



Tutorial

How to modify a QIAseq myeloid workflow to detect KIT D816V at low frequencies

February 3, 2022

— Sample to Insight —

How to modify a QIAseq myeloid workflow to detect KIT D816V at low frequencies

Existing workflows, such as template workflows distributed with the Biomedical Genomics Analysis plugin can be easily customized to suit specific analysis requirements. In this tutorial we modify the QIAseq analysis workflow **Identify QIAseq DNA Somatic Variants (Illumina)** to detect KIT D816V down to 0.4% variant allele frequency.

The following is an overview of the changes we will make to this workflow:

- **The frequency cutoff for variant detection will be lowered** We want to detect the KIT D816V variant down to 0.4%, which is lower than the frequency cutoff used by the original workflow. Thus we will lower the frequency cutoff in the Low Frequency Variant Detection step to support this. As this modification could lead to many false positives in the list of all variants detected, we will make other modifications to account for this.
- **Filters specific for KIT D816V will be added.**
- **A Workflow Output element will be added to save KIT D816V variants detected.**
- **Filters will be added to remove likely false positives from the overall list of variants detected** By decreasing the frequency cutoff for variant detection, the number of false positives is likely to increase. These added filters will be configured to remove very low frequency variants from the overall results, to adjust for this.

Together, these workflow modifications should generate the outputs expected from the original workflow, including a variant track with variants detected at frequency 0.5% and above, with an additional variant track dedicated to KIT D816V variants.

Prerequisites

For this tutorial, you should be working with *CLC Genomics Workbench* 22.0 or higher and have the Biomedical Genomics Analysis plugin installed.

Minimum recommended machine specifications for working with human data sets are listed at <https://digitalinsights.qiagen.com/technical-support/system-requirements/>.

If you are connected to a CLC Server via your Workbench, you will be asked where you would like to run the analysis. We recommend that you run the analysis on a CLC Server when possible. However, the data supplied for this tutorial is a reduced set, so a standard desktop computer/laptop with 4 GB RAM is sufficient.

When focusing on a particular variant or set of variants, a target region track that specifies their locations on the genome is needed. A target region track containing the location of the KIT D816V variant is provided with the tutorial data.

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 where they appear, for example in wizards or in editors.

Download and import the tutorial data

1. Download the sample data from http://resources.qiagenbioinformatics.com/testdata/KITD816V_tutorial_data.zip.

To import the tutorial data:

2. Open the *CLC Genomics Workbench*.

3. Go to:

File | Import (📁) | Standard Import (📁)

4. Choose the zip file you just downloaded, leave the Import type set to **Automatic import** and click on **Next**

5. Select a folder to save the imported data to and click on **Finish**

After import, you will have a folder called "KIT D816V workflow tutorial", containing the following three items:

KIT_D816V_reads_chr4 (📄)

A sequence list containing Illumina reads from a sample that has the KIT D816V mutation.

KIT_D816V_Target_Region_Ch4_55599321_hg19 (📁)

A target region track for KIT D816V (Chr4:55599321), based on hg19.

Identify QIAseq DNA Somatic Variants (Illumina) with KIT_D816V_at_0.4% (🔧)

A copy of the **Identify QIAseq DNA Somatic Variants (Illumina)** workflow with the modifications described in this tutorial already applied to it. You can run this workflow and use the results for comparison with the results you obtain with the modified workflow you create. Select the reference data set as described below (the "QIAseq DNA Panels hg19, Ensembl v87"), the "DHS-003Z_target_regions" and "DHS-003Z_panel_primers" from the drop-down menus. The "KIT_D816V_Target_Region_Ch4_55599321_hg19" from the imported data must be used in the **KIT D816V** step.

Downloading the reference data

We will use the QIAGEN reference set called "QIAseq DNA Panels hg19, Ensembl v87", which can be downloaded using the Reference Data Manager. If you have not already downloaded this, we recommend that you do this now, as this can take some time.

If you have already downloaded the "QIAseq DNA Panels hg19, Ensembl v87" reference set, you can go directly to the section called **Modifying the workflow**.

How to download reference data is described in the [manual](#). We step quickly through this process below.

1. Click on the **References** button in the top toolbar to open the Reference Data Manager.

- Click on the **QIAGEN Sets** tab.
- With the Reference Data Sets section open, tab, scroll down until you find the set named "QIAseq DNA Panels hg19, Ensembl v87". Click on it to select it.
- If you will be running the analysis on a *CLC Genomics Server*, change the setting in the upper right corner to "On Server" as shown in figure 1.
- Click on the **Download** button above the list of data elements to download this reference data set.

A checkmark icon next to the reference set or reference data elements indicates they have been downloaded.

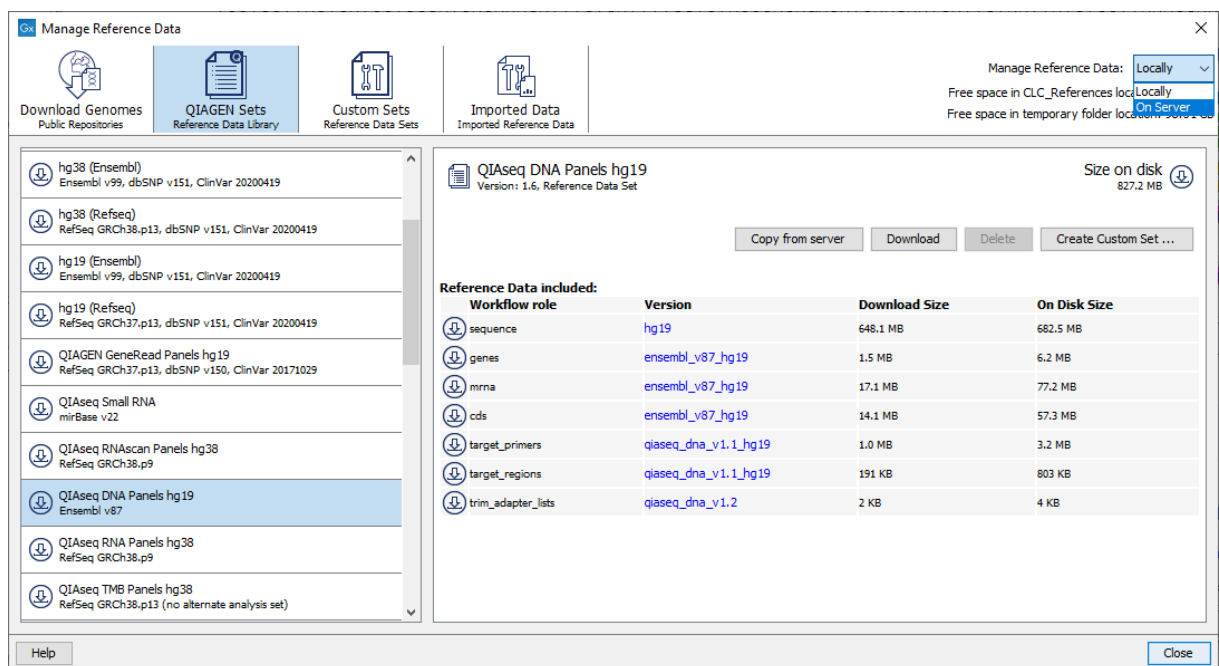


Figure 1: "QIAseq DNA Panels hg19, Ensembl v87" is listed under QIAGEN Sets. No reference data elements have yet been downloaded, as indicated by the arrow icons to the left of the element names. If downloading to a **CLC Genomics Server**, the setting in the top right should be set to "On Server".

