QUEEN'S UNIVERSITY BELFAST

SCHOOL OF
ELECTRONICS,
ELECTRICAL
ENGNIEERING AND
COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 – 2019/2020

The dataset of the individual coursework 1 (ICW1)

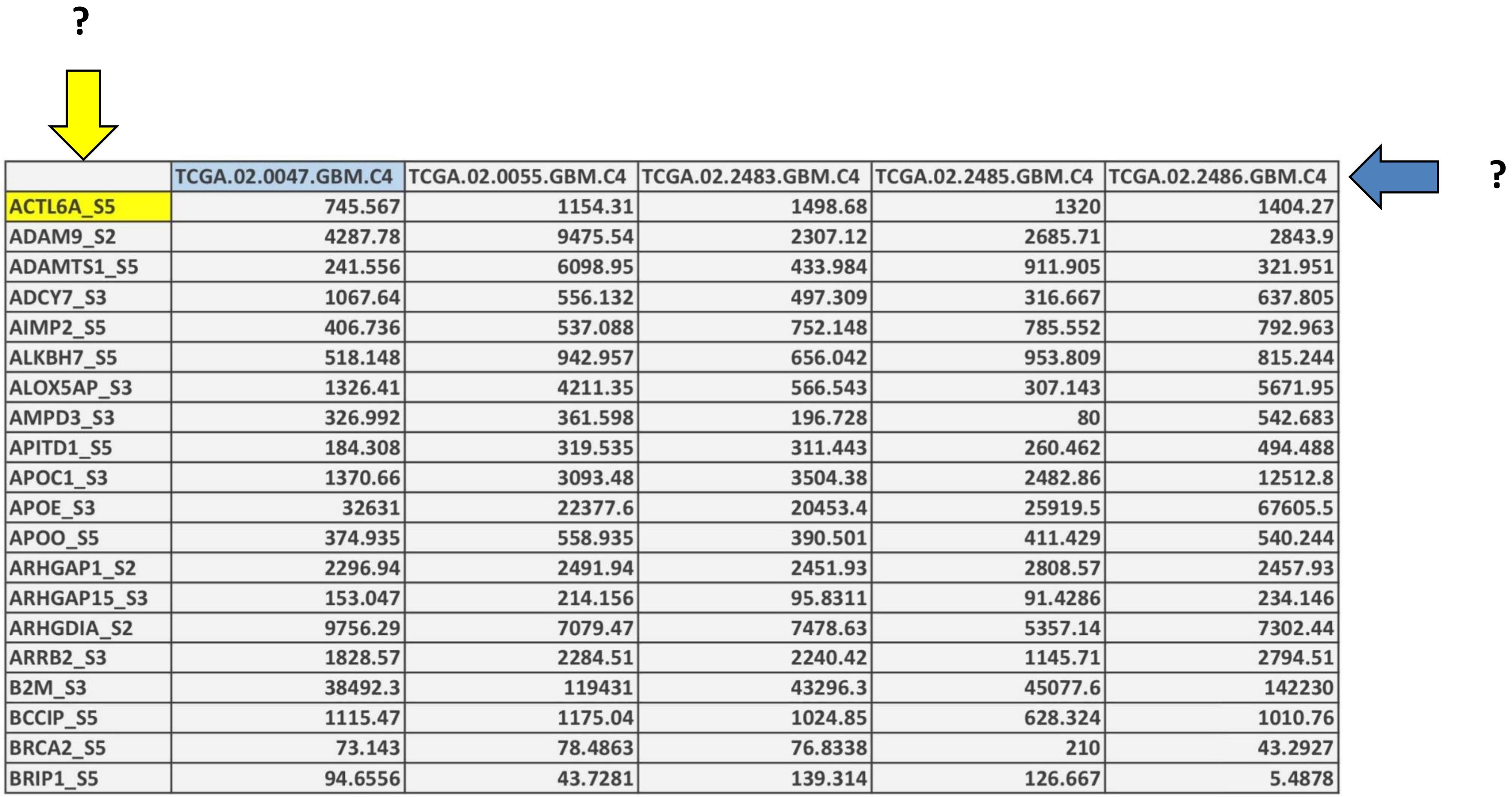
You are required to use a **biological dataset** for some of the practical assignments and all individual courseworks. This type of dataset can be originally downloaded from TCGA network[1], however, you could easily download it from a **GitHub** repository via the following link. The dataset was already compressed (7z file type) which you may simply unzip using WinZip or similar app.

