Tutorial

Identification of somatic variants in a matched tumor-normal pair

March 2, 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

Tutorial

Identification of somatic variants in a matched tumor-normal pair

This tutorial will guide you through the process of identifying somatic variants from a matched tumor/normal sample pair from one patient.

We will use paired-end exome sequencing data from a massive acinic cell carcinoma sample. The sample was sequenced using the Illumina 2000 platform and published by A. C. Nichols et al. in Case reports in Oncological Medicine in 2013 [Nichols et al., 2013] (http://www.hindawi.com/crim/oncological.medicine/2013/270362/).

The example data used in this tutorial include only reads mapping to chromosome 22.

This tutorial includes images, workflows, and tools from Biomedical Genomics Workbench 3.0.

Prerequisites

- Machine meeting the system requirements of the Biomedical Genomics Workbench: http://www.qiagenbioinformatics.com/system-requirements/.
- Biomedical Genomics Workbench 3.0 or higher
- Zip file with example data

Overview The steps carried out in this tutorial include:

- Setting up the correct References
- Identification and annotation of candidate somatic variants
- Inspection of results
- Removal of variants common in a population
- Setting up/modifying an automatic analysis workflow

Introduction Identification of somatic variants from a matched tumor/normal pair is often a bottleneck in bioinformatics, with specific tools or self-made scripts being used for variant detection.

A common approach to identification of somatic variants from matched tumor/normal pairs is to remove all variants identified in the normal sample from the list of variants identified in the tumor sample. However, due to differences in the coverage of sequencing reads mapping to the human reference sequence in the normal sample and in the tumor sample, all germline variants are not always detected in the normal sample.

We will take a different approach in this tutorial. Assuming there are no tumor cells in the normal sample, we will remove all variants found in the tumor sample if they are present in a certain number of mapped sequencing reads from the normal sample. In this way we can achieve a very high sensitivity for removal of germline variants. Next, we will identify known mutations that are present in clinical databases.

Data import and configuration

First, we need to download and import the example data.

1. Download the sample data from our web site: http://resources.qiagenbioinformatics.com/testdata/Example_data_tumor_normal.zip.
2. Start the *Biomedical Genomics Workbench*.
3. Import the data by going to:

   File | **Import (**  **)** | **Standard Import (**  **)**
4. Choose the zip file called *Example_data_tumor_normal.zip*. Leave the Import type set to **Automatic import**.
5. Click on the button labeled **Next**.
6. Choose where to save the example data in the **Navigation Area** and click on the button labeled **Finish**.

The data set includes the following files:

normal reads
: Sequencing reads from the normal sample

target regions - chr22
: Target regions on chromosome 22. Please note that when you start doing your own targeted experiments with your own data, you can obtain the relevant target region tracks from the vendor of the amplicon or hybridization kit.

tumor reads
: Sequencing reads from the tumor sample

Data configuration Ready-to-use analysis workflows in the *Biomedical Genomics Workbench* are bundled together with Human reference data. The reference data just have to be downloaded once when the Biomedical Genomics Workbench is started for the first time. However, we will use here a Reference Data set that was created specifically for this tutorial.

1. Open the **Data Management** tool by clicking on the button with that name in the top right corner of the Workbench (figure 1).
2. Select the Tutorial Reference Data Set called **Identification of Somatic Variants in a Matched Tumor-Normal Pair** under the QIAGEN Reference Data Library tile, and click on the button labeled **Download**.
3. Wait until all downloads are finished and click on the button **Apply**. you can now close the Data Management window.

