Tutorial

An Introduction to Workflows

September 15, 2016

Sample to Insight

CLC bio, a QIAGEN Company · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.clcbio.com · support-clcbio@qiagen.com

An Introduction to Workflows

This tutorial takes you through how to set up and run a workflow in the *CLC Genomics Workbench*. A workflow consists of a series of tools in which the output of one tool is connected to and used as the input of another tool. A workflow is a convenient way to automate your own analysis pipeline or to distribute it to colleagues. Also, workflows can be installed on a CLC Genomics Server making it available to a larger group of users.

We will be putting together a workflow that includes mapping reads to a reference, detection of variants and filtering of these variants for common ones.

Workbench versions To create a workflow, you must be working with the *CLC Genomics Workbench*, version 5.5 or higher. This tutorial uses tools available in the Genomics Workbench 7.5. If you are working through this tutorial with a *CLC Genomics Workbench* other than version 7.5, the precise locations and names of buttons and tools may be slightly different than described in this document.

Overview Setting up a workflow entails:

- Choosing the tools to be included
- Connecting the tools
- Selecting points for output
- Configuring the individual tools - that is, setting tool parameters and selecting dependent items
- Performing a test run
- Creating a workflow installer
- Installing the workflow in the workbench

Once the workflow as been set up we will use it to launch the analysis of two samples, using the batch mode to minimize the hands-on effort.

Downloading and importing the data

First, we need to download and import the data, which will be used for configuring and running the workflows.

1. Download the sample data from our web site: `http://download.clcbio.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 named chrM-tutorial-data.zip. Leave the Import type set to **Automatic**.

The data set includes two sequencing data files (normal tissue reads and cancer tissue reads) as well as a list of tracks.

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

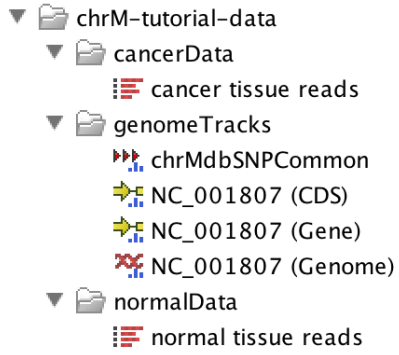

Figure 1: *Navigation area upon import of files.*

The tracks include the human mitochondrial genome from the hg18 build, **NC_001807 (Genome)** sequence track as well as CDS, Gene and mRNA tracks for this reference. Also included are the **chrMdbSNPCommon** track, which contains the dbSNP common variants for the mitochondrial sequence.

Creating a workflow

Selecting the tools In this section we select the tools to be included in our workflow.

1. To start building a workflow go to:

   File | **New** | **Workflow (** **)**

2. Click the button labeled **( )Add Element...**.

3. Select the following tools:

   - **Map Reads to Reference**
   - **Local Realignment**
   - **Fixed Ploidy Variant Detection**
   - **Filter against Known Variants**.

   To select multiple entries please hold down the Ctrl key (⌘ on Mac) (figure 2). If you need to add additional tools later, just click on the **Add Element...** button at the bottom of the workflow editor.

4. Click on the button labeled **OK**. In the window you will now see boxes representing the tools (figure 3). You can click on each box and drag it around in the workflow editor to position the tasks where you want them.

Each tool is displayed as a set of boxes:

