How to Import TCGA Data

Contents

1. FOREWORD

The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/abouttcga/overview) is a large collaboration between 20 institutes across the United States and Canada. Altogether, TCGA has collected 2.5 petabytes of data describing 33 different tumor types across 11,000 patients. Significant findings include identification of therapeutic targets and a better understanding of the molecular basis of cancer as well as improved ability to classify cancer subtypes.

With the amount of data available, TCGA offers an invaluable resource to mine for new insights or support current research. However, with many data types and data quality levels available, understanding how to obtain the data is challenging. Therefore the Broad Institute's GDAC, http://gdac.broadinstitute.org/, resource provides access to analysis ready TCGA data compiled for each cancer cohort. GDAC maintains many data files organized as tables of patients and variables and is an ideal resource for analyzing TCGA with Qlucore Omics Explorer.

2. OPTION 1 - TEMPLATE

There are two options for import of TCGA data into Qlucore. Option 1 is to use the "TCGA RSEM" template. The other option (described in section 3 and onwards below) is to manually download the data set(s) of choice.

Start the **Template Browser** by pressing the Quick launch button "Templ." or by selecting **File -> Template Browser.**

Then select the "TCGA RSEM" template and follow the instructions.

3. OPTION 2 - MANUAL DOWNLOAD

3.1. REQUIREMENTS AND SETUP

The majority of tasks may be completed with only Qlucore Omics Explorer (QOE). To follow along you will need to make sure that you have an internet connection, can access the GDAC resource, http://gdac.broadinstitute.org/, and have the QOE installed. This guide was prepared using the Windows 64bit version of QOE, but is applicable to all other versions. It will be clearly noted where there are differences between the versions. An advanced section is included toward the end of the guide that details how to add TCGA annotations to the QOE. This section is entirely optional and will require you to code in R. We have selected R-Studio (found here, https://www.rstudio.com/) to help us with this task as it is easy to install.

3.2. SUPPORTED DATA FORMATS AND DATA TYPES

The GDAC resource supplies analysis ready tables of TCGA data in plain text files. Columns usually describe patients (samples) and are separated by tab characters. Rows usually describe variables such as gene expression values. This document will only cover these tabular data types and will therefore make use of QOE's wizard tool for importing a wide range of TCGA data.

3.3. GENERAL PROCEDURE

Qlucore QOE can easily import many TCGA data types. However, it is important to understand the separation of data within QOE. In QOE the bulk of the data import will be on a numeric

data. Clinical annotations, and descriptions and categories in QOE are called sample and variable annotations and can be imported in various ways.

Some GDAC tables use a block column format to store several pieces of information for the same patient. For example, in RNA sequencing, a single gene (row - variable) may have both raw read counts as well as normalized RPKM values. Block formatted tables are regular in the sense that the blocks always repeat, one for each patient in a predictable manner. The QOE wizard can import this more complex data table with ease, an example of which will be shown later.

This guide will teach you the general procedure to identify what TCGA-GDAC tables may be imported to QOE, to launch the QOE wizard, and to follow the wizard's prompts. This will import the basic information necessary for analysis in QOE.

To summarize:

1. *Go to GDAC*
2. *Browse and select the data set of your choice*
3. *Select browse in the data column*
4. *Download the file*
5. *Unpack the file*
6. *You would now normally have two files. One is a manifest file which you can ignore.*
7. *Open the file with the Qlucore Omics Explorer Import Wizard. Follow the steps in the Wizard.*

4. BY EXAMPLE

The following examples will make use of the Breast Invasive Carcinoma cohort as it is one of the most feature complete cohorts in TCGA. You may follow along by visiting the Broad Institute's GDAC resource here, http://gdac.broadinstitute.org/, then click on "browse" under the data column for the BRCA cohort.

We will make use of the QOE wizard for all examples shown. You will find both written documentation and video demonstrations of the wizard on qlucore.org, here, http://www.qlucore.com/documentation.