Modifying the workflow

The workflow we will modify can be found here:

Toolbox | Template Workflows | Biomedical Workflows () | QIAseq Sample Analysis | QIAseq Analysis Workflows | Identify QIAseq DNA Somatic Variants (Illumina)

We will make a copy of this workflow, and then modify that copy.

- In the Toolbox at the bottom, left side of the Workbench, right-click on the name of the **Identify QIAseq DNA Somatic Variants (Illumina)** workflow and click on **Open Copy of Workflow** (see figure 2).

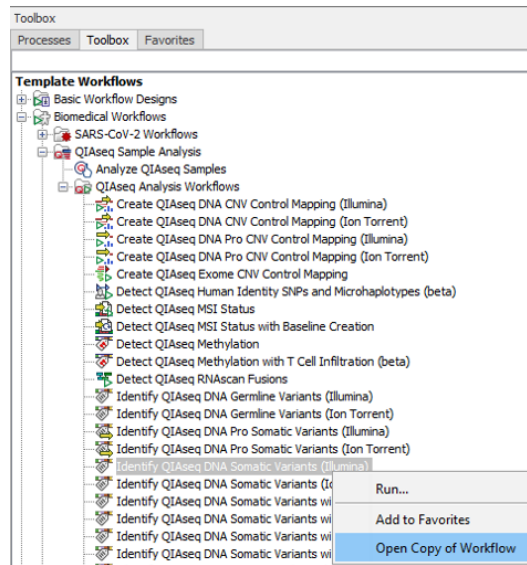


Figure 2: Open a copy of the workflow "Identify QIAseq DNA Somatic Variants (Illumina)".

2. Save the workflow copy by going to the menu at the top of the Workbench and choosing:

File | Save as... Select the folder with the imported data. The default name when saving a copy of a workflow is "Copy of (name of workflow)" - so in this case it will be "Copy of Identify QIAseq DNA Somatic Variants (Illumina)". As for all clc objects you can always change the name if needed.

The saved workflow is now located in the Navigation Area in the "KIT D816V tutorial" folder (see figure 3).

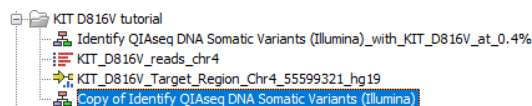


Figure 3: The "Copy of Identify QIAseq DNA Somatic Variants (Illumina)" saved in the Navigation Area.

Adjusting variant detection settings

We will first modify the variant allele frequency cutoff in the variant caller, and then we will adjust some quality filter settings.

3. If not open, then open the copy of the workflow (simply double-click on the newly saved copy to open the workflow). When the workflow is open you can double-click on the Workflow Editor tab (at the top, where the workflow name can be seen) to maximize the view to full size. Double-click on the tab again to return to the previous view size.
4. Find the **Low Frequency Variant Detection** element in the workflow using one of these methods:
 - Use the scroll bar on the right hand side of the view and find the element in the design.

- Use the mini map in the side panel to navigate to the part of the workflow where the Low Frequency Variant Detection tool is found.
- Type "Low Frequency Variant Detection" into the search field in the side panel and click on **Find**. This will put the focus on the Low Frequency Variant Detection element and bring it into view, if it is not already (see figure 4).

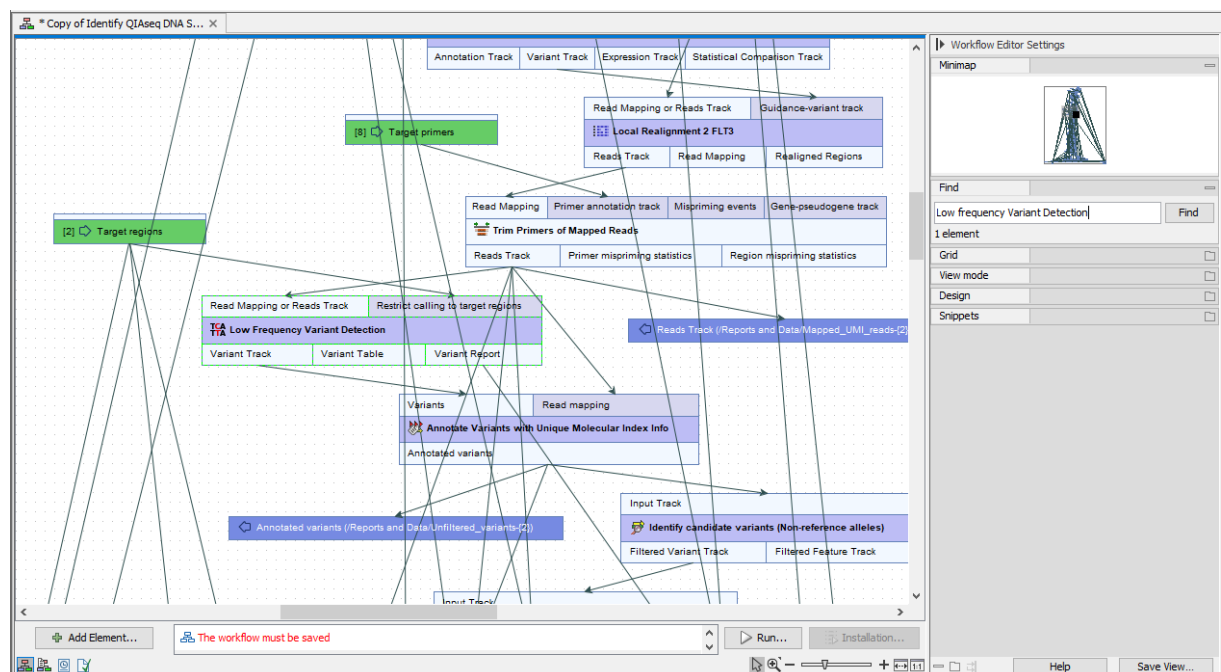


Figure 4: The Low Frequency Variant Detection element can be located using the search function in the Side Panel. Clicking on the Find button highlights the element in the workflow with green dotted lines.

5. Double-click on the name of the Low Frequency Variant Detection element.
This brings up a dialog where you can see the "Low frequency variant parameters" for the Required significance.
6. Leave these settings unchanged and click on **Next**.
7. In the "General filters" dialog, change the value for the "Minimum count" to 1 and "Minimum frequency" to 0.2. The dialog should now look like that shown in figure 5.
8. Click on **Next**.
9. Enable the **Base quality filter** by putting a check in the box next to it in the Quality filters section.
10. Adjust the remaining settings to match those shown in figure 6 and click on **Finish**.

