



Tutorial

An Introduction to Workflows

April 6, 2020

— Sample to Insight —

An Introduction to Workflows

A workflow consists of a series of connected tools where the output of one tool is used as input for another tool, making it possible to analyze many samples using a standardized pipeline. This tutorial covers:

- **Setting up a workflow** that maps reads to a reference sequence, detects variants in the data relative to the reference, and filters away variants that are also present in a database of common variants
- **Launching a workflow in batch mode** to analyze multiple samples
- **Workflow management**, including creating a workflow installer, for installing the workflow on your own or another *CLC Workbench*, or on a *CLC Server*, where all or a selected group of server users could make use of it

Steps where action should be taken are numbered.

Prerequisites: We recommend using *CLC Genomics Workbench* 20.0 or higher when working through this tutorial. Some functionality referred to is not available in earlier versions.

Download and import data for the tutorial

The example data for this tutorial includes two files of sequencing reads and several reference tracks: the human mitochondrial genome from the hg18 build, NC_001807 (Genome), corresponding CDS and Gene tracks, and a chrMdbSNPCommon track containing the dbSNP common variants for this mitochondrial sequence.

1. Download the example data from <http://resources.qiagenbioinformatics.com/testdata/chrM-tutorial-data.zip>.
2. Start the *CLC Genomics Workbench*.
3. Import the data by going to:
File | Import (📁) | Standard Import (📁)
4. Choose the zip file you just downloaded. Leave the Import type set to **Automatic**.

After import, the files listed in the **Navigation Area** should look like those shown in figure 1.

General tips

- You can access the manual by clicking on **Help** buttons, available in the Workflow Editor and in workflow element configuration wizards. It can also be accessed by selecting the "Help" option under the "Help" menu.
- To select multiple items simultaneously in lists, or elements in a workflow editor, hold down the Ctrl key (⌘ on Mac) when selecting each item.

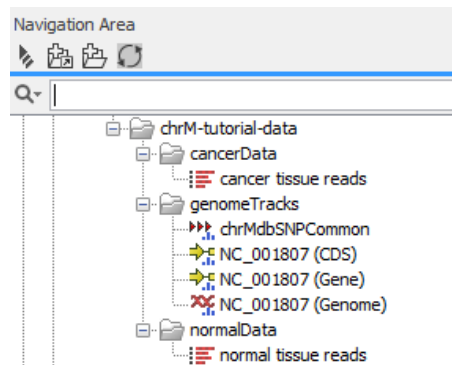


Figure 1: Navigation area upon import of files.

- The layout of a workflow in the Workflow Editor can be adjusted manually by clicking on workflow elements to select them, and then with the mouse button depressed, dragging them to the location you want. Alternatively, the workflow layout can be updated by right-clicking anywhere in the Workflow Editor and selecting the option **Layout** from the menu that appears.

Create a workflow

Setting up a workflow involves the following steps, which we describe in detail in the tutorial:

- Open the Workflow Editor.
- Add tools to the workflow and connect them in the relevant order.
- Configure the individual tools with the desired settings.
- Connect Input elements for inputs to the workflow.
- Connect Output elements for outputs that should be saved from a workflow run.

Open the Workflow Editor

1. Open the Workflow Editor by going to **File | New | Workflow** (🔗) or clicking on the **Workflows** (🔗) button in the top toolbar and choose the option **New Workflow**.

Add tools to the workflow

Two ways to add tools to a workflow are:

- Open the Toolbox tab in the bottom left side of the Workbench, click on a tool and, keeping the mouse depressed, drag it into the Workflow Editor.
- Use the **Add Elements** dialog, launched from the Workflow Editor.

Steps below use the second of these methods, but you can use either.

2. Click the button labeled **(+)Add Element...** in the bottom left corner of the Workflow Editor.
3. Under the **Resequencing Analysis** folder, select the following tools, as shown in figure 2:
 - **Map Reads to Reference**
 - **Local Realignment**
 - **Fixed Ploidy Variant Detection** from within the **Variant Detection** subfolder
 - **Filter against Known Variants** from within the **Variant Filtering** subfolder
4. Click on **OK**.

