



Tutorial

Quantify gene expression using QIAseq UPX 3' panels

May 12, 2021

Sample to Insight

Quantify gene expression using QIAseq UPX 3' panels

This tutorial uses the capabilities of *CLC Genomics Workbench* and the Biomedical Genomics Analysis plugin to analyze sequences generated using the QIAseq UPX 3' Transcriptome Kit. The primary focus is illustrating how to launch analyses, carry out quality control, and identify differentially expressed genes.

In more detail, during this tutorial, we will:

- Make use of the Reference Data Manager to download reference data.
- Import and use metadata, to aid in organization of data, efficient selection of relevant data elements when launching analyses, and to define information about an experiment relevant to differential expression analysis.
- Analyze QIAseq UPX 3' panel data, making use of the **Analyze QIAseq Panels** tool.
- Carry out quality control checks of results generated using the Human UPX 3' analysis workflow.
- Run expression and visualization analyses using a workflow launched from the Workflow Editor.
- Review downstream expression analysis results and plots.
- Extracting sequence IDs of interest from expression results to use in third party, online tools.
- Run GO enrichment analysis in the *CLC Genomics Workbench*.

Steps where action should be taken are numbered.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 20.0 (or higher) with the Biomedical Genomics Analysis plugin installed. How to install plugins is described [in the CLC Genomics Workbench manual](#).

General tips

- Within wizard windows you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or by selecting the "Help" option under the "Help" menu.
- If you are connected to a *CLC Genomics Server* via your *CLC Genomics Workbench*, we recommend that you analyses on the *CLC Genomics Server* when you are offered this option.
- Information about running analyses in batch mode is available in the [the CLC Genomics Workbench manual](#).

Download and import data for this tutorial

Download the sample data

For this tutorial, we use datasets derived from single cell libraries. While single cell datasets are not the best fit for the analysis undertaken here, the relatively small size makes them suitable as tutorial example data. It is unknown whether these samples are from the same cell type sequenced under different conditions or from 3 different cell types. For the type of analysis covered here, whole transcriptome data, potentially from ultra-low input, would generally be expected as input.

The library and barcode (well-identifiers) for the tutorial sample data are listed below. We will later import an expanded form of this information as metadata for use throughout this tutorial.

Library	Barcode
A	C1-C10
B	C13-C22
C	C25-C34

1. Download the sample data from our [website](#).

The data set includes a file containing the sequencing reads, an excel file containing information about the barcodes and library each barcode is associated with, and a workflow for running expression analyses using samples that pass quality control checks.

2. Unzip the data file to a location of your choice.

Import the sample reads and expression analysis workflow

3. Start the CLC Workbench if it is not already running.
4. Import the sample reads and expression analysis workflow, which are in CLC format, by going to:

File | Import (📁) | **Standard Import** (📁).

5. From among the files just unzipped, choose those with names ending in ".clc". Leave the import type set to "Automatic", as shown in figure 1.

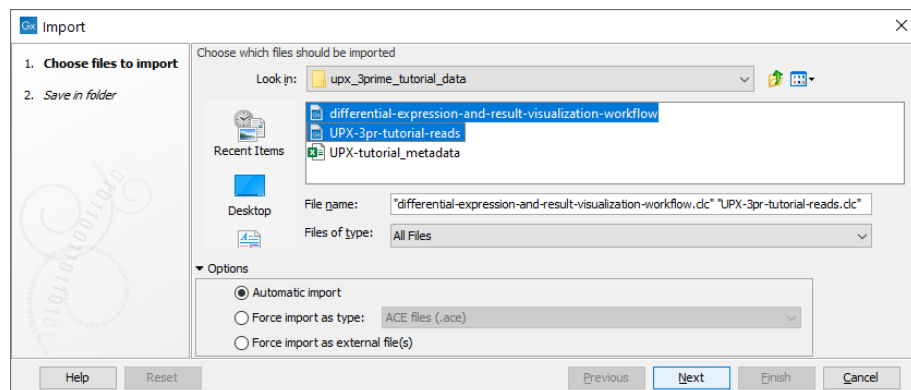


Figure 1: Standard import of ".clc" files

6. Click on **Next** and select a folder to save to.

A new folder can be created by clicking on the "New Folder" option at the top of the window.

Once the import is completed, you should see the folder and files in the Navigation Area as shown in figure 2.

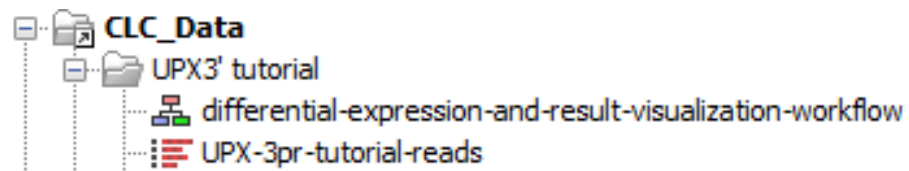


Figure 2: Data imported into selected folder in the navigation area

Download the reference data

The "QIAseq UPX 3' Panels hg38" QIAGEN Reference Data Library is needed for analyses we carry out in this tutorial. In addition, as we will run a GO enrichment analysis, we also need a corresponding GO data track. In this section, we download these using the **Reference Data Manager**.

7. Click on the **References** button in the top toolbar.

8. Select the **QIAGEN Reference Data Library** tab, and then select the **QIAseq UPX 3' Panels hg38** item under Reference Data Sets, on the left hand side.

