



Tutorial

RNA-Seq Analysis of Breast Cancer Data

June 20, 2019

— Sample to Insight —

RNA-Seq Analysis of Breast Cancer Data

This tutorial gives a brief overview of how to analyze RNA-Seq data using *CLC Genomics Workbench*. To reduce the processing time required by the workbench to run the RNA-seq analysis and thereby be able to complete the tutorial, we will work with a subset of publicly available RNA-Seq data. The subset of the data maps to chromosome 17 of the human genome. This tutorial covers the following:

- Creating a metadata table and associating samples to it.
- RNA-Seq analysis with a description of the generated results.
- Demonstration of a Ready-to-Use Workflows that can identify and annotate differentially expressed genes and pathways.

Prerequisites For this tutorial, you must have installed the Biomedical Genomics Analysis plugin. Minimum recommended machine specifications for working with human data sets are listed at <http://www.qiagenbioinformatics.com/system-requirements/>, but for this tutorial a standard desktop computer/laptop with 4 GB RAM should be sufficient.

Download and import data

In this tutorial we will use a subset of the data reported in : <http://www.ebi.ac.uk/ena/data/view/SRP032789>. The original data contains mRNA profiles of 17 breast tumor samples of three different subtypes (TNBC, non-TNBC and HER2 positive) and normal human breast epithelium samples that were sequenced using Illumina HiSeq. We will use only 3 triple negative breast cancer (TNBC) samples and 3 HER2 positive breast cancer samples that mapped to chromosome 17. Go through the following steps to download and import the data into the Workbench.

1. Download the sample data from our website: http://resources.qiagenbioinformatics.com/testdata/HumanChr17dataset_25.zip.
2. Start the workbench.
3. Import the data via the toolbar: **File | Import**  | **Standard Import** 
4. Choose the zip file called HumanChr17dataset_25.zip. Leave the import type set to Automatic.
5. Save the imported data.

Once the data has been downloaded and imported, you should have the folders and data in the Navigation Area as shown in figure 1.

Creating a metadata table

A metadata table indicating to which group belong each sample will be needed later in order to compare gene level expression between the six samples.

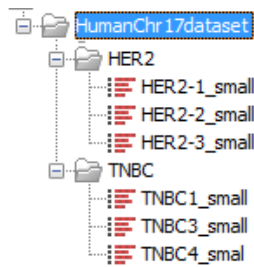


Figure 1: The data set includes the reads from the TNBC and HER2 tissue samples that were mapped to chromosome 17.

While it is possible to create a metadata table in a spreadsheet and import it using the **Import | Metadata** functionality of the workbench, we will show in this section how to create a metadata table directly in the workbench.

1. Go to **New | Metadata Table**.
2. Click **Set Up Table...** at the bottom of the view that opened in the View Area.
3. Click on the **Add column** button (highlighted in red in figure 2) and write "Sample" as the name of the first column. Check the option "Key column", as the values from this column will be matched to the sample names during the association step.

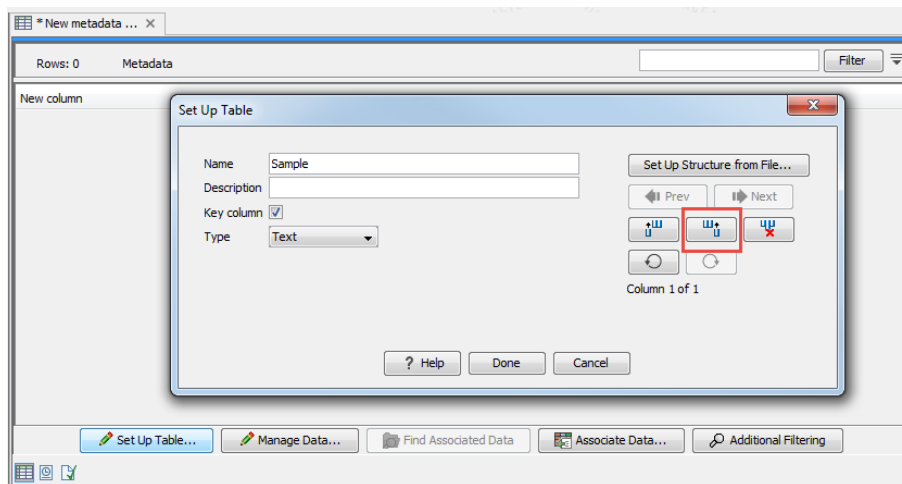


Figure 2: Set up the columns of the metadata table.