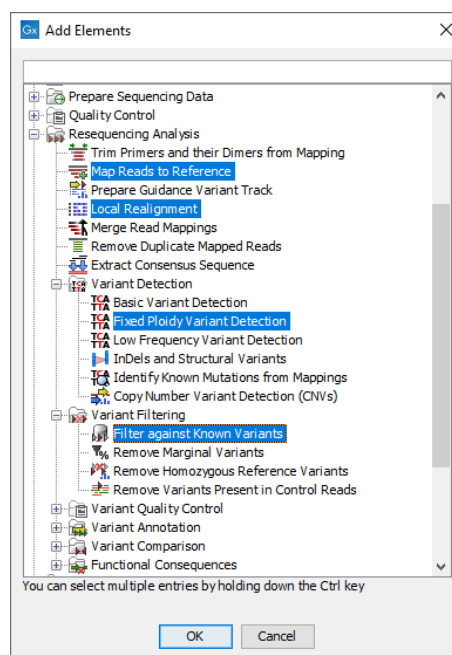


Figure 2: Tools selected for inclusion in a workflow using the Add Elements dialog.

If you need to add more tools later on, just click on the **Add Element...** button again, or drag them from the Toolbox.

To move workflow elements around in the editor, select them, and keeping the left mouse button depressed, drag them to the desired location. Figure 3 shows the workflow elements after they have been moved around in the Workflow Editor.

Each of these workflow elements consists of 3 areas:

- **Input channels**, along the top. Light grey input channels must be connected to another workflow element. For the first element of a workflow, the light grey input channel must be connected to an Input element (described in a later section). Light purple input channels can be connected to other elements, or left as they are. These input channels can also be pre-configured if desired.
- **Output channels**, along the bottom. These can be connected to input channels of other elements, connected to **Output** elements, or left unconnected. At least one output channel must be connected to an Output element for a workflow to be considered complete.

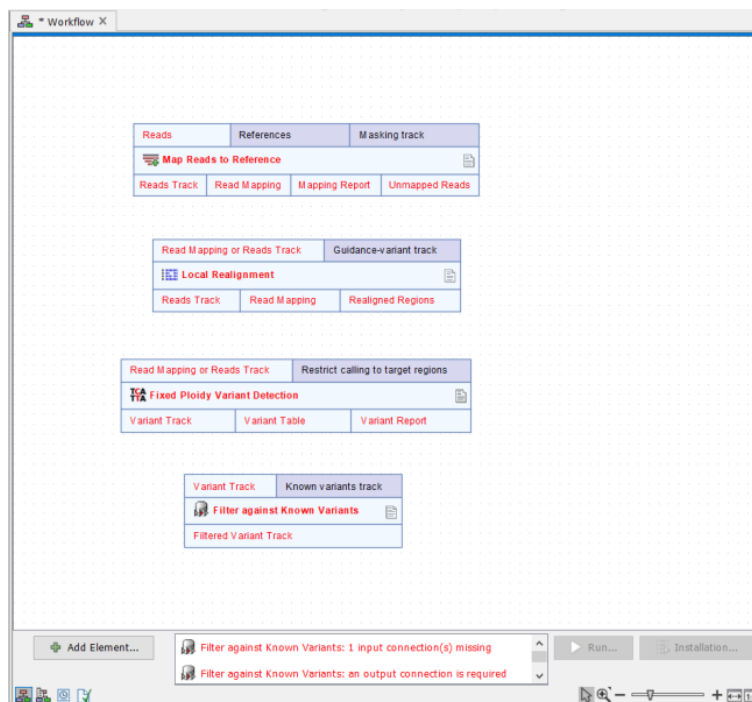


Figure 3: The selected workflow elements in the Workflow Editor

- **The core section**, in the middle (where the element name is). Double-clicking on this area of an element with a small page symbol to on the right hand side opens a configuration wizard.

Right-clicking on any of these areas brings up a menu with options relevant to it.

Detailed information about workflow elements and their configuration is provided in the manual at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflow_elements.html.

Connect the workflow elements

Workflow elements are connected from their output channels to input channels of downstream elements. In addition, Input elements are connected to input channels where the input should be specified when launching the workflow, and Output elements are connected to output channels where results should be saved.

5. The input to the **Map Reads to Reference** element needs to be specified when the workflow is launched, so connect an Input element to the light grey input channel:
 - (a) Right-click on the **Reads** input channel.
 - (b) Select the option **Connect to Workflow Input (🔌)**.