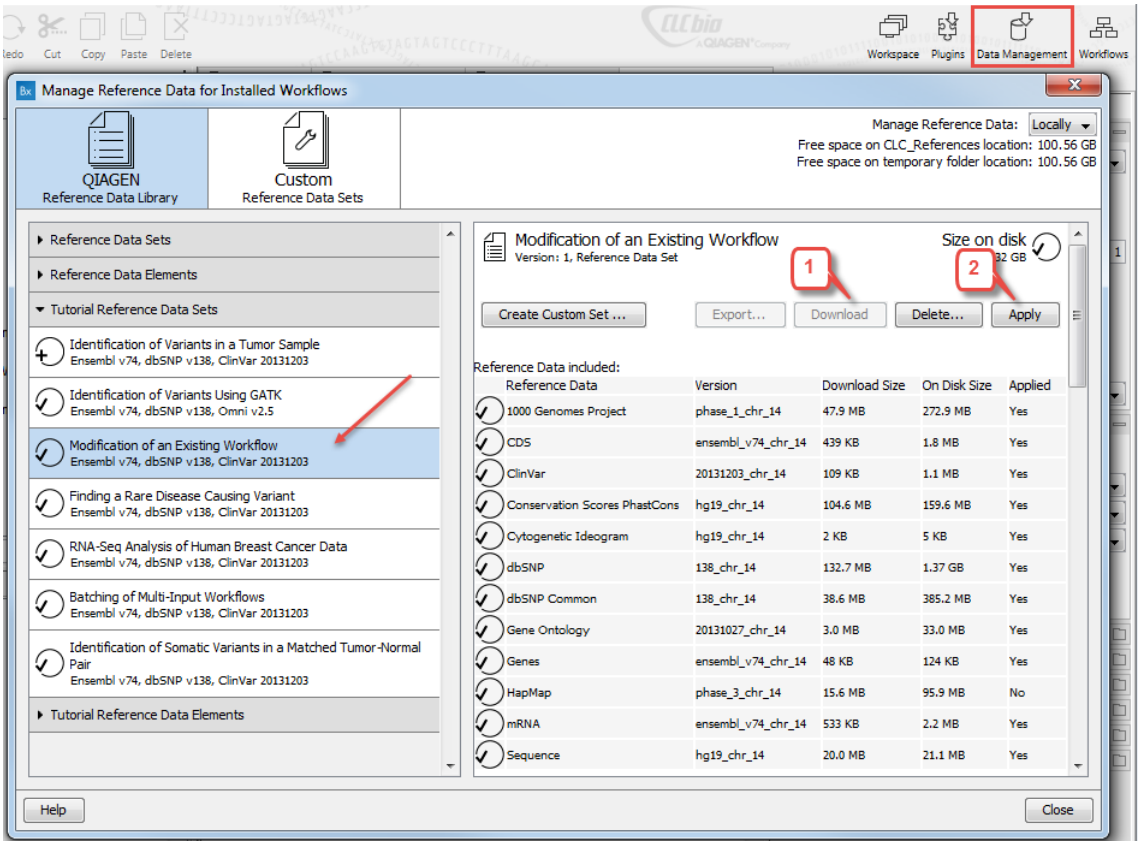

Figure 1: *Click on the "Data management" button and download the relevant reference databases.*

Identification and annotation of candidate somatic variants

The first thing we will do is to map the reads, identify and annotate the variants, and remove all germline variants from the list of variants found in the tumor sample by using the mapped sequencing reads from the normal sample. Next, we will add information from clinical databases to all remaining (potential somatic) variants.

All these steps can be done in one go with the **Identify Somatic Variants from Tumor Normal Pair (WES)** ready-to-use workflow. The output from the analysis is a Genome Browser View with the identified somatic variants shown together with the human reference sequence, the human genes, and clinical variants found in the public database ClinVar.

1. To run the **Identify Somatic Variants from Tumor Normal Pair (WES)** ready-to-use workflow:

   Toolbox | **Ready-to-Use Workflows** | **Whole Exome Sequencing** ( ) | **Somatic Cancer** ( ) | **Identify Somatic Variants from Tumor Normal Pair (WES)** ( )

   Depending on your local setup, you may be asked where you wish to run the job: in your Workbench, on a Server, or on a Grid. If you are presented with this window, you can choose the appropriate option for your work, and then click on the button labeled **Next**. In this example we will run the analysis locally in the Workbench.