9. Click on the **Download** button on the right, above the list of data elements (figure 3).

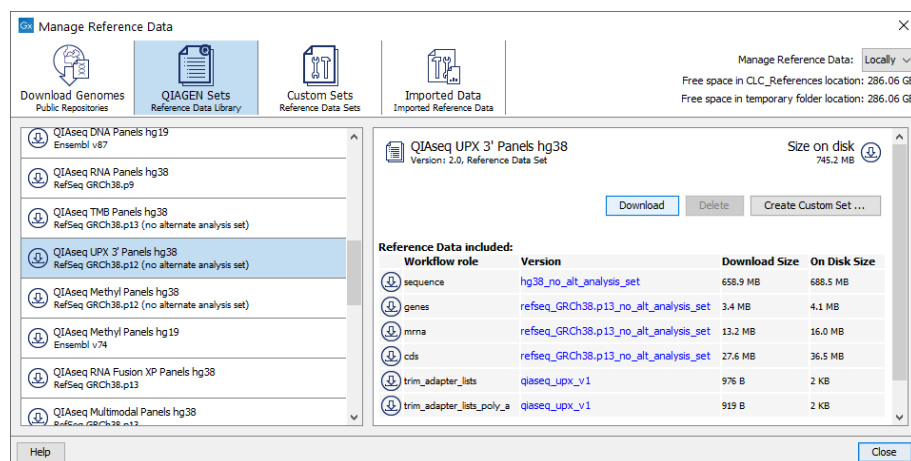


Figure 3: Reference data to be downloaded

If the "Download" button is not enabled, and you see checkmarks by each data element listed, then you already have the reference data in this set.



10. Now click on the Reference Data Elements option in the listing on the left hand side.

11. Scroll to the option "Gene Ontology 20190201_no_alt_analysis_set" and select it.

12. Click on the **Download** button on the right, which will be enabled if you do not already have access to this data element.
13. When you are done, close the **Reference Data Manager**.

Demultiplex the sample data

The sample reads are in a single sequence list. We now demultiplex this list, to give us separate sequence lists for each well.

14. Open the Analyze QIAseq Panels from the Toolbox:
Ready-to-Use Workflows | QIAseq Panel Analysis  | **Analyze QIAseq Panels** 
15. Open the **UPX 3' RNA** tab.
16. Click on **Run** next to the Demultiplex option (figure 4).

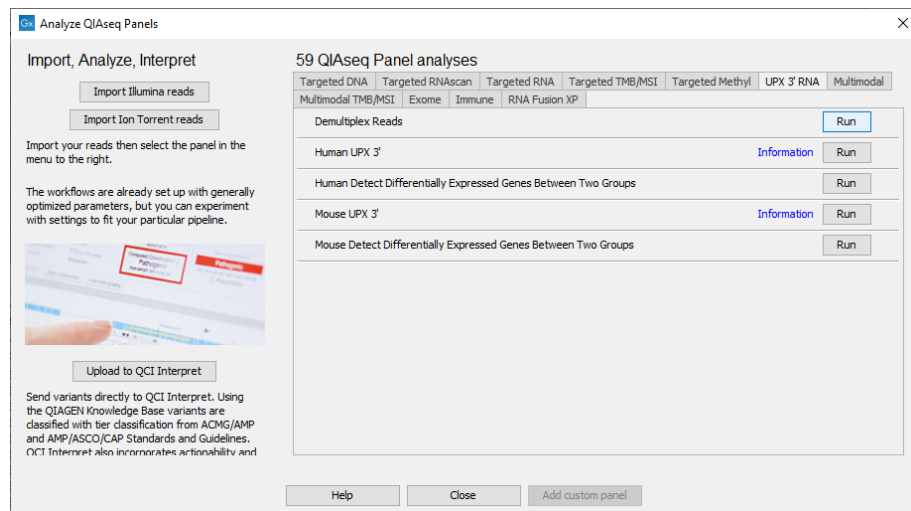


Figure 4: The Demultiplex Reads tool is available under the UPX 3' RNA tab

17. Select the reads you imported earlier.
18. Click on **Next**.

The well picker looks at the construction of the sequencing reads and preselects the sequencing platform and the plate size. It identifies the wells used in the experiment by looking at the barcodes present in the first 10000 reads. These wells are colored in the image of the plate (figure 5).

19. Select **All Mismatches**, so that one mismatch is allowed in the UMI barcodes.

We know that the sample also includes wells C6 and C7 in the design, so we specify manually that these wells were also used.