To move the Input element, click on the thin bar at the top, and keeping the mouse button depressed, drag the element.

6. Configure mapping output from **Map Reads to Reference** as input to **Local Realignment** by clicking on the **Reads track** output channel of the Map Reads to Reference element and, keeping the mouse button depressed, dragging to the **Read Mapping or Reads Track** input channel of the Local Realignment tool.
7. Configure mapping output from **Local Realignment** as input to **Fixed Ploidy Variant Detection** by clicking on the **Reads track** output channel of the Local Realignment element and, keeping the mouse button depressed, dragging to the **Read Mapping or Reads Track** input channel of the Fixed Ploidy Variant Detection element.
8. Configure output from the **Fixed Ploidy Variant Detection** as input to **Filter against Known Variants** by clicking on the **Variant Track** output channel of the Fixed Ploidy Variant Detection element and, keeping the mouse button depressed, dragging to the **Variant Track** input channel of the Filter against Known Variants element.
9. Save the variants generated by **Filter against Known Variants** by connecting an Output element:
 - Right-click the output channel labeled **Filtered Variant Track** and select the menu option **Use as Workflow Output** (🔗).
 - Double-click on the Output element. Hover the mouse cursor over the **Custom output name** field to see the variables you can use to configure the output names.
 - Click in the **Custom output name** field and click on the Shift and F1 buttons. Select the option {input}.
This means the output saved have a name based on the name of input to the workflow. You can also provide specific text to use as the output name or a part of the output name by typing into this field if you wish.
 - Click on **Finish**.

At this point, the workflow should look like the one shown in figure 4.

10. Configure the workflow to save the following outputs by connecting Output elements to:
 - The **Reads Track** and **Mapping Report** output channels of **Map Reads to Reference**
 - The **Reads Track** output channel of **Local Realignment**
 - The **Variant Track** output channel of **Fixed Ploidy Variant Detection**

The workflow should now resemble that shown in figure 5.

Configuring workflow elements

Workflow elements can be configured with reference data or with values for options, such that these are used whenever the workflow is launched. Each setting can be locked, so it cannot be reconfigured when the workflow is launched, or left unlocked, so that it can be. No prompt for locked settings is provided in the wizard when launching.

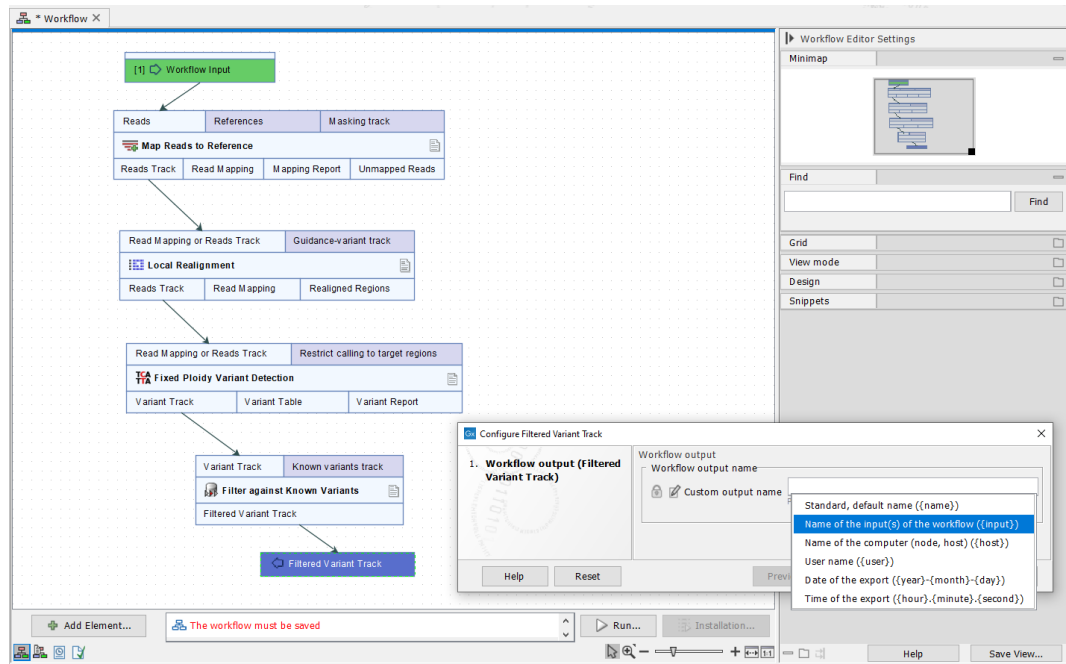


Figure 4: A Workflow Input element and a Workflow Output element have been added, and all the elements in this workflow are connected, as indicated by the arrows between output and input channels.