Add and configure workflow elements specific for KIT D816V detection

We will now add a branch to the workflow containing steps specific for the detection of KIT D816V. This branch will be added below the **Identify candidate variants (Non-reference alleles)** workflow element and will contain two filtering steps and an output.

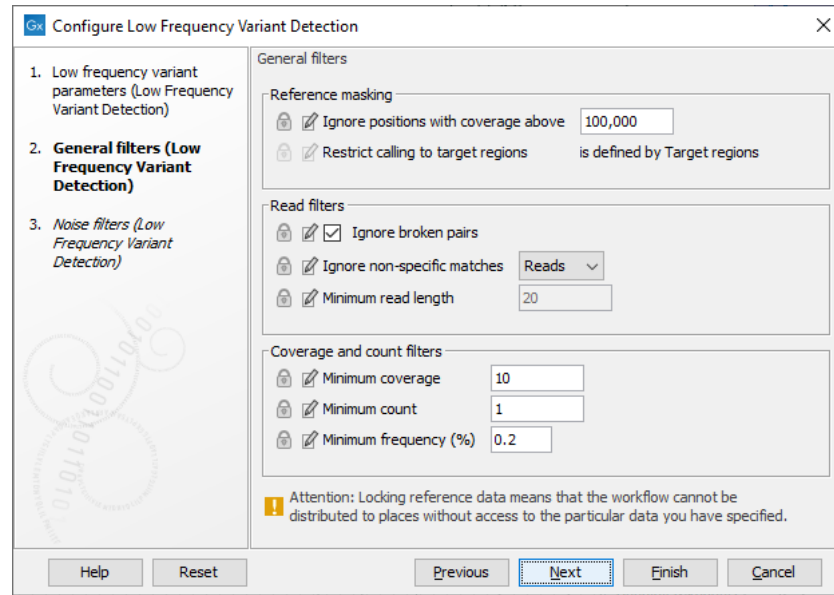


Figure 5: Adjust the settings for the Low Frequency Variant Detection step by reducing the Minimum count to 1 and the Minimum frequency to 0.2%

11. Locate the **Identify candidate variants (Non-reference alleles)** element in the workflow.
We do this so the focus will be in the vicinity of the workflow element this new branch will be connected to.
12. Right-click in a blank area of the canvas, somewhere below the **Identify candidate variants (Non-reference alleles)** element, and choose the option **Add elements...** from the menu that appears, as shown in figure 7.
13. Start typing part of the name **Filter Based on Overlap** in the search field at the top of the **Add Elements** dialog, and select that tool from the list (see figure 8).
14. Double-click on **Filter Based on Overlap** to add it to the workflow, as shown in figure 9.
The text in the newly added tool is red, signaling that further configuration is needed before the workflow can be run.
15. Rename the element you just added by right-clicking on its name and selecting the option "Rename" in the menu that appears, as shown in figure 10.
Add "(KIT D816V only)" to the existing tool name.
This is optional, but renaming elements can make it easier to keep track of the purpose of individual steps in the workflow.
16. Right-click on the "Overlap track" input channel of the **Filter Based on Overlap (KIT D816V only)** element and select the option "Connect to Workflow Input" from the menu that appears, as shown in figure 11.
Workflow inputs are top level inputs. The data provided to these inputs is either configured in the workflow itself, or is requested when launching the workflow. We will provide the target region track for KIT D816V via this Workflow Input.

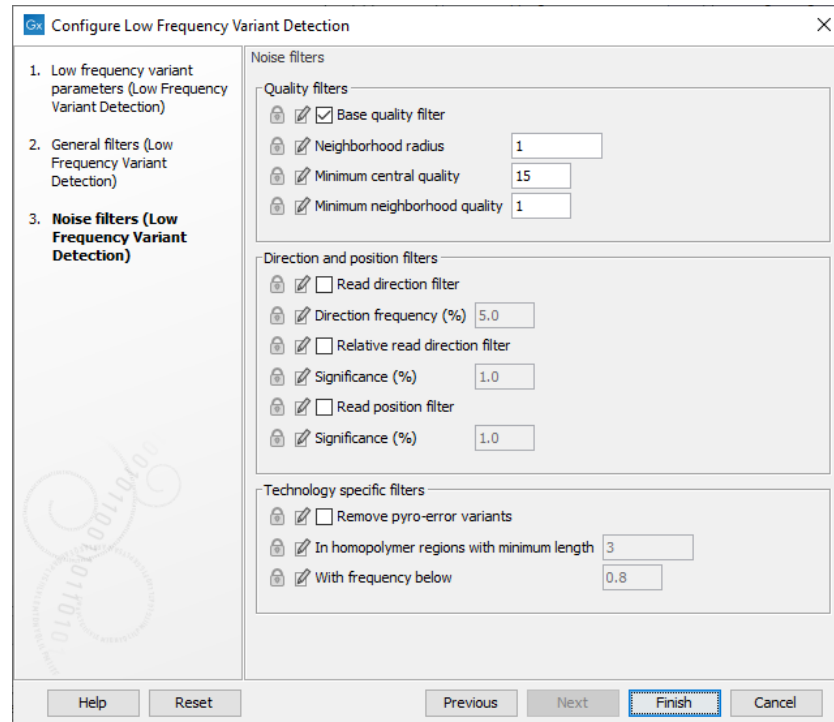


Figure 6: Adjust the quality settings for the Low Frequency Variant Detection step. Here, a base quality filter is enabled, which can be useful when detecting variants present at a low frequency.

17. Double-click on the newly added Workflow Input element, and in the wizard that appears, click on the folder icon to the right of the Workflow Input field.

Select the target region track imported with the tutorial data: KIT_D816V_Ch4_55599321_hg19 and click ok twice to close the wizard.

This track defines the target region of interest: chromosome 4, position 55599321.

We will now configure the filter settings for newly added workflow element.

18. Double-click on the name of the **Filter Based on Overlap (KIT D816V only)** element to get access to the filter settings.
19. Ensure that the option "Keep overlapping" setting is set to "Keep annotations that overlap".
20. Close the dialog by clicking on **Finish**.

This workflow element is now configured so that only variants that overlap with the KIT D816V target region, or in other words, only variants found on chromosome 4 in position 55599321, will be kept.

21. Add another new element to the workflow, this time, **Filter on Custom Criteria**.

With this tool, you can specify additional filter criteria that will be applied to variants detected at the target position, chr4:55599321.

22. Rename this element as described above, naming it "Identify KIT D816V".
23. Double-click on the element name to open a dialog where the filter criteria can be specified.