2. Select the tumor reads (figure 2). Click on the button labeled **Next**.

3. In the next step, select the reads from the normal sample (figure 3). Click on **Next**.

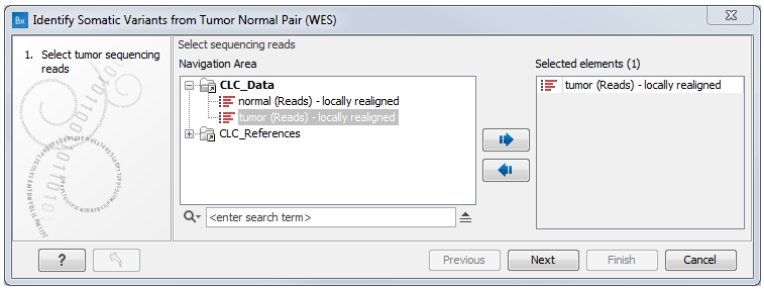

Figure 2: *Select the reads from the tumor sample.*

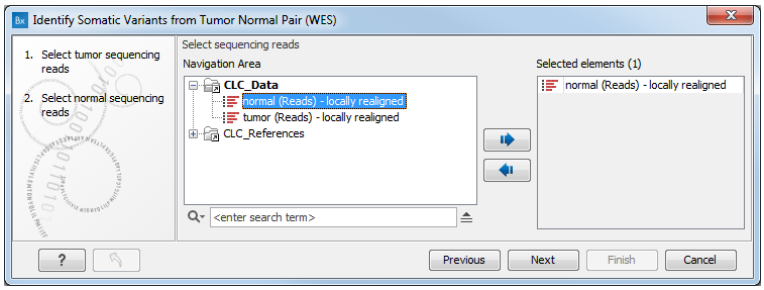

Figure 3: *Select the reads from the normal/control sample.*

4. Select the **target regions - chr22** file provided in this tutorial for the **Indels and Structural Variants** tool for the **tumor** sample. (figure 4). Click on the button labeled **Next**.

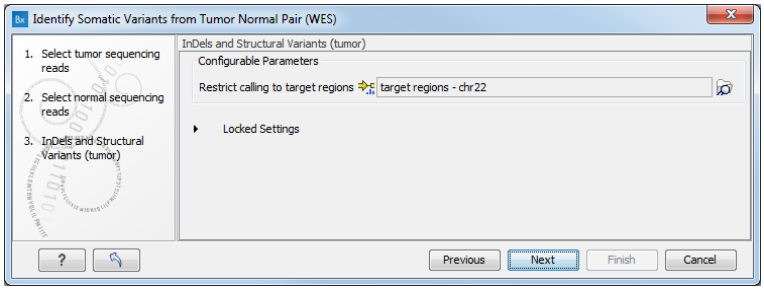

Figure 4: *Select the target file.*

5. Select the same **target regions - chr22** file for the **normal** sample.

6. Select again the **target regions - chr22** for the **Low Frequency Variant Detection** tool, and change the **Minimum frequency** to 15.0 (figure 5).

7. The next 2 wizard steps are a **Quality Check for Target Sequencing** of the tumor and the normal samples. Adjust the minimum coverage to 20 for both and select again the **target regions - chr22** track (figure 6). This means that in the target regions coverage report you will get statistics based on a minimum coverage of 20 e.g., statistics for the fractions of targets with a coverage of at least 20.

8. In the next wizard step you can leave the cutoff for **Removal of Germline Variants** (*Keep variants with control read count below*) to the default value of 2 (figure 7). The value of this parameter is the minimum number of mapped reads in the normal/control sample that support the variant found in the tumor. These variants are likely to be germline variants.