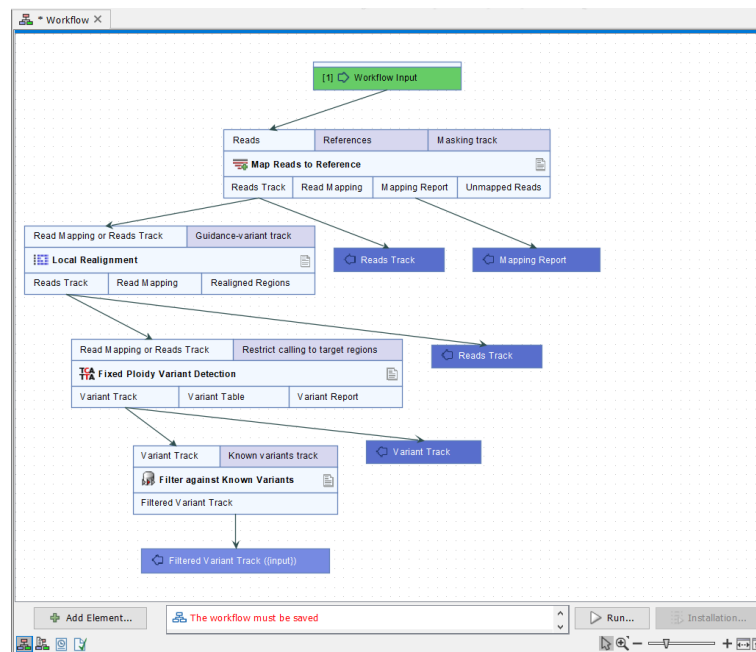


Figure 5: The finished workflow

Configure the Map Reads to Reference element

Specify the reference sequence the reads should be mapped against:

11. Double-click on the **Map Reads to Reference** element to open the configuration wizard.

12. Click the **Browse button** (🔍) to the right of the References field.
A window labelled "Configure references and/or its workflow role" appears, as shown in figure 6.
13. Click on the **Browse button** (🔍) to the right of the "Workflow Input" field and select **NC_001807(Genome)**.
14. Click on **OK** to close this dialog.

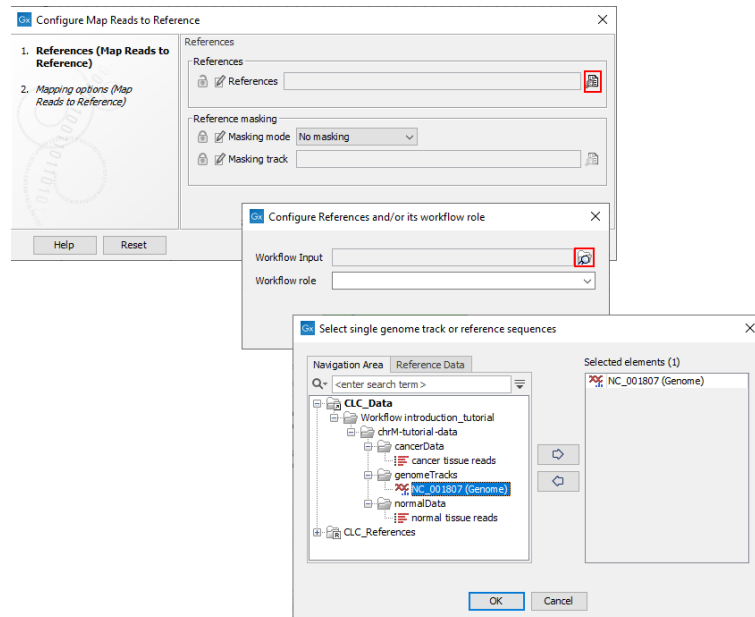


Figure 6: Configuring the Map Reads to Reference element with the reference sequence that the reads should be mapped to.

