Tutorial

Detect Fusions and Actionable Variants Using QIAseq Multimodal Panels

February 25, 2022

Detect Fusions and Actionable Variants Using QIAseq Multimodal Panels

This tutorial uses the capabilities of CLC Genomics Workbench and Biomedical Genomics Analysis plugin to detect fusion events and find actionable variants in reads generated on a custom QIAseq Multimodal panel using SeraCare Seraseq® Myeloid DNA and RNA Fusion Reference Materials.

This tutorial covers the following:

- Importing Illumina paired reads using the on the fly importer during the workflow execution.

- Finding fusion events and low frequency variants using the Perform QIAseq Multimodal Analysis (Illumina) template workflow.

- (Optionally) uploading fusion events and variants to **QCI Interpret**.

Prerequisites

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

For the optional section, "Upload your results to QCI Interpret", you will need an active QCI Interpret subscription (not a trial license).

General tips

- Within the wizard windows presented when launching tools or workflows, 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.

Download the tutorial data

1. Download the sample data from our website and unzip it in a location of your choice on your computer.

Launching QIAseq template workflows

QIAseq template workflows, including the Perform QIAseq Multimodal Analysis (Illumina) workflow, can be launched using the Analyze QIAseq Samples tool. However, running these workflows directly from the Toolbox comes with several advantages, including:

- Settings can be adjusted. By contrast, workflows launched using Analyze QIAseq Samples are run using the default settings.

- Raw data can be imported on the fly when launching the workflow. For example, fastq.gz files can be used directly as input, avoiding the need to import the reads in a separate step.

- When launched directly, you can choose to run the workflow in batch mode.

  Running in batch mode means the job is configured and launched once, and the workflow is then run once per sample, or in other words, once per batch unit, where each batch unit includes the data for a single sample. This allows many samples to be analyzed using identical analysis settings and little hands on effort.

  Launching workflows with more than one changing input in batch mode is described in the CLC Genomics Workbench manual and in a tutorial focused on this topic.

In this tutorial, we want to use on-the-fly import and adjust the settings to enable an option for detecting fusions with novel exon boundaries. Thus, we will launch the workflow from the Toolbox.

Launch the Perform QIAseq Multimodal Analysis (Illumina) workflow

1. Start the Perform QIAseq Multimodal Analysis (Illumina) workflow from the following location in the Toolbox:

   Template Workflows | Biomedical Workflows (📁) | QIAseq Panel Analysis (📁) | QIAseq Analysis Workflows (📁) | Perform QIAseq Multimodal Analysis (Illumina) (📁)

2. In the "Select DNA reads" step (figure 1):

   - Enable the **Select files for import** option and choose **Illumina** from the drop down menu.
   - In the **Select files** field, click on the folder icon to the right and select the two DNA fastq files in the tutorial data.
   - Select the two DNA fastq paired files from the unzipped folder.
   - Enable the **Paired reads**, **Discard read names** options and **Remove failed reads**.
   - Keep the default values for **Read orientation** ("Forward Reverse"), **Minimum distance** (1) and **Maximum distance** (1000).
   - Click on **Next**.

   Using the Illumina importer, each pair of fastq files is imported into a single sequence list containing both members of each read pair.

3. In the "Select RNA Reads" step, repeat the above steps, but select the two RNA fastq files in the tutorial data.

4. In the "Select reference data set" step, select the "QIAseq Multimodal Panels hg38". Click on the **Download to Workbench** button if this data is not already available on your system (figure 2).

5. In the "DNA Primers" step, select the "UHS-009Z_dna_panel_primers" from the drop-down menu (figure 3) and click **Next**.

