Tutorial

Modification of an Existing Workflow

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Modification of an Existing Workflow

Ready-to-use workflows are provided in *Biomedical Genomics Workbench* for different applications and scenarios. However, there may be situations where you would like to extend an analysis or customize it to special needs. In these situations it can be relevant to modify an existing workflow, or to customize a reference data set.

In this tutorial we will first modify the **Identify Somatic Variants from Tumor Normal Pair (WES)** ready-to-use workflow to include the step in which we remove variants present in HapMap. Ready-to-use analysis workflows in the *Biomedical Genomics Workbench* are bundled together with a Reference Data Set that can be downloaded using the **Data Management** function. These Reference Data Sets can also be customized to fit better the data that will be analyzed. We will see how to customize a Reference Data Set in the second part of this tutorial.

Modifying a workflow

1. To modify a workflow we are first going to open the workflow in the **View Area**. This can be done by clicking on the button labeled **Workflows** in the upper right corner of the workbench. Select **New Workflow** from the drop-down list. This will create a new workflow editor.

2. Go to:

   Toolbox | **Ready-to-Use Workflows** | **Whole Exome Sequencing** | **Somatic Cancer**

   Click on and while holding down the mouse key, drag and drop the **Identify Somatic Variants from Tumor Normal Pair (WES)** ready-to-use workflow into the open workflow editor (see figure 1).

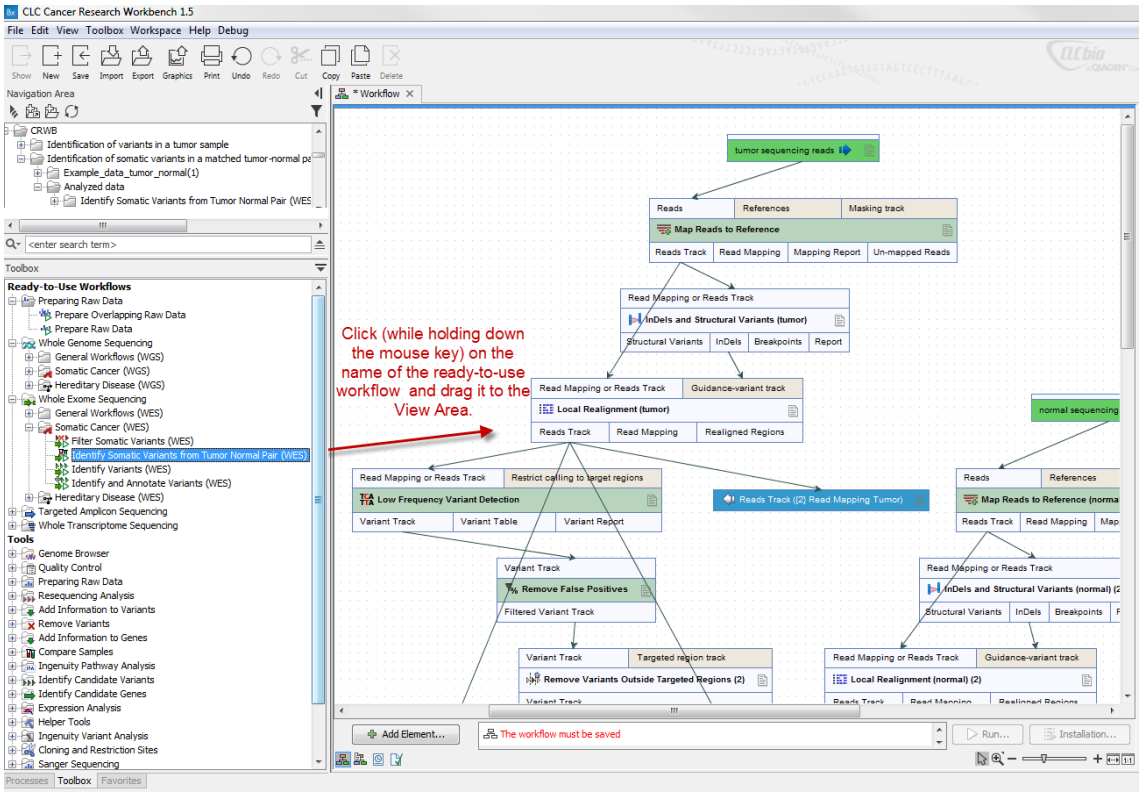

Figure 1: *Drag and drop the workflow in the open workflow editor.*

3. Right-click in the background area in the workflow editor and choose **Layout**. This will rearrange the tools and the arrows connecting them, which will help you get a better overview of the workflow.

