cohort by GDC-Client

Let's build our cohort

# Manifest file

(base) [xianglongtan@REDLBICADM29372:~/media/data/xianglongtan/cancer_project/TCGA_biolinks\$ head MANIFEST.txt			
id filename md5 size state			
021326fa-7b6c-45c0-b078-4646c2260068	021326fa-7b6c-45c0-b078-4646c2260068/97b043a9-8872-4f75-be81-aed7d0bd849d.rna_seq.augmented_star_gene_counts.tsv	847ef91f3552621a994df1289a6fd6e5	422
8056 validated			
056fd446-19f7-4d9f-975b-ca8d2c2e944a	056fd446-19f7-4d9f-975b-ca8d2c2e944a/ecee2642-e793-457d-a325-320a785c917c.rna_seq.augmented_star_gene_counts.tsv	3de5e34c29dcc4e1b0da594a9f9be3ea	424
6549 validated			
09152b74-f690-4dbc-a398-1f8dde28ec06	09152b74-f690-4dbc-a398-1f8dde28ec06/868bc289-3c43-4a0f-bbb1-703f0e9fa33a.rna_seq.augmented_star_gene_counts.tsv	51be60d63f848fcba03454f8a9f7e4e9	426
1356 validated			
0e480f18-0e72-444e-b7e1-956a439471ad	0e480f18-0e72-444e-b7e1-956a439471ad/a21c0490-049b-4451-a714-89007edb62f5.rna_seq.augmented_star_gene_counts.tsv	d1c3182edd30ba92ad70fe402f711deb	422
1130 validated			
0fe3bf01-415a-4945-8606-74b92ae9da44	0fe3bf01-415a-4945-8606-74b92ae9da44/d58fe5dc-9a75-42e7-94af-2ab20a21dab1.rna_seq.augmented_star_gene_counts.tsv	c629fa266a5bf5567c531884c8339292	423
5061 validated			
10bec46a-08c3-4fe7-a65a-d801705fca17	10bec46a-08c3-4fe7-a65a-d801705fca17/90b40e5e-f443-45aa-a810-7026010c41f4.rna_seq.augmented_star_gene_counts.tsv	45fec7e248602354efb4c8e32e6d0a92	424
7739 validated			
12898e2b-86d5-4c0e-8a53-dec1fd528f42	12898e2b-86d5-4c0e-8a53-dec1fd528f42/cb8d1dc4-bfc1-46ba-8e31-af63a0fd5eab.rna_seq.augmented_star_gene_counts.tsv	c6f4933978d266b4df730d353861aaa6	424
2821 validated			
14350006-3c6f-4079-8a35-afd1e5f8a29a	14350006-3c6f-4079-8a35-afd1e5f8a29a/201a8762-54e6-49a1-a9c8-a6c442dcfa0a.rna_seq.augmented_star_gene_counts.tsv	9f980154339d7e94c5e5eb642ff12593	424
7090 validated			
1a80f876-7306-4068-998a-3f32849173ca	1a80f876-7306-4068-998a-3f32849173ca/64499b79-1354-4277-9142-40f388fa6e92.rna_seq.augmented_star_gene_counts.tsv	887d2aadd51bd7b60c61b8ff01814f66	424
7630 validated			

# Universally Unique Identifier (UUID)

## Universally Unique Identifier (UUID)

### Description

A UUID (Universal Unique Identifier) is a 128-bit number used to uniquely identify an object or entity in a system.

### Overview

In the GDC every entity in the data model is assigned a UUID. An entity can be a file, case, project, or other node. Uniquely identifying objects in this manner allows the GDC to have a uniform method of referencing any object in the system<sup>1</sup>.

### References

1. [GDC Entity UUIDs](#)

### External Links

- [!\[\]\(bab4345c1dc8595a37869a09797e1e95\_img.jpg\) Universally unique identifier](#)
- [!\[\]\(53a1175c69e72875804c05c0840bd0e7\_img.jpg\) RFC 4122 UUID](#)

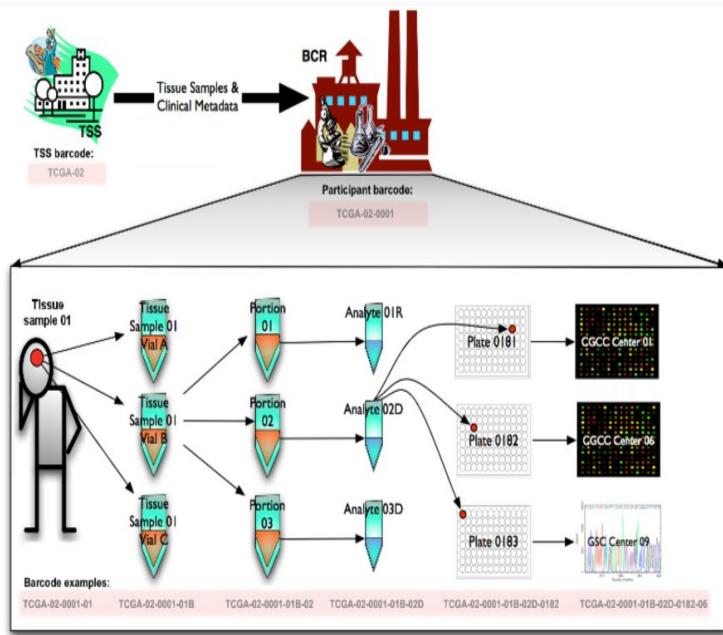
Categories: General

Next: Variant Call Fo

# Clinical datasets (clinical.tsv)

case_id	case_submitter_id	project_id	age_at_index	age_is_obfuscated	cause_of_death	cause_of_death_source	country_of_residence_at_enrollment	days_to_birth	days_to_death	ethnicity	gender	occupation
ae86a89-0377-4080-b16c-408bfbe78687	TCGA-69-7980	TCGA-LUAD	70	'--	'--	'--		-25583	'--	not hispanic or latino	female	'--
ae86a89-0377-4080-b16c-408bfbe78687	TCGA-69-7980	TCGA-LUAD	70	'--	'--	'--		-25583	'--	not hispanic or latino	female	'--
c9c533ee-e154-4a56-bce9-b5af37574b2f	TCGA-55-7913	TCGA-LUAD	61	'--	'--	'--		-22326	561	not hispanic or latino	female	'--
c9c533ee-e154-4a56-bce9-b5af37574b2f	TCGA-55-7913	TCGA-LUAD	61	'--	'--	'--		-22326	561	not hispanic or latino	female	'--
cd9e70e4-8622-4a07-8646-63f8275c1737	TCGA-49-AARE	TCGA-LUAD	51	'--	'--	'--		-18893	1229	not hispanic or latino	female	'--
cd9e70e4-8622-4a07-8646-63f8275c1737	TCGA-49-AARE	TCGA-LUAD	51	'--	'--	'--		-18893	1229	not hispanic or latino	female	'--

# The TCGA barcode system



## Reading Barcodes

A TCGA barcode is composed of a collection of identifiers. Each specifically identifies a TCGA data element. Refer to the following figure for an illustration of how metadata identifiers comprise a barcode. An aliquot barcode, an example of which shows in the illustration, contains the highest number of identifiers.



Label	Identifier for	Value	Value Description	Possible Values
Analyte	Molecular type of analyte for analysis	D	The analyte is a DNA sample	See Code Tables Report
Plate	Order of plate in a sequence of 96-well plates	182	The 182nd plate	4-digit alphanumeric value
Portion	Order of portion in a sequence of 100 - 120 mg sample portions	1	The first portion of the sample	01-99
Vial	Order of sample in a sequence of samples	C	The third vial	A to Z
Project	Project name	TCGA	TCGA project	TCGA
Sample	Sample type	1	A solid tumor	Tumor types range from 01 - 09, normal types from 10 - 19 and control samples from 20 - 29. See Code Tables Report for a complete list of sample codes
Center	Sequencing or characterization center that will receive the aliquot for analysis	1	The Broad Institute GCC	See Code Tables Report
Participant	Study participant	1	The first participant from MD Anderson for GBM study	Any alpha-numeric value
TSS	Tissue source site	2	GBM (brain tumor) sample from MD Anderson	See Code Tables Report

# The GDC Sample Sheet

gdc\_sample\_sheet.2023-11-23

File ID	File Name	Data Category	Data Type	Project ID	Case ID	Sample ID	Sample Type
6feef177-114c-4285-b242-481fe0aea551	77322072-2a7f-49e6-a9d1-c521c25acd70.rna_seq.augmented_star_gene_counts.tsv	Transcriptome Profiling	Gene Expression Quantification	TCGA-LUAD	TCGA-69-7980	TCGA-69-7980-01A	Primary Tumor
4089d037-ab25-47a6-be68-19742473cbc6	50c308c9-922a-4083-ae09-e5e4d8c437af.rna_seq.augmented_star_gene_counts.tsv	Transcriptome Profiling	Gene Expression Quantification	TCGA-LUAD	TCGA-49-AARE	TCGA-49-AARE-01A	Primary Tumor
a0fce6cc-a525-4970-82b7-8512bb0708b3	7d9c7c34-76e0-4622-b5f7-60e12f6d4e07.rna_seq.augmented_star_gene_counts.tsv	Transcriptome Profiling	Gene Expression Quantification	TCGA-LUAD	TCGA-55-7913	TCGA-55-7913-01B	Primary Tumor

# The GDC-client Software

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

```
zhuqian@xianglongtan:Software$ ./gdc-client
usage: gdc-client [-h] [--version] {download,upload,settings} ...

The Genomic Data Commons Command Line Client

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit

commands:
  {download,upload,settings}
    download           for more information, specify -h after a command
    upload              upload data to the GDC
    settings            display default settings

gdc-client error: the following arguments are required: command
```

```
./gdc-client download -m /Users/zhugian/Downloads/gdc_manifest.2023-11-18.txt -d
/path/to/your/desired/directory
```

```
gdc-client download 22a29915-6712-4f7a-8dba-985ae9a1f005
```

```
gdc-client download e5976406-473a-4fb9-8c97-e95187cdc1bd fb3e261b-92ac-4027-b4d9-eb971a92a4c3
```

# Few windows users



xi

You

in windows cmd, how to navigate to the path of my interest



ChatGPT

To navigate to a specific path in Windows Command Prompt (CMD), you can use the `cd` (change directory) command. Here's how you do it:

1. **Open Command Prompt:** You can do this by searching for "cmd" in the Windows search bar and clicking on the Command Prompt application.
2. **Use the `cd` Command:** To change the directory, type `cd` followed by the path to your desired directory. For example:
  - If you want to navigate to a folder on the C drive, such as `C:\Users\YourName\Documents`, you would type:

bash

Copy code

```
cd C:\Users\YourName\Documents
```

- If the folder is on a different drive, you first need to switch to that drive by typing the drive letter followed by a colon. For example, to switch to the D drive:

makefile

Copy code

D:

Then use the `cd` command:

bash

Copy code

```
cd D:\MyFolder
```



### 3. Open Command Prompt:

- You can open the Command Prompt by searching for "cmd" in the Windows search bar.

