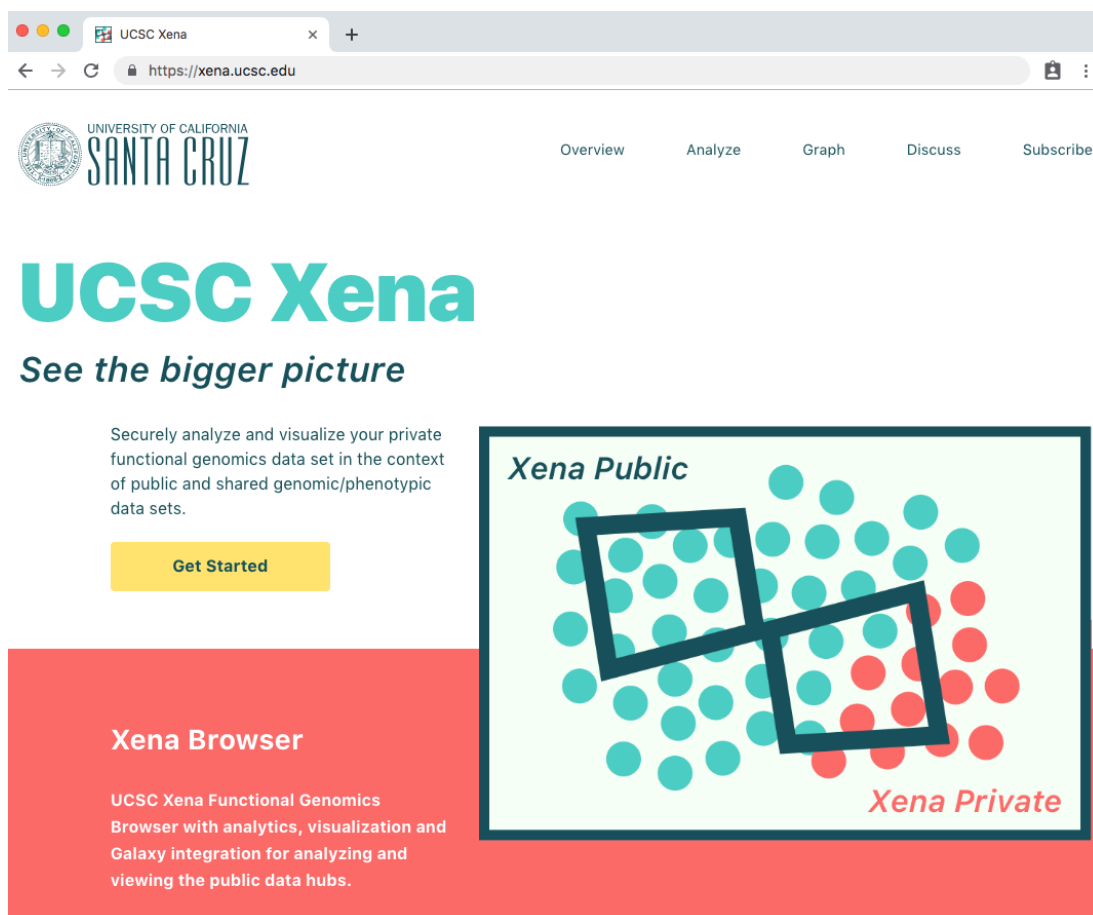


**Tutorial.** How to make a PATHOME-Drug input from TCGA stomach cancer dataset in users' own preferences.

**Task definition:** In the given G2 grade, one group is  $\leq 45$ -year old, and the other group is  $> 85$ -year old. Users make a proper input file by using Excel. The result is in the section "How to make a PATHOME-Drug input from TCGA gastric cancer dataset in users' own preferences" from <http://statgen.snu.ac.kr/software/pathome/?act=gcdatasets>.

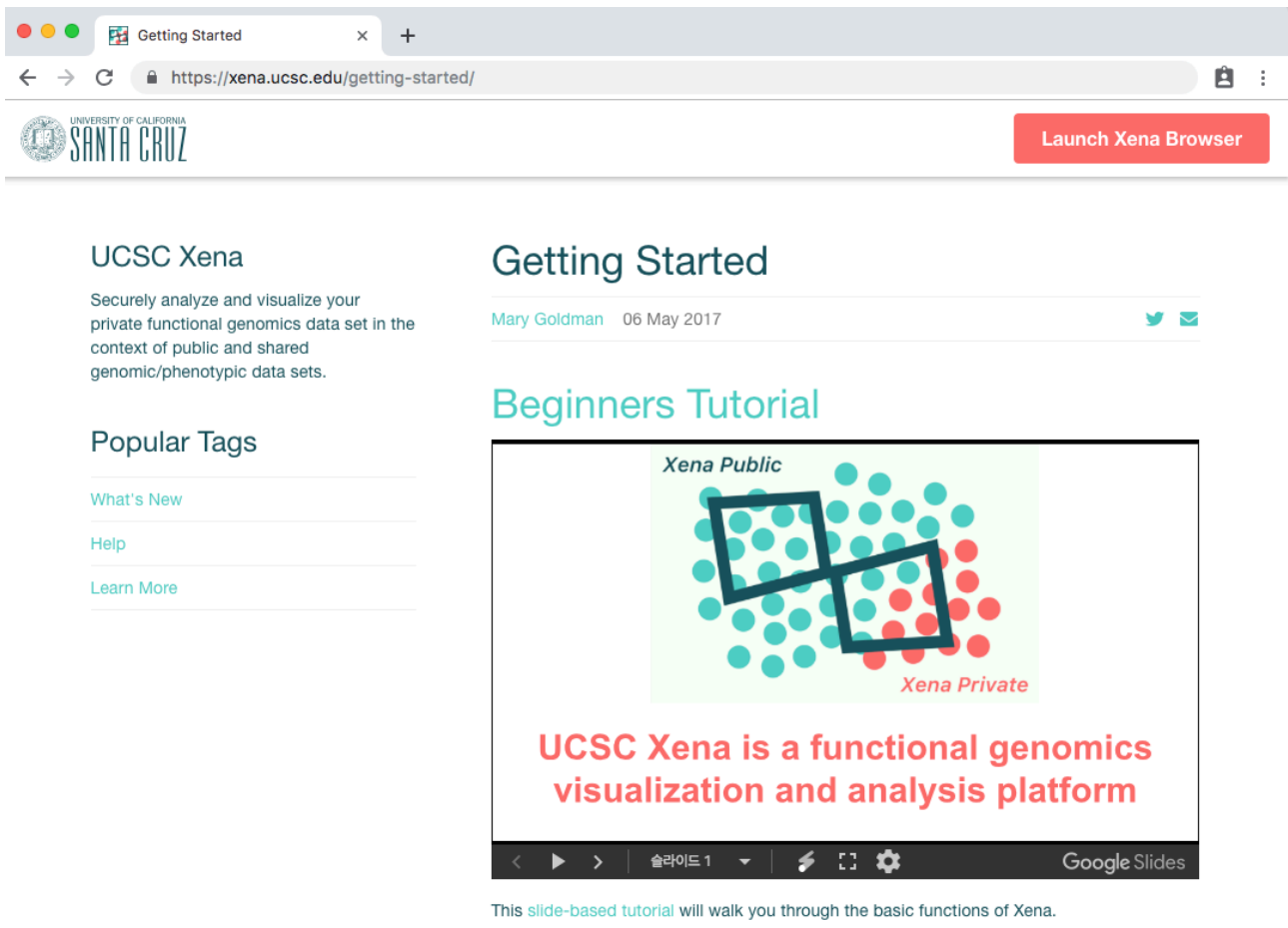
**STEP 1:** Download a compressed TCGA Stomach Adenocarcinoma RNA-Seq data file from the UCSC Xena Browser.