4. Go to the middle of the workflow in the workflow editor and find the tool **Remove Reference Variants**.

5. Remove the arrow between the tool **Remove Reference Variants** and the tool **Add Information about Amino Acid Changes**. You can then delete the tool Remove Reference Variants. You can delete arrows and tools by clicking on it and pressing *Delete* (see figure 2).

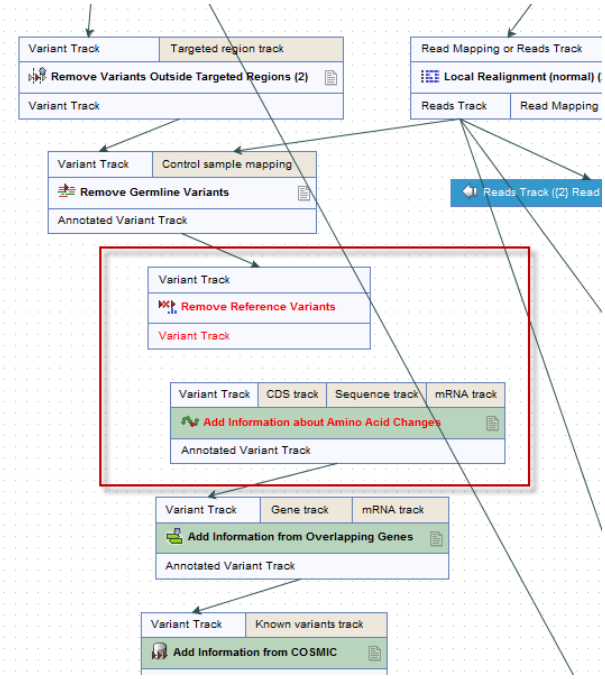

Figure 2: *Remove the arrow between Remove Reference Variants and the Add Information about Amino Acid Changes.*

6. Next we would like to add the "Remove Variants found in HapMap" step. Once again go to the toolbox and locate the tool:

   Toolbox | **Tools** | **Remove Variants** | **From Databases** | **Remove Variants found in HapMap**

7. Click on and while holding down the mouse key, drag and drop the tool **Remove Variants found in HapMap** into the workflow editor.

8. Remove the input (green box: Workflow Input)) and the output (blue box: Filtered Variant Track) from **Remove Variants found in HapMap** by clicking once on the green and the blue box, respectively, and then pressing *Delete* (see figure 3).