4. Click once more on the **Add column** button and call the second column "Group". Click **Done** as the table only contains 2 columns in this tutorial.
5. The metadata table opened in the View Area now has two columns. Click **Manage data...** and add six rows using the button highlighted in red in figure 3. Fill in the six rows with the names of the samples and groups as specified in figure 3, using the **Prev** and **Next** buttons to switch from one row to another. Click **Done** when the table is ready.
6. **Save** the table in the Navigation Area.
7. Click on **Associate Data** at the bottom of the table and choose the "Associate Data Automatically" option.

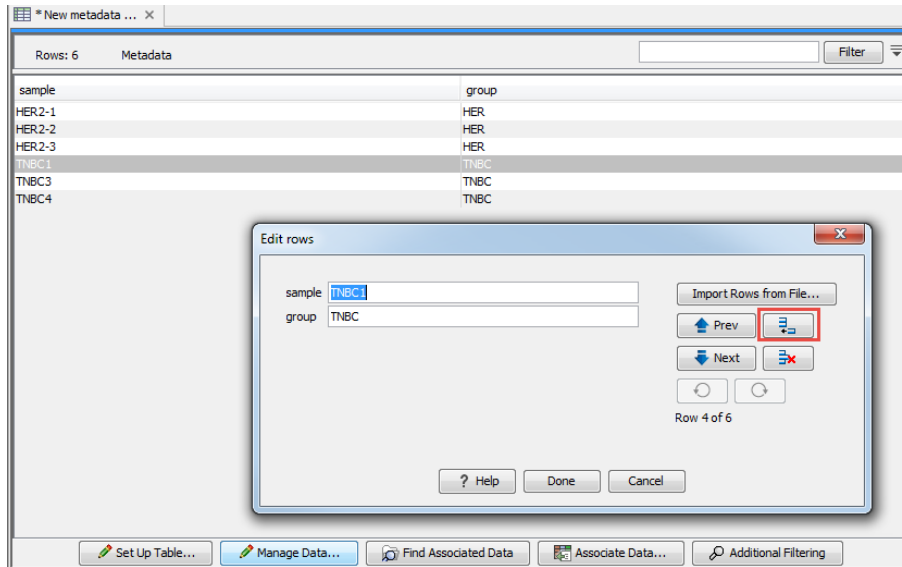


Figure 3: Set up the rows of the metadata table.

- Right-click the folder containing all the initial reads (HumanChr17dataset_25) and select the option "Add folder contents (recursively)" to select at once the six samples (figure 4). Click **Next**.

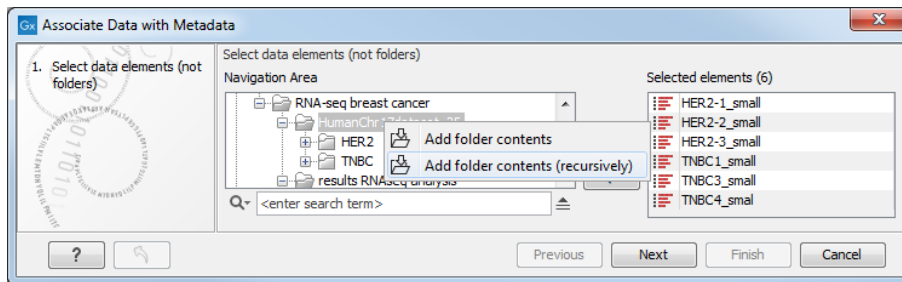


Figure 4: Associate the metadata tables with the samples.

- Leave the "Role of input data" to "Sample data" and click **Next**.
- Choose the option "**Partial**" as we did not write the full name of the samples when filling in the Sample column values. Click **Next**, then click **Finish**.
- You can check that the association was successful by opening a sample (sequence list in the Navigation Area) and choosing to see its "Element Info" view (with the button highlighted in red in figure 5).