### 4. Navigate to the Directory:

- Use the `cd` command to navigate to the directory where you want to download files.

### 5. Run gcd-client:

- To run the gcd-client, simply type `gcd-client` followed by the appropriate command and arguments. For example, to download a file, you would use something like:

Copy code

```
gcd-client download -m manifest.txt
```

Here, `-m manifest.txt` specifies a manifest file that lists the files you want to download.

### 6. Check for Additional Commands:

- The GDC Data Transfer Tool has various commands and options for different purposes (like downloading specific files, resuming interrupted downloads, etc.). You can usually find a list of these commands and how to use them in the tool's documentation or by typing `gcd-client --help` in the Command Prompt.

## Show your current working directory:

2. Once the Command Prompt window is open, you will be in a specific directory by default. To display the current working directory, simply type the following command and press Enter:

shell

Copy code

cd

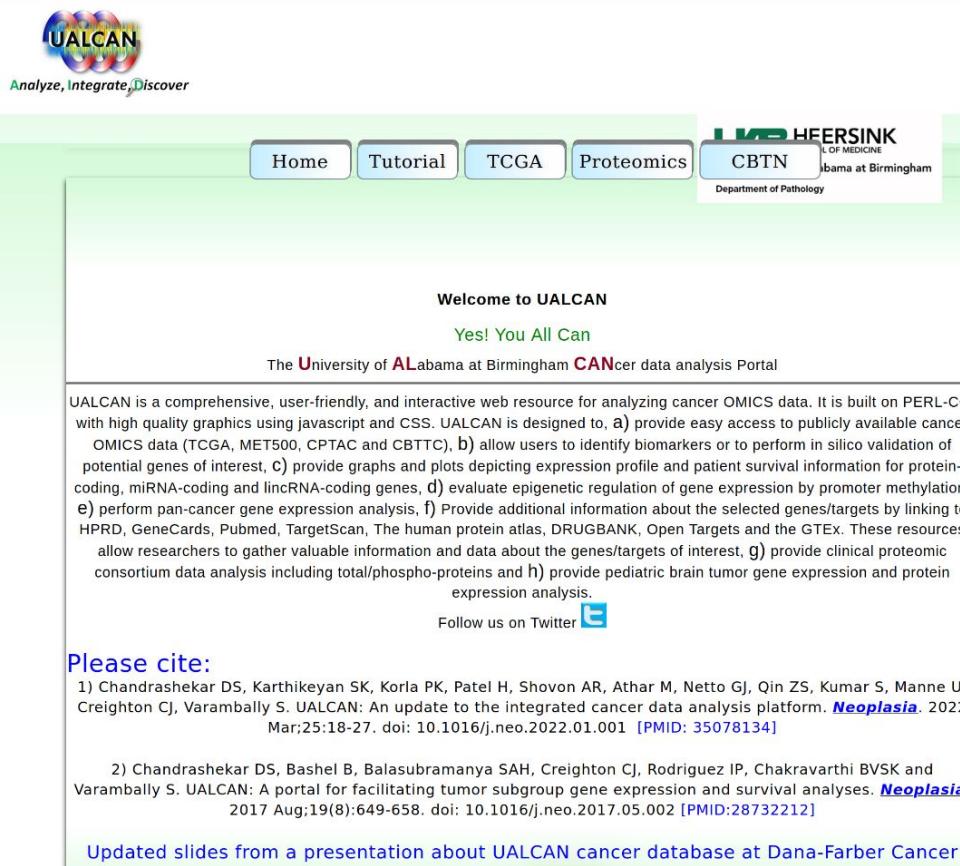
1. The Command Prompt will then display the full path of the current working directory.