9. In the last wizard step you can review and export all parameters that are used in the ready-to-use workflow. Please note that some of the parameters are locked and cannot be

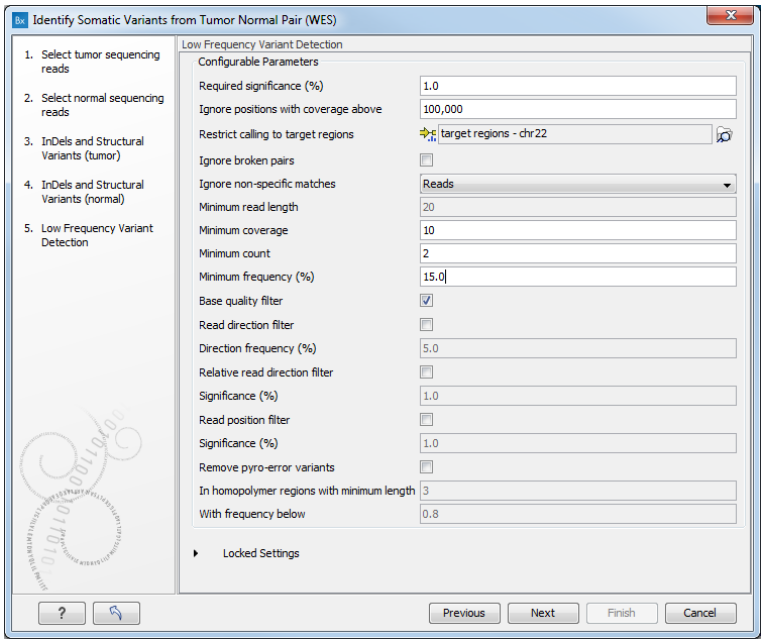

Figure 5: *Select the targeted regions track used to define the region from which the variants should be called.*

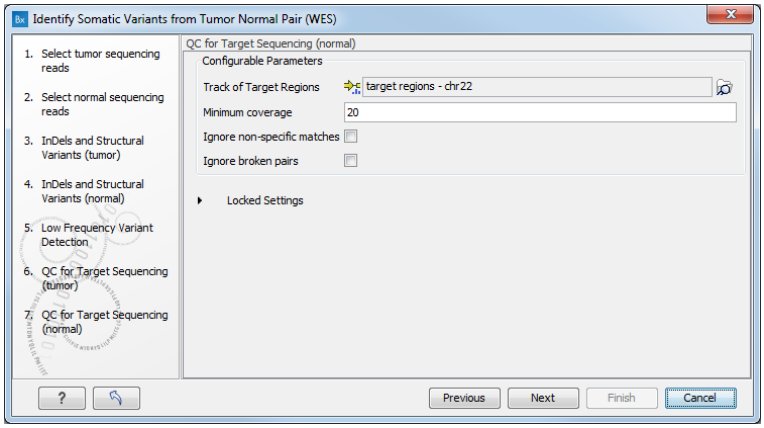

Figure 6: *Select the targeted regions track and adjust the settings for quality check of the targeted sequencing of the tumor sample.*

selected or modified by the user, but at this step you can view all used parameters using *Preview All Parameters*.

10. Choose to **Save** the outputs of the workflow. Click on the button labeled **Next** to specify the location where the outputs should be saved, and click on the button labeled **Finish**.

After you have started the job(s), you can follow the progress in the **Processes** tab that is found in the **Toolbox** in the lower-left corner of the Workbench (figure 8). The first step of the workflow, the mapping of the reads, is the most time consuming one, and it is possible that your workflow remains at 0% for a couple of minutes before showing constant progress.

The results will be placed in the location you have specified. In our example, the results are saved in a folder named *Identify Somatic Variants from Tumor Normal Pair (WES)*. When the workflow is finished, you will see the results shown in figure 9.

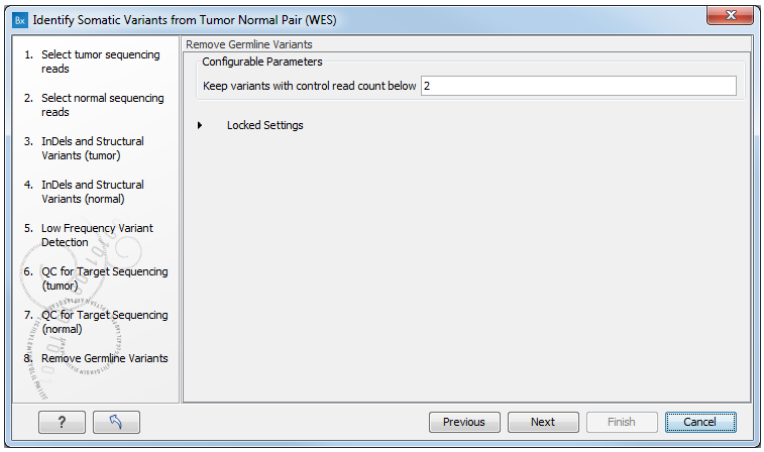

Figure 7: *In this wizard step you can adjust the cutoff for removal of germline variants.*