Running the RNA-seq analysis

In this section we analyze TNBC and HER2 samples using the RNA-Seq analysis tool.

- Go to: **Toolbox | RNA-Seq analysis** (📁) | **RNA-Seq analysis** (🏠)
- Click the **batch** button below the navigation area window and select the TNBC and HER2 folders (see figure 6).

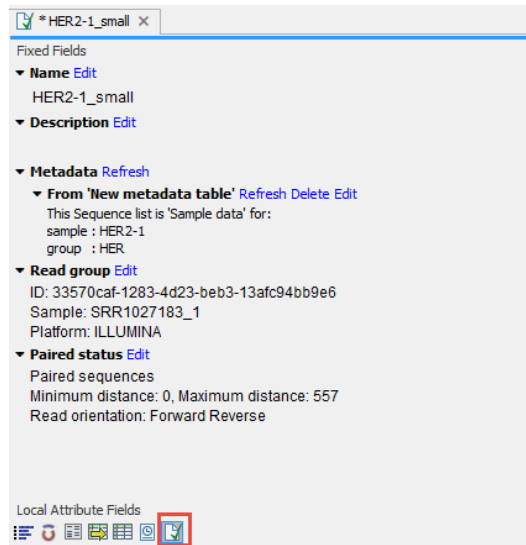


Figure 5: The Element Info view of a sample has now a "Metadata" category showing the association with a metadata table.

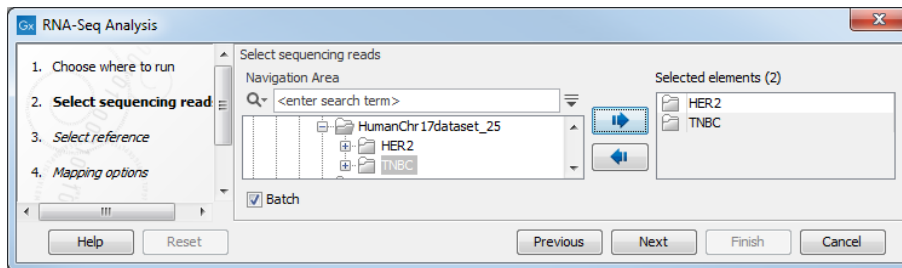


Figure 6: Selecting the TNBC and HER2 samples for RNA-seq analysis.

- This will take you to a dialog where you can see the content of the folders you have selected (figure 7). In some situations you may want to exclude some of the samples found in a folder; this can be done with "Only use elements containing" and "Exclude elements containing" in the lower part of the wizard. In this case we want to use all six samples, so click **Next**.

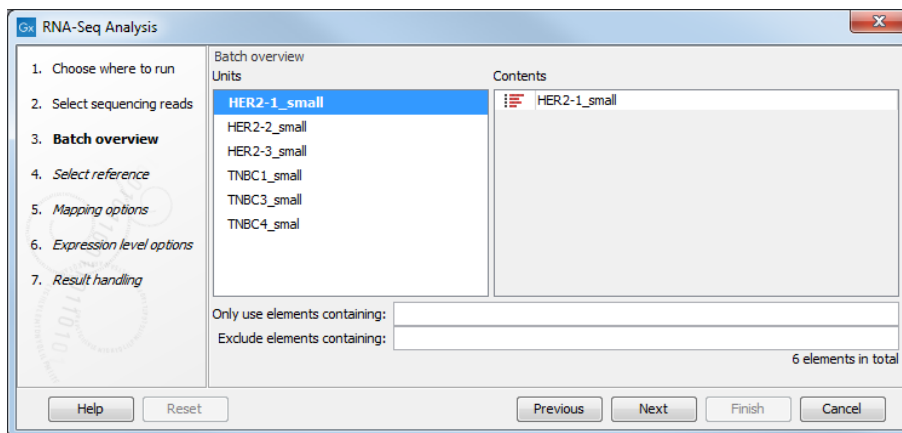


Figure 7: Folders content.