4.1. EXAMPLE-1: SIMPLE DATA IMPORT

For this example, we will make use of the protein abundance data found toward the bottom of the BRCA list presented in figure 1 & 2. Click on "mda rppa core-protein normalization" to download a file labeled "gdac.broadinstitute.org_BRCA.Merge_protein_exp__mda_rppa_core__mdanderson_org__Level_3__protein_normalization__data.Level_3.2016012800.0.0.tar.gz". On Mac OS X, you will be able to double click the downloaded file to unpack it and access the data. On Windows you will need to use a tool such as 7-zip, found here, http://www.7-zip.org/.

Once unpacked, you should find a folder containing two files: the first is a file manifest that you may safely ignore; the second is the data file (about 3 megabytes in size) containing tabularized protein abundance profiles for each sample in the cohort.

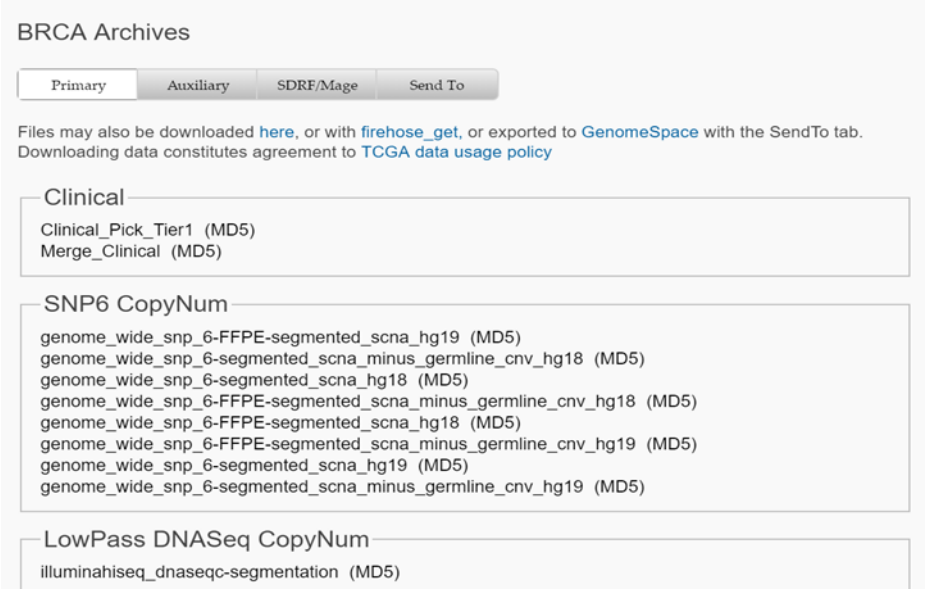


Figure 1. Screenshot of data compiled for the BRCA cohort, from GDAC

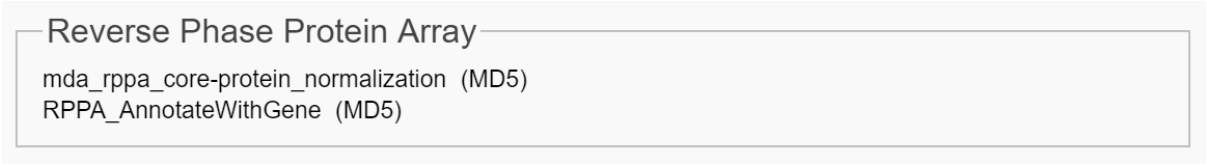
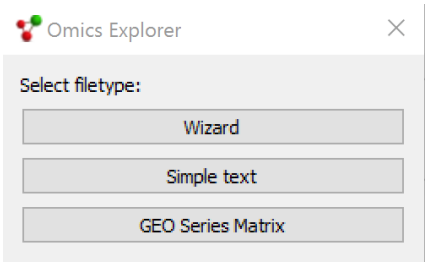


Figure 2. Screenshot of data to download for Example-1: Simple data import

Now that you have some data let's start by running QOE. Once opened, click "File", click "Open with Wizard…", navigate to the data file and click "Open". This will bring up the dialogue shown to the right, click "Wizard" to begin telling QOE how to interpret the data file.