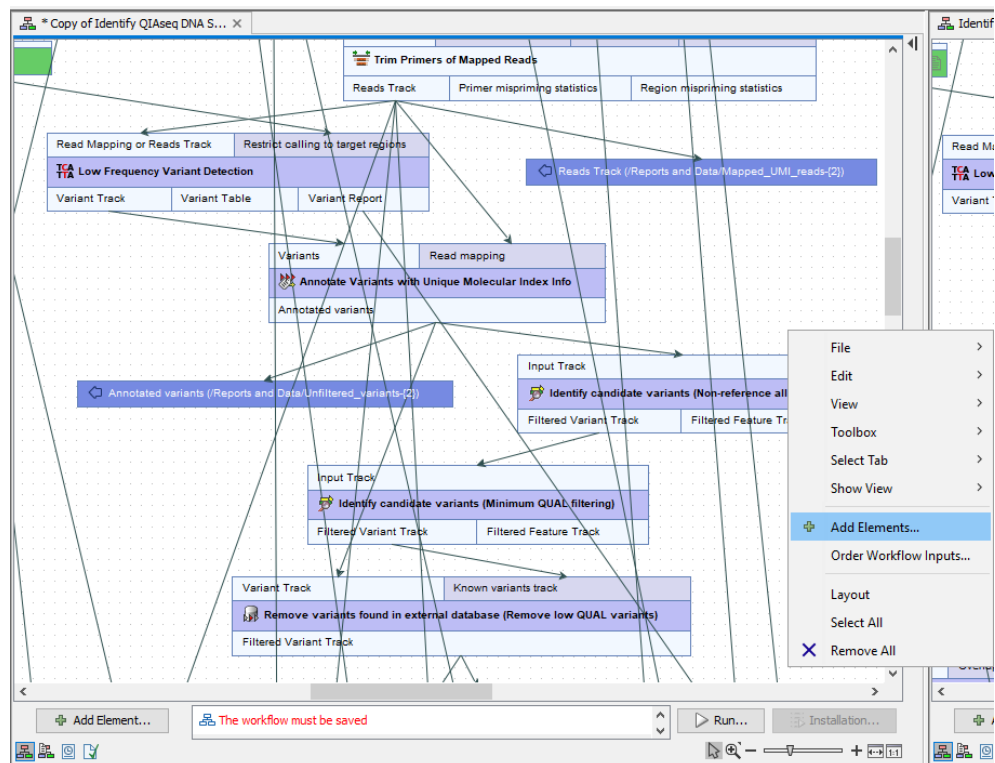


Figure 7: New elements can be added to a workflow using the Add elements dialog.

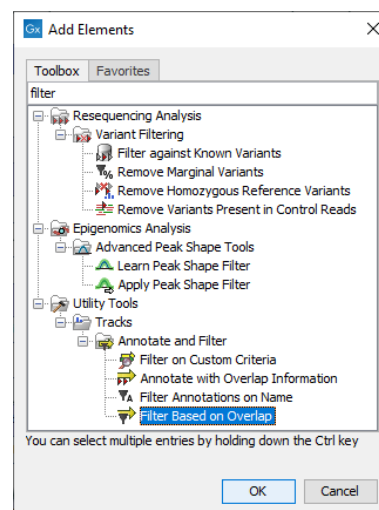


Figure 8: To restrict the list of workflow elements to those of most interest, start typing the name of the tool in the search field at the top of the Add Elements dialog. In this case it was enough to type in the word "Filter" to see "Filter Based on Overlap" listed.

24. Add the filter criteria shown in figure 12.

It is possible to adjust the number of available rows by clicking the green plus or the red x on the right hand side of one of the Criteria. A detailed description of how to add filter criteria can be found by clicking on **Help** in the lower, left corner in the dialog.

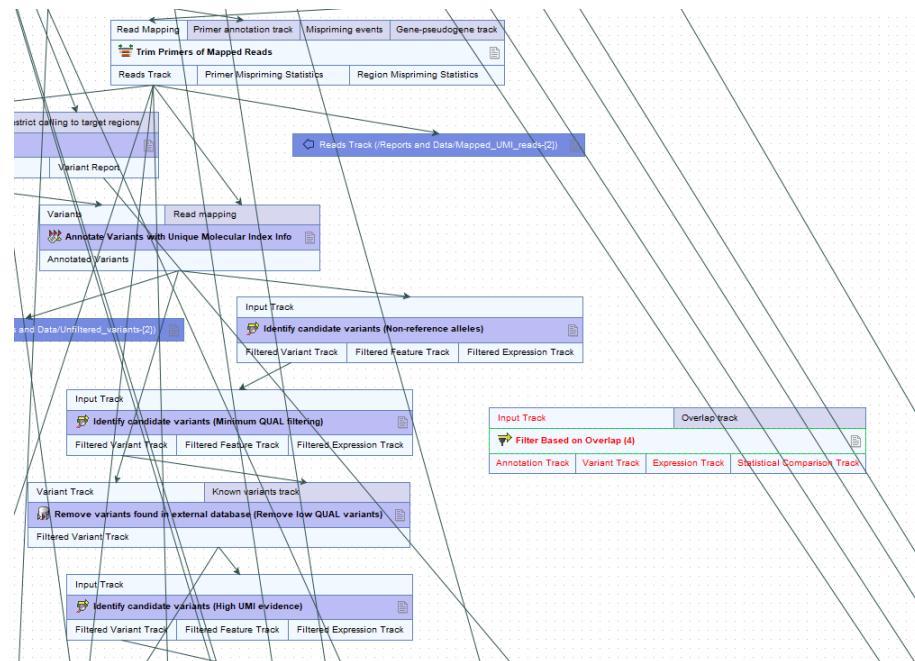


Figure 9: The element *Filter Based on Overlap* has been added to the workflow. The text in the element is red because it needs to be connected to other elements in the workflow before the workflow can be run.

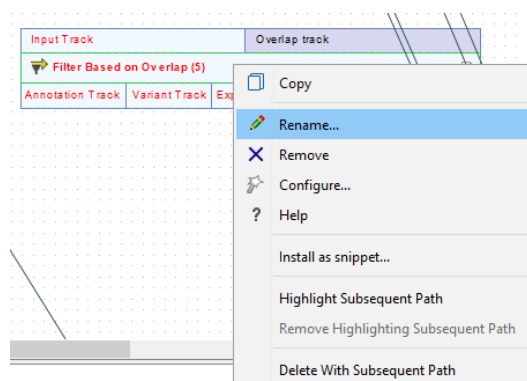


Figure 10: Rename the element "Filter Based on Overlap" to make it easier to keep track of what the aim of this step in the workflow is.

Tips when working with your own data The filter criteria suggested in this tutorial are what should be considered a good starting point. When working with your own data you may find that it will be beneficial to adjust these filter settings. Obviously, to be able to detect a variant down to 0.4% you will need good coverage, and adjusting this threshold to a low number will not increase the chances of detecting KIT D816V. Depending on the quality of your data, you can try adjusting the two quality filters applied, Average quality and QUAL. However, lowering the values for these quality filters too much increases the risk that false positive variants are detected.