- At this point, we will specify which references to map the read sequences to (figure 8). Note that there are two ways to access the references : in the CLC_References tab,

the `homo_sapiens` subfolder has a succession of role subfolders where you can find the relevant items indicated by `chr_17` appendices. Using the Reference Data tab, you can find all the relevant references in QIAGEN Tutorial | RNA-Seq Analysis of Human Breast Cancer Data. You can choose whichever tab is easiest for you to use to find the reference data.

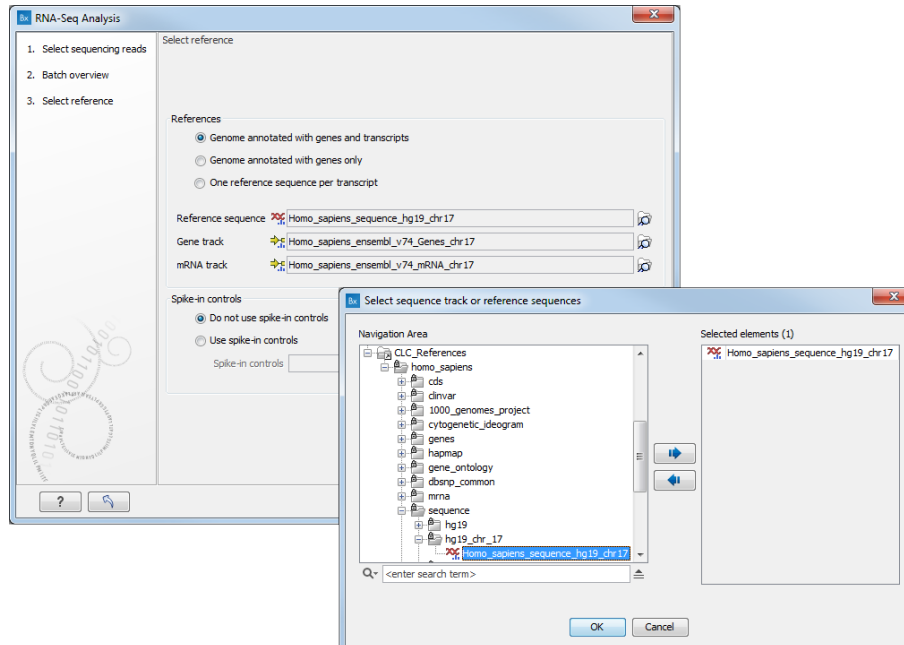


Figure 8: Select reference sequences.

- Choose the **Genome annotated with genes and transcripts** option.
 - For the **Reference sequence**, select **Homo_sapiens_sequence_hg19_chr17**; for gene track, **Homo_sapiens_ensembl_v74_Genes_chr17**; and for mRNA track, **Homo_sapiens_ensembl_v74_mRNA_chr17**.
 - Finally choose the option: **Do not use spike-in controls** and click **Next**.
5. Now the parameters for the mapping of the reads to the reference can be set (figure 9). Leave all the parameters at their default (if you have changed some of the values when running tasks in your workbench and would like to return to the default settings, just click the button **Reset** at the bottom of the wizard view). Click **Next**.

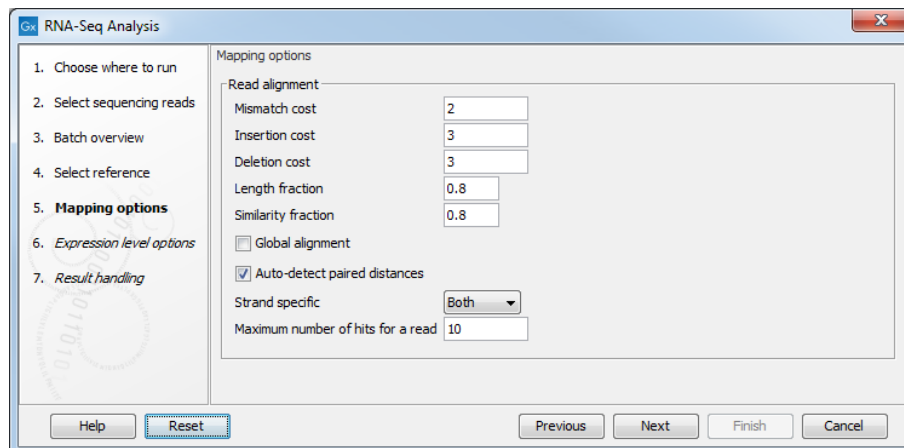


Figure 9: Mapping options.