The first screen of the wizard gives information about the format and anatomy of the data the wizard expects (see figure 3). The basics of the wizard are described more generally in other documentation. In particular, we recommend watching the video tutorial, http://www.qlucore.com/documentation. Click “Next” to proceed.

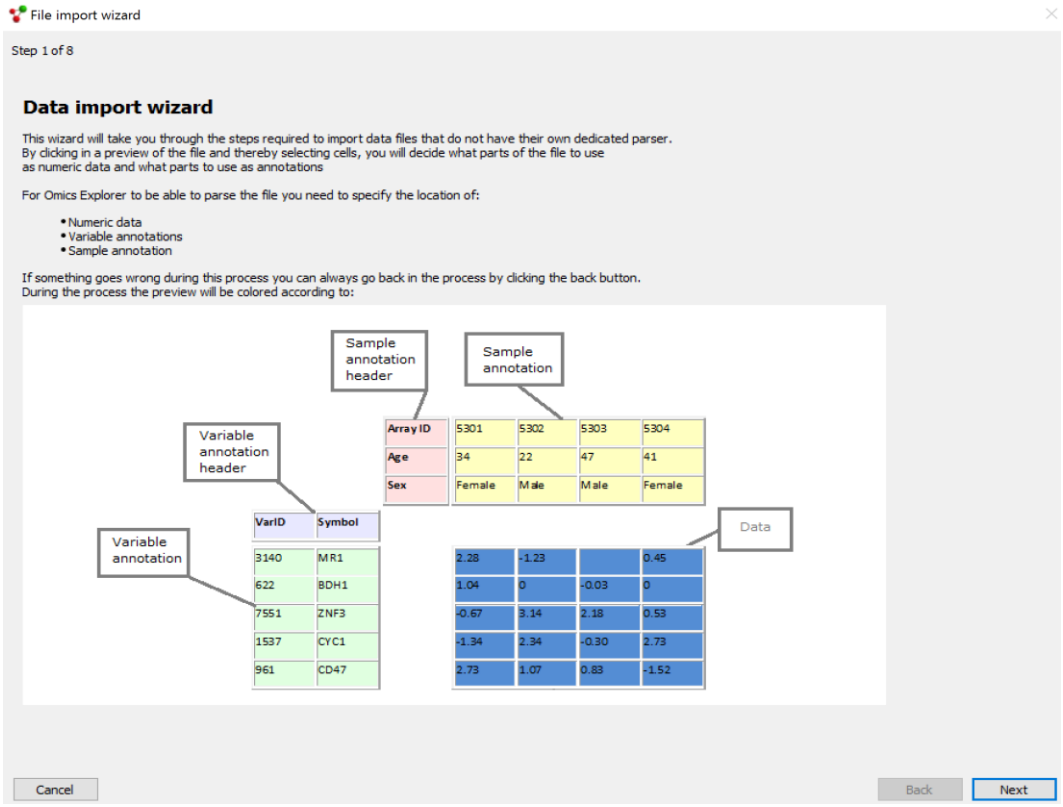


Figure 3: Screenshot of QOE Wizard, step 1 of 8

In the second step, tell QOE what character separates columns in the file. In this case it is a tab character denoted by the special character “\t” and it should be selected for you automatically. Click on other separators to see the difference. Click “Next” to proceed.

In the third step we need to tell QOE how the data is oriented, are the rows variables or samples. By default, Qlucore uses rows as variables, and that too is the default of most GDAC sourced files. Click “Next” to proceed.

In the fourth step we define the boundaries of the numeric data to be imported into QOE. In this step we click the top-left most data value as shown in figure 4. This defines columns to the left as possible row annotations. It also defines rows above, as possible column annotations. In figure 4 we see that column-2 and row-3 was selected since the first row contains sample IDs (barcodes in TCGA) and the second contains non-numeric information. Click “Next” to proceed.