1. Visit the UCSC Xena Browser (<https://xena.ucsc.edu/>)



**Fig. SM1.** The UCSC Xena Browser (<https://xena.ucsc.edu/>).

2. Click “Get Started” button, then click “Launch Xena Browser” button at the top-right corner.



The screenshot shows a web browser window with the URL <https://xena.ucsc.edu/getting-started/>. The page header includes the UCSC Santa Cruz logo and a red button labeled "Launch Xena Browser". The main content area is titled "Getting Started" and features a "Beginners Tutorial" section. The tutorial is a slide-based presentation showing a diagram of data sets categorized into "Xena Public" (green dots) and "Xena Private" (red dots), with two overlapping black rectangles highlighting specific data points. Below the diagram, the text reads "UCSC Xena is a functional genomics visualization and analysis platform". The slide is presented in a Google Slides interface with navigation controls at the bottom.

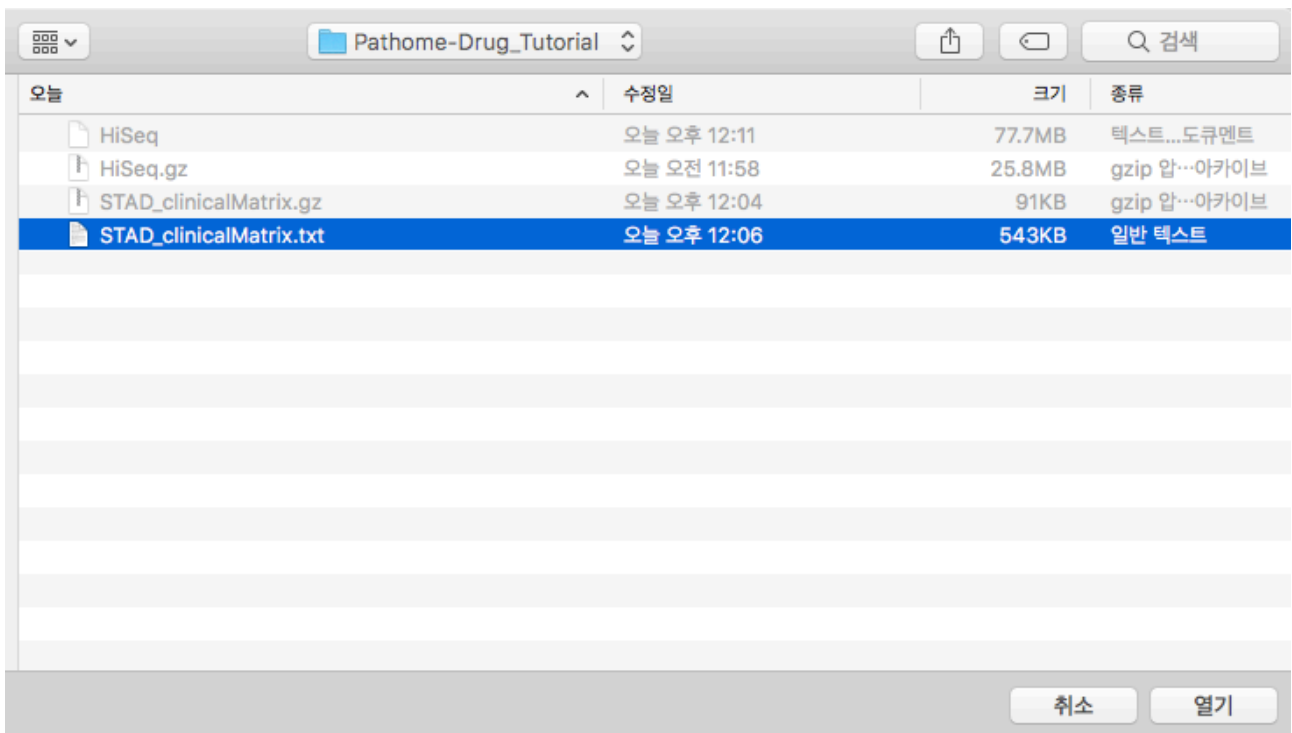
**Fig. SM2.** Launch UCSC Xena browser.

3. Click “DATA SETS” menu at the top of Xena browser, then find and click a link, “TCGA Stomach Cancer (STAD)” at the data page.
4. Click a link of “IlluminaHiSeq BC (n=417) TCGA hub” at the “gene expression RNAseq” section at the “cohort: TCGA Stomach Cancer (STAD)” page.
5. Click “<https://tcga.xenahubs.net/download/TCGA.STAD.sampleMap/HiSeq.gz>” at the “dataset: gene expression RNAseq - IlluminaHiSeq BC” page. In this tutorial we are going to show whole procedure with “TCGA.STAD.sampleMap/HiSeq” (Version 2017-10-13) dataset. When downloading the file is completed, decompress this file with a proper file compression program. (e.g. FSF & GNU zip)

6. Go back to the “cohort: TCGA Stomach Cancer (STAD)” page, click a link of “Phenotypes (n=580) TCGA hub” at the “phenotype” section.
7. Click [“https://tcga.xenahubs.net/download/TCGA.STAD.sampleMap/STAD\\_clinicalMatrix.gz”](https://tcga.xenahubs.net/download/TCGA.STAD.sampleMap/STAD_clinicalMatrix.gz) at the “dataset: phenotype – Phenotypes” page. When downloading the file is completed, decompress this file with a proper file compression program. (e.g. FSF & GNU zip)