Connecting the new workflow elements

The new elements now need to be connected to the rest of the workflow, as shown in figure 13.

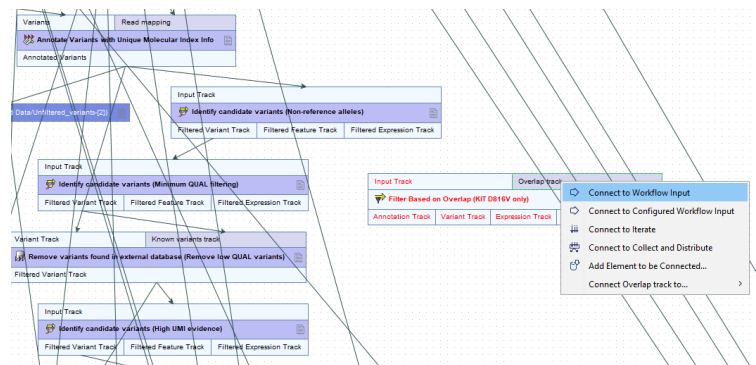


Figure 11: Connect a Workflow Input element to the "Filter Based on Overlap (KIT D816V only)" element. This will be used to specify the KIT D816V target region.

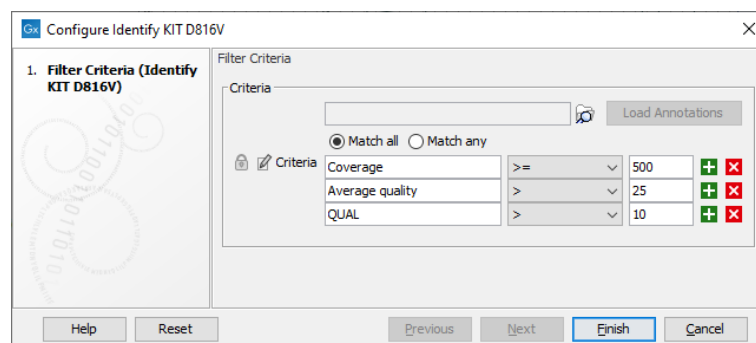


Figure 12: Specify the filter criteria for detection of KIT D816V variants.

Connections are established by connecting an output channel of one element to the input channel of another. Output channels appear at the bottom of workflow elements, and input channels appear at the top.

25. Click on the "Filtered Variant Track" output channel of the **Identify candidate variants (Non-reference alleles)** element, and keeping the mouse button depressed, drag to the "Input Track" input channel of the newly added **Filter Based on Overlap (KIT D816V only)** element.

When the "Input Track" channel is highlighted in bright green, the connection has been made, and you can release the mouse button.

You can move elements around in the workflow by dragging them, which can help ensure they are in locations convenient for connections to be made to other elements.

26. Connect the "Variant Track" output of the **Filter Based on Overlap (KIT D816V only)** to the "Input Track" input channel of the **Identify KIT D816V** element.
27. Right-click on the "Filtered Variant Track" output channel in the **Identify KIT D816V** element and choose the "Use as Workflow Output" option from the menu that appears.

This connects a Workflow Output element to that output channel. Results sent to a Workflow Output element are saved. Results generated when the workflow runs, but which are not sent to a Workflow Output element, or to an export element, are not saved.

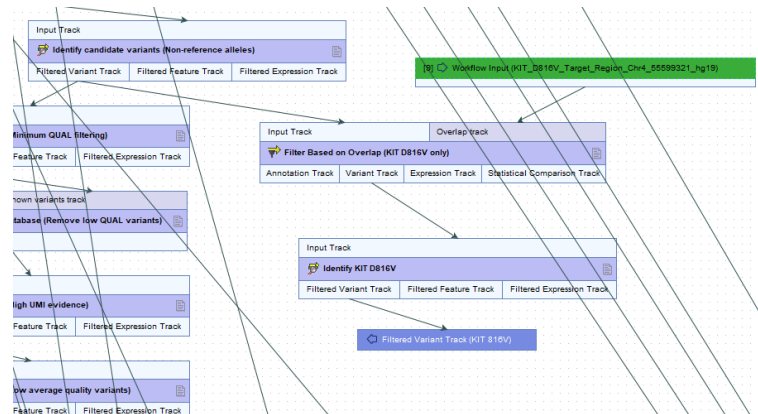


Figure 13: Connect the newly added elements by connecting their input and output channels to other workflow elements.

28. Rename the new Workflow Output element by double-clicking on it and then entering "KIT D816V" in the "Custom output name" field.

The branch for KIT D816V detection is now ready. Before running the workflow, we will add two filters to remove very low frequency variants that are not KIT D816V. Recall that at the start of this tutorial, we decreased the frequency cutoff for the variant detection step to 0.2%. Many very low frequency variants are likely to be false positives, so we will remove them from our results. These two new filters will be placed inline with and just before the "Identify candidate variants (Minimum QUAL filtering)".

Adding elements to minimize reporting of false positive variants

29. First remove the connections (highlight and delete) between **Identify candidate variants (Non-reference alleles)** and **Identify candidate variants (Minimum QUAL filtering)**. Remove also the connection between **Annotate Variants with Unique Molecular Index Info** and **Remove variants found in external database (Remove low QUAL variants)**. Then drag the tools below the **Identify candidate variants (Non-reference alleles)** a bit down to make space for the two new filtering tools to be added as shown in figure 14.
30. Add a new **Filter on Custom Criteria** element to the workflow, following the same method described above.
31. Rename this element to "Identify candidate variants (Minimum frequency filtering)".
32. Configure the filter criteria for this element so it matches that shown in figure 15.
These settings specify a frequency cutoff. Here, we are collecting a list of variants with frequencies below the value set. This is then used for comparison purposes in the next step, allowing us to filter these out of our results.
33. Connect the "Input Track" input channel of this element to the "Filtered Variant Track" output channel of the **Identify candidate variants (Non-reference alleles)** element.
34. Add a new element **Filter against Known Variants** element to the workflow.

Tutorial

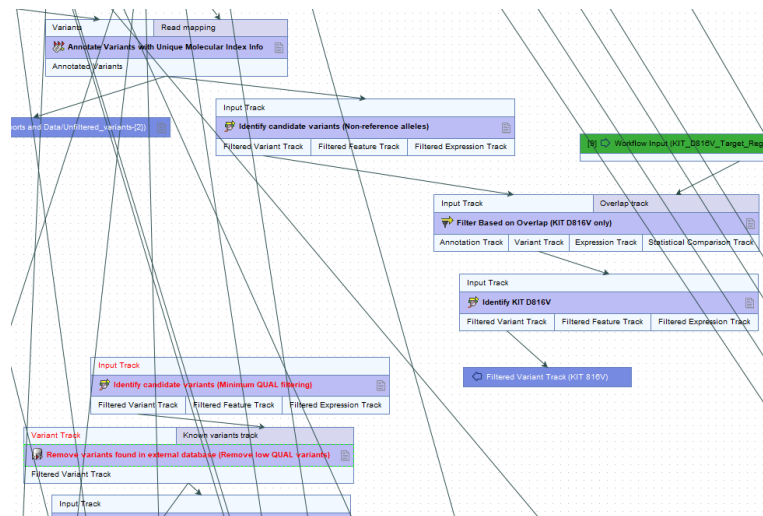


Figure 14: The connection have been removed and tools have been moved down to make space for the two new filtering tools.

