Tutorial

Fusion Detection from RNA-Seq

April 7, 2025

Sample to Insight

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Fusion Detection from RNA-Seq

Introduction

The purpose of this tutorial is to demonstrate how *CLC Genomics Workbench* and the *Biomedical Genomics Analysis* plugin can be used to detect fusions from RNA-Seq data.

We focus on the following:

- Import data.
- Create a `Custom Data Set`.
- Detect fusions using a `template workflow`.
- Interpret the results.
- Find missing fusions.

Data used in this tutorial

This tutorial uses data produced with a QIAseq Targeted RNAscan Panel.

To complete the tutorial in a reasonable amount of time, only a subset of reads from a single sample that map to chromosome 10 are used here.

The data distributed for use with this tutorial also contains the following reference tracks:

- Sequence track with chromosome 10 of the human hg38 genome.
- Genes, mRNA, and CDS annotation tracks.
- Known fusions and primer tracks.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 25.0 and *Biomedical Genomics Analysis* plugin 25.1 or higher. Note that higher versions may produce slightly different results than those shown here.

Installing plugins is described in the `CLC Genomics Workbench manual`.

General tips

- Throughout this tutorial, we provide links to relevant manual pages, which we recommend exploring for additional details.
- Tools and workflows can be found in the Toolbox, but it is often easier to launch them using Quick Launch (), found in the top toolbar (shortcut Ctrl+Shift+T or ⌘ +Shift+T on Mac). Quick Launch displays the full Toolbox path, making it easy to identify the location of the tool or workflow if needed.
- The in-built manual can be accessed by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.
- Within wizards, the **Reset** button can be used to change settings to their default values.
- Columns in tables can be hidden by unchecking their name in the Side Panel.
- Columns in tables can be used to sort the rows, by successively clicking on the column name until the desired order (indicated by an arrow next to the column name) is achieved.
- Most of the tools of *CLC Genomics Workbench* require multiple inputs. When many data elements need to be selected, all elements located under a folder can be added by using the options **Add folder contents** or **Add folder contents (recursively)** found in the right-click menu.
- Many data elements produced by *CLC Genomics Workbench* tools have multiple views, indicated as icons in the lower left corner of elements opened in the View Area. Clicking on one of the view icons while pressing the Ctrl (⌘ on Mac) key will open in split view such that both views are visible at the same time. Often, if viewing a table and a graphical representation in split view, selecting entries in the table will highlight them in the graphical representation. The order of the views can be changed using drag and drop, see Arrange views in View Area.
- Data can be imported prior to starting a workflow, or it can often be imported on the fly when the workflow is launched.

Import the data

We start by downloading and importing the tutorial data.

1. Download the `tutorial data`.
2. Start the *CLC Genomics Workbench*.
3. Import the data using `Standard Import`:
   (a) Launch **Standard Import** ( ) using `Quick Launch` ( ).
   (b) Locate the tutorial data using the **Add files** button and select **Automatic import** (figure 1).

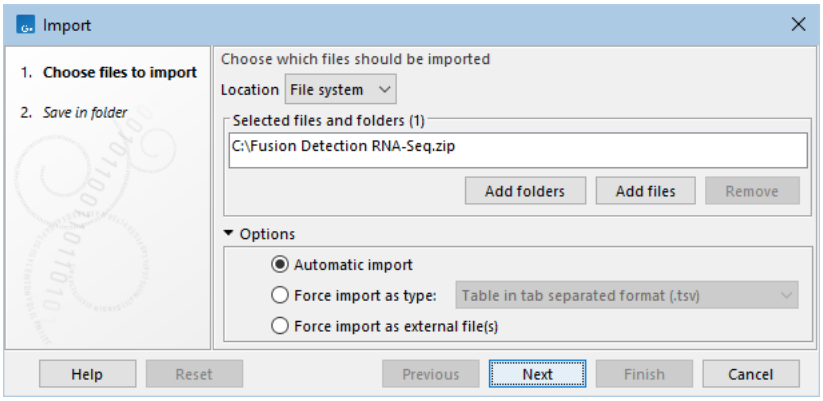

Figure 1: *Standard Import configured to import the tutorial data.*

   (c) In the next step, select a suitable location in the `Navigation Area` to save the imported data and click on **Finish**.