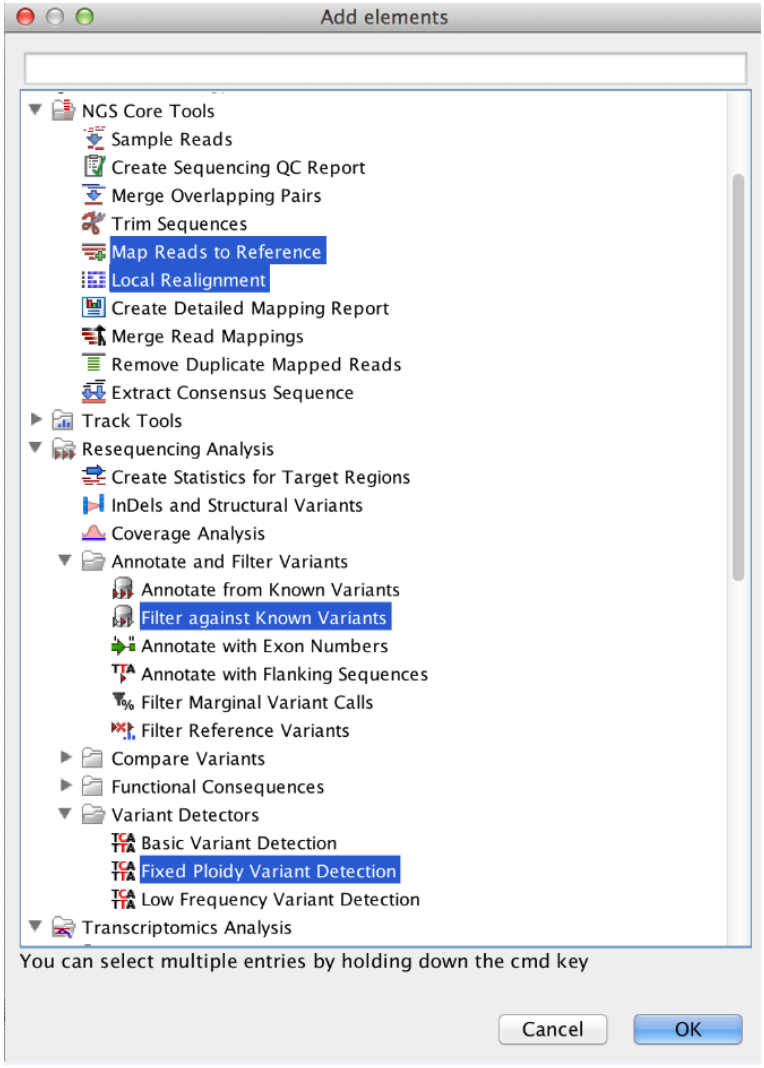

Figure 2: *Selected tools are displayed in the View Area.*

- The top boxes indicate the possible input data types. Pale boxes indicate that an input from a workflow element is required, while beige boxes indicate inputs that will need to be configured. The configuration can be done directly in the View Area, but the tool can also be left unconfigured in the workflow layout. In this case, the user of the workflow will be prompted to supply the configuration through a wizard dialog when launching the workflow.
- The middle box gives the name of the tool. To the right of the name is a small icon that looks like a piece of paper. Clicking on this icon brings up the wizard the user will see for this stage of the workflow. As with all wizards, there is a small question mark icon in the bottom left hand side. Clicking on this brings up the manual information for the tool.
- The lower boxes indicate the different types of output that can be created with the tool.

Connecting the tools We need to connect the tools in the order we wish them to run. We do this by indicating the inputs and outputs for each tool, and linking tools that take outputs from another tool as their input.

Here, we join the tools as follows:

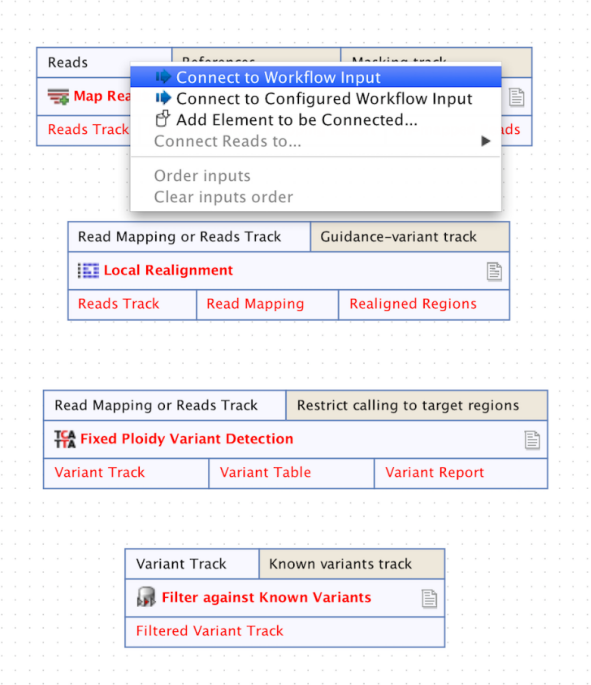

Figure 3: *Add a workflow Input box.*

Map Reads to Reference - **Local Re-alignment** - **Fixed Ploidy Variant Detection** -. **Filter against Known Variants**