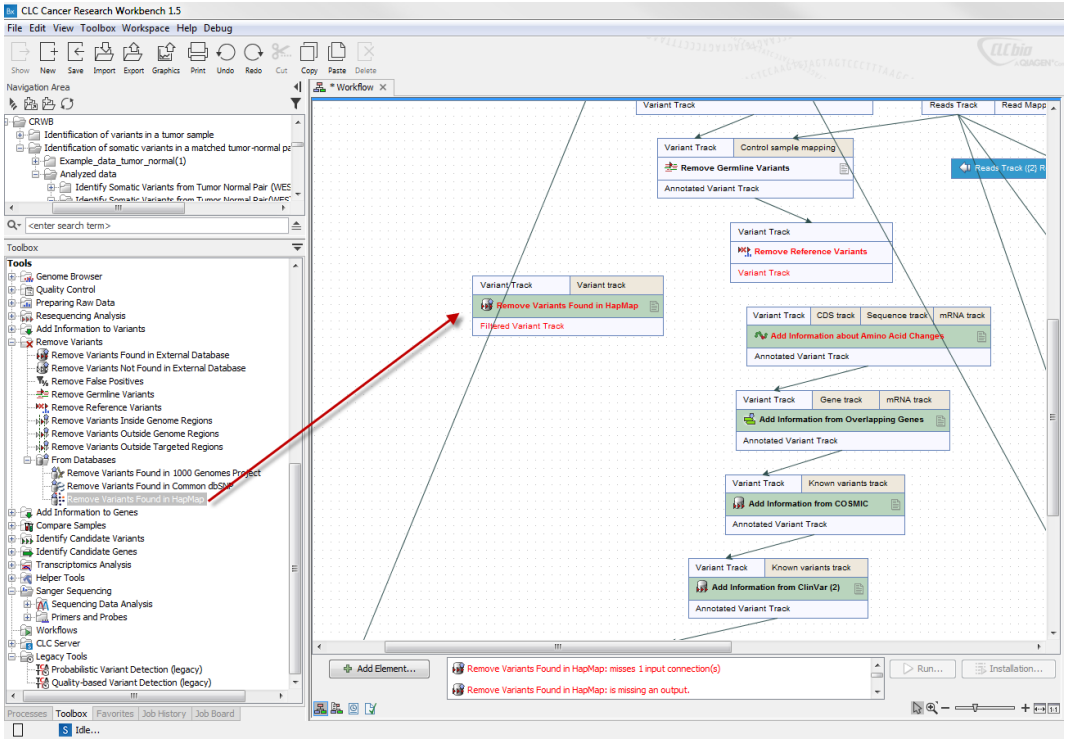

Figure 3: *Drag and drop the tool Remove Variants found in HapMap to the workflow editor.*

9. Connect the output from **Remove Reference Variants** to the input of **Remove Variants found in HapMap** and the output of **Remove Variants found in HapMap** to the input of **Add Information about Amino Acid Changes** (figure 4).

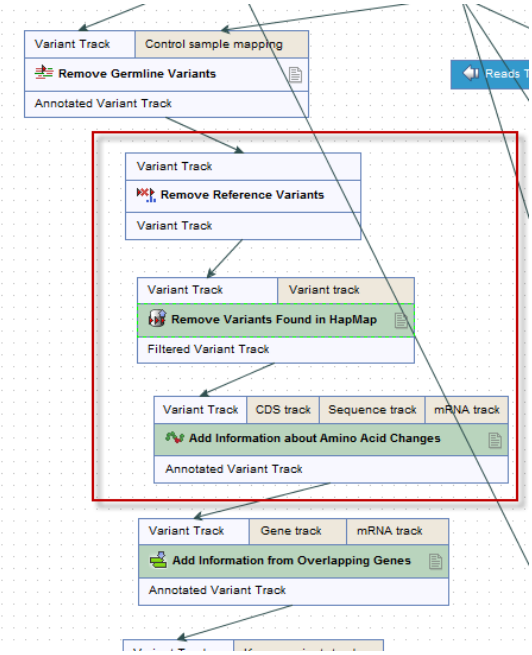

Figure 4: *Combine the inputs and outputs of the tools Remove Reference Variants, Remove Variants found in HapMap and Add Information about Amino Acid Changes.*

This is done by clicking once on the output box (the lower box: Variant Track) from the "Remove References Variants" tool. While holding down the mouse button on this box,

you can drag an arrow from this box to the input box (upper box: Variant Track) in the tool "Remove Variants found in HapMap". When you let go of the mouse button, the two tools will be connected with an arrow.

10. Save the workflow by using **Save** in the top toolbar. The default name is *Workflow*. It can be a good idea to find a more specific name for the workflow. In this case we will call the customized workflow *Identification of Somatic Variants from Tumor Normal Pair (WES) incl remove HapMap variants* (figure 5).