Once the import is completed, the folders and data elements are visible in the Navigation Area (figure 2).

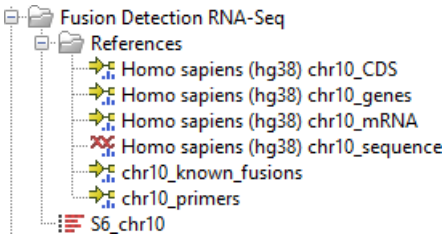

Figure 2: *The imported tutorial data in the Navigation Area.*

Create a Custom Data Set

We will now create a Custom Data Set containing the reference tracks imported in step 3. This is needed later for running the template workflow.

1. Launch `Manage Reference Data` ( ) using Quick Launch ( ).

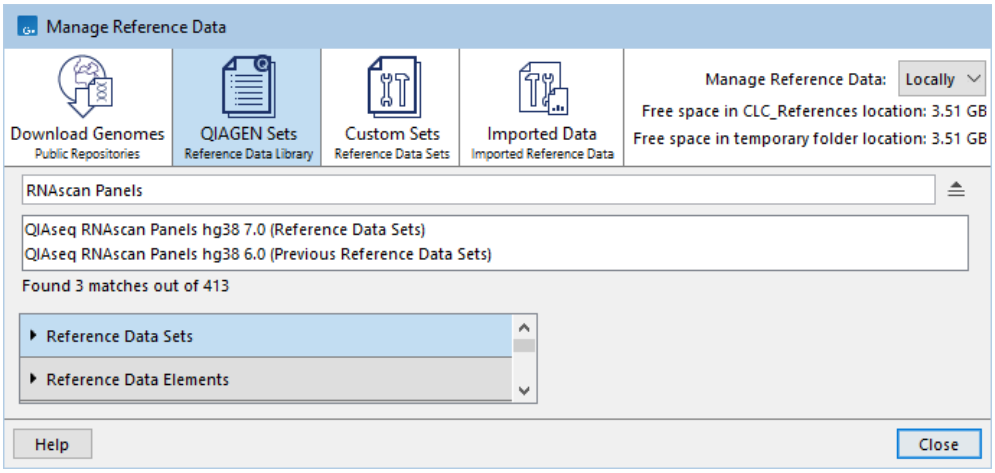

Figure 3: *The Reference Data Manager showing the available RNAscan Reference Data Sets.*

2. On the `QIAGEN sets` tab, enter "RNAscan Panels" in the search field.

   Double-click on the search result that is not a "previous" set (figure 3).

3. Click on the **Download** button located in the top-right corner (figure 4).

   We will create the Custom Data Set while the data is being downloaded. Click on the **Create Custom Set** button at the right-hand side of "Download". A wizard will open.