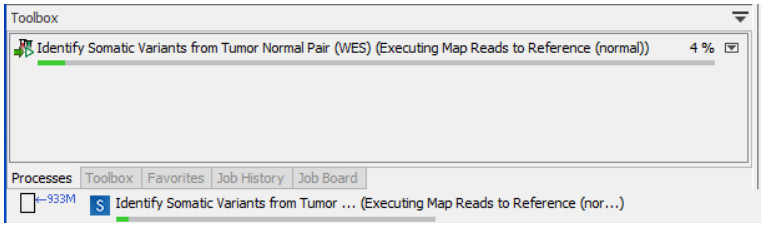

Figure 8: *Follow the progress of the analysis in the Processes tab in the toolbox.*

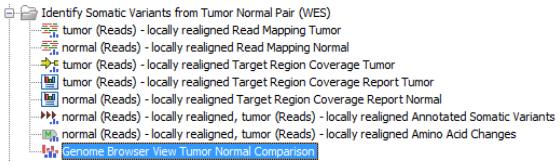

Figure 9: *The output from the workflow. Please note that in this example we have created a folder with the name "Identify Somatic Variants from Tumor Normal Pair (WES)".*

Inspection of potential somatic variants

Double click on the file *Genome Browser View Tumor Normal Comparison* in the **Navigation Area** generated with the ready-to-use workflow. The Genome Browser View opens in a split view as track list and variant table. The Table lists all found variants and related information. You can choose which information you want to display from the menu on the right hand side. The Table and the Genome Browser View are linked together so when you click on an entry (a variant) in the table, this position will automatically be brought into focus in the Genome Browser View. You can also use the zoom function in the Genome Browser View (see figure 10) to zoom in and out on specific variants.

Type *CYP2D6* into the filter field of the Table (highligted in red in figure 10)

- If you look in the table column "Exact match" you can see that one of the two somatic variants detected in the gene CYP2D6 is present in the database ClinVar.
- You can further see the phenotype associated with this variant when hovering on the cell in the table column "CLNDBN ClinVar_20130930". This variant is associated with poor metabolism of Debrisoquine.
- Take a look at the columns *Control Count* and *Control Coverage*. The *Control Count* is the number of mapped sequencing reads from the normal sample that support the identified