15. Leave the "Reference masking" options set to the default values and click on **Next**.
16. Configure the mapping options as shown in figure 7. By default, all these options are locked. Click on the lock icon to unlock any that should be configurable when launching the workflow.
17. Click on **Finish**.

Configure Local Realignment and Fixed Ploidy Variant Detection

For this tutorial, we will use the default values for the parameters for these tools. To look at the individual parameters, double click on the central area of the workflow element and work through the wizard steps.

Configure Filter against Known Variants

Specify the database track to filter against (figure 8):

18. Double-click on the central area of the **Filter against Known Variants** element.

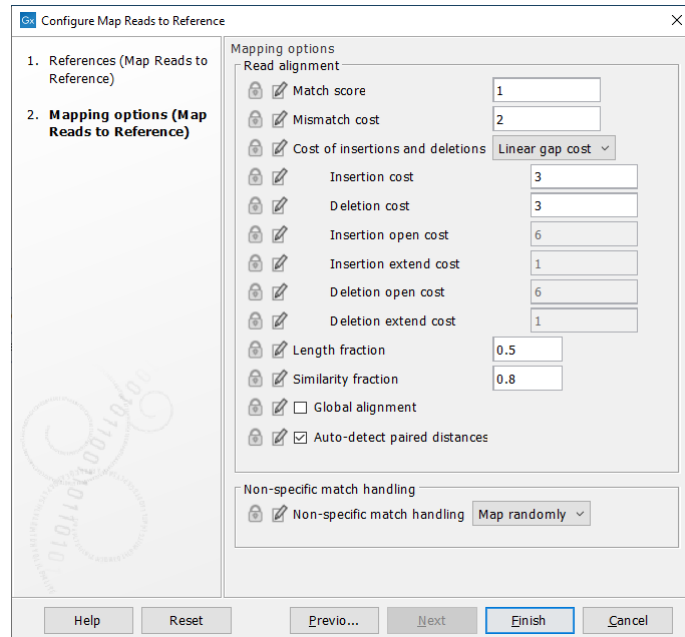



Figure 7: Configuring the mapping options.

19. In the wizard click the **Browse button** () , select **chrMdbSNPCommon** and click on **OK** to proceed.
20. Select the filter option **Keep variants with no exact match found in the track of known variants**.
21. Click on **Finish**.

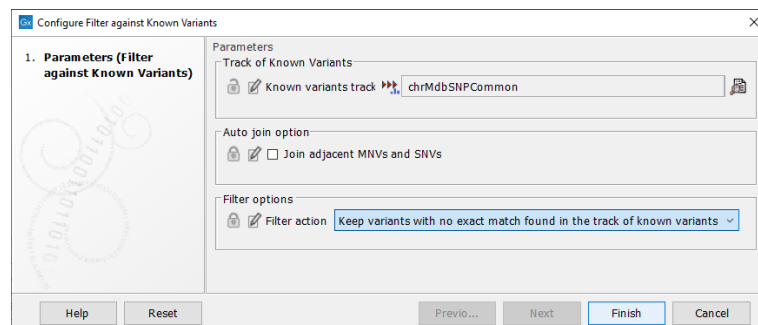


Figure 8: Configuring Filter against Known Variants

Save the workflow

Save the workflow:

22. Go to:

File | Save as ()


and then choose a name for the workflow and a location to save it to.

Other ways to save workflows and other open workbench items, include:

- Click on the tab of the view and drag it in into the folder in the Navigation Area of the workbench where you want to save it, or
- Right click on the tab at the top of the unsaved view, and choose **Save...** or **Save As...** from the menu that appears, or
- Click on the tab at the top of the unsaved view and press **Ctrl-S** on the keyboard.

When a workflow is configured correctly and saved, the message **Validation successful** appears at the bottom of the editor, the workflow name appears in the **Navigation Area**, and the **Run** button is activated.

Perform a test run


1. Click on  **Run...** at the bottom of the Workflow Editor.
2. Select the **normal tissue reads** data element from the normalData folder.
3. Click through the next couple of wizard windows by clicking on **Next**.

The references you configured earlier are preselected in the relevant dialogs. If you had locked the references configured earlier, then the option to configure these would not have been presented here.