In step five we do the same but select the top-right most data cell. This allows us to subset the data with greater control. This step will become more useful in example-2, where we make use of the “Data on every ith column” feature. Click “Next” to proceed.

In the sixth and seventh steps we select row (variable) and column (sample) annotations, respectively (as shown in figure 5). In this case there is only one level of row annotations, the protein name. Click “Finish” to complete the process and import data into QOE. This last step concludes our first example, you should see an image very similar to figure 6. In the next example, we describe how to import some slightly more complex data tables provided by GDAC.

	1	2	3
1	Sample REF	TCGA-3C-AALI-01A-21-A43F-20	TCGA-3C-AALK-01A-21-A43F-20
2	Composite Element REF	Protein Expression	Protein Expression
3	14-3-3_beta	-0.000751765000000182	-0.202251803
4	14-3-3_epsilon	0.0225531484999999	0.0770409575000002
5	14-3-3_zeta	0.021111648	0.153997099
6	4E-BP1	0.101796404	0.299107435

Figure 4: Screenshot of QOE Wizard, step 4 of 8

	1	2	3
1	Sample REF	TCGA-3C-AALI-01A-21-A43F-20	TCGA-3C-AALK-01A-21-A43F-20
2	Composite Element REF	Protein Expression	Protein Expression
3	14-3-3_beta	-0.000751765000000182	-0.202251803
4	14-3-3_epsilon	0.0225531484999999	0.0770409575000002
5	14-3-3_zeta	0.021111648	0.153997099
6	4E-BP1	0.101796404	0.299107435

Figure 5: Screenshot of QOE, step 7

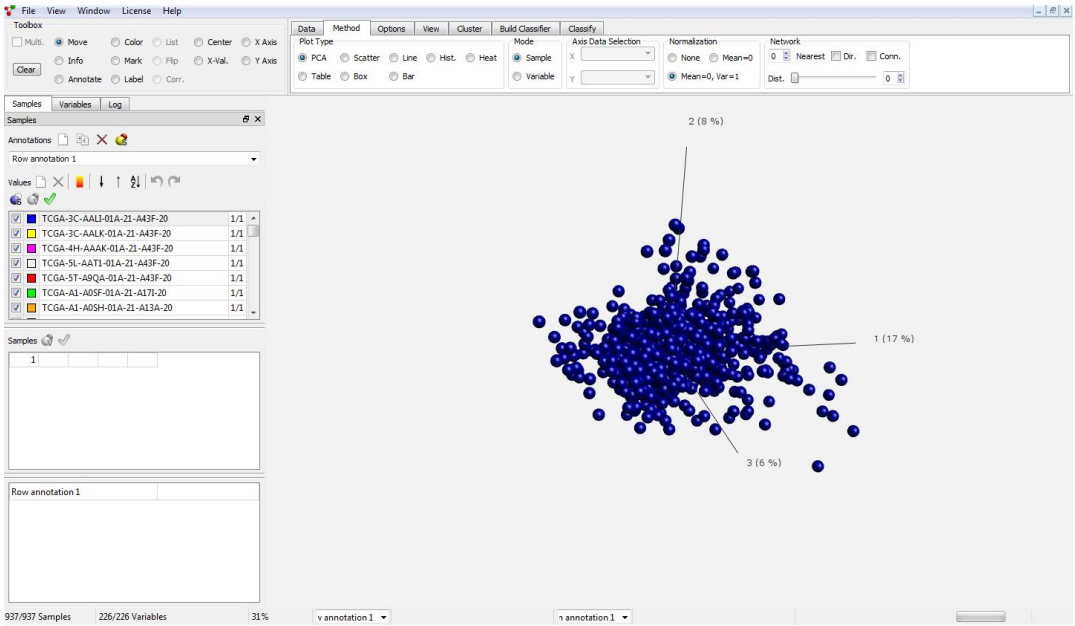


Figure 6: Screenshot of QOE, after importing TCGA data through the wizard

4.2. EXAMPLE-2: IMPORTING FROM BLOCKED COLUMN FORMATS

In this example we will make use of the Illumina HiSeq dataset on miRNA gene expression (again in the BRCA cohort; see figure 7). Follow the same instruction as in example-1 in order to download and unpack this dataset.