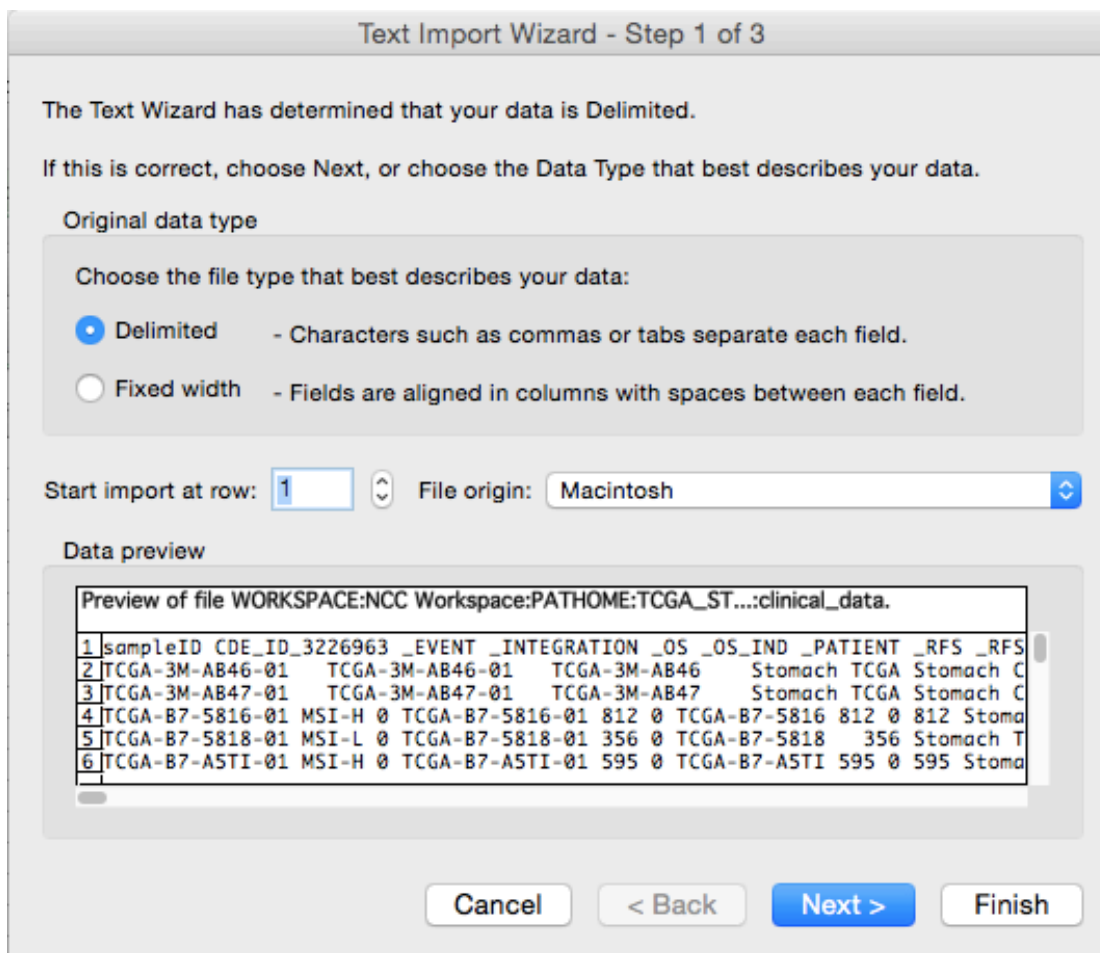
**STEP 2:** Import a phenotype data file into Excel and select data samples under specific criteria. In this tutorial our criteria are “neoplasm\_histologic\_grade” = “G2” and “age\_at\_initial\_pathologic\_diagnosis” <= 45 (control group, C1) or “age\_at\_initial\_pathologic\_diagnosis” > 85 (case group, C2).

8. Launch Microsoft™ Excel, click “File” >> “Open”. Move to the folder where you have decompressed files, and choose a file, “STAD\_clinicalMatrix”. Select an “All Files” item in the “Enable” or “File Format” drop-down box at the bottom of a dialog box. In Excel 2016 for Mac, “STAD\_clinicalMatrix” file is disabled in the file open dialog box, simply, add a text extension (.txt) to the “STAD\_clinicalMatrix” file, to “STAD\_clinicalMatrix.txt”.



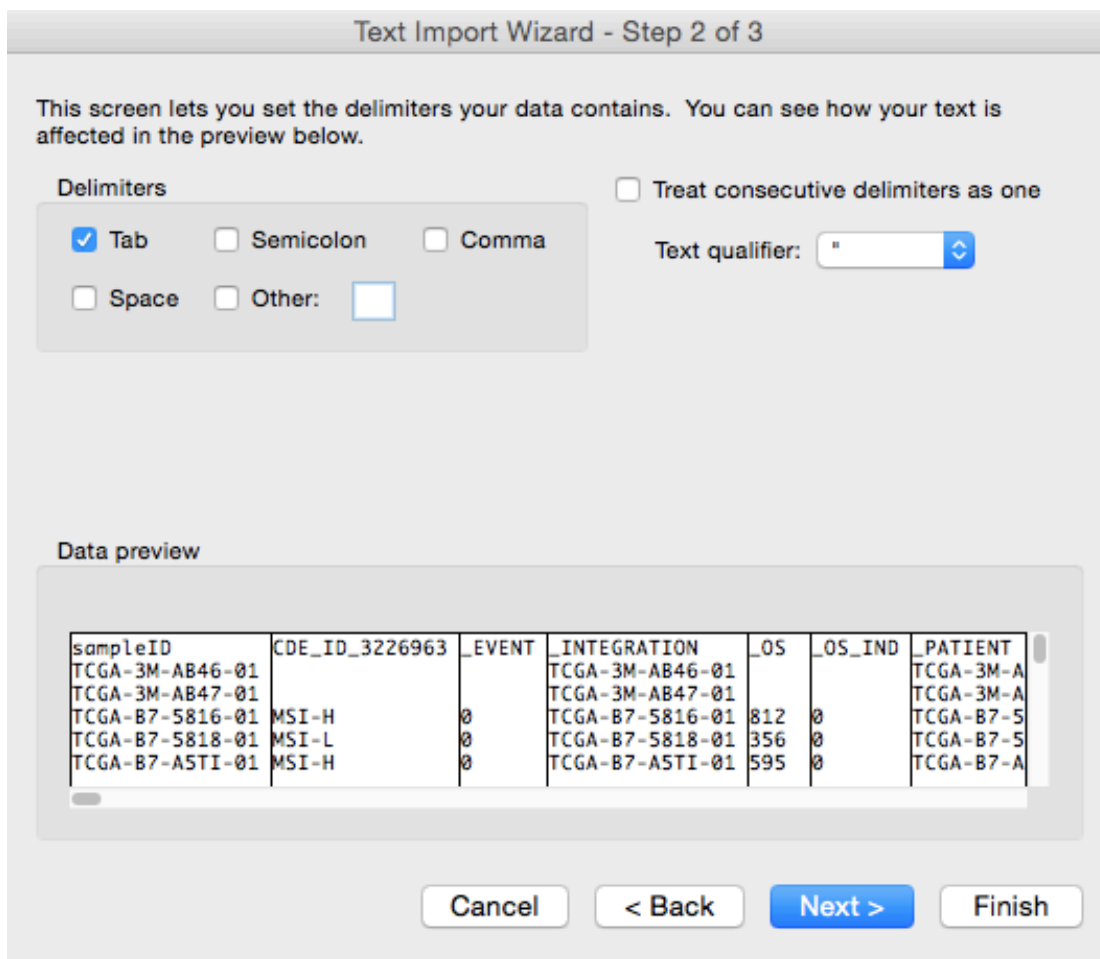
**Fig. SM3.** Open the "STAD\_clinicalMatrix" file from Excel.