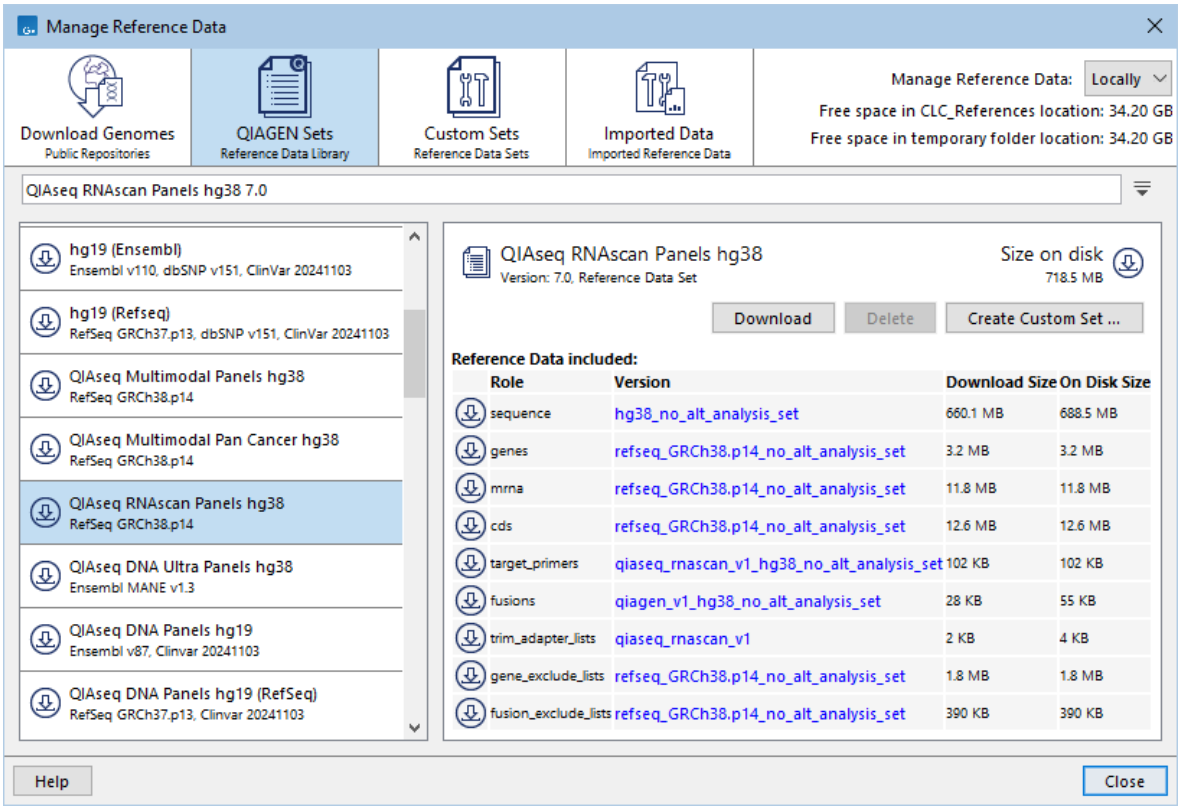

Figure 4: *The Reference Data Elements in the "QIAseq RNAscan Panels hg38" Reference Data Set.*

4. Click on the ( ) button at the right-hand side of **sequence**, **genes**, **mrna**, **cds**, **target_primers**, and **fusions** and select the corresponding imported reference tracks (figure 5).

   Name the set "QIAseq RNAscan Panels hg38_chr10" and click on **OK**.

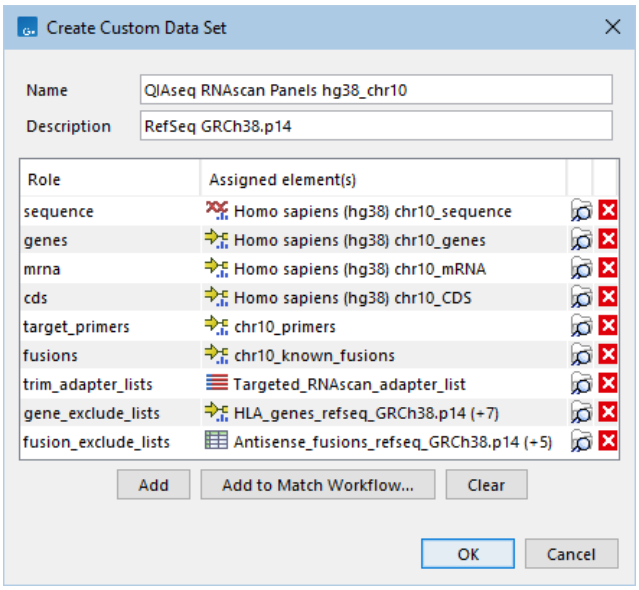

Figure 5: *A Custom Data Set configured with the imported reference tracks.*

5. Click on **Close**.

Detect fusions

We will now use the `Detect QIAseq RNAscan Fusions` template workflow to analyze the tutorial data. This workflow has been designed for data generated using a QIAseq Targeted RNAscan Panel. If you run this workflow on your own data, please note that template workflows are provided as example workflows and may need to be customized to meet the specific requirements of your data.

To see the content of the workflow, locate it in the Toolbox:

> **Workflows** | **Template Workflows** | **Biomedical Workflows** ( ) | **QIAseq Sample Analysis** ( ) | **QIAseq RNA Workflows** ( ) | **Detect QIAseq RNAscan Fusions** ( )

right-click on its name and choose **Open Copy of Workflow**.

We will now run the workflow:

1. Launch the workflow using Quick Launch ( ) or by double-clicking its name in the Toolbox.
2. In the first wizard step, "Select Reads", specify the data to be analyzed by selecting the imported reads (figure 6).
3. In the next step, "Specify reference data handling", select the "QIAseq RNAscan Panels hg38_chr10" Reference Data Set under "Custom" (figure 7).