1. Right-click the box labeled **Reads** at the top of the Map Reads to Reference tool. Click the option **Connect to Workflow Input (** **)** (figure 3). This adds a box labeled **(** **) Workflow Input** to the workflow. The connection between this box and the **Reads** box is indicated by a thin arrow. To move the Workflow input box around in the workflow editor, move the mouse cursor to the thin bar at the top of the box, click and drag.

2. Next, we will configure output from the **Map Reads to Reference** tool as input to the **Local Re-alignment** tool, where the latter takes as input a read mapping or reads track. We will configure it to take in a reads track. Click and hold the mouse while dragging from the **Reads track** box of the Map Reads to Reference tool to the **Read Mapping or Reads Track** box of the Local Re-alignment tool. This will create an arrow indicating that these two tools are connected.

3. Now do the same to connect the **Local Re-alignment** tool to the **Fixed Ploidy Variant Detection** tools. Click and hold the mouse while dragging from the **Reads track** box of the Local Re-alignment tool to the **Read Mapping or Reads Track** box of the Fixed Ploidy Variant Detection tool. The arrow between these two boxes indicate these two tools are connected.

4. Similarly, we will connect the **Fixed Ploidy Variant Detection** and the **Filter against Known Variants** tools. Click and hold the mouse while dragging from the **Variant Track** box of the Fixed Ploidy Variant Detection tool to the **Variant Track** box of the Filter against Known Variants tool. The arrow between these two boxes indicate these two tools are connected.

5. Finally we need to set the result from the **Filter against Known Variants** analysis as workflow output. Right-click the box labeled **Filtered Variant Track**. Click the option **Use as**

Workflow Output ( ). A default output name is provided, but you can easily change it: double click on the output just added. You can change the output name to a static name, which will always be used for this output. Alternatively, you can use the variables provided to get names related to the inputs of the workflow. For this tutorial, click in the Custom output name box and click on the Shift and F1 buttons. Select the second option, so that the output will be named after the name of the workflow, and click on the button labeled **Finish**.

Now that we have connected the individual tools and added input and output, the workflow should look like in figure 4.

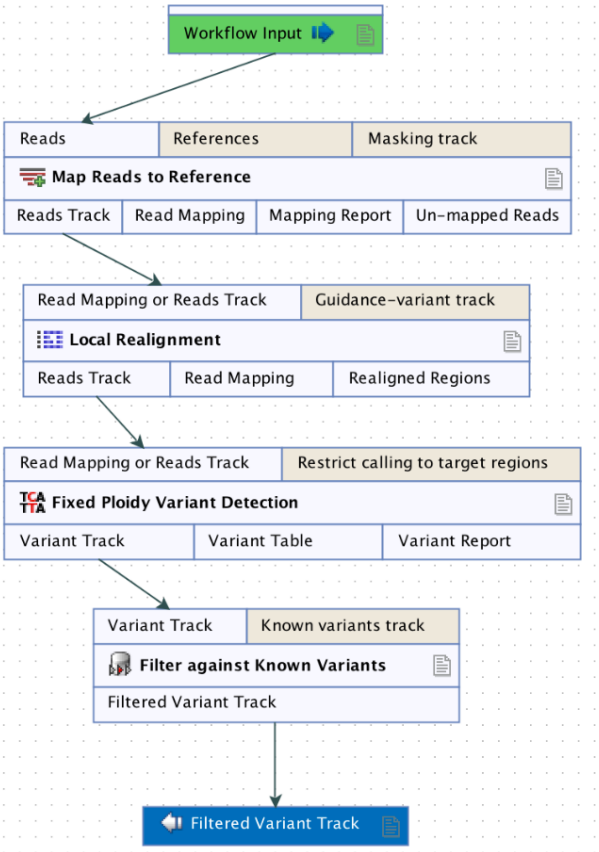

Figure 4: *The workflow tools are now connected via small arrows.*

Configuring additional outputs So far only one workflow output has been configured. This means that only one data object will be saved from this workflow when it is run - in this case, the output of the **Filter against Known Variants** tool.