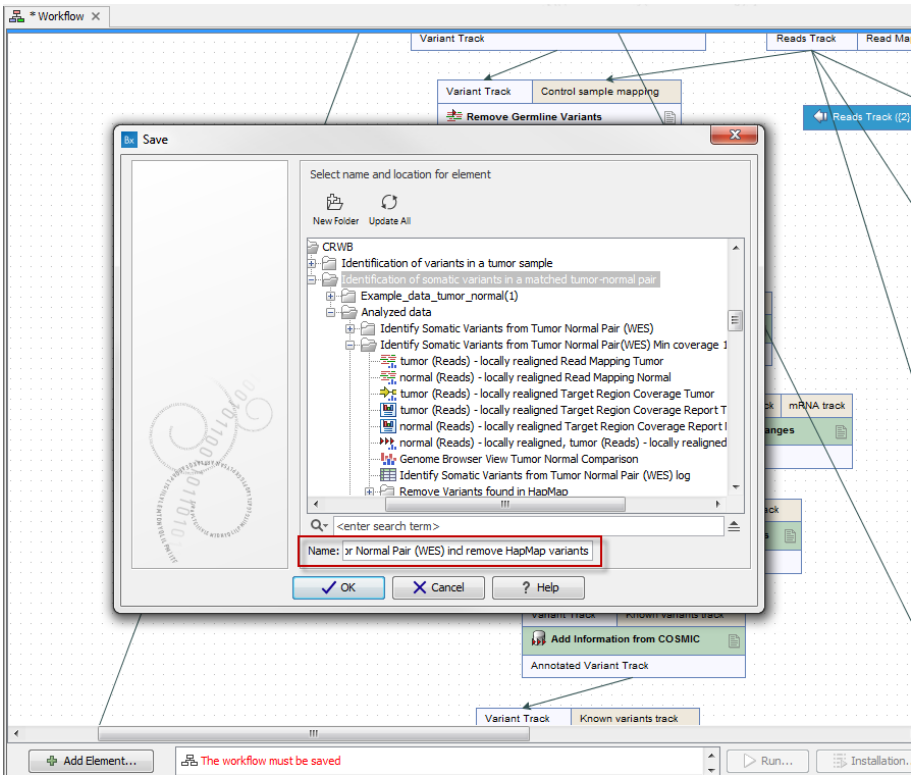

Figure 5: *Rename and save the workflow.*

11. The saved workflow can now be found in the **Navigation Area** as shown in figure 6.

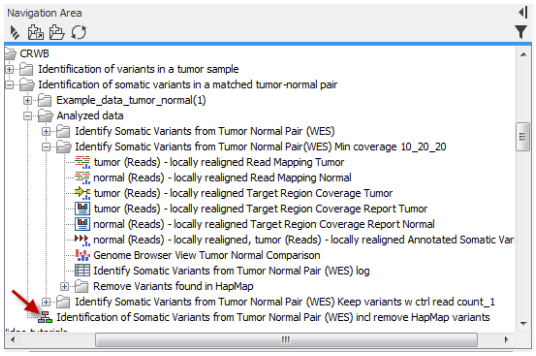

Figure 6: *The saved workflow is found in the Navigation Area.*

12. After the workflow has been saved, click on **Installation** in the lower part of the workflow editor to install the workflow in your workbench (see figure 7).