Figure 7: Screenshot of data to download for example-2

Taking a look at this dataset in spreadsheet software (figure 8) we can see that the columns have a meta structure whereby they represent multiple data types per patient. Note that for each patient, there are three columns: read_count, reads_per_million_miRNA_mapped, and cross-mapped. Also, note that the columns repeat in a predictable way, every 3rd column contains the same data. Finally, note that the columns labeled "cross-mapped" are not numerical and cannot be imported directly into QOE.

Hybridization REF	TCGA-2F-A9	TCGA-2F-A9KO-01A-11R-A38M-13	TCGA-2F-A9KO-(	TCGA-2F-A9I
miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-mapped	read_count
hsa-let-7a-1	20140	5142.198119	N	19131
hsa-let-7a-2	40107	10240.22542	Y	38407
hsa-let-7a-3	20445	5220.071526	N	19520
...	...	...	...	...

Figure 8: Structure of block column format in BRCA Illumina HiSeq dataset on miRNA gene expression

The QOE Wizard provides tools to extract data from this table in a clean way. To proceed, follow example-1's instructions up to the fourth step in the QOE Wizard.

Let's say that we want to extract reads per million data for our analysis. Now, in the fourth step we must defined the top-left value of the table we wish to import. In this case select row-3 and column-3, i.e. the first value under "reads_per_million_miRNA_mapped". This will allow us to use the "Data on every i^{th} column" feature in the next step. Click "Next" to proceed.

In this fifth step we will extract every 3rd column of data after the beginning of the data matrix we defined in the step-4. To do this, change the value in the "Data on every 1 column" widget to 3, because the data (shown in figure 8) repeats every 3 columns. Then define the end of the data matrix by scrolling the horizontal bar all the way to the right and clicking the top-right most value. Click "Next" to proceed and follow the final instructions as outlined for steps 6 and 7 in example-1. Before clicking on "Finish" in step-7 you should see figure 9. When the data is

imported you will see the PCA calculated for miRNA gene expression based on the RPM data provided by GDAC and TCGA (figure 10). This concludes example-2. You should now be able to import many different datasets from the TCGA.

	1	2	3	4	5	6	7	8	9
1	Hybridization REF	TCGA-3C-AAAU-01A-11R-A41G-13	TCGA-3C-AAAU-01A-11R-A41G-13	TCGA-3C...	TCGA-3C...	TCGA-3C...	TCGA-3C...	TCGA-3C...	TCGA-3C...
2	miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-m...	read_co...	reads_p...	cross-m...	read_co...	reads_p...
3	hsa-let-7a-1	95618	8962.996542	N	49201	7739.75...	N	75342	8260.61...
4	hsa-let-7a-2	189674	17779.575039	Y	98691	15524.9...	N	150472	16497.9...
5	hsa-let-7a-3	96815	9075.200383	N	49035	7713.62...	N	76206	8355.34...
6	hsa-let-7b	264034	24749.898857	N	148591	23374.6...	N	99938	10957.3...
7	hsa-let-7c	3641	341.298400	Y	5095	801.487...	N	5799	635.811...
8	hsa-let-7d	4333	406.164781	N	3263	513.297...	N	5658	620.351...