# Download via GDC-client

```
25857230-7253-4929-a790-90815194b213/ 54a00dd0-503f-471f-88e4-8d6fcdfc8483/ 89785c02-2345-4fa3-b73b-4d9da706c972/ b7b898c7-f63f-d3d-864a-406e842db2fe/ e98c7df6-1a65-41a0-9182-c752c89eacf7/ 25f44be3-2e4f-4121-ac23-21bd3866dbce/ 555bdf8-80bb-46db-80f1-421038f5c2b4/ 89bdfb20-8991-4bc4-b112-e9be096b566/ b817c1a8-7e62-423e-b7e0-eb9136fe0455/ e9d3cac1-3a16-4e21-8fe8-a4944ce82cb8/ 269a7c8d-f4d6-4f1a-a85b-874934f214b9/ 55831884-c907-457d-bb2c-7c7b37fc4cd1/ 89bf9c33-c1b4-4786-821f-802fc36eed43/ b91d8293-6331-4f76-98f0-335ebce83539/ ea47f0f6-aad1-4149-abf6-2026f64935f5/ 26c96b62-02b6-4108-9a2d-bec3d880db2a/ 55bbaa27-e618-48d6-bac3-80e9cfcfdcf2c/ 8a2d1te2-4ec1-42b6-8825-6d80ffe4b916/ b9610938-1f47-4556-ae92-794df8ae6401/ eac45bfd-6795-47a4-8d6b-5a7c8bc76cd1/ 26ea6a31-ea33-4bb9-814f-c0840d5afe9e/ 5601dc73-ba9b-4f53-95ef-5fc3de23640c/ 8a518a19-2c85-42ec-b027-47139aa442d4/ b9ac77d6-9a1b-443d-8ef5-e8ef2e3008rd/ eac5ehb0-d788-4908-reb7-976d6b111667/ 270a13b3-158c-4251-8c6a-5bc5275e0ebb/ 5695292b-7c39-4ee6-94e8-987f4a456635/ 8a884668-03c2-44bb-8ea0-5b0d0bc68aa4/ ba3e1die-3193-4770-9c57-b4ff1502af38/ eb29ba2f-e237-403c-ab72-a5dfaet05b05/ 27450388-ffc7-4812-5b54-9819b0a74a0b/ 5669a740-e0e72-4c10-a90d-921034141b42/ 8a898b7d-6f1b-4d2a-9f5d-0751ced1d7bf/ ba8aeab8-07de-46c0-a28f-b2408c36f07f/ ece9f585-2710-4e10-b292-c5fbf588eda9/ 27884724-a2c0-4b3e-921a-54e5f8bf1faa/ 576ad342-ce70-4c3c-b69d-8f1f5c4cd358/ 8af996d7-2c5e-4d17-b1a5-8751f4784eba/ bae11f13-8444-4fe3-b9e4-94db874896af/ ed3bbc-0f04-415f-a083-e34cefaf9e310/ 29bb005f-53d0-44e6-8626-d6bbe0dba8be/ 577f10d0-c9a2-47c3-aa2f-b1bd5bd86d98/ 8b071bf3-0d85-4aba-aa46-d0800dccc8640/ bae501f3-c0ba-4464-8027-0623869176f4/ edfd7d11-4764-43a7-9948-4320dc17b2b1/ 29dd3ad7-f2c6-4abf-8e81-6734e4bd8a6a/ 583900e8-43b2-4517-b57c-ff899594e55a/ 8b26e78c-f162-48fc-8c1b-9628baa3ca62/ bb47dbbf-75f5-4bbf-bd1a-ca81383f211b/ eeacf4ac-09cf-41a8-8066-c38a173e61b4/ 2a7831f2-a5-a6f-92b3-0a7a727d5637/ 583d72d-6cff-4c5c-9317-13d88cd71b79/ 8b55c916-963b-41d6-b7ce-7e8cf958c0/ bb5264e1-0d0bf-4b28-9fbe-a08602c13369/ ef2621f5-ec61-42ee-879f-9d7615372a5e/ 2a9f6c99-7b04-4262-9eb2-d648ae027010/ 58dc4d5a-714f-4a22-b367-813dd41829d/ 8be16a1a-383f-4365-b1e5-139b1625e3ce/ b664143e-c48a-45c7-914b-4ec5bceae927/ ef337612-6a73-4c29-a8b0-85557cbeaff4/ 2ab0f01f-2f66-4489-9de7-2a0d1b6ea04/ 593e91b8-cbcd-4dfb-95ed-9dc8e1a430b/ 8be349e-2407-4f51-809a-ce5b542e0cf1/ b669e798-883c-4c12-b622-3155e1972e6f/ fdb3bf8d-7d14-42d6-8845-e1a9f09d47f/ 2bbe47a1-f055-43e7-94d3-06aaaf5baf6e8/ 5a37129a-9672-4dfa-8576-e7de1db7b2c/ 8cd8b3ed5-5922-4833-b255-30d9322f8fe9/ bbb2d469-806b-4417-a3cd-4dd1b0c8779/ ef4983d-f573-4996-bf5c-5aaa1aafd730/ 2bf191f1-b880-427e-996c-0562c137d3/ 5a7b0d24-3d92-4e64-86bd-b6eea34f6526/ 8d308359-f62a-4c2a-a2e7-b1a3b7bd79e/ b7b70a6e-91c-4c67-9e0c-b33185ab4f8d/ f0890456-1013-4e8a-9132-f05e2a1be5f2/ 2c81c69f-e08a-4586-b3f9-33d4a4f17d1f/ 5b99d571-43d2-4d26-9a3b-0a299a0ce048/ 8dbb93bf-bf45-4148-acdbd-6609d5c4a1e3/ bbd3cd5-cfd3-41b8-ab3a-243de430a5e1/ f19a1a66-ddc9-4567-80db-d009ba455b25/ 2d99f93c-797e-49a5-8892-a4e31e1b79c3/ 5c32fb8b-55ee-4bc9-a0db-d8ac0b04bd4c/ 8ddad1cb-d2ab-4fbf-865e-f134fcce0194/ bbea4239-0ce8-4540-8818-d73d076b3aa6/ f1b1d4ff-2556-44c0-80b0-1fd3b659c227/ 2e0ctcf-7294-4c86-043e-8215c9186d8/ 5caf3127-967f-4d07-b49d-7126728e048/ 8e140ba9-6132-483c-892b-2447047dc38/ b2a644e-3883-4097-af4e-5f68f1c793e/ 2e2eb2d-f002-49ac-86e7-5ed33e9abf3/ 5d5e6761-0c1b3-4e52-b14f-79be5580d4a/ 8e510682-b05c-4792-a149-d713bd6d082d/ f201ba16-5c5b-42d0-9865-e413f7535z/ 2e5071ce-d8cf-46e5-9cc0-91353fd0643c/ 5dd8168d-fe35-4a32-8520-1deae9f53b10/ 8fad6d4-de92-4301-9b6e-7d8857fec978/ f2db8b566-f179-4b8e-8b4d-672149eedcba/ 2f2cc99dc-1d0a-4302-b674-02d562c59ff3/ 5ddf8f4a-eddb-4b97-8d4d-766556910v8c9/ 901c6ff1-19bf-48d8-8949-766f1995a916/ bda198f-aedf-46cc-8156-c1e637ce88f/ 2f6e63bb-17bd-4b53-8666-45c7159a1f4/ 5ece8c-845f-43cf-82c3-73dea58cc53c/ 906a6a3f-91ce-4433-9a0d-e010a1678f0/ beb3712d-76e3-4c3a-becb-c94048ea87f/ 2faa40914-d149-4f67-a7a8-64723253814f/ 5ecf82a3-8658-49a4-14c4-f5d16f65f8/ 911a72ec-dcb3-4454-8f22-7bd4f9d29b21/ b7c02929-0661-4789-871b-d54ba5989786/ 2fe6f165-3396-4abd-a0cd-07986db2828/ 5edfb99-bc9f-84f7-459b-efea8738442c/ 9123a775-d686-49a9-90e7-0dc572c2fcf/ f51ce896-53b9-f4f3b-ac5f-a499b84a7add/ 2ff24c33-39c6-4171-af42-f54e4ee8b021/ 5eefc0d2-90b4-495f-803b-cfe9f158988e/ 9272bedd-0e7e-4695-b744-797f2519389e/ bfecd55-3ed3-43d9-4977-899b-d7c22b6df15b/ 306fbc7-f2b9-493a-9520-b9700d0f34e7/ 5ef157e9-bb3d-4214-946-4f5bc74b746/ 929473bf-f434c-438e-b16c-34b5fb098ea/ c140e8e-45f7-4fb2-21c3c8dde972e/ 5ef157e9-bb3d-4214-946-4f5bc74b746/ 929473bf-f434c-438e-b16c-34b5fb098ea/ 2accf79-6461-4082-b201-cf881bdc6ea2/ c12371d5-868b-4788-91ec-c94bd80805/ 30e51b69-6778-4125-bb18-86694d42a6e9/ 5f2d965a-25df-4736-a675-bef59341a04/ 924b4ef9-2b54-a4ec-9251-8dfb52d9e0/ 2accf79-6461-4082-b201-cf881bdc6ea2/ 313f5ae9-bbea-41d6-82d5-9242b2610d91/ 5f4fa91b-489e-a4c9-b685-53952f77bd7a/ 932ccb56-ad41-40a0-9ab3-21f490c60af5/ c343f781-c1df-4349-aab3-f159329a11da/ 3204eabb-8480-0901-0ce03107490/ 5fa5664e-459c-4b82-8d39-9d560e5049/ 93db566e-8f9g-45bo-a2-0fbed82f580/ c36f84e5-8c59-46fe-af32-77f0684383b2/ 320660cd-100a-4f5a-a604-4de3cd7f042/ 60473723-5657-495a-8e3c-4f806acd3cd8/ 942673a7-5072-466f-985c-42d38aeb5c2c/ c36f84e5-8c59-46fe-af32-77f0684383b2/ 3287f627-f304-4be0-8881-f9842a9fb67/ 614f0d6a-05ff-4914-896c-3ac4280e510d/ 94613d97-0a68-4212-beef-3c97a0b363ef/ c48eecc3-0099-4dd4-b319-f13ad65bdc7e/ 334cc9fa-3e9b-4462-884d-dc595e89d53t/ 62560d5c-0cae-4fc7-820a-9c1236b2c8/ 946929ad-ed52-4b8d-85ac-efe5b178b54/ c4bcccc6-96a1-452b-9169-b4a16bc9403/ 3361d7bc-344b-44b2-97bd-ac0d795b22f1/ 62d2dca7-9a0e-49e7-9423-24540d696049/ 94861d8a-3b40-4334-9f5b-c9df71a1eida/ c568f8c-6942-44ff-a9d9-3f7a03fdc62a/ 3362c701-3e3-4947-983a-6039b51269d/ 63a8839e-0f52-4c5b-4b85-72e095b51269d/ 94913b2f-256e-4487-90d-2c93869ca6b/ c5a1245e-f3c5-4f42-af6d-9b1f201a6445/ 347d7b78b-ddfc-46ed-9fe4-0ca52a5c432ef/ 63f5b69e-2307-4b6e-98e4-b0ff57b6394d0/ 948db81e-0198-4f40-987a-4469b1b303f/ c698e8f1-57a7-439a-9b05-66f14b019322/ 3563bc2e-6547-4081-883e-1e7e498c83be/ 64101504-305f-4c62-bc27-40493e3c9575/ 951f2356-17f4-4f4f-a8ec-855c7e93dc1/ c79162f6-be6e-46b1-acca7-438423fed6c3/ 359d07bf-8e93-49d7-8237-723fb299cfb/ 648b002-f264-428d-8997-baed8f000a60/ 968b4ca2-2b4d-4823-b4f73-7bd4f9d25/ c48eecc3-0099-4dd4-b319-f13ad65bdc7e/ 3604ee2d-3fdc-40fa-95ab-a3f4203343ac/ 64e53ca3-5b79-42e4-831d-3b7d2f213e/ c79162f6-be6e-46b1-acca7-438423fed6c3/ 360ebab-859c-4286-b613-befa0fa306f0/ 6508912f-5f91-43e6-8e63-9233bf3e81eb/ 9726aa3d-8a66-4568-92e5-c092c5a9a5d/ c97d6c2a-9866-4893-8cff-ee747c4a4550/ 36101746-9c10-478e-96fb-847d057a0d4/ 6525f4e0-ba79-4576-8983-df0a0b6e45309/ 975121d2-ba5d-4b2c-d05b-278311c7144a/ ca34be5f-799c-4df4-af91-ec8a9e958d6c/ (base) xianglongtan@REDLBICADM29372:/media/data/xianglongtan/cancer_project/RNA_seq/LUAD/_LUSC_New/GDCdata/TCGA-LUAD/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification$ ls */*tsv | head -d 0052a83-7ae5-470a-a125-5cd94a9fa9e9/a6a6b9c6-9db7-42b3-a09f-77b07e126fb.b RNA_seq.augmented_star_gene_counts.tsv 023a34d9-c000-4053-b695-5b984b4a6fc1/06770623-6a10-4874-9eaa-1497077f18ac RNA_seq.augmented_star_gene_counts.tsv 030d778f-ecb4-44b6-9bba-4286e18c9fd/84595c09-6f85-4a1d-98fa-7d9b9cd2c28 RNA_seq.augmented_star_gene_counts.tsv 0344d3bb-8bb2-459e-bc1f-ba31c7470d08/39807893-979e-44c6-ad49-444c6bb863c3 RNA_seq.augmented_star_gene_counts.tsv 03680aea-84a2-4775-bb9c-24f8c5907247/04e0b83-7afa-475b-8586-e9e79e31d6d2 RNA_seq.augmented_star_gene_counts.tsv 04778fc1-82a1-4029-3d887-1f3b3ee/22c9d1fc-5c45-43e0-987a-19a4129271 RNA_seq.augmented_star_gene_counts.tsv 055fcce9-7448-4a16-ad18-d235dc4tce97/b95a1889-2ef5-4d43-b5e9-14b678c4d71a RNA_seq.augmented_star_gene_counts.tsv 05fa4573-6ca5-40dc-a48b-ec847b7c58ad/58681163-9f9d-4f74-a1bb-3abcb4856d6 RNA_seq.augmented_star_gene_counts.tsv 06a639d9-ad5c-45c-9b10-a4149e4ce4f79/6531e137-3d23-44e7-af10-51e9cc517192 RNA_seq.augmented_star_gene_counts.tsv 06df5913-5f62-4200-8a51-26c8f2052b4/aae809a-97-730b-4180-93d65dbbd26429419 RNA_seq.augmented_star_gene_counts.tsv (base) xianglongtan@REDLBICADM29372:/media/data/xianglongtan/cancer_project/RNA_seq/TCGA_LUAD/_LUSC_New/GDCdata/TCGA-LUAD/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification$ ls */*tsv | head -d
```