9. On the "Text Import Wizard", click a "Delimited" at the Original data type panel then click a "Next" button.



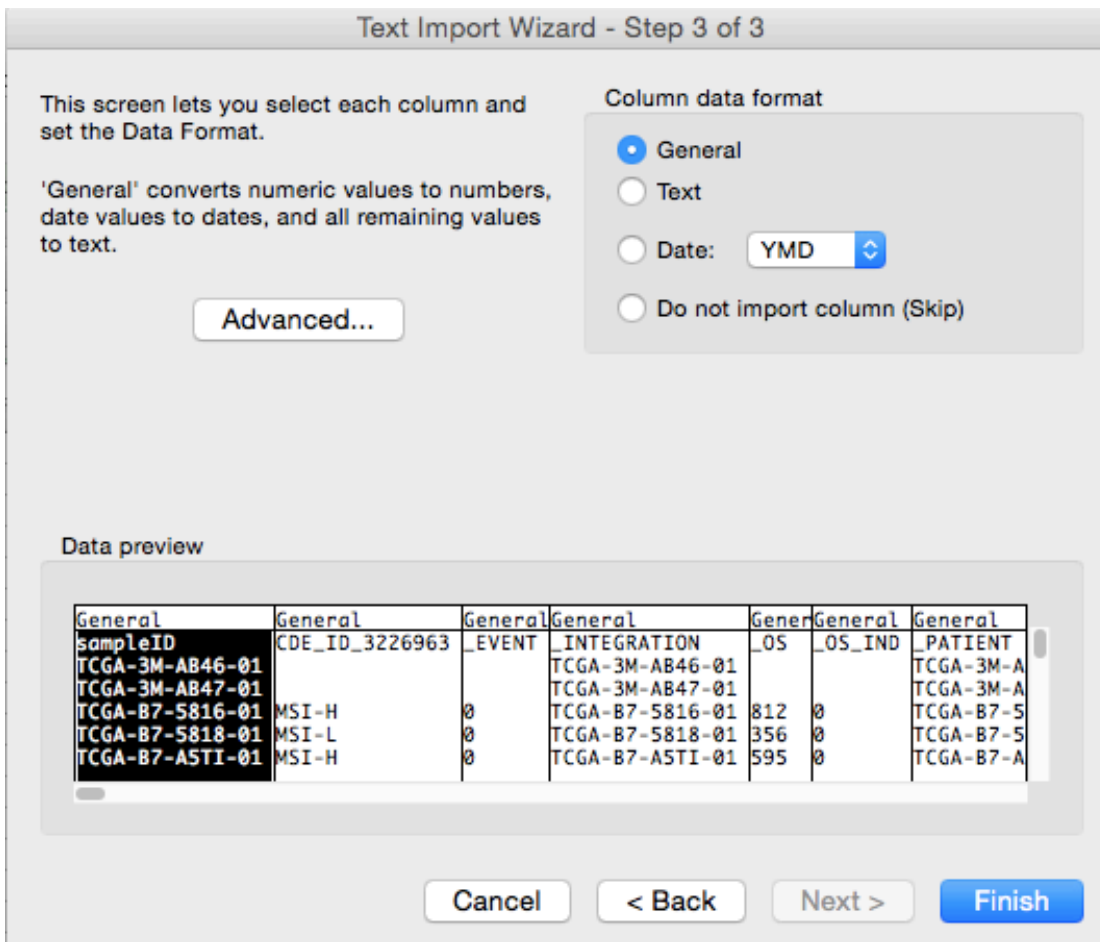
**Fig. SM4.** A step 1 of 3 in the text wizard of Excel. Choose the file type a "Delimited".

10. Check a "Tab" checkbox at a "Delimiters" panel, and then click a "Next" button.



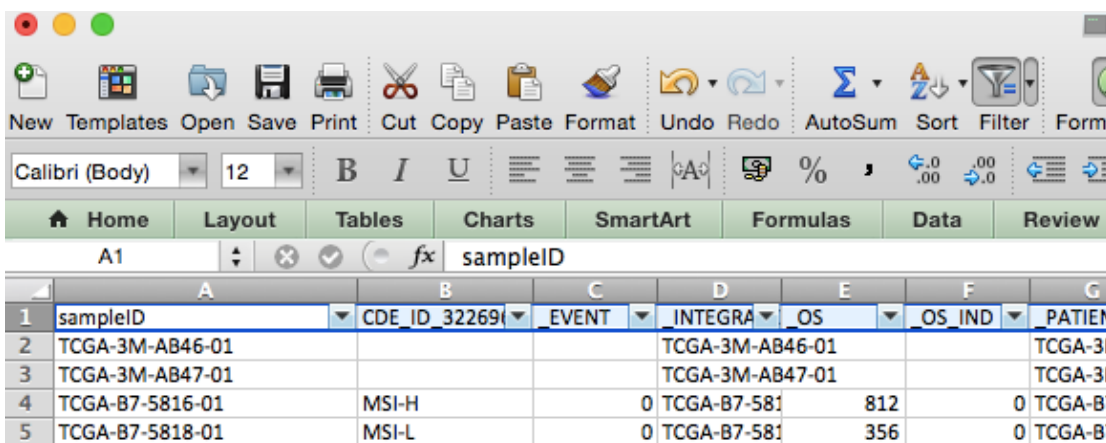
**Fig. SM5.** A second step of the text import wizard. A delimiter is "Tab". in this tutorial.

11. At the 3<sup>rd</sup> step of the Text Import Wizard, you can set a cell data type (format) of an each column. In this tutorial, we are not going to specify any column data type, click a "Finish" button.



**Fig. SM6.** A final step of text import wizard. Just click a "Finish" button.

12. Select whole columns in Excel™. Then Click a "Filter". (On MS Windows, "Data" >> "Filter" at ribbon menu tab.)



**Fig. SM7.** Set a filter with "Filter" button on a right-top side of Excel's menu bar.

13. Click a drop-down button on the "neoplasm\_histologic\_grade" column and uncheck all items and check a "G2" again.

AX	AY	AZ	BA	BB	BC
ph_no	neoplasm	new_neo	new_neo	new_neo	new_tum
	G2				
8	G2				
8	G2				
8	G2				
8	G2				
8	G2				
8	G2				
8	G2				
8	G2				
	G2				
	G2				
	G2				
	G2				
	G2				
	G2				
	G2				
6	G2				
6	G2				
	G2				
	G2				
6	G2				
6	G2				
8	G2				

**neoplasm\_histologic\_grade**

**Sort**

Ascending       Descending

By color:

---

**Filter**

By color:

Choose One

Search

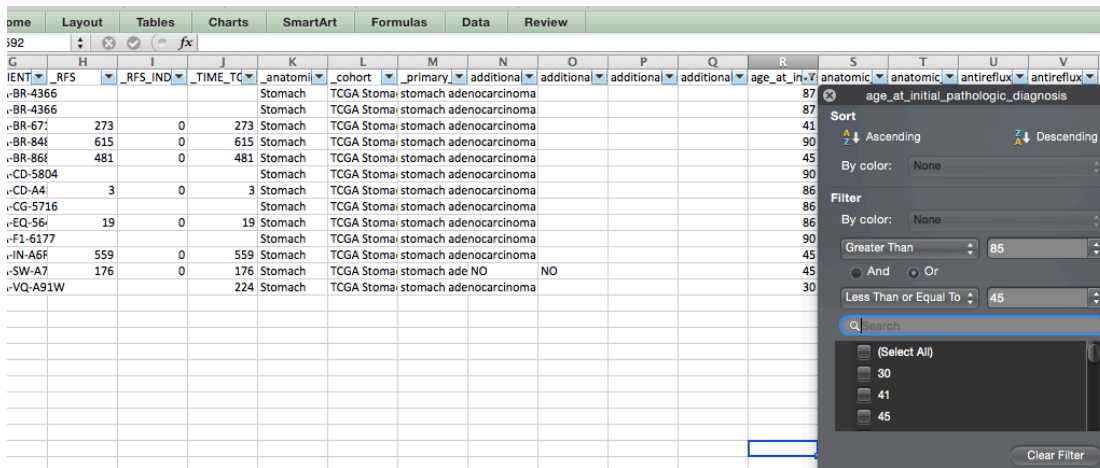
- (Select All)
- G1
- G2
- G3

Clear Filter

**Fig. SM8.** Set a filter value to "G2" at "neoplasm\_histologic\_grade" column.

14. Click a drop-down button on the "age\_at\_initial\_pathologic\_diagnosis" column and click a "Choose One" drop-down box on a Filter Panel. Select a "Greater than" and set its value as "85". When two radio "And" / "Or" buttons and filter option items are appeared, click an "Or" button, and then select a "Less Than or Equal To", and type "45" at a next input box. The filter option setting is shown at Fig. SM9.