20. Click on **Select Detected**.

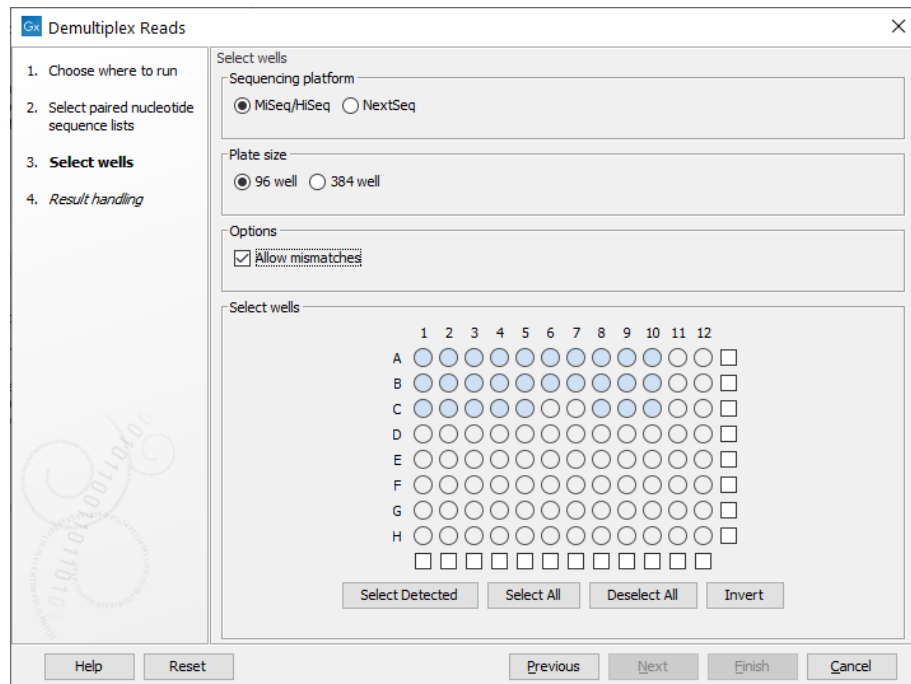


Figure 5: Allow mismatches

21. Click on wells C6 and C7 to select them.
22. Click on **Next**.
23. Leave the output options set to the defaults, choose to **Save** the results and then click on **Next**.
24. Specify a new folder to save the results to.
25. Click on **Finish**.
26. Close the **Analyze QIAseq Panels** tool.

It may take a few minutes for the demultiplexing task to complete. You can follow its progress by opening the **Processes** tab at the bottom left of the *CLC Genomics Workbench*.

Review the demultiplexed data

In the folder the results were saved to, you should find 30 sequence lists with names that include a well number, a sequence list containing reads that did not have the expected barcodes and therefore could not be grouped, and a report.

27. Go to the Navigation Area and open the Demultiplex Reads report.

The number of reads and the percentage of reads are reported for each barcode. The percentages are also displayed in a histogram. (figure 6). There are substantial differences between the numbers of reads for each barcode.

For context, recall that library A consists of C1 to C10, library B consists of C13 to C22 and library C consists of C25 to C34.

1.2 Reads per barcode

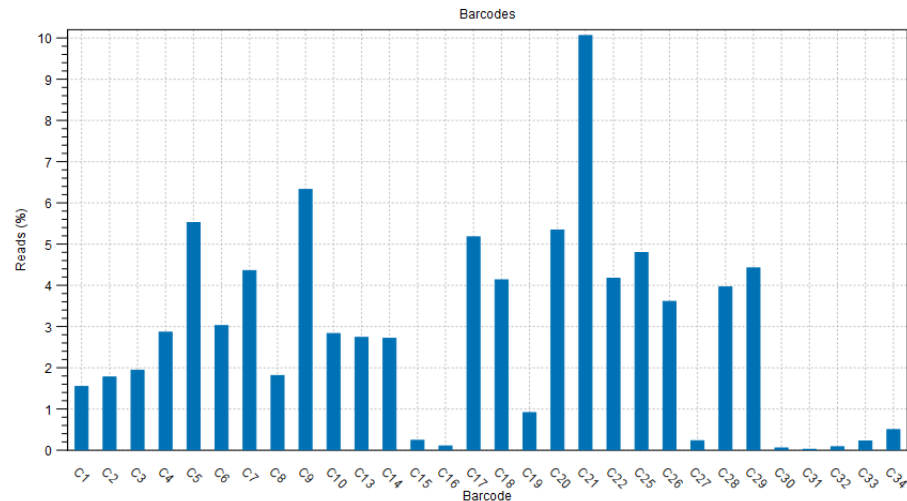


Figure 6: Histogram from the report showing reads per barcodes

Import the metadata and make associations to it

We now import information about the source library for each well from an Excel file. We then associate each of the sequence lists just created with a relevant row of the resulting metadata table.

After these associations are in place, results from analyses that use these sequence lists will inherit the relevant metadata association. This can serve multiple purposes, such as:

- Serving as a useful record.
- Aiding us in locating and efficiently selecting data for downstream analyses.
- Providing the experimental information required for differential expression analyses.

1. Import the metadata by going to:

File | Import (📁) | **Import Metadata** (📊)

2. Choose the file "UPX-tutorial_metadata.xlsx" from the tutorial data you unzipped earlier and click on **Next**.

3. Browse for and select the sequence lists that were just created by the demultiplexing job (figure 7).

4. Select the option Suffix, to indicate the term to match on is the final term of the sequence list names.

In the Data association preview pane, you should now see that all 30 elements have been associated with the metadata we are importing.

Tutorial

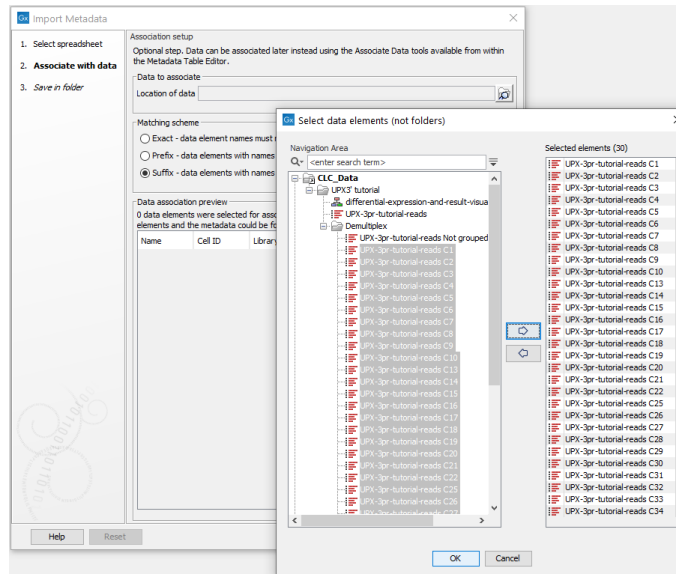


Figure 7: Select the data that should have associations to the metadata.