# Other web-based softwares

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM



The image shows the homepage of the UALCAN cancer data analysis portal. At the top, there is a navigation bar with links for Home, Tutorial, TCGA, Proteomics, CBTN, and HEERSINK. Below the navigation bar, the text "Welcome to UALCAN" is displayed, followed by "Yes! You All Can". A subtext below reads "The University of Alabama at Birmingham CANcer data analysis Portal". The main content area describes UALCAN as a comprehensive, user-friendly, and interactive web resource for analyzing cancer OMICS data. It lists several features: a) easy access to publicly available cancer OMICS data (TCGA, MET500, CPTAC and CBTTC), b) biomarker identification and in silico validation, c) expression profile and patient survival information, d) epigenetic regulation analysis, e) pan-cancer gene expression analysis, f) gene-target linking, g) clinical proteomic consortium data analysis, and h) pediatric brain tumor gene expression analysis. Below this text is a "Follow us on Twitter" link with a small Twitter icon. A section titled "Please cite:" provides two references from the journal *Neoplasia*. The first reference is for an update to the integrated cancer data analysis platform, published in March 2022, with a PMID of 35078134. The second reference is for a portal facilitating tumor subgroup gene expression and survival analyses, published in August 2017, with a PMID of 28732212. At the bottom of the page, a green banner states "Updated slides from a presentation about UALCAN cancer database at Dana-Farber Cancer".

<https://ualcan.path.uab.edu/>

# Day 2 – TCGAbiolinks package

# Two examples of using TCGAbiolinks

1. Build system environment
2. Build a **TCGA-READ** cohort, retrieve  
**somatic mutation** datasets, load into R
3. Using **maftools** package to analyze  
mutational datasets
4. Query, retrieve and prepare all **TCGA-READ**  
**mRNA transcription** datasets.
5. Do differential expression analysis using  
**DESeq2**.

e.g.1

e.g.2

# Data mining by R package TCGAbiolinks



- [News](#)
- [Citation](#)
- [Other useful links](#)
- [Installation](#)
- [Question and issues](#)
- [Required libraries](#)
- [Session info](#)

## Introduction

6 June 2023

TCGAbiolinks is able to access The National Cancer Institute (NCI) Genomic Data Commons (GDC) thorough its [GDC Application Programming Interface \(API\)](#) to search, download and prepare relevant data for analysis in R.

## News

- April 2022:
  - Started to add support for GENCODE v36 pipelines
  - Add stemness score functions
- December 2019:
  - Added support to non TCGA/TARGET projects - [https://rpubs.com/tiagochst/TCGAbiolinks\\_RNA-seq\\_new\\_projects](https://rpubs.com/tiagochst/TCGAbiolinks_RNA-seq_new_projects)
  - Added support to linked Omics data retrieval - <https://rpubs.com/tiagochst/linkedOmics>
  - Included Glioma classifier from GUI to the main package - <https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/classifiers.html>
- Added support to BCR Biotab clinical files
- Workshop materials (ATAC-seq, ELMER and TCGAbiolinks) were added to vignette (<https://bioconductor.org/packages/devel/bioc/vignettes/TCGAbiolinks/inst/doc/index.html>)

## Citation

If you use TCGAbiolinks, please cite:

- Colaprico, Antonio, et al. "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data." Nucleic acids research 44.8 (2015): e71-e71.
- Silva, Tiago C., et al. "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages."

<https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/index.html>

# TCGAbiolinks 3 Main Functions

- **GDCquery**: This function serves as a protocol to interact with the Genomic Data Commons (GDC) API. It allows users to specify desired parameters and identify appropriate datasets hosted within the GDC repository. These parameters could include project identifiers, data categories, workflow types, and other metadata details, facilitating a streamlined and targeted data retrieval process.
- **GDCdownload**: Following the successful creation of a query using GDCquery, GDCdownload is the function responsible for initiating and managing the data transfer process from the GDC servers. It downloads the specified data associated with the prior query, managing potential issues such as data volume and network interruptions, and storing the data locally for further processing and analysis.
- **GDCprepare**: This function takes the locally downloaded data and transforms it into a structured format suitable for downstream analyses. It performs necessary preprocessing tasks such as data normalization, error handling, and data type conversion, ultimately generating a coherent data object that can be easily manipulated using conventional bioinformatic tools and pipelines.

# Installation for TCGAbiolinks package

```
# Install necessary packages
if(!requireNamespace('BiocManager', quietly = TRUE))
  install.packages("BiocManager")

library('BiocManager')

BiocManager::install("BioinformaticsFMRP/TCGAbiolinksGUI.data")
BiocManager::install("BioinformaticsFMRP/TCGAbiolinks")
BiocManager::install('SummarizedExperiment')

BiocManager::install('maftools')

BiocManager::install('DESeq2')
BiocManager::install('EnhancedVolcano')
BiocManager::install('umap')

BiocManager::install('survival')
BiocManager::install('survminer')
```

# Ask ChatGPT!

ChatGPT 4 ▾

programming languages make it a valuable tool for a wide range of applications in data storage, transmission, and configuration.



You

```
any(colnames(clinical_query) %in% c('vital_status','days_to_last_follow_up','days_to_death'))  
I want to check if all in colnames instead of any, which function should I use
```



ChatGPT

To check if all the elements of a vector are present in another vector in R, you can use the `all()` function combined with the `%in%` operator. In your case, if you want to verify whether all of the specified column names ('vital\_status', 'days\_to\_last\_follow\_up', 'days\_to\_death') are present in the column names of the `clinical\_query` data frame, you would write the code like this:

R

Copy code

```
all(c('vital_status', 'days_to_last_follow_up', 'days_to_death') %in% colnames(clinic
```

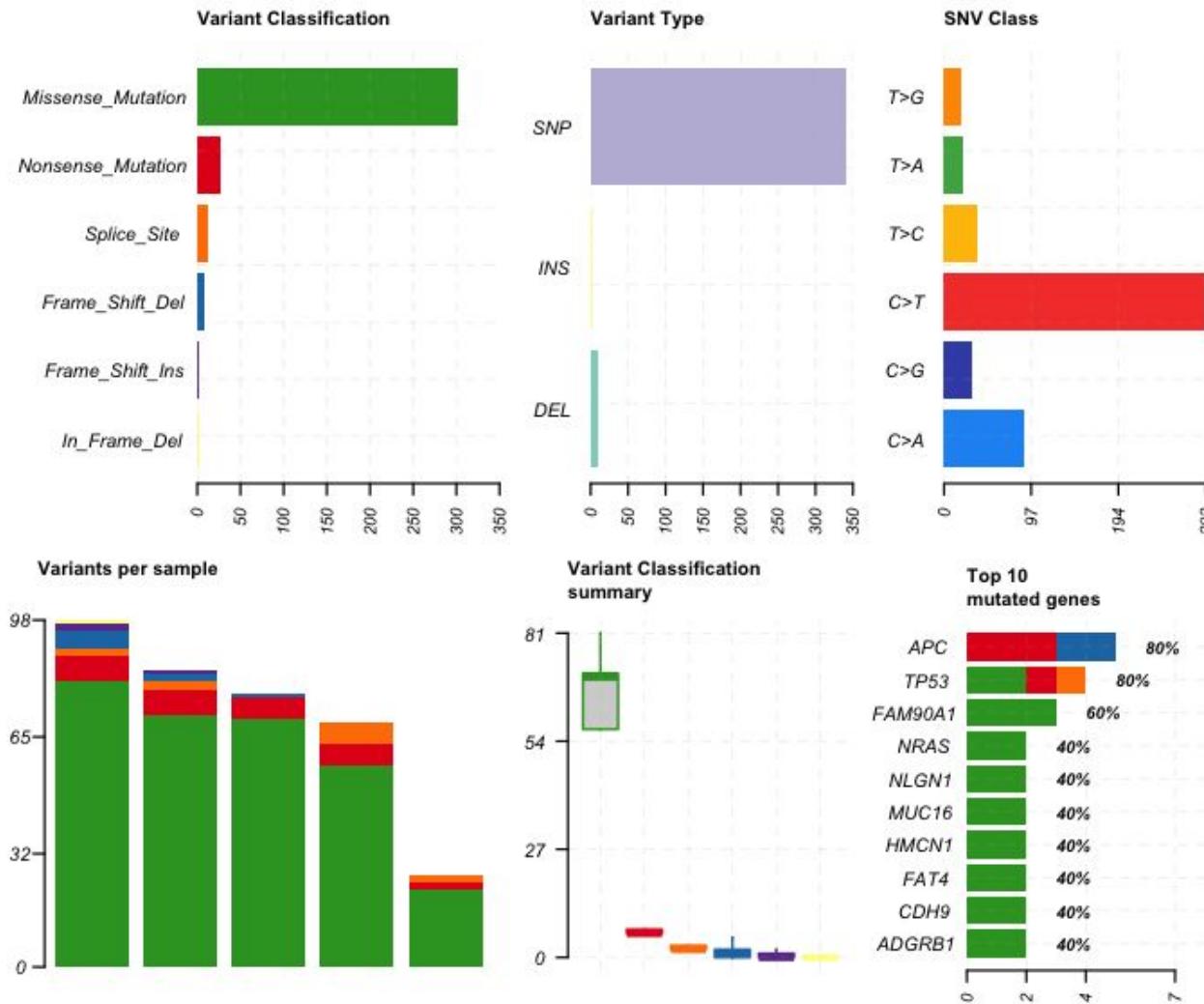
This code works as follows:

- `c('vital\_status', 'days\_to\_last\_follow\_up', 'days\_to\_death')`: Creates a vector containing the column names you want to check.
- `%in% colnames(clinical\_query)` : This checks each element of the vector against the column names of `clinical\_query`. It returns a logical vector where each element is `TRUE` if the corresponding column name is present in `clinical\_query`, and `FALSE` otherwise.
- `all(...)` : Checks if all elements of the resulting logical vector are `TRUE`. It returns `TRUE` only if all specified columns are present in `clinical\_query`, and `FALSE` if any are missing.