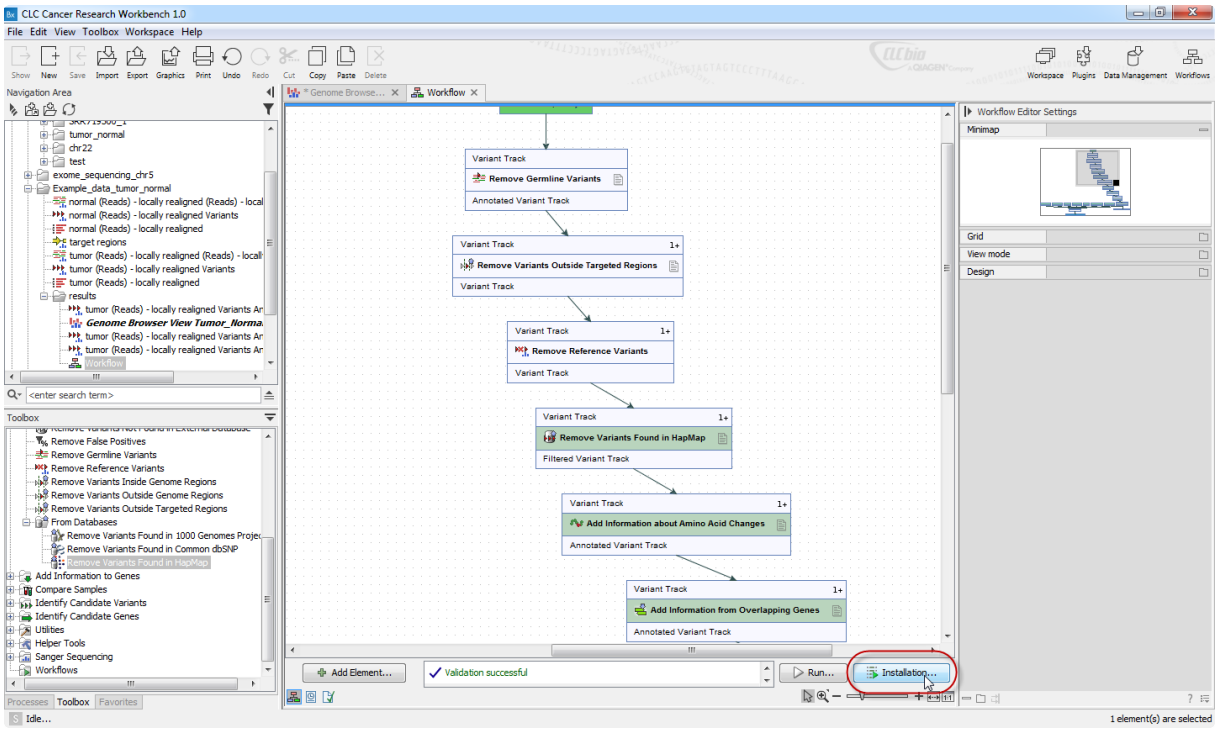

Figure 7: *Click on Installation to install the workflow or share it with a colleague.*

13. Fill in your affiliation details as in figure 8: name, email, institution etc. Under "Workflow description" it is a good idea to write a detailed workflow description. Click on the button labeled **Next**.

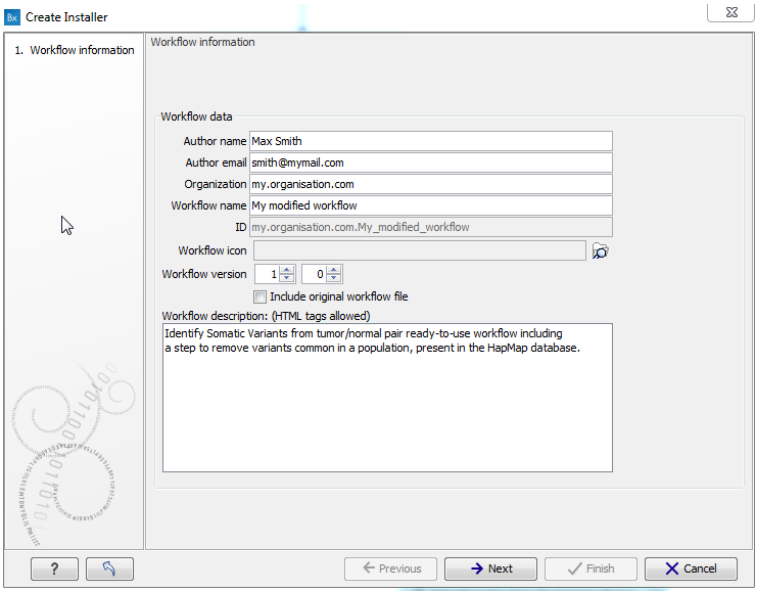

Figure 8: *Fill in your email adress and give the workflow a name.*

14. In the Reference Data window, you can choose between "Ignore", "Reference" or "Bundle" the references. When selecting **Bundle** the reference data are packed into the installer and the workflow will be configured to use exactly these data. This is in contrast to **Ignore** where no specific reference data are packed into the installation file of the workflow. Instead, the workflow will use the reference data that you have downloaded and specified under "Data

Management".

15. Choose to **Install the workflow on your own computer**.
16. Click on the button labeled **Finish**.

You can now find your new workflow in the toolbox in the "Workflows" folder:

Toolbox | **Tools** | **Workflows** .

You have now created a workflow that includes the two analyses "Identify Somatic Variants from Tumor Normal Pair (WES)" and "Remove Variants found in HapMap". We are not going to test the workflow in this tutorial, but you can try it out yourself. You should get the exact same results as when running the two workflows separately.

Preparing a custom Reference Data Set

A workflow is always bundled with a particular Reference Data Set, usually hg19 or hg38. The reference data just have to be download once when the Biomedical Genomics Workbench is started for the first time. Whenever a new version of the database is available, the user automatically gets notified. The Data Management function also allows the user to switch between different versions whenever more than one version is available.