5. Click on **Next**.
6. Choose a location to save the metadata table to.

Note: If you are using *CLC Genomics Workbench 20.x* the newly created metadata table is given the default name "Samples". Rename to "UPX-tutorial_metadata".

Run the Human UPX 3' workflow to analyze the samples

We now analyze the data using the Human UPX 3' workflow. We will launch the workflow in batch mode, such that the analysis will be run once for each sequence list supplied as input.

For each sample, the workflow will produce QC related reports, a Gene Expression track, an RNA-Seq Report and a Combined Report, containing a summary of information from the other reports.

Further information about this workflow can be found in the [Biomedical Genomics Analysis plugin manual](#).



While it is reasonably straightforward to find the input data in the Navigation Area, we will use the metadata table to aid in selecting the input data. This method is particularly useful for some later analyses, when only a subset of the samples are used as input.

1. Open the metadata table called "UPX-tutorial_metadata".
2. Select all the rows and click on **Find Associated Data**.

You should now see the sequence lists we wish to analyze in a panel below the metadata.

3. Select all the sequence lists in the bottom panel and click on **Find in Navigation Area**.

The data elements to analyze are now pre-selected.

4. Open the Analyze QIAseq Panels from the Toolbox:
Ready-to-Use Workflows | QIAseq Panel Analysis  | **Analyze QIAseq Panels** 
5. Open the **UPX 3' RNA** tab.
6. Click on **Run** next to the Human UPX 3' option (figure 8).

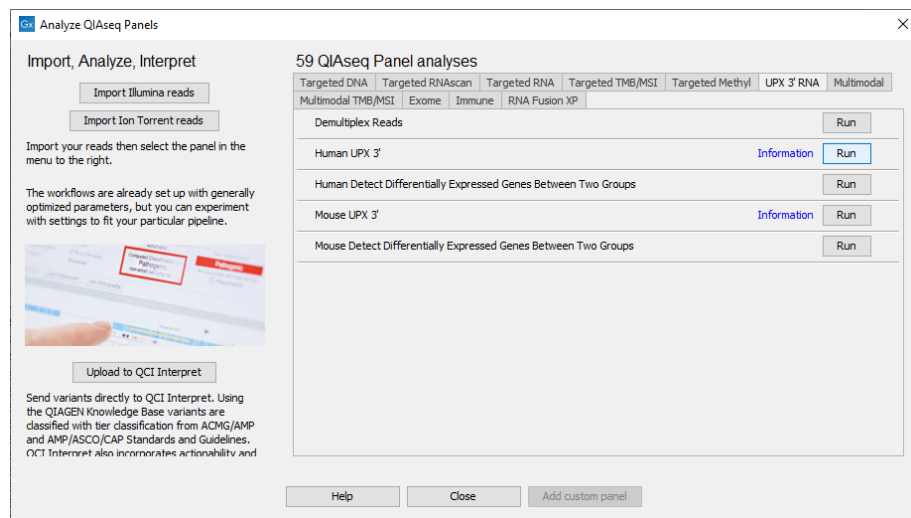


Figure 8: Run the human UPX 3' from Analyze QIAseq Panels

The pre-selected sequence lists should be present in the Selected elements list on the right hand side.

We will chose to run the workflow in batch mode, where the workflow is run once per sequence list.

7. Check the **Batch** option below the data selection area (figure 9).
8. Click on **Next**.
9. Specify a new folder to save the results to and click on **Finish**.

It will take some time for the analysis to run. You can follow its progress by opening the Processes tab at the bottom left of the *CLC Genomics Workbench*.

In our hands, the full analysis took approximately 1 hour on a 2014 MacBook Pro with 16 GB RAM. However, the results are saved per run. So after the first sequence list has been analysed, you can review the results without waiting for all the analyses to complete.

When an analysis is complete, you should see results saved, as shown in figure 10.

Investigating the results of the Human UPX 3' workflow

We recommend using both Combined Reports of QC results as well as PCA plots for assessing sample quality when analyzing deeply sequenced whole transcriptome sequencing data.