In most cases, other outputs should be saved, such as the mappings, reports, and so on. We do this by configuring more workflow outputs.

For each output type (**Reads Track** from the mapping phase, **Reads Track** from the Local Re-alignment phase, **Mapping Report** from the mapping phase, and the **Variant Track** from the Fixed Ploidy Variant Detection phase):

1. Right-click on the relevant box.

2. Click the option **Use as Workflow Output ( )**.
3. Set up a name or variable to use for the naming of the output, if you do not like the default naming for the output.
4. You can move workflow elements around in the editor by clicking on them and dragging them to where you want them to be. You can also allow the workbench to set up an orderly layout for you by right clicking anywhere in the workfow editor area and selecting the option **Layout** from the menu that appears.

Now the workflow should look like in figure 5.

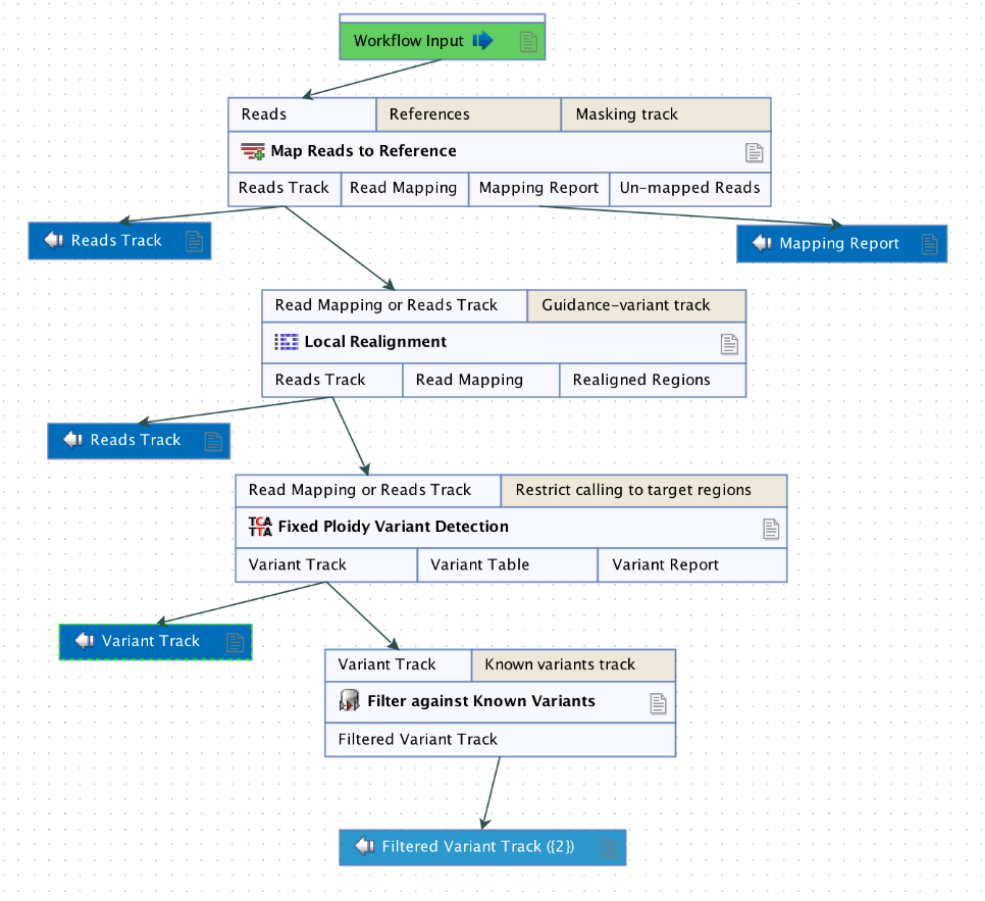

Figure 5: *The finished workflow.*

The workflow can be saved and tested at this point, if desired. If this is done, the user will be prompted for all the data inputs that have not yet been provided, while many parameters will be locked by default.