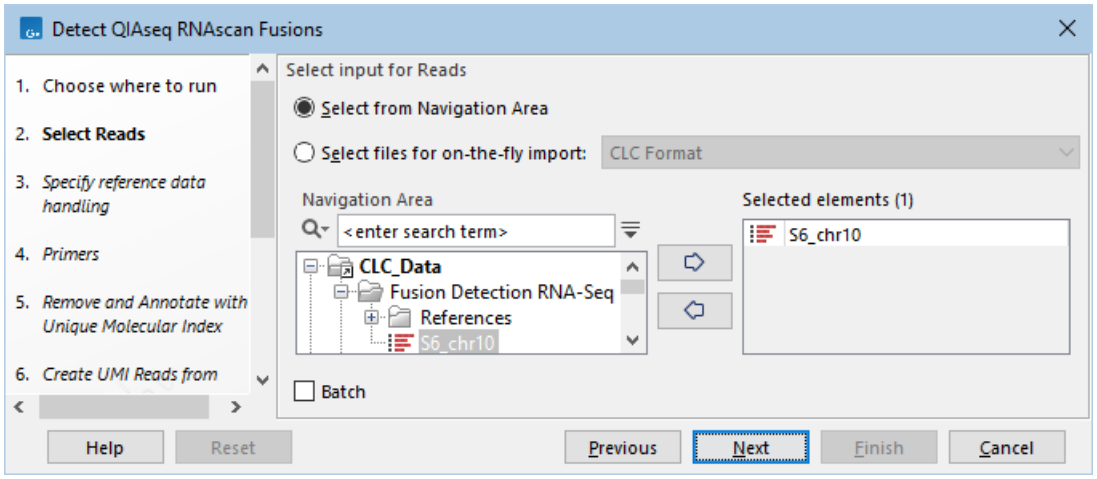

Figure 6: *The "S6_chr10" reads are used as input.*

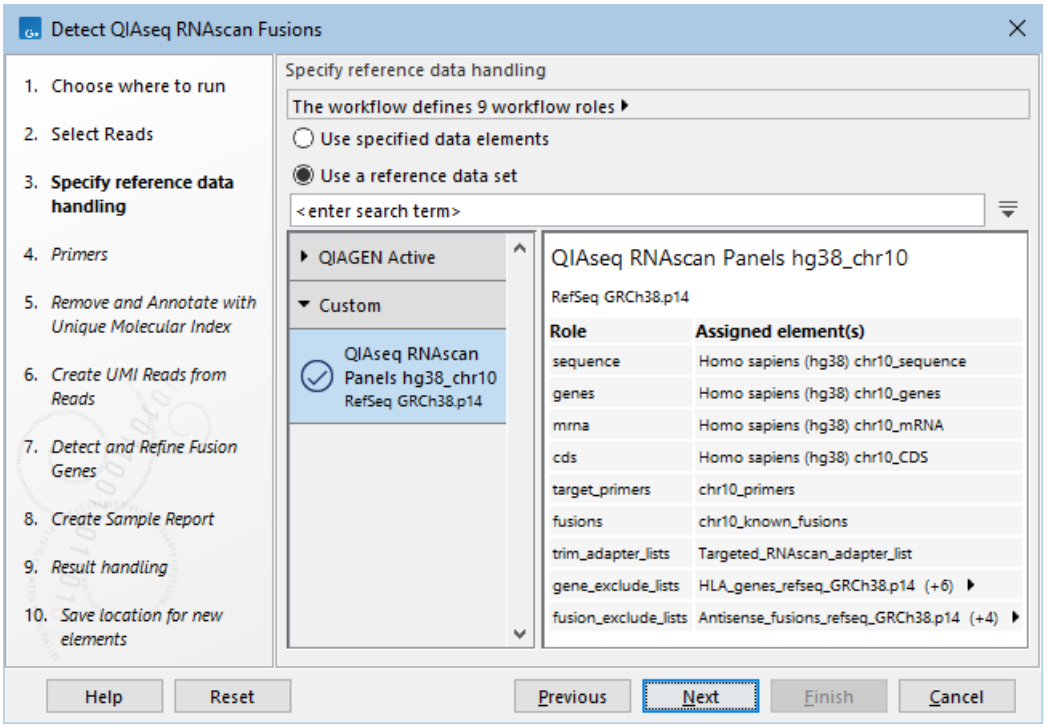

Figure 7: *The "QIAseq RNAscan Panels hg38_chr10" Custom Data Set is selected.*

4. In the next steps, "Primers", "Remove and Annotate with Unique Molecular Index", "Create UMI Reads from Reads", "Detect and Refine Fusion Genes", "Create Sample Report", and "Result handling", keep the default settings.