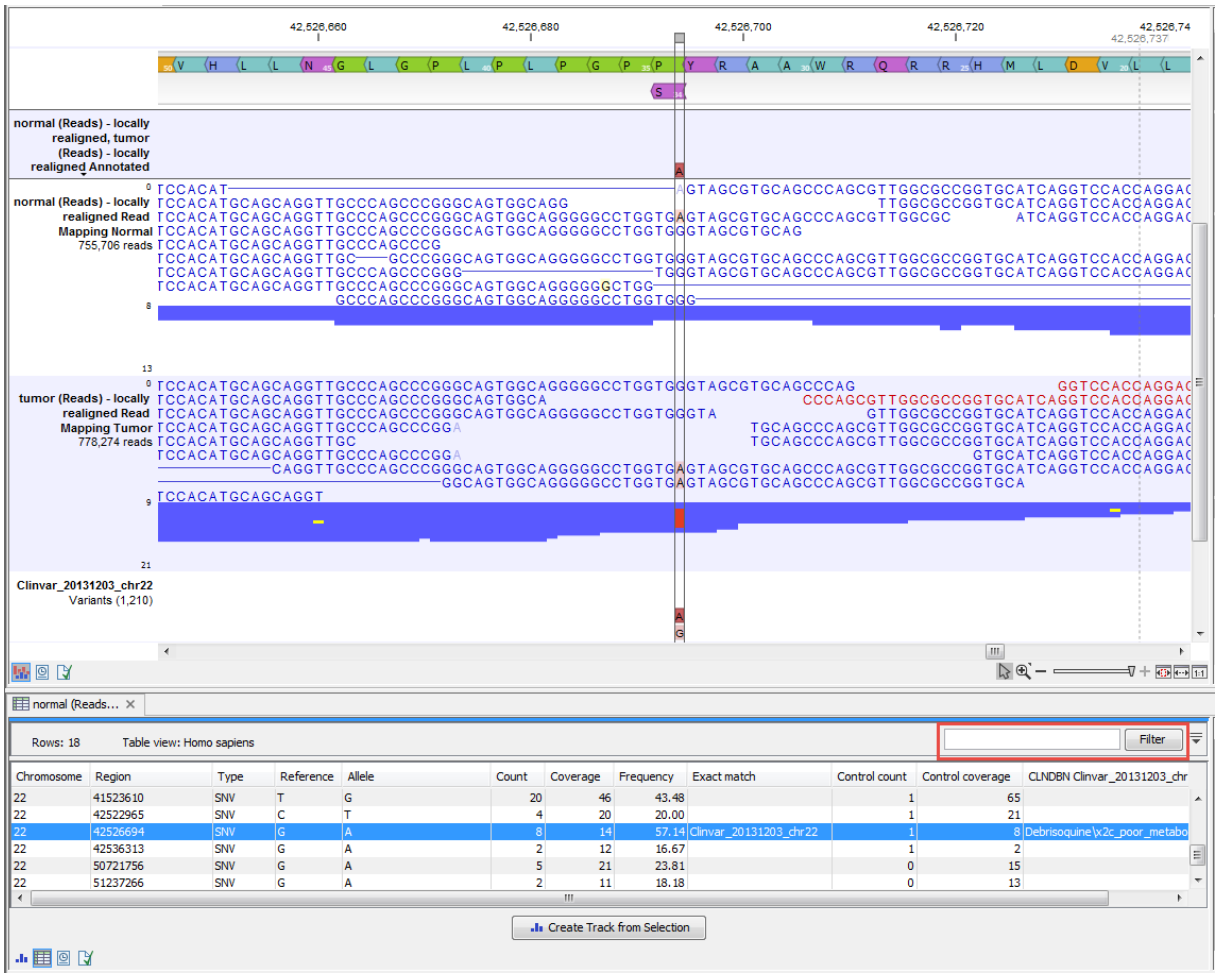

Figure 10: *When you click on the name of the track "normal (Reads) - locally realigned, tumor (Reads) - locally realigned Annotated Somatic Variants" (in the left side of the Genome Browser View) you can open this track in table view. Please note that This means that when you*

somatic variant. *Control Coverage* is the number of total sequencing reads in the read mapping of the normal sample in the position of the variant.

We will focus on the CYP2D6 variant that is found in the ClinVar database. For this variant the *Control Count* is in our example a number below 2. This is expected as all variants identified in 2 or more sequencing reads in the normal sample were deleted from the list of variants found in the tumor sample (see the parameter settings in the analysis workflow *Identify Somatic Variants from Tumor Normal Pair*).

The *Control Count* for the identified somatic variant in gene CYP2D6 is 1, which means that one read supports the variant in the normal sample. As the coverage at this position in the read mapping of the normal sample is very low with just 8 reads (see column *Control Coverage*), there is a high chance that this variant is actually germline and not somatic. In contrast to this you can see that the other variant found in CYP2D6 has a control coverage of 20, which makes it more plausible that this is a true somatic variant.

Removal of variants that are common in a population

Even when very stringent parameters are used to filter out variants present in the normal sample, there is still a high chance that some of the somatic variants in our list are actually germline. The read coverage in the read mapping of both tumor and normal is with 40x coverage what you would call low and it is likely that some of the alleles were not sequenced at all. In the next step we would like to exclude these germline variants from our list of somatic variants. The best way to do this would be to use an in-house database that includes common variants found in the population that the patient belongs to.