Configuring the tools

In this section we set the parameters for the individual tools as well as set up the reference sequences.

Configuring the Map Reads to Reference tool For this tool we need to choose the reference to map the reads against.

1. Double click on the box labeled **References**.
2. In the wizard click the **Browse button ( )** to the right of the References box in the wizard and select **NC_001807(Genome)**, adding it to the right side of the selection window by clicking on the arrow pointing to the right. Then click on the button labeled **OK**.
3. Click on the button labeled **Next**.
4. Set the mapping parameters to match those in figure 6. Notice the small lock symbol on the left hand side. Clicking on the icon locks or unlocks the parameter. Unlocked parameters will be presented in the workflow wizard and can be changed by the user when the workflow is launched.

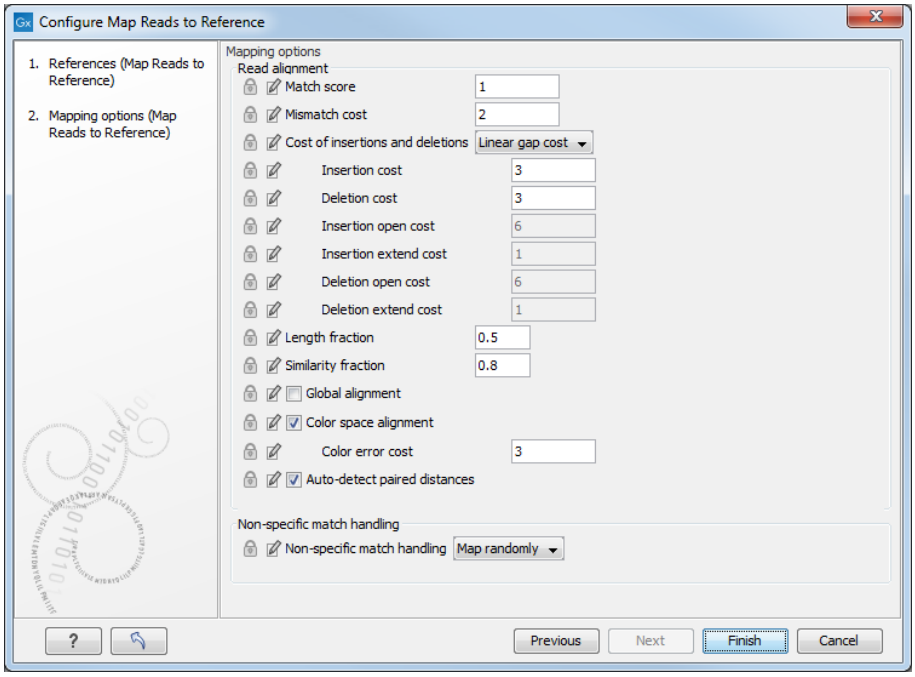

Figure 6: *Setting the mapping parameters.*

5. Click on the button labeled **Finish**.

Configuring the Local Re-alignment and Fixed Ploidy Variant Detection tools We will leave the parameters for these tools as the defaults. If you wish to have a look at the individual parameters please follow the below steps. If not, just skip this subsection.

1. Double click on the box labeled with the name of the tool in the workflow editor.
2. View the settings.
3. Click on the button labeled **Finish** when you're done.

Configuring the Filter against Known Variants tool For this tool we will choose the database track to filter against and the action to perform.

1. Double click the box labeled **Filter against Known Variants**

2. In the wizard click the **Browse button (**) and select **chrMdbSNPCommon** adding it to the right handside and click **OK**.
3. For **Filter options** click on the option **Keep variants with no exact match found in the track of known variants**.
4. Click on the button labeled **Finish**.

Performing a test run

1. Save the workflow:

   File | **Save as (**)

   There are a number of ways to save things open in the workbench, including workflows:

   (a) 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

   (b) Right click on the tab at the top of the unsaved view, and choose **Save...** or **Save As...** from the menu that appears, or

   (c) 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 area and the button labeled **Run** becomes activated.
2. Name the workflow **chrM workflow**.
3. Click on the button labeled **OK**.