5. In the last step, make a new subfolder in "Fusion Detection RNA-Seq" called "Results" and choose to save the workflow results there.

   Click on **Finish**.

   The workflow will now execute. The progress can be monitored under the `Processes` tab in the Toolbox.

Interpret the results

Results from the workflow are placed in the "Results" folder (figure 8).

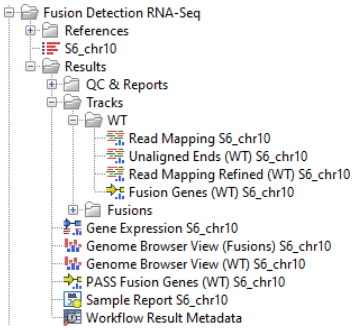

Figure 8: *The "Results" folder in the Navigation Area.*

The folder contains, among other things:

- A **QC & Reports** subfolder containing all reports produced by the workflow.
- A **Tracks** subfolder containing the read mapping from RNA-Seq Analysis and results from Detect and Refine Fusion Genes.
- Two Genome Browser Views: one for the wildtype (WT) chromsomes and one for the artificial fusion chromosomes created by Detect and Refine Fusion Genes.
- A Sample Report summarizing information from all the reports located in the QC & Reports folder.

Quality control

It is important to first verify that the data quality is satisfactory.

Open the **Sample Report S6_chr10**, found in the "Results" folder. The Quality control section contains different summary items that can be used to assess the quality of the reads (figure 9). These summary items can be configured in the "Create Sample Report" wizard step when launching the workflow.

1.2 Quality control

Summary item	Report type	Value	Threshold
% of R1 reads with average quality score ≥ 30	QC for sequencing reads	98.79	≥ 80.00
% of R2 reads with average quality score ≥ 30	QC for sequencing reads	91.83	≥ 80.00
Percentage of paired reads mapped in pairs	RNA-Seq analysis	92.36	≥ 50.00
DNA contamination ratio	QC for RNAscan panels (WT)	N/A	< 0.12
Mean read coverage per reference gene control primer	QC for RNAscan panels (WT)	N/A	≥ 300.00

Figure 9: *The sample report contains summary items for assessing the quality of the reads.*

QIAseq Targeted RNAscan Panels typically include control primers to check for DNA contamination. However, the reference primer track used here does not include control primers, so the "DNA contamination ratio" and "Mean read coverage per reference gene control primer" cannot be calculated. As a result, the values are marked in yellow. This does not affect the analysis here.

Review the identified fusions

To get a quick overview of the identified fusions, open the **Fusion Report S6_chr10**, found in the "QC & Reports" subfolder.

The `report` summarizes all identified fusions found to be adequately supported: *NCOA4-RET* and *KIF5B-RET*. We will now investigate *KIF5B-RET* (figure 10) to determine whether it might be a false positive call. See `Interpretation of fusion results` for more information.

3.2 KIF5B-RET

Fusion name	KIF5B-RET
5' gene	KIF5B
5' chromosome	10
3' gene	RET
3' chromosome	10
Reported transcript 5'	NM_004521.2
Reported transcript 3'	NM_020975.4
Translocation name	KIF5B{NM_004521.2}:r.1_3231_RET {NM_020975.4}:r.2070_5617
P-value	0.00
Z-score	14.97
Fusion crossing reads	28
Fusion spanning reads	0
5' read coverage	1,936
5' spanning read coverage	203
3' read coverage	28
3' spanning read coverage	0

5' position	3' position	Translocation Name	Count
32,017,142	43,114,479	KIF5B{NM_004521.2}: r.1_3231_RET {NM_020975.4}:r. 2070_5617	28