6. In the Expression level options dialog, the expression value selected is **Total counts** (figure 10). Click **Next**.

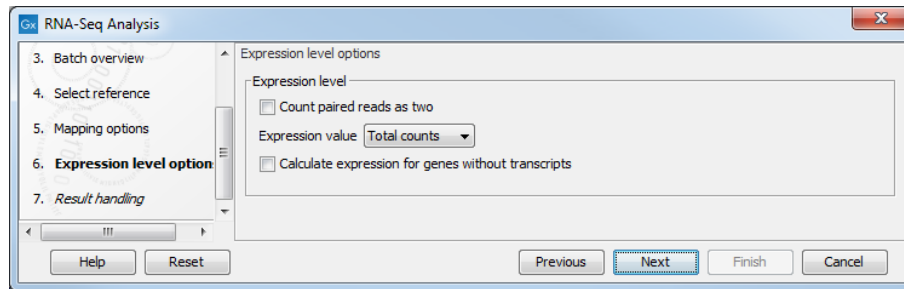


Figure 10: Expression level options.

The options in this wizard step are described here http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Calculating_expression_values_from_RNA_Seq.html.

7. Make sure that the box next to **Create report** is checked, and choose to **Save** your results (as in figure 11 for example) before clicking **Finish**.

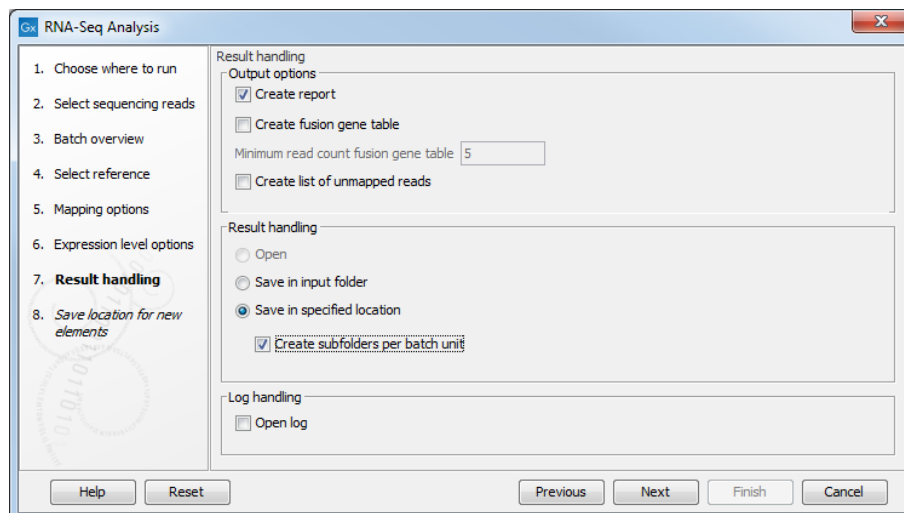


Figure 11: Selecting the output of the RNA-seq analysis.

The RNA-seq algorithm will always produce a reads track. It will also automatically produce gene, transcript, or "reference" level expression tracks, depending on which option was selected as "References" earlier. In this case we have specified both a gene and an mRNA track, and the algorithm will generate both a gene and a transcript level expression track. In addition to these automatically generated tracks, it is possible to get a report, a fusion-gene table, and a list of the unmapped reads.

The RNA-seq analysis is now running. If you want to watch its progress, click on the **Processes tab** in the bottom left side of the Workbench. For complex tasks like RNA-seq analysis, text above the progress bar will let you know what stage is running: Mapping, Counting matches, Building mappings, Finalizing mappings and Saving results. After the RNA-seq analysis is finished, the four output types requested will be available for each sample analyzed (see figure 12).