In this tutorial we will use the publicly available variant database HapMap that includes the variants that are common in the European/Caucasian population.

1. Open **Data Management** by pressing the button in the top right corner of the workbench.
2. Select the **Tutorial Reference Data Set** made for this tutorial and called Identification of Somatic Variants in a Matched Tumor-Normal Pair and click on 'Create Custom Set'.
3. This opens a new window in which you can select from a drop down menu the Hapmap version you wan to work with. Click on the Hapmap phase_3_chr_22 version in the drop down menu (see figure 11).

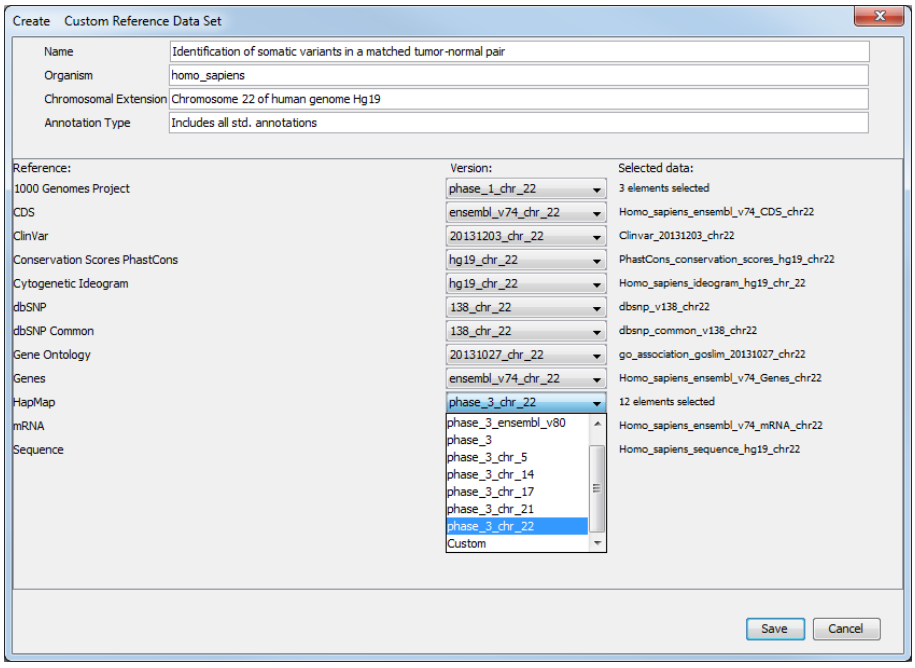

Figure 11: *Choose the file with the identified somatic variants.*

4. This opens another window in which you can select the specific population you are working with **HAPMAP_phase_3_CEU_chr_22** (see figure 12).
5. Clicking **OK** will close the 'Select Data for HapMap window and take you back to the 'Create Custom Reference Data Set' window. Edit the name of the new custom data set (for example to Identification of Somatic Variants in a Matched Tumor-Normal Pair - CEU) and click on the button **Save**.
6. You know have your new custom reference data set listed under the tile 'Custom Reference Data Set'. Click on the button labeled **Apply** and close the Data Management wizard.

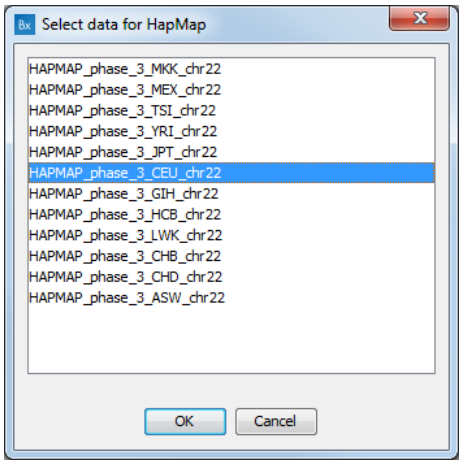

Figure 12: *Choose the file with the identified somatic variants.*

7. Run the tool **Remove Variants found in HapMap** to remove all variants found in the HapMap database by:

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

8. Select the result from the first analysis that you performed (**normal (Reads)- locally realigned, tumor (Reads) - locally realigned Annotated Somatic Variants** as input for the tool (see figure 13).