Figure 9: Screenshot of QOE, step 7, showing the extraction of data from every 3rd column

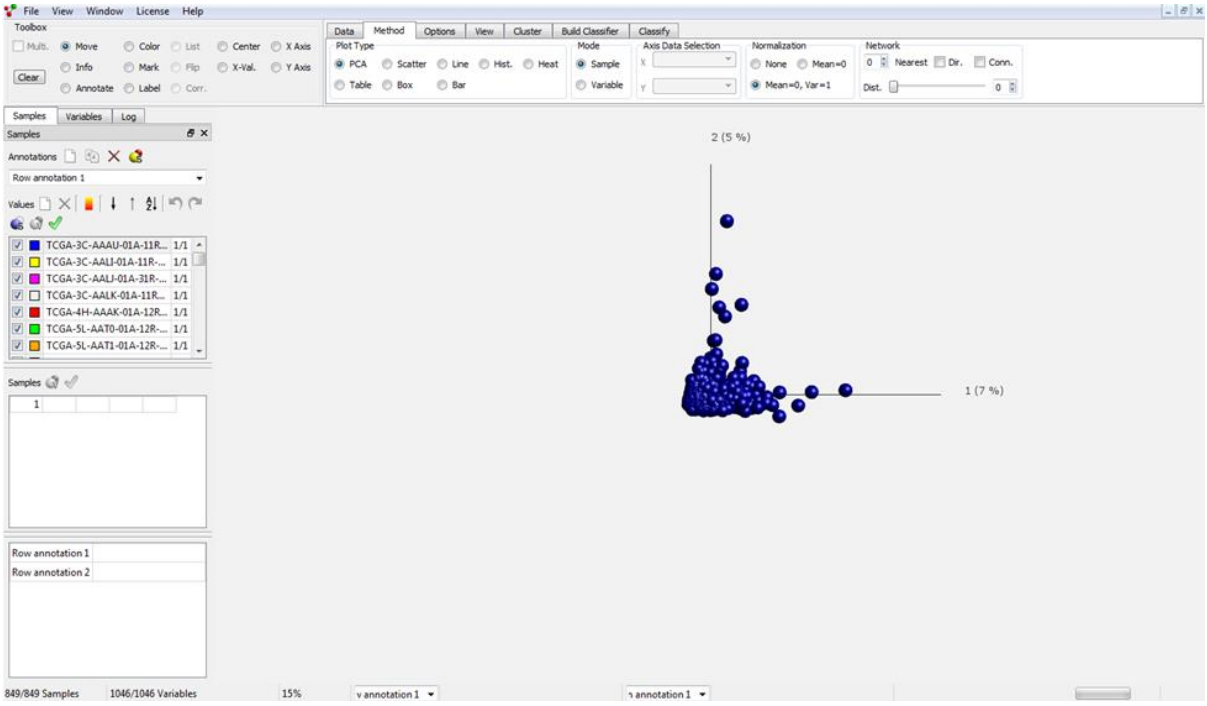


Figure 10: Screenshot of QOE, after the successful import of example-2's dataset

4.3. EXAMPLE-3: IMPORTING CLINICAL ANNOTATIONS

This is an advanced example that will require you to code in R and manipulate data. The task is to import clinical data on patients (often categorical) and map it to sample data such as protein arrays. Annotation in QOE is a powerful way to gain a better understanding of your data. For this task we will make use of the same type of data as presented in Example-1, the reverse phase protein array data. Only, this time we will use the bladder cancer dataset, http://firebrowse.org/?cohort=BLCA&download_dialog=true, to give you experience with a different cohort and demonstrate the continuity of data formats found throughout the GDAC resource.

The first few steps are to (1) create a folder specific to our task (we will use "./Example-3/"), and (2) download the datasets and unpack them into our new folder. This time, you will also

download the clinical data labeled “Merge_Clinical” from the BLCA cohort. Once you have unpacked the datasets, we suggest to rename the protein dataset to something a little easier. We’ve labeled ours “./Example-3/MY_GDAC_BLCA_PROTEIN_DATA/” and “./Example-3/MY_GDAC_BLCA_CLINICAL_DATA/”. For the purposes of this example we will use these folder names, but feel free to use your own. Make sure to record the path to both datasets as we will need them in the following steps.