Open a Gene Expression track, and check the "Element Info" view. You can see that the association from the input sample to the metadata table has carried over to the outputs of the

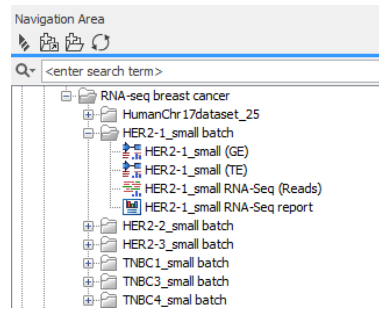


Figure 12: The outputs of the RNA-seq analysis.

RNA-Seq analysis: The tracks are thus associated to the metadata as well.

Ready-to-use Workflows

In this part, we will learn about using a ready-to-use workflow called Identify and Annotate Differentially Expressed Genes and Pathways (see figure 13 to find it in the toolbox). A workflow consists of tools in which the output of one tool is connected and used as the input of another tool. To check the layout, select and then right-click on the name of the workflow and choose to **Open Copy of Workflow**.

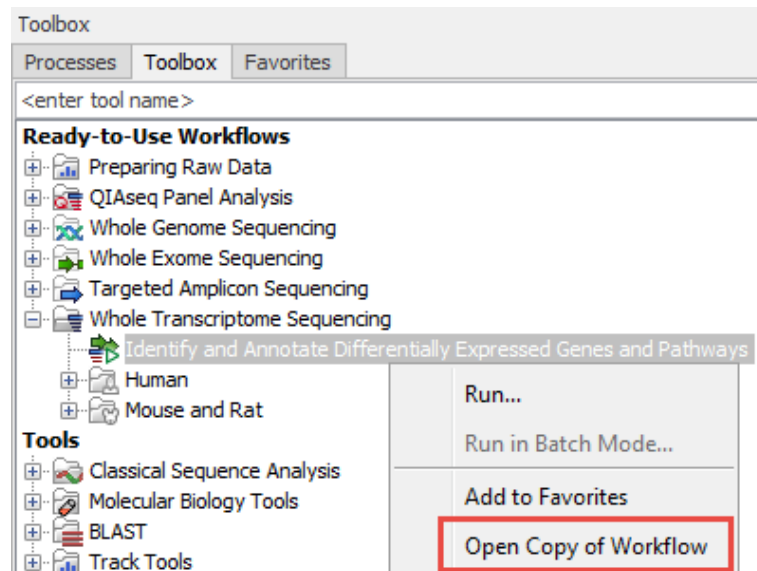



Figure 13: Opening the layout of a workflow.

The inputs to the workflow are shown in green boxes and outputs are shown in blue boxes. In this particular workflow, expression tracks are used as input to run several tools and in turn generate various outputs: a heat map, a PCA plot, an expression browser, statistical comparisons and an enrichment analysis table. Finally, the statistical comparison track along with the reference sequence, mRNA and genes track are used to create a Track List of differentially expressed genes and pathways.

1. Launch the workflow **Identify and Annotate Differentially Expressed Genes and Pathways** .
2. Select all the associated GE tracks generated earlier by the RNA-Seq Analysis (figure 14).

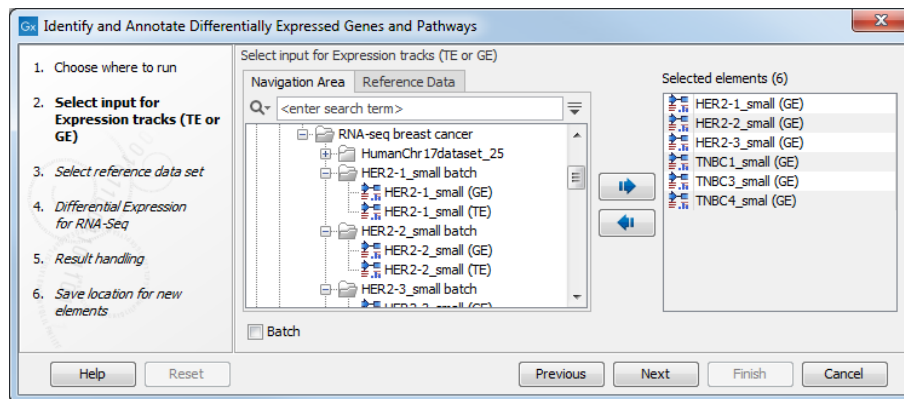


Figure 14: Select GE tracks.