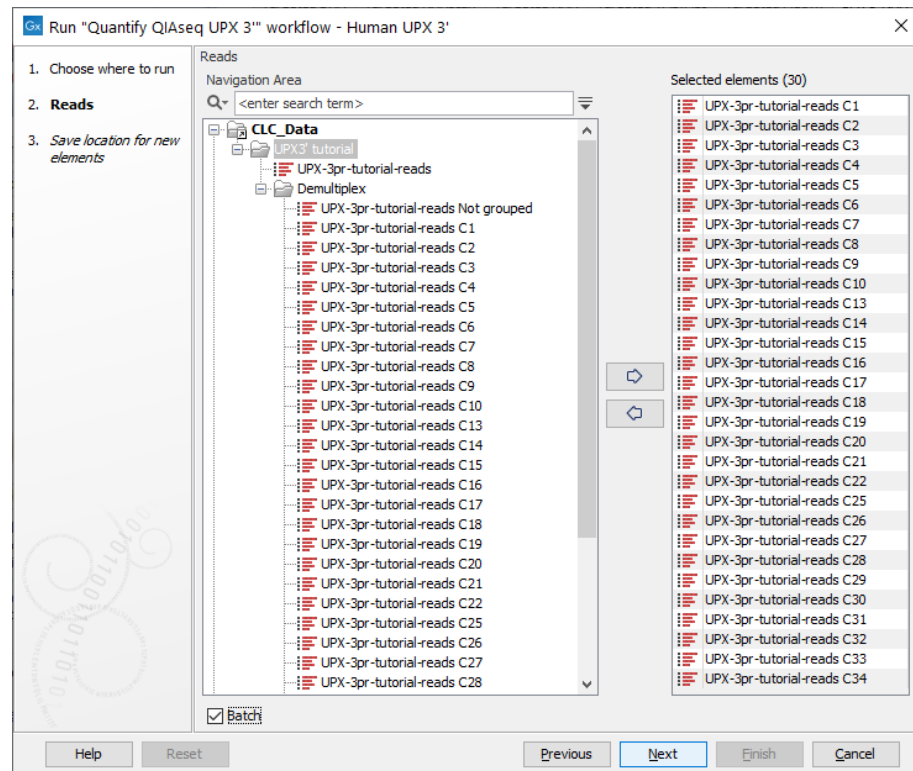


Figure 9: Reads to be analyzed in batch mode

Below we generate both of these for these single cell data sets and review the results to decide which samples to keep for downstream processing.

Inspecting a single RNA-Seq report

1. Click on a folder of results for one of the samples in the Navigation Area to expand its contents.
2. Open the RNA-seq report in that folder.

In Section 1.1 Sequence Reads, the entry in the "Paired" column is "no". The QIAseq UPX 3' kits are designed such that only R1 of each pair contains biological signal. Thus the data is analyzed as single reads as R2 is processed and removed prior to RNA-Seq Analysis.

In section 4 Mapping statistics, the Single reads table provides information on the number and percentage of reads mapped to the reference genome. This value is reduced when spike-ins or phiX are present.

In section 7 Transcript length coverage, it is possible to see the expected 3' bias of this type of sequencing, where the polyA tail is targeted for capture. The peak in normalized transcript coverage is situated on the right of the plot.

Inspecting a summarized report for a single sample

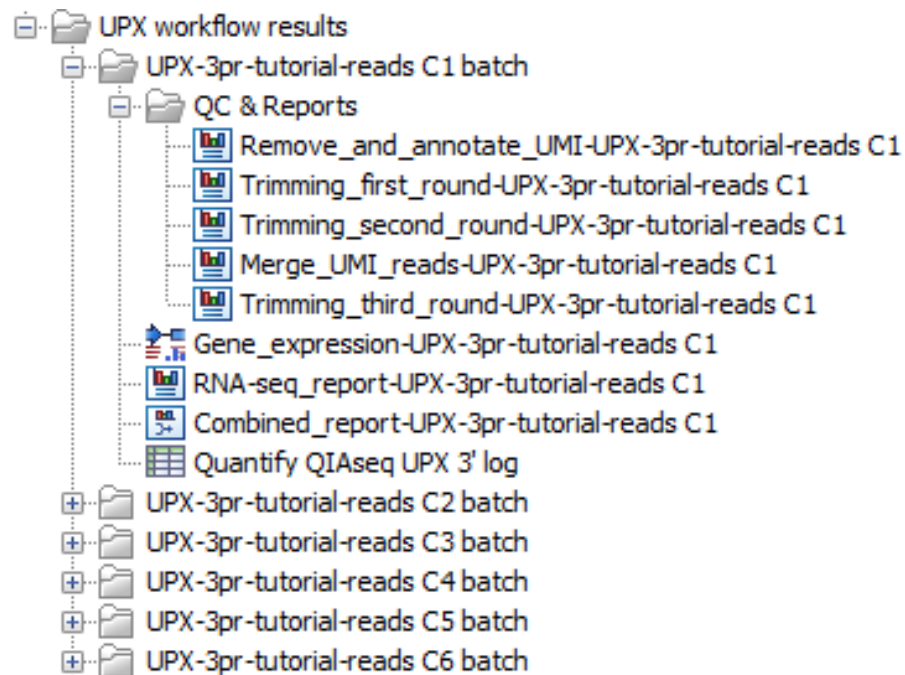


Figure 10: The output from the workflow are saved in one folder for each sample

Various tools have been run in the workflow, many of them generating reports. The contents of those reports were then summarized in a single "Combined report" for that sample.

3. Click on a folder of results for one of the samples in the Navigation Area to expand its contents.
4. Open the report in that folder with a name starting with "Combined_Report".

There are multiple tables in the report, providing information about this particular sample. **Cells colored pink** indicate there is a potential problem. Information under the table states the criteria used for particular columns to decide if a cell should be colored pink.

Having looked at two reports for a single sample, it is clear that it would be very time-consuming to inspect the various reports for each sample individually, and difficult to look for outliers that way. Instead, we can combine the information from these reports into a single, summary report and review that.

Creating a combined report for all the samples

5. To create a Combined Report, summarizing the QC results for each sample, go to:
Toolbox | Quality Control (📄) | Combine Reports (📄)
6. Search for the term "Combined_report-UPX-3pr-tutorial-reads" using the search field above the Navigation Area on the left hand side, as shown in figure 11.
7. Select these reports and click on **Next**.