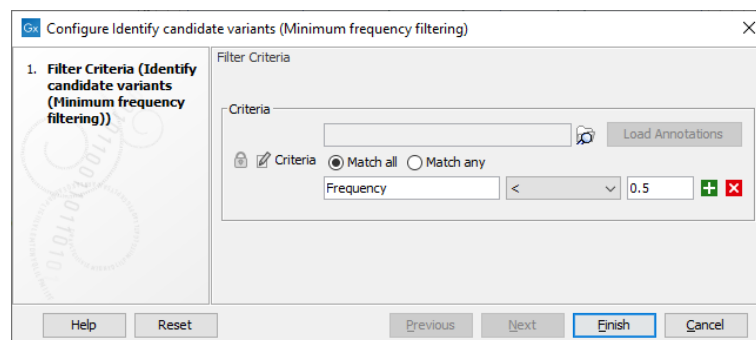


Figure 15: Add a frequency cutoff to filter for variants with a frequency below 0.5%.

35. Rename this element to "Remove variants found in external database (Remove low frequency variants)".
36. Connect the "Filtered Variant Track" output channel of the **Identify candidate variants (Minimum frequency filtering)** element to the "Known variants track" input channel of the **Remove variants found in external database (Remove low frequency variants)** element.
37. Connect the "Annotated Variants" output channel of the **Annotate Variants with Unique Molecular Index Info** element to the "Variant Track" input channel of the **Remove variants found in external database (Remove low frequency variants)** element.
38. Connect the "Filtered Variant Track" output channel of the **Remove variants found in external database (Remove low frequency variants)** to both the "Input Track" input channel of **Identify candidate variants (Minimum QUAL filtering)** and to the "Variant Track" input channel of **Remove variants found in external database (Remove low QUAL variants)**.
39. Finally configure the filter criteria for the **Remove variants found in external database (Remove low frequency variants)** element, enabling the "Join adjacent MNVs and SNVs"

and selecting the Filter action: "Keep variants with no exact match found in the track of known variants". The configuration should match that shown in figure 16. Click **Finish**

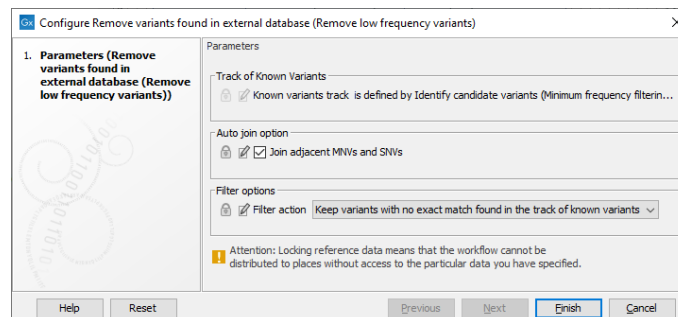


Figure 16: By selecting "Keep variants with no exact match found in the track of known variants", you filter away variants that have a frequency below the frequency filter cutoff that you specified in the previous tool (0.5%).

The area of the workflow including the newly added element should now look like that shown in figure 17.

40. Save the workflow by going to the menu at the top of the Workbench and choosing:

File | Save

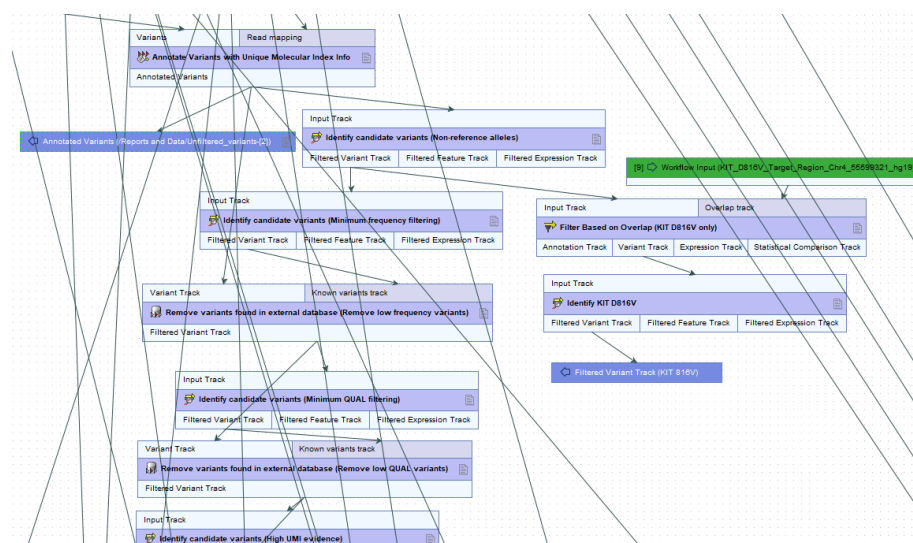


Figure 17: When the two newly added elements have been renamed and connected, the workflow will look like this.

The result of these latest additions to the workflow is a that list of variants with frequency below 0.5% is compared to a list all detected variants present at this point, and the low frequency variants are removed.

Running the workflow

You are now ready to run the workflow using the sample reads provided with the tutorial data.

Tutorial

We will launch the workflow from the Workflow Editor. This is done by clicking on the **Run** button in the bottom, right hand side of the editor, shown in figure 18.