**Fig. SM9.** Set filter values "Greater Than 85" or "Less Than or Equal To 45" at an "age\_at\_initial\_pathologic\_diagnosis" column.

15. We show TCGA RNA-Seq sample IDs of control and case groups at Table SM1. In summary, we selected the two groups, in the give G2 grade. One is <= 45-year old, and the other is > 85-year old.

**Table SM 1.** Selected Sample IDs under criteria, " neoplasm\_histologic\_grade " is "G2" and ("age\_at\_initial\_pathologic\_diagnosis" is greater than "85" or less than or equal to "45"). Note: RNASeq data of TCGA-EQ-5647-01, TCGA-BR-4366-01, and TCGA-BR-4366-11 are not available currently.

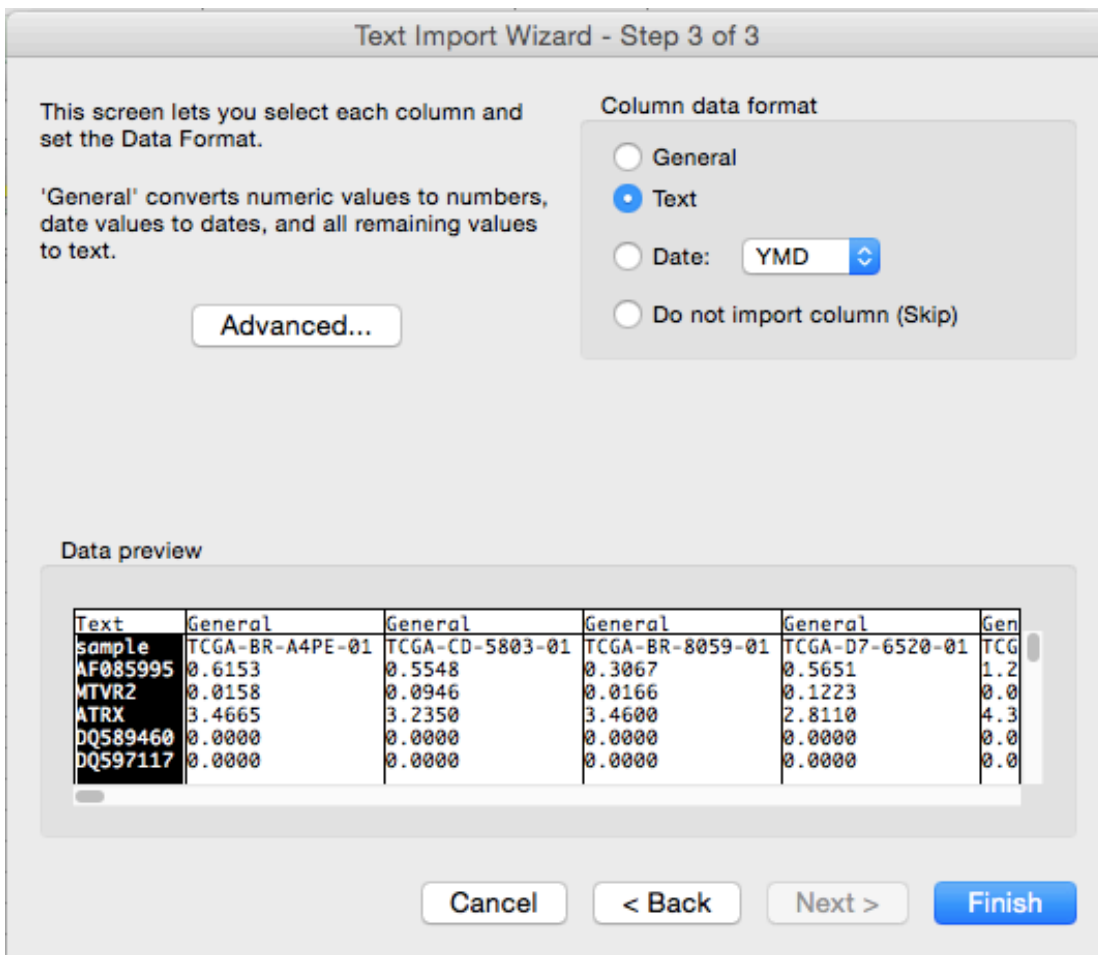
Group	Sample ID	MSI status	Age at diagnosis	Family history of stomach cancer	Gender	TNM tage
Control	TCGA-VQ-A91W-01	MSI-L	30	NO	M	Stage IIIA
Control	TCGA-VQ-AA69-01	MSS	35	NO	M	Stage IIIA
Control	TCGA-BR-6710-01	MSS	41	NO	M	Stage IB
Control	TCGA-BR-8680-01	MSS	45	NO	M	Stage IV
Control	TCGA-IN-A6RI-01	MSS	45	N/A	M	Stage I

Control	TCGA-SW-A7EB-01	MSS	45	NO	M	Stage IIIA
Case	TCGA-CD-A4MH-01	MSS	86	NO	F	Stage IIA
Case	TCGA-CG-5716-01	MSS	86	N/A	M	Stage IV
Case	TCGA-EQ-5647-01	MSS	86	NO	F	Stage IV
Case	TCGA-BR-4366-0	MSS	87	N/A	M	N/A
Case	TCGA-BR-4366-11	MSS	87	N/A	M	N/A
Case	TCGA-BR-8486-01	MSS	90	NO	F	Stage IA
Case	TCGA-CD-5804-01	MSS	90	NO	M	[Discrepancy]
Case	TCGA-F1-6177-01	MSI-H	90	NO	M	Stage I

**\*N/A : Not available.**

**STEP 3:** Extract RNA-Seq. expression data with selected sample IDs from a “HiSeq” file. A “HiSeq” has expression values of an each gene of all samples, is about 77.7Mbytes. It is possible to open this compact size file with Excel without any high performance computer. The UCSC Xena browser already matched gene symbols with its chromosomal positions and merged duplicated gene symbols and expression values as preprocessing.

16. Import a “HiSeq” file with same procedures importing a “STAD\_clinicalMatrix” to Excel above (from 6 to 9 in the STEP 2). It is important to set “sample” column’s data format as “Text” to prevent Excel changes some gene symbols to date symbols (for example, MARCH15 could change to Mar.15 automatically in Excel).



**Fig. SM10.** A third step of a text import wizard for "HiSeq" file. Be aware to set the first column, "sample" as "Text" in a "Column data format" panel at the right-top side of a dialog box.