   The workflow now appears in the **Navigation Area**.
4. Click the button labeled **(**)**Run...** at the bottom of the workflow editor area.
5. In the wizard select the **normal tissue reads** from within the normalData folder, adding the file to the right side or the selection window, and click on the button labeled **OK**.
6. Click on the button labeled **Next**.
7. Click through the next couple of wizard windows by clicking on the button labeled **Next**.
8. At the Result Handling phase of the wizard, choose to **Save** the results and to open the log. Opening the log allows you to see the progress of the workflow.
9. Click on the button labeled **Next**.
10. Click the ( ) button to create a new folder. Name the new folder **Test run** and choose to save the results into this.
11. Click on the button labeled **Finish**.

Once the test run is done you will see 5 files in the **Test run** folder (figure 7).

If you wish you can open up the individual files to have a look at the results.

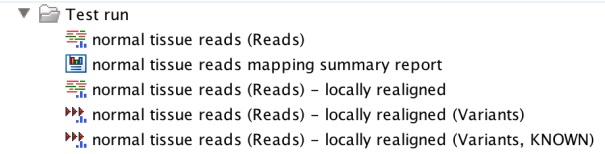

Figure 7: *The test run creates 4 files.*

What we have worked with so far is a workflow design. To use this workflow via the workbench menu system, you need to install the workflow. Once installed, you'll be able to launch the tool to run single jobs or to run jobs in batch mode.

To distribute the workflow to others, you will want to create an workflow installer and that installer can then be used to install the workflow in a CLC Workbench or a CLC Server. Installing the workflow on a CLC Server makes it available to all users of that Server.

Installing and managing workflows

Installing the workflow on your machine

1. Click the button labeled **Installation** at the bottom of the workflow editor.
2. Fill out the fields with your name etc. Note that you can make a workflow description for future reference.
3. Click the button labeled **Next** and work through the rest of the wizard windows.
4. At the Install Location step, choose the option to **Install the workflow on your local computer** and click on the button labeled **Finish**.

When this is done, go to the Workflows section of the Toolbox. You will find your workflow available to use from there.

Managing workflows You can see information about the workflows you have installed by launching the Workflows Manager. If someone shares a workflow installer file with you, this is also the tool you would use to install that workflow.

1. To launch the Workflows manager, click on the Workflows button on the toolbar and choose **Manage Workflows**.

   The workflow is now listed on the left hand side. On the right hand side, pressing the preview tab gives you a graphical overview of the workflow (figure 8).
2. Click the button labeled **Close**.
3. Now have a look at **Toolboox** | **Workflows (** **)** - this now features the **chrM workflow (** **)**.

Running the installed workflow Now we will start the **chrM workflow** from the Toolbox and run the two datasets, which we have imported, in batch mode. From this point, with a few clicks by the mouse you are able to run a total of 8 analysis, with 4 analysis steps for each dataset.