4. At the Result Handling step, choose to:

- **Save the results.**
- **Open the log.** This allows you to see the progress of the workflow as it runs.
- **Create workflow result metadata**

Workflow Result Metadata tables list the outputs generated by a workflow run and can be used to locate and access results of interest, as described in the manual at http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflow_outputs_workflow_result_metadata_tables.html

5. Click on **Next**.
6. Click on the  button at the top of the window to create a new folder to save the results into. Name the new folder **Test run** and select that folder.
7. Click on **Finish**.

The workflow has now been launched. When it completes, outputs are saved to the **Test run** folder, as shown in figure 9.

If you wish, open up the individual files to have a look at the results. You can also access the results through the Workflow Result Metadata table. Just highlight a row or rows in the table, and then click on the "Find Associated Data" button. The relevant data elements will be listed in a table below. Double clicking on a data element opens it.

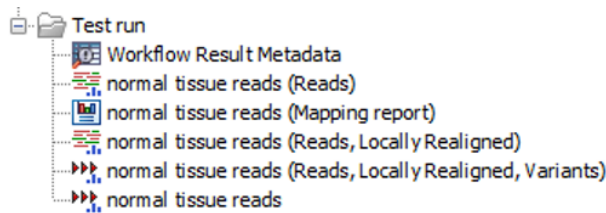


Figure 9: Outputs saved to the "Test run" folder after the workflow has completed.

Installing and managing workflows

As well as launching workflows from the Workflow Editor, we can install the workflow, and launch it from the Toolbox. This essentially locks down the version that is launched, helping ensure that the same analysis (same tools and settings) is run each time.

We step through this process here. The same initial steps are used to create a workflow installer for installing the workflow on a CLC Server or to distribute to others.

Create a workflow installer

1. Click on the **Installation** button at the bottom of the Workflow Editor.
2. Click on the **Help** button at the bottom left side to see information about the fields in this wizard.
3. Fill out the mandatory fields "Workflow name" and "Organization", and any other fields you wish to.
4. Click on **Next**.

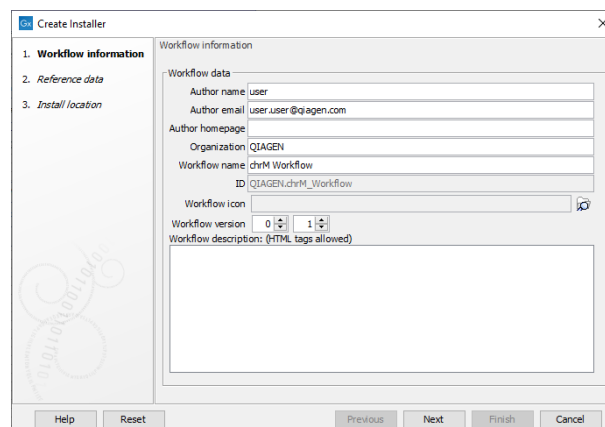


Figure 10: Name and describe the workflow you want to install.

5. Specify whether to "Ignore" or "Bundle" the reference data the workflow has been configured with (figure 11).

The option "Bundle" will include a copy of the data with the workflow installer. Click on the **Help** button at the bottom left side to see further information about these options.

6. Click on **Next**.

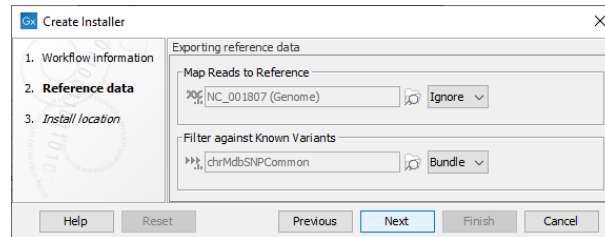


Figure 11: Reference data can be bundled with the workflow installer if desired.

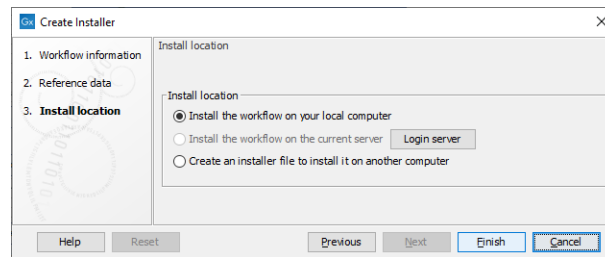


Figure 12: You can install the workflow on your local machine, the server, or you can create an installer to share with a colleague.