6. In the "RNA Primers" step, select the "UHS-009Z_rna_panel_primers" from the drop-down menu and click **Next**.

7. In the "DNA target regions" step, select the "UHS-009Z_target_regions" from the drop-down menu and click **Next**.
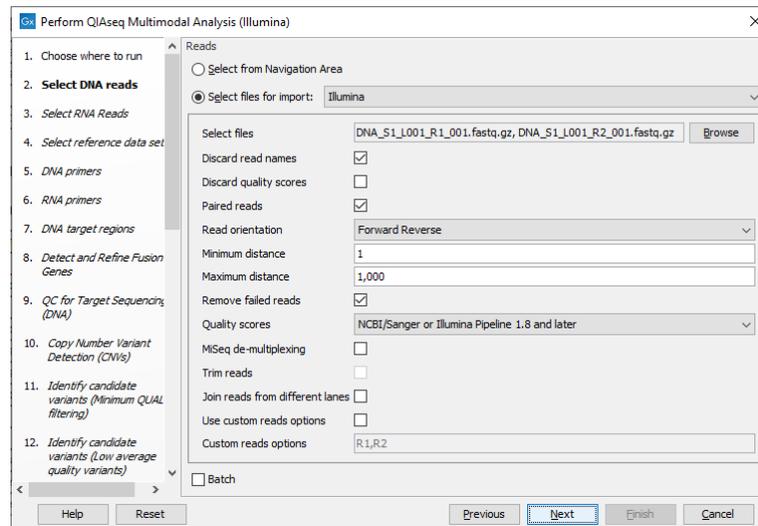
Figure 1: *Choose the "Select files for import" option to import data using a specified importer before the analysis steps of the workflow are run.*
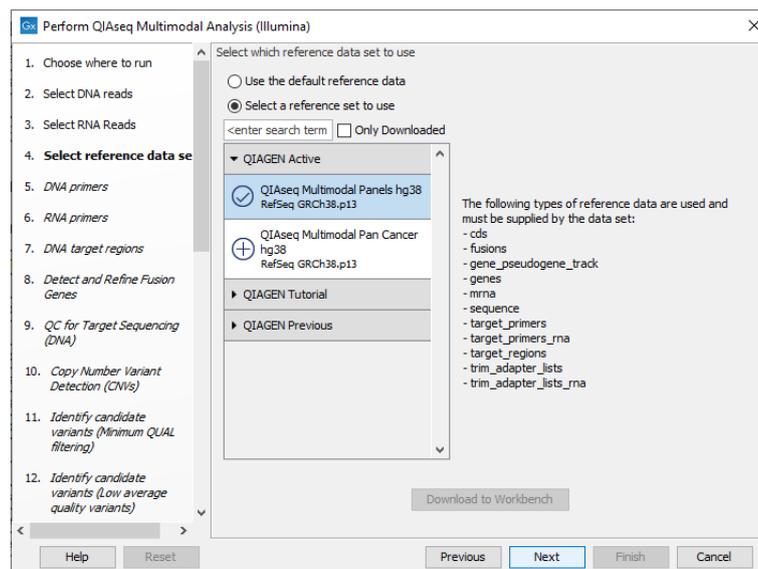
Figure 2: *The relevant Reference Data Set is highlighted. In the text to the right, the types of reference data needed are listed.*

8. In the "Detect and Refine Fusion Genes" step, click in the checkbox next to "Detect fusions with novel exon boundaries" to enable that option (figure 4), and click on **Next**.

9. Click on **Next** for all the following steps until you get to the last step, called "Save location for new elements".

10. In the "Save location for new elements" step, select a folder to save the results to and then click on **Finish**.

This analysis takes approximately 60 minutes on a 2019 Windows 10 with 24GB of RAM with no other heavy tasks running.
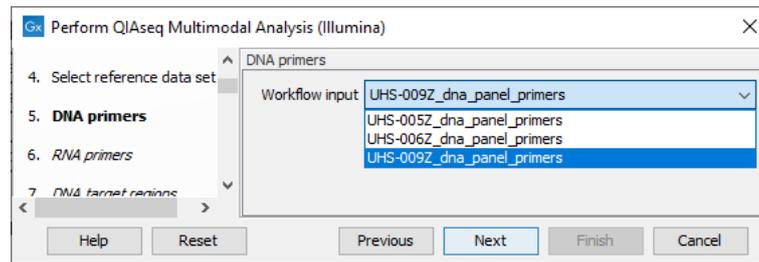
Figure 3: *The right primers are selected from the drop-down menu.*
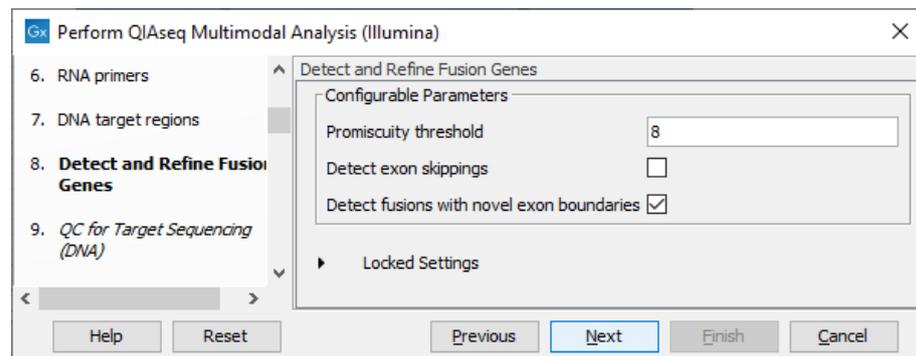
Figure 4: *Enable the option for detecting fusions with novel exon boundaries.*

Outputs from the Perform QIAseq Multimodal Analysis (Illumina) workflow

The Perform QIAseq Multimodal Analysis (Illumina) workflow produces the following top-level outputs, as shown in figure 5:

- A DNA Combined QC Report ( ) summarizing important QC values pertaining to the DNA read data.