https://github.com/RRafiee/Data-Analysis-and-Visualisation/blob/master/PanCanAtlas_9126RNASeqSamplesWithImmuneSubtypes_440Genes_SampleIdsOrdered_RR020718_RownamesGenesWithSignature.7z

The **password** to unzip this file is **CSC3062**.

[1] TCGA Network is a data repository and portal including biological information (i.e., molecular data)

Which one is a feature name (yellow or blue)?



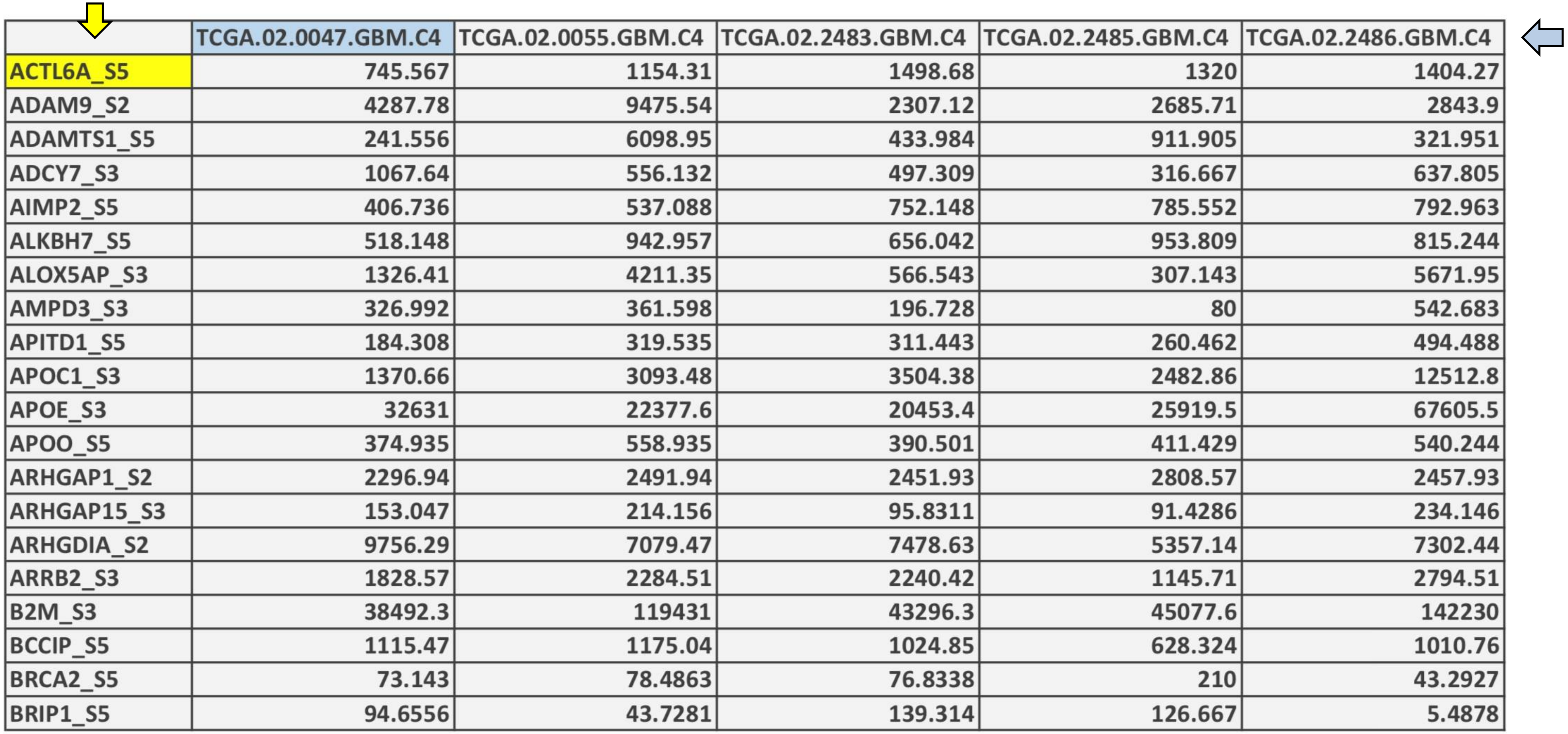
?