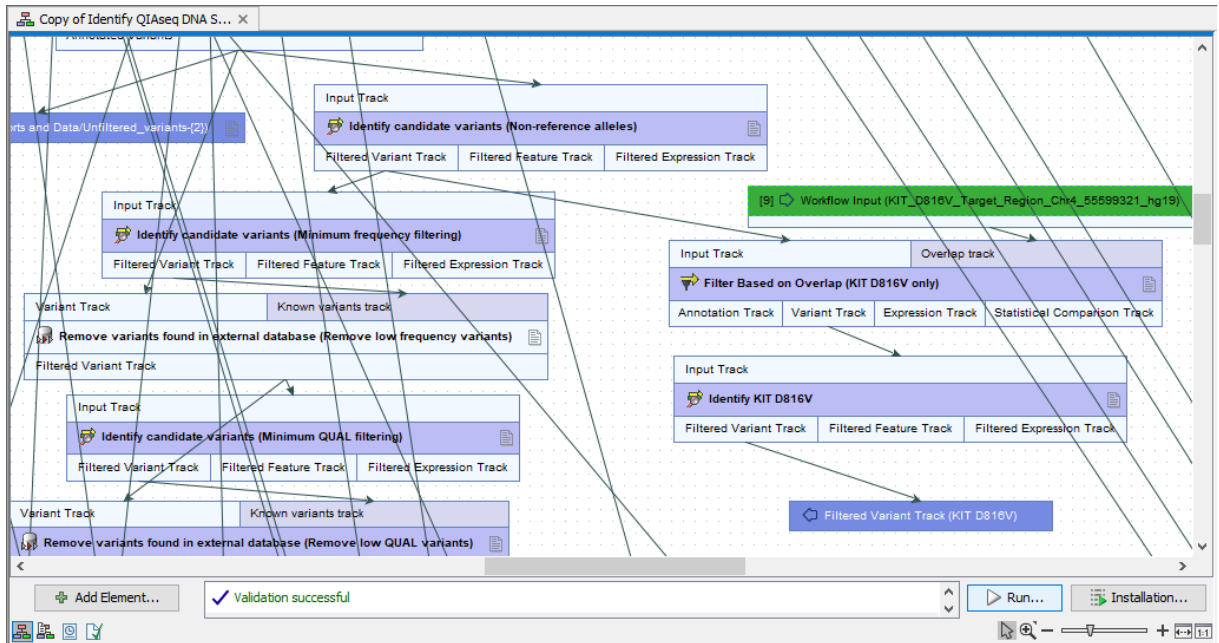


Figure 18: The workflow can be launched by clicking on the Run button in the lower, right hand corner of the Workflow Editor.

1. Click on **Run** in the lower, right corner of the Workflow Editor to launch the workflow.
2. If you are connected to a CLC Genomics Server, the first dialog will allow you to choose where to run the workflow. If so, select where it should be run and click on **Next**.
3. Select the sequence list called KIT_D816V_reads_chr4, as shown in figure 19, and click on **Next**.

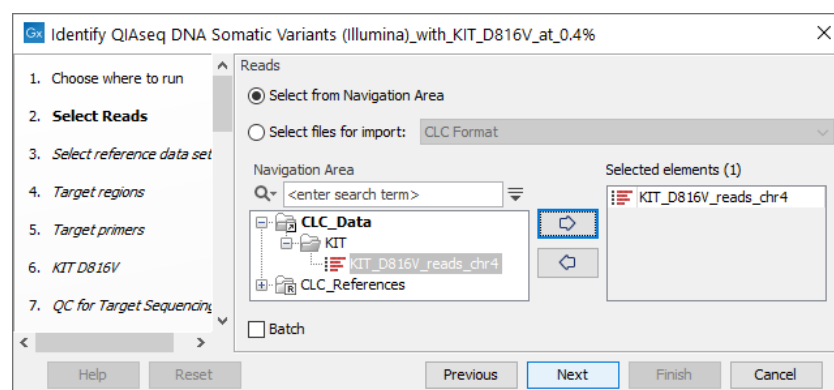


Figure 19: Select the sequence list imported from tutorial data by double-clicking on the file in the Navigation Area or by selecting the data element and clicking on the arrow pointing to the right in the middle of the dialog.

4. Choose the option "Select a reference set to use" and click on "QIAseq DNA Panels hg19 Ensembl v87".

5. Click on **Next**.

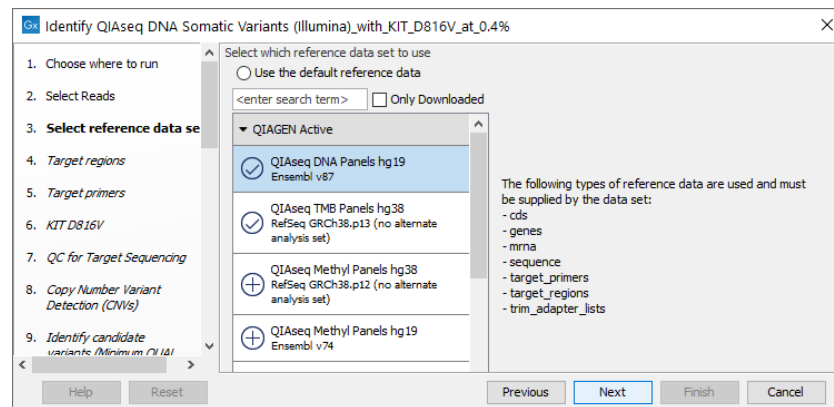


Figure 20: Select the "QIAseq DNA Panels hg19 Ensembl v87" reference set.

6. In the drop-down list of target regions, select "DHS-003Z_target_regions", as shown in figure 21 and click on **Next**.

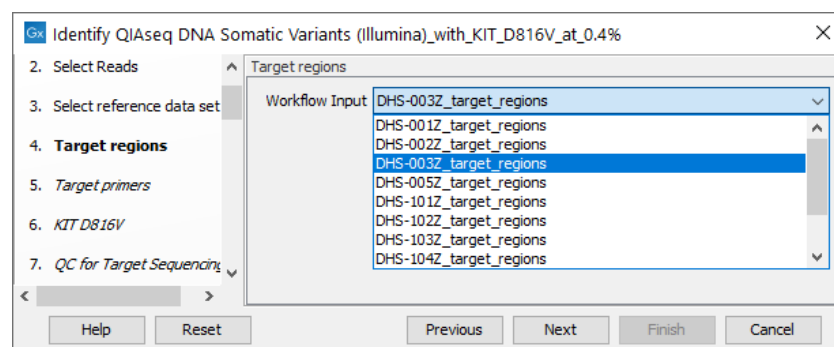


Figure 21: Select the DHS-003Z target regions.

7. In the drop-down list of target primers, select "DHS-003Z_panel_primers", as shown in figure 22 and click on **Next**.

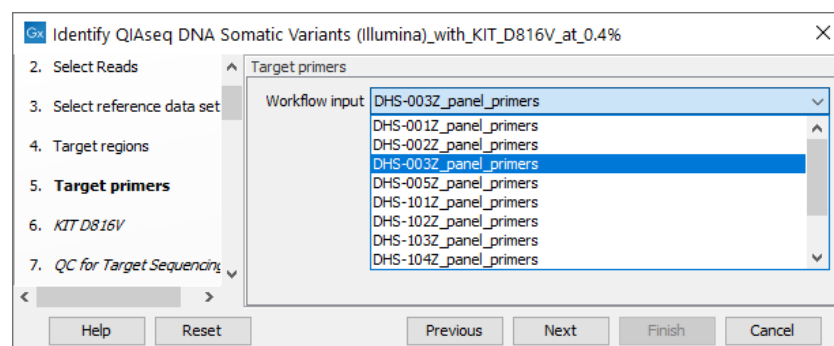


Figure 22: Select the DHS-003Z panel primers.

8. Click on **Next** in all the following dialogs until you reach the step where you are asked where to save your data.

9. Select a location to save the results to and click on **Finish**.

The workflow will now run. You can monitor the progress of the workflow in the "Processes" tab in the bottom, left side of the Workbench.