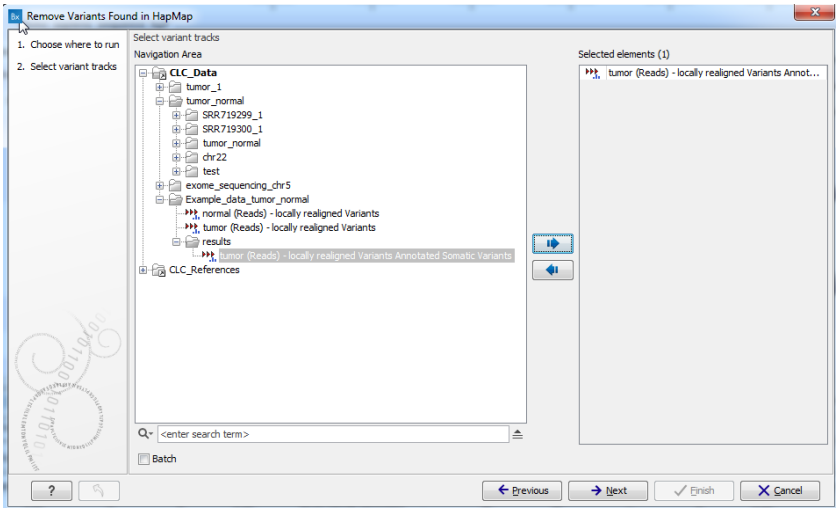

Figure 13: *Choose the file with the identified somatic variants.*

9. Choose to save the result. In our example we save the results in the folder called "results".

Open the newly generated file (it has the same name as the input file) and view it as table. If you compare the output from this analysis with the input data, you can see (in the upper left corner of the table) that in this case one variant was removed (from 18 variants to 17).

Additional tips

Adding and removing tracks in the Genome Browser View More tracks can easily be added to a track list by dragging and dropping track objects from the **Navigation Area** into the opened Genome Browser View track list.

Tracks can be removed from a track list by right clicking on the track you wish to remove. Then select the option **Remove Track** from the menu that pops up.

Saving changes If the name of a data object in the **Navigation Area** appears in bold, italicized text, it means your changes have not yet been saved.

There are two ways to save data objects that are open in a view:

1. Right click on the tab at the top of the unsaved view, and choose **Save As** from the menu that appears, or
2. Click on the tab at the top of the unsaved view and press **Ctrl-S** on the keyboard.

Once saved, the name of the data object should appear in standard font in the **Navigation Area**.

History - check what happened earlier All data within the Workbench has history information associated with it. The history includes information about how the data was created, which parameter settings were used, which version of the software was used and so on. You can view the history information for any data by opening it in the Viewing area of the Workbench and clicking on the History view button () at the bottom.

The history is very useful for double-checking the source data and parameters used in a given analysis that led to the generation of any particular data or results in the Workbench.

Change the reference data back to the default settings Before you leave this tutorial, remember to change the reference data set back to a whole genome version.

To do this:

1. Open the **Data Management** tool by clicking on the button with that name in the top right corner of the Workbench.
2. Select the **Hg19** or **Hg38** Reference Data Set.
3. Click on the button labeled **Apply** before clicking on the button labeled **Close**.

Bibliography

[Nichols et al., 2013] Nichols, A. C., Chan-Seng-Yue, M., Yoo, J., Agrawal, S. K., Starmans, M. H. W., Waggott, D., Harding, N. J., Dowthwaite, S. A., Palma, D. A., Fung, K., Wehrli, B., Macneil, S. D., Lambin, P., Winquist, E., Koropatnick, J., Mymryk, J. S., Boutros, P. C., and Barrett, J. W. (2013). A case report and genetic characterization of a massive acinic cell carcinoma of the parotid with delayed distant metastases. *Case Rep Oncol Med*, 2013:270362.