- A RNA Combined QC Report ( ) summarizing important QC values pertaining to the DNA read data.

- A Fusion Report ( ) that includes graphical representations of the fusions found. Double-clicking on a fusion plot in the report will open the plot in a view that allows it to be exported as a high resolution image.

- A Gene Expression Track ( ) containing gene expression counts.

- A Genome Browser View (WT) ( ) that contains DNA and RNA read mapping tracks and variant tracks.

- A Genome Browser View (Fusions) ( ) that contains only the fusion chromosomes, which are used for refining the fusions.

- A Workflow Result Metadata table that records all the outputs generated by the workflow run.

Four subfolders (QC & Reports, Tracks (WT), Tracks (fusion), and VCF Exportable Tracks) contain outputs that are either referenced by a Genome Browser View, or provide more detail for the top-level outputs.
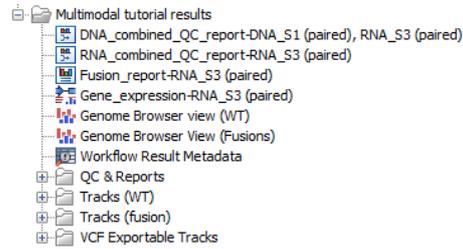
Figure 5: *Outputs from the Perform QIAseq Multimodal Analysis (Illumina) workflow.*

Reviewing QC Reports   First, it is important to check the Combined QC Reports to see if there are problems with the samples. There is one report section for each type of tool in the workflow, and the order of sections matches the order in which the tools are run. One way of reading the DNA Combined QC report might be as follows:

- "Trim Reads | Detailed Trim Results | Trim on Quality": Few reads were trimmed for low read quality.

- "Reads mapping summary | Reads summary": Almost all reads mapped to the reference.

- "Reads mapping summary | Mapped paired reads": The average fragment size was close to 200nt, with a standard deviation close to 100nt.

- "Calculate unique molecular index groups | Group sizes | Median": Most of the reads had only a single UMI. Ideally there would be 2-4 reads per UMI such that most reads could be error-corrected by making a consensus UMI read. Extremely low frequency variant calls on this sample are less likely to be correct.

- "Create UMI reads from grouped reads | Median Q scores for UMI reads": This value is under 41.5 (the default cut off). Stringent filtering in the workflow filters away variant calls that are supported by many consensus UMI reads where the average Q score is <41.5. This sample may have missing variant calls and it can then be recommended to make the Q score filtering less stringent by reducing the "Average quality" from the default 41.5 to 38 in the **Identify candidate variants (Low average quality variants)** dialog.

- "QC for targeted sequencing | Fractions of targets with coverage above threshold": Approximately half the target regions have >100x coverage over 90% of the target region.

A similar review of the RNA Combined QC Report should be carried out. The contents of this report are different, as the analysis uses different tools. A cause of concern is that a cell in the "Strand specificity" section is `pink`, indicating a suspected problem. The warning shows that >90% of the RNA reads mapped sense to annotated transcripts, which might suggest that the data was generated with a strand-specific protocol and should be analyzed accordingly. In the case of Multimodal panels, however, this number is determined by primer design and it is correct to ignore the warning. Another cause of concern is a pink cell in the "Adapter read-through" section telling that >10% of the paired reads are estimated to conatin read-through adapters and it can be necessary to remove these adapters. Find more information about how to creating new adapter list to be used to trim your reads `here`.