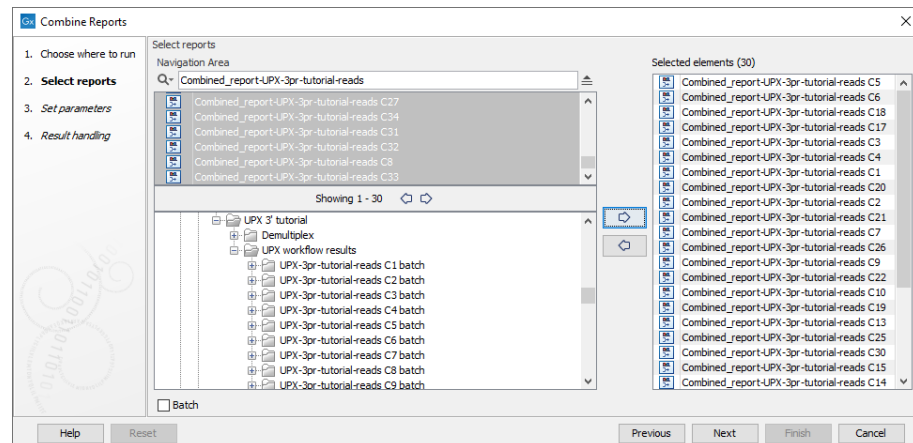


Figure 11: Select reports to be combined

8. Check the **Include tables for outliers** option and click on **Next**.
9. Choose to **Save** the results and click on **Next**.
10. Specify the location to save the combined report to and click on **Finish**.

Inspecting the cross-sample combined report

11. Find the combined report you just created in the Navigation Area and open it.
12. Look for the yellow highlighted squares in Table 1.1, where information about the UMI sequence annotation and removal is presented. These are outliers: their value lies outside the range: lower quartile - 1.5 IQR, upper quartile + 1.5 IQR.

The individual samples identified as outliers are listed in the table for outliers, just under the main table.

13. Scroll down to section 5 RNA-Seq Analysis.

In table 5.1 Read count statistics, some cells are colored pink, indicating potential problems. It can also be seen that several samples are highlighted as outliers in multiple tables. Issues highlighted include the read length distribution, short lengths after trimming, lower than average Q scores, and high proportions of reads not been mapped.

The cause of these problems is most likely too little RNA input. It is likely that the input has been too fragmented and most of what is sequenced is phiX and ERCC spike-ins.

We will bear these things in mind as we consider further which samples should be included in the analysis.

Visualizing library effects using a PCA plot

We will use the **PCA for RNA-Seq** tool to look at library effects.

As we did earlier, we will pre-select the data we wish to analyze, making use of the metadata table.

14. Open the metadata table called "UPX-tutorial_metadata".
15. Select all the rows and click on **Find Associated Data**.

In addition to the original sequence lists, we now see that results of the Human UPX 3' workflow also have associations to the metadata. We will use the advanced filtering functionality to list just the data we wish to use as input to the **PCA for RNA-Seq** tool.

16. Click on the small arrow beside the Filter button in the bottom pane, to see the advanced filtering options.
17. Choose Role from the first drop down menu, then = from the second, and type in "Gene expression" in the text field (figure 12).

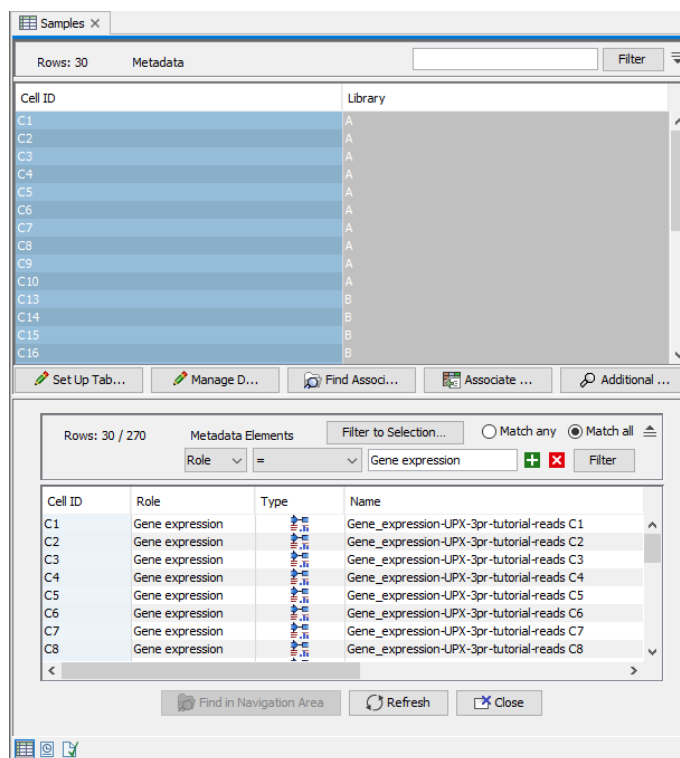




Figure 12: Select the data that should have associations to the metadata.

18. You should now see all the gene expression tracks we wish to use as input to the **PCA for RNA-Seq** tool.
19. Select all rows and click on **Find in Navigation Area**.
20. Go to:

Toolbox | RNA-Seq and Small RNA Analysis () | **PCA for RNA-Seq** ()
21. You should see all the gene expression tracks of interest already selected for this analysis.