?

	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

The feature and sample in this dataset

The downloaded dataset, which is a tabular data (it's a csv file), has the following format:



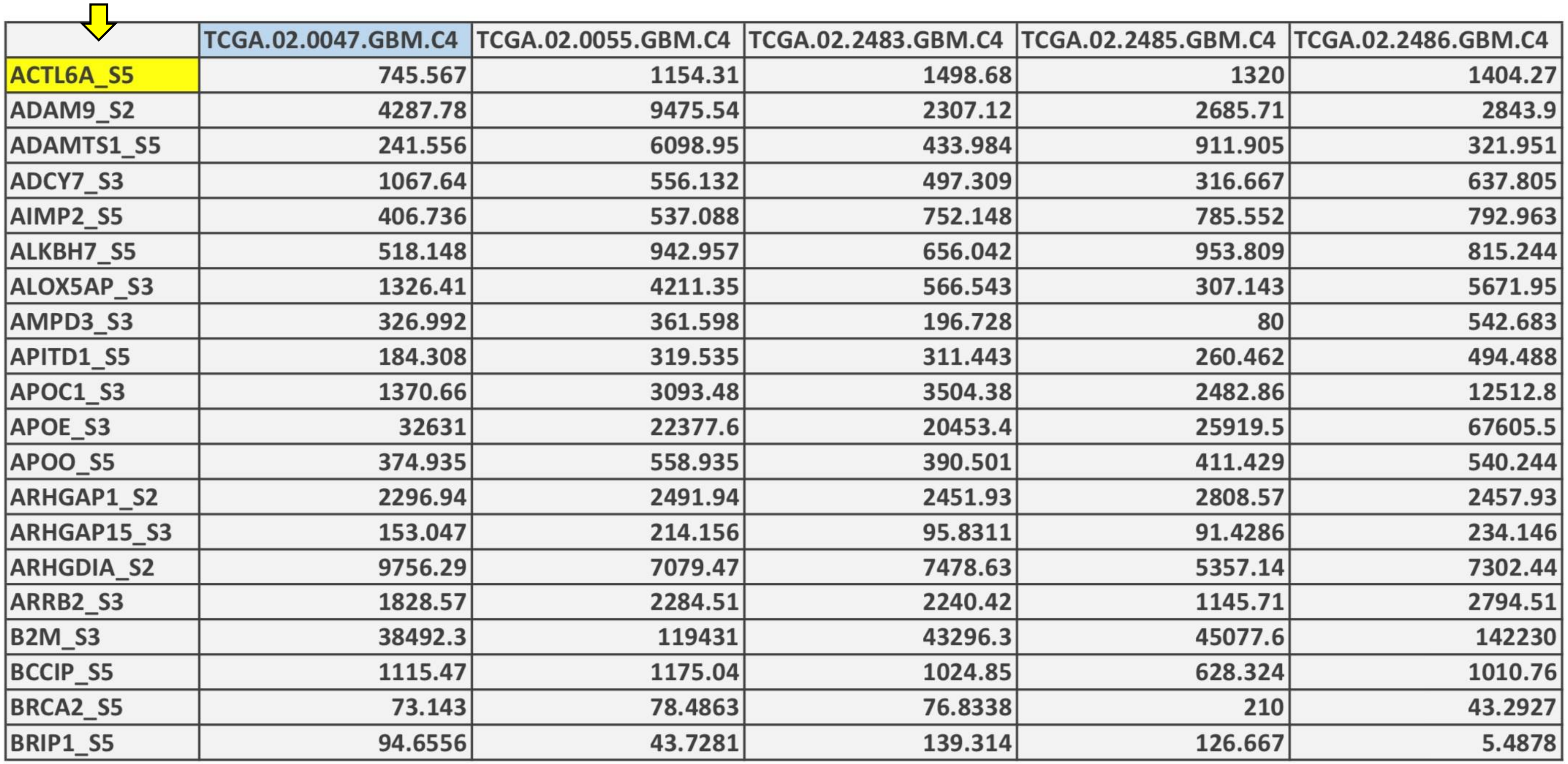
	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