Reviewing variant calls

To investigate variant calls, open the Genome Browser View (WT) output. The file opens as a split screen with both Track List and the corresponding variant table. In the table, filter for variants for which the allele is not identical to the reference: open the advanced filter by clicking on the arrow in the top right corner of the table. Fill in the filtering fields with "**Reference allele**", "**=**" and type "**No**" (figure 6).
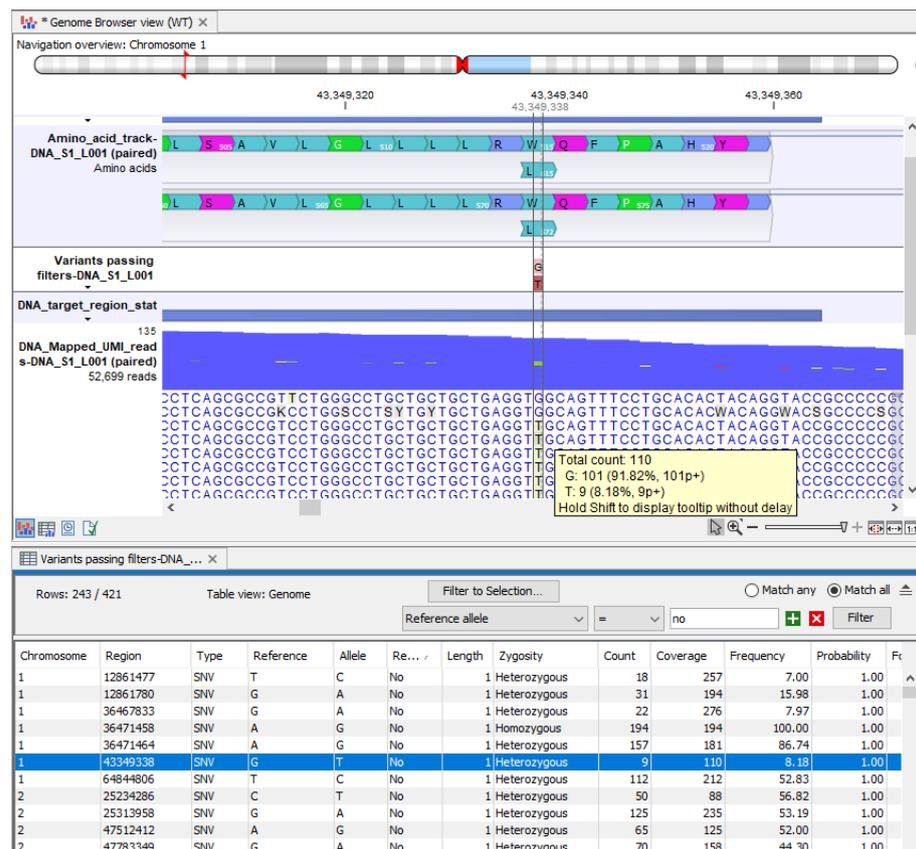
Figure 6: *Genome Browser View after filtering to show only alt alleles. The selected variant is from SeraCare Seraseq® Myeloid DNA and has an expected frequency of 5%.*

Clicking on a row in the variant table causes the Genome Browser to zoom in to the nucleotide level with the focus on that variant position. This allows you to view the coverage and the context of this variant in the various tracks, e.g., Mapped UMI Reads, Amino acid track, and so on.

This sample consists of a germline background supplemented by synthetic spike-ins of variants at somatic frequencies. For some regions of the genome, the calls in the germline background have been well-characterized by the GIAB project (Zook et al. 2019). To compare your calls with the background set, right-click within the Track List, choose **Include More Tracks...**, and select the Expected-HG002 track. There are 148 expected background variants. Of these, approximately 130 will be present in your calls. The missing calls are mostly filtered out due to the "Median Q scores for UMI reads" being lower than ideal.

Your call set will contain additional variants corresponding to i) the germline background in regions that are not yet well-characterized (outside GIAB high confidence regions), ii) the synthetic spike-ins, and iii) some false positive calls, which are almost exclusively at frequencies <5%.

An example of a synthetic spike-in call is a 28nt duplication in FLT3 with an additional "GCCCC" between the duplicated and native sequence. This is represented as a single insertion of length 33nt (figure 7). To find this in your own data, sort the "**Length**" column in the variant table in descending order by clicking on the header of that column.

Figure 7: *An example of a synthetic spike-in call: a 28nt duplication in FLT3 with an additional "GCCCC" between the duplicated and native sequence.*

Calls down to a frequency of 0.5% are produced by the workflow, but the lower the frequency, the greater the importance of having several reads per UMI to correct for error. In this sample, there is usually only one read per UMI and so most reads are uncorrected. To find low frequency calls, sort the "**Frequency**" column in the variant table in ascending order by clicking on the header of that column.

Reviewing fusion calls   To investigate fusion calls, open the Fusion Report output. The report has a section for each detected fusion that passes quality filters, and the sections are ordered from the most certain to the least certain calls.