7. Choose to **Install the workflow on your local computer** (figure 12).
8. Click on **Finish**.
9. If you have chosen to bundle the data, you will now be prompted for the location to save the bundled reference data.

This workflow will now be available in the Installed Workflows folder of the Toolbox. If you bundled reference data, you should see it in the location you specified (figure 13).

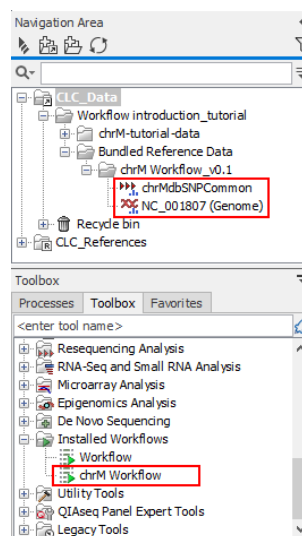


Figure 13: Installed workflows are available from "Installed workflows" folder of the Toolbox.

Managing workflows

The Workflow Manager can be launched by clicking on the **Workflows** button in the top toolbar and choosing the option **Manage Workflows** (figure 14).

Information about installed workflows, and Ready-to-Use workflows, distributed with plugins, are available in the Workflow Manager. This is also where you can install workflows using an installer file and update workflows.

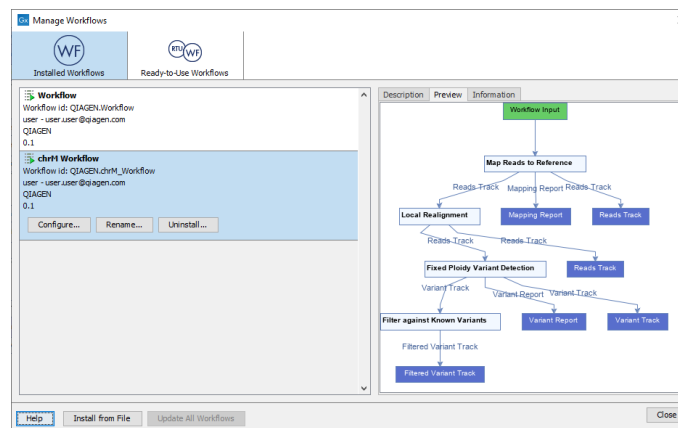


Figure 14: Installed workflows are listed in the Workflow Manager. Information about the selected workflow is shown in the right hand panel. Here, the Preview tab is open.

Launching an installed workflow and running it in batch mode

Here, we launch the **chrM workflow** from the Toolbox, and analyze two datasets serially, by running the workflow in batch mode.

1. Launch the installed workflow by going to:

Toolbox | Workflows () | chrM workflow ()

2. Check the **Batch** option in the bottom left side of the wizard.
3. Select the **chrM-tutorial-data** folder for analysis.
4. Click on **Next**.
5. Choose to "Use organization of input data" as the basis for defining the batch units.

The workflow will be launched once, but run twice, once using relevant data elements in one of the subfolders as input, and then again, using the relevant data elements in the other subfolder as input. In other words, there are 2 "batch units", the "cancerData" subfolder and the "normalData" subfolder.

6. Click on **Next**.

The Batch overview step allows you to review the batch units. If you select the cancerData batch unit, you can see that one sequence list, "cancer tissue reads" will be used as input to the workflow when this batch unit is processed.

7. Click on **Next** in the wizard until you get to the Results handling step.

8. Choose the options to **Save in input folder**, **Create workflow result metadata**, and **Open log**.
9. Click on **Finish**.

While the analysis is running, you can watch its progress in the Processes tab, located in the bottom left side of the Workbench. Alternatively, open the log file for more detailed progress information.

When the analysis is complete, you should see the results of the analyses within the cancerData and normalData folders. The Workflow Result Metadata table in each folder can be used to see a listing of the results and access individual results, as described earlier in the tutorial.