3. Select the tutorial specific Reference Data Set (figure 15).

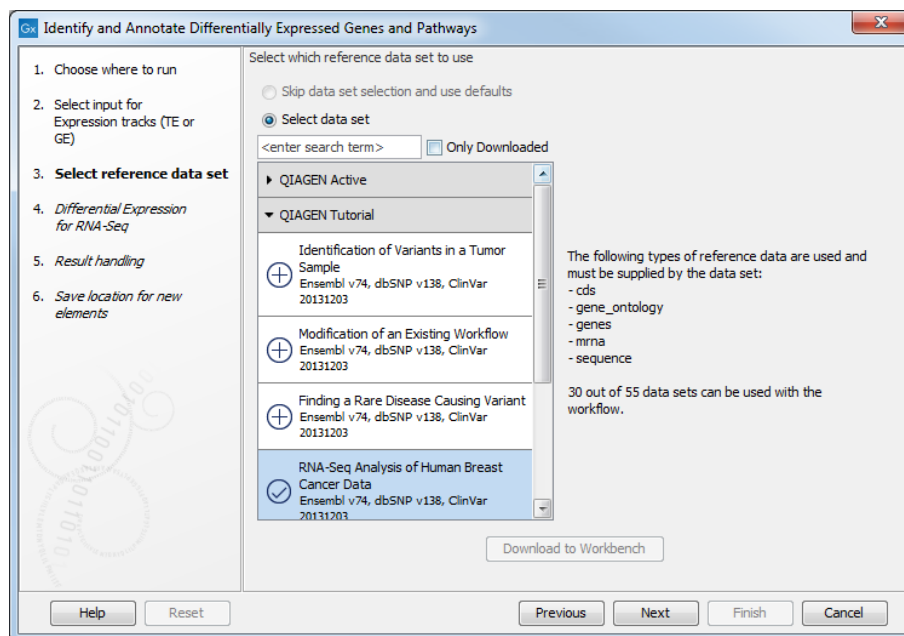


Figure 15: Select the tutorial data set.

4. Set up the Differential Expression as seen in figure 16: select the **New metadata table** created earlier, and choose to "Test differential expression due to" **Group**. Choose to compare samples "Against a control group", and choose **TNBC** as control. Click **Next**.
5. **Save** the results and click **Finish**.

The ready-to-use workflow **Identify and Annotate Differentially Expressed Genes and Pathways** generates the outputs shown in figure 17.

Open the statistical comparison HER vs. TNBC: in the table of differential expression, click on the P-value header to sort entries as in figure 18. Right-click on a cell where the P-value is equal to 0.01, and choose the option "Table filters" followed by "P-value <= 0.01". The table is now displaying only the difference in expression with a P-value lower or equal to 0.01, including ERBB2. Its overexpression is associated with numerous cancer types, including breast cancer.

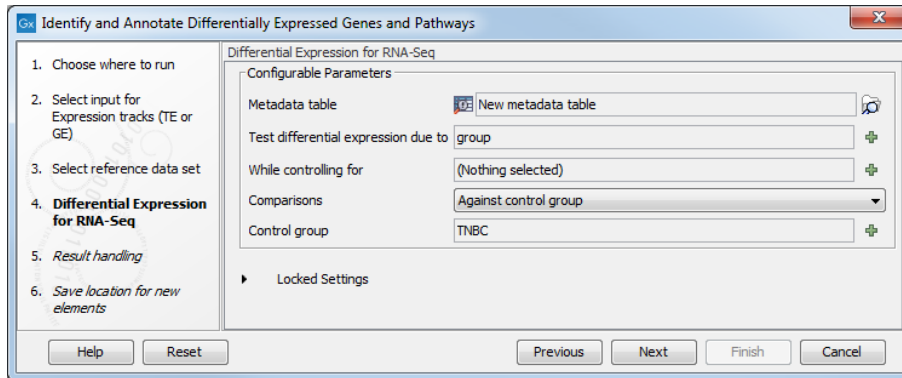


Figure 16: Extract Differentially Expressed Genes.