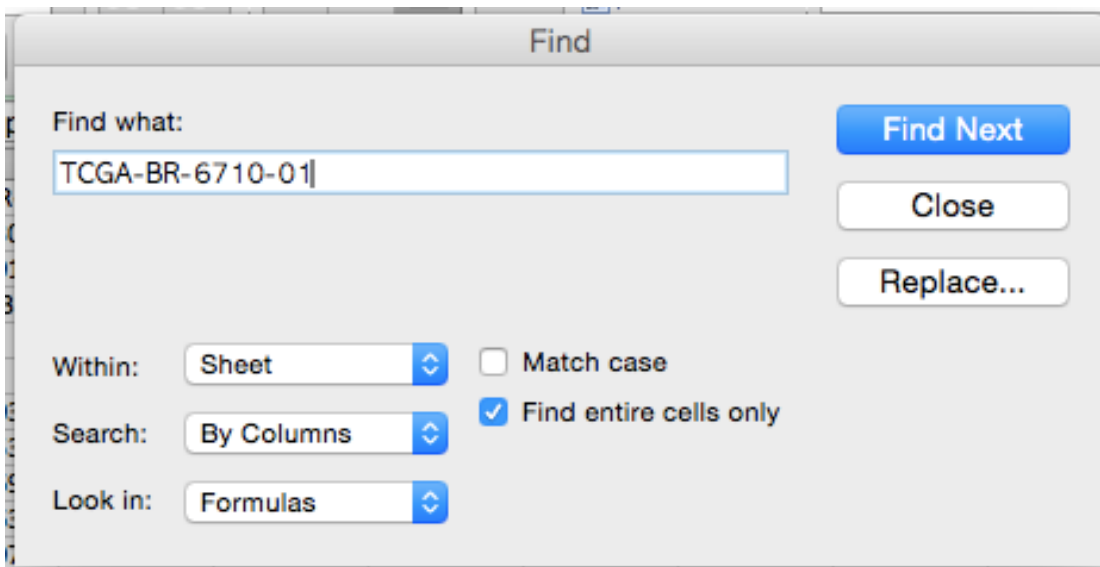
**STEP 4:** Filter patient samples and save it as the PATHOME's file format, a "sip".

17. An imported "HiSeq" file is shown in a Fig. SM13.

	A	B	C	D	E	F	G	H	I	J	K	L
1	sample	TCGA-BR-A4	TCGA-CD-58	TCGA-BR-80	TCGA-D7-65	TCGA-VQ-A8	TCGA-BR-65	TCGA-IN-78	TCGA-EQ-81	TCGA-FP-86	TCGA-R5-A8	TCGA-VQ-A
2	AF085995	0.6153	0.5548	0.3067	0.5651	1.2358	0.8888	1.1136	0.5685	2.0181	0.9244	0.958
3	MTVR2	0.0158	0.0946	0.0166	0.1223	0.0301	0.0641	0	0.3463	0.4333	0.4103	0.113
4	ATRX	3.4665	3.235	3.46	2.811	4.3868	3.1916	3.5604	3.7774	4.1118	3.7136	4.344
5	DQ589460	0	0	0	0	0	0	0	0	0.5144	0	0
6	DQ597117	0	0	0	0	0	0	0	0	0	0	0
7	LOC147670	0	0.0613	0.0318	0.0281	0.1304	0.0514	0.2399	0.0911	0.2895	0.0176	0.010
8	LOC1005068	0.8246	0.9225	0.5372	0.8129	1.0914	0.7069	0.8031	0.2346	0.1845	1.0929	0.766
9	LOC441204	0.2862	0.6972	0.6967	0.871	0.2905	0.2892	0.3298	0.4398	0.3337	0.2423	0.043
10	TCOF1	3.8807	4.2941	3.6339	3.6475	3.7229	4.7272	4.0327	4.3553	4.5557	3.8355	4.059
11	NSRP1	3.8323	3.2836	4.0706	3.2664	3.9653	3.5349	4.169	5.0449	4.2069	4.4636	4.042
12	SPPL3	3.7821	4.0686	3.6807	4.8475	4.0623	3.8636	4.0799	4.1949	4.1454	4.0856	4.075
13	OPA3	1.9568	2.0153	2.2366	1.9438	2.2211	2.3099	2.3163	2.1919	1.9384	2.357	2.254
14	OPA1	4.5098	3.3183	3.8034	3.2591	4.4445	3.7635	3.3	3.5652	3.8923	4.0654	4.115
15	ITGA8	0.7388	1.072	2.6139	0.6682	2.8608	1.4988	3.6632	0.812	1.2637	2.2816	1.475
16	ITGA9	1.2669	3.2068	2.276	1.766	2.6133	3.3276	2.1199	1.8385	1.69	2.0626	1.120
17	ITGA1	2.7323	3.8885	3.836	2.8007	4.0771	3.1258	3.2344	3.5014	3.0111	4.1962	3.302
18	ITGA2	3.4125	2.5892	4.978	2.5931	2.5829	3.3905	2.8298	3.4569	3.6047	4.5956	4.058
19	ITGA3	4.4317	5.1412	7.2258	6.6427	5.7148	4.9068	5.6138	6.1839	6.4863	6.3206	5.39

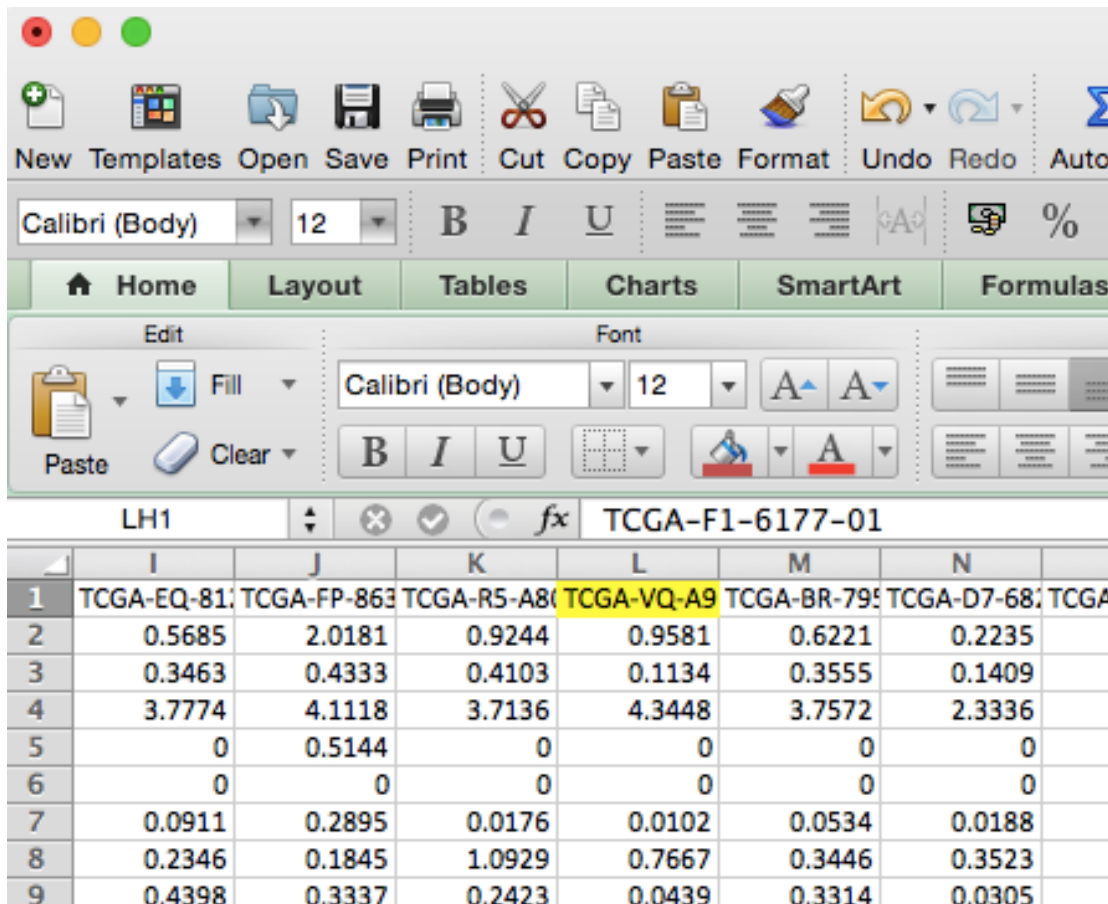
**Fig. SM11.** An imported " HiSeq " RNA-Seq expression values file in Excel.