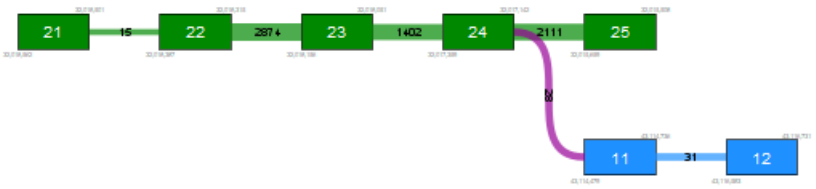

Figure 10: *Fusion report showing the KIF5B-RET fusion.*

There are 28 fusion crossing reads, meaning that 28 reads cover the exact fusion breakpoints, providing high-confidence support for the fusion. This is also reflected by the very low p-value and

high Z-score. *KIF5B-RET* has breakpoints at exon boundaries (figure 10), which is expected for most fusions.

To inspect the fusion more closely, open the **Genome Browser View (Fusions) S6_chr10**, found in the "Results" folder and double-click on the **Fusion Genes (Fusions) S6_chr10** track. The track will open as a table in split view, below the Genome Browser View (figure 11).

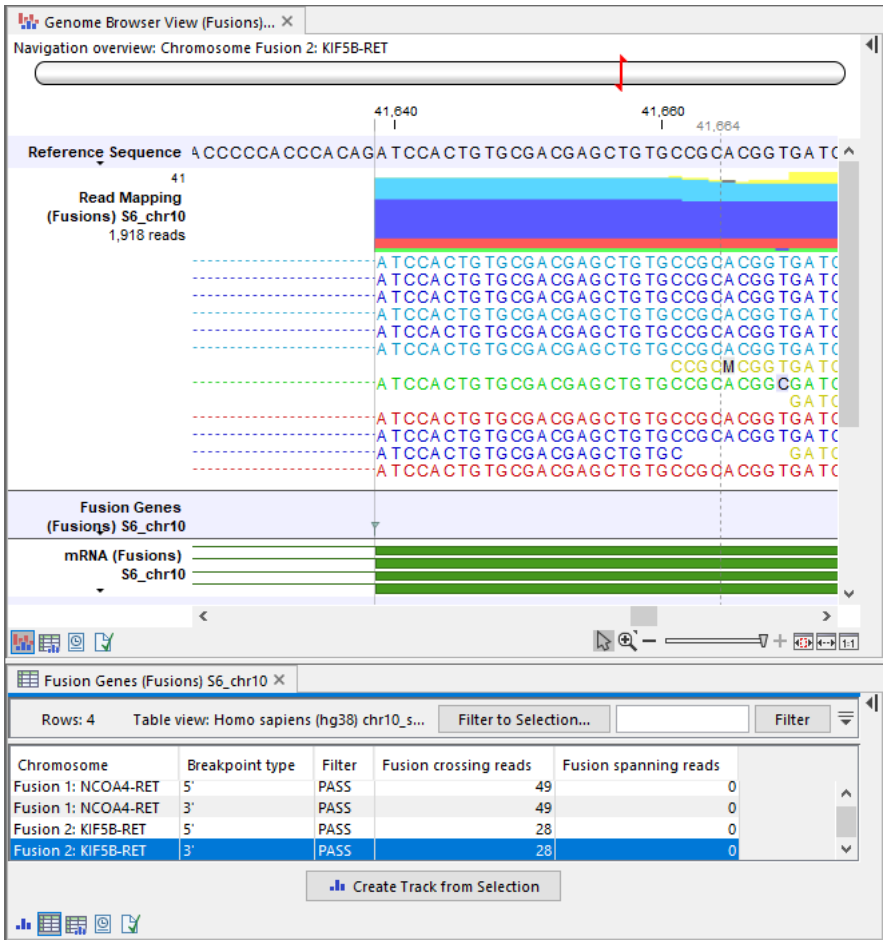

Figure 11: *Split view showing the genome browser at the top and the table view of the fusion genes at the bottom. Note that not all table columns are shown.*

Each row in the table contains information about one breakpoint. A fusion is represented by two breakpoints with the same "Fusion number" and "Fusion pair" values. A fusion must be supported by reads with spliced alignments, where the splice donor site is at the 5' breakpoint and the splice acceptor site is at the 3' breakpoint. Splicing is represented as dashed lines in the read mapping track. Note that the fusion chromosome also contains intronic regions, as shown by the mRNA track in the genome browser (figure 11). There can also be multiple breakpoints for each fusion chromosome.