Reviewing the results

1. When the workflow run has completed, open the output variant track called "KIT D816V" by double-clicking on the data element name in the Navigation Area.

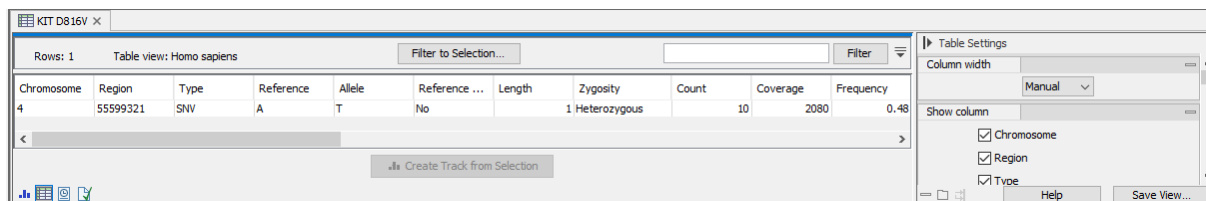
If you have trouble finding the workflow results:

- (a) Open the "Processes" tab in the bottom, left side of the Workbench.
- (b) Find the entry corresponding to the workflow run, and click on the small black triangle on the right hand side.
- (c) Choose the option "Find Results" in the menu that appears.

The results generated by the workflow will be highlighted in the Navigation Area.

2. Click on the table icon in the lower left corner of the variant track view.

You should now see what is shown in figure 23, where the KIT variant has been detected with a frequency of 0.48%.



Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Coverage	Frequency
4	55599321	SNV	A	T	No		1 Heterozygous	10	2080	0.48

Figure 23: Double-clicking on the KIT D816V variant track opens it in track view. Click on the table icon in the lower left corner of the view to open the table view of this data element.

The KIT variant can be viewed in context, that is, together with the mapped reads and the reference data, by opening the workflow output called Genome_Browser_View and adding the KIT D816V variant track to it. To do this:

1. Open the workflow output called Genome Browser View by double clicking on its name in the Navigation Area.
2. Click on the name of the KIT D816V track in the Navigation Area, and keeping the mouse button depressed, drag it into the Genome Browser view, and then let go of the mouse button.
3. When the KIT D816V variant track has been added to the Genome Browser view, double-click on the name of the KIT D816V track in the left-hand side of the Genome Browser view to open the variant track. The variant track will open in table view below the Genome Browser view. As the two views are connected, clicking on the variant in the table view will automatically bring the KIT D816V variant into focus (see figure 24).

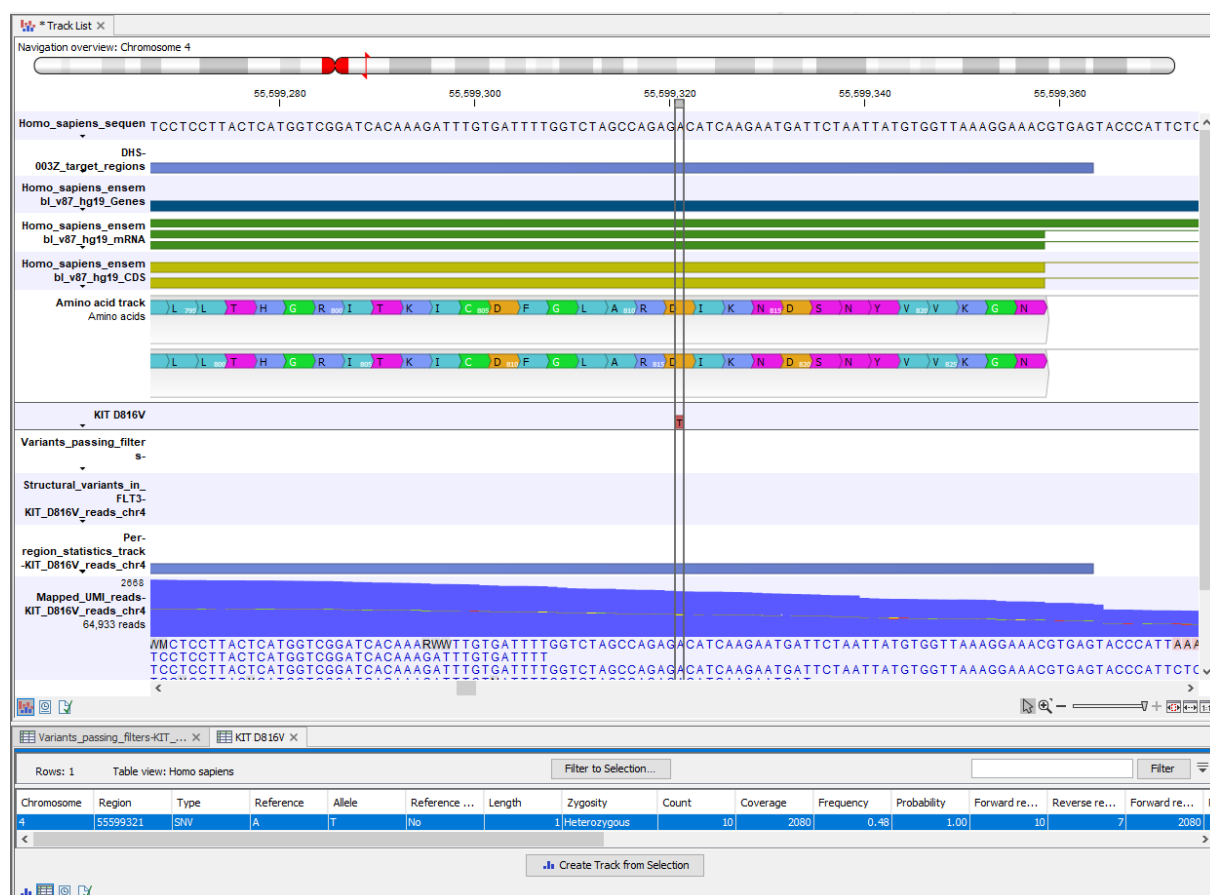


Figure 24: To view the KIT D816V variant together with the reference data and the mapped reads, drag the variant track into the Genome Browser view.

The approach outlined in this tutorial can be used to edit any existing workflow to create analysis pipelines addressing specific analysis needs, and it is particularly useful for customizing complex workflows already available, such as those distributed with the Biomedical Genomics Analysis plugin.

The specific customizations undertaken in this tutorial, with use of the target region track KIT_D816V_Target_Region_Ch4_55599321_hg19, allowed us to focus on the KIT D816V variant. By adapting the modified workflow to use a different target region track, the focus could, of course, be put on other low frequency variants of interest.

In addition, workflows can be launched from the Workflow Editor, as described in this tutorial, but they can also be installed on your Workbench or CLC Genomics Server. Once installed, they can be launched from the Toolbox. Workflow installers can also be shared with others, for installation on other systems. For further information about installing workflows, please refer to the [manual](#).