For tutorials we use smaller data set, and we analyze them using chromosome-specific Tutorials References Data Set that are faster to download. Just remember to change back the set of references applied once you are done with this tutorial by following the instructions given at the end of this tutorial.

To download a Tutorial Reference Data Set (figure 9):

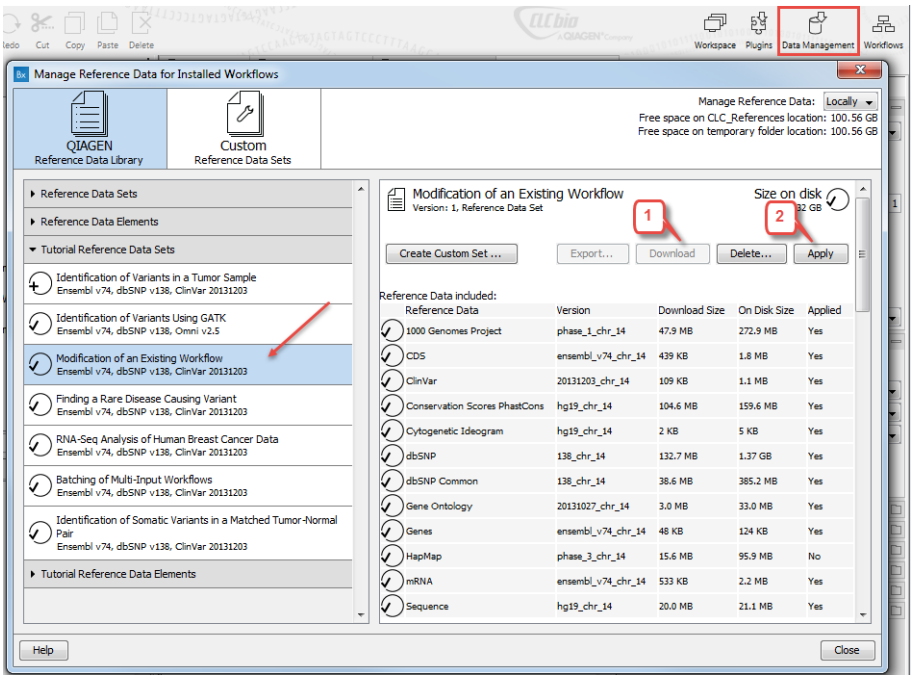

Figure 9: *Click on the "Data management" button and download the relevant reference databases.*

1. Click on the button labeled **Data Management** ( ) in the top right corner of the Workbench.

2. Open the tab "Tutorial References Data Set" and select the Reference Data Set "**Modification of an Existing Workflow**"
3. Click on the button labeled **Create Custom Set**.
4. A pop up window opens (figure 10). Edit the "Name" of the new Custom Reference Data Set to "Custom Hapmap chr14 Europe" for example. You can also edit the "Chromosomal Extension" field to mention chromosome 14.
5. Choose the option "Custom" from the drop down menu for the item "Hapmap".