Click on the 3' breakpoint row of the *KIF5B-RET* fusion. This will zoom in to its location in the genome browser and the reads supporting the fusion can be reviewed. Key aspects to verify are:

- The majority of reads map uniquely across the breakpoint. Yellow reads are not uniquely mapped and may indicate a false positive fusion. Here, only few reads are yellow.

- There is no sign of incomplete poly-A trimming. Reads that are A-rich on either side of the breakpoint may indicate a false positive fusion. Here, there is no sign of A-rich reads.

The Detect QIAseq RNAscan Fusions template workflow annotates the identified fusions with information about typical fusions occurring in cancer samples, found in the imported known fusions track. You can import your own known fusions track in the *CLC Genomics Workbench*.

Open the annotated **PASS Fusion Genes (WT) S6_chr10** track, found in the "Results" folder. Switch to the table view by clicking on the ( ) button at the bottom-left corner (figure 12).

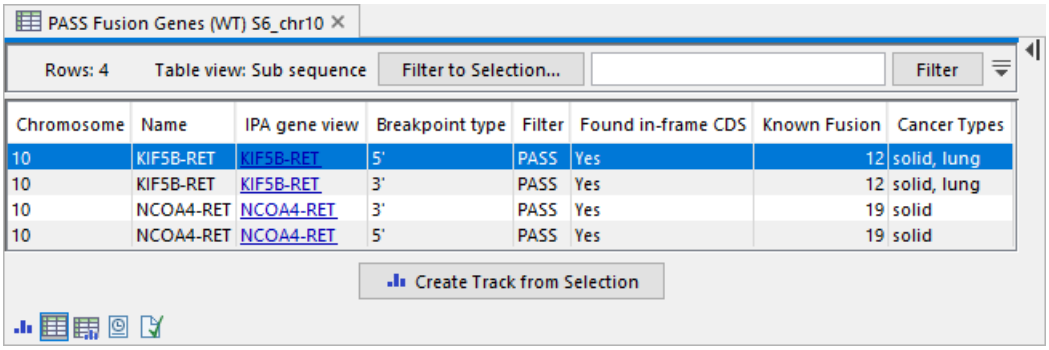

Figure 12: *Table view of the fusion genes (WT) track. Note that not all columns are shown.*

The WT table is very similar to **Fusion Genes (Fusions) S6_chr10** (figure 11). The main difference is that the WT table uses the wildtype chromosomes, whereas the fusion table uses artificial fusion chromosomes. The WT table shows that *KIF5B-RET* is a known fusion found in solid tumors and it corresponds to the fusion with number 12 in the known fusion track. Additionally, the fusion is in-frame.

The table view includes a link to QIAGEN Ingenuity Pathway Analysis (IPA) for fusions where additional information is available. Accessing the link requires access to IPA. You can request a free trial here. *KIF5B-RET* contains an IPA gene view link, indicating that this fusion is also recorded in IPA.

Altogether, the investigation demonstrates strong support for the *KIF5B-RET* fusion, making it highly unlikely to be a false positive call.

Finding missing fusions

The sample analyzed in this tutorial is known to contain the following fusions:

- *NCOA4-RET*
- *KIF5B-RET*
- *DNA2-TET1*
- *TM9SF3-OIT3*
- *EML4-ALK*

However, we observed that only the *NCOA4-RET* and *KIF5B-RET* fusions were adequately supported. We will now investigate why the remaining fusions were not identified.

Open the **Fusion Genes (WT) S6_chr10** track in the "Tracks/WT" subfolder. This track contains all detected fusions, including those with little support. To `filter` the table for the three missing fusions (figure 13):

1. Switch to the table view by clicking on the ( ) button at the bottom-left corner.
2. Click on the arrow ( ) at the top-right corner.
3. Select "Name" and "is in list" in the drop-down menus.
4. Enter "DNA2-TET1, TM9SF3-OIT3, EML4-ALK" in the search field.
5. Click on **Filter**.