22. Click on **Next**.
23. Choose to Open the results and click on **Finish**.

The PCA plot opens when the analysis is done. By default it will open in a 2D view. By clicking on the second small icon at the bottom (Show Principal Component Plot (3d)), you can view it in 3D. In both the 2D and 3D view, you can choose to look at different components using option in the right hand panel.

Samples are colored according to their source library. This information is taken from the "UPX-tutorial_metadata" table. It is available because the expression tracks used are associated with rows in this metadata table.

It is noticeable that C21 is distant from all the other samples.

Based on our review of the QC reports and the PCA plot, we will not include sample C21 or any samples with a read count below 10000 reads (C15, C16, C27, C30, C31, C32 and C33) in our downstream analyses.

Note that there seem to be quite a few problems with many samples in Library C. Under normal circumstances, this might lead us to further investigate the quality of this library before including any samples from it. For the purposes of this tutorial however, we will proceed with some of the samples from library C.

Using metadata to support easy and error free data selection

There are various ways you could select just the datasets you are interested in analyzing. Here we will add a column to the metadata table, recording which samples passed our quality control checks. This provides us with a record of this information, as well as making it easy to select the relevant data elements for downstream analysis.

This section is optional. You are welcome to skip this step and, instead, manually select the elements to analyze in later tutorial steps, instead of using the metadata table to facilitate data selection.

Add quality check information to the metadata table

24. Open the metadata table called "UPX-tutorial_metadata".
25. Add a new column to the metadata table:
 - (a) Click on **Set up Table...** You are initially viewing information about the first column in the table, "Cell ID".
 - (b) Click on the **Next** button on the right hand side of this dialog. You are now viewing information about the second column in the table "Library".
 - (c) Make a new column to the right of Library by clicking on the middle icon on the right hand side.
 - (d) Give it the name "Passed quality checks" and leave the data type as Text.
 - (e) Click on **Done**.

Now the column is in place, we will enter information for each sample indicating if it passed or failed our quality control checks.

26. Click on **Manage Data**.

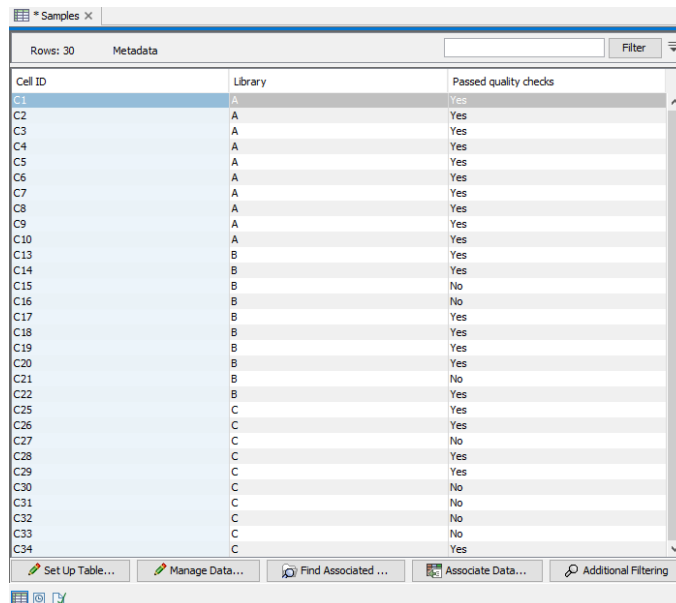
You are now viewing the information in the first row of the metadata.

27. Enter "Yes" into the "Passed quality checks" field for C1 and then click on **Next** in the right hand area.

28. Step through each row of the table entering "Yes" or "No" as relevant for the given row. Enter "No" for the rows with Cell IDs C15, C16, C21, C27, C30, C31, C32 and C33.

29. After information is entered in in the "Passed quality checks" field for all rows, click on **Done**.

The "UPX-tutorial_metadata" metadata table should now look like that shown in figure 13.



Cell ID	Library	Passed quality checks
C1	A	Yes
C2	A	Yes
C3	A	Yes
C4	A	Yes
C5	A	Yes
C6	A	Yes
C7	A	Yes
C8	A	Yes
C9	A	Yes
C10	A	Yes
C13	B	Yes
C14	B	Yes
C15	B	No
C16	B	No
C17	B	Yes
C18	B	Yes
C19	B	Yes
C20	B	Yes
C21	B	No
C22	B	Yes
C25	C	Yes
C26	C	Yes
C27	C	No
C28	C	Yes
C29	C	Yes
C30	C	No
C31	C	No
C32	C	No
C33	C	No
C34	C	Yes

Figure 13: The metadata now includes information about the results of the quality checks carried out.

For the analyses we carry out from this point, we can make use of this new information to preselect as input only the samples that have passed the quality checks.

Preselect gene expression tracks for samples that passed quality checks

30. Click on the small arrow beside the Filter button in the top pane, to expose the advanced filtering options.

31. Choose "Passed quality checks" from the first drop down menu, then = from the second, and type in "Yes" in the text field.

32. Select the 22 rows and click on **Find Associated Data**.

In the bottom pane, the data associated with samples that passed our quality checks are listed. We wish to restrict that list to just Gene Expression tracks, so we filter so that only these are listed.

33. Click on the small arrow beside the Filter button in the bottom pane, to expose the advanced filtering options, if they are not already visible.
34. Choose Role from the first drop down menu, then = from the second, and type in "Gene expression" in the text field.
35. Select all the Gene Expression tracks listed in the bottom panel and click on **Find in Navigation Area**.