18. Select sample IDs under our criteria (see Table SM1). Press a "COMMAND + F" for Mac or "Control + F" for Windows. Type a sample ID in a text box, select "By Columns" in "Search" drop-down box and check "Find entire cells only" box.



**Fig. SM12.** Find a sample ID under the criteria. Be sure select options to search column-wise and find entire cells only.

19. In the case of control group (“age\_at\_initial\_pathologic\_diagnosis” <= 45), we marked yellow cell background colors. And for case group (“age\_at\_initial\_pathologic\_diagnosis” > 85), we colored cells with red.



**Fig. SM13.** Mark a sample ID with a specific color to distinguish its group or deleting cells out of criteria later.

20. Select all columns out of criteria. Click the first column header, then click the last one pressing “Shift” key. Click a second (right) mouse button on a selected (highlighted) area then press a “Delete” menu item.

A	B	C	D	E	F	G
sample	TCGA-VQ-A9	TCGA-FP-821	TCGA-3M-AE	TCGA-D7-65	TCGA-BR-86	TCGA-BR-A4
AF085995	0.9581	0.5687	0.882	0.4732	1.3416	0.8522
MTVR2	0.1134	0.0973	0.2552	0.0909	0.0751	0.4233
ATRX	4.3448	4.0697	4.4263	3.3621	3.9638	4.7291
DQ589460	0	0	0	0	0.1262	0
DQ597117	0	0	0	0	0	0
LOC147670	0.0102	0.0846	0.1529	0.1062	0.1795	0.0596
LOC1005068	0.7667	0.7591	0.9452	0.6391	0.7978	0.7699
LOC441204	0.0439	1.2295	0.9085			
TCOF1	4.0594	4.5074	3.9294			
NSRP1	4.0422	4.2399	4.3729			
SPPL3	4.0756	3.3819	3.8298			
OPA3	2.2548	1.7225	2.3757			
OPA1	4.1151	4.4592	3.9274			
ITGA8	1.4753	1.1554	1.9774			
ITGA9	1.1206	1.0035	2.4703			
ITGA1	3.3028	2.3949	4.3386			
ITGA2	4.0582	4.5452	3.1688			
ITGA3	5.392	6.8997	4.3739			
ITGA4	3.0719	1.7206	2.6516			
ITGA5	4.6894	3.7627	5.1107			
ITGA6	6.1883	6.0488	5.3911			
ITGA7	1.535	1.0458	3.1234			
TRHR	0.0122	0.0192	0			
UFSP1	0.9467	2.2784	1.2766	1.1427	1.3139	1.7325

- Cut ⌘X
- Copy ⌘C
- Paste ⌘V
- Paste Special... ^⌘V
- Insert
- Delete
- Clear Contents
- Format Cells... ⌘1
- Column Width...
- Hide
- Unhide

Fig. SM14. Delete ruled out cells.

21. Rearrange columns with “cut” and “insert cut cell” menu items. Yellow colored cells are a control group and reds are a case group.

The screenshot shows the Microsoft Excel ribbon with the 'Home' tab selected. The 'Edit' group is active, showing 'Paste' and 'Clear' options. The 'Font' group shows 'Calibri (Body)' font and size '12'. The 'Alignment' group shows text alignment options. Below the ribbon, the active cell is E1, containing the text 'TCGA-CD-A4MH-01'. A data table is visible below the cell, with columns A through E. The table contains sample names and numerical values. A context menu is open over the table, showing options like 'Cut', 'Copy', 'Paste', 'Paste Special...', 'Insert Cut Cells', 'Delete', 'Clear Contents', and 'Format Cells...'. The table data is as follows:

	A	B	C	D	E
1	sample	TCGA-VQ-A9	TCGA-BR-86	TCGA-SW-A7	TCGA-CD-A4MH-01
2	AF085995	0.9581	1.3416	1.5411	1.1427
3	MTVR2	0.1134	0.0751	0.1371	0.0909
4	ATRX	4.3448	3.9638	3.5368	4.7291
5	DQ589460	0	0.1262	0	0.1262
6	DQ597117	0	0	0	0
7	LOC147670	0.0102	0.1795	0.2329	0.1062
8	LOC1005068	0.7667	0.7978	0.9581	0.6391
9	LOC441204	0.0439	0.2044	0.7586	
10	TCOF1	4.0594	5.7791	3.1097	
11	NSRP1	4.0422	4.2399	4.3729	

Fig. SM15. Rearrange columns.

22. Change a column name “sample” to “#NAME”. Then select whole a first row clicking the first row header (“1”) then click an “insert” menu item. Then type “c1” and “c2” respectively above each column, and “#Class” at the “A1” cell.

	A	B	C	D	E	F	G	H	I	J	K	L
1	#Class	c1	c1	c1	c1	c1	c1	c2	c2	c2	c2	c2
2	#NAME	TCGA-VQ-A9	TCGA-VQ-AA	TCGA-BR-671	TCGA-BR-866	TCGA-IN-A6F	TCGA-SW-A7	TCGA-CD-A4	TCGA-CG-57	TCGA-BR-846	TCGA-CD-58	TCGA-F1-617
3	AF085995	0.9581	1.4123	0.6836	1.3416	1.4639	1.5411	1.7395	0.3812	1.0144	0.3985	0.907
4	MTRV2	0.1134	0.0126	0	0.0751	0.1141	0.1371	0.0853	0	0.2176	0.1675	0.0719
5	ATRX	4.3448	4.5946	2.577	3.9638	3.4634	3.5368	4.2524	2.6532	4.6334	3.2344	2.8606
6	DQ589460	0	0	0	0.1262	0	0	0	0	0	0	0
7	DQ597117	0	0	0	0	0	0	0	0	0	0	0
8	LOC147670	0.0102	0.0161	0.0277	0.1795	0	0.2329	0.3554	0.0141	0	0.0986	0
9	LOC1005068	0.7667	0.6514	0.6476	0.7978	0.3167	0.9581	0.9639	0.442	0.8872	0.479	0.5254
10	LOC441204	0.0439	0.0694	0.1426	0.2044	0.1781	0.7586	0.28	0.1743	0.3724	0.2523	0.0507
11	TCOF1	4.0594	4.6129	2.4841	5.7791	3.7611	3.1097	3.9714	3.6218	4.7589	4.3561	4.7643
12	NSRP1	4.0422	3.7704	2.7652	4.2801	4.0405	2.9349	4.8242	3.1987	4.0175	4.6103	3.4266
13	SPPL3	4.0756	4.2889	3.2188	4.3327	4.3639	4.7257	3.9449	2.9885	3.9815	3.6215	4.1048
14	OPA3	2.2548	2.4324	1.543	2.1623	1.7468	2.5449	1.7107	2.0098	2.2733	2.0592	1.95
15	OPA1	4.1151	4.3624	3.1948	4.4157	4.3388	3.3316	3.877	3.6291	3.6783	3.4903	3.6639
16	ITGA8	1.4753	0.2142	3.0912	0.1716	0.2915	1.1299	0.7407	1.6797	2.8069	1.0711	0.258
17	ITGA9	1.1206	1.0504	2.5395	2.0779	0.3411	1.7675	1.2805	1.4146	2.3939	2.0087	0.8261
18	ITGA1	3.3028	4.7626	2.0422	2.6387	1.9408	2.2401	2.6572	1.5973	4.3652	2.8892	3.1425
19	ITGA2	4.0582	4.3872	1.5048	4.7602	2.6517	2.677	2.7262	2.5554	4.3491	1.911	3.4669
20	ITGA3	5.392	4.7929	4.595	5.1187	6.3294	4.9039	5.4393	5.4752	5.3086	5.8317	4.6889

Fig. SM16. Add meta tags for PATHOME web server data formats.