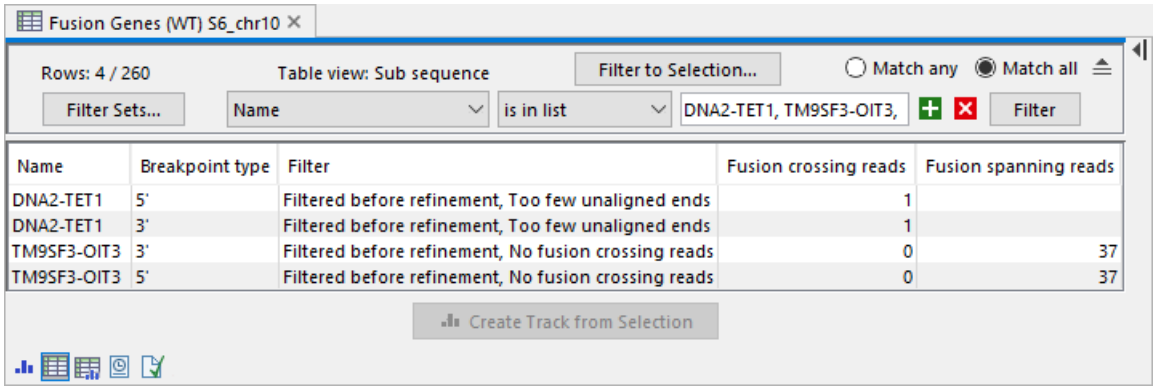

Figure 13: *Table view of the fusion genes (WT) track filtered to show the missing fusions. The "Filter" column provides details explaining why the fusions have been filtered. Note that not all columns are shown.*

The results show that two of these fusions have in fact been detected. The `Filter` column provides additional details:

- *DNA2-TET1* was filtered due to `Too few unaligned ends`. Only 1 unaligned end (i.e., 1 fusion crossing read, as seen in the "Fusion crossing reads" column) has been identified for this fusion (figure 13), while "Minimum unaligned end count" is set to 4 in the template workflow.

  To potentially recover this fusion:

  1. `Open a copy of the workflow`.
  2. Double-click on the **Detect and Refine Fusion Genes** element.
  3. Click on **Next** to navigate to the `Filter` wizard step.
  4. Set "Minimum unaligned end count" to 1.
  5. Click on **Finish**.
  6. Click on **Run** to run the workflow, as described previously.

     Note that workflows can also be `installed`, to provide and use a controlled version of the workflow.

- *TM9SF3-OIT3* was filtered due to `No fusion crossing reads`. The precise breakpoint location could not be identified, but the fusion is supported by 37 fusion spanning reads (figure 13), meaning 37 paired reads mapped as broken pairs on each of the two genes.

  To potentially recover this fusion, the following options in the `Detect` wizard step of Detect and Refine Fusion Genes can be changed in a workflow copy, as described previously:

  - "Minimum length of unaligned ends" is set to 20 in the template workflow, but can be decreased to 15. A lower value increases the chance of identifying fusion crossing reads.
  - "Maximum distance to known exon boundary" is set to 10 in the template workflow, but can be increased to 15. A higher value increases the chance of identifying fusion crossing reads when "Detect novel exon boundaries" is not checked, as is the case in the template workflow. Alternatively, checking "Detect novel exon boundaries" and "Detect novel exon boundaries in both genes" leads to identifying fusions with breakpoints outside exon boundaries. However, checking these two options also increases the tool's runtime and can lead to a higher number of false positive fusion calls.

- *EML4-ALK* is not in the table (figure 13), indicating that the fusion was not detected and there is no support from broken paired reads either. Adjusting the options in the `Detect` wizard step of Detect and Refine Fusion Genes, as described previously, could potentially help recover this fusion.

Both *TM9SF3-OIT3* and *EML4-ALK* lack fusion crossing reads. The key difference is that *TM9SF3-OIT3* is supported by fusion spanning reads, whereas *EML4-ALK* is not. It is possible that no reads in the sample cover the fusion breakpoints, making it impossible to detect the fusion regardless of the options chosen. This could be due to low expression of the fusion gene, which may not be captured if the sample is not sequenced deeply enough, or it could be due to some error in the experimental procedure.

Upload to QCI Interpret

To classify pathogenicity and actionability of the identified fusions, we recommend `QCI Interpret`. If you have an active subscription (not a trial license), you can use your account credentials and `Upload to QCI Interpret` to upload the **PASS Fusion Genes (WT) S6_chr10** track.