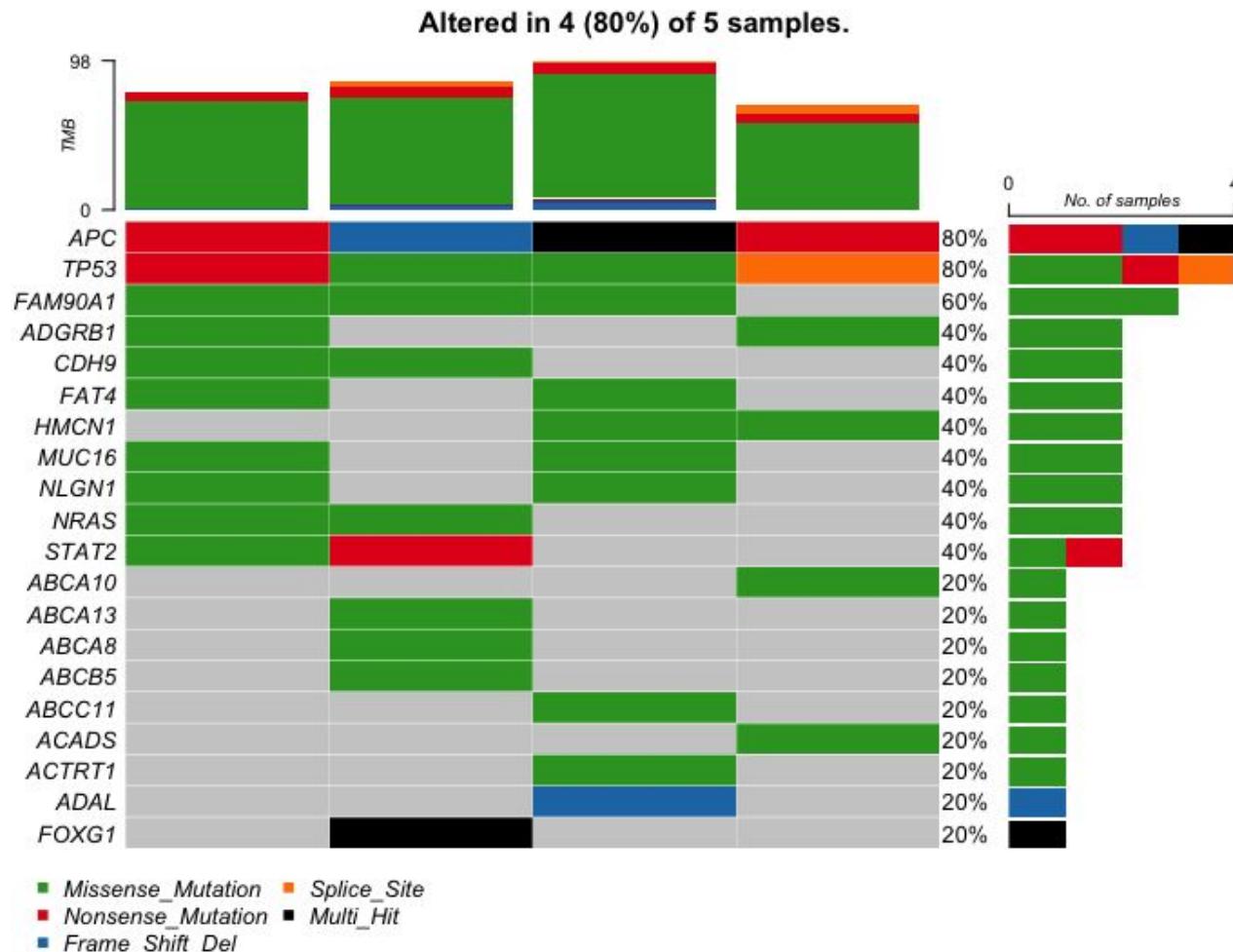
So, this line of code will return `TRUE` if and only if all of 'vital\_status', 'days\_to\_last\_follow\_up', and 'days\_to\_death' are column names in the `clinical\_query` data frame.



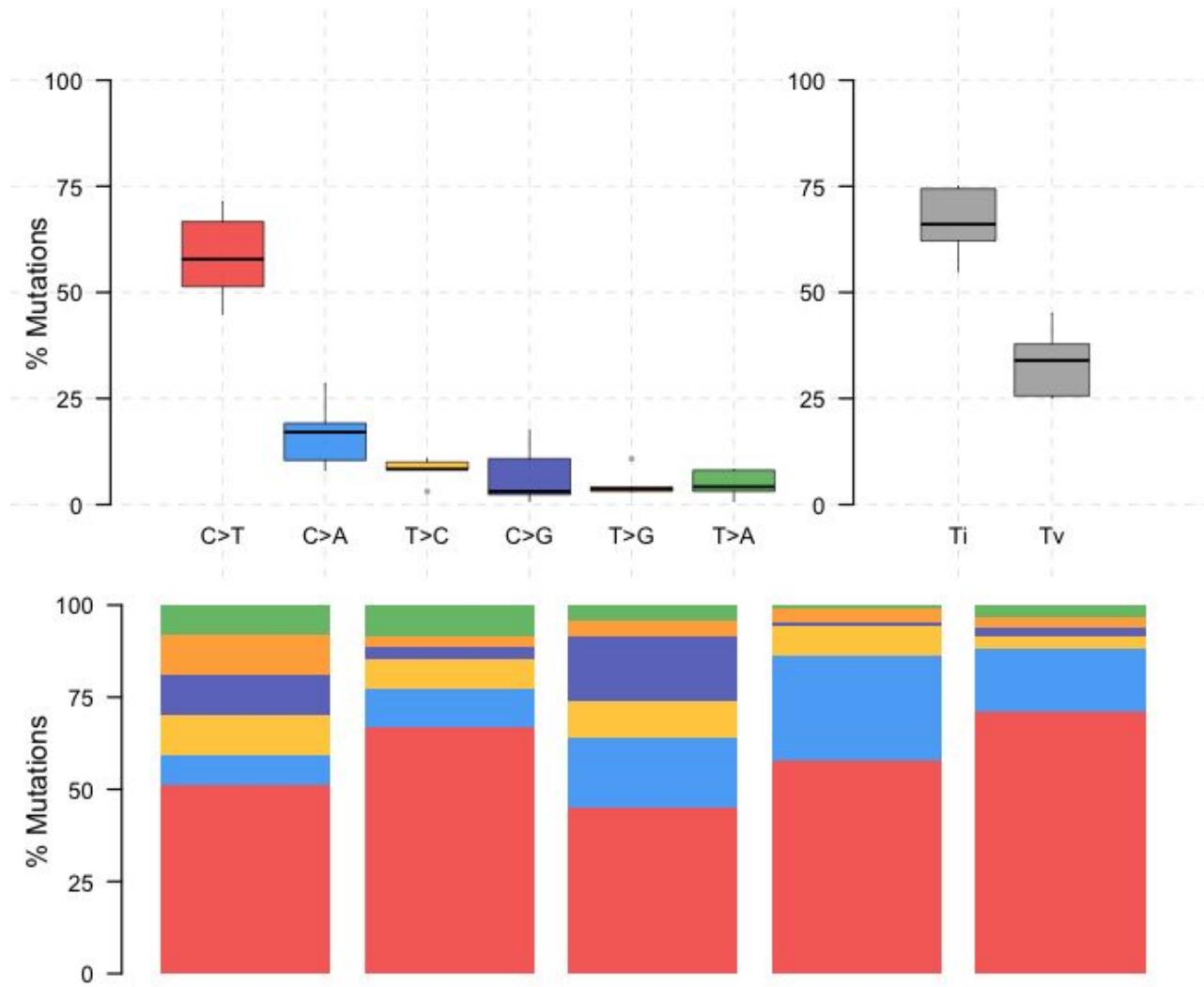
# Somatic Mutation Analysis (plotmafSummary)