You have now pre-selected exactly the data you wish to use as input in the next analysis we run.

While this seems a bit long winded to set up, it will save time when launching analyses using these data elements, and it is less error prone than manually selecting the intended inputs.

Differential expression analysis and plots of expression data

In this section, we use the expression data from samples that passed our quality control checks in a differential expression analysis, and we generate plots to help evaluate the expression results. For this section, we will use a small workflow. This will allow us to reproducibly launch the analyses quickly and easily.

View and run the expression analysis workflow

1. Open the differential-expression-and-result-visualization-workflow, which you imported earlier.

This opens the workflow in the Workflow Editor. Here, you can view and edit the configuration of the workflow elements, as well as launch the workflow.

The workflow has 4 analysis elements and results are saved from all of these. The results of the Across groups (ANOVA-like) differential expression analysis is also being used as input to the heatmap.

2. Double click on each of the workflow elements and click through the wizards to view the settings.

Only the options that have an unlocked symbol beside them will be offered to you to configure when you launch the workflow.

3. Click on the **Run** button on the bottom right to launch the workflow.
4. The Gene Expression tracks for all samples except C15, C16, C21, C27, C30, C31, C32 and C33 should be selected for use.

If you are preselecting data using the metadata related functionality described above, then these should already be selected as inputs, although you may have to click on the arrow pointing right to put these 22 elements in the right hand, Selected elements, area.

If you are not preselecting data using the metadata related functionality, then you could search for the term "Gene_expression-Tutorial" using the search field above the Navigation Area on the left hand side, and select the Gene Expression tracks for only the samples we wish to include in the analysis. This is all the samples except for C15, C16, C27, C30, C31, C32 and C33.

In the following steps, you are prompted to configure options for some steps of the workflow.

5. To configure the Pairwise differential expression step:
 - (a) For the Metadata table field, browse for the "Samples" metadata table and select it.
 - (b) For the **Test Differential Expression due to** option, click on the + symbol, select "Library" and move it to the right hand, Selected area.
 - (c) For the comparison type, "All groups pairs" is already selected.
 - (d) Click on **Next**.
6. To configure the Across groups (ANOVA-like) differential expression step:
 - (a) For the Metadata table field, browse for the "Samples" metadata table and select it.
 - (b) For the **Test Differential Expression due to** option, click on the + symbol, select "Library" and move it to the right hand, Selected area.
 - (c) For the comparison type, "Across groups" is already selected.
 - (d) Click on **Next**.
7. Leave the Result handling options as the defaults and click on **Next**.
8. Create a new folder to save the results to and click on **Finish**.

The outputs from the workflow are 3 pairwise statistical comparison tracks, an across group comparison statistical track, a heatmap, and a Venn Diagram. There is also a Workflow Metadata Result table, which lists these items. For a small set of outputs, this table is not that informative, but it can be very useful when running large workflows or batches of workflows, with many outputs.

Reviewing the results

The output of the pairwise differential expression analysis is 3 Statistical Comparison tracks, one for each of the 3 pairwise comparisons.

If you open these files and review them, you would find that different numbers of genes are differentially expressed between the 3 groups.

The Venn Diagram contains information from the 3 statistical comparison tracks containing the pairwise differential expression results. That diagram allows further investigation of the cross-library expression results.

9. Open the Venn Diagram output by the differential-expression-and-result-visualization-workflow

It can be seen that different numbers of genes are differentially expressed between the 3 groups, with no genes being identified as differentially expressed across all 3 comparisons.

The heatmap is another aid for investigating the groups of differential expressed genes.

10. Open the Heat Map for RNA-seq output by the differential-expression-and-result-visualization-workflow, which can be found in the Expression Analysis output folder
11. Select the option "Library" under "Metadata layer #1" in the Heat Map Settings panel on the right.

The Libraries are now indicated by color on the sample tree. The individual libraries cluster closely together, which is expected.

12. Zoom in to see the gene names.
13. Remove the sample names from the tree by unticking **Show names above** option in the Samples tab of the Heat Map Settings panel.
14. You can select genes of interest in the heatmap view. Right-click over the selected area and choose the **Copy Names to Clipboard** option.
15. Open a web browser and go to <http://geneontology.org/> and paste the gene names from the clipboard into the box at the top right that prompts for gene ids.

GO Enrichment Analysis

We now look for over-represented or under-represented GO terms in the differential expression results by running a GO enrichment analysis. The analysis will be run on each pair-based differential analysis table in turn, by launching it in batch mode.

You need GO reference data for this section. If you have not already done so, please download this as described near the start of this tutorial.

1. Go to:
Toolbox | Expression Analysis (📁) | Gene Set Test (📊)
2. Check the **Batch** option below the list of data elements.
3. Select the 3 pairwise statistical comparison tracks as input and click on **Next**.
4. Review the batch units in the Batch overview step and then click on **Next**.
5. Select the GO ontology reference data by clicking the Browse button and searching for "goa" in the Navigation Area. Select the GO reference data element called "goa_human_20190201_no_alt_analysis_set".

6. Keep the default settings, which should mean the following are selected: "GO biological processes", "Exclude computationally inferred GO terms" and "Ignore gene name capitalization".
7. Click on **Next**.
8. Keep the default filtering parameter settings and click on **Next**.
9. Choose to "Save in input folder" and click on **Finish**.