Visualize track results in a Track List

Workflows can include Track List elements, where reference and result tracks can be gathered for easy visualization and comparative analysis. Output channels connected to a Track List element must also be connected to an Output element. In figure 15, we show the workflow used earlier in the tutorial after a Track List element has been added. Outputs from several of the analysis elements have been used as inputs to the Track List. A couple of Input elements have also been connected, configured with reference data tracks. An example of a Track List generated using the tutorial data and this analysis workflow is shown in figure 16.

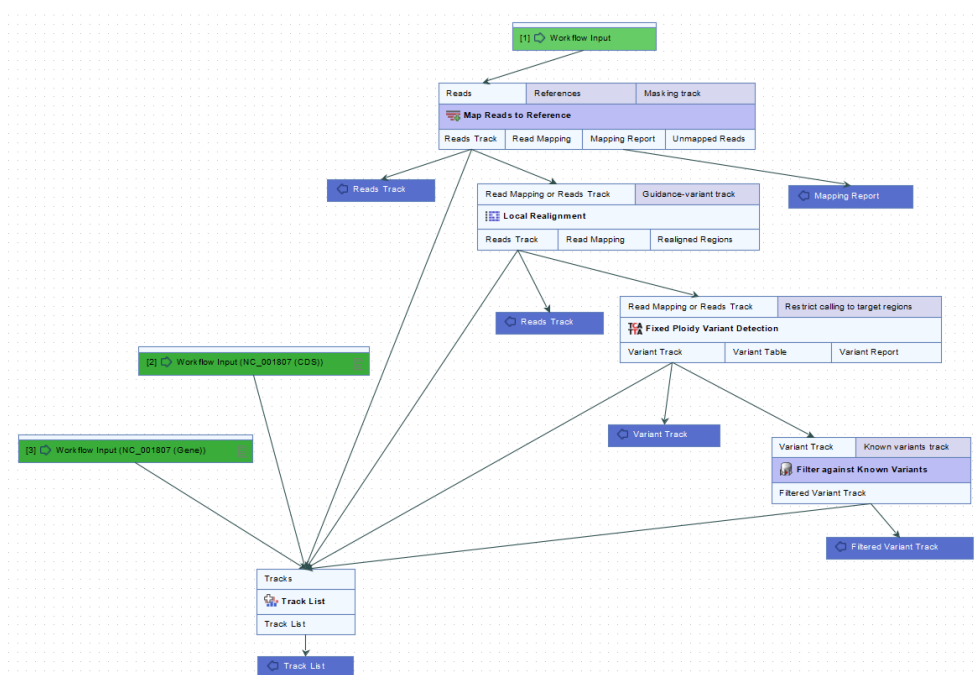


Figure 15: A *Track List* element has been added, taking analysis outputs and reference tracks as inputs.

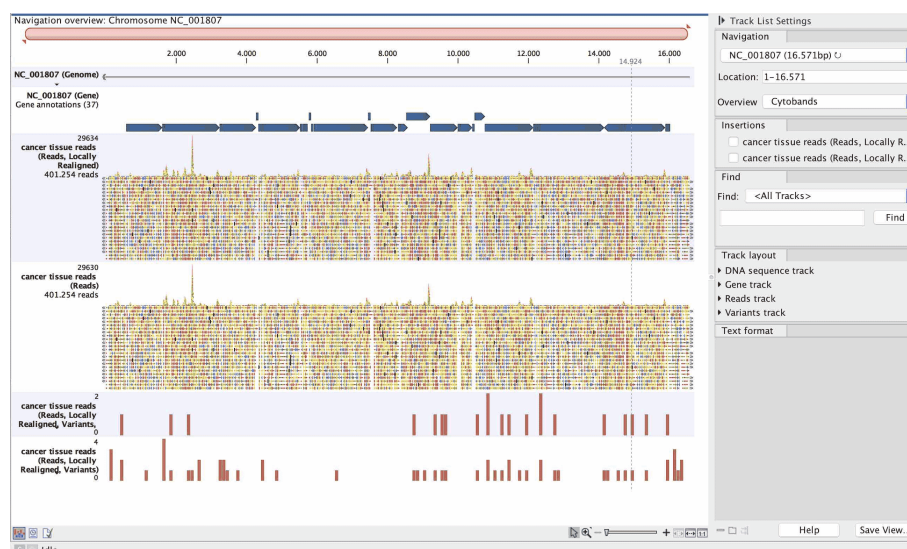


Figure 16: A track list created using the workflow shown in figure 15.