Quality control of fusion calls is described in detail in the Biomedical Genomics Analysis manual.

In the report, the first 8 calls are all the fusions expected to be present in this SeraCare sample (figure 8). The last call containing TAF15 and RNR1 (ribosomal RNA) is a false positive call. You might also see a false positive call containing CD74 and RNA18SN4. There are several indications for both being false positives.

- Fusion breakpoints are most commonly observed at exon boundaries, but in the false positives most reads extend from the middle of an exon to the middle of an exon, as shown in figure 9.

- The false positive fusion is the last one reported, and therefore have least evidence among the fusion passing filter.

Figure 8: *Contents of the fusion report. Each fusion has its own section.*

- TAF15 is the most highly expressed gene in the sample, and ribosomal RNA is also typically highly expressed. This gives a possible mechanism for the two being in close spatial proximity during transcription.

- There is no primer for ribosomal RNA in the panel. This is not conclusive, but suggests that such fusions are unexpected.
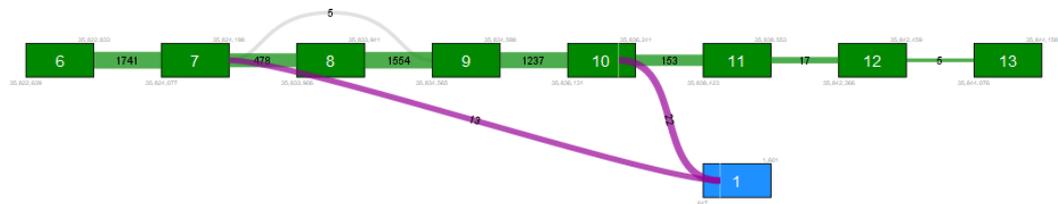
Figure 9: *The TAF15-RNR1 fusion is likely a false positive. Fusion breakpoints are most commonly observed at exon boundaries, but here most fusion reads (purple) have both breakpoints in the middle of exons (white vertical lines).*

(Optional) Upload your results to QCI Interpret

It is possible to upload your results to QCI Interpret. For this section, you need an active QCI Interpret subscription (not a trial license).

1. To start the upload, open the Analyze QIAseq Samples and click on **Upload to QCI Interpret**. Alternatively, open **Upload to QCI Interpret** directly from the toolbox or via the **Launch** button.

2. In the first dialog (figure 10), select the variant track and fusion track from the "VCF Exportable" subfolder of your results.

3. In the next step it is optional to type in a Sample name, SubjectID and/or Project as seen within QCI Interpret. Select the reference sequence to hg38. This sequence can be found in
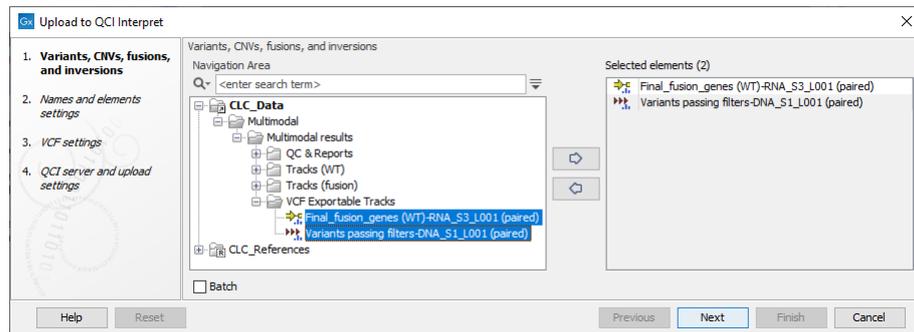
Figure 10: *Select the variants and fusion files you would like to upload to QCI Interpret.*