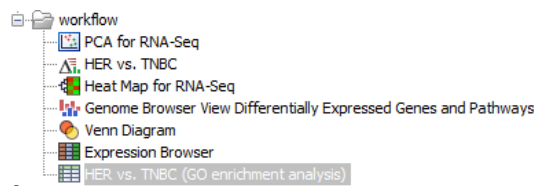


Figure 17: Table of significantly enriched pathways.

Name	Chromosome	Region	Max group mean	Log ₂ fold change	Fold change	P-value	FDR p-v...	Bonferroni
ERBB2	17	37844167..37886679	25,912.48	4.22	18.65	6.21E-7	1.80E-3	1.80E-3
GRB7	17	37894180..37903544	5,113.61	3.58	11.99	5.69E-6	8.25E-3	0.02
KRT42P	17	complement(39782579..39796451)	86.68	-4.78	-27.49	8.32E-5	0.06	0.24
STARD3	17	37793318..37819737	3,665.07	2.78	6.86	8.35E-5	0.06	0.24
STAC2	17	complement(37366789..37382125)	2,145.41	-3.85	-14.38	1.51E-4	0.09	0.44
HNF1B	17	complement(36046435..36105237)	26.25	-4.36	-20.48	9.11E-4	0.44	1.00
THRA	17	38214543..38250120	674.57	-1.57	-2.96	2.04E-3	0.84	1.00
KRT13	17	complement(39657233..39661957)	68.34	-3.14	-8.81	2.64E-3	0.96	1.00
AC003958.2	17	39558668..39581301	59.09	-6.58	-95.69	3.98E-3	0.98	1.00
KRT15	17	complement(39669995..39678781)	7,031.14	-3.90	-14.98	4.17E-3	0.98	1.00
CTB-131K11.1	17	37558046..37562486	3,830.58	2.15	4.43	4.45E-3	0.98	1.00
LINC00974	17	complement(39705858..39710747)	76.58	-5.96	-62.45	4.88E-3	0.98	1.00
IKZF3	17	complement(37921198..38020441)	1,192.99	1.55	2.94	8.74E-3	0.98	1.00
MIEN1	17	complement(37884749..37887040)	11,240.07	2.71	6.52	0.01	0.98	1.00
KRT9	17	complement(39722096..39728310)	114.69	-2.52	-5.74			
C17orf96	17	complement(36827961..36831187)	625.65	1.79	3.45			
TMEM99	17	38975358..38992522	231.15	1.34	2.53			
TCAP	17	37820440..37822808	164.05	2.32	4.98			
AC124789.1	17	complement(36606638..36608688)	59.00	-4.81	-28.01			
GJD3	17	complement(38517235..38520067)	112.31	-5.35	-40.92			
KRTAP16-1	17	complement(39463952..39465505)	44.54	-4.81	-28.00			
ZFPBP2	17	38024417..38034149	29.62	-3.22	-9.35			
KRT222_2	17	complement(38810917..38821433)	23.32	-4.91	-30.05			
LRRC3C	17	38097727..38101000	78.57	-3.01	-8.08	0.03		
PPP1R1B	17	37782993..37792879	8,905.65	2.58	5.99	0.03		

Figure 18: Filtering results from the gene expression analysis.

Open the Track List, then double click on the **Statistical Comparison** track name to open the corresponding table below the Track List view. Select for the gene of interest (for example ERBB2) and the Track List will automatically **zoom to selection**. At this zoom level you can see the individual reads mapped. Reads that are mapped in pairs are represented by a bold line, and the individual reads in the pair are connected by a thin line. For some reads a thin dotted line connects each end of the read. These are the reads that have been mapped across exon-exon boundaries.

Open in turn the other outputs. While a Venn diagram does not make a lot of sense in case of a two groups comparison, the PCA plot (2D and 3D) and the heat map provide interesting visualization options for understanding the data.



Tutorial