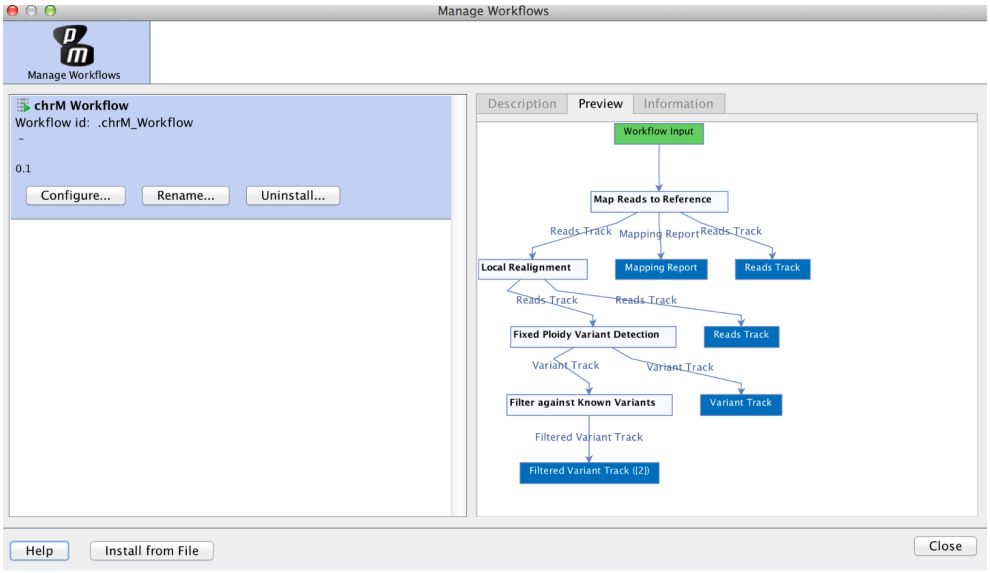

Figure 8: *Upon installation the workflow appears in the Workflow manager. The right handside tabs give the workflow description and preview.*

1. To start the workflow go to:

   Toolbox | **Workflows (**) | **chrM workflow (**)

2. Select the **Batch** option in the bottom left of the wizard.

3. Add the folder called chrM-tutorial-data to the right hands panel.

   The folders under this will be looked into for appropriate data objects. In this case, sets of reads. Here, the two folders, cancerData and normalData each contain one read set. The reason to set things up this way is because when running analyses via the batch functionality, the results are written to the same folder as the input data. Having the results organized within different folders can help when finding and working on the outputs later.

4. Click on the button labeled **Next** on each of the wizard windows until you get to the Results handling step.

5. Choose to Save your data. and to open the log.

6. Click on the button labeled **Finish**.

As the job is running, you can watch its progress by opening the tab labeled Processes at the bottom left side of the workbench, as well as by opening the log file for a more detailed view. After the jobs are finished, you should see the results within the cancerData and normalData folders.

You can also configure a Workflow to create a track list for easy comparison and further downstream processing. Here, it is not entirely appropriate as it is likely we would wish to create a track list that includes the reference data and both of the sample results. However, if we did wish to set up a workflow that output a track list for each sample processed, one could set up a Workflow like the one shown in figure 9. Note that a condition for tracks to be included in a tracklist within a workflow is that they must be configured as workflow outputs.

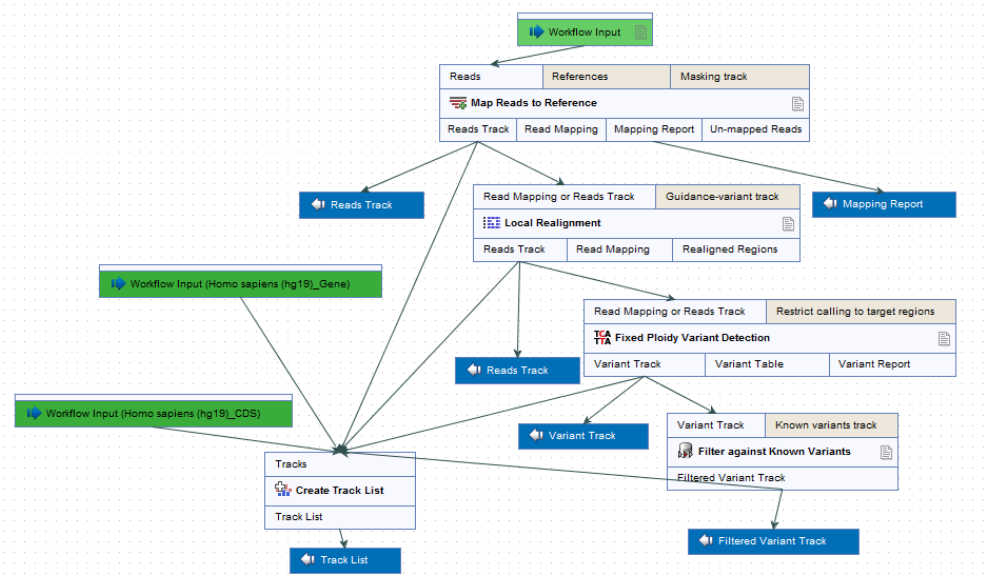

Figure 9: *Here, a track list is created from the outputs of the Workflow as well as reference data.*