the "Reference Data tab", under the "QIAGEN Active". Find here the QIAseq Multimodal Panels hg38 and under "sequence", the Homo_sapiens_sequence_hg38_no_alt_analysis_set (figure 11).
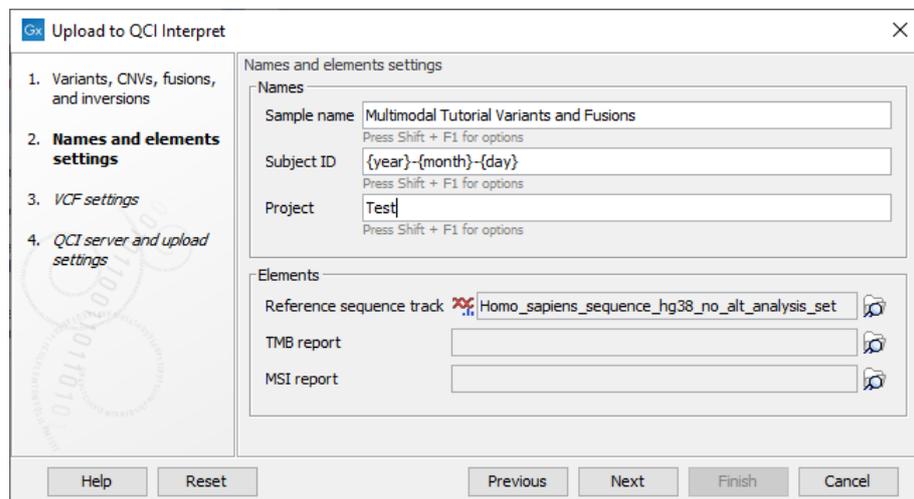
Figure 11: *Provide names for the QCI Interpret sample and select the reference sequence.*

4. In the next step you can specify VCF export settings. For this upload, choose "Somatic" under "Prefill VCF Settings".

5. In the Server and upload settings dialog (figure 12), select your QCI region. Under QCI Interpret login, select "Browser" and click "Log in". This will open a new browser window where username and password associated with your QCI Interpret account can be typed in. When you are logged in, you will be told to return to the CLC Workbench.

6. Click on **Finish**.

Note: Clicking on **Finish** in the wizard is the start of the steps needed to transfer variants and/or fusions to QCI Interpret. The tool will output a report containing the link to the QCI Interpret sample list page (this page will also open when the tool is complete) (figure 13).

To analyze the uploaded data in QCI Interpret, create a new test selecting the uploaded sample and fill in the QCI Interpret Interview pages.
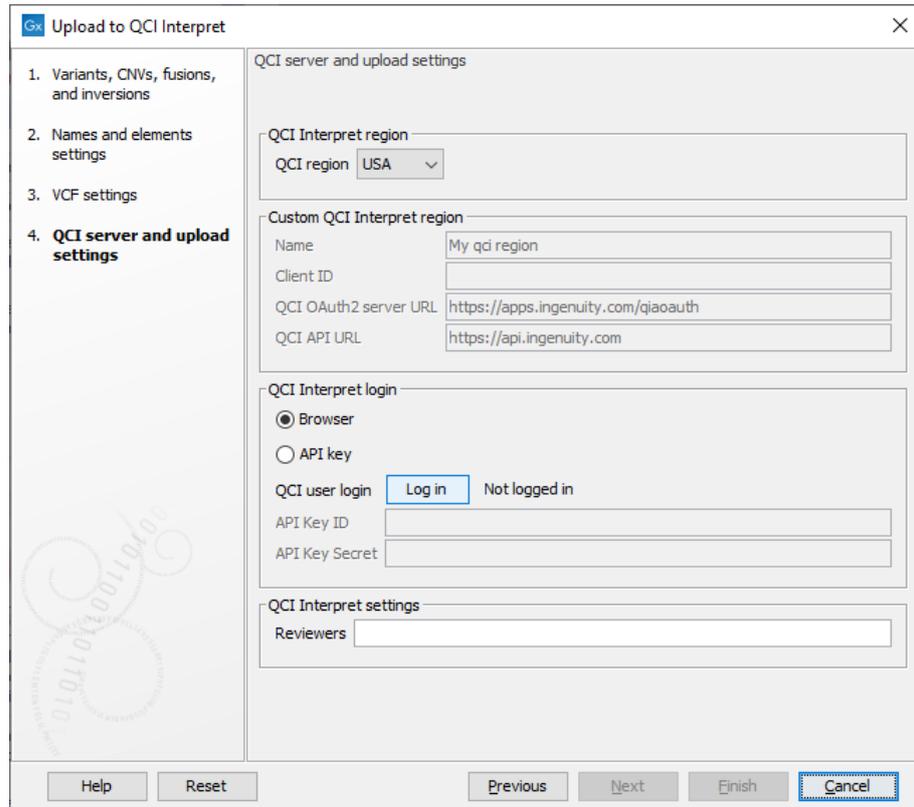
Tutorial

Figure 12: *Configure the connection to the QCI Interpret server dialog.*

Please refer to the QCI Interpret documentation to learn more about the interpretation of the results (figure 14).

Figure 13: *Example of a report containing a link to the QCI Interpret sample list.*

Figure 14: *Results of the interpretation analysis from the QCI Interpret website.*