Feature and sample in this dataset

The downloaded dataset, which is a tabular data (it's a csv file), has the following format:

Features

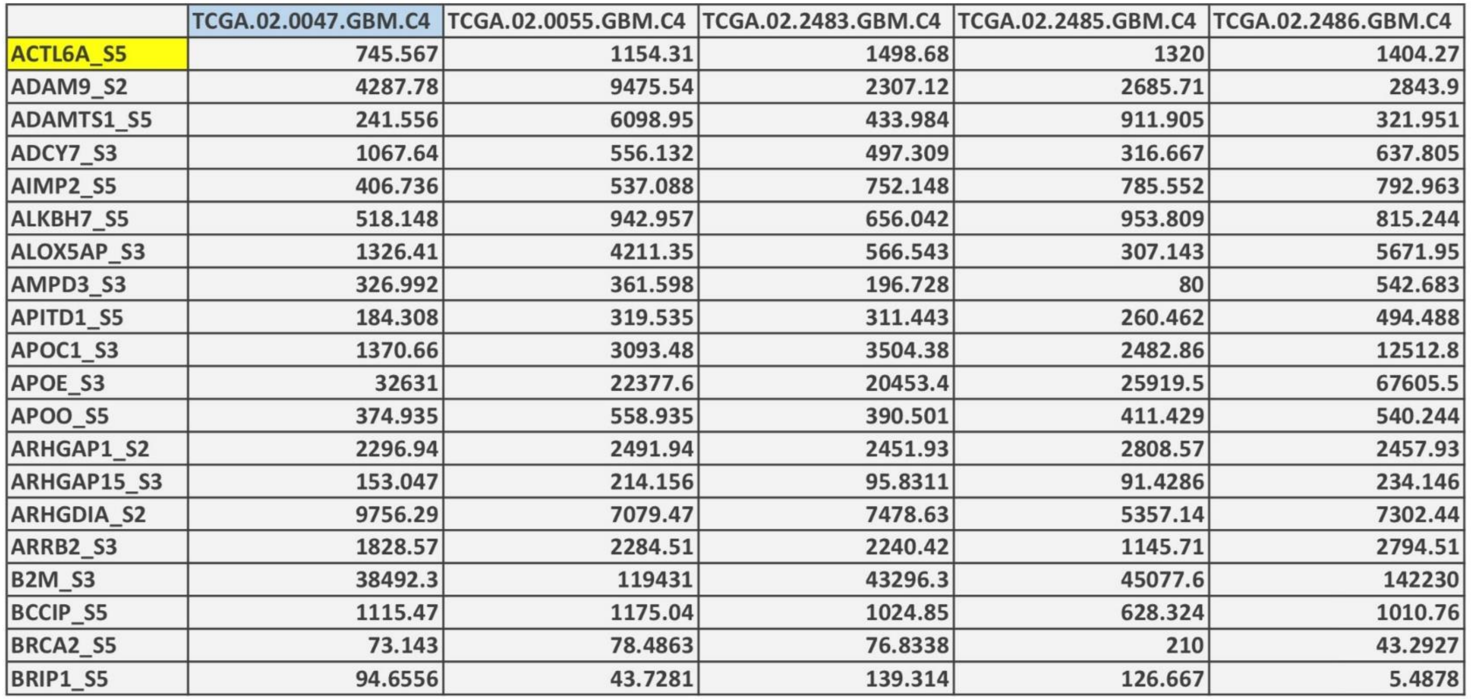
Samples



	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

Feature and sample in this dataset