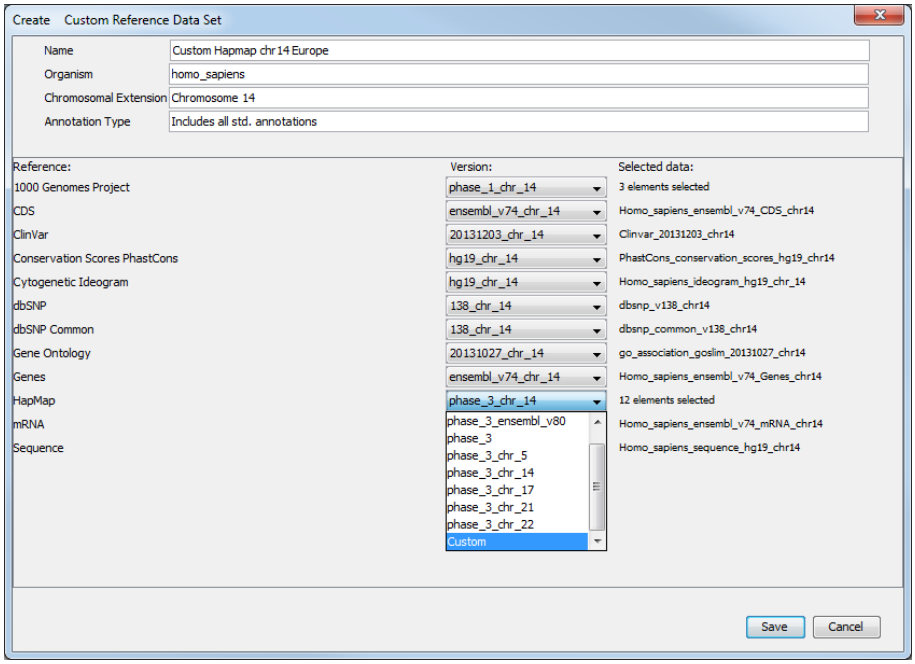

Figure 10: *Edit the Custom Data Set name and select the option "Custom" for each item you want to edit.*

6. When you select the option "Custom", a second pop up window opens. Look for the different Hapmap populations in the CLC_References folder that can be found in the Navigation Area (see figure 11), and double-click on the element(s) to select. We choose here to work with the European population HAPMAP_phase_3_CEU_chr14. Click on the button labeled **OK**.
7. The population(s) chosen is now visible in the first pop up window, and you can click on the button **Save** (figure 12).
8. The new data set will be saved in the Data Management window, under the "Custom Reference Data Sets" tab. Click on the button **Download**. You can check progress of the download in the 'Processes' tile of the toolbox in the lower left corner of the workbench. Once all elements are downloaded ( ), you can click on the button labeled **Apply** (figure 13).
9. Close the Data Management window.

Note that when applying a Custome Reference Data Set where the Hapmap population or the 1000 Genomes Project population have been customized, the workflow wizard window that usually allows you to define the Hapmap/1000 Genomes populations will not appear anymore.

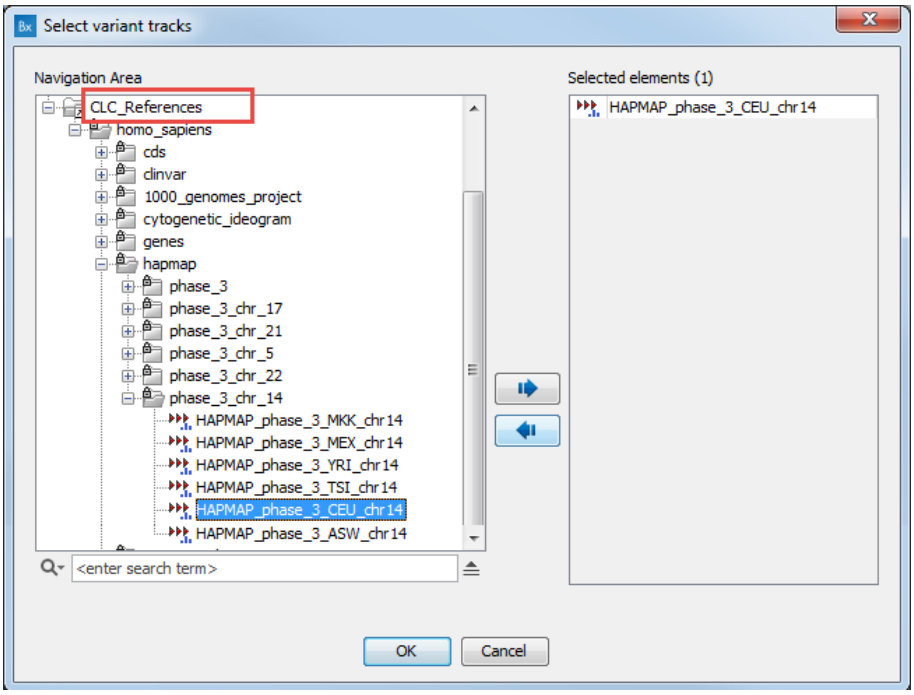

Figure 11: *Choose the relevant sequence for each item from your Navigation Area.*