23. Save a “sip” formatted file into a folder with a file extension, tab delimited text file, “txt” (say, PATHOME\_TCGA\_GC\_45vs85.sip.txt).

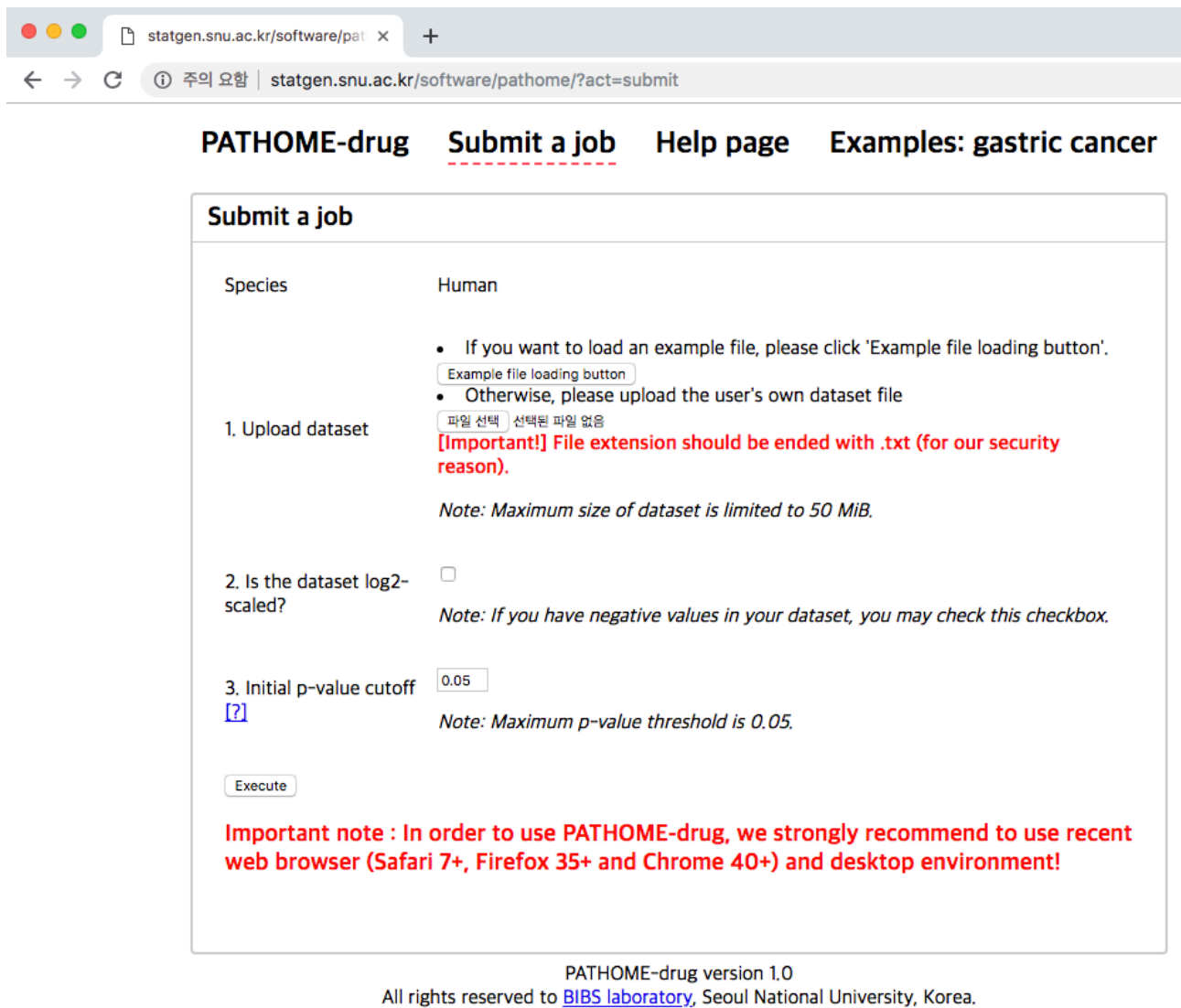
[Note] If you are using Excel in Mac OS, you need to convert to proper end of line (EOL) characteristics (details in <http://en.wikipedia.org/wiki/Newline>). Open the Terminal and type the following command (Note that the first \$ sign represents the prompt of the Terminal, which is not included in the command).

```
$ tr '\r' '\n' < PATHOME_TCGA_GC_45vs85.sip.txt > PATHOME_TCGA_GC.sip.txt
```

#### STEP 5: PATHOME Web service.

24. Visit the PATHOME-Drug Web site (<http://statgen.snu.ac.kr/software/pathome/>).





**Fig. SM17.** The PATHOME web (<http://statgen.snu.ac.kr/software/pathome/>).

25. Click a “File choose” button, and then select a sip file (PATHOME\_TCGA\_GC\_45vs85.sip.txt). If you converted EOL characters in Mac OS, upload PATHOME\_TCGA\_GC.sip.txt instead.
26. Check the check box, next of “log2-scaled?”. The UCSC Xena browser provides TCGA dataset with log2 transformed expression values ( $\log_2(x + 1)$ ).



## Submit a job

Species	Human
1. Upload dataset	<ul style="list-style-type: none"> <li>If you want to load an example file, please click 'Example file loading button'.  <input type="button" value="Example file loading button"/></li> <li>Otherwise, please upload the user's own dataset file.  <input type="button" value="파일 선택"/> <span style="margin-left: 10px;">선택된 파일 없음</span></li> </ul> <p style="color: red; margin: 0;"><b>[Important!] File extension should be ended with .txt (for our security reason).</b></p> <p style="margin: 0;"><i>Note: Maximum size of dataset is limited to 50 MiB.</i></p>
2. Is the dataset log2-scaled?	<input type="checkbox"/> <p style="margin: 0;"><i>Note: If you have negative values in your dataset, you may check this checkbox.</i></p>
3. Initial p-value cutoff <a href="#">[?]</a>	<input style="width: 50px;" type="text" value="0.05"/> <p style="margin: 0;"><i>Note: Maximum p-value threshold is 0.05.</i></p>

**Important note : In order to use PATHOME-drug, we strongly recommend to use recent web browser (Safari 7+, Firefox 35+ and Chrome 40+) and desktop environment!**

**Fig. SM18.** PATHOME-Drug web server option.

27. Click the “Execute” button. (The result is in the section “How to make a PATHOME-Drug input from TCGA gastric cancer dataset in users' own preferences” from <http://statgen.snu.ac.kr/software/pathome/?act=gcdatasets>).