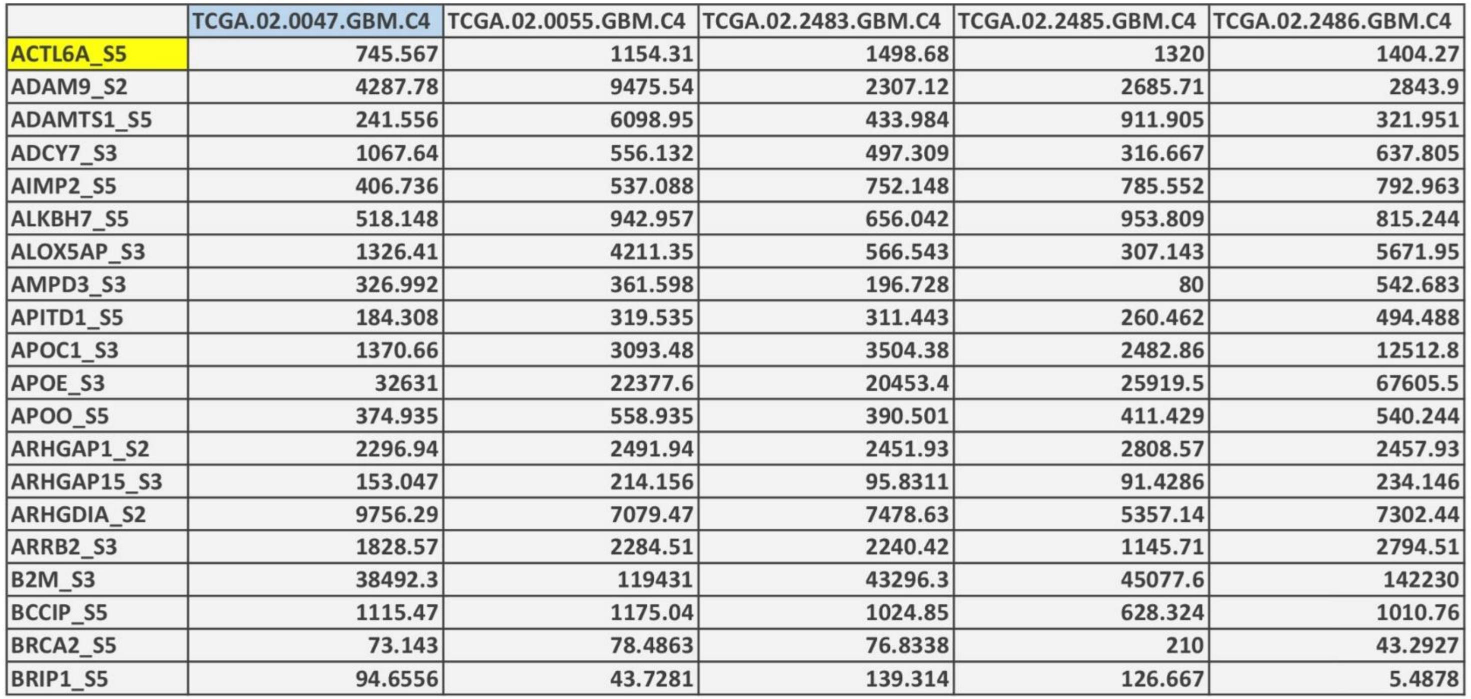
- There are **9126 cancer patients (which we call “samples” from now)**. Here, it illustrates only 5 samples and 20 features.
- Each individual sample has **440 “features”**.
- A row in this dataset represents the values of a **feature** across all samples.