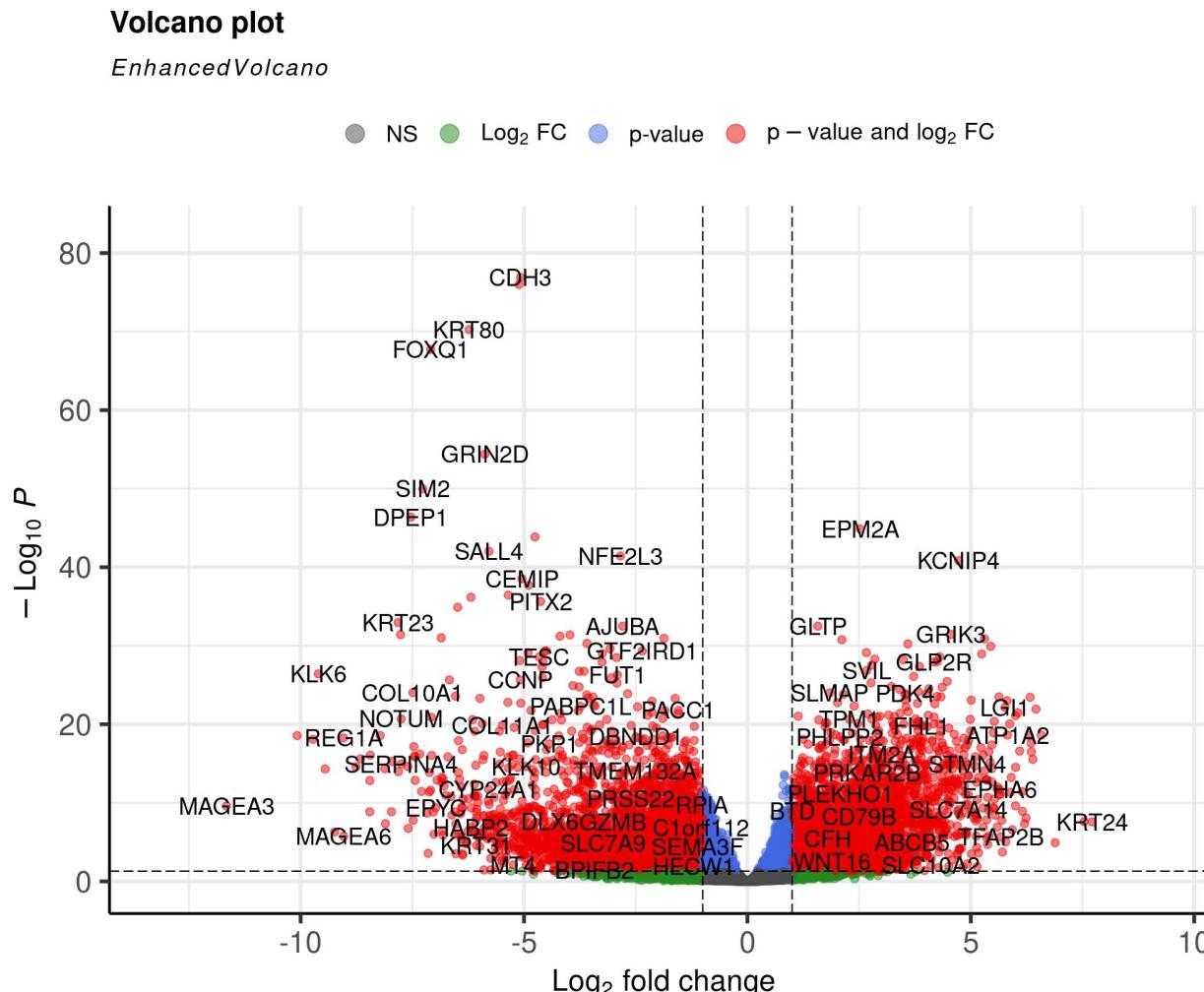
# Somatic Mutation Analysis (oncoplot)



# Somatic Mutation Analysis (plotTiTv)



# Differential mRNA Expression Analysis



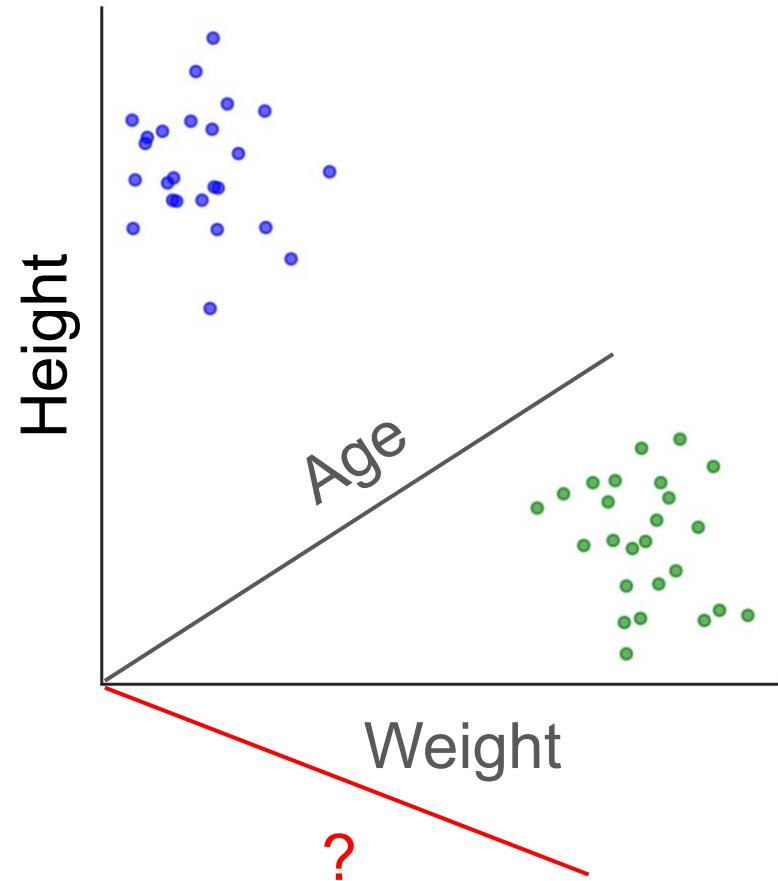
# Day 3 – Advancing Downstream Data Analysis

# Two examples of downstream analysis

- e.g.1 {
  - 1. Review DESeq analysis on all TCGA-READ cohort
  - 2. Build a downstream **UMAP projection** (pseudo-single cell) on TCGA-READ cohort
  
- e.g.2 {
  - 3. Retrieve and neat **clinical datasets**
  - 4. Do a survival analysis by **Kaplan-Meier Curve** on a select oncogene txn level

# Dimensionality Reduction Principle

Height	Weight
167	55
175	68
195	82
...	...



# Introduction of Dimensionality Reduction Techniques

- **Principal Component Analysis (PCA)**: PCA is a technique for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables (Principal Components) that successively maximize variance. The idea is to find the directions (or vectors) in the high-dimensional space along which the original data is highly variable. These new vectors represent the principal axes, and the data is projected onto these axes to reduce its dimensionality.
  -
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: t-SNE is a technique for dimensionality reduction that is particularly well-suited for the visualization of high-dimensional datasets. It uses a probabilistic approach: it computes probabilities that translate the relationships (similarities) between points in a high-dimensional space and then seeks to minimize the divergence between these probabilities and the ones computed for the points in the low-dimensional space. Unlike PCA, t-SNE is not a linear projection; it uses a non-linear approach that tends to maintain the structure at different scales.
  -
- **Uniform Manifold Approximation and Projection (UMAP)**: UMAP is another dimension reduction technique that operates in a manner somewhat similar to t-SNE, but it starts with a different mathematical framework, based on Riemannian geometry and algebraic topology. UMAP tends to preserve more of the global structure than t-SNE, making it more suitable for general non-linear dimension reduction tasks, not just visualization. The technique aims to capture both local and global aspects of the data in a low-dimensional space.

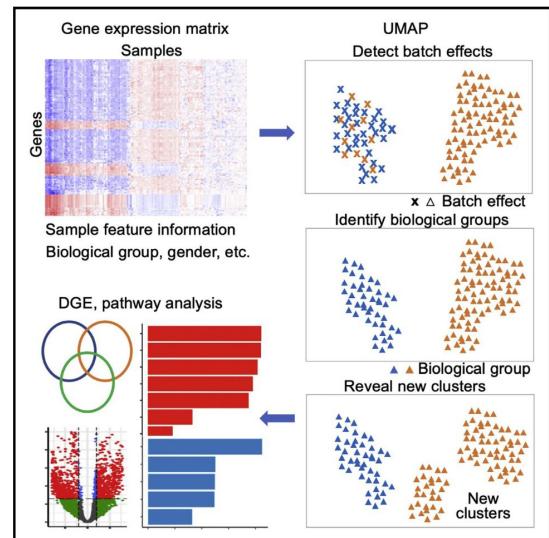
<https://www.youtube.com/watch?v=eN0wFzBA4Sc>