Let’s start by opening R-Studio and creating a project. Since we’ve already created the folder “Example-3” we will use it to store our R-Studio project details. In R-Studio, select “File > New Project” and then “Existing Directory”. Navigate to and select the folder named “Example-3”. R-Studio will now setup the folder as an R-Studio project. It is still a normal folder as far as Windows or Mac OS X is concerned. However, R-Studio will have many convenient features such as a memory of what you did in the project and setting the working directory to “Example-3” so that you don’t have to keep writing long file paths when reading in data.

We should now create a text file to store our R script in. We will use it to format our annotation data correctly for QOE. Click “File > New File > R Script”. A tab named “Untitled1” should appear. Click “File > Save”, and give the file an appropriate name such as “process_annotations.R”.

We will begin by annotating our protein samples with the identifiers of the patients they came from. We need to do this because the clinical data is patient centric. The protein data is labeled with full TCGA barcodes that contain the patient IDs within them (read about the TCGA barcode format here, https://wiki.nci.nih.gov/display/TCGA/TCGA+Barcode). Therefore, our initial task is to import these barcodes from the protein data file and then extract the patient IDs from them before finally creating a mapping file for import into the QOE.

To do this we will make use of the ‘readr’ and ‘stringr’ packages. Write the following code to the top of the script file:

```
# Attach required packages

library(readr) #- for fast reading of data files

library(stringr) #- for string manipulation
```

Next, we will read in the protein data file. To do this, we will define a working directory rooted by our project, “Example-3”, and then use read_delim from the readr package to load the data into the R environment. Use the following code, but take note and make sure you use the paths you have chosen otherwise you will get “file not found” errors.

```
##################################

####### Convert barcodes to patient IDs

workingDirectory = "./MY_GDAC_BLCA_PROTEIN_DATA"
```

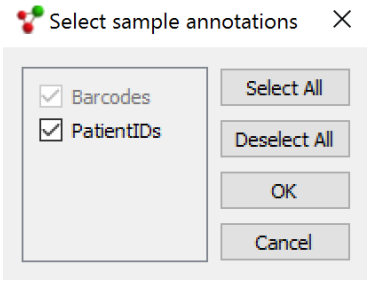
```
proteinFile =
"protein_exp__mda_rppa_core__mdanderson_org__Level_3__protein_normalization__data.txt"

pathToMyFile = paste(workingDirectory, proteinFile, sep = "/")

mydata = read_delim(pathToMyFile, "\t", col_names = FALSE) #- read in the data file
```

We now have our data in R-Studio in the variable called 'mydata'. If you click on 'mydata' in the Environment tab, R-Studio will display it like a spreadsheet for you to browse. The next step is to extract the barcodes from this dataset and then the patient IDs. Use the following code:

```
barcodes = as.character(mydata[1,-1]) #- Extract the barcodes from the file

patientIDs = str_extract(barcodes, "(.+?)-(.+?)-(.+?)-") #- Extract patient information from
barcodes

patientIDs = str_replace(patientIDs, "-$", "") #- Clean the trailing "-" character
```

With this last code, we have extracted a list of barcodes and from that, a list of patient IDs. We now need to put them together as a table and save it to disk. Since the procedure kept the order between the lists intact we can simply use the following code:

```
myDF = data.frame(Barcodes = barcodes, PatientIDs =
tolower(patientIDs)) #- Create dataframe

outFileName = "my_annotation_file.txt"

write.table(myDF, file = paste(workingDirectory,
outFileName, sep= "/"), quote = FALSE, sep = "\t",
row.names = FALSE)
```



There should now be a filename "my_annotation_file.txt" in your Example-3 project folder. Let's import this into the QOE and make the first annotation of our protein samples. To do this you will need to import the BLCA protein data as you first did in Example-1. Once you have done so, in the QOE click on "File > Import > Sample Annotations" and select "my_annotation_file.txt". A dialogue will appear and you will notice that Barcodes is greyed out while PatientIDs is not. Here, it is important to realise that the QOE considers the first column in an annotation file as the KEY identifier that will be used to map patient IDs to the data already imported into the QOE. If this KEY does not match, the QOE will not find anything.