	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

Feature and sample in this dataset

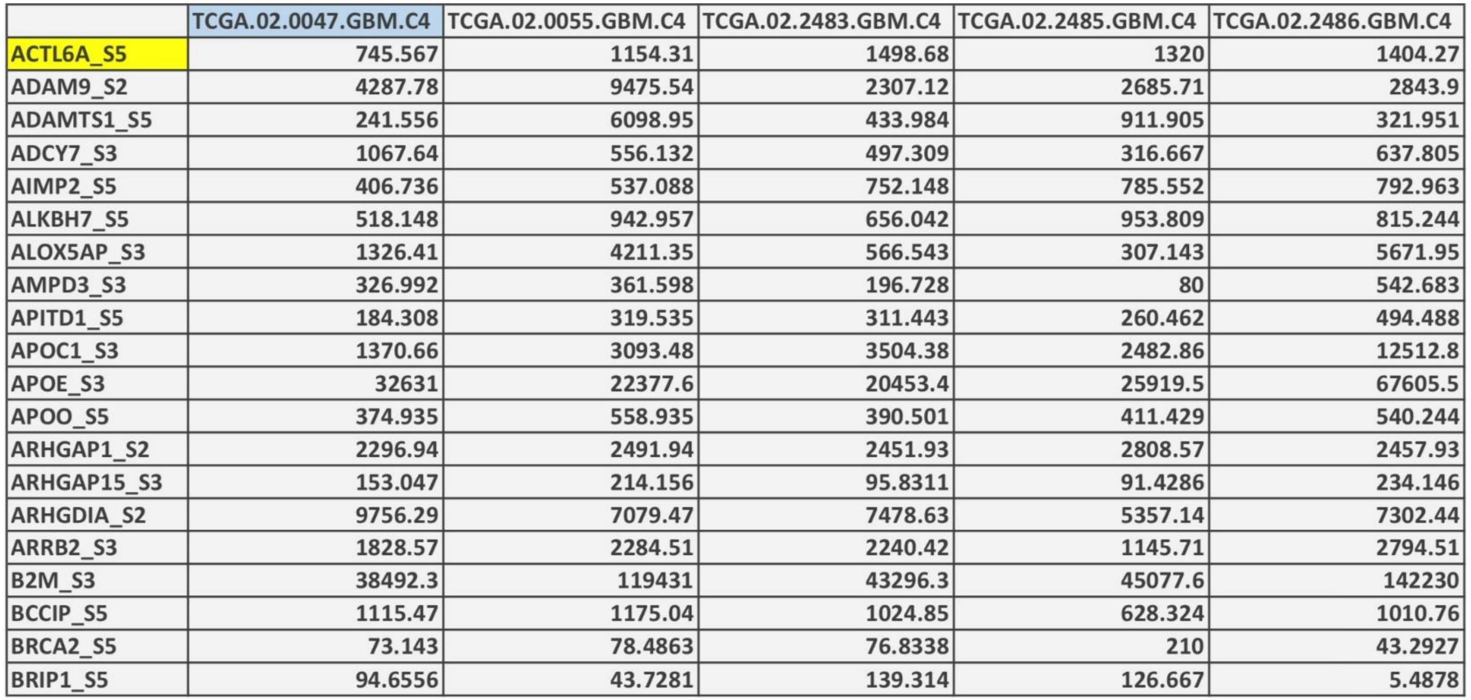
- **Feature name** (e.g., the yellow cell) comprising a “gene name” (e.g., ***ACTL6A***), underscore character (“_”) and a “signature group” (e.g., **S5**).
- A column in this dataset represents all features’ values for a one specific sample.