# Previous Research: Utilizing UMAP for TCGA Data Analysis

## Cell Reports

### Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data

#### Graphical abstract



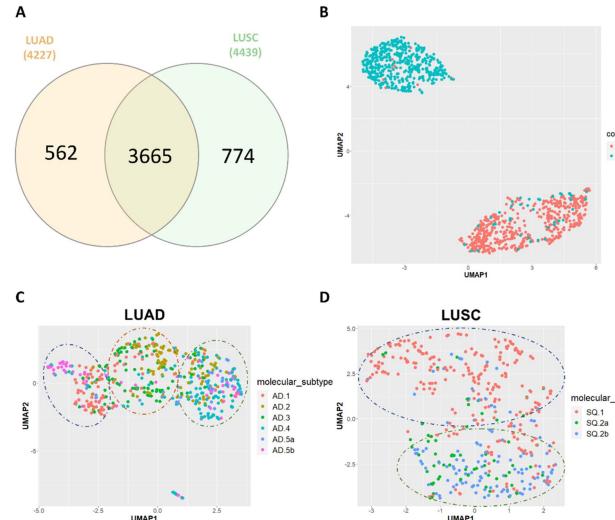
## Resource



#### Article

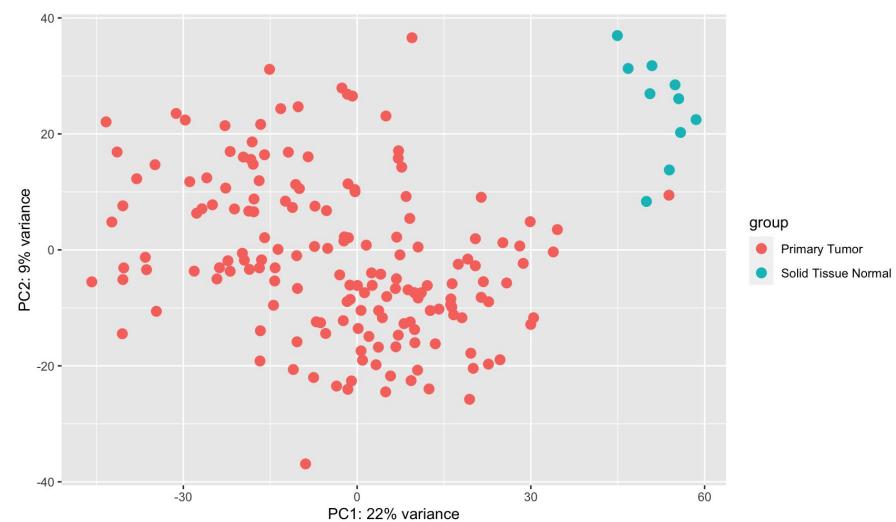
### Transcriptome Analysis of Human Endogenous Retroviruses at Locus-Specific Resolution in Non-Small Cell Lung Cancer

Alessandro La Ferlita <sup>1,\*†</sup>, Rosario Distefano <sup>1,†</sup>, Salvatore Alaimo <sup>2</sup>, Joal D. Beane <sup>3</sup>, Alfredo Ferro <sup>2</sup>, Carlo M. Croce <sup>1</sup>, Philip N. Tsichlis <sup>1</sup>, Alfredo Pulvirenti <sup>2,‡</sup> and Giovanni Nigita <sup>1,\*‡</sup>