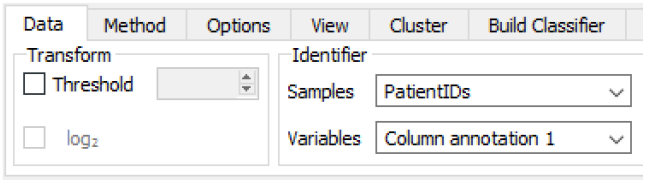
Congratulations, you have now imported patient IDs into Qlucore Omics Explorer and mapped them to the barcodes of the protein dataset. But, we are only half way. We've told QOE which

samples belong to which patient, now we need to annotate those patients with clinically relevant information. For this we will run through a similar process as above. Typically, every TCGA study contains patient level data ideally suited for import into QOE, this is the data type we will next focus on and is labeled “Merge_Clinical” on the online GDAC resource. Use the code below to import clinical data into R and get it ready for QOE:

```
workingDirectory = "./MY_GDAC_BLCA_CLINICAL_DATA"

pathToMyFile = paste(workingDirectory, "All_CDEs.txt", sep = "/")

myClinicalData = read_delim(pathToMyFile, "\t", col_names = FALSE) #- read in clinical data

myClinicalData = t(myClinicalData) #- Transpose data for Qlucore QOE

myClinicalData[1,1] = "PatientIDs"

outFileName = "my_clinical_annotations.txt"

write.table(myClinicalData, file = paste(workingDirectory, outFileName, sep = "/"),

            quote = F, sep = "\t", row.names = F, col.names = F)
```

In this last code we (1) changed the working directory, (2) set the file path to “All_CDEs.txt”, (3) read in the data and transposed it, (4) changed the name of the identifier to match what we previously used “PatientIDs”, and (5) write the table to “my_clinical_annotations.txt”.

Now that we have a clinical annotations file in the correct orientation we can import it into the QOE. But, first, we need to tell QOE that the identifier used for the mapping should be the “PatientIDs” that we first mapped, and not the barcodes. To do this click on the “Data” tab in QOE and select the drop-down list labeled “Sample” within the “Identifier” section. Select “PatientIDs”. Now that QOE knows to use PatientIDs as an identifier, go ahead and import the clinical data as before: “File > Import > Sample Annotations” and choose the file named “my_clinical_annotations.txt”. You’ll notice the dialogue that appears contains a lot more annotations to import than before. For now just select all of them. QOE will warn you that some patients in the clinical dataset are not represented in the protein dataset, this is OK and you can safely ignore this warning. But, it is good to realise that some datasets may not overlap perfectly.



This concludes the final example in this guide. You should now be able to import complex datasets into the QOE from the TCGA study. Annotations offer a powerful way to explore the relationships within your data, make sure to review the other How-To guides and video tutorials available on the Qlucore website, http://www.qlucore.se/home.

5. USAGE AND ACKNOWLEDGEMENT

Qlucore has no direct affiliation with TCGA. This how-to guide makes use of TCGA data and therefore requires that you accept the Data Use Policy and Publication Guidelines to promote the responsible use of TCGA data sets. All investigators, and their institutions, seeking access

and use of TCGA data must acknowledge their agreement with TCGA policies and procedures. Please note that downloading data from the Broad Institute GDAC constitutes an acknowledgement that you, and any collaborators who use this data with you, will conduct research and publish in accordance with TCGA guidelines on responsible use of data, as informed by The Fort Lauderdale Agreement. You will be able to acknowledge your understanding of this data use policy and TCGA guidelines online whenever downloading TCGA data.

6. DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.