	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

Feature and sample in this dataset

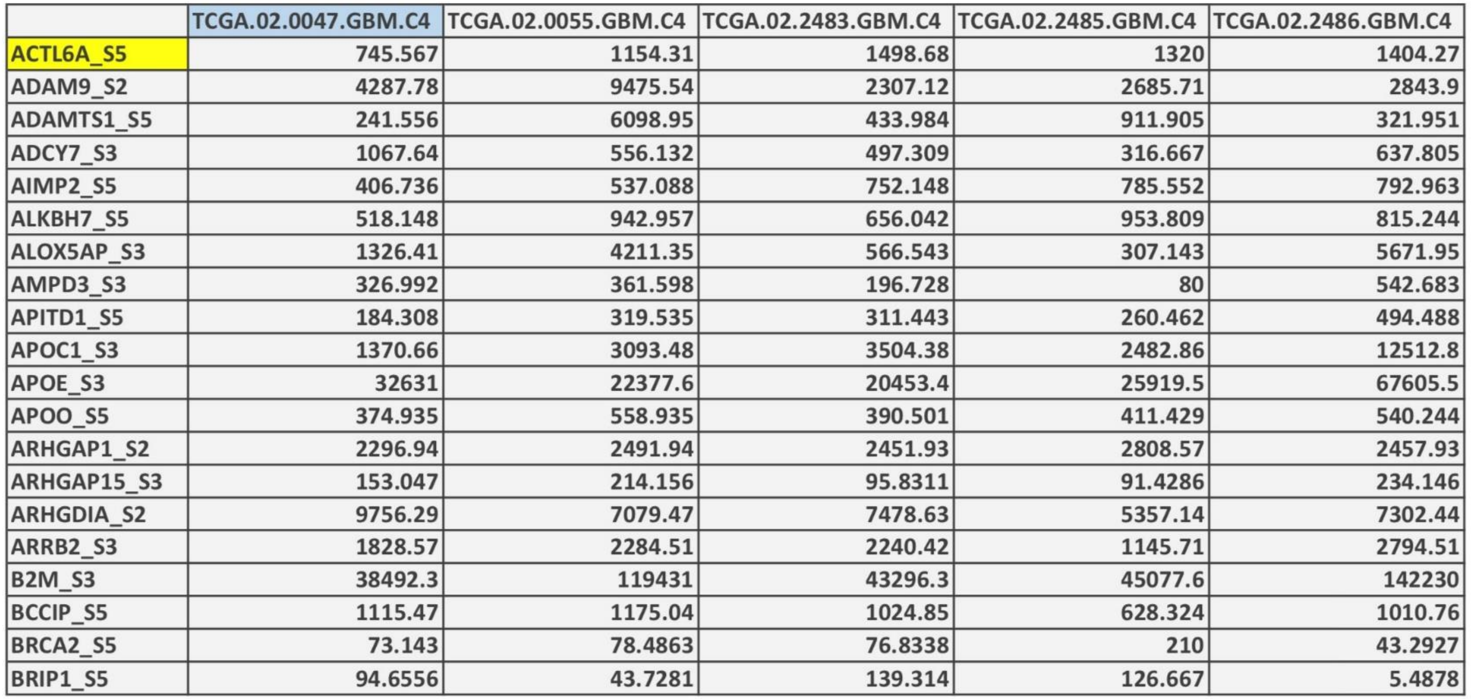
- **Sample name** (e.g., the blue cell) comprising a “sample id” (e.g., **TCGA.02.0047**), a “cancer name” (e.g., **GBM**) and an “immune subtype” (e.g., **C4**) which all are attached together using a “**.**”.
- As it has been illustrated in the following tabular figure, biological data in this dataset are represented by real numbers (e.g., 745.567). These data are known as RNA-Seq gene expression data.



	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

Feature and sample in this dataset

ACTL6A is a gene name (or gene id) and S5 is the signature group of this gene. **TCGA.02.0047** is called sample id which is a part of a TCGA barcode. There are 6 immune subgroups (i.e., C1, C2, C3, C4, C5 and C6) for all 9126 samples in this dataset[2].



	TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4
ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27
ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9
ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951
ADCY7_S3	1067.64	556.132	497.309	316.667	637.805
AIMP2_S5	406.736	537.088	752.148	785.552	792.963
ALKBH7_S5	518.148	942.957	656.042	953.809	815.244
ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95
AMPD3_S3	326.992	361.598	196.728	80	542.683
APITD1_S5	184.308	319.535	311.443	260.462	494.488
APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8
APOE_S3	32631	22377.6	20453.4	25919.5	67605.5
APOO_S5	374.935	558.935	390.501	411.429	540.244
ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93
ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146
ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44
ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51
B2M_S3	38492.3	119431	43296.3	45077.6	142230
BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76
BRCA2_S5	73.143	78.4863	76.8338	210	43.2927
BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878

[2] Thorsson *et al.*, 2018. **The Immune Landscape of Cancer**

Availability of ICW1 questions

ICW 1 will be available on Saturday (5th Oct.)