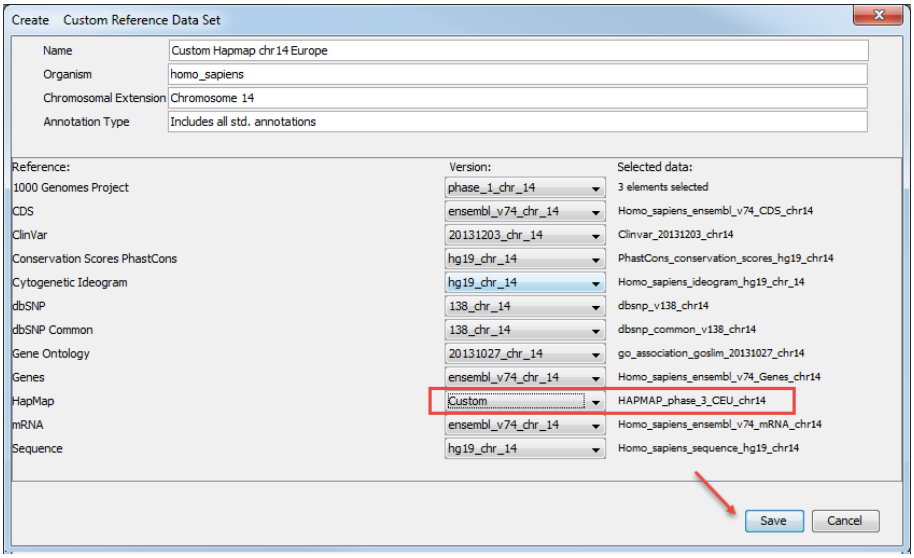

Figure 12: *Save the Custom Reference Data Set.*

Once applied, this Custom Reference Data Set is the one that will be used with the workflows you will run next, whether it is the one you just modified or one of the Ready-to-Use workflow from the toolbox. It will remain so until you decide to apply another Reference Data Set. It is important to always be aware of what Reference Data Set is applied when running a workflow, as it can influence the results dramatically.

To apply a less specific Reference Data Set:

1. Open the **Data Management** tool by clicking on the button with that name in the top right corner of the Workbench.
2. If you have a server, choose whether you want to apply the Reference dataset locally or on

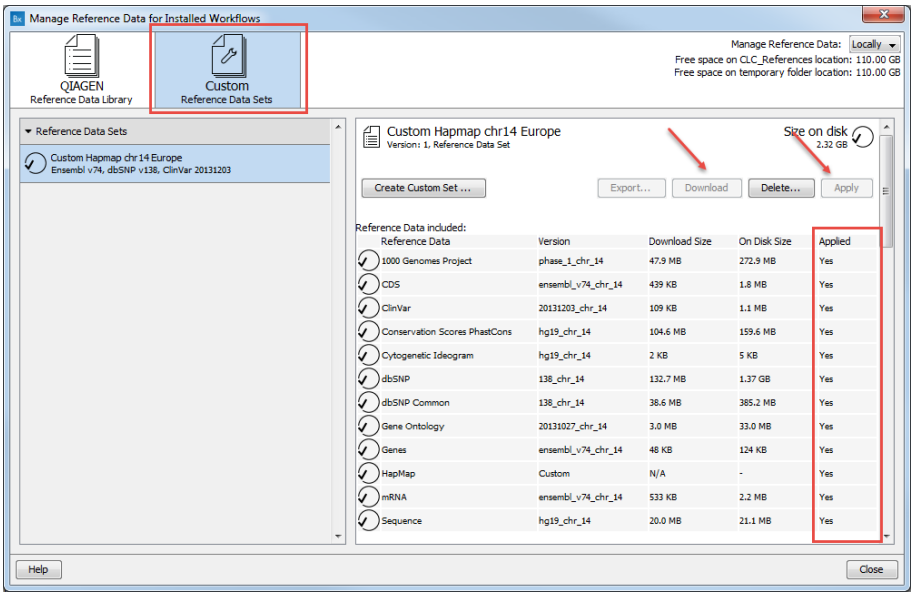

Figure 13: *Download all Data Set elements if it was not already downloaded before, and click on "Apply" once the download is finished.*

the server.

3. Select the **hg19** or **hg38** Reference Data Set.
4. Download the Reference Data Set if you had not done so previously.
5. Click on the button labeled **Apply**.
6. **Close** the Data Manager.