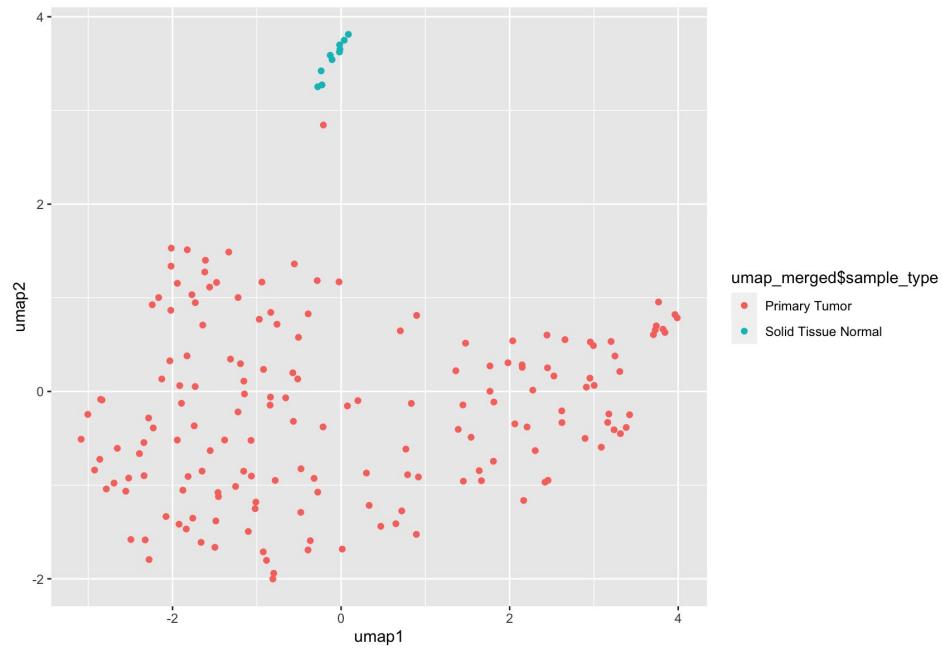


# TCGA compare between PCA with UMAP for same batch

PCA

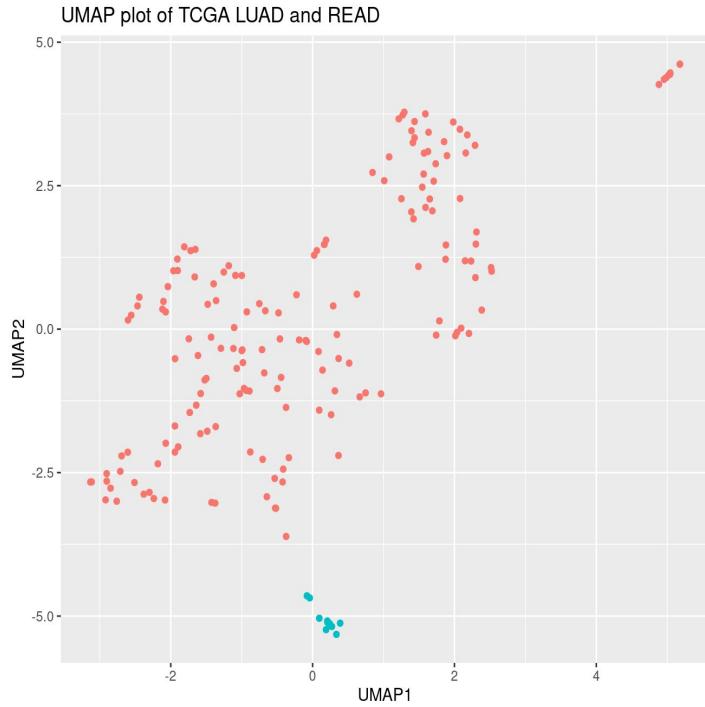


UMAP

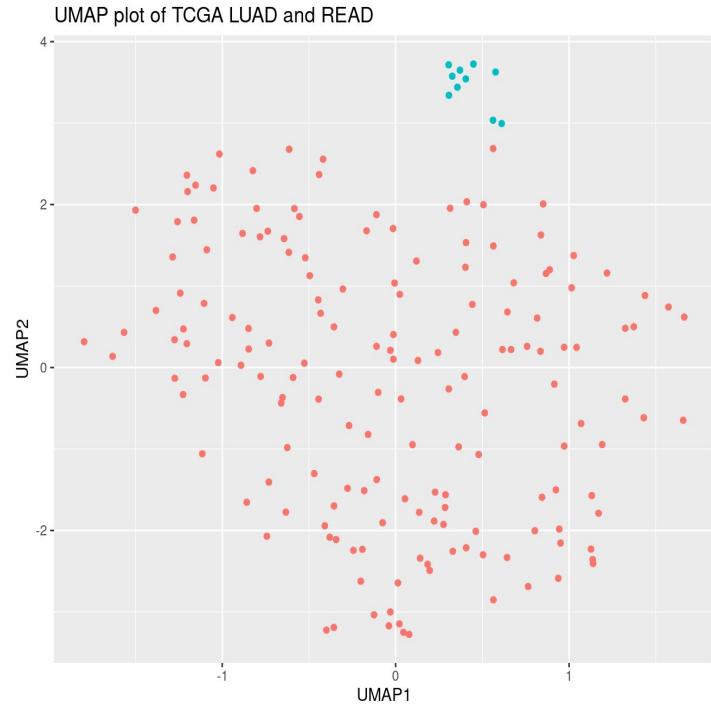


# Plotting UMAP on the TCGA-READ project

# UMAP clustering Practice: TCGA-READ



n\_neighbor=5

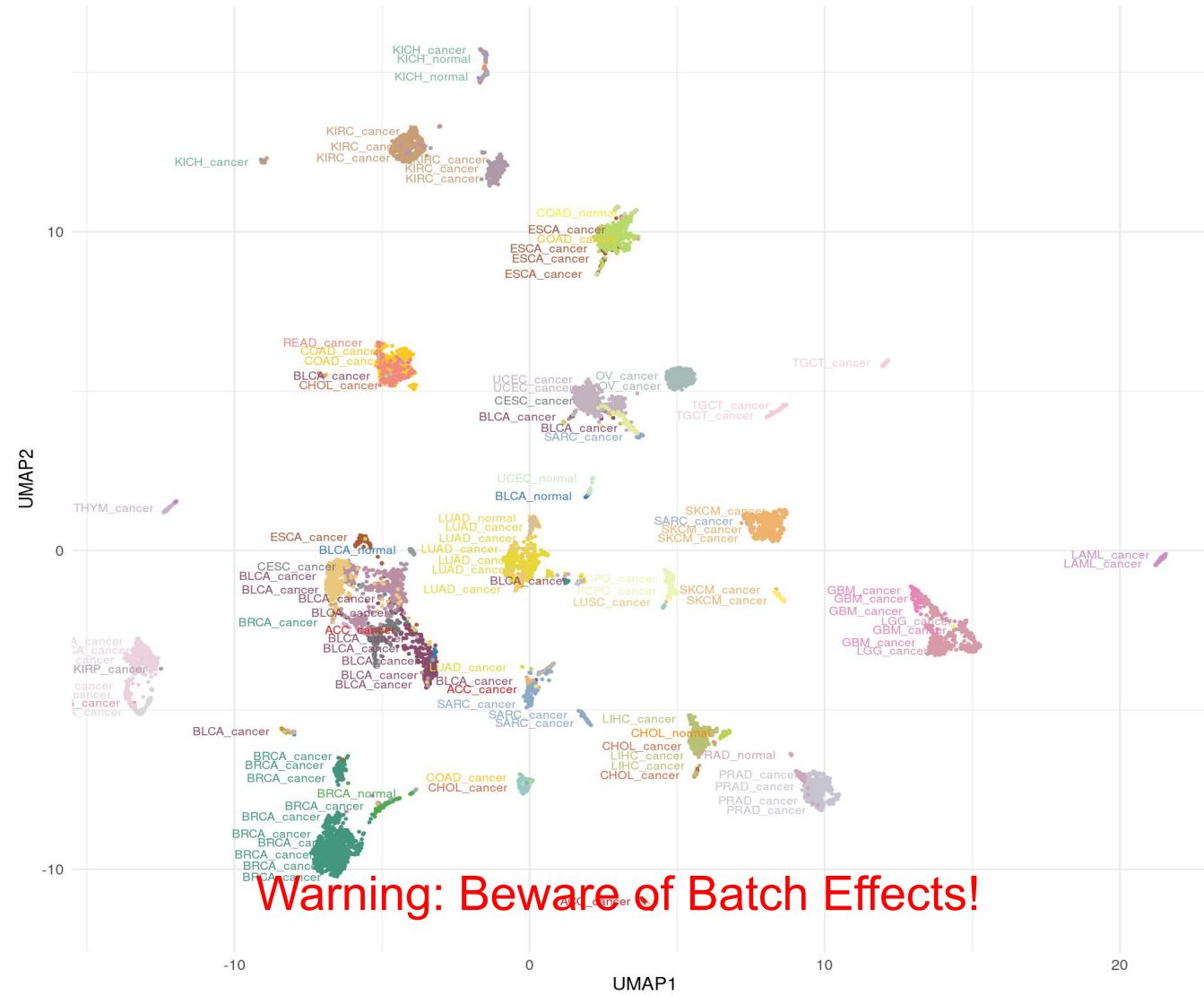


n\_neighbor=50

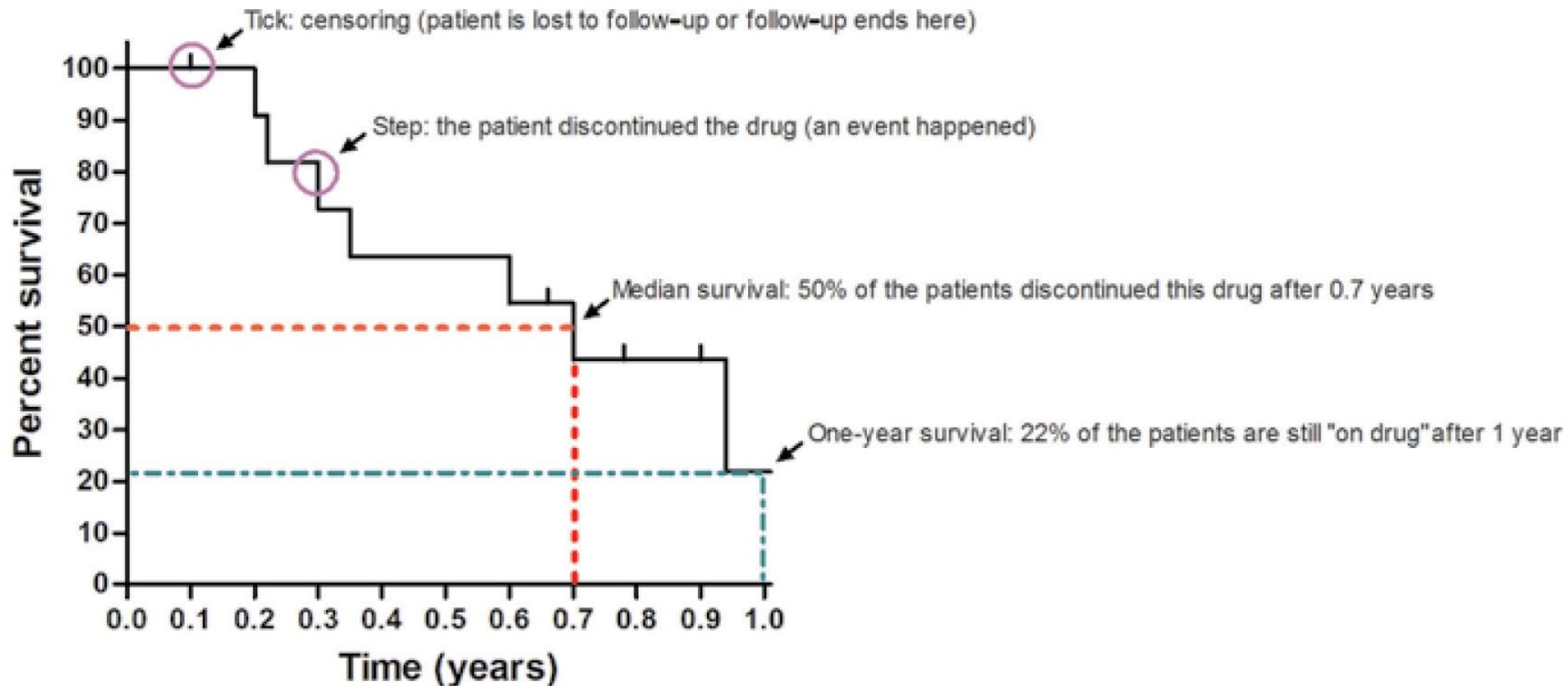
N\_neighbor: Nearest neighbor  
Large number : capture global patterns  
Small number : focus on finer details

# All TCGA samples (33 cancer types, n=10986)

## UMAP plot of all TCGA cancer



# Survival Analysis By Kaplan-Meier Curve

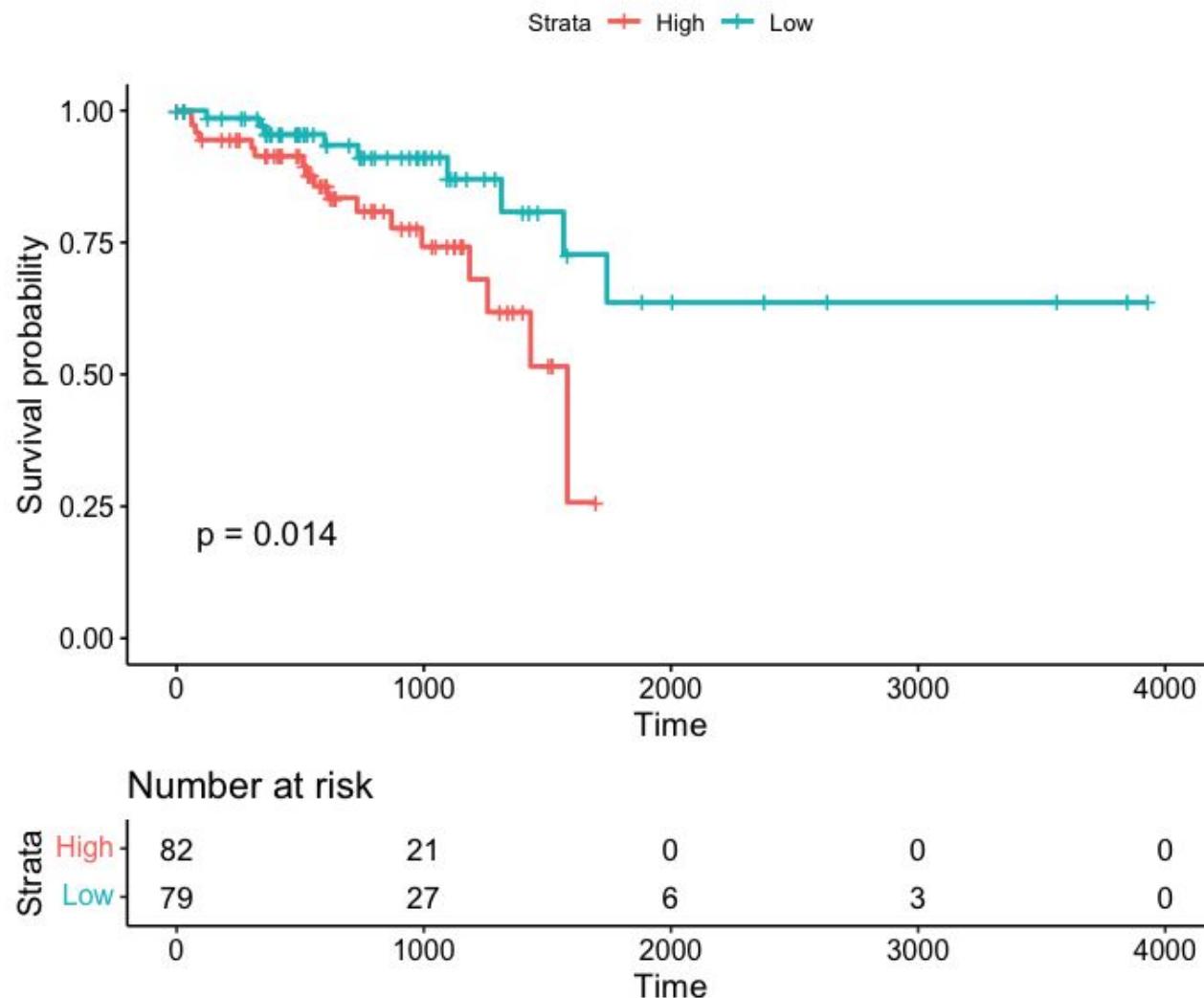


# Concepts for survival analysis

**Event:** An "event" in survival analysis is the occurrence of the specific outcome of interest (e.g., death or disease recurrence) during the study period.

**Censoring:** "Censoring" in survival analysis refers to incomplete or truncated follow-up data, where the exact time of the event for some participants is unknown because they have not experienced the event by the end of the study.

# Kaplan-Meier Curve of CDKN1A



Thank you!